Trabajo de Fin de Grado

Desarrollo de un modelo estadístico de predicción de la velocidad del viento para el área del vertedero de Bailín

Autora

Carolina Pelegrín Cuartero

Directores

Jesús Asín Lafuente
Jesús Fernández Cascán

Ponente

Beatriz Lacruz Casaucau

Ciencias Ambientales
Escuela Politécnica Superior de Huesca
Junio de 2014

ÍNDICE

RESUMEN

Este Trabajo Fin de Grado se desarrolla en el marco del proceso de desmantelamiento del vertedero de residuos organoclorados de Bailín (Huesca) y su transferencia a un depósito seguro, llevado a cabo durante el verano de 2014. El fin último del trabajo ha sido crear una herramienta que permita anticipar posibles eventos dispersivos de la contaminación provocados por el viento, capaces de agravar la problemática ambiental existente en la zona.

Para ello, en primer lugar se ha realizado un análisis exploratorio del régimen de vientos en el vertedero de Bailín para el periodo de verano de 2012, así como un estudio que demuestra que el modelo de predicciones de AEMET existente no resulta útil de forma directa. Por ello, se han desarrollado modelos estadísticos de predicción de la velocidad del viento a 24 horas en forma de modelos de regresión (MR) que han permitido el *downscaling* del modelo de predicciones de AEMET, adaptándolo a las condiciones específicas del área y el periodo de estudio.

El proceso de construcción de MR se ha llevado a cabo en el entorno del software estadístico R mediante los siguientes pasos: 1) Estudio de parámetros potencialmente útiles (de predicción AEMET y otros), y construcción de una base de datos con ellos; 2) Desarrollo de una estrategia de construcción y selección de modelos (sin y con interacciones), haciendo uso de métodos *stepwise* de selección de parámetros, así como de criterios de selección de modelos como bondad de ajuste, desviación típica residual de validación cruzada o realización de test ANOVA; 3) Estudio de la adecuación de los modelos mediante el análisis de residuos; 4) Estudio del funcionamiento de los modelos candidatos finales; 5) Selección del modelo óptimo e implementación para su uso operativo.

El modelo finalmente seleccionado ha aumentado el porcentaje de explicación de la variabilidad de la velocidad del viento hasta un 57%, respecto al 14% obtenido con las predicciones de AEMET. La metodología creada también pretende servir de base para la resolución de problemas de predicción similares.

ABSTRACT

This study was developed within the framework of the Bailín landfill (Huesca, Spain) dismantlement and its waste transfer to a new and safe location. The aim of the work has been to create a tool that will allow to anticipate wind events, which could produce pollution dispersion, aggravating the environmental situation at the area.

First, an exploratory analysis of the wind regime at the area has been done. This first studies showed that the AEMET (*Agencia Estatal de Meteorología*) prediction model didn't reflect the real wind speed at the area and period of interest (Bailín, summer of 2012). That's the reason why a regression model (RM) has been developed; it has allowed us to downscale the AEMET wind speed predictions, adapting them to the real wind conditions in Bailín. The model provides the wind speed prediction at the Bailín area, 24 hours in advance.

The construction of the RM was carried out within the R environment, through the following steps: 1) Study of the useful parameters to be included at the model, and construction of a database with them; 2) Developing of a model-construction and model-selection strategy (with and without interactions), using *stepwise* methods, as well as model selection criteria (R^2 adjusted, residual deviation of cross validation or ANOVA test); 3) Study of the models' fit, by studying its residuals; 4) Study of the finally-selected-models' running at the area and period of interest; 5) Final best-model selection.

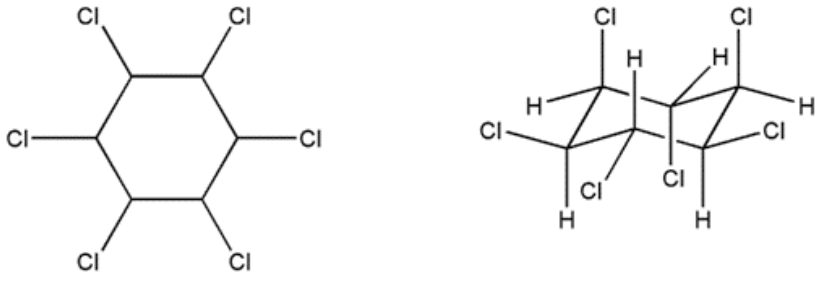
The model finally selected has been able to increase the explanation of the real wind speed variability up to a 57%, with regard to the 14% explained by the AEMET predictions. The methodology developed also pretends to provide a tool for solving similar prediction problems.

1. INTRODUCCIÓN

HCH TÉCNICO Y LINDANO

Desde la antigüedad, la práctica de la agricultura ha estado unida al uso de compuestos destinados a mejorar la producción y evitar las plagas. A comienzos del siglo XX, y especialmente tras la Segunda Guerra Mundial, comenzó la utilización masiva de productos químicos organoclorados[1] como pesticidas, tanto en la agricultura como en el control de parásitos (Tadeo, 2008). Uno de los que se introdujo más rápidamente en el mercado mundial de productos fitosanitarios fue el hexaclorociclohexano (HCH), gracias a sus propiedades insecticidas, su rápido efecto y su fácil obtención industrial (Li et al., 1998).

El HCH es un insecticida organoclorado de amplio espectro que se usó principalmente desde la Segunda Guerra Mundial hasta la década de 1990 (Forter, 1995; Breivik et al., 1999). Su nomenclatura es 1,2,3,4,5,6-hexaclorociclohexano, pues todos sus hidrógenos se encuentran sustituidos por átomos de cloro (ver Figura 1 izda.). Lo componen 8 isómeros diferentes, y durante los primeros años, se utilizó el conjunto de todos ellos (conocido como HCH técnico) como un insecticida económico. Más adelante, se descubrió que su actividad insecticida radicaba únicamente en el isómero gamma (γ-HCH, ver Figura 1 dcha.), el cual comenzó a refinarse a partir del HCH técnico, y se comercializó con el nombre de *lindano* (Walker et al., 1999). El lindano representa únicamente entre el 8% y el 15% del HCH técnico (Vijgen et al., 2011), por lo cual su fabricación genera una gran cantidad de residuos, formados por el resto de isómeros "inactivos", principalmente α-HCH, β-HCH y δ-HCH (Bodenstein, 1972).



Hexaclorociclohexano (HCH) Lindano (γ–HCH)

Figura 1. Composición química del Hexaclorociclohexano (HCH técnico, izda.) y del lindano (γ-HCH, dcha.), el isómero que posee las propiedades insecticidas. (Fernández, 2004).

Los compuestos derivados del HCH pertenecen al grupo de los pesticidas persistentes por su baja biodegradabilidad, la cual, junto a su toxicidad, supone una grave amenaza tanto para el medio ambiente como para la salud (Willett et al., 1998). Sin embargo, durante muchos años se

[1] Compuestos químicos orgánicos, es decir, con un esqueleto de átomos de carbono, en el cual alguno de los átomos de hidrógeno han sido sustituidos por átomos de cloro.

consideró que dichos compuestos eran insolubles en el agua e inocuos en general, y fueron vertidos al medio ambiente sin ningún control e incluso utilizados en la construcción (Forter, 1995; Vijgen 2006) o pavimentación (Torres et al., 2012). Esto provocó importantes focos de contaminación a nivel mundial, los cuales se han detectado a lo largo de los últimos años (Götz et al., 2013; Jit et al., 2010; Vijgen et al., 2011; Weber and Varbelow 2012; Wycisk et al., 2012).

En la década de 1970 comenzaron a detectarse ciertos problemas ambientales ocasionados por los derivados del HCH, lo que propició su prohibición en muchos países (Vijgen et al., 2011; Voldner and Li, 1995; Walker et al., 1999) Sin embargo, no fue hasta el año 2009, cuando los isómeros α-HCH, β-HCH y γ-HCH (lindano) se incluyeron en la Convención de Estocolmo sobre Contaminantes Persistentes Orgánicos (COPs), cuyo principal objetivo es el de proteger la salud humana y el medio ambiente frente a este tipo de contaminantes. Desde entonces, el problema de la contaminación derivada del uso y la producción del HCH y el lindano, junto con sus depósitos de residuos, se reconoció como un asunto de preocupación a nivel mundial (Vijgen et al., 2011) debiendo identificarse los lugares contaminados, así como adoptar las medidas necesarias para gestionar dichos residuos.

En resumen, la producción de lindano ha generado el mayor almacén mundial de COPs, estimado entre 4.8 y 7.2 millones de tm (Vijgen, 2006; Vijgen et al., 2011). El tratamiento de las áreas contaminadas debe ser abordado como parte de la implementación de la Convención de Estocolmo.

ÁREA DE ESTUDIO. EL BARRANCO DE BAILÍN

En España, uno de los principales focos de contaminación por lindano se localiza en los alrededores de Sabiñánigo (Huesca), donde se calcula que la empresa INQUINOSA (Industrias Químicas del Noroeste, ver Figura 2), que operó en la zona entre 1975 y 1988, vertió en los alrededores 115 000 tm de residuos peligrosos organoclorados, en gran parte isómeros de HCH (Vijgen, 2006; Fernández et al., 2012) sin ningún tipo de política medioambiental.

Uno de los principales puntos de depósito fue el barranco de Bailín, nuestra zona de estudio (ver Figura 3). Según Fernández et al. (2012), en él se vertieron, entre los años 1984 y 1988, entre 30 000 y 80 000 tm^3 de residuos sólidos de HCH, así como 2000 toneladas de residuo líquido DNAPL (*dense nonaquous phase liquid,* ver Figura 4). Estas estimaciones lo sitúan como uno de los mayores depósitos de HCH a nivel mundial (Vijgen, 2006).

En el año 2005 se descubrió que el DNAPL estaba percolando hacia el río Gállego, que se encuentra a tan solo 800 metros. Por ello, en 2007 se decidió el desmantelamiento y transferencia del vertedero de Bailín a otro emplazamiento que tuviese las medidas de seguridad oportunas. Como nuevo emplazamiento, y teniendo en cuenta preocupaciones locales y consideraciones de seguridad, se eligió un sitio cercano en el mismo valle, el cual se ha construido con importantes medidas de aislamiento (ver Figura 5). La capacidad del nuevo vaso es de 250 000 m^3, suficiente para almacenar el contenido del vertedero de Bailín, los residuos que hay dentro de INQUINOSA y los suelos afectados alrededor del vertedero de Sardás, el segundo punto de mayor depósito de residuos, después de Bailín (Fernández et al., 2012).

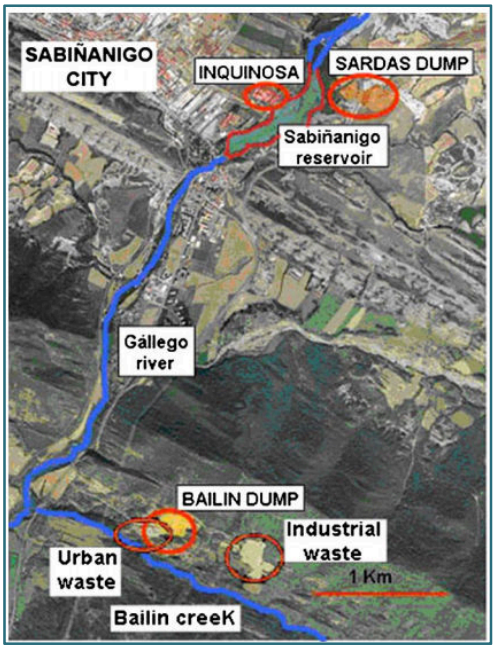


Figura 2. Antigua fábrica de INQUINOSA, hoy en día abandonada. *Fotografía extraída del Diario del Alto Aragón.*

Figura 3 Localización geográfica de los puntos de contaminación por residuos de HCH alrededor de Sabiñánigo, incluido el vertedero de Bailín (*Bailín dump*) (Fernández et al., 2012).

Figura 4. Residuo líquido DNAPL en el área del vertedero de Bailín. *Fotografía de Alfonso Pardo.*

Figura 5. Nuevo vaso construido para albergar los residuos organoclorados del antiguo vertedero de Bailín. *Fotografía de Jesús Fernández.*

Según Fernández et al. (2012), la construcción de dicha infraestructura comenzó en 2010 y las obras de desmantelamiento y transferencia se prevén para el verano de 2014, entre los meses de mayo a octubre, cuando las precipitaciones son menores. Aunque esto no supone una eliminación final de los residuos de HCH, el desmantelamiento es necesario dado el impacto negativo que las percolaciones de DNAPL suponen para el río Gállego.

Según lo dispuesto en la Convención de Estocolmo, en un futuro se deberán eliminar los residuos almacenados en el nuevo vaso. En Bailín, este proceso se llevará a cabo cuando esté disponible un tratamiento asequible.

DISPERSIÓN DE PARTÍCULAS EN VERTEDEROS Y PROTOCOLO DE ACTUACIÓN EN BAILÍN

Hasta hace no muchos años no existían normas sobre deposición de residuos en vertederos y, tanto empresas como núcleos urbanos, depositaban sus residuos en los alrededores de su actividad sin ningún tipo de control. Como ya se ha mencionado anteriormente, lo mismo ocurrió con los residuos de fabricación del lindano. Hoy en día, el principal método de eliminación de residuos en Europa sigue siendo su deposición en vertederos (Giusti, 2009), pero la UE controla esta actividad gracias a reglas muy estrictas en su legislación (Directiva 2008/98/EC).

Se espera que durante el proceso de desmantelamiento y transferencia de los residuos contenidos en el vertedero de Bailín al nuevo vaso, se emita una gran variedad de contaminantes al aire, principalmente gases y partículas (Chalvatzaki et al., 2010; Kopanakis et al., 2011), a causa de los procesos que se llevarán a cabo de descomposición, resuspensión y operaciones típicas de vertedero, como carga y descarga de residuos, o transporte de los mismos. Estas emisiones estarán también influenciadas por las condiciones climáticas locales, especialmente por la presencia e intensidad del viento (Chalvatzaki et al., 2012). Por ello, las operaciones llevadas a cabo en el vertedero de Bailín se suspenderán cuando se prevean lluvias e intensidades de viento elevadas (Fernández et al., 2012).

En resumen, uno de los mayores retos en el desmantelamiento y transferencia de los residuos contenidos en el vertedero de Bailín, consiste en evitar la dispersión atmosférica de partículas de lindano en la medida de lo posible, por los riesgos ambientales y para la salud que ésta supondría (Pope et al., 1995). Para ello, y según el protocolo de actuación de las obras, éstas se detendrán cuando se prevean vientos superiores a 40 km/h. Esta parada no será instantánea, es decir, una vez que esté sucediendo o haya sucedido el evento, sino que deberá anticiparse en la medida de lo posible, atendiendo a un modelo de predicción de la velocidad de viento.

Actualmente y para tal fin, se dispone de datos de predicción del viento “por horas” para municipios, calculados por AEMET a partir de los modelos numéricos de predicción europeos del ECMWF (*European Centre for Medium-Range Weather Forecasts*). Estas predicciones se ofrecen con un horizonte de hasta 48 horas después de la hora de actualización de dichos modelos (00, 06, 12 y 18 UTC).

Para este estudio, la Diputación General de Aragón (DGA) ha facilitado una recopilación de datos horarios de predicción de AEMET a 24 horas, de velocidad y dirección de viento, para el año 2012 y para el municipio de Sabiñánigo. No ha sido posible disponer de predicciones para otras variables meteorológicas, ni para otros periodos de tiempo.

En la zona del vertedero de Bailín se ubican una serie de instalaciones pertenecientes a la DGA, destinadas principalmente al estudio y control de la contaminación. Entre ellas, y en la cota menor del vertedero, se encuentra una estación meteorológica fija, gestionada por la empresa GEONICA S.A (ver Figura 6). La estación registra datos cada cinco segundos y calcula distintas medidas de velocidad de viento y dirección asociada, que proporciona en intervalos de 10 minutos. La cantidad de registros es muy amplia y por tanto, los datos en estudio disponibles son muy representativos del régimen de vientos real para la zona. La DGA también ha facilitado los datos recogidos en dicha estación durante el año 2012.

Figura 6. Estación meteorológica fija, situada a los pies del vaso viejo de Bailín.

Cuando se trabaja con datos de tipo meteorológico, hay que tener en cuenta la dificultad de su predicción, dados los complejos procesos de formación y evolución que sufren los fenómenos meteorológicos. Esta complejidad es especialmente patente en áreas reducidas con características locales especiales, como puede ser el barranco de Bailín. Recordemos que los datos de predicción por horas disponibles de AEMET, vienen dados para el área de Sabiñánigo y por tanto, pueden no coincidir con las observaciones en Bailín.

OBJETIVOS DEL TRABAJO

El objetivo general de este Trabajo Fin de Grado (TFG) es generar un modelo estadístico de predicción de la velocidad del viento para el área específica del vertedero de Bailín y para el periodo entre mayo y septiembre, que es cuando previsiblemente se llevará a cabo el grueso de las obras de desmantelamiento y transferencia del vertedero de Bailín. Este nuevo modelo pretende mejorar las predicciones disponibles, que son las que AEMET proporciona para el área del municipio de Sabiñánigo, para ajustarlas a la realidad del vertedero de Bailín. El nuevo modelo deberá ser capaz de mostrar en un instante t_0, la velocidad de viento que se registrará 24 horas después (es decir, en el instante t_0+24).

Para ello, se analizará en primer lugar el comportamiento del régimen real de vientos en el entorno de Bailín, estudiando cuestiones como las direcciones preferentes o posibles comportamientos estacionales del viento. A continuación, se compararán las predicciones proporcionadas por AEMET con los datos observados en el vertedero. Después, se construirán modelos de regresión que se desarrollarán a partir de las predicciones horarias de AEMET disponibles para Sabiñánigo y para el año 2012, y consistirán en el *downscaling* de dichas predicciones, adaptándolas al área y período de interés. Finalmente, se seleccionará el modelo óptimo.

Las hipótesis de partida, previas al desarrollo de dicho modelo, son las siguientes:

- El modelo de predicción de AEMET existente tiene capacidad predictiva para los datos en estudio.

- El modelo de AEMET está suavizado y por tanto, no recoge elevadas intensidades de viento de carácter extremo que puedan darse en realidad, y que por su capacidad dispersiva, resultan de interés para este estudio.
- Para adaptar el modelo de predicciones de AEMET a las condiciones locales específicas del área en estudio, es necesario realizar un modelo estadístico que reescale dichas predicciones.

El hecho de disponer de un modelo de predicciones que se ajuste mejor a la realidad supondrá una herramienta útil en la toma de decisiones, en lo que respecta a si las condiciones de viento previstas para la siguiente jornada (o las horas siguientes) son adecuadas para trabajar.

Así, el fin último del modelo de predicciones a desarrollar a lo largo este TFG es predecir posibles eventos de velocidad elevada con el tiempo suficiente para que se puedan parar las obras, evitando una potencial dispersión atmosférica de la contaminación, capaz de agravar la problemática ambiental existente en la zona.

Este estudio de la predicción y la planificación de los trabajos, viene contemplado en la Autorización Ambiental Integrada (AAI) para el proyecto de obras de la fase B del vertedero de HCH de Bailín, relativa al desmantelamiento y transferencia del mismo.

DESCRIPCIÓN DEL CONTENIDO DE LA MEMORIA

La memoria se ha organizado, tras la anterior introducción sobre la problemática en estudio y los objetivos de este TFG, en los siguientes capítulos:

- El **capítulo 2** contiene la presentación de las bases de datos disponibles y una introducción a la metodología llevada a cabo en los análisis exploratorios realizados. La finalidad de dichos análisis es, en primer lugar, estudiar el régimen real de vientos en el entorno de Bailín, profundizando en cuestiones como las direcciones preferentes o el comportamiento estacional del viento. También se presentan las herramientas para la comparación de las predicciones de AEMET con los datos observados.
- El **capítulo 3** expone la metodología desarrollada para la construcción del modelo estadístico óptimo de predicción de la velocidad del viento, mediante el diseño de procedimientos que permiten comparar una gran variedad de MR (lineales, no lineales, sin y con interacciones), así como de las herramientas de verificación del MR útiles para su control.
- El **capítulo 4** contiene los resultados obtenidos al aplicar la metodología de los dos capítulos anteriores a los datos en estudio, así como la discusión de dichos resultados. Como resultado final, se obtendrá el modelo óptimo buscado, y se estudiará su capacidad operativa.
- El **capítulo 5** contiene las principales conclusiones extraídas a lo largo de este TFG.

Además, la memoria se completa con tres **anexos**, que contienen gráficos y tablas referenciados en el capítulo 4 (*Anexo I*), la revisión de algunos conceptos estadísticos de interés mencionados a lo largo del trabajo (*Anexo II*), así como estudios adicionales (*Anexo III*).

2. DATOS DISPONIBLES Y ANÁLISIS EXPLORATORIOS

Antes de comenzar con la modelización del conjunto de datos en estudio, es importante llevar a cabo una serie de análisis previos, cuya finalidad será estudiar el comportamiento del régimen de vientos en el entorno de Bailín y de las predicciones de AEMET correspondientes.

El análisis exploratorio se ha realizado en tres fases:

1. *Análisis y pretratamiento de la base de datos:*
    a. Análisis de las bases de datos (BD) disponibles (reales y predichos).
    b. Pretratamiento de las BD reales, en vistas a posibilitar su comparación con la BD de predicción existente.

2. *Estudio de los regímenes de vientos y adecuación del modelo de predicciones de AEMET:*
    a. Estudio de la adecuación del modelo de predicciones de AEMET:
        i. mediante análisis simultáneo de datos reales y predichos de velocidad del viento, y
        ii. mediante rosas de vientos, para datos de velocidad y dirección.
    b. Estudio del ciclo diario de velocidad de viento mediante rosas de vientos y gráficos de variación de la velocidad media horaria.
    c. Estudio de la estacionalidad de las series de datos.

3. *Análisis de correlación.*
    a. Análisis de correlación, correlación cruzada y autocorrelación para las variables disponibles (velocidad y dirección del viento, reales y predichas)

Se pretende que los resultados de este análisis sirvan de guía a la hora de plantear la metodología de desarrollo de un adecuado modelo de predicciones.

2.1 ANÁLISIS Y PRETRATAMIENTO DE LA BASE DE DATOS

2.1.1 ANÁLISIS DE LAS BASES DE DATOS DISPONIBLES

Como punto de partida, se dispone de dos BD. Una corresponde a datos observados en el área de Bailín, recogidos por una estación meteorológica fija (ver Figura 6); contiene datos cada diez minutos. La otra contiene datos horarios de predicción para Sabiñánigo, procedentes del modelo de AEMET.

A su vez, se dispone de dos tipos de observaciones de viento:

- Velocidad media (en km/h) cada diez minutos. Es el dato que resulta de promediar, cada diez minutos, todos los datos de velocidad de viento registrados por la estación fija

durante ese periodo (recordemos que la estación registra datos cada cinco segundos). También se dispone de la dirección del viento media (en grados) para dicho intervalo.
- Velocidad máxima (en km/h) cada diez minutos. Es el dato máximo de velocidad registrado por la estación fija en cada intervalo. Se dispone también de su dirección instantánea correspondiente (es decir, recogida en el mismo instante de tiempo).

Se ha trabajado con la BD de mayo a septiembre de 2012. El hecho de que haya datos disponibles cada diez minutos, implica que se ha trabajado con un total de 22 032 registros observados.

En la BD predichos, se dispone de la predicción instantánea esperada para un intervalo de diez minutos centrado en cada hora en punto, de velocidad del viento (en km/h) y de dirección del viento correspondiente (en grados). Por tanto, comprende un total de 3672 registros de predicción.

2.1.2 PRETRATAMIENTO DE LA BASES DE DATOS OBSERVADOS

Como se ha visto, se dispone de una BD en intervalos de 10 minutos, tomados en el área de estudio, y de una BD de predicciones en intervalos de 60 minutos. Por tanto, será necesario realizar un pretratamiento de la BD observados con el objeto de obtener datos horarios que permitan la comparación entre ambas. Para ello, se han planteado dos opciones:

- Seleccionar y utilizar únicamente el dato de velocidad de viento real para cada hora en punto, así como su dirección correspondiente, obviando todo el resto de datos disponibles. A este tratamiento se le llamará P1 (pretratamiento 1).
- Seleccionar y utilizar el dato máximo de entre los tres datos reales de velocidad de viento disponibles alrededor de cada hora (10 minutos antes, hora en punto, 10 minutos después). A este tratamiento se le llamará P2 (pretratamiento 2).

Se cree que mediante la aplicación de P2 se obtendrá un dato real más semejante a los datos de predicción, ya que de esa forma, ambas BD mostrarán datos instantáneos, centrados alrededor de la hora en punto. Además, el hecho de trabajar con tres datos por cada hora y no solo con uno, permitirá recoger una mayor información acerca del régimen real de vientos.

Nuestra BD observados es bastante extensa, así que sería poco eficaz tratarla mediante una hoja de cálculo (tipo Excel). Por ello, se diseñará un procedimiento en R.

R es un lenguaje y un entorno de programación de libre descarga, que permite manipular bases de datos, realizar cálculos y visualizar gráficos, mediante una gran variedad de técnicas gráficas y estadísticas. El lenguaje de programación de R es simple e intuitivo, y permite al usuario definir condiciones, bucles y funciones recursivas. En los llamados “paquetes de R”, también de libre descarga, los usuarios ponen a disposición del público funciones desarrolladas por ellos mismos, que se pueden implementar para BD particulares.

En este caso, el procedimiento diseñado en R permitirá:

- Seleccionar el dato que corresponde a la velocidad máxima real, de las medidas comprendidas entre la hora en punto y los intervalos de 10 minutos adyacentes.

- Seleccionar el dato de dirección de viento correspondiente al instante en que se produce dicha velocidad. En caso de que dos (o los tres) datos de velocidad en estudio sean idénticos, el procedimiento deberá hallar la media de las direcciones correspondientes a cada una de dichas velocidades.

Este procedimiento corresponde al P2 anteriormente explicado. Una alternativa a dicho procedimiento permitirá seleccionar únicamente los datos que corresponden a la hora en punto, tanto de velocidad como de dirección de viento; es decir, permitirá aplicar P1 a los datos observados.

Mediante el pretratamiento de las BD observados (de velocidad media y máxima, y direcciones correspondientes), se obtendrán cuatro BD diferentes:

- BD.med1, que resultará de aplicar P1 a la base de datos medios. Estará formada por los datos de velocidad media y su dirección correspondiente para horas en punto.
- BD.max1, que resultará de aplicar P1 a la base de datos máximos.
- BD.med2, que resultará de aplicar P2 a la base de datos medios. Por tanto, contendrá los datos más elevados de velocidad media, para el intervalo de 10 minutos antes y después de la hora en punto, así como su dirección correspondiente.
- BD.max2, que resultará de aplicar P2 a la base de datos de velocidad máxima.

La base de datos de predicciones de AEMET se denominará BD.p.

Se estudiará la comparación entre las BD observados y la BD.p, seleccionando los datos correspondientes al periodo de interés (entre mayo y septiembre). Como se ha mencionado anteriormente, se cree que las BD obtenidas tras aplicar P2 contendrán datos más representativos de la variabilidad real del régimen de vientos en Bailín. La hipótesis de partida es que se debería trabajar con BD.max2, ya que según los objetivos del estudio (evitar la propagación de contaminantes por efecto del viento), será más adecuado trabajar con datos máximos de velocidad del viento. Sin embargo, y dado que los modelos de predicción se encuentran suavizados, serán los datos de velocidad media los que se ajustarán mejor a las predicciones de AEMET.

Para comprobar las hipótesis anteriores, se llevarán a cabo una serie de análisis gráficos en forma de histogramas que contendrán las discrepancias para cada instante de tiempo (es decir, las diferencias entre las cuatro BD reales obtenidas y BD.p). También se estudiará la magnitud de discrepancias (de ±5 km/h y de ±10 km/h) en forma de porcentajes.

Las discrepancias halladas tendrán signo negativo para situaciones en las que el modelo de AEMET predice velocidades superiores a las que realmente se observan, y positivo en caso de predecir velocidades inferiores a las realmente observadas (situaciones que llamaremos de *falsa alarma* y *alarma* respectivamente). Según el objetivo del trabajo, las situaciones problemáticas serán las *situaciones de alarma*.

2.2 ESTUDIO DE LOS REGÍMENES DE VIENTOS, REAL Y PREDICHO

La finalidad de este estudio será tener una idea más precisa del régimen de vientos en el área del vertedero de Bailín, así como de la adecuación del modelo de predicciones de AEMET, es decir, si dicho modelo es realmente capaz de recoger el comportamiento real del viento para el área y el periodo de interés, así como su comportamiento estacional.

- **Adecuación del modelo de predicciones de AEMET**

Para el análisis de la adecuación del modelo de predicciones de AEMET, en términos de velocidad del viento, se llevará a cabo un análisis simultáneo de datos reales y predichos mediante un gráfico de dispersión. En caso de que el modelo de predicciones fuese adecuado, dicho gráfico debería contener puntos alrededor de la línea identidad, o al menos, de una recta de pendiente positiva.

Para el análisis de la adecuación del modelo de predicciones, en términos tanto de intervalos de velocidad como de dirección correspondiente, se utilizarán rosas de los vientos, herramienta que se ha introducido brevemente en el Anexo II, y que se realizarán mediante el programa de análisis de viento *WRPlot view*^{TM} *– Freeware* (Lakes Environmental, 2011).

Se representarán rosas de vientos para tres intervalos diferentes de velocidad, los cuales se calcularán mediante el uso de percentiles, que permitirán tener en cuenta la hipótesis de suavizado del modelo de predicciones de AEMET. Los intervalos representarán situaciones de *calma* (velocidades bajas), *intensidad media de viento* e *intensidad de viento elevada*.

La comparación del régimen de vientos para ambas BD se realizará “hora a hora” (es decir, para el mismo instante de tiempo) y de dos formas. Primero, se estudiarán los datos reales correspondientes a cada intervalo y la situación de su predicción instantánea (*situación 1*); después, se estudiará la predicción para cada intervalo, y el grado de similitud de la observación instantánea (*situación 2*).

Para poder aceptar como adecuado al modelo de predicciones de AEMET, la hipótesis de partida es que siendo que se trabajará con percentiles, las rosas de vientos deberían tener un aspecto similar, para cada situación y para cada intervalo.

- **Estudio de la estacionalidad y el ciclo diario**

Para el estudio del ciclo diario se utilizarán de nuevo rosas de vientos, en este caso realizadas para los datos correspondientes al horario diurno (considerando como tal al periodo comprendido entre las 8 y las 19 horas, pues se piensa que éste será, aproximadamente, el principal horario de trabajo) y al nocturno (de 20 a 7 horas), por separado. Las diferencias significativas entre las rosas en ambos horarios, indicarán la existencia de un ciclo diario en la intensidad del viento. Para el estudio del ciclo diario también se realizará un gráfico de variación horaria de las velocidades medias, reales y predichas.

Para estudiar la estacionalidad se llevarán a cabo dos tipos de gráficos. En primer lugar, de variación de la velocidad media a lo largo del periodo en estudio (para cada mes), y en segundo lugar, de dicha variación para distintos cuantiles.

Como resultado de los estudios anteriores, se comprobará también la adecuación o no de las predicciones de AEMET al comportamiento real del viento.

2.3 ANÁLISIS DE CORRELACIÓN

Para finalizar el análisis exploratorio del conjunto de datos, se llevará a cabo un estudio de la correlación entre las covariables de interés, cuyo objetivo es el análisis de las relaciones existentes entre ellas. En el Anexo II se ha incluido una introducción acerca de los conceptos que se utilizan a continuación.

Para los pares de variables de velocidad (real y predicha) y dirección (real y predicha), se estudiará el coeficiente de correlación de Pearson y la correlación cruzada; la autocorrelación se estudiará para cada variable por separado. Los análisis se llevarán a cabo mediante gráficos de correlación (correlogramas), herramientas útiles para estudiar la existencia de una correlación retardada (es decir, una correlación mayor con otro instante de tiempo), una autocorrelación, o para analizar la amplitud de los ciclos existentes.

3. METODOLOGÍA. CONSTRUCCIÓN DE MODELOS DE REGRESIÓN

El fin último de este TFG es realizar un modelo estadístico de predicción de la velocidad del viento que sea capaz de adaptar las predicciones de AEMET al área y periodo de interés. Esta adaptación se ha conseguido mediante la realización de modelos de regresión (MR); en el Anexo II se ha incluido una introducción acerca de los mismos.

En este apartado se va a presentar la metodología desarrollada a lo largo del TFG para la construcción de dichos MR. El fin último será encontrar el modelo óptimo de entre todos los construidos, es decir, aquel que ajustándose mejor a la BD en estudio, ofrezca un mejor pronóstico.

A lo largo de todo el proceso, y para la implementación de la metodología desarrollada, se utilizará el programa R.

Previo al comienzo de la búsqueda del modelo candidato óptimo, se desarrollarán modelos de regresión lineal simple (MRLS), que únicamente incluirán como variable regresora a las predicciones de la velocidad del viento de AEMET. Su objetivo será estudiar la adecuación del modelo de AEMET al régimen real de vientos, de una forma más precisa a la estudiada mediante el análisis exploratorio de datos, y para distintos periodos.

Para obtener el MR óptimo se seguirán los pasos que se indican a continuación:

1. *Procedimiento de construcción de modelos candidatos*, que incluye:
    a. Estudio y construcción de variables potencialmente útiles.
    b. Estudio de la transformación de variables.
    c. Estrategia de construcción de modelos.
    d. Estudio de la linealidad de ciertas covariables.
    e. Criterios de bondad de ajuste (en términos de R^2 ajustado).
    f. Selección de términos (semiautomática y manual) candidatos a formar parte del modelo óptimo.

2. *Selección de los "mejores" modelos candidatos* una vez desarrollados, atendiendo a distintos criterios.

3. *Estudio de la adecuación de los modelos seleccionados* mediante el análisis de residuos.

4. *Estudio del funcionamiento de los modelos finalmente seleccionados para la realización de pronósticos:*
    a. Estudio gráfico del pronóstico, para distintos umbrales de velocidad.
    b. Estudio del funcionamiento mediante tablas de contingencia para datos reales y de predicción instantáneos, y estudio de tablas con datos de acumulación horaria para distintos umbrales de velocidad de viento.

5. *Selección final del modelo óptimo*. *Desarrollo de un procedimiento en R* capaz de obtener pronósticos del modelo óptimo seleccionado a nuevos datos de predicción de AEMET.

3.1 PROCEDIMIENTO DE CONSTRUCCIÓN DE MODELOS CANDIDATOS

3.1.1 ESTUDIO DE VARIABLES POTENCIALMENTE ÚTILES. CONSTRUCCIÓN DE LA BASE DE DATOS

En este apartado, se va a explicar cómo desarrollar la BD "final" que contendrá las covariables usadas en la construcción de los MR. A partir de ellas, también podrán considerarse interacciones. Se ha llamado *términos* al conjunto de covariables e interacciones que formarán parte de distintos modelos candidatos. Se llamará *variable* o *covariable* a las variables "aisladas" (es decir, que no forme parte de una interacción).

Las covariables que compondrán la BD final son las siguientes:

- Covariables que expresen la predicción de la velocidad del viento.
- Covariables que expresen la predicción de la dirección del viento.
- Covariables que expresen el ciclo diario de brisas.
- Covariables que expresen la estacionalidad de la serie de datos.

Se tendrán en cuenta las características que las anteriores covariables deban poseer para que puedan ser introducidas al modelo, siendo capaces de reproducir de forma adecuada el comportamiento del fenómeno que con ellas se pretende estudiar.

Para la construcción de la BD final, se introducirán en primer lugar los datos de predicción de velocidad de viento de AEMET (instantáneos), así como datos construidos en una ventana temporal alrededor del instante de interés estudiado (t_0-4, t_0+4). A dichos datos se les denotará como:

- x a la predicción instantánea de velocidad de viento de AEMET, para t_0.
- $x1, x2, x3, x4$ a los datos de predicción de la velocidad del viento con un desfase horario entre t_0-1 y t_0-4 horas ("retardos", ver Figura 7 izda.).
- $x.p1, x.p2, x.p3, x.p4$ a la predicción con un desfase horario entre t_0+1 y t_0+4 horas ("adelantos", ver Figura 7 dcha.).

Se procederá del mismo modo para los datos de predicción de la dirección, teniendo en cuenta que es una variable circular y viene dada en grados. Por tanto, se trabajará con ella en forma del seno y coseno del primer armónico, aplicando las siguientes expresiones, donde dx expresa la dirección predicha en el instante t_0.

$$sendv = sin\frac{dx}{360}2\pi \quad ; \quad cosdv = cos\frac{dx}{360}2\pi$$

Para los armónicos de dirección, se utilizará el mismo código de retardos y adelantos mostrado para las covariables de velocidad.

x	x1	x2	x3	x4
5.40	NA	NA	NA	NA
7.20	5.40	NA	NA	NA
9.36	7.20	5.40	NA	NA
11.16	9.36	7.20	5.40	NA
12.60	11.16	9.36	7.20	5.40
12.60	12.60	11.16	9.36	7.20
9.36	12.60	12.60	11.16	9.36
10.80	9.36	12.60	12.60	11.16
10.08	10.80	9.36	12.60	12.60
10.44	10.08	10.80	9.36	12.60

x	x.p1	x.p2	x.p3	x.p4
5.40	7.20	9.36	11.16	12.60
7.20	9.36	11.16	12.60	12.60
9.36	11.16	12.60	12.60	9.36
11.16	12.60	12.60	9.36	10.80
12.60	12.60	9.36	10.80	10.08
12.60	9.36	10.80	10.08	10.44
9.36	10.80	10.08	10.44	11.52
10.80	10.08	10.44	11.52	11.88
10.08	10.44	11.52	11.88	12.96
10.44	11.52	11.88	12.96	11.16

Figura 7. Muestra de la construcción de las covariables de predicción de la velocidad del viento alrededor de la ventana temporal (t_0-4, t_0+4). A la izda. se muestra el desfase horario existente entre $x, x1, x2, x3, x4$; a la dcha. el desfase entre $x.p1, x.p2, x.p3, x.p4$. En ambas figuras, *NA* indica que no existen datos.

También se utilizarán armónicos para expresar la estacionalidad y los ciclos diarios existentes. Para el ciclo diario se calcularán las siguientes fórmulas, donde $hora$ es una componente de la BD que contiene los datos de la hora en punto para cada día, correspondientes a cada observación.

$$sen.hora = sin\frac{hora}{24}2\pi \quad ; \quad cos.hora = cos\frac{hora}{24}2\pi$$

Por su parte, los términos que representan la estacionalidad se calcularán mediante la siguiente fórmula, donde $dia.año$ nos indica el día del año en que se ha registrado cada dato. Esta covariable tendrá en cuenta desde el día 1 de mayo (que corresponde al día 123 del año 2012) hasta el día 30 de septiembre.

$$sen.dia = sin\frac{dia.año}{366}2\pi \; ; \; cos.dia = cos\frac{dia.año}{366}2\pi$$

Con todas las covariables anteriores (31 en total, de 4 tipos diferentes) se construirá la BD final que se usará en la construcción de los MR (ver Tabla 1).

Tipo de covariable	Nº cov.	Código en BD
Predicción de velocidad de viento	**9**	
instantánea	1	x
retardos	4	$x1, x2, x3, x4$
adelantos	4	$x.p1, x.p2, x.p3, x.p4$
Predicción de dirección de viento	**18**	
instantánea	2	$sendv, cosdv$
retardos	8	$send1, cosd1, send2, cosd2, send3, cosd3, send4, cosd4$
adelantos	8	$send.p1, cosd.p1, send.p2, cosd.p2, send.p3, cosd.p3, send.p4, cosd.p4$
Ciclo diario	**2**	$sen.hora, cos.hora$
Ciclo anual	**2**	$sen.dia, cos.dia$
TOTAL	**31**	

Tabla 1. Base de datos final, usada para el desarrollo de MR candidatos a ser el modelo de predicción óptimo.

3.1.2 ESTUDIO DE UNA POSIBLE TRANSFORMACIÓN DE VARIABLES

El estudio de la transformación de la respuesta será necesario cuando los residuos de los MR en estudio no cumplan las hipótesis de distribución normal y varianza constante (ver apartado 3.3).

Para ello, se estudiará la transformación Box Cox, procedimiento estadístico que se resume en el Anexo II.

Se considerará también una transformación de tipo polinómico para las covariables regresoras de predicción de la velocidad de viento, las únicas susceptibles de precisar de dicha transformación para explicar su relación con la variable respuesta.

3.1.3 ESTRATEGIA DE CONSTRUCCIÓN DE MODELOS

El elevado número de potenciales covariables hace inviable la construcción exhaustiva de todos los modelos posibles. Por ello, se ha diseñado una estrategia que ha consistido en el desarrollo de modelos sucesivos, desde los más sencillos posibles (con tan solo un tipo de covariables y sin interacciones) hasta los más complejos, con todos los tipos de covariables, transformaciones polinómicas e interacciones de todo tipo.

- **Desarrollo de modelos de regresión lineal sin interacciones**

En primer lugar se desarrollarán modelos sucesivos, a los cuales se irán añadiendo, en cada paso, covariables de diferente tipo. En el proceso se estudiará cuáles de dichas covariables son más significativas para explicar la respuesta. Las covariables añadidas serán, en este orden, covariables de predicción de la velocidad del viento, armónicos de predicción de la dirección del viento, armónicos de estacionalidad y armónicos de ciclo diario.

A su vez, las covariables se añadirán mediante dos métodos diferentes. El *método 1* consistirá en añadir un nuevo tipo de covariable al "mejor" modelo hallado en el paso anterior; el *método 2* estudiará directamente el modelo que incluya todas las covariables potenciales, en un único paso. Para hallar el "mejor" modelo en ambos casos, se seguirán los criterios de bondad de ajuste y de selección de variables, posteriormente explicados en el apartados 3.1.4.

Mediante cada método se obtendrá un "mejor modelo", susceptible de contener los cuatro tipos de covariables.

Se realizará entonces un estudio de la linealidad de los "mejores modelos", es decir, de la necesidad de expresiones polinómicas para las covariables de predicción de la velocidad del viento. El proceso consistirá en realizar distintos modelos anidados (es decir, donde los predictores de los modelos más sencillos sean un subconjunto de los que forman los modelos más complejos) con la función *gam*, *generalized additive models* (Hastie and Tibshirani, 1991), y comparar dichos modelos con un test ANOVA (Fox and Weisberg, 2011), comprobando cuál de ellos resulta más significativo. En el Anexo II se ha recogido una breve introducción acerca del test ANOVA.

Tras el estudio de la linealidad, se desarrollarán nuevos modelos en los cuales se incluirán los términos polinómicos que hayan resultado significativos.

- **Desarrollo de modelos de regresión con interacciones**

Una vez que se hayan obtenido los "mejores" modelos posibles sin interacciones, lineales y polinómicos, y con el fin de aumentar la bondad de ajuste del modelo, se estudiará la introducción de interacciones a los modelos candidatos, para encontrar aquellas que puedan resultar significativas.

La introducción de interacciones a los modelos podrá realizarse mediante bucles, lo cual permitirá estudiar un espectro de combinaciones de covariables e interacciones mucho mayor que mediante la introducción manual.

Las interacciones que se ensayarán serán las que a priori, y dados los análisis exploratorios anteriores, se crea que resultarán más significativas.

La búsqueda mediante bucles será de dos tipos:

- *Bucle 1* → Considerará covariables e interacciones guardadas en un vector, una en cada repetición del bucle, y desarrollará un modelo con cada una mediante un procedimiento "paso a paso" (*stepwise*, ver apartado 3.1.4), partiendo de uno de los "mejores" modelos obtenidos en el apartado anterior. En cada repetición, seleccionará y guardará en otro vector todos aquellos términos que hayan resultado significativos. Una vez el bucle termine, se ajustará manualmente un MR que incluya todos los términos guardados.
- *Bucle 2* → Considerará covariables e interacciones guardadas en el mismo vector anterior, y desarrollará un modelo con ellas. La diferencia es que en este caso, el modelo se "actualizará", es decir, en cada repetición del bucle el modelo añadirá nuevos términos y eliminará aquellos que no resulten significativos, también en un procedimiento de tipo *stepwise*. Por tanto, el resultado del bucle será directamente un MR candidato, resultado de aplicar todos los términos en estudio, en todas las repeticiones.

A lo largo del desarrollo de los modelos es importante tener en cuenta que habrá que seleccionar, tanto términos a incluir en el modelo, como modelos propiamente dichos, mediante criterios de selección de variables y de bondad de ajuste respectivamente, los cuales se estudian a continuación.

3.1.4 CRITERIOS DE BONDAD DE AJUSTE Y SELECCIÓN DE VARIABLES

Los criterios de bondad de ajuste permiten medir la capacidad del modelo para explicar el comportamiento de la variable respuesta. De entre los existentes, se ha seleccionado el R^2 ajustado. En el Anexo II se puede encontrar más información sobre el mismo.

Anteriormente, se ha explicado cómo construir la BD final, constituida por 31 covariables, que junto con las interacciones construidas entre ellas, serán candidatas a formar parte del MR óptimo que se busca. Sin embargo, probablemente solo un subconjunto de ellas estará relacionado con la respuesta. Por tanto, se debe realizar un proceso de selección, cuyo fin será encontrar la "mejor ecuación de regresión" (Montgomery et al., 2012), es decir, un MR con suficientes términos como para capturar el comportamiento de la variable respuesta, pero que

sea fácil de usar e interpretar y por tanto, que tenga el menor número posible de ellos. La selección se puede llevar a cabo de forma exhaustiva o automática.

La aproximación exhaustiva es la más directa a la hora de decidir qué términos son importantes para el modelo, y consiste en estudiar todas las combinaciones de modelos posibles para encontrar aquellas más predictivas. Este estudio solo resulta abordable si el número de términos en estudio es pequeño, ya que las posibles combinaciones con p parámetros es del orden de 2^p (James et al., 2013).

Por tanto, cuando se dispone de un número elevado de términos, se necesita de una aproximación semiautomática. En el Anexo II se puede encontrar una introducción acerca del método semiautomático *stepwise* de selección mixta, utilizado en este trabajo.

Cuando se lleva a cabo la selección mixta puede ocurrir que el modelo resultante contenga covariables no significativas, porque formen parte de interacciones que sí lo sean. Esto ocurre por el llamado *principio de jerarquía*, que expone que si se incluye una interacción en el modelo, también deben incluirse las covariables que la forman, para que dicho modelo se pueda interpretar (James et al., 2013). Sin embargo, se ha comprobado que los modelos candidatos son demasiado complejos en términos de interpretabilidad. Además, se ha establecido como criterio que todos los parámetros regresores incluidos en el modelo deberán ser significativos para el mismo, es decir, tener un p-valor asociado a sus parámetros inferior a 0.05.

Esto supone que tras la aplicación de la selección mixta, se deberá realizar una limpieza manual del modelo de términos no significativos para el mismo. Esta actualización del modelo se debe hacer "paso por paso", es decir, se debe eliminar (o añadir) un único término cada vez, estudiando cómo afecta este cambio a la significación, tanto del modelo como de los términos que lo forman, hasta obtener un modelo en el cual todos los términos sean significativos.

En caso de que existan varios modelos anidados, con un ajuste similar y con distinto número de variables, se estudiará su adecuación mediante su comparación con un test ANOVA.

3.2 SELECCIÓN DE MODELOS CANDIDATOS

Una vez se haya desarrollado un conjunto de modelos candidatos, sin y con interacciones, se hará una selección de los "mejores" atendiendo a los siguientes criterios:

- La bondad de ajuste de cada modelo, expresada por su R^2 ajustado y por su desviación típica residual.
- La composición de los mismos, es decir, el número de términos que lo forman, el grado de significación de dichos términos a la hora de predecir la respuesta, la existencia de parámetros no deseables o el número de interacciones triples.
- La desviación típica residual de validación cruzada, técnica posteriormente introducida.
- La bondad de ajuste de cada modelo a cada mes por separado y en cada una de las 24 horas.

El proceso de selección de los modelos candidatos consiste en hallar un equilibrio entre los criterios anteriores; los criterios que primarán serán la bondad de ajuste y el resultado de la validación cruzada, en términos de desviación típica residual. En especial, si ambos criterios tienen similar resultado para los modelos en estudio, será importante atender a su composición, teniendo preferencia los modelos más sencillos, evitando con ello posibles sobreajustes. El estudio de la composición también aportará información sobre posibles deficiencias en el comportamiento del modelo, que deberán subsanarse.

Se atenderá asimismo al grado de significación de los términos, prefiriendo modelos cuyos términos sean muy significativos a la hora de explicar la respuesta (es decir, cuyo p-valor sea inferior a 0.01).

Por último, la bondad de ajuste de cada modelo a cada mes por separado, indicará si el modelo funciona correctamente para todo el periodo, o si por el contrario será necesario el desarrollo de un modelo adaptado a algún mes o periodo del día en particular.

3.3 ESTUDIO DE LA ADECUACIÓN DE LOS MODELOS SELECCIONADOS

La adecuación de los modelos seleccionados se estudiará mediante el análisis de sus residuos, que permiten comprobar si el modelo tiene realmente capacidad predictiva, es decir, si es capaz de recoger el comportamiento de la variable respuesta en estudio. Para considerar que la estimación de los parámetros de los modelos es adecuada, se deben suponer, al menos, las siguientes hipótesis:

- Los errores tienen media cero.
- Los errores tienen varianza constante (situación de homocedasticidad).
- Los errores se distribuyen según una normal.

El cumplimiento de las hipótesis anteriores es condición básica para considerar como apropiado el test de hipótesis, que determina si existe relación lineal entre la respuesta y un subconjunto de variables regresoras, siendo:

$$H_0: \beta_1 = \beta_2 = \cdots = \beta_k = 0$$

$H_1: \beta_j \neq 0$, para al menos un valor de *j*

Así, el rechazo de la hipótesis nula implica que al menos una de las variables regresoras contribuye al modelo de forma significativa.

El análisis de los residuos se realizará mediante gráficos de diagnóstico que permitirán el estudio de su linealidad, homocedasticidad, distribución y puntos influyentes. La distribución se estudiará también mediante histogramas. Todos los gráficos anteriores vienen explicados en el Anexo II.

En caso de encontrarse deficiencias en la normalidad de los residuos de los modelos candidatos seleccionados, se llevarán a cabo los siguientes estudios:

- Transformación Box Cox de la respuesta.
- Gráfico secuencial de residuos.
- Gráfico de residuos sobre variables regresoras incluidas en el modelo (velocidad y dirección del viento).
- Gráficos de residuos en forma de diagramas de caja, para meses y horas.
- Gráficos residuales parciales, obtenidos en R mediante la función *crPlots* (Fox and Weisberg, 2011), que evalúan el efecto de cada predictor sobre la variable respuesta. Por tanto, permiten estudiar deficiencias en el modelo (p.e., necesidad de transformaciones polinómicas).
- Gráficos de variable añadida, obtenidos en R mediante la función *avPlots* (Fox and Weisberg, 2011), en los cuales se estudiará la pendiente hallada en cada gráfico, pues ésta representa el coeficiente parcial de regresión para el parámetro en estudio.

3.4 ESTUDIO DE LA CAPACIDAD PREDICTIVA DEL MODELO

Anteriormente, se han citado los criterios de selección de modelos estudiados. Uno de ellos era la desviación típica residual de validación cruzada.

Se llama validación a la comprobación de la capacidad predictiva del modelo, es decir, si los modelos ajustados para la BD en estudio, funcionarían también para un nuevo conjunto de observaciones, o si por el contrario el modelo está sobreajustado y por tanto, solo es válido para la BD que se ha usado en su ajuste.

Antes de estudiar cómo realizar dicha validación, se deben tener en cuenta dos conceptos importantes: el error de validación (*test error*) y el error de entrenamiento *(training error*). Ambos se han explicado en el Anexo II.

Cuando se quiere validar un MR existen dos alternativas. La primera consiste en estudiar cómo se comporta el modelo en una BD que no se haya utilizado en su ajuste, con lo que se obtiene directamente el error de validación.

La alternativa, cuando no existe una nueva BD, es aplicar herramientas estadísticas de remuestreo (*resampling*) que permiten validar un modelo usando las observaciones previamente utilizadas para su ajuste. De los métodos de remuestreo existentes, se ha seleccionado el método de validación cruzada para *k* subconjuntos, que se da en dos pasos:

1. División de la BD en *k* subconjuntos, de forma consecutiva (ver ejemplo en Figura 8), o aleatoria (ver ejemplo en Figura 9).
2. Reajuste del modelo, en *k* iteraciones; en cada iteración se toma un subconjunto de prueba, y un subconjunto de entrenamiento.

En el Anexo II se ha incluido una explicación más desarrollada acerca del proceso de validación cruzada para *k* subconjuntos.

La validación cruzada se llevará a cabo en R mediante el paquete *cvTools* (Alfons, 2012).

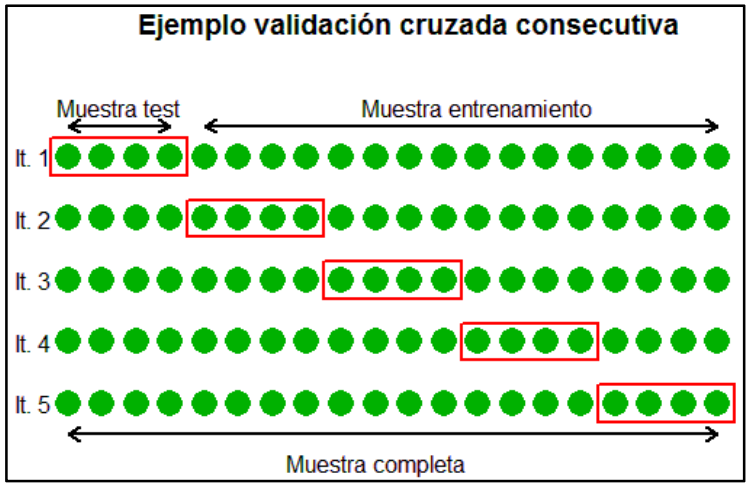


Figura 8. Ejemplo gráfico del método consecutivo de validación cruzada, en 5 iteraciones (It.).

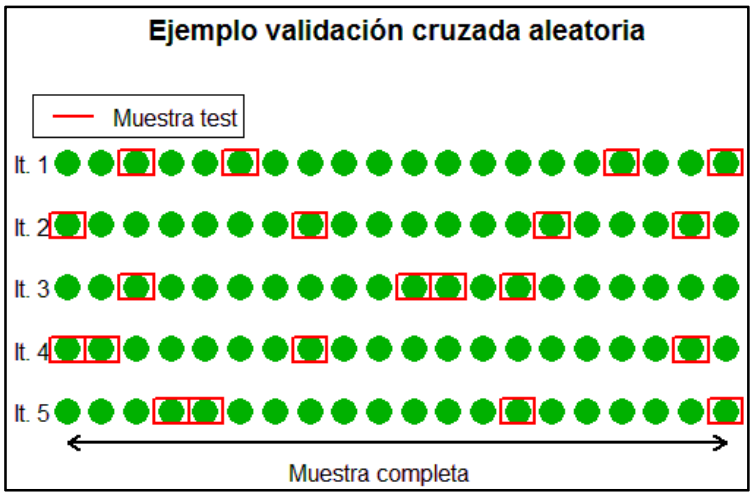


Figura 9. Ejemplo gráfico del método aleatorio de validación cruzada, en 5 iteraciones (It.).

3.5 ESTUDIO DEL FUNCIONAMIENTO OPERATIVO DE LOS MODELOS SELECCIONADOS. REALIZACIÓN DE PRONÓSTICOS

Una vez se ha estudiado la adecuación de los modelos seleccionados, comprobando que se cumplen las hipótesis sobre los residuos y que los modelos se ajustan a "nuevas" BD, y previo a la selección final del modelo óptimo, se estudiará el funcionamiento real de los modelos.

Resulta muy complicado que un modelo de tipo meteorológico acierte en sus predicciones de forma instantánea. Es decir, que se cumplan los valores que el modelo ha predicho para el momento exacto para el cual los ha predicho.

Por tanto, no interesa tanto estudiar las predicciones instantáneas de los modelos desarrollados, sino su comportamiento en términos de probabilidad de ocurrencia del fenómeno en estudio (pronósticos). En este caso, se estudiará la probabilidad de que el valor de velocidad de viento

predicho por el modelo, que se considera tiene distribución normal y desviación típica residual $\hat{s}_R$, supere a un determinado umbral y_0, dado por el propio valor observado.

$$P(N(\hat{y}, \hat{s}_R) > y_0) \quad \text{(Ec. 1)}$$

Los intervalos considerados para la probabilidad serán P<0.25, P>0.25, P>0.50, P>0.75, considerándose P>0.50, una probabilidad alta de ocurrencia de un cierto umbral de velocidad.

En primer lugar, se estudiará el comportamiento de los modelos en los "días conflictivos" de entre el periodo estudiado, considerando como tales a aquellos en los que se registraron datos de velocidad de viento superiores a 40 km/h ("situaciones de alarma"), ya que como se ha visto al comienzo del trabajo, es el umbral de velocidad al cual el protocolo de actuación recomienda detener las obras. Dicho estudio se realizará mediante gráficos que contendrán:

- Las observaciones de velocidad del viento registradas para cada día conflictivo seleccionado.
- En el eje izquierdo, la escala de velocidad del viento real observada.
- El eje derecho contendrá la escala de medida correspondiente a la probabilidad de superar distintos umbrales. El gráfico contendrá líneas discontinuas de referencia para P=0.25, P=0.50 y P=0.75.
- Los gráficos contendrán líneas (con códigos de color contenidos en la leyenda) que indicarán la probabilidad, ajustada a cada hora, de superar un umbral de velocidad de viento determinado.

Se realizarán gráficos para el estudio del pronóstico de cada modelo seleccionado y para cada día problemático.

Se parte de la hipótesis de que una obra de gran envergadura no se puede parar para un único instante, si no que de pararse, se haría durante un periodo amplio o incluso durante un día entero, en caso de existir alarmas de velocidades elevadas. Lo que el gráfico de pronósticos explicado permitirá, es estudiar si en el caso de existir un registro de velocidad del viento elevada, el modelo ha sido capaz de predecir dichas alarmas con una probabilidad suficiente (P>0.50), quizá no instantáneamente, pero sí en horas adyacentes o a lo largo del día.

Como parte del estudio del pronóstico de los modelos finalmente seleccionados se desarrollarán también tablas de contingencia que permitirán estudiar el comportamiento del modelo de forma instantánea, para todo el periodo en estudio. Estas tablas mostrarán el número de observaciones pertenecientes a cada intervalo de velocidades (indicado en la parte superior de la tabla), y la probabilidad con la que los modelos pronostican de forma instantánea dichas observaciones.

Además, y para comprobar si los modelos han sido capaces de pronosticar velocidades elevadas, si no instantáneamente, en al menos una hora a lo largo de cada día en que se han dado dichas velocidades, se realizarán tablas que incluyan el número de días en que se han dado un número determinado de observaciones que superan el umbral de velocidad en estudio (indicado en la parte superior de la tabla), así como la probabilidad con que el modelo ha pronosticado la superación de dicho umbral.

En el estudio del comportamiento de los modelos, y en términos de observaciones problemáticas, se diferenciará entre dos tipos de "alarmas".

- *Casos de alarma*→ Datos reales registrados, con velocidades superiores a 40 km/h. En caso de no parar las obras, estas situaciones llevarían a la dispersión de las partículas contaminantes, lo cual se traduciría en un elevado coste ambiental.
- *Casos de falsa alarma*→ Casos en los que el modelo predice velocidades elevadas de viento, cuando en realidad no se observan dichas velocidades. Esto supondría una parada de obras sin un riesgo real de dispersión de la contaminación, y por tanto, pérdidas económicas.

3.6 SELECCIÓN DEL MODELO ÓPTIMO Y ESTUDIO OPERATIVO

El modelo final óptimo se seleccionará según los resultados obtenidos en el apartado anterior.

Para finalizar el trabajo, se realizará un estudio operativo del modelo óptimo seleccionado. Este estudio consistirá en primer lugar, en una comparación gráfica entre las observaciones y las predicciones instantáneas, tanto las ajustadas por el modelo seleccionado, como las predichas por el modelo de AEMET existente.

También se implementará una utilidad en R para utilizar el modelo estadístico de predicción. Para ello se desarrollará un *script* en R que permitirá obtener pronósticos diarios (en forma de datos instantáneos y gráficos de probabilidad), mediante la lectura de una hoja Excel (en formato .csv) a la cual se introduzcan nuevas predicciones de AEMET.

4. RESULTADOS

En este apartado se van a exponer los resultados obtenidos, para el análisis exploratorio de datos (ver apartado 2), y tras la aplicación de la metodología de construcción de MR candidatos a ser el modelo estadístico óptimo de predicción de la velocidad del viento, para el área de Bailín y el periodo de verano, desarrollada en el apartado 3.

4.1 ANÁLISIS EXPLORATORIO

Los análisis exploratorios a que se han sometido los datos disponibles (reales y predichos) han sido los siguientes, siguiendo con lo expuesto en el apartado 2:

1. *Análisis, pretratamiento y selección de la BD observados más adecuada:*
    a. Pretratamiento de las distintas BD de velocidad real de que se disponen.
    b. Selección de la BD más adecuada, según el objetivo de este TFG.

2. *Estudio del régimen de vientos:*
    a. Estudio de la adecuación del modelo de AEMET mediante análisis simultáneos (gráficos de dispersión) y rosas de vientos.
    b. Estudio de la existencia de estacionalidad y ciclo diario:
        i. Estudio del ciclo diario, mediante gráficos de variación de la velocidad media diaria y rosas de vientos para horarios diurnos/nocturnos.
        ii. Estudio de la estacionalidad mediante gráficos de variación de la velocidad media mensual, para el periodo en estudio.

3. *Análisis de la correlación, autocorrelación y correlación cruzada* entre pares de variables (velocidad del viento real y predicha, dirección del viento real y predicha).

4.1.1 ANÁLISIS, PRETRATAMIENTO Y SELECCIÓN DE LA BASE DE DATOS

Como se ha expuesto en el apartado 2.1.2, se han pretratado las BD reales disponibles, con los procedimientos 1 (P1) y 2 (P2), obteniendo BD.med1, BD.med2, BD.max1 y BD.max2; dichas bases de datos se han comparado con BD.p mediante el uso de histogramas que contienen las discrepancias entre ellas, y mediante el estudio del porcentaje de discrepancias de ± 5 km/h y de ± 10 km/h.

Recordemos que la hipótesis inicial dice que BD.p se ajustará mejor a los datos medios, dada la hipótesis de suavizado del modelo de predicciones de AEMET. Por ello, el procedimiento idóneo se obtendrá de la comparación entre las bases de datos medios y BD.p.

Los histogramas producto de la comparación entre BD.med1 y BD.med2 con BD.p, recogidos en la Figura 10 izda. (para BD.med1) y Figura 10 dcha. (para BD.med2), muestran una distribución de frecuencias similar y asimétrica, con valores negativos de la media y la mediana, con distribución ampliada hacia la izquierda, indicando que las predicciones son mayores a las

observaciones. Los valores de media y mediana son inferiores para BD.med2, indicando un mejor ajuste de esta base de datos a BD.p (ver Figura 10 dcha.).

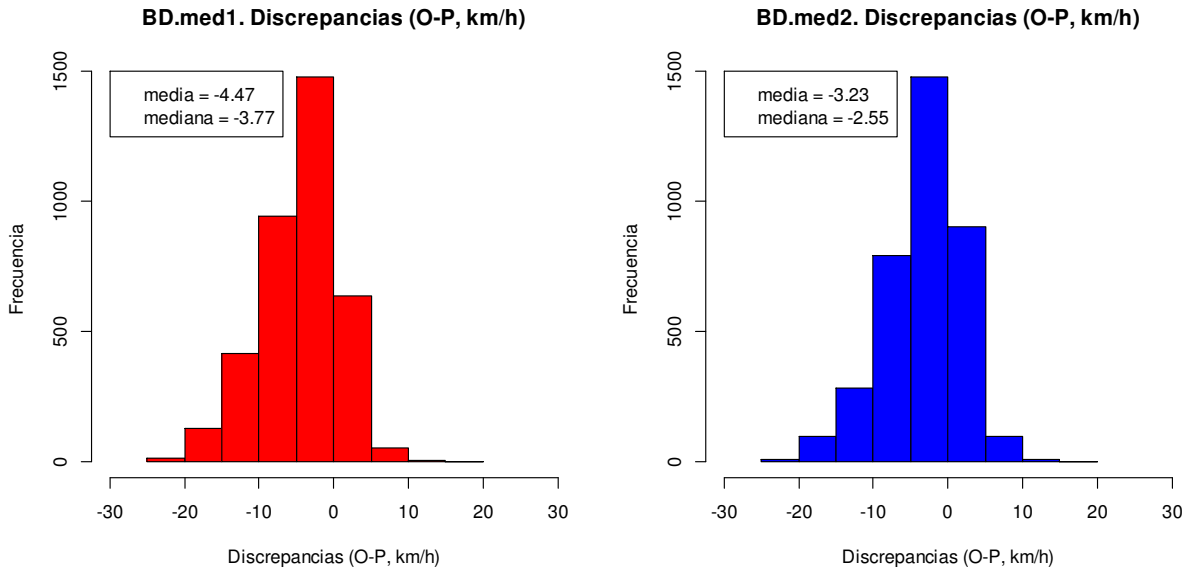


Figura 10 Histogramas que muestran las discrepancias, en km/h, entre las bases de datos observados de velocidad de viento media (sometidas a los pretratamientos P1 y P2) y la base de predicciones de AEMET, BD.p. En rojo se muestran los resultados para P1, en azul para P2.

El porcentaje de discrepancias de ± 5 km/h y de ± 10 km/h obtenido es visiblemente inferior al aplicar P2, 35.21% y 10.92 % respectivamente, que cuando se aplica P1, 42.40% y 15.25% respectivamente (ver Tabla A. 1). Por tanto, se puede suponer que el modelo de predicciones se ajusta mejor a las velocidades instantáneas (procedentes de aplicar P2), probablemente porque recogen una mayor variabilidad del comportamiento real de viento y porque además, se ajustan mejor a las características de las predicciones, que también son datos instantáneos centrados alrededor de la hora en punto.

El mismo estudio se ha realizado entre BD.max1, BD.max2 y BD.p. En este caso, las predicciones se ajustarán peor a los datos observados que para el caso anterior.

Los histogramas de discrepancias, recogidos en la Figura 11 izda. para BD.max1 y en la Figura 11 dcha. para BD.max2, muestran de nuevo una distribución asimétrica, en este caso ampliada hacia la derecha, con valores positivos de la media y la mediana, especialmente para BD.max2. El porcentaje de discrepancias más allá de ± 5 km/h y de ± 10 km/h, es inferior al aplicar P1, 52.75% y 17.78 % respectivamente, que cuando se aplica P2, 61.25% y 24.18% (ver Tabla A.2).

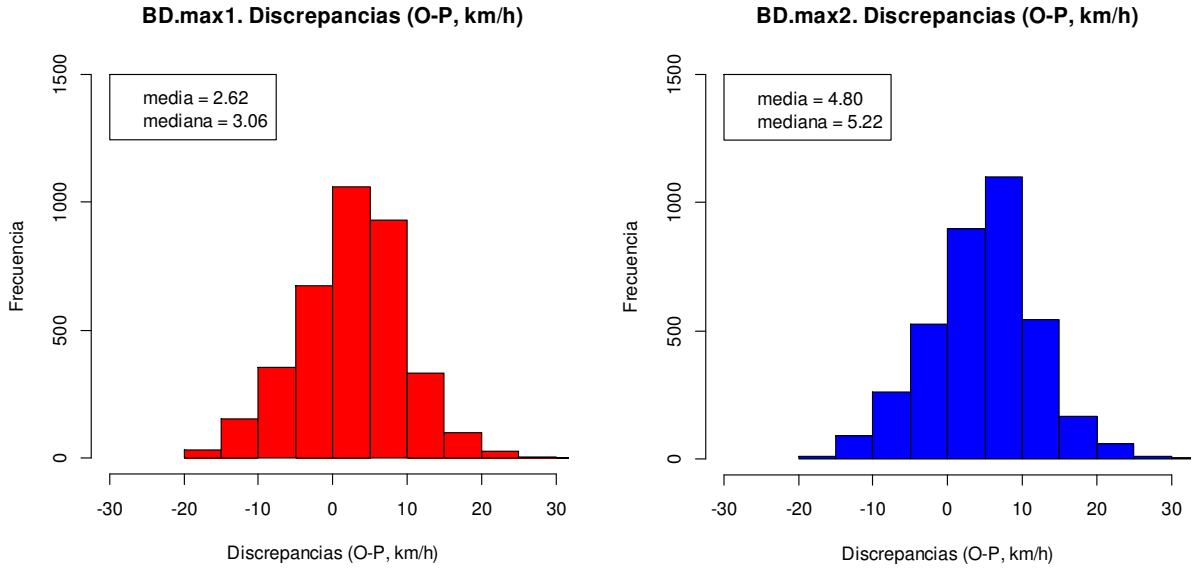


Figura 11 Como Figura 10, para la comparación de bases de datos observados de velocidad máxima

En resumen, se ha seleccionado como mejor pretratamiento a P2, por recoger una mayor variabilidad del régimen real de vientos, y por ajustarse mejor a la base de datos medios. Sin embargo, ya que los datos de interés para nuestro estudio son los datos máximos de velocidad, se trabajará con BD.max2. Los elevados porcentajes de discrepancias obtenidos para el estudio de la base de datos máximos, indican que el modelo de predicciones AEMET está suavizado y no es capaz de reproducir correctamente el régimen de vientos máximos observado.

4.1.2 ESTUDIO DEL RÉGIMEN DE VIENTOS EN BAILÍN Y DE LA ADECUACIÓN DEL MODELO DE PREDICCIONES DE AEMET

El análisis del régimen de vientos para el área de Bailín se ha realizado siguiendo lo expuesto en el apartado 2.2., donde los principales objetivos son el estudio de la adecuación del modelo de AEMET, así como el estudio del comportamiento del viento en términos de estacionalidad y ciclos diarios, dado que se trabaja con una serie temporal.

ADECUACIÓN DEL MODELO DE AEMET

- **Análisis gráfico de la relación simultánea**

Como primer paso orientativo en el estudio de la adecuación entre las predicciones de AEMET y el régimen real de vientos, se ha realizado un análisis gráfico de la dispersión entre los datos simultáneos de ambas BD (es decir, datos correspondientes para los mismos instantes de tiempo). Mediante este análisis se quiere comprobar si entre ellas existe algún tipo de relación, lineal o no lineal.

El gráfico de dispersión entre ambas covariables, recogido en la Figura 12, muestra las siguientes características:

- Hay situaciones de velocidad real del viento nula o muy reducida, para las que sus predicciones instantáneas llegan hasta los 25 km/h (círculo verde en el gráfico); como situación opuesta, existen datos reales de hasta 35 km/h cuyas predicciones son frecuentemente inferiores a 10 km/h (ver círculo rojo en gráfico).
- Solo se observan valores de predicción de hasta ≈30 km/h; sin embargo, las velocidades reales del viento llegan a alcanzar los ≈60 km/h. Por tanto, se descarta el uso directo de las predicciones de AEMET, justificando la realización de este TFG y demostrando que el modelo de predicciones de AEMET está suavizado.
- A pesar de ello, se observa una relación lineal en la cual los datos de predicción aumentan conforme aumentan los datos reales (ver recta de regresión en el gráfico).

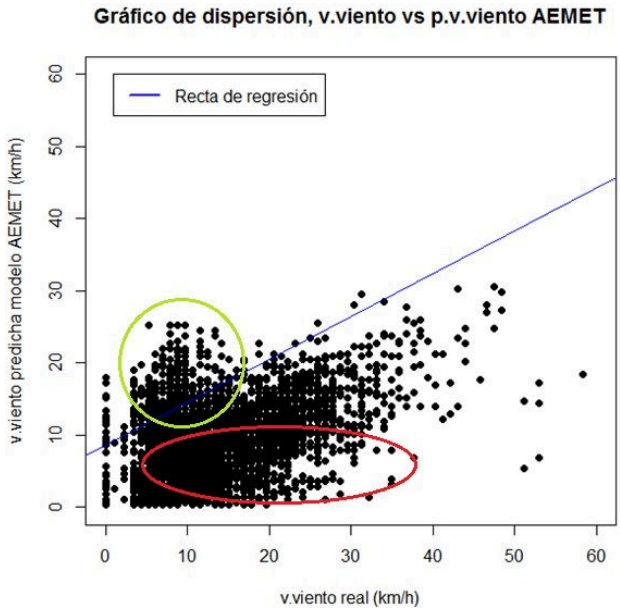


Figura 12. Gráfico de dispersión entre datos de BD.max2 y BD.p. En azul se ha dibujado la recta de regresión obtenida para el MRLS desarrollado entre los datos de predicción de AEMET (predictor) y los datos reales de velocidad (respuesta). Dicho modelo se calculará en el apartado 4.2.

- **Estudio mediante rosas de vientos**

La adecuación del modelo de predicciones de AEMET al régimen real de vientos de Bailín se ha estudiado también mediante la realización de rosas de vientos para BD.max2 y BD.p. Como se ha explicado en el apartado 2.2, se han usado percentiles, que han permitido tener en cuenta el suavizado de la BD.p.

Se ha tomado v=30 km/h como umbral orientativo de intensidad elevada de viento. Este umbral corresponde al percentil 96.1, que a su vez corresponde a una velocidad predicha de 19.44 km/h. En este caso se ha decidido desechar el uso de 40 km/h como umbral, ya que el número de datos con los que trabajar sería demasiado reducido.

Los intervalos de velocidad seleccionados para estudiar el régimen de vientos, han sido los correspondientes a los percentiles entre 0 y 50 (para situaciones de *calma*), entre 50 y 96.1 (para situaciones de *intensidad media de viento*) y entre 96.1 y 100 (para situaciones de *intensidad de viento elevada*). El percentil 50 corresponde 12.42 km/h para velocidades de viento observadas y 9.36 km/h para velocidades predichas.

En las rosas de vientos, la escala de velocidad de viento estudiada ha ido desde 0 y hasta el percentil 96.1 correspondiente a cada BD, en intervalos de 5 km/h para los datos reales y de 3.24 km/h para los predichos; de esta forma ambas tienen el mismo número de intervalos y el mismo código de color y son comparables gráficamente.

La comparación por intervalos se ha hecho para las llamadas *situaciones 1 y 2*. Recordemos que la *situación 1* estudiaba los datos reales para cada intervalo y sus predicciones instantáneas correspondientes, mientras la *situación 2* estudiaba los datos predichos pertenecientes a cada intervalo y las observaciones instantáneas correspondientes.

En la Figura 13 se recogen las rosas de vientos elaboradas para el estudio de la *situación 1,* y para los intervalos de velocidad del viento correspondientes a intensidades altas y medias. Para el intervalo de intensidades elevadas de viento, los datos reales muestran una clara componente WNW (ver Figura 13a), al igual que sus predicciones instantáneas, las cuales en general muestran velocidades elevadas (ver Figura 13b).

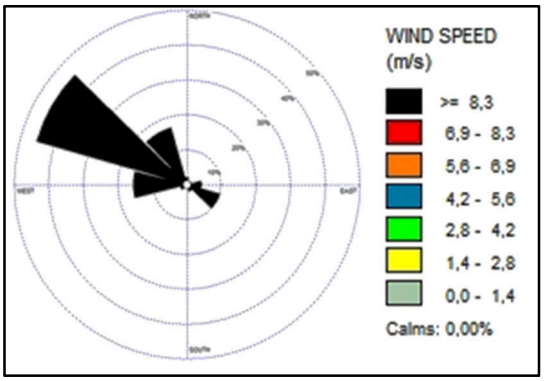


Figura 13a

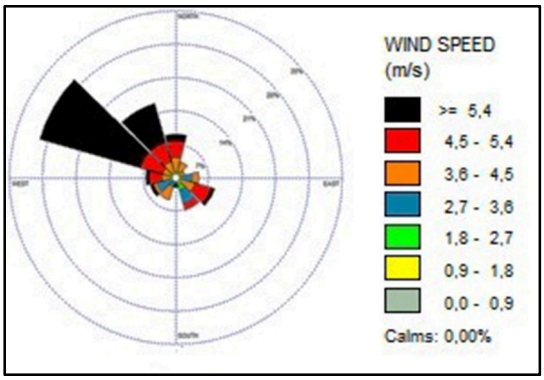


Figura 13b

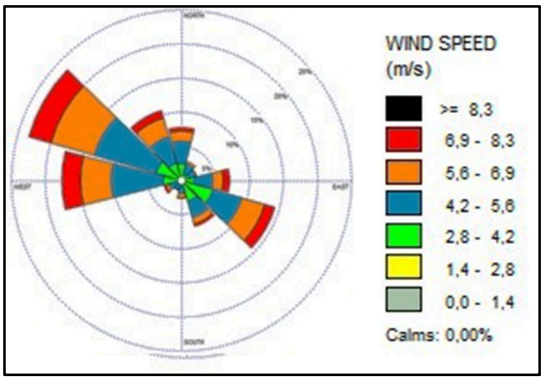


Figura 13c

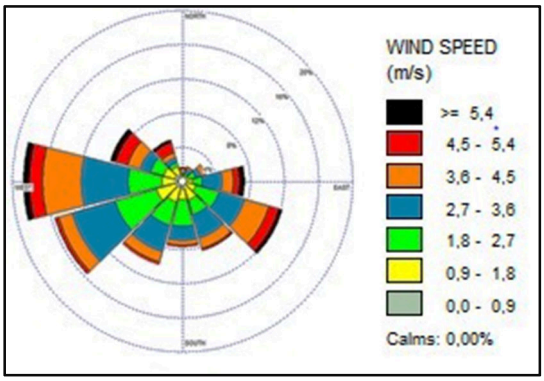


Figura 13d

Figura 13. Rosas de vientos correspondientes al estudio de la *situación 1* para intervalos de intensidades elevadas y medias de viento. La Figura 13a. corresponde a datos observados en el intervalo de intensidad de viento elevada y la Figura 13b muestra sus datos de predicción correspondientes. La Figura 13c muestra datos observados en el intervalo de intensidad de viento media, mientras la Figura 13d. recoge sus datos de predicción correspondientes.

Para intensidades de viento medias, los datos reales siguen mostrando una componente WNW predominante, aunque aparecen nuevas componentes, en especial W y ESE (ver Figura 13c). Por

su parte, los datos de predicción correspondientes muestran una componente W predominante, además de muchas otras que no se observan en el régimen real de vientos (ver Figura 13d). Se predicen velocidades más altas de las que realmente se observan.

Para situaciones de calma, el régimen real de vientos cambia radicalmente de componente principal, pasando a ser SSE (ver Figura A. 3a). Sin embargo, se observan velocidades muy elevadas para la predicción instantánea correspondiente (ver Figura A. 3b). La componente principal de la predicción de la dirección también es diferente en este caso, destacando la componente E.

En el estudio de la *situación 2*, recogido en la Figura 14, se observa que para intensidades elevadas de viento, las velocidades predichas muestran una componente E mayoritaria, siendo importante también la componente WNW (ver Figura 14a). Los datos reales correspondientes muestran, al igual que en la *situación 1* para el mismo intervalo, una marcada componente WNW, aunque con un mayor rango de velocidades (Figura 14b). No se observa la componente E en las velocidades reales.

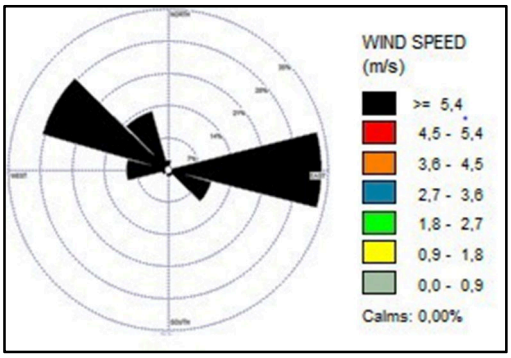


Figura 14a

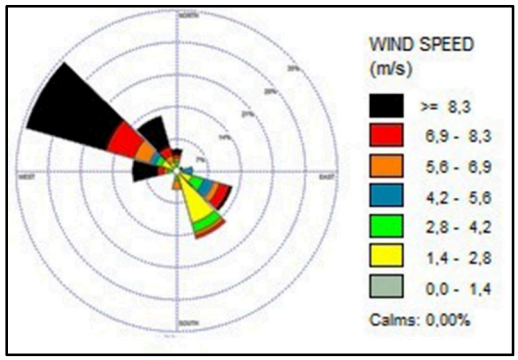


Figura 14b.

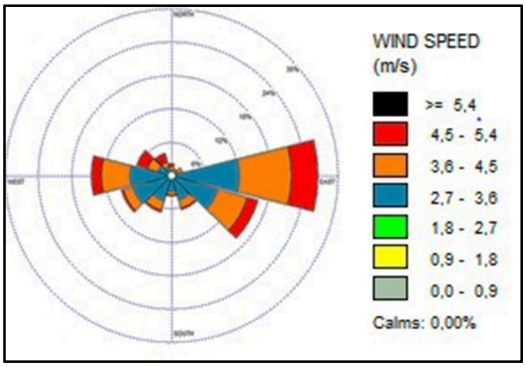


Figura 14c.

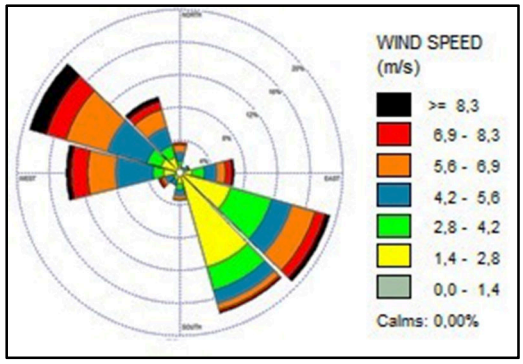


Figura 14d.

Figura 14. Como Figura 13, para estudio de rosas de vientos correspondientes a *situación 2*. En este caso Figuras 14 a y c muestran predicciones, y Figuras 14 b y d observaciones correspondientes

Para intensidades medias de viento, la dirección preferente del modelo de predicciones de AEMET es E de nuevo (Figura 14c), mientras que los datos reales correspondientes muestran una clara dirección WNW y SSE (Figura 14d). Se observa que el modelo de predicciones de AEMET, aparte de no poder capturar el comportamiento del régimen real de vientos en términos

de dirección, tampoco predice las velocidades elevadas que se observan en realidad para los mismos instantes de tiempo.

Para el intervalo de calmas, se observan predicciones para prácticamente todas las direcciones contenidas en la rosa de vientos (Figura A. 4a), mientras las observaciones correspondientes instantáneas muestran una dirección preferente SSE (Figura A. 4b), al igual que ocurría en la *situación 1* para el mismo intervalo de vientos. También se observa que cuando el modelo de AEMET predice situaciones de calma, las observaciones reales de velocidad de viento no se encuentran necesariamente en dicho intervalo, sino que muestran un rango mucho más amplio, llegando a observarse velocidades de viento muy elevadas que como decimos, no quedan reflejadas en las predicciones de AEMET.

En general, se han encontrado grandes diferencias entre las rosas de vientos de datos observados y datos instantáneos predichos por AEMET, para cada intervalo y situación en estudio. Por tanto, de nuevo se puede concluir que el modelo de predicciones de AEMET no se adecua correctamente al régimen real de vientos.

ESTUDIO DE LA ESTACIONALIDAD Y DEL CICLO DIARIO DEL RÉGIMEN DE VIENTOS

Siguiendo con la metodología desarrollada en el apartado 2.2, se ha realizado un estudio de la estacionalidad y del ciclo diario de las BD en estudio. Se ha descartado profundizar en el estudio de la tendencia de la BD.max2, ya que un análisis previo de tipo descomposición de serie de datos, no ha arrojado resultados significativos sobre la existencia de dicha tendencia (ver Figura A. 43 en Anexo III, sobre estudios adicionales).

- **Estudio del ciclo diario**

El comportamiento del ciclo diario se ha estudiado mediante un gráfico de variación horaria de las velocidades medias durante el periodo en estudio, reales y predichas, y mediante la comparación de las rosas de los vientos para los horarios diurnos (entre 8 y 19 horas) y nocturnos (entre 20 y 7 horas).

La Figura 15 muestra el gráfico de variación horaria de las velocidades medias. En él se observan claras diferencias entre datos reales y predichos: el régimen real de velocidad muestra la existencia de un ciclo diario, el cual tiene un máximo alrededor de las 15 horas (máximo que también se observa para el modelo de predicciones) y disminuye, siguiendo una forma sinusoidal hasta alcanzar un mínimo a las 8 de la mañana. Se observa un ciclo menos marcado para el modelo de predicciones de AEMET, manteniéndose el nivel medio más homogéneo, con un rango en la velocidad media (entre los valores mínimo y máximo) de hasta 6 km/h; este rango aumenta hasta los 16 km/h para los datos reales (ver Figura 15).

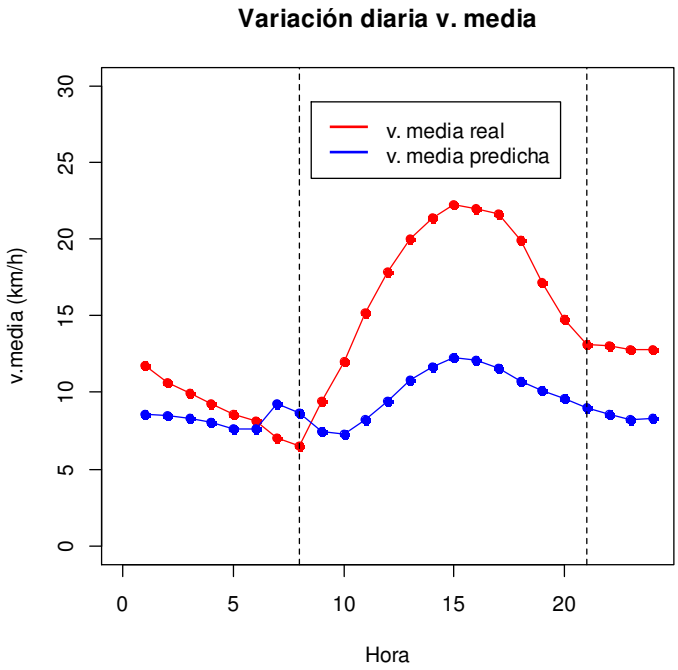


Figura 15. Estudio de la variación de las velocidades medias real y predicha a lo largo del día. Las líneas discontinuas indican los puntos de cambio del ciclo diario existente en el régimen real de velocidades.

Para el horario diurno, las rosas de los vientos muestran cómo el régimen real de vientos tiene una clara componente WNW y una ligera componente ESE (ver Figura 16 izda.), con velocidades más o menos elevadas. Por su parte, el modelo de predicciones de AEMET muestra un régimen de vientos con predominancia de la componente W (ver Figura 16 dcha.), así como una distribución similar de frecuencias de velocidad de viento.

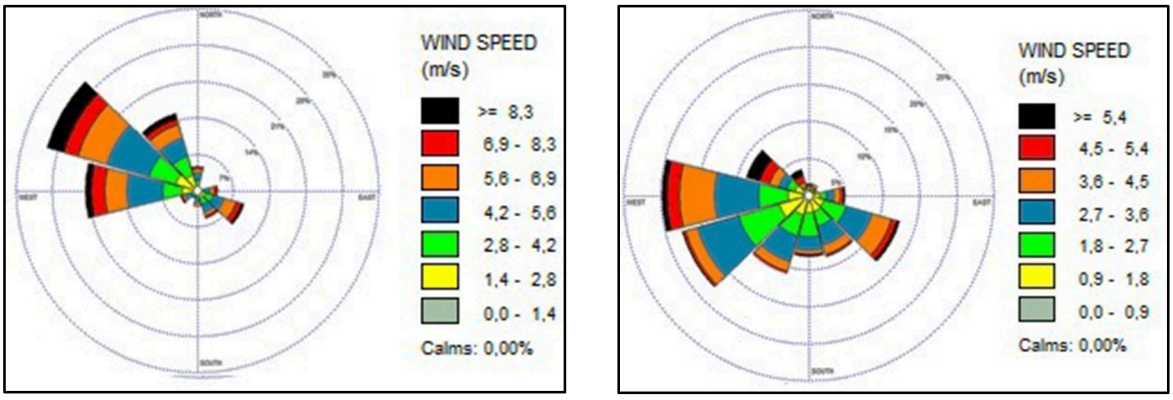


Figura 16. Rosas de vientos realizadas con datos correspondientes al horario diurno. A la izda. Se muestra la rosa para BD.max2 y a la dcha., la correspondiente a BD.p, cada una con una escala adecuada.

La distribución de frecuencias de BD.max2 para el horario diurno es de tipo campana de Gauss (ver Figura A. 5a), con una frecuencia máxima del 28% para el intervalo de 15-20 km/h. Las velocidades superiores a 20 km/h suponen un 34.5% en total, siendo superiores a 30 km/h un 6.8%. La distribución de las frecuencias BD.p para el mismo horario (ver Figura A. 5.b) tiene una forma similar, pero se encuentra desplazada hacia la izquierda, mostrando el suavizado del

predicciones de AEMET con respecto a los datos observados. En este caso, los porcentajes máximos de frecuencias están concentrados en los intervalos entre 5-10 km/h (34.2%) y 10-15 km/h (35.7%). El intervalo de velocidades de 15-20 km/h agrupan un 11.4% de los datos, mientras que solo un 3.7% de los datos son superiores a 20 km/h, y un 0.7% a 30 km/h. Es decir, para velocidades superiores a 20 km/h, las frecuencias del modelo de predicciones son aproximadamente una décima parte de las realmente observadas (ver Tabla 2).

HORARIO DIURNO	**FRECUENCIA**		**HORARIO NOCTURNO**	**FRECUENCIA**	
Intervalo velocidad (km/h)	**Observado**	**Predicho**	***Intervalo velocidad (km/h)***	**Observado**	**Predicho**
5 a 10	-	34.2%	5 a 10	51.7%	27.1%
10 a 15	-	35.7%	10 a 15	24.7%	25.2%
15 a 20	28%	11.4%	-	-	-
> 20	34.5%	3.7%	>15	14.7%	14.4%
> 30	6.8%	0.7%	> 30	1%	-

Tabla 2. Estudio del porcentaje de frecuencias de acumulación de datos de velocidad del viento para cada intervalo de viento estudiado, horarios diurno (izda.) y nocturno (dcha.). El símbolo "-" indica que los datos para ese intervalo no son relevantes para el estudio.

El comportamiento real del viento durante la noche cambia radicalmente, siendo la dirección preferente SSE, y registrando bajas velocidades en comparación con las observadas durante el día (ver Figura 17 izda.). El modelo de predicciones de AEMET muestra un régimen de vientos con una marcada componente E (Figura 17 dcha.), y rangos de velocidades mucho mayores de los que se dan realmente.

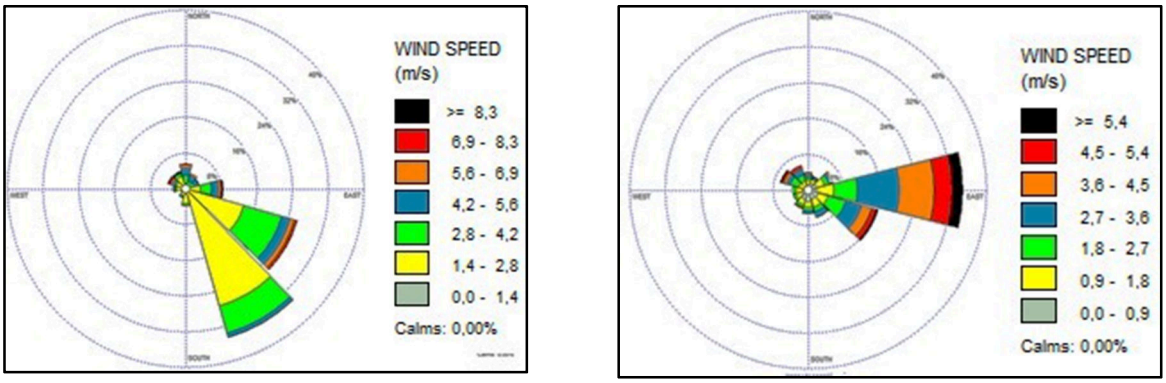


Figura 17. Rosas de vientos realizadas con datos correspondientes al horario nocturno. A la izda. Se recoge la rosa de vientos para BD.max2; a la dcha., la correspondiente a BD.p.

Para el horario nocturno, la distribución de frecuencias de intervalos de velocidad de viento cambia respecto al diurno, siendo en este caso distinta para ambas BD (ver Figura A. 5 c y d). Para BD.max2 (ver Tabla 2), un 51.7% de las observaciones pertenecen al intervalo de 5-10 km/h, un 24.7% de 10-15 km/h, y un 14.7% entre 15-30 km/h; únicamente un 1% de las observaciones superan los 30 km/h en comparación con el 6.8% obtenido para el horario diurno. Las predicciones instantáneas muestran una distribución más homogénea de frecuencias (ver Figura A. 5.d), 27.1% para 5-10 km/h, 25.2% para 10-15 km/h y un total de 14.4% para velocidades superiores a 15 km/h (ver Tabla 2).

En resumen, para el horario nocturno y en cuanto a la distribución de la velocidad, el modelo de predicciones de AEMET se comporta correctamente para velocidades superiores a 10 km/h, probablemente dada la menor variabilidad de las observaciones registradas durante la noche; para el horario diurno y para velocidades observadas superiores a 20 km/h, el modelo de predicciones de AEMET recoge únicamente un ≈10% de ellas.

- **Estudio de la estacionalidad**

La estacionalidad se ha estudiado mediante gráficos de variación de la velocidad media a lo largo del periodo en estudio, y de dicha variación para distintos cuartiles.

En el gráfico de variación de la velocidad media mensual, recogido en la Figura 18, se observa que la evolución que siguen las medias de los datos de velocidad real y predicha es muy diferente: la velocidad media real varía de forma estable a lo largo del periodo, siguiendo una curva que tiene un máximo en agosto para posteriormente descender, mientras que la velocidad media predicha para cada mes distingue dos mínimos relativos en julio y septiembre.

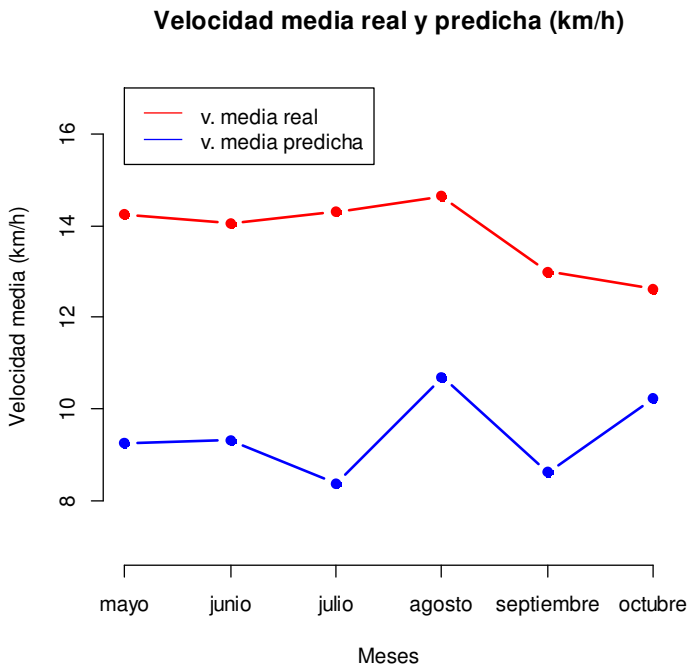


Figura 18. Estudio de la variación de los datos medios de velocidad real y predicha para cada mes del periodo entre mayo y octubre.

Esta diferencia aparente entre la velocidad de viento real y predicha, también aparece en el estudio de los distintos cuartiles para ambas BD. En la Figura 19 izda. se observa que la tendencia que siguen los percentiles 0.25 y 0.50 de las velocidades medias es bastante similar, aunque no lo es para el percentil 0.75. Por el contrario, si estudiamos la tendencia de los cuartiles para la velocidad predicha (ver Figura 19 dcha.) todos ellos muestran una evolución casi paralela, simetría que no se observa en el gráfico de cuartiles de velocidad real.

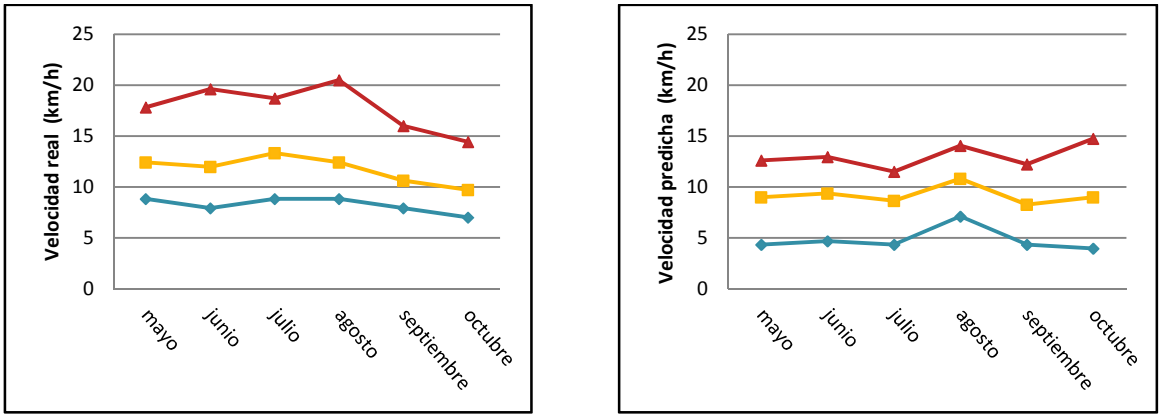


Figura 19. Estudio de las velocidades real (izda.) y predicha (dcha) correspondientes a los 3 cuartiles (0.25, 0.5 y 0.75).

4.1.3 ANÁLISIS DE CORRELACIÓN

Como último paso del análisis exploratorio de datos, se ha estudiado la correlación entre las variables reales y predichas según lo especificado en el apartado 2.3. El fin de este análisis es valorar la existencia de una relación lineal entre las variables en estudio, así como comprobar si existe algún tipo de retardo en dicha relación, es decir, si los datos de predicción podrían estar desfasados por algún motivo.

- **Velocidad del viento real y predicha**

Para la pareja de covariables formada por la velocidad del viento real y la predicha, el coeficiente de correlación de Pearson hallado es de 0.41 y es significativa a nivel 0.05.

Atendiendo a la Figura 20, que muestra el gráfico de correlación cruzada entre ambas covariables, vemos que el valor máximo de correlación corresponde al instante 0 (*lag=0* en el gráfico), estableciendo que no hay un desfase sistemático entre las covariables a lo largo del tiempo. El comportamiento de la correlación entre ambas covariables sigue una forma sinusoidal, con un ciclo completo de aproximadamente 24 horas, tras el cual la correlación cruzada es de nuevo máxima y positiva (ver línea verde en el gráfico), dividido en ciclos de 12 horas en los cuales la correlación pasa a tener signo negativo (línea discontinua roja en el gráfico). Esta situación expresa el ciclo diario de brisas existente (la diferencia en la velocidad del viento entre el día y la noche, ya observada en análisis anteriores), que se refleja en ambas series de datos, aunque menos nítidamente en la predicha (ver Figura 21).

Los gráficos de autocorrelación de las variables muestran cómo la velocidad real del viento cambia el signo de su autocorrelación cada 12 horas, retomando el signo en ciclos de 24 horas (ver Figura 21.izda). Sin embargo, las predicciones de AEMET de velocidad del viento muestran un ciclo completo cada 24 horas, pero no un cambio en el signo de su autocorrelación (ver Figura 21.dcha).

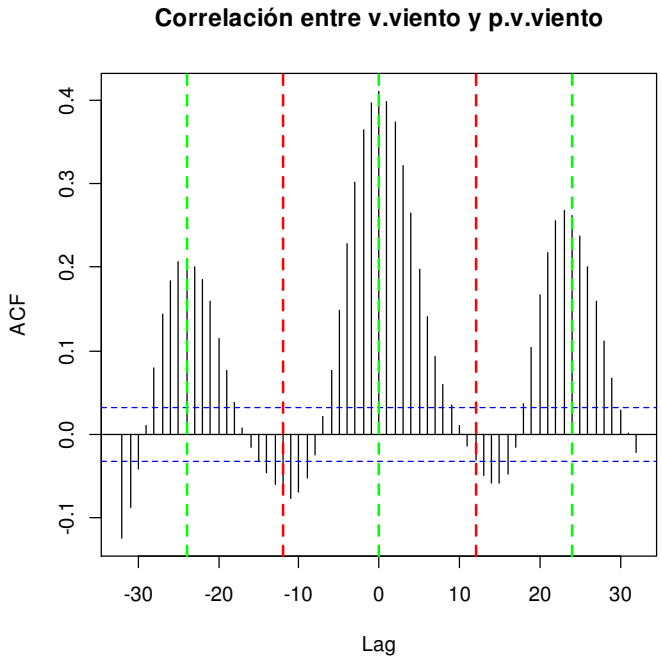


Figura 20. Gráfico de correlación cruzada entre velocidad del viento real y predicha.

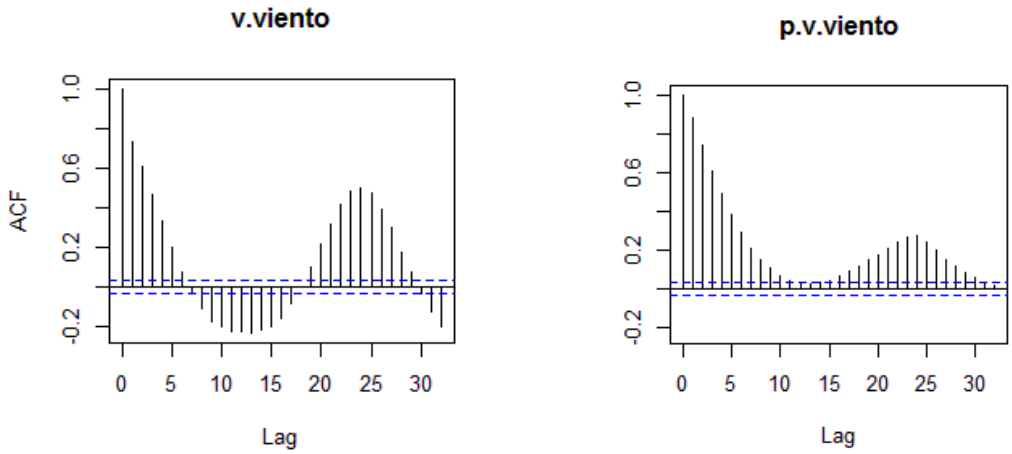


Figura 21. Conjunto de correlogramas para variables de velocidad del viento real (*v.viento,* izda.) y predicha (*p.v.viento,* dcha.).

- **Dirección del viento real y predicha**

Se ha realizado el mismo estudio comparando las variables de dirección del viento real y su predicción correspondiente. En este caso, aunque se trata de una variable circular, la rosa de vientos reales ha indicado la ausencia de viento de componente Norte (ver Figura 13), por lo que se ha concluido que se puede trabajar con la correlación como una herramienta descriptiva de la relación.

El coeficiente de correlación de Pearson hallado para ambas variables es de tan solo un 0.26.

El gráfico de correlación cruzada de ambas variables (ver Figura 22), muestra un máximo en lag=-2 (≈0.35), es decir, con un desfase horario de 2 horas. Al igual que ocurría con la velocidad

del viento, el gráfico tiene una clara forma sinusoidal que cambia la tendencia de la correlación en ciclos de aproximadamente 12 horas, retomando el signo positivo en ciclos de 24 horas.

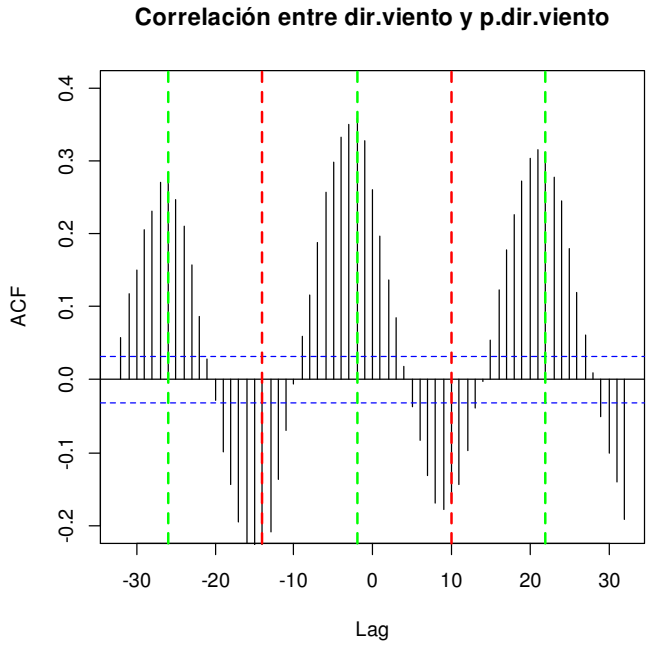


Figura 22. Gráfico de correlación cruzada entre dirección del viento real y predicción de la dirección del viento .

Si se comparan los autocorrelogramas para ambas covariables, se observa que la dirección real muestra un patrón de giro del signo de la autocorrelación cada 12 horas (ver Figura 23.izda), mientras que las predicciones son capaces de mostrar autocorrelación nula en el retardo 12, pero no negativa (ver Figura 23 dcha.).

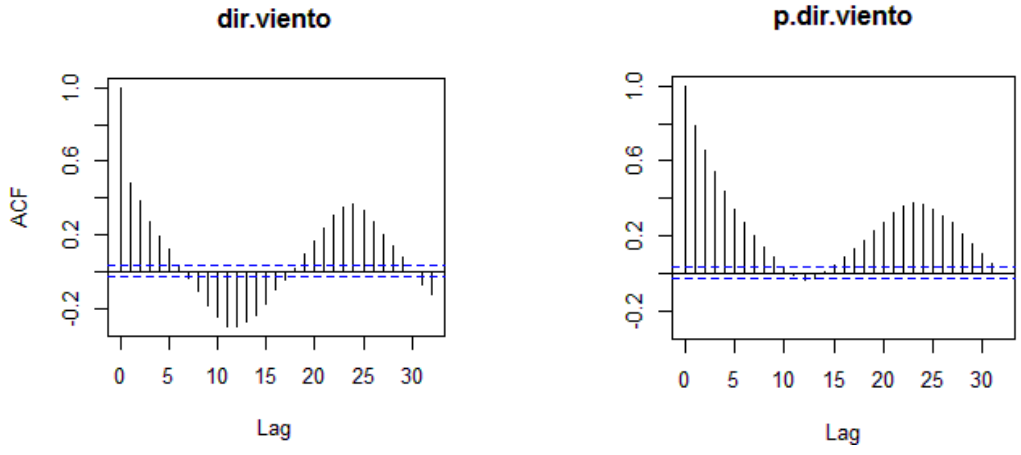


Figura 23. Conjunto de gráficos de autocorrelación para variables de dirección real del viento (*dir.viento,* izda.) y predicha (*p.dir.viento,* dcha.)

Los resultados obtenidos en el análisis de correlación de nuevo muestran la existencia de un régimen de brisas diferenciado para el horario diurno y nocturno, el cual el modelo de predicciones no es capaz de reflejar satisfactoriamente.

4.1.4 PRINCIPALES CONCLUSIONES DEL ANÁLISIS EXPLORATORIO

Los análisis exploratorios realizados hasta el momento sugieren descartar el uso directo del modelo de predicciones de AEMET, por las carencias que dicho modelo ha mostrado para el área del vertedero de Bailín y el periodo de verano.

Será necesario por tanto, desarrollar un nuevo modelo estadístico de predicciones de la velocidad del viento. Dicho modelo deberá ser capaz de reescalar las predicciones de AEMET, incluyendo componentes que reflejen la estacionalidad y el ciclo diario existente en el régimen real de vientos, características ausentes en las predicciones de AEMET. Además, el nuevo modelo deberá incluir términos de interacción entre las covariables anteriores, dada la relación observada entre por ejemplo, la velocidad y la dirección del viento, la velocidad del viento y el ciclo diario de brisas, o la velocidad del viento y la estacionalidad.

En el análisis de la correlación se ha observado una situación en la cual ésta es mayor para una diferencia temporal de -2 horas (entre las direcciones de viento reales y predichas). Descartando que esta diferencia pudiese deberse a que los datos de predicción se diesen de forma decalada, surgió la idea de crear "nuevas" covariables a incluir en el modelo, a partir de los datos de predicción de AEMET (tanto de velocidad como de dirección), en una ventana temporal de 4 horas alrededor del instante de interés estudiado (t_0-4,t_0+4), ya que en dicho intervalo la correlación cruzada se mantiene superior a 0.2, tanto para la velocidad como para la dirección.

Por último, en este nuevo modelo estadístico de predicciones deberá estudiarse la conveniencia de la inclusión del mes de octubre, ya que se ha observado que posee un régimen de vientos más homogéneo que el resto de periodo en estudio.

4.2 DESARROLLO DE MODELOS DE REGRESIÓN

Según lo expuesto en el apartado 3.1 de la metodología, el tipo de modelos seleccionados para realizar la adaptación (*downscaling*) de las predicciones de AEMET al periodo y área de interés, han sido los modelos de regresión (MR).

Los principales pasos seguidos para la construcción de los modelos han sido los siguientes:

1. Estudio de la capacidad para distintos periodos, del modelo de regresión lineal simple (MRLS) construido a partir de las predicciones de AEMET.
2. Desarrollo de MR candidatos a ser el modelo estadístico óptimo buscado.
3. Selección de modelos candidatos.
4. Adecuación de los modelos seleccionados.
5. Estudio del funcionamiento operativo de los modelos seleccionados. Pronóstico.

4.2.1 ESTIMACIÓN DEL MODELO DE REGRESIÓN LINEAL SIMPLE PARA LOS DATOS DE PREDICCIÓN EXISTENTES

En primer lugar se ha realizado un MRLS, entre los datos de predicción AEMET de la velocidad del viento y las observaciones reales instantáneas, correspondientes a BD.max2, seleccionando los datos correspondientes a distintos periodos:

- Año 2012 completo (BD.anual).
- Periodo de verano (periodo de interés), comprendido entre los meses de mayo y septiembre (BD.verano).
- Periodo de mayo a octubre, dadas las dudas surgidas anteriormente respecto a la inclusión de octubre en el modelo.
- Mes de octubre (BD.octubre), por la misma razón.

En todos los casos, los MRLS desarrollados han mostrado que el dato de predicción de la velocidad del viento es altamente significativo a la hora de predecir la velocidad real (respuesta), pues su p-valor es inferior a 0.05 en todos los casos. El grado de ajuste de los modelos, recogido en la Tabla A. 6, ha dependido del periodo considerado, siendo la bondad de ajuste máxima para BD.octubre (R^2 ajustado de 0.51), seguida de BD.anual (R^2 ajustado de 0.31), BD de verano con octubre (R^2 ajustado de 0.23), y por último, BD.verano, con un R^2 ajustado de únicamente 0.17 (ver Tabla A.6).

Los resultados obtenidos son indicativos de que los meses no veraniegos poseen un régimen de vientos más fácilmente predecible. El bajo ajuste del modelo de predicciones de AEMET a la BD.verano muestra, en concordancia con los análisis exploratorios anteriormente realizados, que las predicciones AEMET no son capaces de reproducir correctamente el comportamiento real del viento para el periodo de verano (el periodo de interés, con más efectos locales) y por tanto, que es necesario desarrollar un nuevo modelo estadístico de predicción de la velocidad del viento, que mejore al modelo de AEMET en la medida de lo posible.

A pesar de que Fernández et al. (2012) contempla que la fase de desmantelamiento y transferencia del vertedero de Bailín puede ampliarse hasta el mes de octubre, se ha decidido no incluir dicho mes en el nuevo modelo, ya que el ajuste del modelo de predicciones de AEMET para octubre se considera suficientemente bueno. Además, los datos de octubre podrían influir en el modelo construido, limitando su capacidad predictiva para el resto de meses.

ESTUDIO DE LA ADECUACIÓN DEL MODELO DE PREDICCIONES DE AEMET PARA EL PERIODO DE VERANO

Previo al comienzo del desarrollo de un nuevo modelo para el periodo de verano, se ha realizado un análisis de la adecuación del MRLS construido para dicho periodo. El fin del estudio ha sido encontrar posibles deficiencias que ayuden en el desarrollo del nuevo modelo mejorado.

Como se ha visto en el apartado 3.3, el estudio de la adecuación del modelo se lleva a cabo a partir de la inspección de los residuos del mismo, los cuales en teoría, deben verificar las hipótesis de linealidad, varianza constante, independencia y distribución normal.

- **Estudio de la linealidad**

En la Figura 24 se observa cómo la hipótesis de media cero para los residuos parece admisible, pues la curva suavizada de medias (línea roja en el gráfico) es paralela al eje de abscisas. Además, no se observa ningún patrón característico, siendo la distribución de los residuos más o menos homogénea. Por tanto, no hay indicios de falta de linealidad.

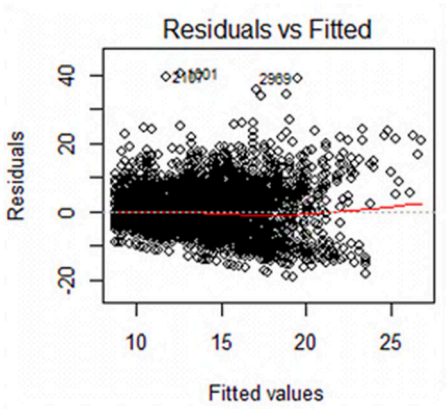


Figura 24. Gráfico *residuals vs fitted* de estudio de la linealidad del MRLS desarrollado a partir de las predicciones de AEMET para BD.verano.

- **Estudio de la distribución**

En el gráfico de probabilidad normal se observa que los residuos no se adecúan a la normal. Las desviaciones observadas se dan en la cola derecha de la distribución, y corresponden a medidas elevadas de velocidad de viento real que el MRLS en estudio no es capaz de predecir. También existen varios *outliers* para esta situación (ver Figura 25dcha.). El histograma muestra que la distribución de los residuos se amplía hacia la derecha. Este tipo de distribuciones son habituales cuando se trabaja con velocidades de viento, las cuales nunca pueden ser menores que 0, pero pueden ampliarse en esta dirección (ver Figura 25 izda.).

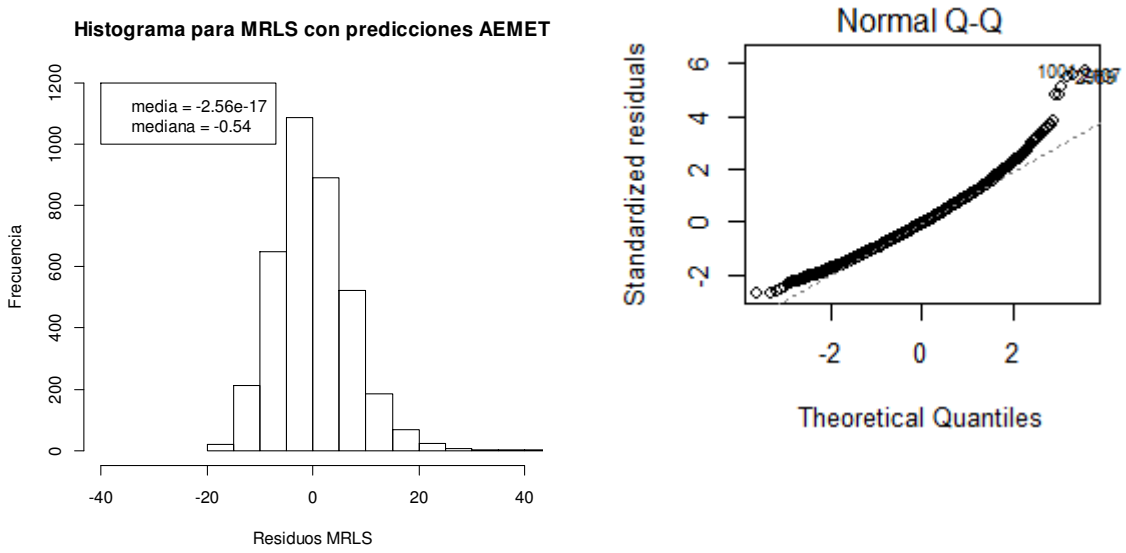


Figura 25. Histograma (izda.) y gráfico de probabilidad normal (dcha.) de residuos del MRLS construido con predicciones de AEMET.

- **Varianza constante**

El gráfico *scale-location* (ver Figura 26 izda.) muestra que no se verifica la hipótesis de varianza constante, si no que ésta presenta tendencia a incrementarse conforme lo hace el valor de la respuesta. Por tanto, para el MRLS en estudio no se verifica la hipótesis de homocedasticidad.

- **Puntos de alto leverage**

Por último, en el gráfico *Residuals vs Leverage* (ver Figura 26 dcha.) no se observan puntos especialmente influyentes y que por tanto, deban eliminarse de la BD.

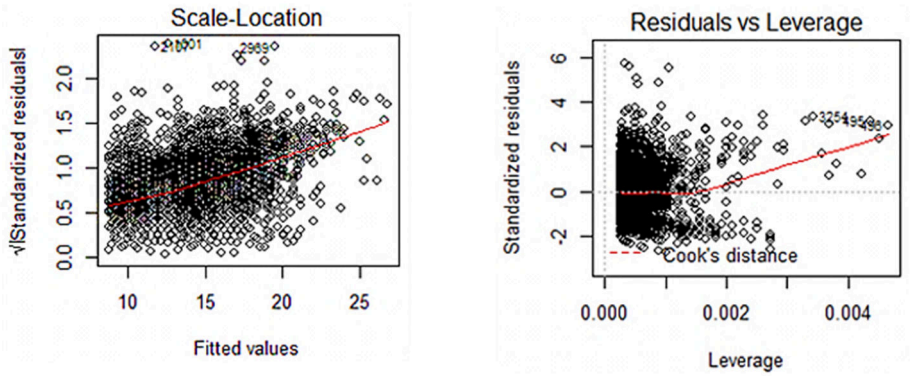


Figura 26. Gráficos *scale-location*, de estudio de la varianza (izda.), y *residuals vs leverage*, de estudio de puntos influyentes (dcha.) para los residuos del MRLS construido a partir de las predicciones de AEMET.

En resumen, para el MRLS desarrollado, con los datos de predicción de AEMET como variable predictora y los datos reales de la BD.verano como respuesta:

- Se explica un 17% de la variabilidad de la respuesta.
- Se acepta la hipótesis de linealidad y media 0 de los residuos.
- Se rechaza la hipótesis de normalidad.
- Se rechaza la hipótesis de homocedasticidad.
- No se observan puntos de alto leverage.

Para mejorar la adecuación del modelo, en especial en lo que respecta a normalizar la distribución de sus residuos y a cumplir la hipótesis de homocedasticidad, se ha estudiado la transformación Box Cox de la respuesta (ver apartado 3.1.2 y Anexo II).

En la práctica, el principal problema a la hora de estudiar la transformación Box Cox es que no admite valores del conjunto de datos iguales a 0. Por tanto, se ha modificado ligeramente la BD.verano, sustituyendo las velocidades iguales a 0 (tan solo un 1.55% del total) por un 1/3 del valor mínimo registrado por la estación meteorológica (1.14 km/h).

Al aplicar la transformación Box Cox a nuestros datos, se observa la necesidad de transformación raíz cuadrada de la respuesta (ver en Figura 27 que el intervalo de confianza del parámetro λ, prácticamente incluye al valor 0.5).

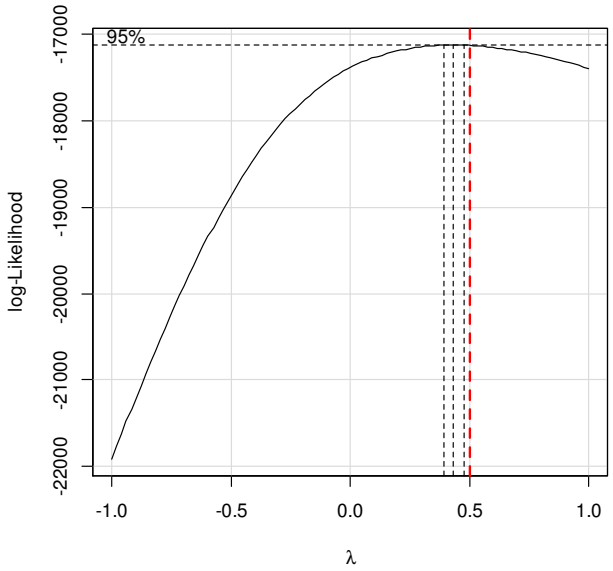


Figura 27. Estudio gráfico de la transformación Box Cox para el MRLS realizado a partir de los datos de predicción de AEMET y la BD.verano

Por tanto, se ha realizado un nuevo MRLS entre los datos de velocidad real y predicción de AEMET, utilizando como respuesta a la raíz cuadrada de la velocidad observada. Dicho modelo tiene una menor bondad de ajuste (R^2 ajustado de 0.14 frente al 0.17 del MRLS con respuesta sin transformar). Sin embargo, el histograma del nuevo modelo muestra un aspecto simétrico (ver Figura 28 izda.); también el gráfico de probabilidad normal muestra una normalización en la distribución de los residuos, y la desaparición de los *outliers* de la cola derecha de la distribución (ver Figura 28 dcha.).

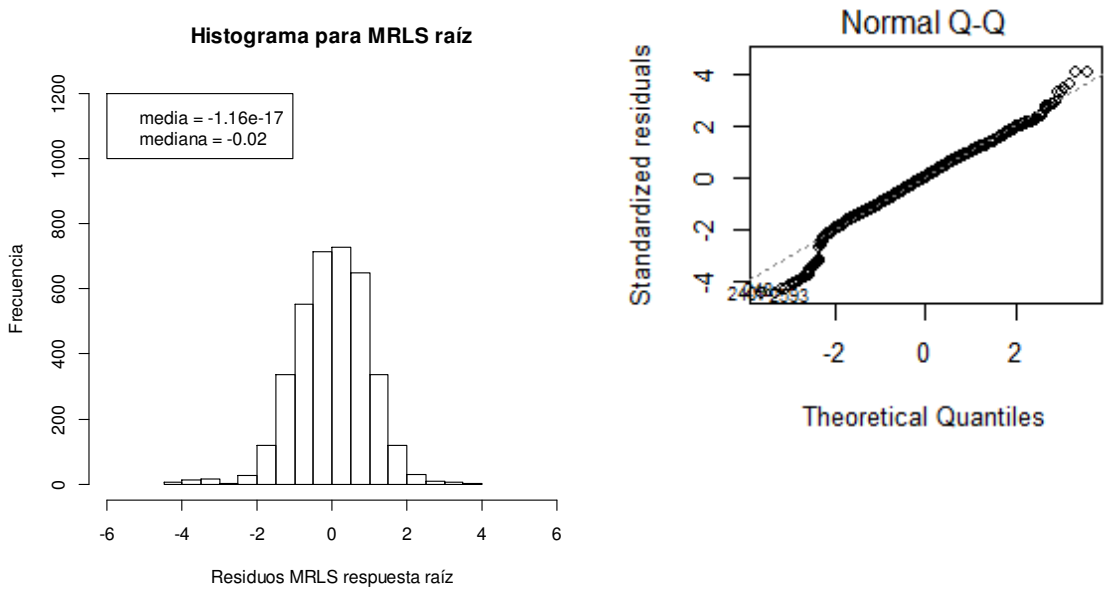


Figura 28. Histograma (izda.) y gráfico de probabilidad normal (dcha.) de residuos del MRLS construido con predicciones de AEMET y con respuesta transformada en raíz cuadrada.

En resumen, la transformación raíz de la respuesta ha permitido normalizar la distribución de los residuos del modelo.

Hay que tener presente que, para llegar a seleccionar un modelo final óptimo, se deben desarrollar un conjunto de modelos "por pasos", de entre los cuales se seleccionará el "mejor" en cada paso. Para ello, los modelos deberán ser comparables.

En estudios preliminares (realizados únicamente para el mes de junio de 2012 y no incluidos en esta memoria) se comprobó que los modelos obtenidos diferían según se hubiesen desarrollado con la respuesta normal o transformada; por ello se ha decidido trabajar con la respuesta raíz cuadrada desde un principio y para todos los modelos. No obstante, se comprobará que la transformación es adecuada en los modelos óptimos finalmente seleccionados, y se obtendrán resultados en la escala original de la respuesta para facilitar el uso del modelo estadístico.

4.2.2 DESARROLLO DE MODELOS DE REGRESIÓN CANDIDATOS A SER EL ÓPTIMO BUSCADO

Como se ha establecido en el apartado 3.1, se ha desarrollado una metodología de búsqueda de MR candidatos en dos etapas, siendo la primera el desarrollo de MR *sin* interacciones, y la segunda, el desarrollo de MR *con* interacciones a partir de los anteriores.

DESARROLLO DE MODELOS DE REGRESIÓN SIN INTERACCIONES

Como punto de partida, se tiene una variable respuesta en estudio (raíz de la velocidad real del viento), así como un total de 31 covariables regresoras (ver Tabla 1) de cuatro tipos diferentes: asociadas a la predicción de la velocidad, a la predicción de la dirección, y a los ciclos diario y anual (estacionalidad).

En primer lugar, se han desarrollado modelos de regresión lineal (MRL), según la metodología expuesta en el apartado 3.1.3, con el requerimiento de que las covariables que formen parte del modelo deben ser significativas a nivel de 0.05.

Como resultado se han obtenido diversos MRL (resumidos en la Tabla 3) que contienen entre 3 covariables (*ensayo1,* con solo covariables de predicción de velocidad del viento), y 14 covariables (*ensayo4.1*, con los 4 tipos de covariables añadidas al mismo tiempo).

En términos de bondad de ajuste, los *métodos 1* (considerar el mejor modelo anterior como base e incluir un nuevo tipo de covariable, aplicando *stepwise* en cada paso) y *2* (construir el mejor modelo con todas las covariables disponibles, aplicando *stepwise* en un solo paso) muestran resultados similares de R^2 ajustado.

Aunque aparentemente el mejor resultado corresponde al *ensayo4*, desarrollado con el *método 1* (con un R^2 ajustado de 0.47 y únicamente 10 covariables), la realización de un test ANOVA entre el *ensayo4* y el *ensayo4.1* (resultado del *método 2*, con 14 covariables y R^2 ajustado de 0.47), muestra que el *ensayo4*.1 es más adecuado (ver Figura A. 7).

Por tanto, a la hora de seguir desarrollando el modelo óptimo, se ha decidido tomar los dos modelos como candidatos.

MÉTODO 1	***Ensayo anterior + nuevas covariables***			***MÉTODO 2***	***Todas las covariables cada vez***	
	n.cov	**R^2 ajustado**	**Covariables**		**n.cov**	**R^2 ajustado**
m. AEMET	1	0.14	Velocidad (v) instantánea	**m. AEMET**	1	0.14
ensayo1	3	0.16	v (con retardos y adelantos)	***ensayo1.1***	3	0.16
ensayo2	10	0.34	v + dirección (dir, con retardos y adelantos)	***ensayo2.1***	11	0.34
ensayo3	12	0.34	v + dir + estacionalidad	***ensayo3.1***	12	0.34
ensayo4	10	0.47	v + dir + estacionalidad + ciclo	***ensayo4.1***	14	0.47

Tabla 3. Resultado del desarrollo de MRL sin interacciones, mediante los dos métodos 1 y 2. n.cov=número de covariables (de que se compone el modelo en estudio)

- **Estudio de la linealidad de los modelos candidatos seleccionados en el paso anterior**

Los dos modelos candidatos seleccionados contienen covariables de predicción de la velocidad del viento (3 en el caso del *ensayo4* y 4 para el *ensayo4.1*), las cuales son susceptibles de mostrar un efecto no lineal en el propio modelo.

El estudio de estos posibles efectos polinómicos se ha realizado según lo establecido en el apartado 3.1.3, y teniendo en cuenta que las covariables cambian su comportamiento (significación) dependiendo del resto de covariables que componen el modelo; por tanto, el efecto lineal de las covariables se debe estudiar en el propio modelo y no de forma independiente. Este efecto se ha estudiado aplicando la función *gam*.

En primer lugar, se ha aplicado la función *gam* a cada covariable de la velocidad del viento por separado, con el fin de hallar indicios acerca del grado del polinomio que expresa el efecto de cada covariable. Los resultados de este estudio para el *ensayo4* se recogen en la Figura A.8, donde se observa que la covariable $x2$ parece ajustarse a un comportamiento lineal o cuadrático (por tanto, se espera que tenga grados 1 ó 2), mientras que para las covariables $x, x.p2$, toda su relación con la respuesta podría estar recogida por los grados 3 ó 4.

En segundo lugar, la función *gam* se ha usado para desarrollar modelos anidados, que mediante su posterior comparación con un test ANOVA, han permitido estudiar qué grado es significativo para cada variable en el modelo, obteniendo que el modelo mejora hasta considerar los grados 2 para $x2$, 3 para x y 4 para $x.p2$ (ver Figura A.9).

El mismo estudio realizado para el *ensayo4.1*, muestra que el polinomio en x tiene de nuevo grado 3, $x.p2$ grado 4, $x2$ grado 2, mientras que $x4$ resulta ser lineal (ver Figura A.10).

Se han desarrollado dos nuevos MR que incluyen los polinomios estudiados (*ensayo4.poly* y *ensayo4.1.poly*). Los nuevos modelos polinómicos suponen un incremento mínimo en el R^2 ajustado (de aproximadamente un 0.01), mientras aumenta considerablemente el número de predictores: en 6 covariables respecto al *ensayo4* y en 2 covariables respecto al *ensayo4.1* (ver Tabla 4). Sin embargo, los test ANOVA realizados entre los modelos anidados, sin y con polinomios, resultan en una mejora significativa de los ensayos que contienen polinomios (ver Figura A.11 para modelos desarrollados a partir del *ensayo4* y Figura A.12 a partir del *ensayo4.1*).

Por tanto, se ha decidido trabajar también con los dos modelos que incluyen polinomios.

MODELO	Estudio no-linealidad	R^2 ajustado	Nº predictores
modelo.original	*	0.13	1
ensayo1	*	0.16	3
ensayo2	*	0.34	11
ensayo3	*	0.33	12
ensayo4		0.47	10
	ensayo4.poly	0.48	16
ensayo4.1		0.47	14
	ensayo4.1.poly	0.49	16

Tabla 4. Resumen de los modelos obtenidos mediante el desarrollo de MR sin interacciones.

En general, se puede concluir que para el estudio de los MR sin interacciones, se da una mejora significativa en el R^2 ajustado, llegando éste a alcanzar en torno al 0.48, cuando se añaden:

- Covariables ligadas a la dirección del viento, que aumentan la explicación de la variabilidad de la respuesta en un 18%.
- Covariables que expresan el ciclo diario existente en el régimen de brisas, que mejoran la explicación de la respuesta en un 13%.
- Los términos polinómicos únicamente aumentan la explicación de la variabilidad de la respuesta en un 1%, aunque según los test ANOVA realizados, resultan significativos.

Los modelos candidatos que se utilizarán para seguir desarrollando el modelo "óptimo" son de tipo lineal (*ensayo4* y *ensayo4.1)* y polinómico (*ensayo4.poly* y *ensayo4.1.poly).*

DESARROLLO DE MODELOS DE REGRESIÓN CON INTERACCIONES

Se pretende incrementar el grado de explicación de la respuesta (velocidad observada) considerando la introducción de interacciones entre componentes que intervienen en el MR. El efecto de las interacciones se ha considerado construyendo modelos dirigidos por las covariables existentes y también mediante procedimientos semiautomáticos.

- **Introducción manual**

Siguiendo la metodología, y dados los indicios hallados en el análisis exploratorio, las interacciones evaluadas han sido las siguientes:

- Interacciones entre covariables de predicción de la velocidad del viento y estacionalidad, ya que el régimen de vientos cambia a lo largo del año.
- Interacciones entre covariables de predicción de la velocidad y predicción de la dirección del viento, ya que se ha detectado una relación entre ambas.
- Interacciones entre covariables de predicción de la velocidad y ciclo diario, ya que se ha comprobado la existencia de un régimen diario de brisas diferenciado.
- Interacciones entre ciclo diario y estacionalidad.

El test ANOVA muestra como modelos más significativos aquellos a los cuales se introducen todas las interacciones en estudio (ver Figura A.13 para modelos lineales y Figura A.14 para polinómicos). Dichos modelos tienen mayor R^2 ajustado y número de parámetros.

Sin embargo, la interacción que produce el mayor aumento en la bondad de ajuste del modelo con el menor aumento en el número de parámetros (la "mejor ecuación de regresión" en este caso), es aquella formada por las covariables de predicción de velocidad de viento y ciclo diario (ver casillas coloreadas en Tabla 5). Este aumento es de un 2.5% con solo 1 parámetro más para los dos modelos lineales, y de un 2.5% para los polinómicos, aunque con un mayor aumento en el número de parámetros (entre 6 y 11, ver Tabla 5).

MODELOS LINEALES	N.par.	R^2 ajustado	Interacción añadida	MODELOS POLINÓMICOS	N.par.	R^2 ajustado
Modelo original ensayo4	10	0.471	- -	***Modelo original ensayo4.poly***	16	0.485
ensayo4.a	13	0.473	p.v.viento*ciclo.anual	*ensayo4.poly.a*	26	0.492
ensayo4.b	11	0.496	p.v.viento*ciclo.diario	*ensayo4.poly.b*	27	0.511
ensayo4.c	12	0.497	p.v.viento*(ciclo.diario + ciclo.anual)	*ensayo4.poly.c*	32	0.515
ensayo4.d	14	0.499	ensayo4.c + ciclo.diario*ciclo.anual	*ensayo4.poly.d*	34	0.517
Modelo original ensayo4.1	14	0.474	- -	***Modelo original ensayo4.1.poly***	16	0.486
ensayo4.1.a	17	0.477	p.v.viento*ciclo.anual	*ensayo4.1.poly.a*	23	0.493
ensayo4.1.b	15	0.500	p.v.viento*ciclo.diario	*ensayo4.1.poly.b*	22	0.511
ensayo4.1.c	15	0.501	p.v.viento*(ciclo.diario + ciclo.anual)	*ensayo4.1.poly.c*	31	0.517
ensayo4.1.d	19	0.503	ensayo4.1.c + ciclo.diario*ciclo.anual	*ensayo4.1.poly.d*	23	0.512

Tabla 5. Modelos lineales y polinómicos a los que se han introducido interacciones de forma manual. El símbolo (- -) indica que el modelo no tiene ninguna interacción. Se han coloreado los modelos a los que se han introducido interacciones que producen una mejora más significativa.

- **Introducción mediante procedimientos semiautomáticos**

El segundo ensayo se basa en la introducción de interacciones a los modelos mediante el uso de los bucles desarrollados en R para tal fin (*b1* y *b2*, ver apartado 3.1.3), que han permitido estudiar una gran cantidad de modelos diferentes de forma automática y rápida. Los modelos obtenidos se han resumido mediante el nombre del modelo de partida seguido del bucle utilizado en cada caso (*b1, b2*).

Las interacciones (hasta de orden 3) se han estudiado entre las covariables asociadas a velocidad y dirección predichas y armónicos de los ciclos diarios y estacionales; esto corresponde a aproximadamente 5000 términos potenciales distintos.

Si se comparan los modelos obtenidos mediante *b1* y *b2*, tanto a partir de modelos lineales como polinómicos (ver Tabla 6), se observa cómo el uso *b2* obtiene modelos con menor número de parámetros y similar bondad de ajuste.

La realización de un test ANOVA, muestra resultados más significativos para el *ensayo4.1.b2* de entre los modelos lineales (ver Figura A.15) y para el *ensayo4.poly.b2* de entre los polinómicos (ver Figura A.16), aunque en este caso, no se ha podido comparar con el *ensayo4.1.poly.b2* por no estar anidados.

MODELOS LINEALES			**MODELOS POLINÓMICOS**		
BUCLE 1	**R^2 ajustado**	**Número parámetros**	**BUCLE 1**	**R^2 ajustado**	**Número parámetros**
ensayo4.b1	0.5276	24	*ensayo4.poly.b1*	0.5522	41
ensayo4.1.b1	0.5442	35	*ensayo4.1.poly.b1*	0.5586	47
BUCLE 2			**BUCLE 2**		
ensayo4.b2	0.5227	16	*ensayo4.poly.b2*	0.5511	39
ensayo4.1.b2	0.5404	22	*ensayo4.1.poly.b2*	0.5578	34

Tabla 6. Modelos resultantes de introducir interacciones mediante bucles *1* y *2*, a modelos candidatos seleccionados, de tipo lineal y polinómico.

En resumen, las interacciones que parecen resultar más significativas son aquellas formadas por las covariables de predicción de la velocidad del viento y las que expresan el ciclo diario existente en el régimen de brisas; es decir, que la intensidad del viento dependerá en general, de la hora del día en que nos encontremos.

El uso de bucles optimiza el tiempo invertido en la construcción de modelos, permitiendo estudiar una cantidad de interacciones entre covariables que de otro modo resultaría inviable. En particular, el bucle 2 es mucho más sencillo, tanto en eficiencia de los cálculos como en "limpieza" posterior del modelo, y obtiene modelos con menor número de parámetros y similar bondad de ajuste.

4.2.3 SELECCIÓN DE MODELOS CANDIDATOS

En la Tabla 7 se han recopilado los modelos más interesantes desarrollados hasta el momento, con el fin de estudiarlos con más profundidad y seleccionar aquellos con los que seguir trabajando. Los criterios de selección utilizados han sido los expuestos en el apartado 3.2 de la metodología.

De entre los modelos de tipo lineal, finalmente se han seleccionado los ensayos *4.1.b1* y *4.1.b2,* que muestran los mejores resultados en términos de bondad de ajuste, desviación típica residual y desviación típica de validación cruzada (ver modelos en negrita en Tabla 7).

Para guiar la elección se han realizado distintos test ANOVA, en primer lugar para la colección de ensayos lineales desarrollados a partir del *ensayo4* (ver Figura A.17) y del *ensayo4.1* (ver Figura A.18), para posteriormente comparar los "mejores" ensayos obtenidos en ambos casos; el test ANOVA realizado entre ellos (ensayos *4.b2, 4.1.b1* y *4.1.b2*), muestra como preferible al último (ver Figura A.19).

De entre los modelos de tipo polinómico, los mejores resultados corresponden a los ensayos *4.1.poly.b1* y *4.1.poly.b2* (ver Tabla 7). En este caso, el test ANOVA muestra como más significativo al más sencillo, el *ensayo4.1.poly.b2* (ver Figura A.20).

Los ensayos finalmente seleccionados se muestran en negrita en la Tabla 7.

MODELO	$\hat{S}_R$ método CV aleatorio	$\hat{S}_R$ método CV consecutivo	R^2 ajustado	Número parámetros	$\hat{S}_R$
Lineal					
ensayo4	0.747	0.747	0.471	10	0.745
ensayo4.1	0.746	0.746	0.474	14	0.743
ensayo4.b	0.730	0.729	0.496	11	0.728
ensayo4.1.b	0.727	0.727	0.500	15	0.724
ensayo4.b1	0.708	0.708	0.528	24	0.704
ensayo4.1.b1 (mc.l1)	0.697	0.697	0.544	35	0.692
ensayo4.b2	0.711	0.711	0.523	16	0.708
ensayo4.1.b2 (mc.l2)	0.698	0.698	0.540	22	0.695
Polinómico					
ensayo4.poly	0.738	0.738	0.485	16	0.736
ensayo4.1.poly	0.737	0.737	0.486	16	0.734
ensayo4.poly.b	0.721	0.722	0.511	27	0.717
ensayo4.1.poly.b	0.720	0.720	0.511	22	0.716
ensayo4.poly.b1	0.692	0.691	0.552	41	0.686
ensayo4.1.poly.b1 (mc.p3)	0.688	0.687	0.559	47	0.681
ensayo4.poly.b2	0.692	0.692	0.551	39	0.686
ensayo4.1.poly.b2(mc.p4)	0.686	0.686	0.558	34	0.681

Tabla 7. Resultados de los criterios de selección estudiados para los modelos seleccionados hasta el momento: bondad de ajuste (R^2 ajustado), número de parámetros, desviación típica residual ($\hat{S}_R$) y desviación típica resultante del proceso de validación cruzada (CV $\hat{S}_R$) con procedimiento de ventanas aleatorias y consecutivas. En negrita se muestran los modelos seleccionados; entre paréntesis se indica el nuevo nombre asignado. Las casillas coloreadas corresponden a los mejores resultados para cada criterio de selección en estudio.

El R^2 ajustado de los ensayos seleccionados oscila entre 0.559 (*ensayo4.1.poly.b1,* de 47 parámetros) y 0.540 (*ensayo4.1.b2*, de 22 parámetros). Todos ellos proceden del *ensayo4.1*, y a todos se han introducido las interacciones de forma automática. Se ha modificado el nombre a los modelos seleccionados, para hacer su manejo más fácil e intuitivo (ver Tabla 8).

ensayo4.1.b1	*mc.l1*
ensayo4.1.b2	*mc.l2*
ensayo4.1.poly.b1	*mc.p3*
ensayo4.1.poly.b2	*mc.p4*

Tabla 8. Nuevos nombres asignados a los modelos candidatos seleccionados, donde "mc" hace referencia a *modelo candidato*, "l" a *lineal* y "p" a *polinómico*.

En resumen, en la Tabla 7 se han coloreado las casillas que muestran los mejores resultados para cada uno de los criterios estudiados. Atendiendo a la bondad de ajuste y a la desviación típica residual, el mejor modelo candidato es *mc.p3*, seguido de *mc.p4* (ambos polinómicos), con igual desviación residual, y que ofrece los mejores resultados en términos de desviación típica de validación cruzada. Por último, *mc.l2* es interesante por su sencillez, ya que solo tiene 22

parámetros. Comparado con los otros tres, el modelo *mc.l1* no parece aportar mejoras en ningún criterio.

- **Estudio de la composición de los modelos candidatos**

Los modelos candidatos *mc.p3* y *mc.p4* incluyen interacciones triples en su composición, muy complicadas de interpretar; también aparecen algunas interacciones "no deseadas", por ejemplo entre términos de un armónico, que indican que los armónicos introducidos son insuficientes para recoger el comportamiento estacional y diario del régimen de vientos. Por ello, se ha estudiado la sustitución de dichas interacciones "no deseadas" por los segundos armónicos de los ciclos diarios y estacionales.

El segundo armónico del ciclo diario se ha calculado de la misma forma que se calculó el primer armónico, permitiendo reproducir el comportamiento de dos ciclos durante el periodo de 24 horas.

$$sen.hora2 = sin\frac{hora}{24}2\pi * 2 \quad ; \quad cos.hora2 = cos\frac{hora}{24}2\pi * 2$$

Por su parte, el segundo armónico para el ciclo estacional se ha calculado:

$$sen.dia2 = sin\frac{dia.año}{366}2\pi * 2 \quad ; \quad cos.dia2 = cos\frac{dia.año}{366}2\pi * 2$$

Mediante sucesivas pruebas (no incluidas en la memoria por la extensión que éstas supondrían), se ha comprobado que a la hora de explicar el comportamiento del viento a lo largo del día, resultan necesarios también los terceros y cuartos armónicos del ciclo diario. Dichos armónicos se han calculado de la misma forma.

Para los modelos candidatos lineales (*mc.l1* y *mc.l2*) sin interacciones "no deseadas", se ha estudiado el efecto de añadir los nuevos armónicos a los modelos, actualizando el modelo resultante; estos modelos se han renombrado a partir del modelo de origen (*mc.l1.2, mc.l2.2*), y se incluyen en la Tabla 9. Gracias a la introducción de los nuevos armónicos se ha conseguido eliminar ciertas interacciones triples de *mc.l1*, mientras el número de parámetros, bien ha aumentado en 1 (para *mc.l2*), bien ha disminuido (para *mc.l1*); por su parte, la bondad de ajuste ha mejorado ligeramente en ambos casos. También se observa cómo el porcentaje de parámetros muy significativos para el modelo aumenta en ambos casos.

La introducción de los nuevos armónicos en los modelos candidatos polinómicos (*mc.p3.2* y *mc.p4.2*) resulta beneficiosa para el modelo, pues elimina la necesidad de interacciones "no deseadas". Además, en ambos casos permite disminuir el número de parámetros a la vez que aumenta su significación, con una bondad de ajuste y una desviación típica residual muy similar (en torno al 0.57-0.58, y entre 0.665 y 0.668 respectivamente), mientras el número de interacciones triples disminuye. También se obtienen mejores resultados en términos de desviación típica residual de validación cruzada para *mc.p3* (ver casillas azules en Tabla 9).

MODELO	N.par	%.par.sig	R^2 ajustado	$\hat{S}_R$	N.int. triples	$\hat{S}_R$ CV método aleatorio	$\hat{S}_R$ CV método consecutivo
Lineal							
mc.l1	35	60%	0.544	0.692	9	0.697	0.697
mc.l1.2	27	88.9%	0.570	0.672	4	0.676	0.676
mc.l2	22	86.4%	0.540	0.695	1	0.698	0.698
mc.l2.2	23	100%	0.569	0.674	0	0.676	0.676
Polinómico							
mc.p3	47	61.7%	0.559	0.681	7	0.688	0.687
mc.p3.2	40	62.5%	0.579	0.665	5	0.670	0.670
mc.p4	34	85.3%	0.558	0.681	3	0.686	0.686
mc.p4.2	32	78.1%	0.575	0.668	2	0.672	0.672

Tabla 9. Estudio de la variación en los criterios de selección de los modelos candidatos, tras la introducción de nuevos armónicos de ciclo y estacionalidad. Las casillas coloreadas corresponden a los mejores resultados obtenidos para cada criterio. %.par.sig= porcentaje de parámetros con alta significación a la hora de predecir la respuesta (p-valor <0.01). N.int= número de interacciones. Los modelos en negrita corresponden a los finalmente seleccionados.

Con el fin de seleccionar el modelo lineal óptimo de entre los estudiados en la Tabla 9, se ha realizado un test ANOVA que muestra como más significativo al modelo *mc.l2.2* (ver Figura A. 21). Para el grupo de los modelos polinómicos, el test ANOVA no aporta gran cantidad de información, ya que éstos no están anidados (ver Figura A. 22).

Antes de la selección final de los modelos óptimos, se ha estudiado el ajuste de los modelos candidatos para cada mes del periodo en estudio. Los resultados de dicho estudio vienen recogidos en la Tabla 10, donde se observa que los modelos se ajustan de forma similar a cada mes por separado, con R^2 ajustados que oscilan entre 0.63 para septiembre y 0.51 para julio. Por tanto, se considera que no será necesario desarrollar un nuevo modelo específico para ningún mes en particular.

MODELO	N.par	R^2 ajustado	R^2 ajustado				
Lineal			**Mayo**	**Junio**	**Julio**	**Agosto**	**Sept.**
mc.l1	35	0.544	0.519	0.537	0.514	0.558	0.603
mc.l1.2	27	0.5701	0.546	0.553	0.528	0.596	0.630
mc.l2	22	0.540	0.517	0.523	0.517	0.552	0.596
mc.l2.2	23	0.5685	0.543	0.543	0.533	0.594	0.630
Polinómico							
mc.p3	47	0.559	0.543	0.548	0.534	0.572	0.613
mc.p3.2	40	0.579	0.571	0.563	0.545	0.602	0.624
mc.p4	34	0.558	0.536	0.55	0.529	0.568	0.616
mc.p4.2	32	0.5754	0.562	0.561	0.540	0.595	0.623

Tabla 10. Ajuste de los modelos candidatos (con y sin nuevos armónicos) a meses por separado. Las casillas coloreadas muestran los mejores ajustes.

De entre los modelos candidatos lineales, no hay ninguno que se adapte mejor a todos los meses. Por ello, se ha seleccionado el modelo *mc.l2.2*, conforme al test ANOVA realizado anteriormente (ver Figura A. 21) y dada su sencillez: está formado por únicamente 23 parámetros y no contiene interacciones triples. En vistas a facilitar su análisis, se le ha denominado *ms1* ("modelo seleccionado 1").

De entre los modelos candidatos polinómicos, el que mejor se ajusta, tanto al periodo en general como a los meses por separado, es *mc.p3.2*. Sin embargo, las diferencias con *mc.p4.2* en términos de bondad de ajuste (tanto global como por meses) son mínimas. Por ello, y por ser un modelo mucho más sencillo, se ha preferido *mc.p4.2*, el cual se ha renombrado *ms2*.

- **Estudio comparativo de la composición de los modelos seleccionados**

Comparativamente, los dos modelos seleccionados utilizan 13 *covariables*, de las cuales 7 coinciden (ver casillas coloreadas en Tabla 11); ambos modelos incluyen covariables de alta significación. Estas covariables son de tipo predicción de la dirección del viento (*cosd4*), ciclo diario (*sen.hora, sen.hora2, cos.hora2, sen.hora4, sen.hora4, cos.hora4)* y estacionalidad (*sen.dia2*). Las principales diferencias entre los modelos seleccionados, son la presencia en *ms2* de una covariable polinómica de segundo grado de predicción de la velocidad del viento.

covariables	***ms1***		**sig.**	***ms2***		**sig.**
intercept			***			***
pred.vel	2	x4	***	3	x2	***
		x.p2	***		poly(x,2)1	***
		-			poly(x,2)2	***
pred.dir	2	cosd4	**	1	cosd4	***
		send3	**		-	
ciclo	6	sen.hora	***	7	sen.hora	***
		-			cos.hora	***
		sen.hora2	***		sen.hora2	***
		cos.hora2	***		cos.hora2	***
		sen.hora3	**		sen.hora3	***
		sen.hora4	***		sen.hora4	***
		cos.hora4	**		cos.hora4	*
estacionalidad	3	cos.dia	***	2	sen.dia	***
		sen.dia2	**		sen.dia2	***
		cos.dia2	**		-	
TOTAL	13			13		

Tabla 11. Resumen de las covariables contenidas en los modelos seleccionados *ms1* y *ms2*, y su significación (sig.) en el modelo, a la hora de predecir la respuesta: "***" indica p-valor<0.001, "**" p-valor<0.1 y "*" p-valor <0.5.

En la Tabla 12 se ha recogido el estudio de las *interacciones* que contienen ambos modelos: *ms1* muestra únicamente 13, sin que ninguna de ellas sea triple; contiene 18, de las cuales 2 son triples. Solo 3 interacciones son comunes a ambos modelos, construidas entre predicción de la dirección y ciclo diario (*send3:sen.hora, cosd4:sen.hora*) y predicción de la dirección y estacionalidad (*cosd4:cos.dia*). De nuevo la mayor diferencia entre ambos modelos es la presencia de términos que incluyen covariables polinómicas de predicción de la velocidad del viento (de segundo y tercer grado) en *ms2*.

interacciones	*ms1*		sig.	*ms2*		sig.
p.v*p.v	1	x:x2	***	1	x3:x.p3	***
p.v*p.dir.v	2	x:send.p1	***	3	poly(x,3)1:send3	**
		x:cosd.p4	**		poly(x,3)2:send3	***
		-			poly(x,3)3:send3	*
p.v*ciclo	2	x2:cos.hora	***	3	poly(x,3)1:sen.hora	-
		x.p2:cos.hora	***		poly(x,3)2:sen.hora	-
		-			poly(x,3)3:sen.hora	**
p.v*estacionalidad	0	-		3	poly(x,3)1:cos.dia	-
		-			poly(x,3)2:cos.dia	***
		-			poly(x,3)3:cos.dia	***
p.dir.v*ciclo	3	send3:sen.hora	***	5	send3:sen.hora	***
		cosd4:sen.hora	***		cosd4:sen.hora	***
		sendv:sen.hora	**		send.p1:sen.hora	**
		-			poly(x,2)1:cos.hora	-
		-			poly(x,2)3:cos.hora	*
p.dir.v*estacionalidad	1	cosd4:cos.dia	***	1	cosd4:cos.dia	***
ciclo*estacionalidad	1	cos.hora:sen.dia	***	0	-	
interacciones triples	0			2		
TOTAL	10			18		

Tabla 12. Resumen de las interacciones contenidas en los modelos seleccionados *ms1* y *ms2*, y su significación (sig.) en el modelo, a la hora de predecir la respuesta: "***" indica p-valor<0.001, "**" p-valor<0.1 y "*" p-valor <0.5.

4.2.4 ADECUACIÓN DE LOS MODELOS SELECCIONADOS

Siguiendo la metodología desarrollada en el apartado 3.3, se ha realizado un estudio de la adecuación de los dos modelos óptimos seleccionados, basado en el análisis de sus residuos.

No se ha podido realizar una validación "real" en este caso, ya que no ha sido posible disponer de una nueva BD que pudiera usarse para tal fin. Por tanto, a la hora de validar el modelo, únicamente se dispone de los resultados del proceso de validación cruzada (desviación típica residual de validación cruzada), llevada a cabo en el apartado anterior.

Para los modelos seleccionados, la diferencia obtenida entre los valores de desviación típica y los de desviación típica de validación cruzada, ha sido inferior al 0.7% (ver Tabla 9) mostrando que los modelos funcionan, no solo para la BD usada en su ajuste, si no que funcionarían también para una nueva BD.

ADECUACIÓN DEL MODELO CANDIDATO MS1

Los gráficos del análisis de residuos del modelo *ms1* (ver Figura A.23) muestran lo siguiente:

- Se verifica la hipótesis de linealidad, pues no se observa ningún patrón en el gráfico *residuals vs fitted.*
- Se verifica la hipótesis de homocedasticidad, pues no existe evidencia de incremento o cambio de la variabilidad de la varianza.
- No existen puntos influyentes que deban eliminarse del modelo (gráfico r*esiduals vs leverage*).

- Existen ciertas desviaciones de la normalidad (gráfico *Normal QQ plot).*

Se ha descartado que el problema de la normalidad de los residuos se deba a una incorrecta transformación de la respuesta (ver estudio de la transformación Box Cox en Figura A.24). El histograma para los residuos de *ms1*, recogido en la Figura A.25, presenta un aspecto simétrico, con valores de media y mediana aproximadamente iguales y en torno al 0.

El estudio de la distribución de los residuos estandarizados para cada mes mediante diagramas de cajas (ver Figura 29), muestra que la distribución es simétrica para todos los meses (mediana en torno al 0 y en el centro de la caja), y que por tanto, el modelo es capaz de recoger la variabilidad del comportamiento del viento de modo similar para cada mes. Los valores atípicos que superan el rango de ±3 (línea discontinua verde) no se agrupan en ningún mes en particular, siendo predominantemente de signo negativo. Se observan aproximadamente 20 datos atípicos de signo positivo en total (4 en mayo, 5 en junio, 4 en julio, 2 en agosto y 5 en septiembre). Estos residuos son de mayor interés que los de signo negativo, pues indican situaciones de alarma en las cuales el modelo no es capaz de predecir velocidades elevadas que se dan en realidad.

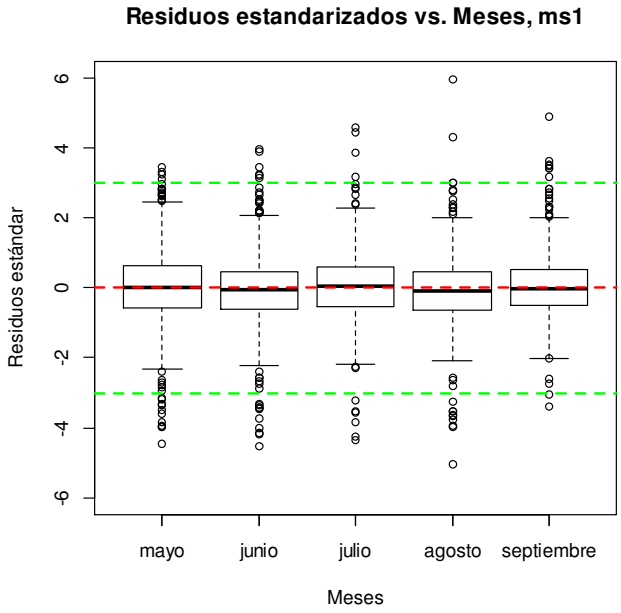


Figura 29. Diagramas de cajas para estudiar la variación de los residuos estándar del modelo *ms1* a lo largo del periodo considerado (mayo-septiembre).

Los diagramas de cajas para estudiar la distribución de los residuos estándar de *ms1* a lo largo del día, muestran que en concreto para el horario diurno (el horario de mayor interés, ver Figura 30), se observan 9 medidas de residuos estandarizados muy elevados, a las 13 horas (1 medida), las 16 horas (2 medidas), a las 18 horas (3 medidas) y a las 19 horas (3 medidas). También en este caso las cajas son simétricas, y la mediana se encuentra en torno al 0, mostrando un correcto comportamiento de *ms1* para el horario diurno.

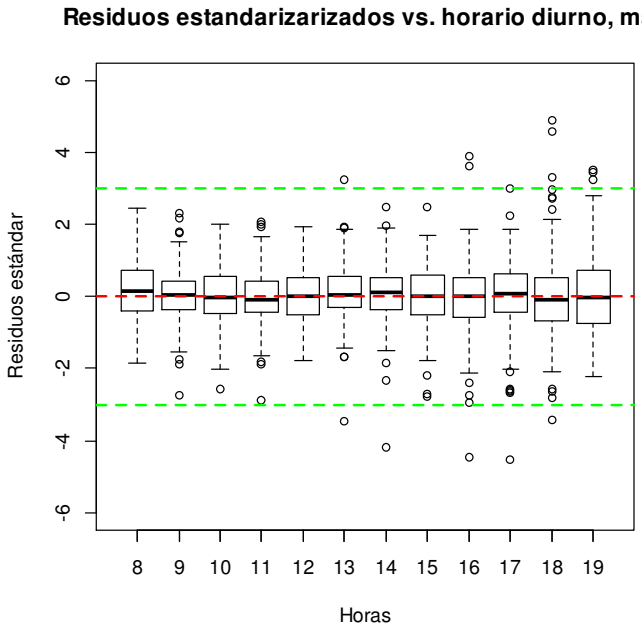


Figura 30. Estudio de la variación de los diagramas de cajas para residuos estándar del modelo *ms1* por horas para el horario diurno

Para el horario nocturno, recogido en la Figura A. 26 del Anexo I, se observan residuos estándar mayores a +3 durante todo el periodo, con un total de 14 medidas no capturadas por el modelo (en contraste con las 9 del horario diurno). Para este horario, el modelo se ajusta peor a los datos: no se observa simetría en las cajas, y su mediana se desplaza ligeramente respecto al 0.

El gráfico secuencial de residuos (recogido en la Figura A.27) no muestra ningún tipo de pauta o tendencia a lo largo del tiempo, luego no existen indicios contra la adecuación de *ms1* en dicho gráfico; tampoco aportan indicios contra ella los gráficos de variables incluidas en el modelo versus residuos del modelo (ver Figura A. 28 izda. para *p.v.viento* y dcha. para *p.dir.viento*).

El resultado de los gráficos *crPlots* para las covariables del modelo en estudio, recogido en la Figura A.29, muestra que en general todas las covariables se adecúan bien, pues la línea verde, que indica el efecto ajustado, se superpone a la roja, que expresa el suavizado de la tendencia de los datos. Por último, el estudio de los gráficos *avPlots* de variable añadida (recogidos en la Figura A.30) muestra pendientes relevantes en el comportamiento de la covariable $sen.hora$ y para la interacción x: $x2$, en cuyo caso la pendiente viene influenciada por un número escaso de puntos. La situación de estudio de mayor interés sería la existencia de grupos de residuos de signo positivo para algún parámetro, que podría dar pie a una investigación más exhaustiva sobre dichos grupos. Sin embargo, solo existen agrupaciones de residuos de signo negativo; por tanto, los gráficos *avPlots* tampoco aportan indicios sobre cómo mejorar el modelo *ms1*.

ADECUACIÓN DEL MODELO CANDIDATO MS2

Para el caso del modelo *ms2*, que contiene covariables de tipo polinómico, los gráficos en estudio (ver Figura A.31) muestran que se recoge la linealidad y la homocedasticidad del modelo, que no existen puntos influyentes que deban eliminarse, y que existen ciertas desviaciones en

las colas de distribución. Es decir, al igual que ocurría con *ms1*, el modelo *ms2* presenta alguna dificultad con la normalidad de los residuos, que de nuevo no se debe a una incorrecta transformación de la respuesta (ver Figura A.32). En este caso, el histograma de residuos también presenta un aspecto simétrico (ver Figura A.33).

El estudio de los diagramas de cajas para meses por separado (ver Figura 31), muestra que en general, *ms2* se adapta bien a todos los meses. En este caso, se observan alrededor de 20 residuos atípicos positivos, frente a los 14 observados en *ms1*.

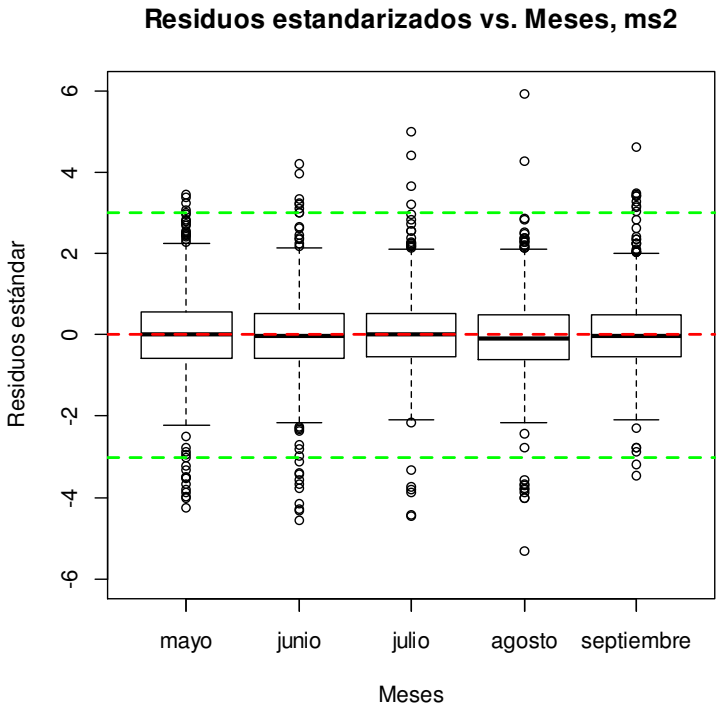


Figura 31 Diagramas de cajas para estudiar la variación de los residuos estándar del modelo *ms2* a lo largo del periodo en estudio (mayo-septiembre).

El estudio de los diagramas de cajas de residuos estándar para los horarios diurno (ver Figura 32) y nocturno (ver Figura A. 34), muestran resultados similares a los hallados para *ms1*, con un menor número de datos atípicos positivos (7 en *ms2* frente a 9 en *ms1*), agrupados entre las 16 y las 19 horas en este caso (2 en 16 horas, 3 en 18 horas y 2 en 19 horas). Esto hace pensar que quizá los términos polinómicos que contiene *ms2* ayudan a mejorar la distribución o a resolver algunas de las observaciones que no se predicen adecuadamente.

Tampoco los gráficos secuencial de residuos (ver Figura A.35), de predicción de la velocidad y de predicción de la dirección versus residuos (ver Figura A. 36 izda. y dcha. respectivamente), los gráficos *crPlots* (ver Figura A.37) y *avPlots* de variable añadida (ver Figura A.38), muestran pautas a lo largo del tiempo para *ms2*, así que no aportan información sobre posibles carencias del modelo y opciones de mejora.

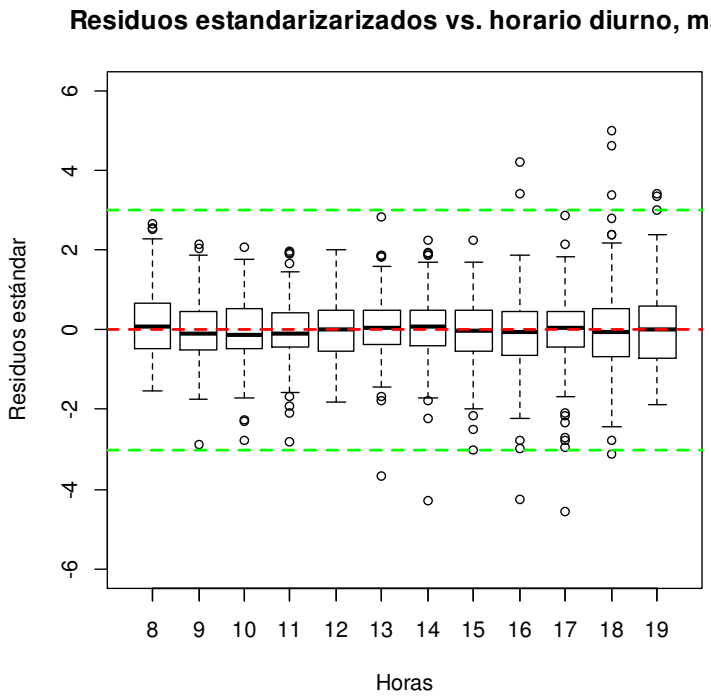


Figura 32 Estudio de la variación de los diagramas de cajas para residuos estándar del modelo *ms2* por horas, para el horario diurno (entre las 8 y las 19 horas).

En resumen, se ha observado que ambos modelos seleccionados (*ms1* y *ms2)* se ajustan correctamente y de forma similar, tanto para cada mes por separado, como por horas, especialmente durante el horario diurno. No se han observado comportamientos anómalos en el análisis de residuos realizado que sugieran la necesidad de incluir más armónicos, o de ajustar otros modelos.

4.2.5 ESTUDIO DEL FUNCIONAMIENTO OPERATIVO DE LOS MODELOS SELECCIONADOS. PRONÓSTICO

El protocolo de actuaciones de las obras de desmantelamiento y transferencia del vertedero de Bailín, postula que las mismas deberán detenerse, cuando el modelo de predicciones indique velocidades de viento superiores al umbral de 40 km/h. Durante el periodo en estudio se registraron velocidades superiores a 40 km/h en 27 ocasiones, durante 17 días diferentes (ver Tabla A.39), a los cuales se ha denominado como “días parada” por la necesidad de parar las obras que dichas velocidades supondrían.

Los registros problemáticos se dieron en todos los meses, predominando septiembre (que concentra al 44.5% del total), seguido de mayo (18.5%), agosto (14.8%) y junio y julio (ambos con un 11.1%). Un 92.6% de dichas observaciones se registraron durante el horario diurno, especialmente entre las 13 y las 18 horas (70.4%). Esta situación es lógica, pues a lo largo del análisis exploratorio se ha observado que las mayores velocidades se dan durante el día, y que durante la noche la velocidad observada es por lo general de intensidad reducida.

Se ha realizado un estudio del pronóstico de los dos modelos seleccionados (*ms1*, lineal, y *ms2*, polinómico) para cada "día parada", de forma gráfica y por intervalos de probabilidad de ocurrencia de distintos intervalos de velocidad del viento, según lo establecido en el apartado 3.5 de la metodología. Se ha prestado especial atención, como en análisis anteriores, al funcionamiento de los modelos para el horario diurno (entre las 8 y las 19 horas); por ello, se han eliminado a priori dos de los días en estudio, en los cuales las velocidades problemáticas se registraron durante la noche (26 de julio y 28 de agosto, ver Tabla A.39).

Posteriormente, se ha estudiado el comportamiento del modelo mediante dos tipos de tablas: tablas de contingencia para el estudio del pronóstico instantáneo, y tablas que han permitido estudiar el pronóstico de los modelos por días, según el número de observaciones correspondientes a distintos umbrales de velocidad del viento que se superan durante cada día.

ESTUDIO GRÁFICO DE LOS PRONÓSTICOS

Tras el estudio gráfico de los pronósticos de los modelos para los "días parada", se ha observado la existencia de situaciones en las que el funcionamiento de los modelos ha sido correcto, y situaciones en las que los modelos no han sido capaces de pronosticar correctamente. Dentro de las primeras, se distinguen tres tipos de situaciones:

- Días en los que el modelo ha pronosticado correctamente, en los cuales la velocidad del viento observada sigue un marcado ciclo diario.
- Días en los que se considera que el modelo ha funcionado correctamente, con observaciones aisladas originadas por fenómenos convectivos puntuales (tormentas).
- Días en los que se considera que el modelo ha funcionado correctamente, ya que es capaz de pronosticar velocidades elevadas (superiores a 30 km/h).

Se muestran a continuación algunos ejemplos de cada situación; el resto de días estudiados se han recogido en el Anexo I.

- **Estudio de "días parada" en los que el modelo *sí* es capaz de predecir**

- <u>Días conflictivos en que el modelo pronostica correctamente</u>

Los modelos en estudio *ms1* y *ms2*, han sido capaces de predecir velocidades superiores a 40 km/h con P>0.75 (*P* denota probabilidad), como efectivamente ocurre, para por ejemplo el 21 de mayo. Para este día, cuyo estudio gráfico se recoge en la Figura 33, las observaciones de velocidad del viento siguen una marcada trayectoria a lo largo del día. Ambos modelos pronostican también con P≈1, que se superará el umbral de los 30 km/h.

Es importante que el modelo sea capaz de recoger este tipo de comportamientos de fuerte velocidad, en los que se dan situaciones de viento sostenidas durante varias horas, es decir, que no se producen de forma puntual.

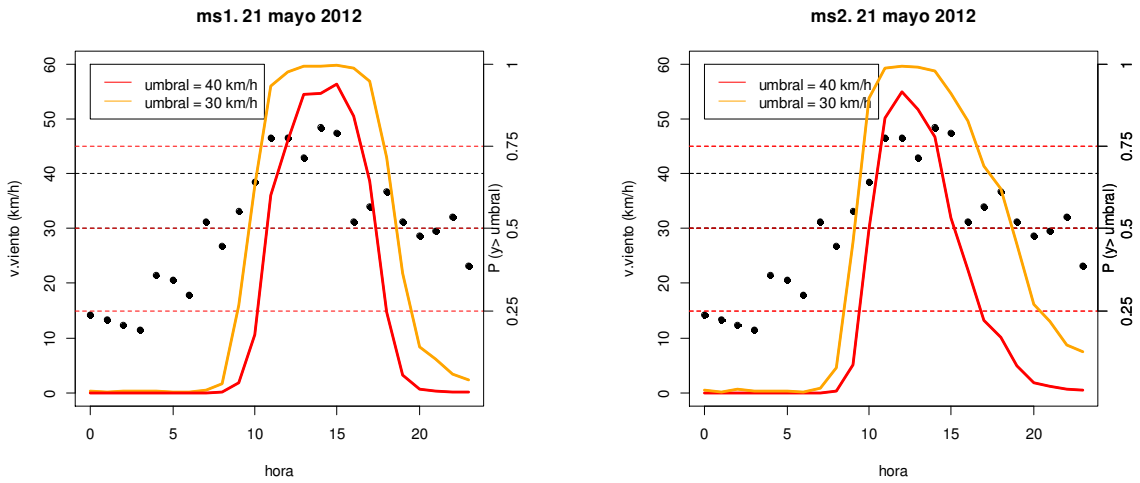


Figura 33. Estudio gráfico del pronóstico de los modelos en estudio (*ms1* a la izda. y *ms2* a la dcha.) para el 21 de mayo de 2012. Los puntos representan la velocidad observada en cada hora y las líneas discontinuas rojas los intervalos estudiados, de probabilidad estimada por los modelos de superar cierto umbral, definido por el código de color. La escala vertical derecha corresponde a probabilidades; la izquierda, a velocidad en km/h.

- Días en que se dan una (o varias) observaciones aisladas problemáticas

Existen varios días conflictivos en los cuales se han registrado observaciones aisladas de intensidad de viento superior al umbral de 40 km/h. Ocurre por ejemplo el 17 de mayo, cuando se registran velocidades inferiores a 30 km/h durante todo el día, excepto para un momento determinado en que se encuentra una observación aislada que supera el umbral de los 40 km/h y que como vemos en la Figura 34, ninguno de los dos modelos candidatos seleccionados es capaz de predecir.

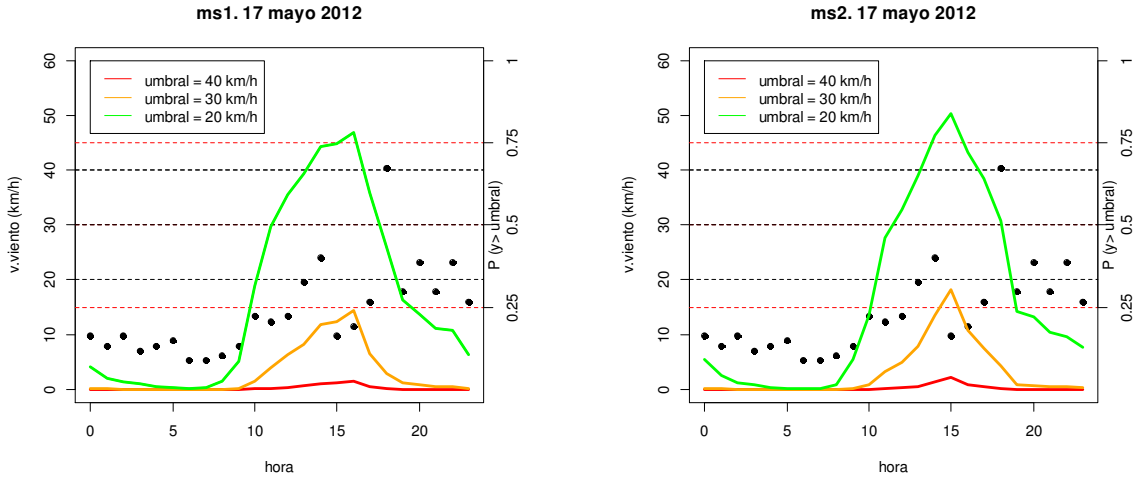


Figura 34. Como Figura 33 para pronóstico de *ms1* (izda.) y *ms2* (dcha.) el 17 de mayo de 2012.

Partiendo de la base de que el modelo de predicciones de AEMET no tiene un elevado grado de resolución a nivel local, y que los nuevos modelos, a pesar de estar adaptados al área y periodo en estudio, son un reescalado del mismo, se concluye que los nuevos modelos no serán capaces

de pronosticar eventos puntuales de carácter local que no queden previamente reflejados en las predicciones de AEMET. Por tanto, se deberán tener en cuenta también otras fuentes de predicción, en especial relativas a tormentas, capaces de originar grandes ráfagas puntuales de viento.

- Días en que se pronostican velocidades elevadas, aunque no mayores a 40 km/h

Se han observado varias situaciones para las cuales los modelos en estudio no han sido capaces de predecir umbrales de v> 40 km/h, pero sí de v>30 km/h. Esta situación se da por ejemplo el 25 de agosto (ver Figura 35), donde se observan velocidades de viento elevadas durante toda la mañana, que siguen una tendencia creciente hasta alcanzar un máximo que sobrepasa el umbral de los 40 km/h a las 17 horas, para después decrecer. Para este día, se observa que P(v>40 km/h) es de, a lo más, P=0.25. Sin embargo, P(v>30 km/h) llega a ser de 0.75 a las 15 horas.

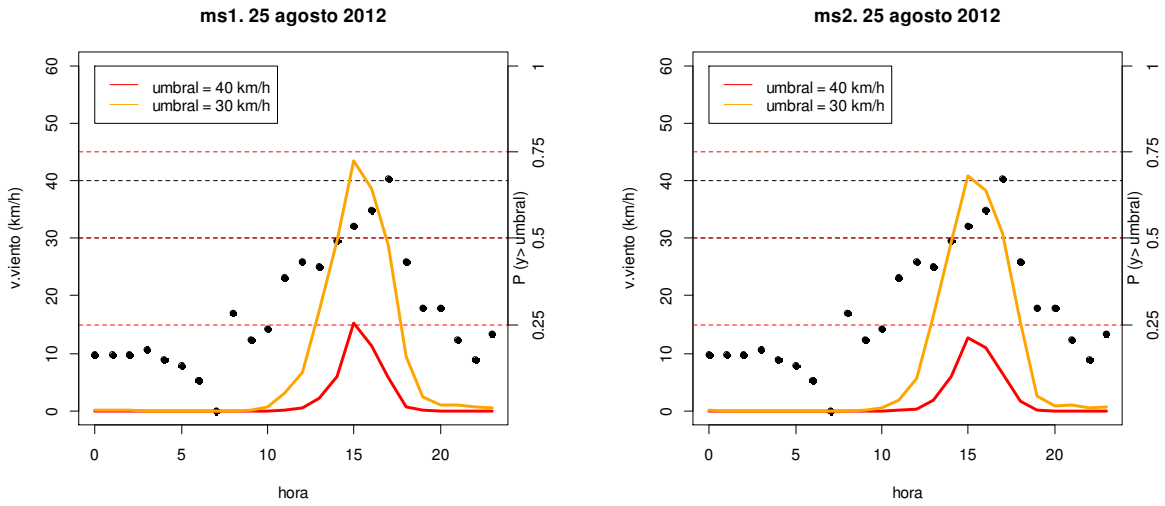


Figura 35. Como Figura 33 para pronóstico de *ms1* (izda.) y *ms2* (dcha.) el 25 de agosto de 2012.

En este caso no se observan eventos puntuales de carácter local (tormentas) y las predicciones de AEMET son capaces de reflejar la existencia de un ciclo marcado de velocidades elevadas. Por tanto, se cree que estudiando predicciones de AEMET actualizadas, por ejemplo a las 0 horas o a las 6 horas del propio día de interés, los modelos podrían llegar a predecir situaciones como la que se acaba de estudiar.

- **"Días parada" en los que el modelo *no* es capaz de predecir**

Por último, existen días conflictivos en los cuales, a pesar de observarse una cierta tendencia en el conjunto de datos y no darse las velocidades elevadas de forma aislada, los modelos no son capaces de recoger el comportamiento real del viento. Ocurre por ejemplo el 1 de julio (ver Figura 36), día en el que se observan elevadas velocidades (superiores a 30 km/h) durante un periodo amplio, sin que dichas velocidades se pronostiquen.

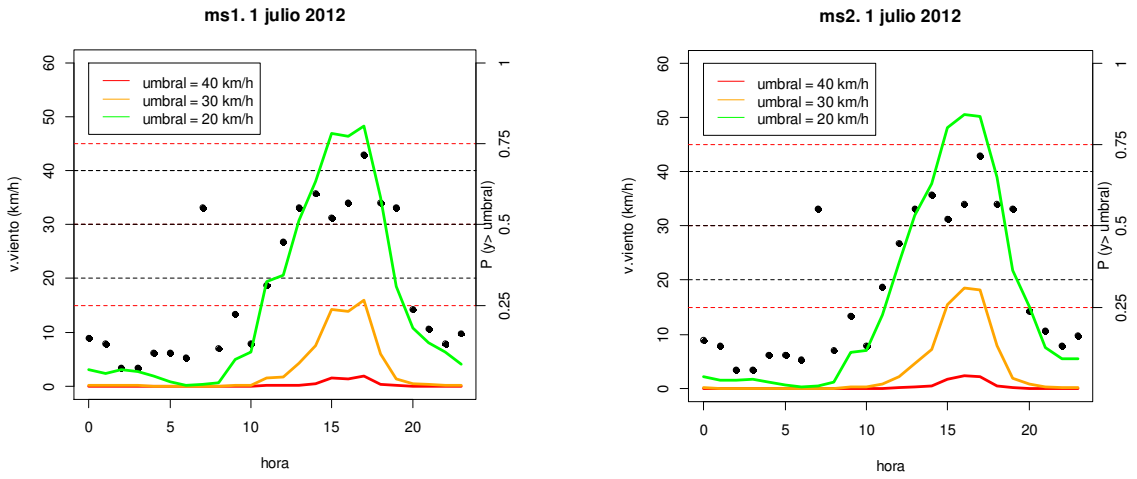


Figura 36. Como Figura 33, para el pronóstico de *ms1* (izda.) y *ms2* (dcha.) el día 1 de julio de 2012.

Ya que no se encuentra causa aparente para este fallo en el pronóstico, de nuevo surge la duda de si los modelos serían capaces de predecir dicha situación en caso de disponer de actualizaciones de las predicciones de AEMET.

- **Estudio del pronóstico del modelo en días aleatorios**

Como último paso en el estudio gráfico del pronóstico de los modelos candidatos, y previo al estudio por tablas de contingencia, se ha observado el funcionamiento de los modelos para días seleccionados de forma aleatoria. Los días propuestos son 12 y 21 de mayo, 21 y 24 de junio, 11 y 28 de julio, 1 y 18 de agosto, 11 y 23 de septiembre.

Como resultado, se ha obtenido que el modelo ha pronosticado correctamente un 50% de los casos estudiados (es decir, ha pronosticado con P>0.50, los intervalos de velocidad que se han dado en realidad). De entre ellos, para un 80% de los días la probabilidad ha superado el 0.75, y para los días restantes, el 0.5.

A continuación, se muestran como ejemplo los días 21 de junio, 18 de agosto y 11 de septiembre, recogidos en la Figura 36 *superior*, *centro* e *inferior* respectivamente. Para los tres días se observan velocidades superiores a 20 km/h e inferiores a 30 km/h, que los modelos han pronosticado correctamente, con P>0.75 para los días 21 de junio y 18 de agosto, y con P>0.5 para el 11 de septiembre.

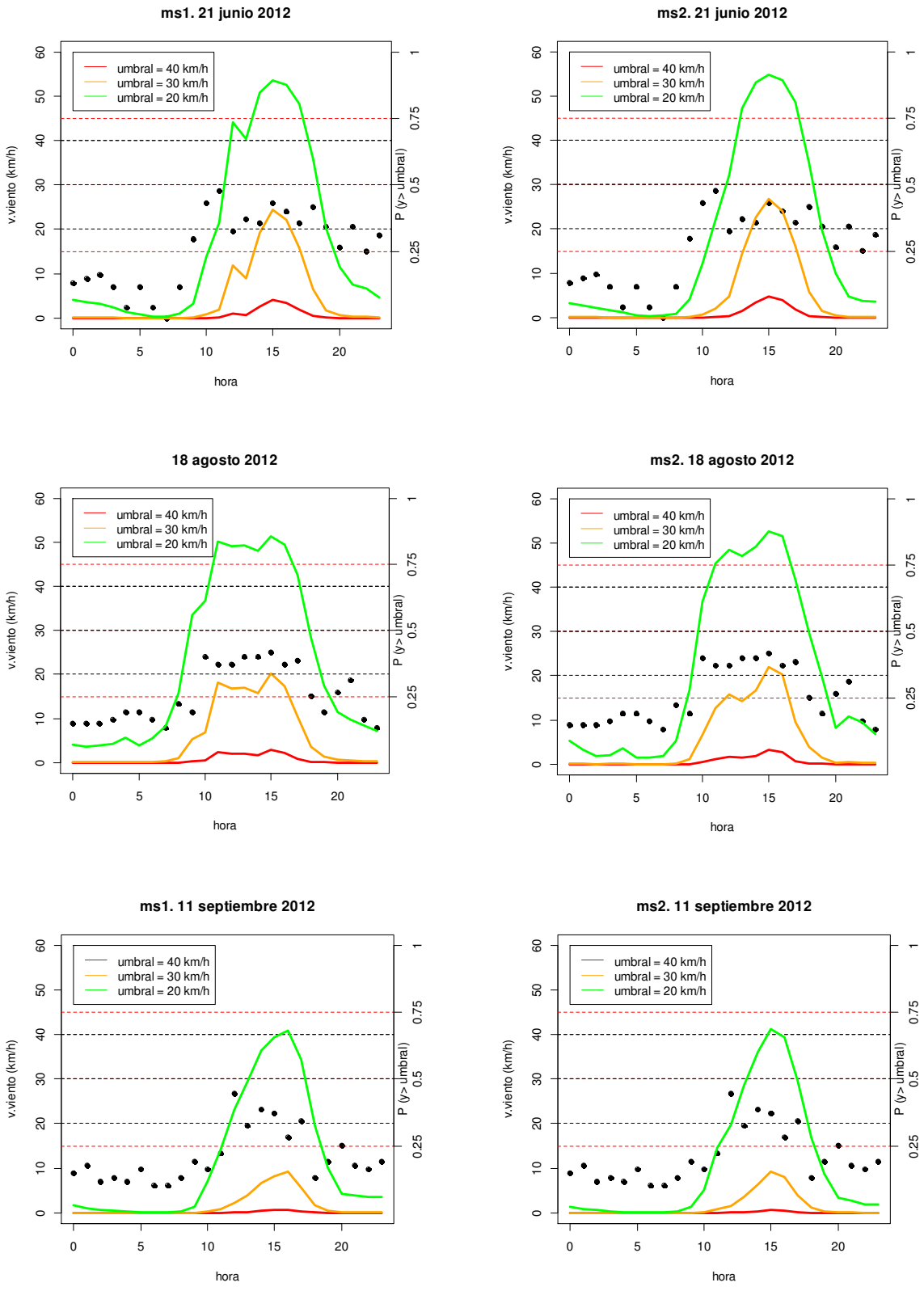


Figura 37. Velocidad observada instantánea y probabilidad estimada de superar viento umbral, a partir de los modelos *ms1* y *ms2*. Estudio de días seleccionados de forma aleatoria (21 de junio de 2012, *superior*; 18 de agosto de 2012, *central*; 11 de septiembre de 2012, *inferior*).

ESTUDIO DEL PRONÓSTICO MEDIANTE TABLAS DE CONTINGENCIA Y TABLAS DE OBSERVACIONES POR DÍAS

Siguiendo la metodología expuesta en el apartado 3.5, se han desarrollado dos tipos de tablas como parte del estudio del pronóstico de los modelos finalmente seleccionados, *ms1* y *ms2*.

En primer lugar se han realizado tablas de contingencia, para el estudio del comportamiento de los modelos de forma instantánea. En segundo lugar, y para estudiar su comportamiento "global" por días, es decir, si han sido capaces de pronosticar velocidades elevadas con P>0.5 en al menos un momento a lo largo del día, se han realizado un segundo tipo de tablas, que se han denominado "tablas de observaciones por días".

Recordemos que se ha considerado que el pronóstico de los modelos es correcto si son capaces de predecir un umbral de velocidad con una probabilidad superior a 0.5.

- **Estudio de tablas de contingencia**

Se han realizado tablas de contingencia para el estudio del comportamiento instantáneo de *ms1* y *ms2*, y para umbrales de velocidad de 40 y 30 km/h. El estudio del umbral de 20 km/h se ha recogido en el Anexo III sobre estudios adicionales. Se ha obviado el estudio de umbrales inferiores de velocidad, ya que no resultan de interés para este trabajo.

- Umbral > 40 km/h

El umbral de 40 km/h de velocidad real se supera en 27 ocasiones a lo largo del periodo en estudio, de las cuales el modelo *ms1* ha sido capaz de pronosticar 6 de forma instantánea, lo que supone un 22.22% del total de casos registrados; este porcentaje aumenta hasta el 33.33% para *ms2* (ver recuadros verdes en Tabla 13). Para ambos modelos, se han dado falsas alarmas en el intervalo de velocidades de 30 a 40 km/h, en un 4.3% del total de observaciones de dicho intervalo para *ms1,* y en un 0.8% para *ms2*. No existe ninguna falsa alarma para el resto de intervalos de velocidad, en ninguno de los dos modelos (ver recuadros azules en Tabla 13).

ms1	**Intervalos velocidad (km/h)**				
P(v>40 km/h)	**(0,10]**	**(10,20]**	**(20,30]**	**(30,40]**	**> 40**
(0, 0.25]	1357	1529	600	101	18
(0.25,0.50]	0	0	1	10	3
(0.50,0.75]	0	0	0	4	2
(0.75,1]	0	0	0	1	4
ms2					
(0, 0.25]	1357	1529	601	109	16
(0.25,0.50]	0	0	0	6	2
(0.50,0.75]	0	0	0	1	5
(0.75,1]	0	0	0	0	4

Tabla 13. Tabla de contingencia para pronósticos instantáneos de los modelos *ms1* (superior) *y ms2* (inferior), con umbral de velocidad > 40 km/h, y para la velocidad observada. En azul se muestran las falsas alarmas; en verde, las alarmas correctamente pronosticadas.

- Umbral v> 30 km/h

Existen 143 observaciones en las que se ha superado el umbral de velocidad de 30 km/h durante el periodo en estudio. El modelo *ms1* pronostica correctamente un 28.67% de ellas y el modelo *ms2* un 27.27% (ver recuadros verdes en Tabla 14); *ms1* pronostica una mayor cantidad de observaciones con P>0.75, lo cual se traduce un pronóstico más acertado. Las falsas alarmas (recuadros azules en las tablas) han sido de 19 para *ms1* y de 15 para *ms2*.

De las 27 observaciones superiores a 40 km/h, los modelos han pronosticado instantáneamente velocidades superiores a 30 km/h (por tanto, velocidades elevadas), en un 44.44% de los casos para *ms1*, y en un 48.15% de los casos para *ms2*. Es decir, este tipo de alarma puede ayudar a identificar la posible ocurrencia de "días parada".

ms1	**Intervalos velocidad (km/h)**				
P(v>30 km/h)	**(0,10]**	**(10,20]**	**(20,30]**	**(30,40]**	**40**
(0, 0.25]	1356	1520	538	70	9
(0.25,0.50]	0	7	46	17	6
(0.50,0.75]	1	2	16	17	3
(0.75,1]	0	0	1	12	9
ms2					
(0, 0.25]	1355	1517	530	68	8
(0.25,0.50]	2	11	57	22	6
(0.50,0.75]	0	1	14	21	2
(0.75,1]	0	0	0	5	11

Tabla 14. Como Tabla 13, para probabilidad de superar los 30 km/h.

- **Estudio de tablas de observaciones por días**

En las tablas propuestas a continuación, se recogen el número de días del periodo en estudio en que se han registrado un número determinado de velocidades que superan el umbral de estudio en cada caso (40, 30 km/h), divididas entre 0, 1-2, o más de 3. Cada uno de los días en estudio se coloca en la tabla según la probabilidad con la que los modelos han pronosticado observaciones superiores al umbral en estudio, en al menos una hora a lo largo del día.

De nuevo, en las tablas se han dibujado recuadros verdes, que contienen las alarmas correctamente pronosticadas por los modelos, así como recuadros azules cuando el modelo ha originado falsas alarmas.

- Umbral v>40 km/h

En la Tabla 15 se ha estudiado la probabilidad de superar el umbral de 40 km/h de los modelos *ms1* y *ms2.* Existe una falsa alarma para *ms1* (es decir, observaciones que el modelo ha pronosticado, serán superiores a 40 km/h, cuando en realidad no lo son), y ninguna para *ms2*. Para días en los que se han registrado una o dos observaciones superiores a 40 km/h (situación que ocurre en 14 ocasiones) ninguno de los modelos ha sido capaz de recoger dicho comportamiento con P>0.5. Sin embargo, ambos modelos son capaces de predecir correctamente un 66.66% de los días que han registrado entre 3 y 5 observaciones problemáticas.

ms1	**Nº horas/día con observaciones > 40 km/h**		
Intervalos P(v>40 km/h)	**0**	**1-2**	**3-5**
(0, 0.25]	134	11	0
(0.25,0.50]	1	3	1
(0.50,0.75]	1	0	1
(0.75,1]	0	0	1
ms2			
(0, 0.25]	136	13	0
(0.25,0.50]	0	1	1
(0.50,0.75]	0	0	1
(0.75,1]	0	0	1

Tabla 15. Clasificación según probabilidad de superar los 40 km/h, de los modelos *ms1* y *ms2,* para número de días con un número determinado de observaciones > 40 km/h

Dados los resultados del estudio gráfico del funcionamiento de *ms1* y *ms2*, los días con una o dos observaciones corresponderán en general a situaciones de tormenta. También atendiendo dichos resultados, se puede suponer que algunas de las observaciones que han superado el umbral de los 40 km/h, habrán sido pronosticadas por los modelos con al menos, P(v>30 km/h)>0.50. Se ha estudiado dicha hipótesis en la Tabla 16, donde se observa que el porcentaje de pronóstico para 1-2 observaciones aumenta, pasando de un 0 a un 28.6% para ambos modelos, mientras que para más de 3 observaciones, el porcentaje de acierto pasa del 66.66% al 100%, también para ambos modelos.

En resumen, los modelos han sido capaces de pronosticar velocidades elevadas (v>30 km/h) en al menos una hora, para todos los días en que se dan 3 observaciones o más superiores al umbral de 40 km/h.

En la Tabla 16, también se observan 8 falsas alarmas para *ms1* y 7 para *ms2*; en 5 se supera realmente el umbral de los 30 km/h, aunque no para las otras 3 (ver Tabla 17).

ms1	**Nº horas/día con observaciones >40 km/h**		
Intervalos P(v>30 km/h)	**0**	**1-2**	**3-5**
(0, 0.25]	113	5	0
(0.25,0.50]	13	5	0
(0.50,0.75]	8	3	0
(0.75,1]	0	1	3
ms2			
(0, 0.25]	111	5	0
(0.25,0.50]	18	5	0
(0.50,0.75]	6	3	0
(0.75,1]	1	1	3

Tabla 16. Pronóstico de los modelos *ms1* y *ms2* para número de días con un número determinado de observaciones > 40 km/h, estudiando P(v>30 km/h).

- Umbral v>30 km/h

En la Tabla 17 se ha recogido el estudio del pronóstico de los modelos *ms1 y ms2*, por días y número de alarmas, para un umbral de velocidad de 30 km/h. Como se ha comentado anteriormente, ambos modelos pronostican 3 falsas alarmas para esta situación. Los modelos

son capaces de pronosticar con P>0.5, días con 1-2 observaciones con un 11.8% *ms1*, y con un 5.9% *ms2*. Para días con más de 3 observaciones, los porcentajes aumentan hasta un 62.5% en *ms1* y un 56.25% en *ms2*. Por tanto, se ve que para este umbral de velocidad, el pronóstico de *ms1* es ligeramente mejor.

ms1	**Nº horas/día con observaciones > 30 km/h**		
Intervalos P(v>30 km/h)	**0**	**1-2**	**3-13**
(0, 0.25]	92	25	1
(0.25,0.50]	8	5	5
(0.50,0.75]	3	3	5
(0.75,1]	0	1	5
ms2			
(0, 0.25]	89	25	1
(0.25,0.50]	11	7	6
(0.50,0.75]	3	2	4
(0.75,1]	0	0	5

Tabla 17. Pronóstico de los modelos *ms1* y *ms2* para número de días con un número determinado de observaciones > 30 km/h

Como resumen final del estudio del pronóstico de los modelos óptimos seleccionados (*ms1* y *ms2*), y en especial atendiendo a las observaciones que han superado el umbral de los 40 km/h a lo largo del periodo en estudio, se puede concluir lo siguiente:

- Cuando el modelo pronostica velocidades superiores a 40 km/h, en forma de un marcado ciclo diario, puede suponerse que dicha predicción se cumplirá.

- Se ha hallado que para velocidades observadas superiores a 40 km/h, los modelos han sido capaces de pronosticar, con P(v>30 km/h)>0.5, entre un 44.44% (*ms1*) y un 48.15% (*ms2*) de ellas, de forma instantánea; y entre un 28.6% (para 1-2 observaciones) y un 100% (para 3 o más), de forma diaria, es decir, para al menos una hora del día en que se han producido dichas observaciones. Por tanto, se concluye que si los modelos pronostican velocidades superiores a 30 km/h, habrá que prestar atención a dicho día y a ser posible, actualizar el modelo finalmente seleccionado con nuevas predicciones de AEMET (disponibles por ejemplo a las 0 horas del propio día pronosticado), para estudiar si éstas podrían llegar a superar el umbral de los 40 km/h.

- Las dificultades que han presentado los modelos en la predicción de 1-2 observaciones diarias superiores a 40 km/h (e incluso 30 km/h), hacen pensar que éstas corresponden a eventos súbitos, tipo tormentas estivales, de rápida aparición y que producen grandes ráfagas de viento. Por su carácter local, y teniendo en cuenta que los modelos se han construido a partir del modelo de predicciones de AEMET, el cual no tiene resolución espacial y temporal suficiente, dichas situaciones de viento no podrán ser reflejadas por los modelos construidos. Por ello, se deberá prestar especial atención a los pronósticos de tormentas que operativamente pueda establecer AEMET.

4.3 SELECCIÓN DEL MODELO ÓPTIMO, ESTUDIO DE SU CAPACIDAD OPERATIVA E IMPLEMENTACIÓN DEL PROCEDIMIENTO

No se han encontrado diferencias relevantes entre los pronósticos aportados por *ms1* y *ms2*, a excepción de que en general, *ms1* calcula probabilidades más elevadas, y obtiene mejores resultados en el pronóstico de velocidades superiores a 30 km/h. Por último, *ms1* funciona mejor para el día 30 de agosto (uno de los días conflictivos, ver Figura A. 40 en el Anexo I), siendo de entre los "días parada" estudiados, el que ofrece una diferencia más relevante entre *ms1* y *ms2*.

Se concluye así que la mayor complejidad del modelo *ms2* (respecto a número de términos y existencia de polinomios) no ayuda a mejorar la predicción de la respuesta. Por tanto, se ha considerado a *ms1* como modelo óptimo final.

Para finalizar el trabajo, y conforme al apartado 3.6, se ha realizado un estudio operativo del modelo óptimo seleccionado, consistente en un estudio gráfico, y en la realización de una función en R que permitirá aplicar el modelo seleccionado operativamente a partir de las predicciones de AEMET.

Para la comparación gráfica, se han realizado dos gráficos de dispersión. En la Figura 38 (izda.), se recoge el gráfico realizado entre los datos reales de velocidad (transformados en raíz) y los datos ajustados por el MRLS desarrollado a partir del modelo de AEMET, para la respuesta transformada. En la Figura 38 (dcha.) se recoge el gráfico de dispersión entre datos reales (también transformados) y datos ajustados por el modelo seleccionado (*ms1*).

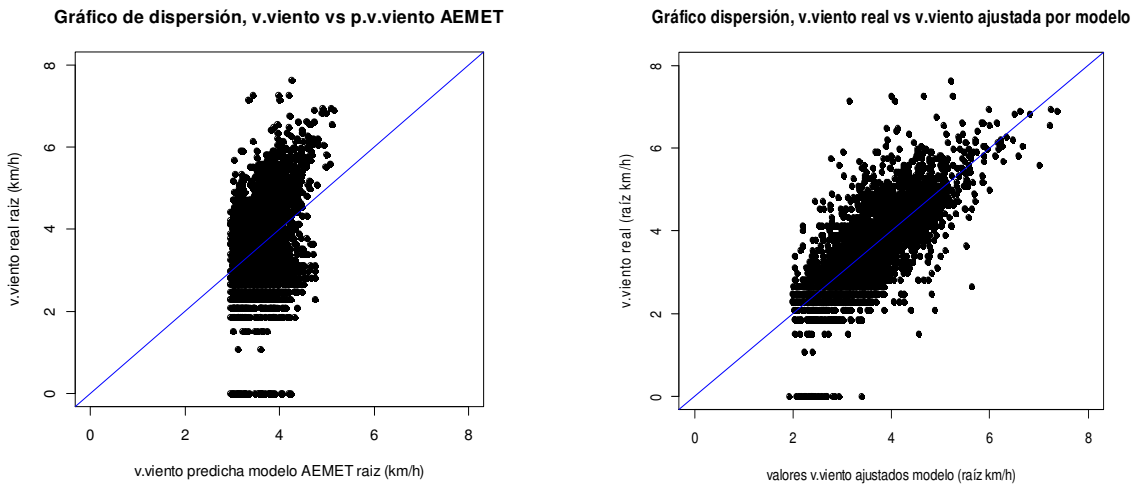


Figura 38. Gráfico de dispersión realizado entre datos iniciales y finales. A la izda. se muestra el gráfico de dispersión entre datos reales y ajustados por MRLS de AEMET; a la dcha., el correspondiente a datos reales y ajustados por *ms1.*

Se puede verificar que el nuevo modelo de predicciones *ms1* obtiene un ajuste mucho más apropiado, concentrado alrededor de la recta de identidad. Por el contrario, el ajuste del MRLS desarrollado únicamente con predicciones de AEMET resulta en un gráfico mucho más disperso

cuyos puntos no siguen la identidad, reflejando así la deficiente reproducción de la realidad observada y la escasa variabilidad que el modelo AEMET es capaz de recoger.

Como finalización de este trabajo, y para permitir el uso real del modelo seleccionado *ms1* se ha desarrollado un procedimiento en R. Dicho procedimiento necesita de una hoja Excel en formato .csv, en la cual se deben incluir las nuevas predicciones de AEMET a partir de las cuales se quiere obtener el pronóstico a 24 horas. En esta hoja Excel deben incluirse tres columnas:

- Fecha y hora.
- Predicción de AEMET de velocidad de viento para la fecha y hora determinadas.
- Predicción de AEMET de dirección de viento para la fecha y hora determinadas.

Para que el procedimiento funcione correctamente, las predicciones deberán corresponder a un periodo de 23 horas, desde las 3 am del día para el que se tienen las predicciones. Por supuesto, el procedimiento puede modificarse para trabajar con periodos distintos de tiempo.

Como resultado, se obtendrán predicciones para el periodo entre las 7 y las 21 horas del día de interés. Los datos ajustados por el modelo (es decir, las predicciones instantáneas), vendrán incluidas en un vector. El procedimiento creará también un gráfico de pronóstico, del tipo de los estudiados en el apartado anterior (ver Figura 33). En dicho gráfico vendrán incluidas las predicciones instantáneas ajustadas por el modelo para cada hora (en el periodo entre las 7 y las 21 horas), así como las curvas de representación de probabilidad de superación de umbrales estudiados.

En el Anexo III se recoge el *script* de R que permite el funcionamiento del procedimiento desarrollado.

5. CONCLUSIONES

Se ha realizado a lo largo de este TFG, en primer lugar, un análisis exploratorio del régimen de vientos en el vertedero de Bailín, para el periodo de mayo a septiembre de 2012, así como un estudio que demuestra la falta de adecuación a dicho régimen de vientos del modelo de predicciones de AEMET existente.

Posteriormente, y conforme al objetivo general de este TFG, se han desarrollado modelos estadísticos de predicción de la velocidad del viento en forma de modelos de regresión (MR), que han permitido el *downscaling* del modelo de predicciones de AEMET, adaptándolo a las condiciones específicas del área y el periodo de estudio.

De los análisis exploratorios se han extraído las siguientes conclusiones:

1) Se ha caracterizado en los datos observados de velocidad de viento, un ciclo diario de brisas y una estacionalidad, aunque no una tendencia.

2) Se ha establecido que el modelo de predicciones de AEMET existente funciona mejor para el periodo de invierno. Sin embargo, se ha comprobado que no resulta útil de forma directa para verano, ya que además de proporcionar valores con menor dispersión, no es capaz de reflejar la estacionalidad ni el ciclo diario existentes. Las obras de desmantelamiento y transferencia del vertedero de Bailín pueden ampliarse hasta el mes de octubre. Sin embargo, se ha decidido prescindir de dicho mes en el ajuste del nuevo modelo, ya que se considera que las predicciones de AEMET para dicho mes son apropiadas; además, los datos de octubre podrían influir en el modelo construido, limitando su capacidad predictiva para el resto de meses.

3) Los análisis exploratorios han descrito las características del viento observado y la reproducción que de ellas hacen las predicciones de AEMET. Esta fase ha proporcionado los argumentos para la construcción de la base de covariables utilizada en el desarrollo del MR óptimo buscado, y que contiene covariables de predicción de velocidad y dirección del viento (instantáneas y decaladas), así como covariables que reflejan ciclos diarios y estacionalidad, a partir de las cuales se han construido interacciones.

De la metodología desarrollada para la construcción de MR, se han extraído las siguientes conclusiones:

1) El uso general de procedimientos desarrollados en R, y en particular de algoritmos para la construcción de modelos de predicción candidatos, ha optimizado el tiempo invertido, permitiendo estudiar una gran cantidad de modelos con diferentes estructuras que de otra forma habría resultado imposible.

2) Para la selección de modelos candidatos, se ha intentado hallar un equilibrio entre la complejidad de los mismos y otros criterios de selección estudiados, como la bondad de

ajuste, la desviación típica residual, la desviación típica de validación cruzada, o el resultado de los test ANOVA para la selección entre modelos anidados.

Las conclusiones sobre los resultados obtenidos en la construcción de los modelos de predicción de la velocidad del viento candidatos, han sido las siguientes:

1) Las covariables que han resultado más significativas para explicar la variabilidad de la velocidad del viento real, han sido, como era de esperar por los análisis exploratorios, las covariables de predicción de la dirección del viento, tanto instantánea (extraída del modelo de AEMET) como decaladas, así como los armónicos de ciclo diario, que han aumentado la explicación de la variabilidad en un 18% y un 13% respectivamente.

2) Se ha profundizado en el estudio del comportamiento de los modelos finalmente seleccionados (*ms1* lineal y *ms2* polinómico), para situaciones en que se supera el umbral de los 40 km/h, velocidad a la que según el protocolo de actuación, deberían detenerse las obras. Se ha concluido que si el modelo pronostica velocidades superiores a 40 km/h, dicha predicción se cumplirá. En caso de pronosticarse velocidades superiores a 30 km/h, se deberá prestar atención a dicho día y a ser posible, actualizar el modelo finalmente seleccionado con nuevas predicciones de AEMET (disponibles por ejemplo a las 0 horas del propio día pronosticado), pues dichas observaciones son susceptibles de superar el umbral de los 40 km/h. El comportamiento de los modelos es menos satisfactorio para situaciones puntuales, como tormentas estivales, en las cuales se levantan grandes ráfagas de viento de forma súbita y de escasa duración. Por tanto, se concluye también que paralelamente al estudio del pronóstico de los modelos, se debe prestar especial atención a los pronósticos de tormentas que operativamente pueda establecer AEMET.

3) Se ha comprobado que las diferencias entre los pronósticos ofrecidos por ambos modelos estadísticos son mínimas. Por tanto, se ha concluido que los términos polinómicos, así como en general, la mayor complejidad del modelo *ms2*, no se traduce en una mejor predicción.

4) El modelo *ms1* (lineal) finalmente seleccionado, está compuesto de 23 parámetros (13 covariables y 10 interacciones), y es capaz de explicar un 57% de la variabilidad de la velocidad del viento para el área del vertedero de Bailín. Este porcentaje representa una mejora muy importante si se tiene en cuenta que el MRLS construido a partir de las predicciones de AEMET, únicamente era capaz de explicar el 14% de dicha variabilidad. A la hora de juzgar este porcentaje, se debe tener en cuenta también la dificultad que supone trabajar con datos meteorológicos, cuyas características hacen de su predicción una ardua tarea.

5) Para que el modelo construido pueda ser utilizado, se ha desarrollado un procedimiento en R, que mediante la introducción de las predicciones de AEMET de interés (para un periodo de 23 horas) en una hoja Excel, es capaz de ofrecer un estudio gráfico del pronóstico del modelo, en términos de predicciones instantáneas y de probabilidad de ocurrencia de distintos umbrales de velocidad del viento.

Para resumir, a lo largo de este TFG se ha desarrollado una metodología de construcción de un modelo de predicción de la velocidad del viento, que ha permitido el *downscaling* de la predicción AEMET, el cual se ha aplicado en Bailín con resultados satisfactorios. El modelo desarrollado pretende ser una herramienta más en la lucha contra la contaminación de áreas adyacentes al vertedero de Bailín a causa de la dispersión atmosférica, lo cual incrementaría la problemática ambiental existente.

- **Posibles trabajos futuros**

Teniendo en cuenta el carácter herramental que poseen los modelos de regresión para distintas disciplinas, la metodología desarrollada puede servir para resolver otro tipo de problemas, en los cuales se necesite predecir cualquier respuesta a partir de un conjunto de covariables predictoras.

Para este TFG en concreto, se han resumido una serie de recomendaciones en vistas a una posible realización de trabajos futuros que sigan la línea del estudio realizado. Las recomendaciones que se sugieren son las siguientes:

- En estudios incluidos en el Anexo III, se ha visto cómo la temperatura real muestra una clara relación con la velocidad del viento. Lamentablemente, a la hora de construir el modelo no se ha podido disponer de datos de predicción de la temperatura. La introducción de esas (u otras) variables atmosféricas para la zona, permitirían ajustar mejor el modelo, obteniendo pronósticos más acertados.

- Se cree que trabajando con predicciones AEMET actualizadas a más corto plazo, se obtendrían predicciones más acordes a la realidad (en especial en lo que a velocidades elevadas se refiere), tanto de forma instantánea como para el estudio del pronóstico diario. Es decir, si el modelo trabajase operativamente, nutriéndose de forma continua de nuevas predicciones de AEMET actualizadas, podría mejorar las probabilidades obtenidas.

AGRADECIMIENTOS

Para finalizar, me gustaría dar las gracias a todas las personas que de una manera u otra han hecho posible este proyecto.

A mis directores. A Jesús Asín, por su paciencia y todo el tiempo que ha invertido; porque sin su ayuda este proyecto no habría podido salir adelante; por ser un gran profesor y por haberme transmitido el entusiasmo por la estadística. A Jesús Fernández, por haberme ofrecido este trabajo; agradezco su gran interés y su confianza en mí. A Beatriz Lacruz, por el interés que desde el principio mostró en este proyecto, y por la ayuda que me ha prestado.

A mi familia, por ser un gran soporte en mi vida, por animarme en todo lo que hago, por sus consejos y su apoyo incondicional. A Patrick, por estar siempre a mi lado a pesar de la distancia, por escucharme y comprenderme, y por apoyarme en todas mis decisiones. Y a mis amigos, los que están cerca y los que están lejos; por ayudarme y animarme en los momentos difíciles.

BIBLIOGRAFÍA

Alfons, A. (2012). A toolkit for cross-validation: The R package cvTools. *useR! The 8th International R User Conference*, June 12-15, 2012, Nashville, Tennessee, USA

Autorización Ambiental Integrada de la fase B del vertedero de Bailín, en el término municipal de Sabiñánigo (Huesca). BOA nº 134, de 14 de Julio de 2009.

Bodenstein, G. (1972). Disposal of wastes from Lindane manufacture. In: Uhlmann, E., Verlag, K. (eds) *Lindane monograph of an insecticide*. Schillinger, Freiburg im Breisgau, pp 23-77.

Breivik K., Pacyna J., Munch J. (1999). Use of α-, β-, γ- hexachlorocyclohexane in Europe, 1970-1996. *Sci. Total Environ.*, *239* (1-3)*:*151-163.

Chalvatzaki, E., Kontaksakis, M., Glytsos, T., Kalogerakis, N., Lazaridis, M. (2010). Measurements of particulate matter concentrations at a landfill site (Crete, Greece). *Waste Management*, *30*, 2058-2064

Chalvatzaki, E., Aleksandropoulou, V., Glytsos, T., Lazaridis, M. (2012). The effect of dust emisions from open storage piles to particle ambient concentration and human exposure. *Waste Management*, *32*, 2456-2468

Convención de Estocolmo (2009). Report of the conference of the parties of the Stockholm Convention on persistent organic pollutants on the work of its fourth meeting. UNEP/POPS/COP.4/38, 8 May

Directiva 2008/98/CE del Parlamento Europeo y del Consejo de 19 de Noviembre de 2008, sobre los residuos y por la que se derogan determinadas Directivas. Diario Oficial de la Unión Europea (DOUE L 312/3 de 22-11-2008)

Fernández, A. (2004). *Las sustancias tóxicas persistentes*. Méjico: Secretaría de Medio Ambiente y Recursos Naturales. Instituto Nacional de Ecología (pp 115)

Fernández J., Arjol M., Cacho, C. (2012). POP-contaminated sites from HCH production in Sabiñánigo, Spain. *Environmental Science and Pollution Research*, *20* (4):1937-1950

Forter, M. (1995). Umweltnutzung durch die chemische Industrie am Fall Beispiel der HCH-Fabrik Ugine- Kuhlmann, Hüningen (France), Lizentiatsarbeit, 1995, Universität Basel, Schweiz

Fox, J. and Weisberg, S. (2011). *An R Companion to Applied Regression*. 2nd edition. SAGE Publications Inc.

Giusti, L. (2009). A review on waste management practices and their impact on human health. *Waste Management*, *29* (8):2227-2239

Götz, R., Sokollek, V., Weber, R. (2013). The dioxin/POPs legacy of pesticide production in Hamburg: part 2—waste deposits and remediation of Georgswerder landfill. *Environ. Sci. Pollut. Res. Int., 20* (4):1925-36. DOI: 10.1007/s11356-012-0986-x

Hastie, T., Tibshirani, R. (1990) *Generalized Additive Models*. London: Chapman and Hall.

James, G., Witten, D., Hastie, T., Tibshirani, R. (2013). *An introduction to statistical learning with applications in R*. New York: Springer Science. DOI 10.1007/978-1-4614-7138-7

Jit, S., Dadhwal, M., Kumari, H., Jindal, S., Kaur, J., Lata, P., Niharika, N., Lal, D., Garg, N., Gupta, S.K., Sharma, P., Bala, K., Singh, A., Vijgen, J., Weber, R., Lal, R. (2010). Evaluation of hexachlorocyclohexane contamination from the last lindane production plant operating in India. *Environ. Sci. Pollut. Res. Int., 18* (4):586–597. DOI: 10.1007/s11356-010-0401-4

Kopanakis, I., Chalvatzaki, E., Kalogerakis, N., Lazaridis, M. (2011). Particulate matter concentration and chemical composition during excavation woks for a restoration of an uncontrolled landfill. *Proceedings of the 12th International Conference on Environmental Science and Technology*. Rhodes, Greece.

Lakes Environmental, (2011). WRPLOT View™ – Freeware. Wind Rose Plots for Meteorological Data. Free Software, retrieved from http://www.weblakes.com/products/wrplot/?AspxAutoDetectCookieSupport=1

Li, Y.F., Cai, D.J., Singh, A. (1998). Technical Hexachlorocyclohexane Use Trends in China and Their Impact. *Arch. Environ. Contam. Toxicol*, *35* (4), 688-697.

Montgomery, D.C., Peck, E.A., Vining, G.G. (2012). *Introduction to linear regression analysis*, 5th edition. New York: John Wiley & Sons.

Pepió, M. (2011). Autocorrelación en *Series temporalis,* 2ª edición. Barcelona: Edicions UPC.

Pope, C.A., Bates, D.V., Mark, E. (1995). Health Effects of Particulate Air Pollution: Time for Reassessment? *Environmental Health Perspectives, 103* (5):472-480

R Development Core Team (2012). R: A language and environment forstatistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/

Tadeo, J.L. (2008). *Analysis of Pesticides in Food and Environmental Samples*. FL: Taylor & Francis Group.

Torres, J., Fróes-Asmus, C. I. R., Weber, R., Vijgen, J. M. H. (2012). HCH contamination from former pesticide production in Brazil—a challenge for Stockholm Convention implementation. *Environ. Sci. Pollut. Res. Int., 20* (4):1951-7. DOI:10.1007/s11356-012-1089-4

Venables, W.N., Ripley, B.D. (2002). Modern Applied Statistics with S, 4th edition. Springer.

Vijgen, J. (2006). The legacy of lindane HCH isomer production - Main report. *IHPA.* En http://ew.eea.europa.eu/Agriculture/Agreports/obsolete_pesticides/lindane_production.pdf

Vijgen, J., Abhilash, P.C., Li, Y., Lal, R., Forter, M., Torres, J., Singh, N., Yunus, M., Tian, C., Schäffer, A., Weber, R. (2011). Hexachlorocyclohexane (HCH) as new Stockholm Convention POPs—a global perspective on the management of Lindane and its waste isomers. *Environ. Sci. Pollut. Res. Int., 18* (2):152–162. DOI: 10.1007/s11356-010-0417-9

Voldner, E.C., Li, Y.F. (1995). Global usage of selected persistent organochlorines. *Sci. Total Environ., 160/161*, 201–210. DOI: 10.1016/0048-9697(95)04357-7

Walker, K., Vallero, D.A., Lewis, R.G. (1999). Factors influencing the distribution of lindano and other hexachlorocyclohexanes in the environment. *Environmental Science & Technology, 33* (24):4373-4378.

Weber, R., Varbelow, G. (2012). Dioxin/POPs legacy of pesticide production in Hamburg: part 1—securing of the production area. *Environ. Sci. Pollut. Res. Int., 20* (4):1918-24. DOI: 10.1007/s11356-012-1011-0

Willet, K.L., Utrich, E.M., Hites, R.A. (1998). Differential toxicity and environmental facts of hexachlorocyclohexane isomers. *Environmental Science & Technology*, *32* (15):2197-2207. DOI:10.1021/es9708530

Wycisk, P., Stollberg, R., Neumann, C., Gossel, W., Weiss, H., Weber, R. (2012). Integrated methodology for assessing the HCH groundwater pollution at the multi-source contaminated mega-site Bitterfeld/Wolfen. *Environ. Sci. Pollut. Res. Int, 20*, 1907-1917. DOI:10.1007/s11356-012-0963-4

ANEXO I. FIGURAS

ANÁLISIS EXPLORATORIO

	Residuo	
	± 5 km/h	**± 10 km/h**
P1	42,40%	15,25%
P2	35,21%	10,92%

Tabla A. 1. Porcentaje de residuos de ± 5 km/h y de ± 10 km/h, tras aplicar P1 y P2 a BD de velocidad media.

	Residuo	
	± 5km/h	**± 10 km/h**
P1	52,75%	17,78%
P2	61,25%	24,18%

Tabla A. 2. Porcentaje de residuos de ± 5 km/h y de ± 10 km/h, tras aplicar P1 y P2 a BD de velocidad máxima.

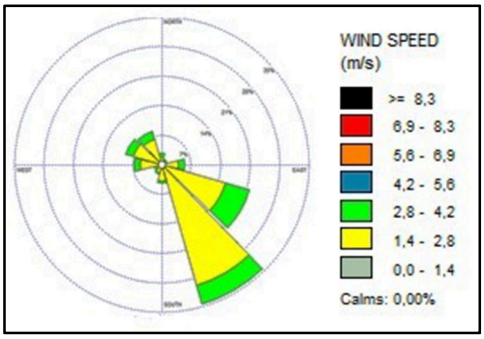


Figura A. 3a.

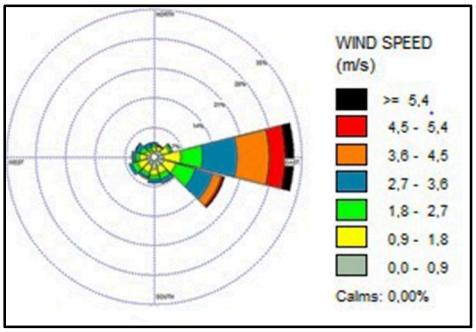


Figura A. 3b

Figura A. 3. Estudio de rosas de vientos para intervalo de calmas. *Situación 1*. Figura A. 3a muestra datos reales y Figura A. 3b datos de predicción instantánea correspondientes.

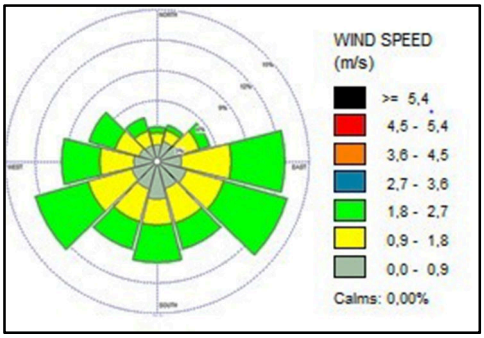


Figura A. 4a

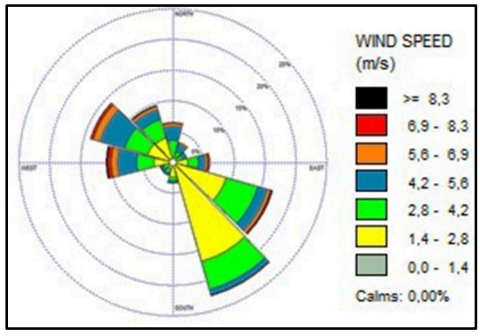


Figura A. 4b

Figura A. 4. Estudio de rosas de vientos para intervalo de calmas. *Situación 1*. Figura A. 4a contiene datos reales y Figura A. 4b datos de predicción instantánea correspondientes.

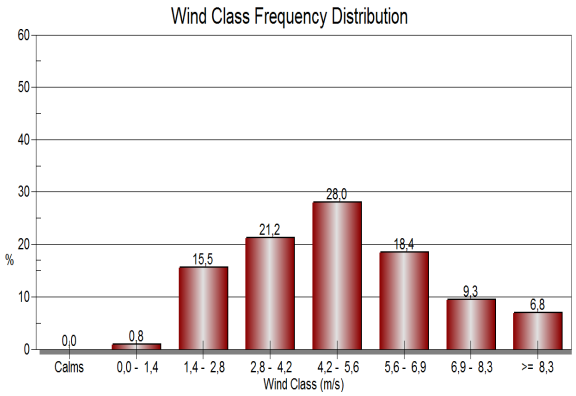


Figura A. 5a

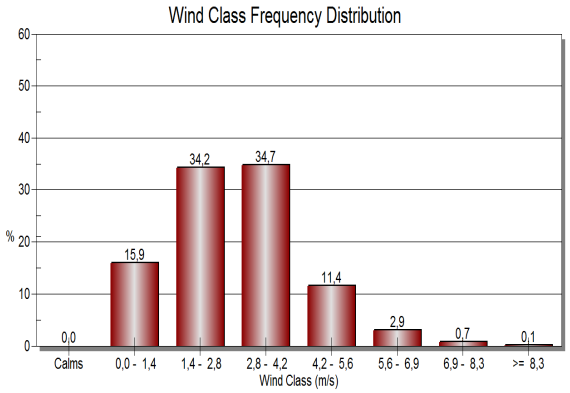


Figura A. 5b.

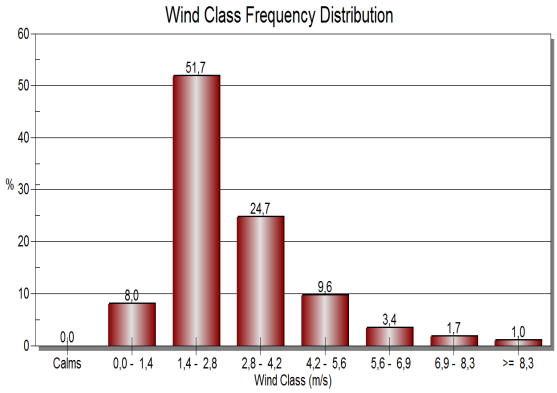


Figura A. 5c.

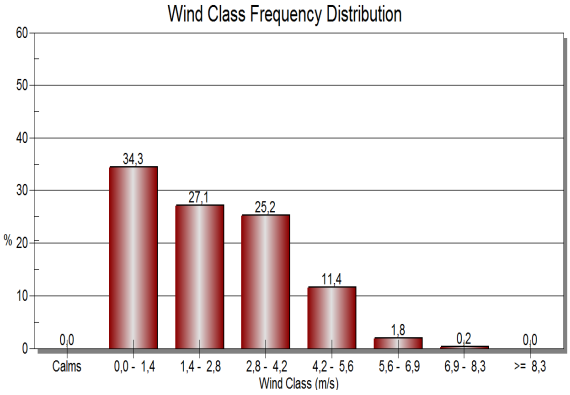


Figura A. 5d.

Figura A. 5. Estudio de la distribución de frecuencias de intervalos de velocidad del viento, para datos reales y de predicción, y para horario diurno y nocturno. Figura A. 5a y Figura A. 5b. muestran distribución de frecuencias para horario diurno, y para datos reales y predichos respectivamente; Figura A. 5c. y Figura A. 5d. muestran distribución de frecuencias para horario nocturno, para datos reales y predichos respectivamente.

DESARROLLO DE MODELOS DE REGRESIÓN

Periodo considerado	**R^2 ajustado**
BD verano (mayo-septiembre)	0.1681
BD verano + octubre	0.2343
BD anual	0.3082
BD octubre	0.5138

Tabla A. 6. Bondad de ajuste de los MRLS desarrollados con el fin de estudiar la adecuación del modelo de AEMET a distintos periodos, ordenados de menor a mayor bondad de ajuste.

```
> anova(ensayo4, ensayo4.1, test='F')
Analysis of Variance Table

Model 1: y.raiz ~ x + x2 + x.p2 + sendv + send.p1 + cosd1 + cosd3 + cosd.p4 +
    cos.dia + sen.hora + cos.hora
Model 2: y.raiz ~ x + x2 + x4 + x.p2 + sendv + send3 + send.p1 + cosd1 +
    cosd4 + cosd.p3 + cosd.p4 + sen.dia + cos.dia + sen.hora +
    cos.hora
  Res.Df    RSS Df Sum of Sq     F    Pr(>F)
1   3652 2029.5
2   3648 2016.3  4     13.13 5.939 9.235e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figura A. 7. Test ANOVA entre ensayos *4* y *4.1* (en orden de aparición). El *ensayo4.1* resulta altamente significativo.

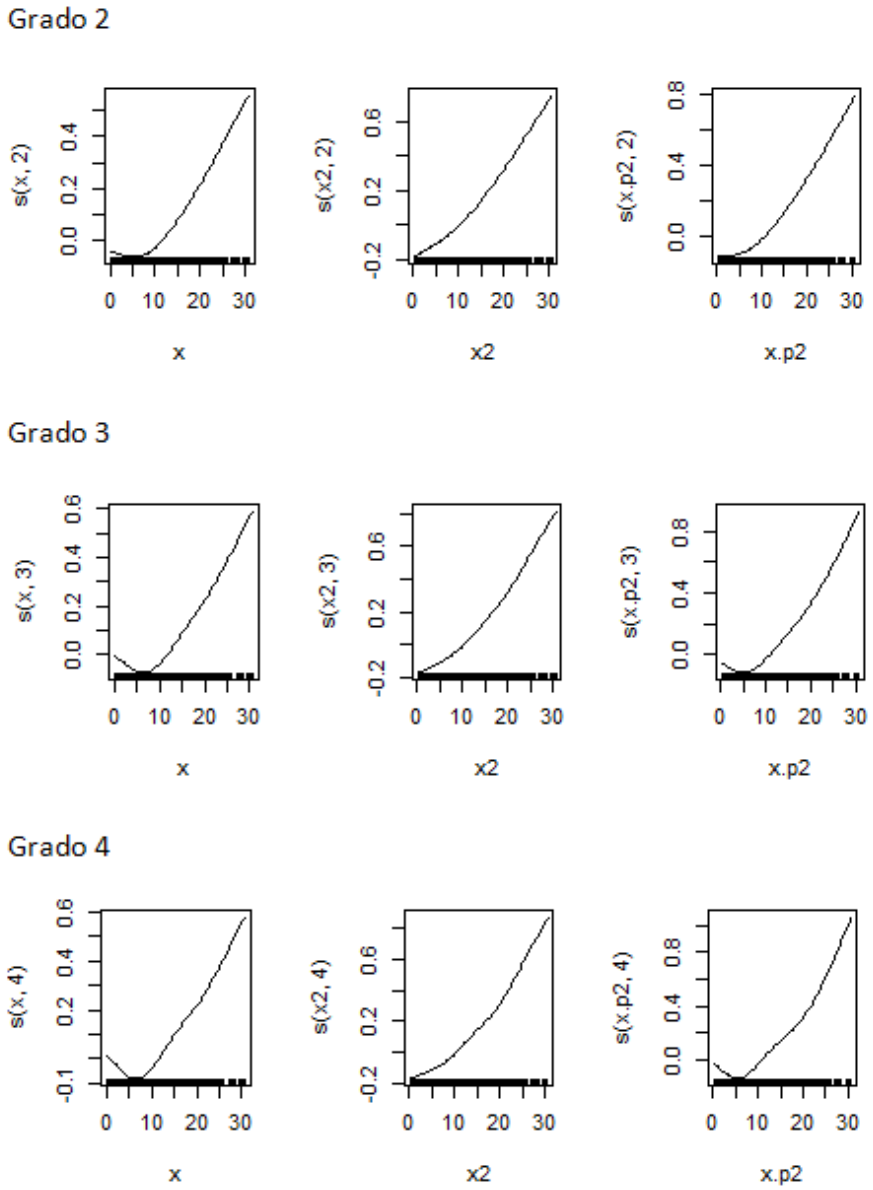


Figura A.8. Estudio gráfico de los grados más apropiados para las covariables de predicción de viento del *ensayo4*, mediante la función *gam* en R.

```
Model 1: y.raiz ~ s(x, 2) + x2 + s(x.p2, 2) + send.p1 + cosd1 + cosd3 +
    cosd.p4 + cos.dia + sen.hora + cos.hora
Model 2: y.raiz ~ s(x, 2) + s(x2, 2) + s(x.p2, 2) + send.p1 + cosd1 +
    cosd3 + cosd.p4 + cos.dia + sen.hora + cos.hora
Model 3: y.raiz ~ s(x, 2) + s(x2, 2) + s(x.p2, 3) + send.p1 + cosd1 +
    cosd3 + cosd.p4 + cos.dia + sen.hora + cos.hora
Model 4: y.raiz ~ s(x, 3) + s(x2, 2) + s(x.p2, 3) + send.p1 + cosd1 +
    cosd3 + cosd.p4 + cos.dia + sen.hora + cos.hora
Model 5: y.raiz ~ s(x, 3) + s(x2, 2) + s(x.p2, 4) + send.p1 + cosd1 +
    cosd3 + cosd.p4 + cos.dia + sen.hora + cos.hora
Model 6: y.raiz ~ s(x, 4) + s(x2, 2) + s(x.p2, 4) + send.p1 + cosd1 +
    cosd3 + cosd.p4 + cos.dia + sen.hora + cos.hora
Model 7: y.raiz ~ s(x, 3) + s(x2, 2) + s(x.p2, 5) + send.p1 + cosd1 +
    cosd3 + cosd.p4 + cos.dia + sen.hora + cos.hora
  Resid. Df Resid. Dev          Df Deviance      F   Pr(>F)
1      3651     1991.8
2      3650     1988.7  1.00004896   3.1106 5.7330 0.016697 *
3      3649     1983.8  0.99979817   4.9427 9.1122 0.002558 **
4      3648     1981.6  0.99984262   2.1573 3.9769 0.046209 *
5      3647     1978.8  1.00009945   2.8289 5.2136 0.022465 *
6      3646     1978.1  1.00009940   0.6792 1.2518 0.263293
7      3646     1977.3 -0.00014008   0.8086
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figura A.9. Estudio de modelos anidados para *ensayo4*, analizando distintos grados para las covariables de predicción de la velocidad del viento. Los modelos se han creado mediante la función *gam* en R, y comparado con la función *anova,* para hallar el más significativo (marcado en amarillo).

```
Model 1: y.raiz ~ s(x, 3) + s(x2, 2) + x4 + s(x.p2, 4) + sendv + send3 +
    send.p1 + cosd1 + cosd4 + cosd.p3 + cosd.p4 + cos.dia + sen.hora +
    cos.hora
Model 2: y.raiz ~ s(x, 3) + s(x2, 2) + s(x4, 2) + s(x.p2, 4) + sendv +
    send3 + send.p1 + cosd1 + cosd4 + cosd.p3 + cosd.p4 + cos.dia +
    sen.hora + cos.hora
  Resid. Df Resid. Dev Df Deviance      F Pr(>F)
1      3643       1964
2      3642       1964  1 0.078109 0.1448 0.7036
```

Figura A.10. Resultado del test ANOVA aplicado para estudiar el grado de *x4* en *ensayo4.1*, obteniendo como resultado que *x4* se adapta mejor al grado 1.

```
Analysis of Variance Table

Model 1: y.raiz ~ x + x2 + x.p2 + sendv + send.p1 + cosd1 + cosd3 + cosd.p4 +
    cos.dia + sen.hora + cos.hora
Model 2: y.raiz ~ poly(x, 3) + poly(x2, 2) + poly(x.p2, 4) + send.p1 +
    cosd1 + cosd3 + cosd.p4 + cos.dia + sen.hora + cos.hora
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1   3652 2029.5
2   3647 1976.3  5    53.156 19.619 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figura A.11 Test ANOVA realizado entre *ensayo4* y *ensayo4.poly*. Como resultado, se obtiene como más significativo al *ensayo4.poly.*

```
Analysis of Variance Table

Model 1: y.raiz ~ x + x2 + x4 + x.p2 + sendv + send3 + send.p1 + cosd1 +
    cosd4 + cosd.p3 + cosd.p4 + sen.dia + cos.dia + sen.hora +
    cos.hora
Model 2: y.raiz ~ poly(x, 3) + poly(x.p2, 4) + send3 + send.p1 + cosd1 +
    cosd4 + cosd.p4 + cos.dia + sen.hora + cos.hora + x2
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1   3648 2016.3
2   3647 1969.2  1    47.156 87.336 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figura A.12 Test ANOVA realizado entre *ensayo4.1* y *ensayo4.1.poly*. Como resultado, se obtiene como más significativo al *ensayo4.1.poly.*

```
  Res.Df    RSS Df Sum of Sq      F   Pr(>F)
1   3650 2020.6
2   3652 1935.7 -2    84.936
3   3651 1931.3  1     4.402 8.4314 0.003710 **
4   3649 1923.3  2     7.957 7.6208 0.000498 ***
5   3646 2005.2  3   -81.901
6   3648 1917.2 -2    88.092
7   3648 1914.6  0     2.568
8   3638 1899.3 10    15.306 2.9318 0.001142 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figura A.13 Resultado del test ANOVA entre los 8 modelos lineales a los que se ha introducido interacciones de forma manual. Los modelos 1 a 8 siguen el orden de aparición de la columna de modelos lineales en la Tabla 5.

```
  Res.Df    RSS  Df Sum of Sq        F    Pr(>F)
1   3637 1942.6
2   3636 1870.8   1    71.849 141.4823 < 2.2e-16 ***
3   3631 1851.9   5    18.826   7.4144 6.274e-07 ***
4   3629 1842.9   2     9.020   8.8812  0.000142 ***
5   3640 1938.9 -11   -95.976  17.1810 < 2.2e-16 ***
6   3641 1870.7  -1    68.153
7   3632 1846.3   9    24.469   5.3538 2.651e-07 ***
8   3640 1868.3  -8   -22.069   5.4322 7.932e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figura A.14. Resultado del test ANOVA entre los 8 modelos polinómicos a los que se ha introducido interacciones de forma manual. Los modelos aquí representados (del 1 al 8) siguen el orden de aparición de los modelos polinómicos en la Tabla 5.

```
  Res.Df    RSS  Df Sum of Sq      F    Pr(>F)
1   3639 1807.3
2   3628 1738.3  11    68.906 13.074 < 2.2e-16 ***
3   3647 1830.1 -19   -91.730 10.076 < 2.2e-16 ***
4   3641 1759.3   6    70.758 24.613 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figura A.15 Resultados del test ANOVA realizado para los modelos lineales desarrollados mediante *b1* y *b2*, estudiados en el orden en que aparecen en la Tabla 6. El modelo más significativo es el *ensayo4.1.b2*

```
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1   3622 1705.1
2   3616 1678.0  6    27.100 9.7331 1.165e-10 ***
3   3624 1710.1 -8   -32.105 8.6480 9.386e-12 ***
4   3629 1686.9 -5    23.207
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figura A.16 Resultados del test ANOVA realizado para los modelos polinómicos desarrollados mediante *b1* y *b2,* estudiados en el orden en que aparecen en la Tabla 6. El modelo más significativo, de entre los que han resultado comparables por estar anidados, es el *ensayo4.poly.b2.* No se ha podido comparar con *ensayo4.1.poly.b2*.

```
  Res.Df    RSS Df Sum of Sq        F    Pr(>F)
1   3653 2030.9
2   3652 1935.7  1    95.201 191.6914 < 2.2e-16 ***
3   3639 1807.3 13   128.450  19.8953 < 2.2e-16 ***
4   3647 1830.1 -8   -22.823   5.7445 2.703e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figura A.17. Resultado del test ANOVA para ensayos construidos a partir del *ensayo4* (4, 4.b, 4.b1 y 4.b2); resulta más significativo el *ensayo4.b2.*

```
  Res.Df    RSS  Df Sum of Sq        F    Pr(>F)
1   3648 2016.3
2   3648 1917.2   0    99.160
3   3628 1738.3  20   178.806 18.6587 < 2.2e-16 ***
4   3641 1759.3 -13   -20.971  3.3667 3.613e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figura A.18 Resultado del test ANOVA para ensayos construidos a partir del *ensayo4.1;* de entre los modelos anidados, resulta más significativo *ensayo4.1.b2.*

```
  Res.Df    RSS  Df Sum of Sq       F    Pr(>F)
1   3647 1830.1
2   3628 1738.3  19    91.730 10.0759 < 2.2e-16 ***
3   3641 1759.3 -13   -20.971  3.3667 3.613e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figura A.19 Resultado del test ANOVA entre *ensayo4.b2*, *ensayo4.1.b1* y *ensayo4.1.b2* (por orden de aparición).

```
> anova(ensayo4.1.poly.b1, ensayo4.1.poly.b2, test='F')
Analysis of Variance Table

Model 1: y.raiz ~ cosd1 + cosd4 + cosd.p4 + cos.dia + sen.hora + cos.hora +
    x.p2 + poly(x, 2) + x2:x + x3:x.p4 + cos.dia:sen.dia + cos.hora:sen.dia +
    poly(x, 3):send3 + send3:poly(x2, 4) + poly(x, 3):cos.dia +
    poly(x, 3):sen.hora + cos.hora:poly(x2, 4) + cosd4:cos.dia +
    send3:sen.hora + cosd4:sen.hora + cosd.p4:sen.hora + sen.hora:cos.hora +
    send.p1:cos.hora + cosd1:cos.hora + cosd.p4:x + sen.hora:poly(x2,
    2) + send.p1:x.p4:x.p1 + send.p1:x3:x.p3 + cosd.p4:x2:x +
    send3:x3:x.p4 + send.p1:x2:x.p3 + sen.hora:x2:x.p3 + cos.hora:x2:x.p3
Model 2: y.raiz ~ sen.hora + cos.hora + cosd4 + sen.dia + x2 + poly(x,
    2) + x3:x.p3 + x.p1:x.p2 + sen.dia:cos.dia + cos.hora:sen.dia +
    poly(x, 3):send3 + poly(x, 3):cos.dia + poly(x, 3):sen.hora +
    cosd4:cos.dia + sen.hora:send3 + sen.hora:send.p1 + sen.hora:cosd1 +
    sen.hora:cosd4 + sen.hora:cosd.p4 + sen.hora:cos.hora + cos.hora:send3 +
    cos.hora:send.p1 + cos.hora:poly(x, 2) + x.p1:send.p1:x.p4 +
    cos.hora:x3:x.p3 + x3:send3:x.p4
  Res.Df    RSS  Df Sum of Sq     F Pr(>F)
1   3616 1678.0
2   3629 1686.9 -13   -8.8979 1.475 0.1184
```

Figura A.20 Resultados del test ANOVA para los ensayos *4.1.poly.b1* y *4.1.poly.b2,* mostrando como más significativo al último.

```
  Res.Df    RSS Df Sum of Sq       F    Pr(>F)
1   3628 1738.3
2   3636 1643.4 -8     94.99
3   3641 1759.3 -5   -115.96  48.403 < 2.2e-16 ***
4   3640 1651.2  1    108.17 225.751 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figura A. 21 Resultados del test ANOVA para los modelos lineales candidatos (en orden de aparición en Tabla 9), para medir la significación de los nuevos armónicos introducidos.

```
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1   3616 1678.0
2   3623 1602.4 -7    75.575
3   3629 1686.9 -6   -84.473 30.339 < 2.2e-16 ***
4   3631 1620.7 -2    66.192
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figura A. 22 Resultados del test ANOVA para los modelos polinómicos candidatos (en orden de aparición en Tabla 9), para medir la significación de los nuevos armónicos introducidos.

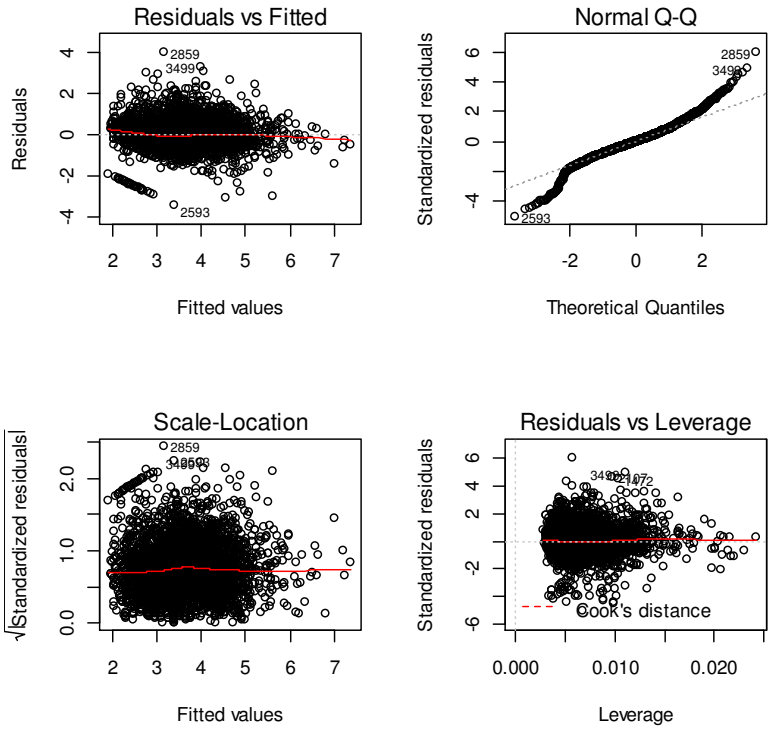


Figura A.23. Resultados de la función *plot* en R, para modelo *ms1*. Comprobación de hipótesis de residuos.

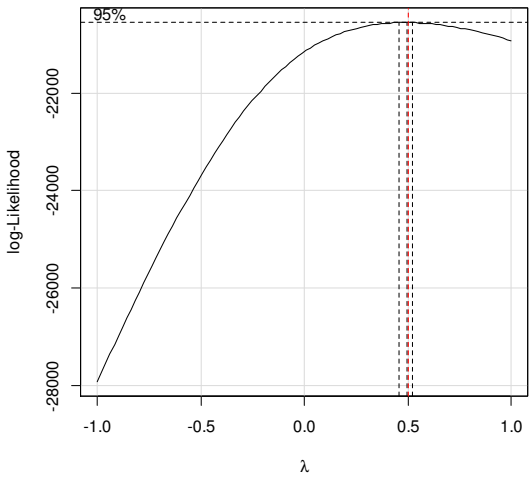


Figura A.24. Resultado del estudio de la transformación Box Cox para *ms1*. La transformación correcta de la respuesta es la transformación raíz cuadrada.

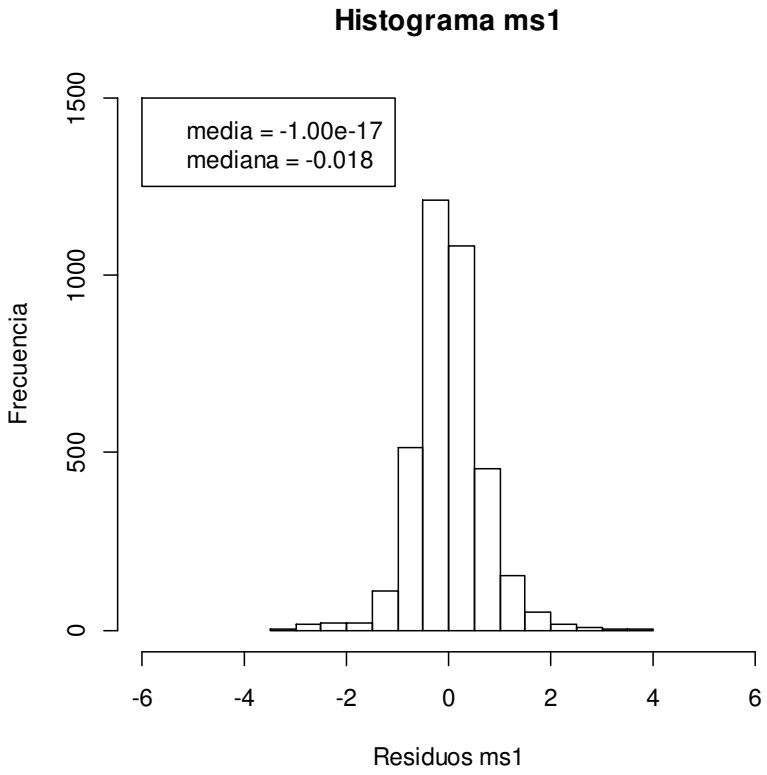


Figura A.25. Histograma de residuos para *ms1.*

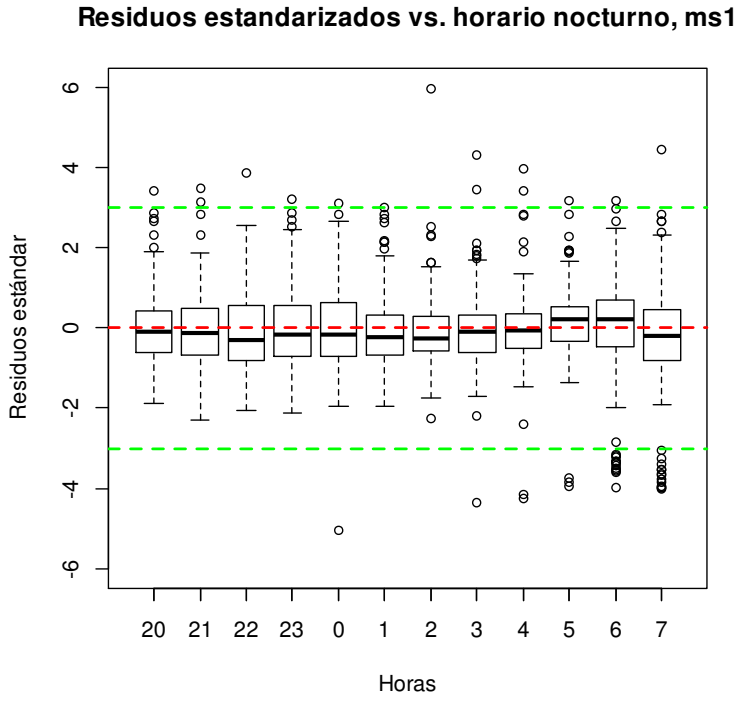


Figura A. 26. Diagramas de cajas para estudiar la variación de los residuos estándar del modelo *ms1* a lo largo del horario nocturno (entre las 20 y las 7 horas, para cada hora por separado).

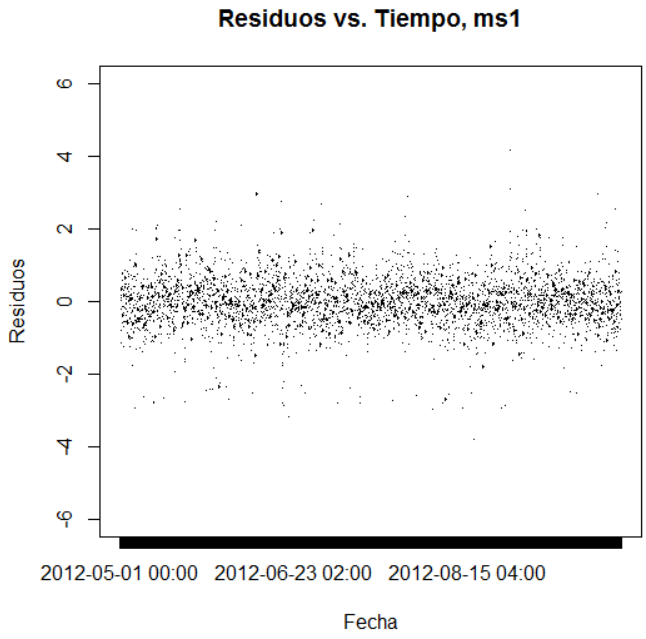


Figura A.27. Gráfico residuos vs tiempo para *ms1*. No se observa ningún patrón especial.

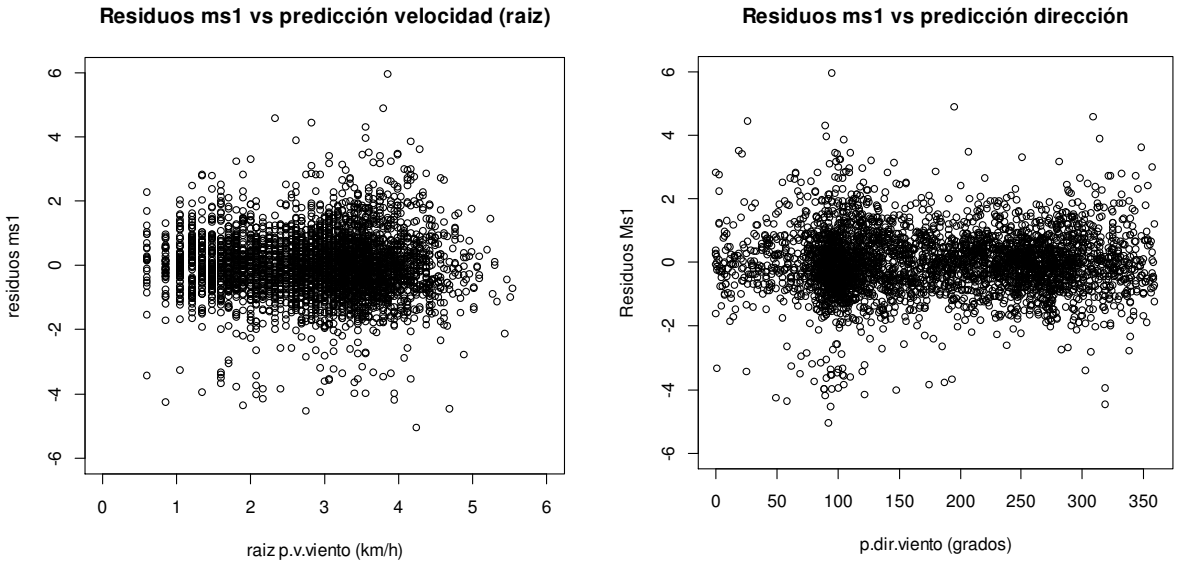


Figura A. 28. Gráfico de residuos del modelo *ms1* versus covariables de predicción de la velocidad del viento del modelo de AEMET (izda.) y de predicción de la dirección del viento del modelo AEMET (dcha.).

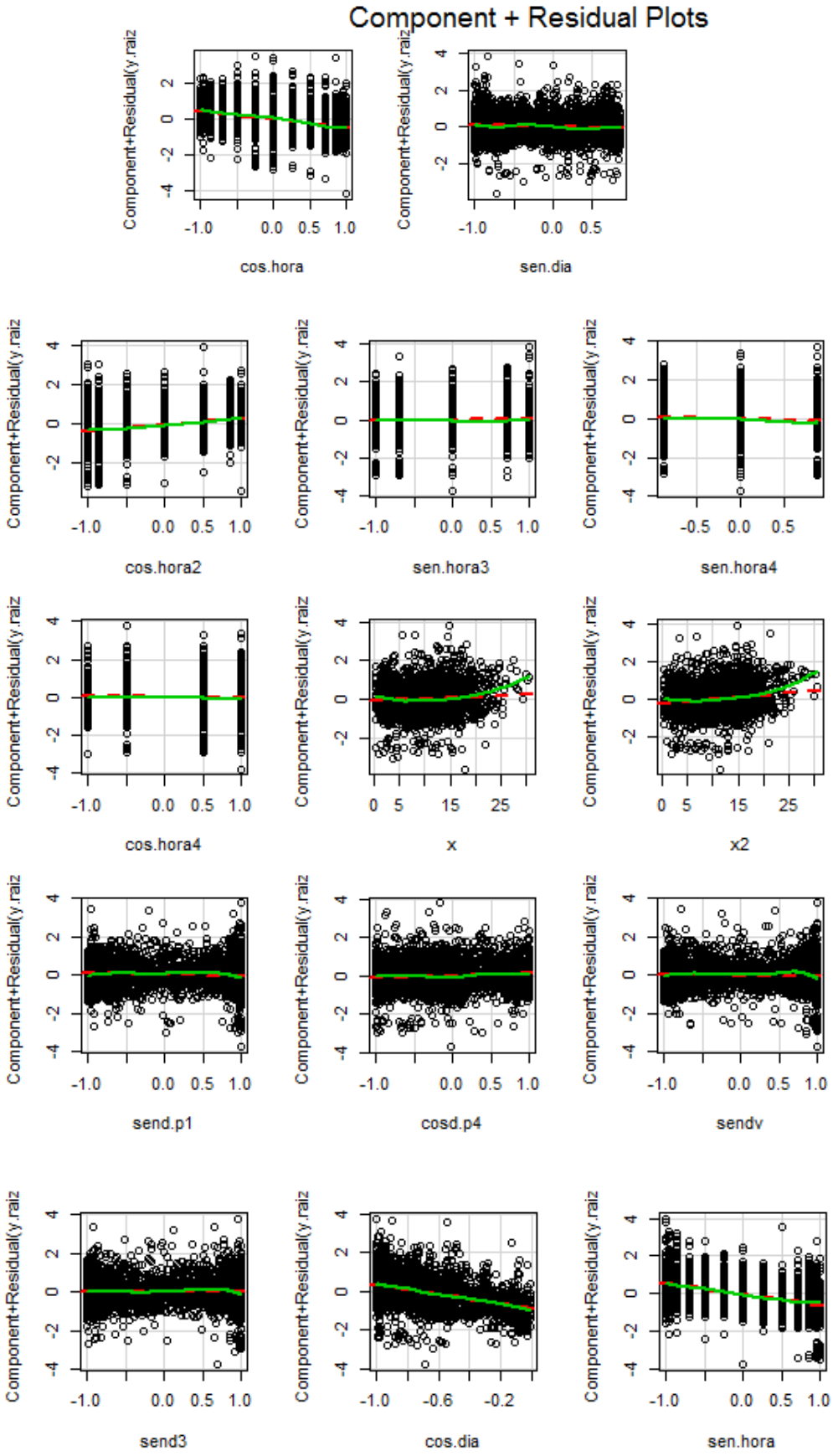


Figura A.29. Gráficos *crPlots* para el estudio de la adecuación de las covariables del modelo *ms1*. *Continúa.*

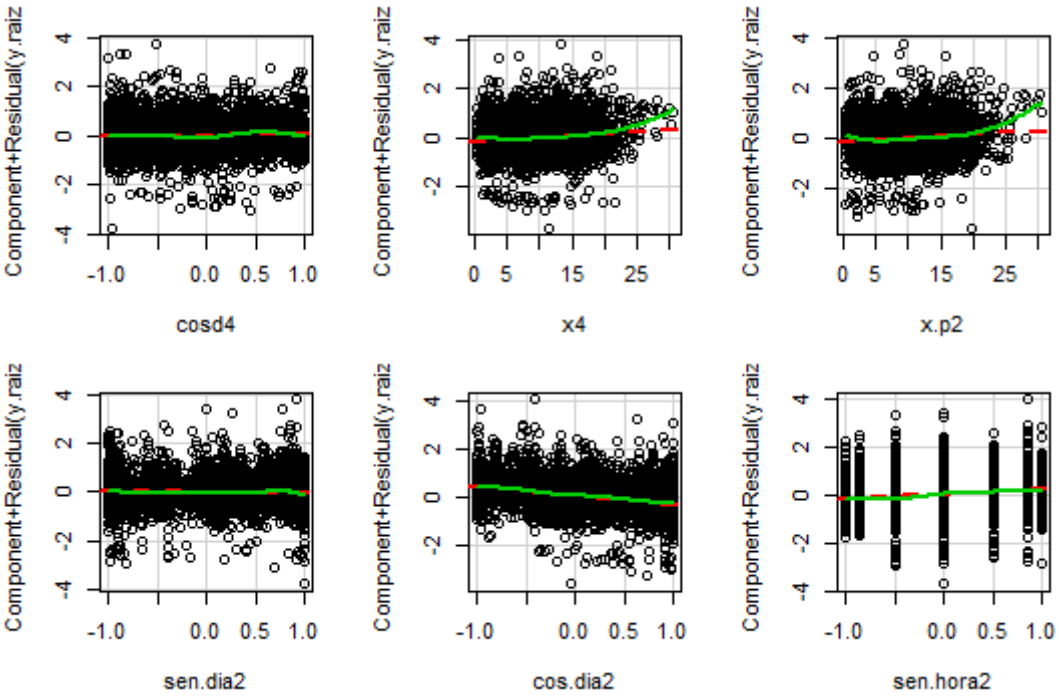


Figura A.29. Gráficos *crPlots* para el estudio de la adecuación de las covariables del modelo *ms1*.

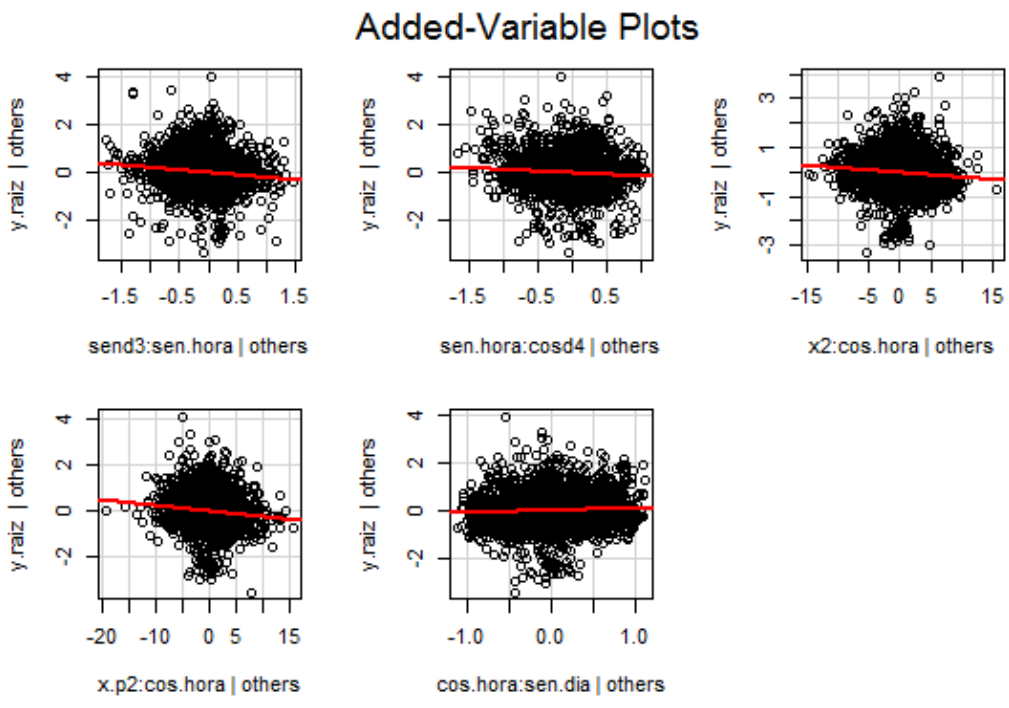


Figura A.30. Gráficos *avPlots* para el modelo *ms1. Continúa.*

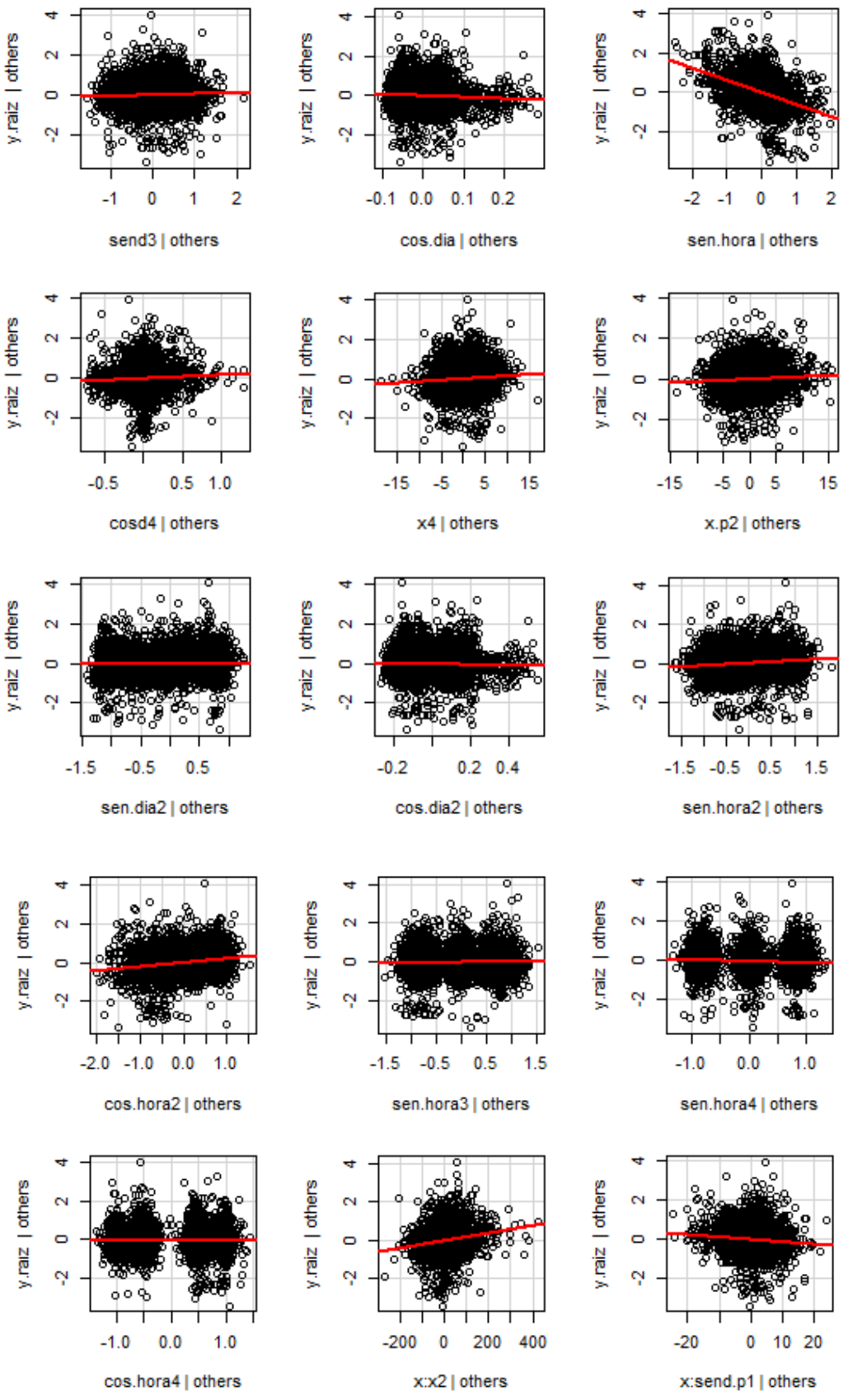


Figura A.30. Gráficos *avPlots* para el modelo *ms1. Continúa.*

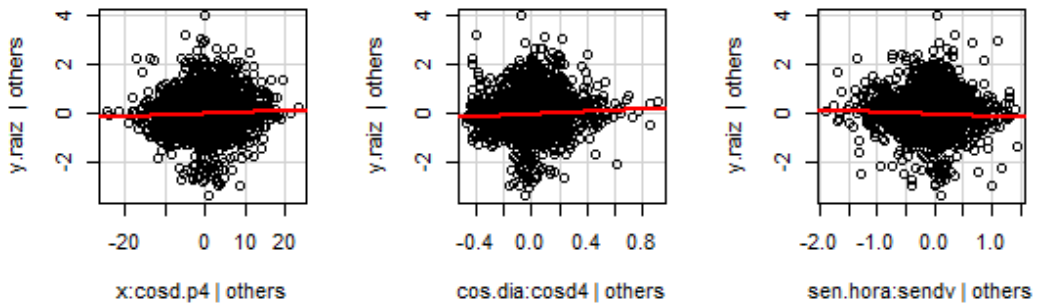


Figura A.30 Gráficos *avPlots* para el modelo *ms1.*

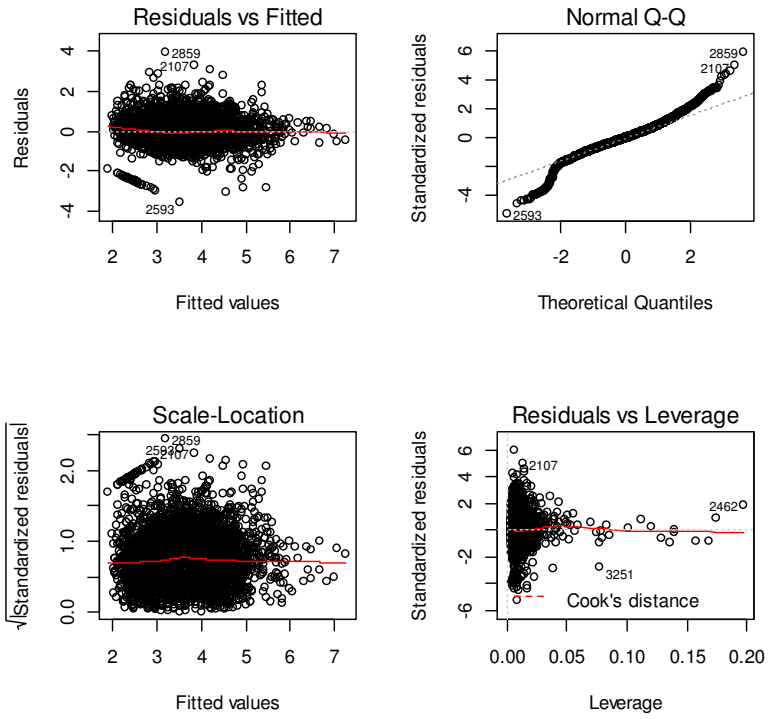


Figura A.31. Resultados de la función *plot* en R, para modelo *ms2*. Comprobación de hipótesis de residuos.

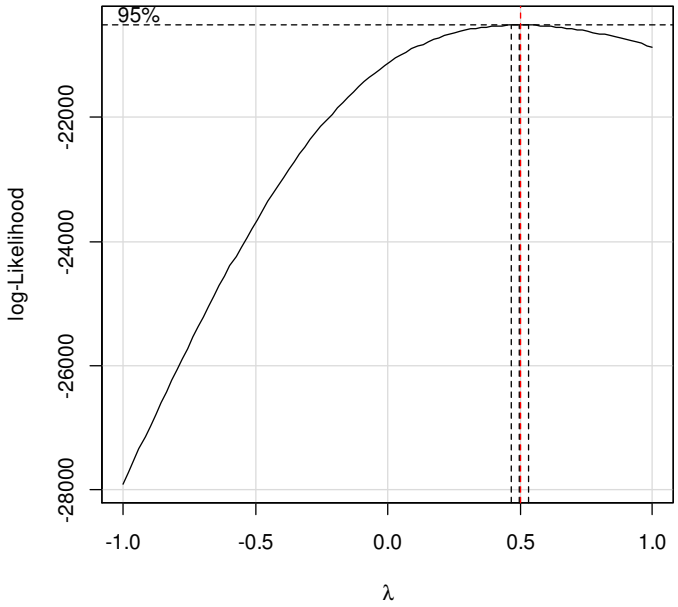


Figura A.32 Resultado del estudio de la transformación Box Cox para *ms2*.

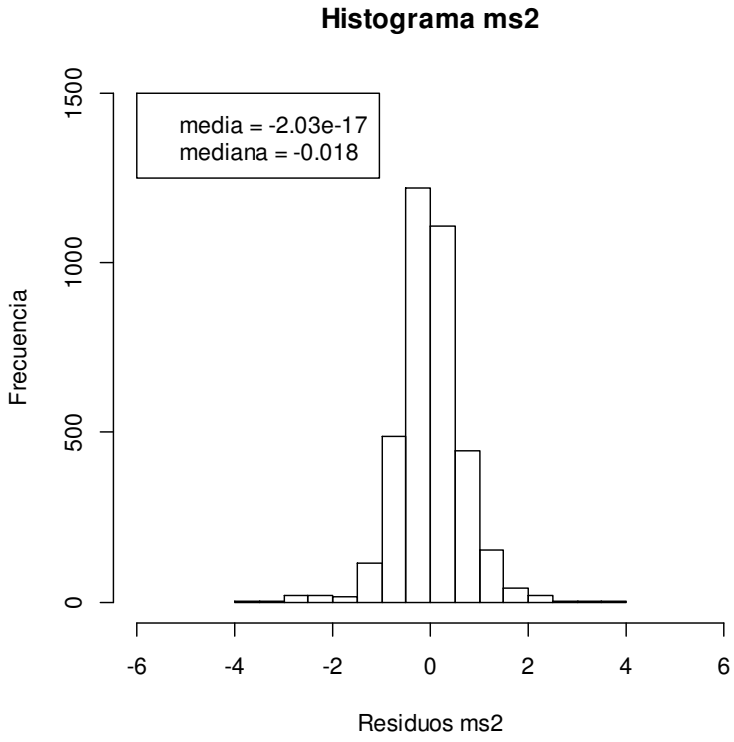


Figura A.33 Histograma de residuos para *ms2.*

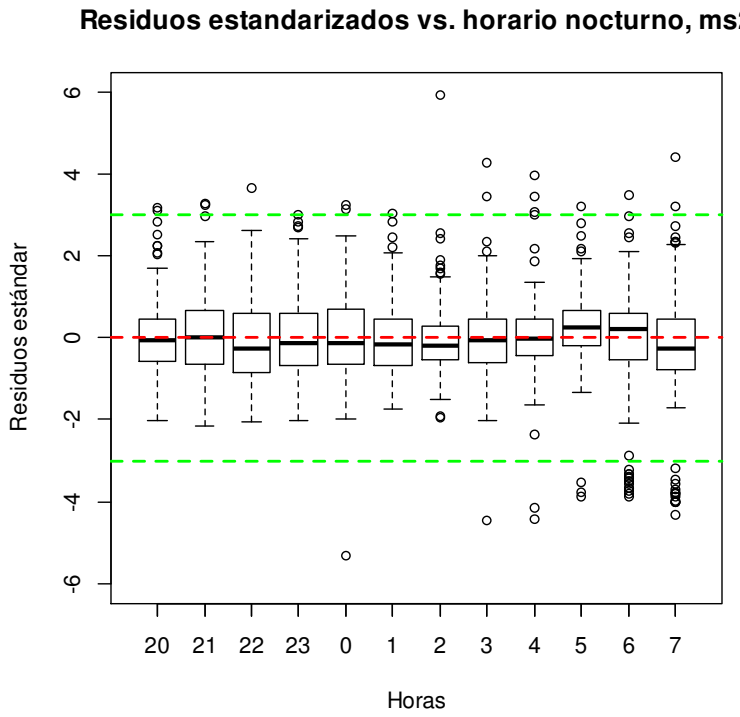


Figura A. 34. Diagramas de cajas para estudiar la variación de los residuos estándar del modelo *ms2* por horas, a lo largo del horario nocturno (entre las 20 y las 7 horas).

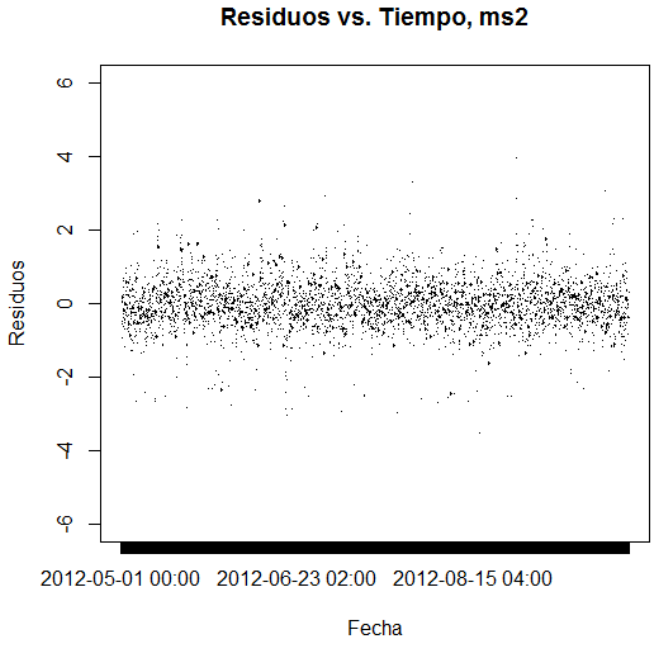


Figura A.35. Gráfico residuos vs tiempo para *ms2*. No se observa ningún patrón especial.

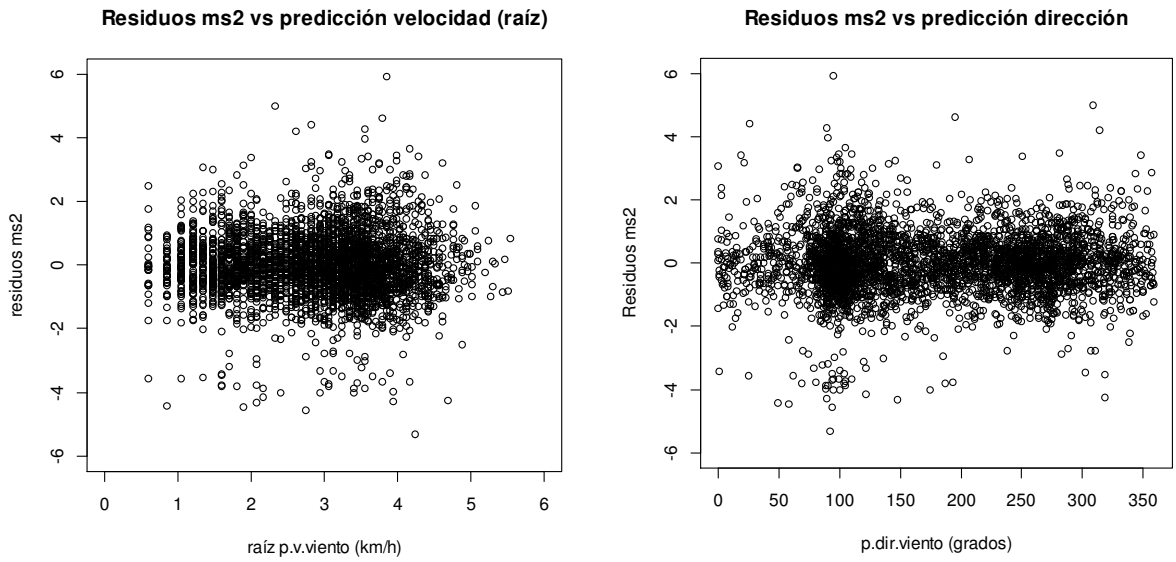


Figura A. 36. Gráfico de residuos del modelo *ms2* versus covariables de predicción de la velocidad del viento del modelo de AEMET (izda.) y de predicción de la dirección del viento del modelo AEMET (dcha.).

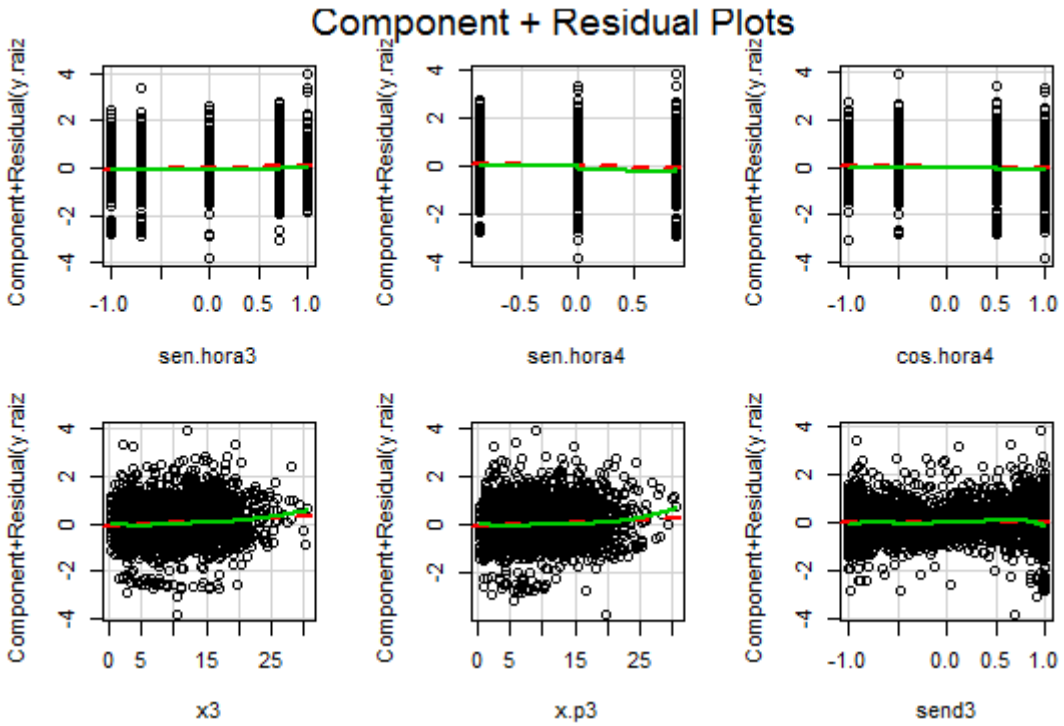


Figura A.37. Gráficos *crPlots* para el estudio de la adecuación de las covariables del modelo *ms2*. *Continúa.*

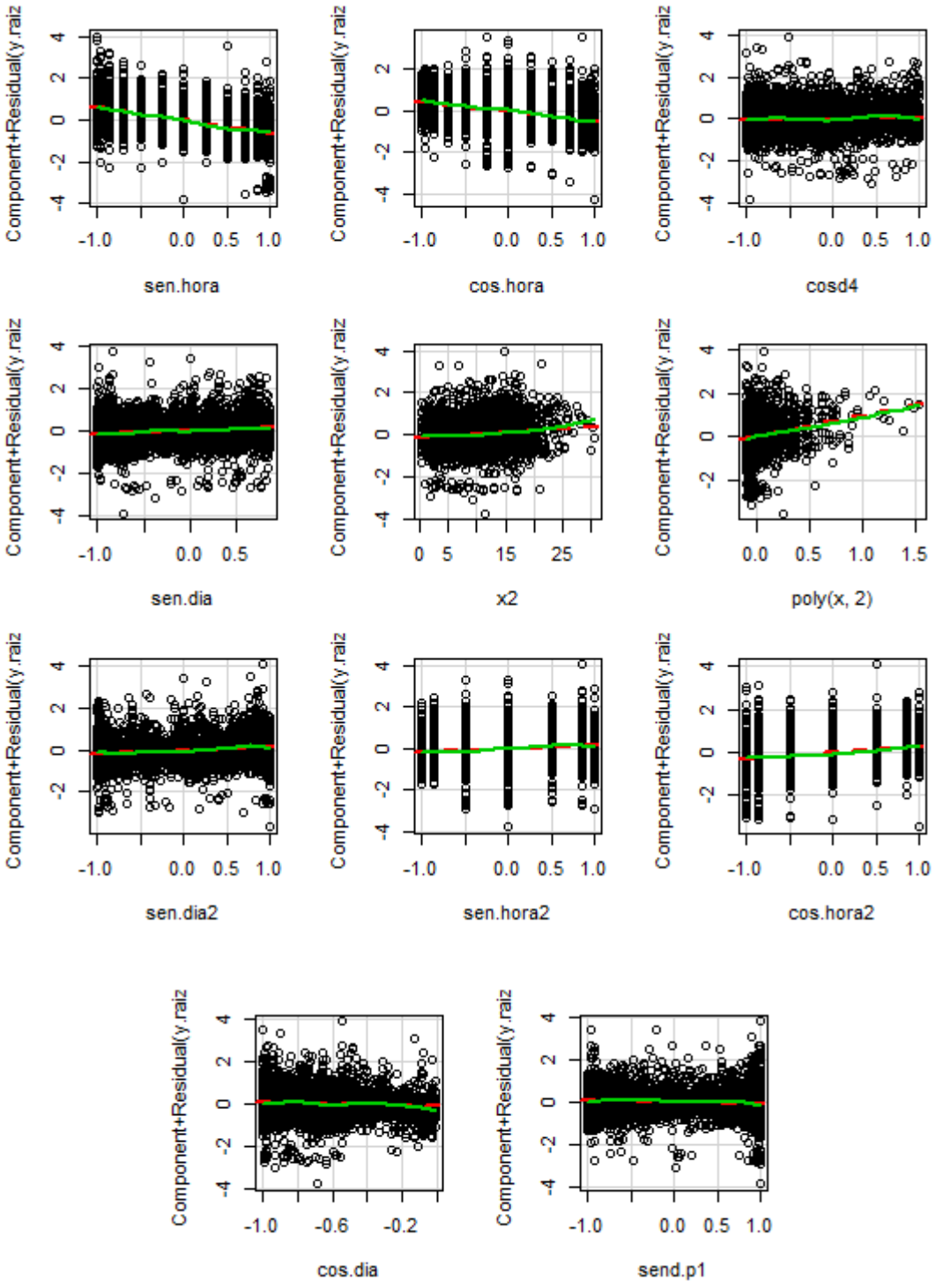


Figura A.37. Gráficos *crPlots* para el estudio de la adecuación de las covariables del modelo *ms2*.

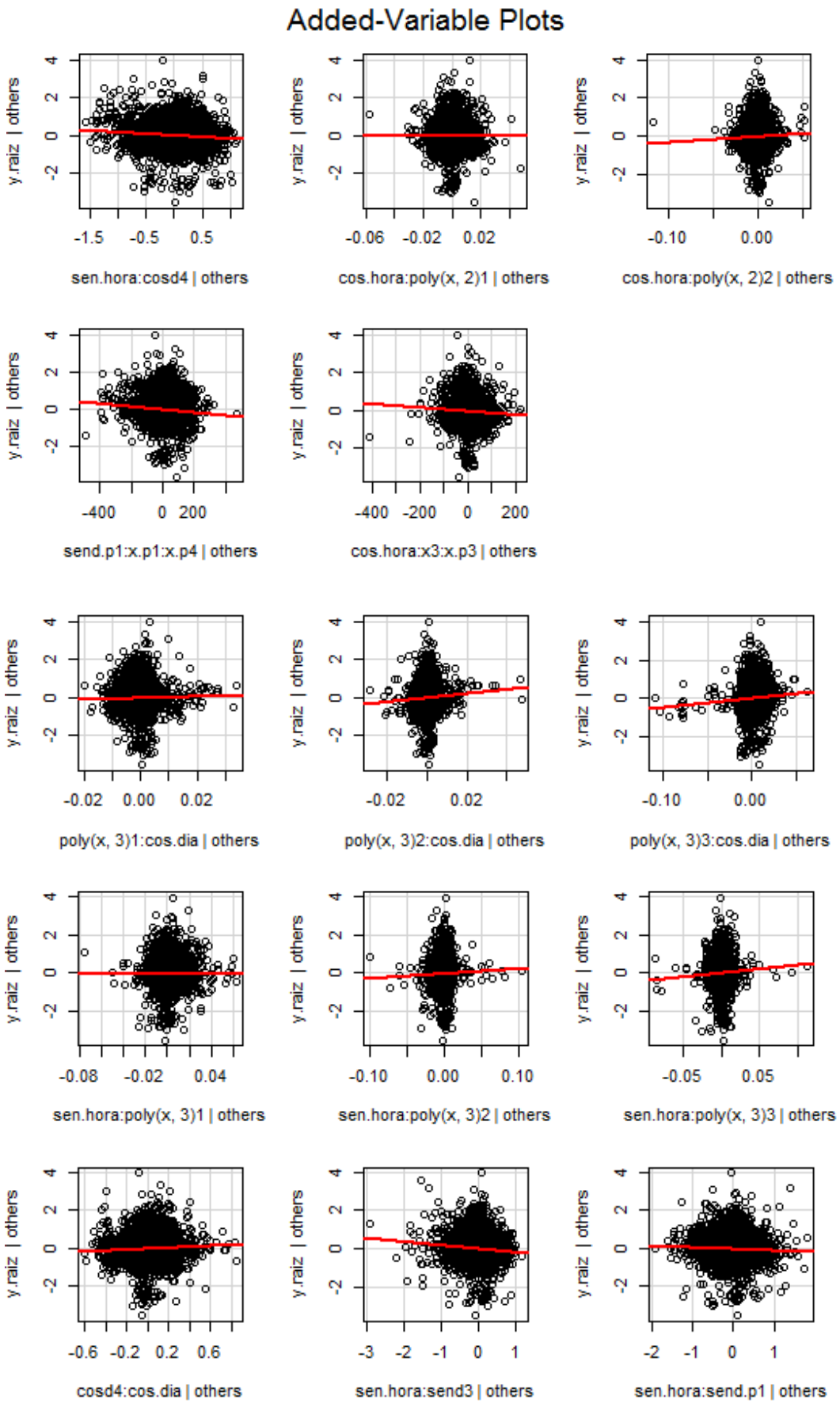


Figura A.38 Gráficos *avPlots* para el modelo *ms2*. *Continúa.*

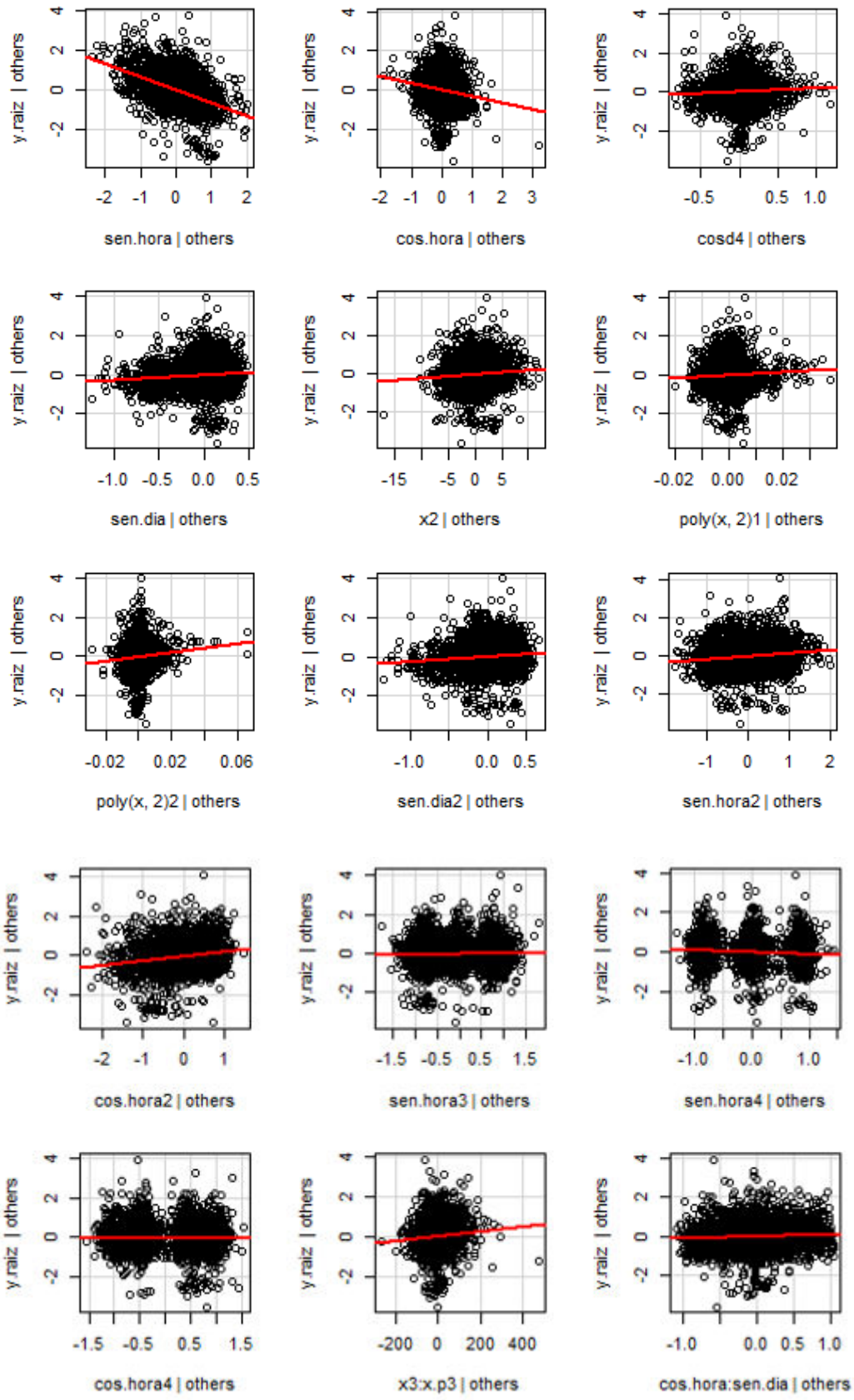


Figura A.38. Gráficos *avPlots* para el modelo *ms2*. *Continúa.*

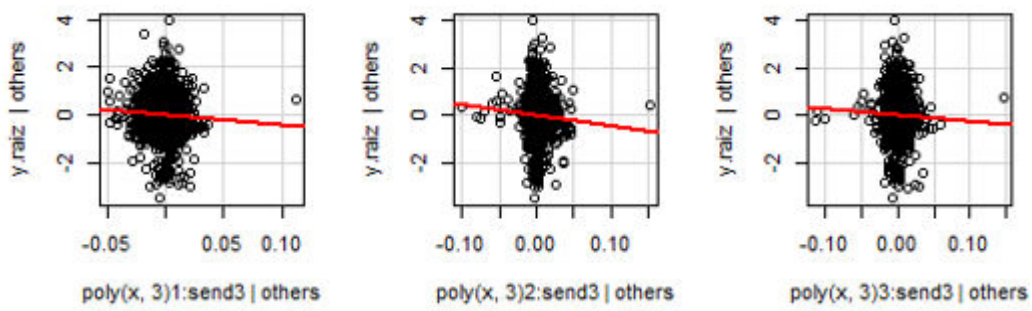


Figura A.38. Gráficos *avPlots* para el modelo *ms2*.

ESTUDIO GRÁFICO DEL PRONÓSTICO DE LOS MODELOS SELECCIONADOS *MS1* Y *MS2*

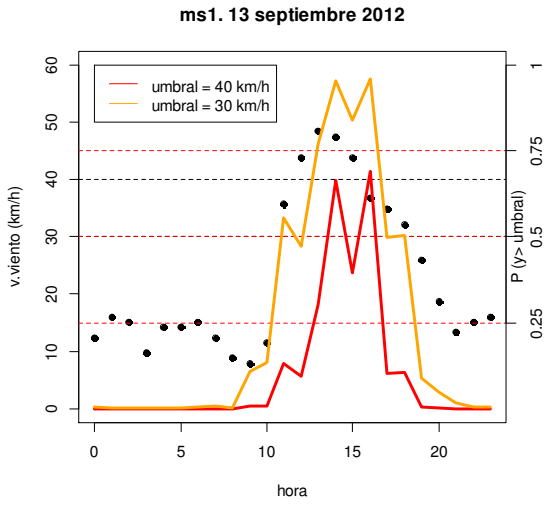
- **Estudio de "días parada"**

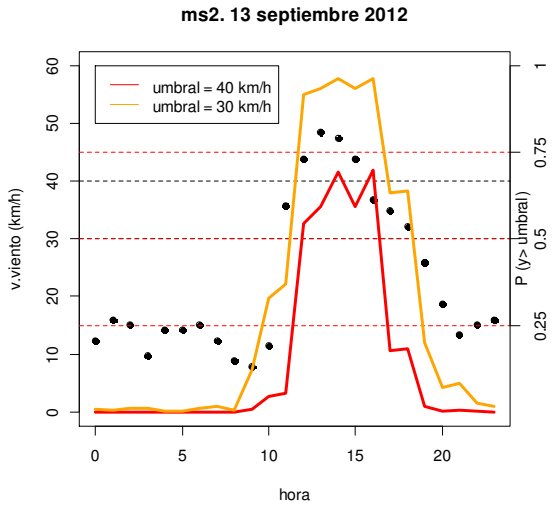
Mes	Día	Hora
mayo	17	18
	21	11
	21	12
	21	13
	21	14
	21	15
junio	11	16
	12	16
	18	19
julio	1	17
	26	22
	27	18
agosto	25	17
	28	2
	30	18
	31	17
septiembre	1	16
	1	19
	3	16
	12	14
	12	15
	12	16
	13	12
	13	13
	13	14
	13	15
	23	18

Tabla A.39. Tabla resumen de "días parada", considerando como tales, aquellos que presentan observaciones de velocidad de viento superiores a 40 km/h.

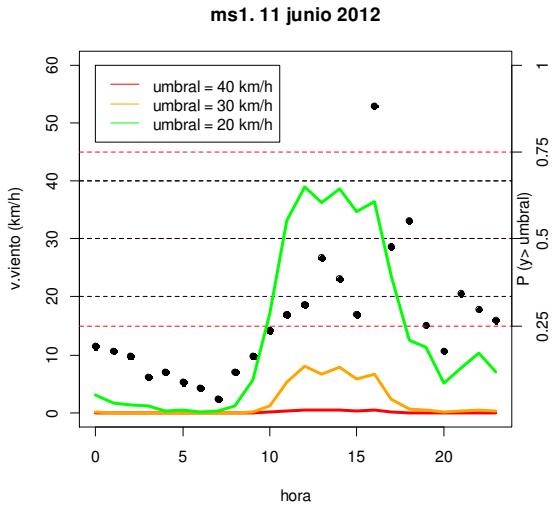
- **"Días parada" en los que el modelo pronostica correctamente**

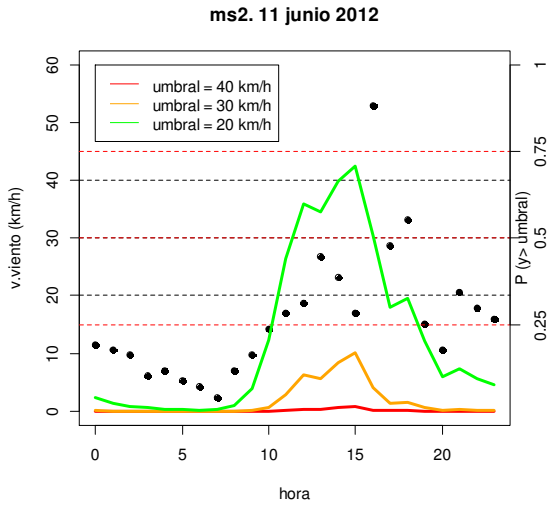
"Días parada" con observaciones problemáticas que el modelo pronostica correctamente



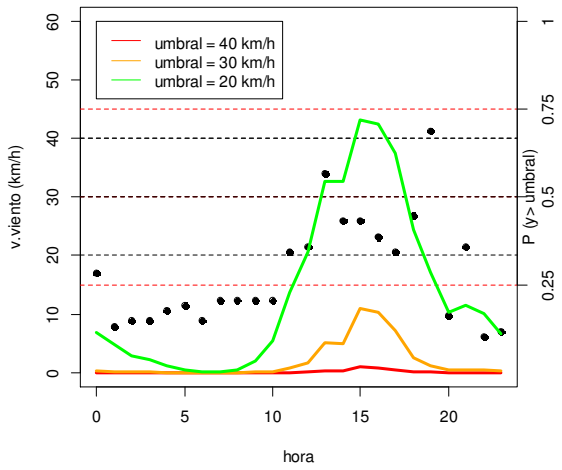


"Días parada" con observaciones aisladas

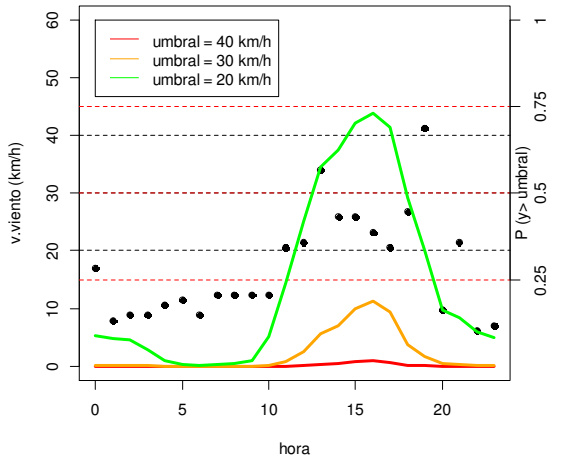




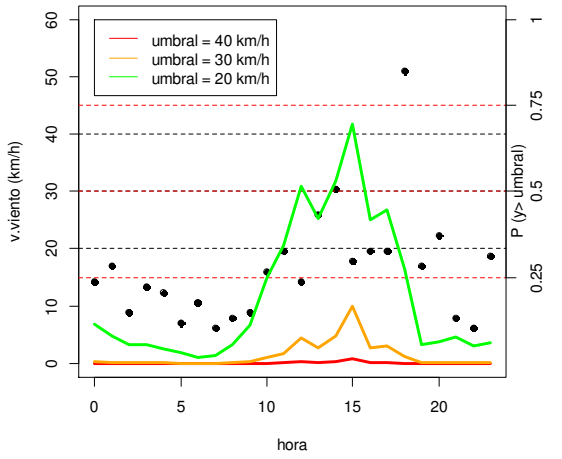
ms1. 18 junio 2012



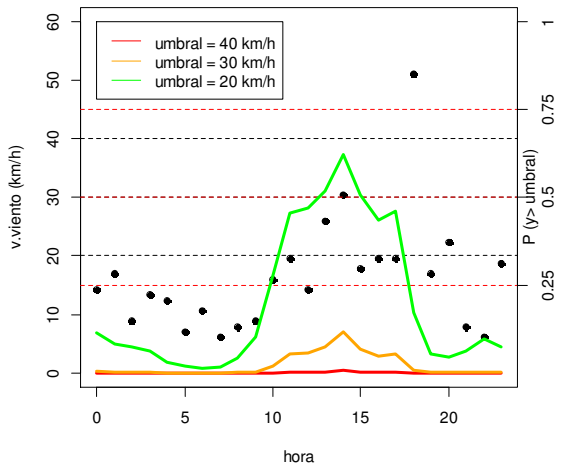
ms2. 18 junio 2012



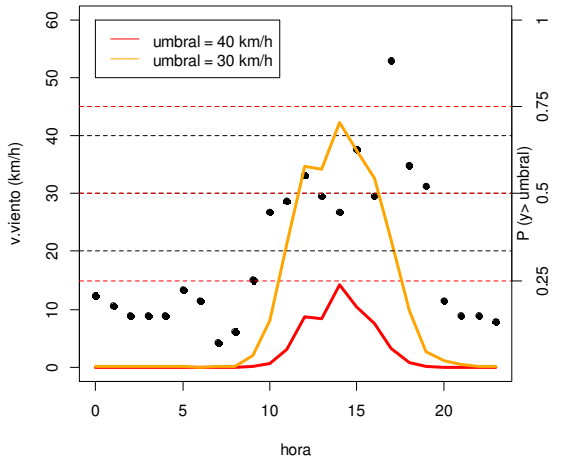
ms1. 27 julio 2012



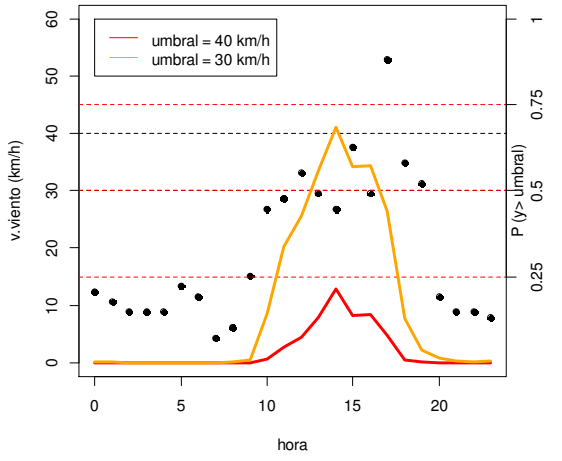
ms2. 27 julio 2012

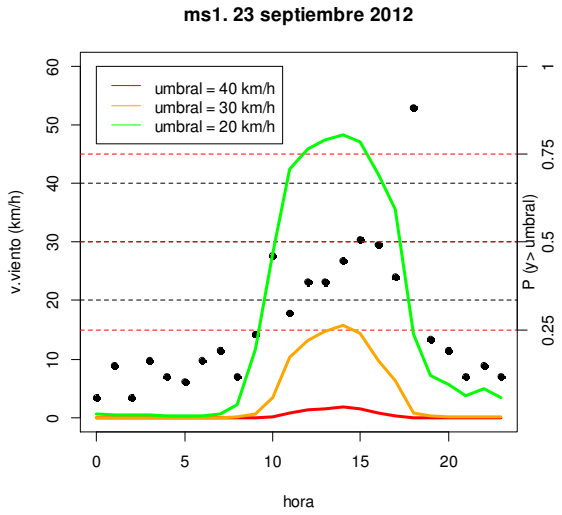


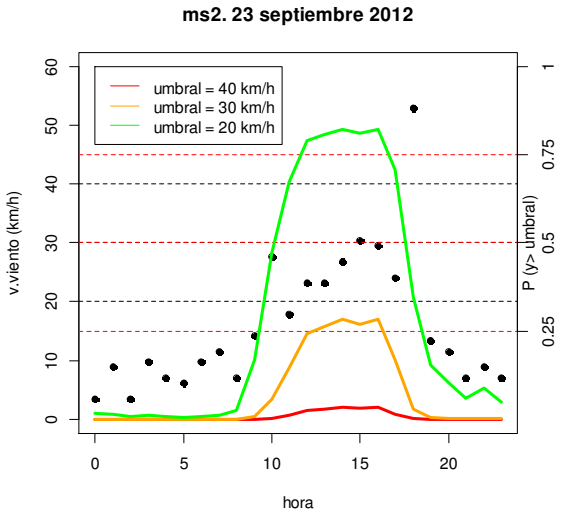
ms1. 31 agosto 2012



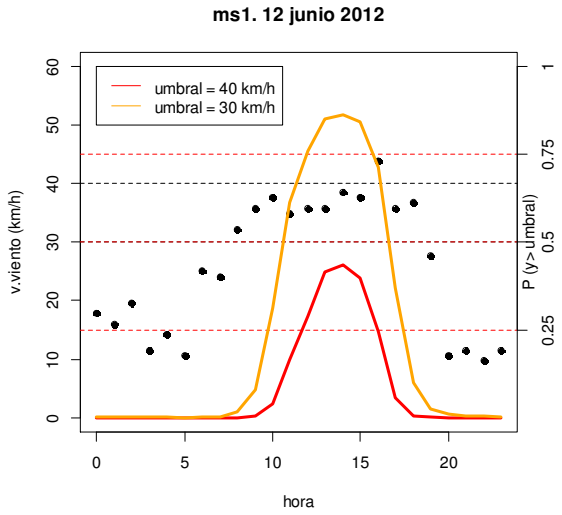
ms2. 31 agosto 2012

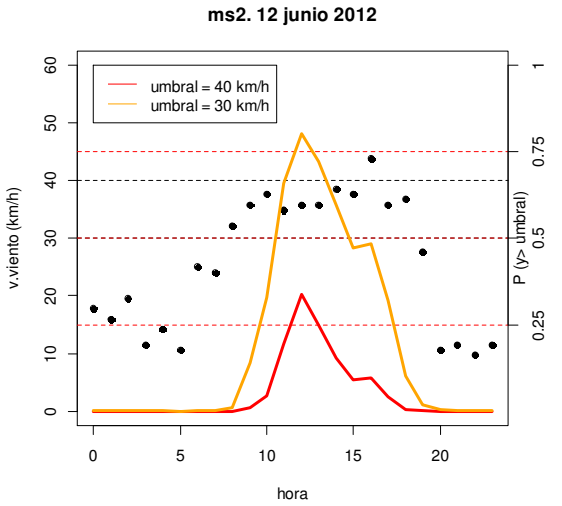


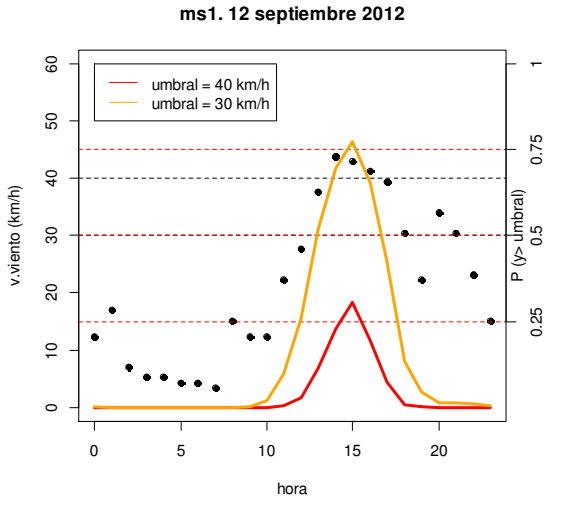


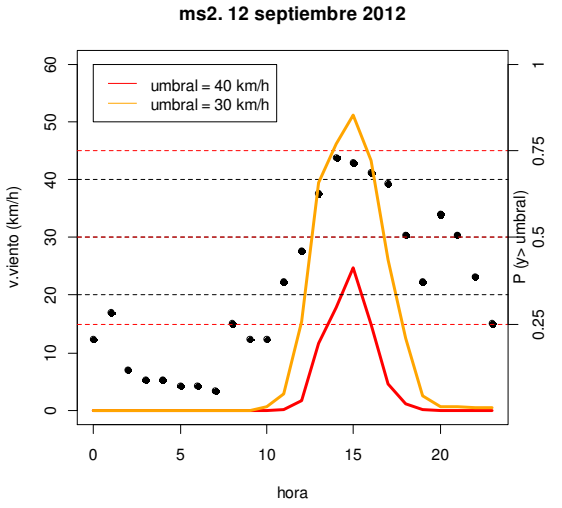


"Días parada" en que el modelo pronostica velocidades elevadas

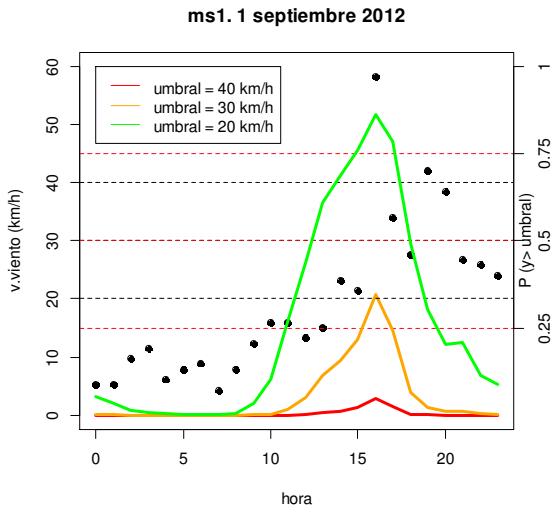


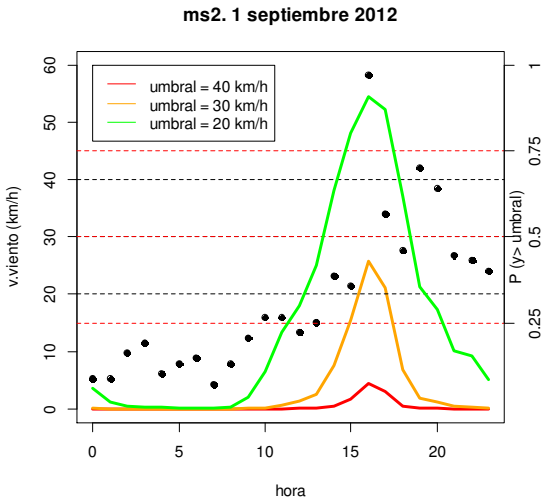


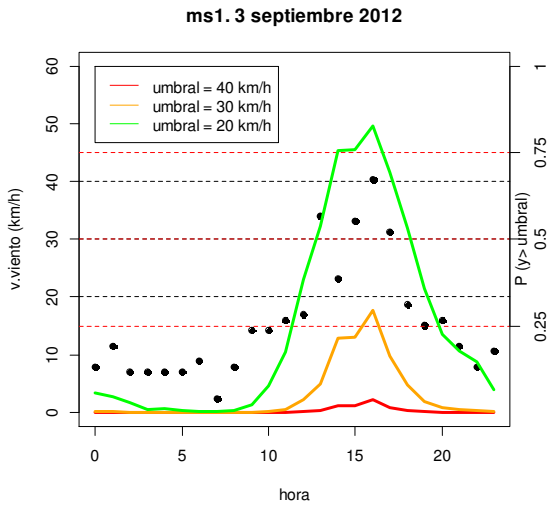


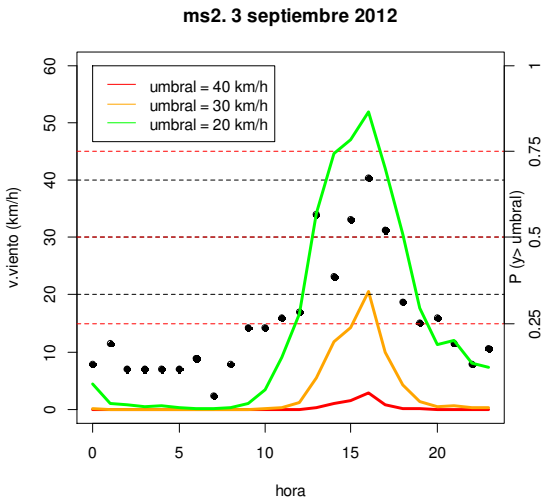


- **"Días parada" en los que el modelo *no* pronostica correctamente, sin causa aparente**









- **Día con diferencias más visibles entre ms1 y ms2**

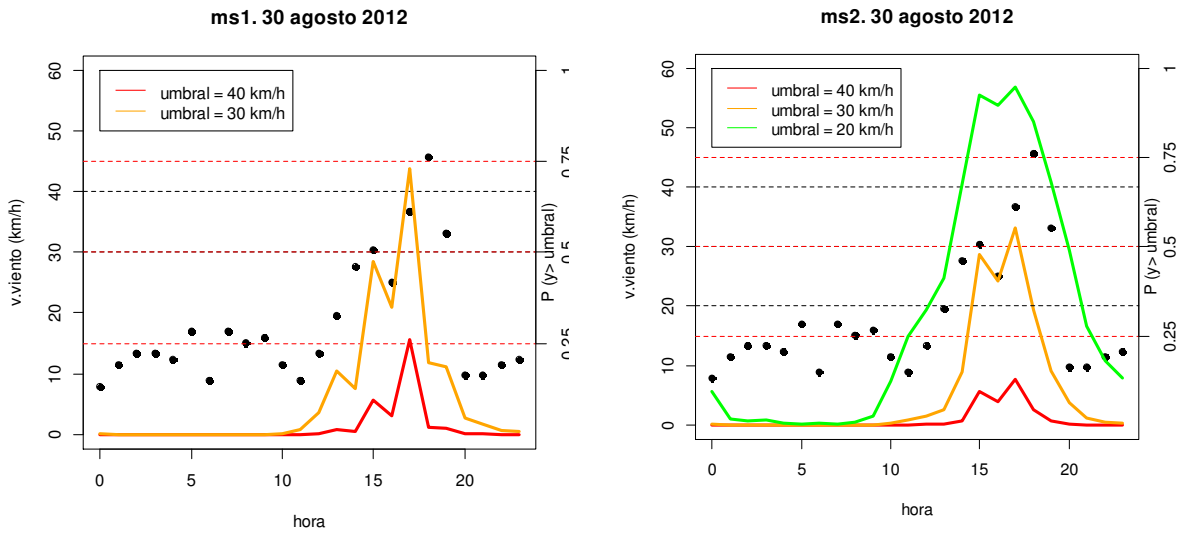


Figura A. 40 Estudio del pronóstico de *ms1* y *ms2* para el 30 de agosto de 2012, día con diferencias más significativas en el pronóstico de ambos modelos.

ANEXO II. CONCEPTOS ESTADÍSTICOS DE INTERÉS

LAS ROSAS DE LOS VIENTOS

La rosa de los vientos (ver ejemplo en la Figura A.41) es una herramienta que permite resumir de forma gráfica el comportamiento del viento, ya que representa la frecuencia con que se ha registrado una determinada velocidad en cada partición de la dirección del viento (en este caso definida por 12 rumbos), mediante códigos de amplitud y de color. El código de color muestra los intervalos de velocidad considerados, mientras la amplitud de la franja coloreada muestra la frecuencia a la que dichos intervalos se han dado, para cada dirección.

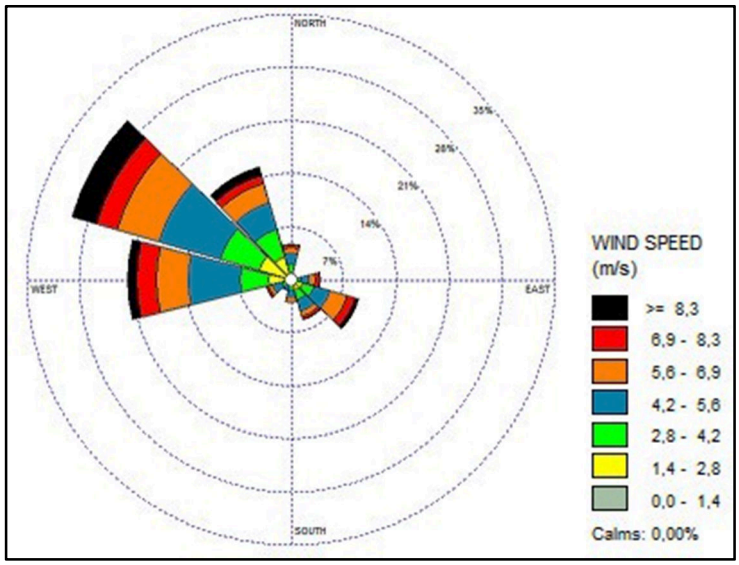


Figura A.41 Ejemplo de rosa de vientos

La realización de rosas de vientos permitirá comparar gráficamente el comportamiento del viento para las BD real y predicha por el modelo de AEMET.

CONCEPTOS SOBRE CORRELACIÓN

El coeficiente de correlación de Pearson indica cuál es la relación lineal entre un par de variables para el instante de tiempo t_0, es decir, indica la correlación instantánea entre ellas. Sin embargo, en ocasiones interesa estudiar también cuál es la correlación para cada instante de tiempo del periodo en estudio, es decir, la correlación cruzada. Este análisis se realiza cuando se quiere calcular cuál sería la correlación para instantes decalados.

La autocorrelación permite estudiar la relación existente entre los valores de una misma variable, observados con una diferencia de *k* unidades de tiempo, denominada "retardo". El estudio de la autocorrelación de las variables aporta información sobre su comportamiento a lo largo del tiempo; es decir, permite estudiar por ejemplo, si los datos anteriores o posteriores de una variable en el tiempo, podrían ser útiles a la hora de explicar el comportamiento de dicha variable.

Así, la existencia de autocorrelación puede expresar la existencia de estacionalidad, lo cual es de esperar si, como en nuestro caso, se trabaja con series de tiempo horarias (Pepió, 2011).

El gráfico que permite estudiar la correlación, tanto entre un par de variables como para una misma variable, se denomina correlograma. El estudio del correlograma resulta útil cuando, por ejemplo, se cree que puede existir una correlación retardada (es decir, una correlación mayor con otro instante de tiempo), una autocorrelación, o para estudiar la amplitud de los ciclos existentes.

El correlograma se compone de:

- En el *eje x*, el periodo considerado para nuestros datos. En nuestro caso, este periodo será de 24 horas.
- En el *eje y*, bien la correlación cruzada en caso de estudiar dos series de datos, bien la autocorrelación en caso de estudiar una única variable.

INTRODUCCIÓN A LOS MODELOS DE REGRESIÓN LINEAL

Según el objetivo general de este TFG, se ha planteado el desarrollo de una serie de modelos de regresión (MR), lineales y no lineales, que permiten explicar y predecir el comportamiento de una variable cuantitativa (Y), llamada respuesta, a partir de una o un conjunto de variables regresoras o predictoras (X), suponiendo que la relación entre ellas es lineal o linealizable.

Además, según James et al. (2013), los MR permiten contestar a cuestiones que se plantean en el trabajo, tales como:

- ¿Existe una relación entre predictores y respuesta? ¿Cuál es su forma? (es lineal, no lineal).
- A la hora de predecir la respuesta, ¿resulta útil todo el conjunto de los predictores, o es solamente un subconjunto de ellos el que ayuda a explicarla?
- ¿Se ajusta el modelo de regresión a nuestro conjunto de datos?, es decir, ¿se ajustan a la realidad las predicciones de nuestro modelo? ¿cómo de precisa es dicha predicción?

Los modelos de regresión lineal (MRL) pueden ser simples (MRLS), o múltiples (MRLM), dependiendo del número de variables regresoras en estudio. Así, si únicamente se dispone de una variable regresora, el modelo será simple y seguirá la expresión recogida en Ec.2, llamada ecuación de regresión, donde Y es la variable respuesta, X la variable predictora, ϵ representa el residuo o error de la regresión, β_0 es el punto de corte de la recta con el eje de abscisas y β_1 representa la pendiente de la recta. β_0 y β_1 se denominan coeficientes de regresión.

$$Y = \beta_0 + \beta_1 X + \epsilon \qquad \text{(Ec. 2)}$$

Sin embargo, normalmente no se dispone de una única variable regresora en estudio, sino de un conjunto de ellas. En ese caso, el modelo en estudio se denomina de regresión lineal múltiple y es una generalización del modelo anterior:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon \qquad \text{(Ec. 3)}$$

El fin del MRLM sigue siendo el mismo: usar el conjunto de variables regresoras X para prededir la respuesta Y. En este caso, la línea de regresión se reemplaza por un hiperplano.

El fin último de cualquier MRL es seleccionar los coeficientes de regresión que minimicen la suma de cuadrados residual (*RSS, residual sum of squares*), es decir,

$$RSS = e_1^2 + e_2^2 + \cdots + e_n^2 \qquad \text{(Ec. 4)}$$

Siendo e_i (error o residuo) la diferencia entre el valor real observado (y_i) y el valor correspondiente predicho por el modelo ($\hat{y}_i$), para cada instante i,

$$e_i = y_i - \hat{y}_i \qquad \text{(Ec. 5)}$$

Esta diferencia es la distancia existente entre el valor observado y, y la recta de regresión estimada (ver ejemplo en Figura A. 42).

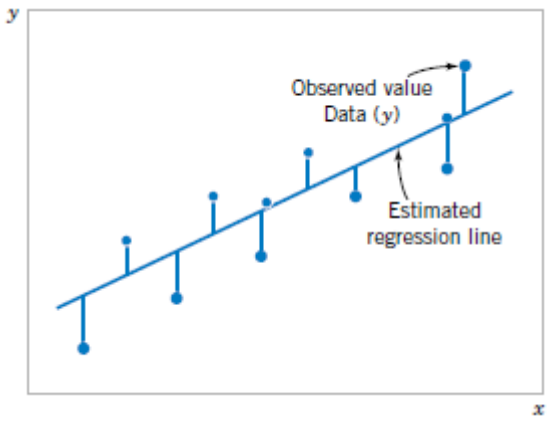


Figura A. 42 *Fuente*: *Montgomery (2012)*. Ejemplo de las desviaciones o residuos de los datos observados (y) del modelo de regresión lineal (simple) estimado. Las líneas de unión entre los datos observados y la recta de regresión, son los residuos estimados para cado dato y.

TRANSFORMACIÓN BOX COX

La transformación Box Cox es un procedimiento estadístico, desarrollado en 1964 por los estadísticos Box y Cox, de estudio de una posible transformación de la respuesta en un MR. Este procedimiento resulta adecuado cuando no hay una razón a priori para elegir una transformación determinada, ya que incluye todas las posibles transformaciones candidatas a normalizar la distribución de los residuos del MR. Esta transformación depende de un parámetro denominado λ, cuyo valor indica el tipo de transformación a realizar (ver Ec.6). Así,

- si λ = 1, no será necesaria ninguna transformación,
- si λ = 0, la transformación más adecuada será logarítmica,
- si λ = ½ la transformación deberá ser de tipo raíz cuadrada, y
- si λ = 1/3 la transformación será de tipo raíz cúbica.

Los distintos valores de λ se aplican a la ecuación de transformación Box Cox, seleccionando finalmente aquel valor con el cual el conjunto de datos transformados ($y^{(\lambda)}$) tenga una menor varianza, que es la situación más adecuada a la hora de conseguir una distribución normal y una varianza constante.

$$y^{(\lambda)} = \begin{cases} (y^\lambda - 1)/\lambda & \lambda \neq 0 \\ log\ y & \lambda = 0 \end{cases} \quad \text{(Ec.6. Venables y Ripley, 2002)}$$

TEST ANOVA

El análisis de la varianza, más conocido como test ANOVA, es un test de hipótesis que compara dos (o más) modelos. Según James et al. (2013), las hipótesis que contrasta el test ANOVA son las siguientes:

> H_0: todos los modelos se ajustan igualmente a los datos (y por tanto, se prefiere el modelo más sencillo).
>
> H_a: el modelo más complejo se ajusta mejor al conjunto de datos.

El resultado del test ANOVA se interpreta mediante el p-valor. Así, cuanto menor sea el p-valor, mayor será la probabilidad de que el modelo más complejo se ajuste mejor al conjunto de datos. Por tanto, la hipótesis nula (de aceptación del modelo más sencillo) se rechazará si el p-valor de alguno de los modelos más complejos en estudio es inferior a 0.05.

CRITERIOS DE BONDAD DE AJUSTE. EL R^2 AJUSTADO

Las herramientas que permiten estudiar el ajuste de los modelos, entendiendo como tal a la capacidad del modelo para explicar el comportamiento de la variable respuesta, se denominan criterios de bondad de ajuste.

De entre los criterios de bondad de ajuste existentes, en este trabajo se ha estudiado el R^2 estadístico, que es una cantidad comprendida entre 0 y 1, la cual mide la proporción de variabilidad de Y que puede ser explicada por X. Así, interesará que el MR tenga un R^2 estadístico lo más cercano posible a 1, pues esto indica que es capaz de reproducir el comportamiento de la variable respuesta y por tanto, también será capaz de predecirla.

Según Montgomery (2012), el R^2 estadístico se define como:

$$R^2 = \frac{SS_R}{SS_T} = 1 - \frac{SS_E}{SS_T} \qquad \text{(Ec. 7)}$$

Donde,

- $SS_R = \sum_{i=1}^{n}(\epsilon_i)^2$,
- $SS_E = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$,
- SS_T es la suma de las anteriores, $SS_T = SS_R + SS_E$

El inconveniente del R^2 estadístico es que siempre se incrementa al añadir más variables al modelo, incluso cuando dichas variables no están relacionadas con la respuesta. Por tanto, solo permite comparar modelos con el mismo número de covariables.

Por ello, se ha decidido usar el R^2 ajustado, ya que éste sí tiene en cuenta el número de variables que contiene el modelo, penalizando su valor conforme se añaden términos al modelo. Esto permite comparar modelos con diferente número de variables, y ayuda a evitar el sobreajuste del modelo.

$$R^2_{adj} = 1 - \frac{SS_E/(n-p)}{SS_T/(n-1)} \qquad \text{(Ec. 8. Montgomery, 2012)}$$

MÉTODOS *STEPWISE* DE SELECCIÓN DE PARÁMETROS

A lo largo del trabajo, se ha establecido una metodología de desarrollo de MR candidatos a ser el óptimo buscado. Para encontrarlo, se deberá seleccionar a los "mejores" de entre los modelos desarrollados en cada paso. La idea de esta selección consiste en realizar una amplia variedad de modelos diferentes, que incluyan distintos subconjuntos de términos, y seleccionar el mejor de ellos, de acuerdo con ciertos criterios. Por ejemplo, interesa que el modelo final tenga suficientes parámetros regresores como para capturar el comportamiento de la variable respuesta, pero también que sea fácil de usar e interpretar y que por tanto, tenga el menor número posible de ellos. Según Montgomery (2012), el compromiso entre dichos objetivos conflictivos se conoce como la búsqueda de la "mejor ecuación de regresión".

A la hora de seleccionar subconjuntos (de covariables e interacciones) en nuestros MR a desarrollar, y debido a la gran cantidad de covariables disponibles, se ha seleccionado el uso de métodos *stepwise* ("paso a paso"), dado que se consideran la mejor alternativa a dicho problema. Estos métodos, al contrario que los métodos exhaustivos, no estudian todas las posibles combinaciones (se estudian aproximadamente p^2 combinaciones, frente a las frente a las 2^p combinaciones del método anterior). Aun así, permiten obtener modelos muy próximos al óptimo posible a la vez que son mucho más eficientes en los cálculos (James et al., 2012).

Existen tres aproximaciones diferentes de selección automática de variables, con distinto funcionamiento. En nuestro caso, se ha seleccionado la aproximación mixta (*mixed selection*), ya que integra el funcionamiento de las otras dos y por tanto, creemos ofrecerá una selección de subconjuntos más significativa.

La selección mixta se realiza en una serie de pasos, comenzando con un modelo "nulo" (sin covariables, únicamente con el punto de corte), al cual en cada paso se añade aquel término que produce el mejor ajuste en el modelo (es decir, el que menor suma de cuadrados tenga, y consecuentemente, produzca un modelo con mayor R^2). Sin embargo, la significación de los parámetros de un modelo depende del resto de parámetros presentes en el mismo y por tanto, al eliminar o añadir un parámetro al modelo, la significación del resto varía en mayor o menor medida, pudiendo dejar de ser significativos algunos de los términos; si esto ocurre, la selección mixta los elimina del modelo. El proceso funciona (añadiendo y quitando covariables) hasta que todos los parámetros incluidos en el modelo tengan un p-valor lo suficientemente reducido, y todos los parámetros fuera del modelo tuviesen un p-valor no significativo si se añadiesen al modelo.

ADECUACIÓN DE LOS MODELOS SELECCIONADOS. ANÁLISIS DE RESIDUOS MEDIANTE GRÁFICOS

- **Hipótesis de linealidad**

La hipótesis de linealidad se estudia mediante el gráfico *Residuals vs Fitted,* un gráfico de dispersión en el cual se representan, en el eje de abscisas, los valores ajustados por el modelo de regresión (*fitted values*) y en el eje de ordenadas, los residuos de dichos ajustes (*residuals*). Este gráfico permite estudiar:

- La hipótesis de media cero, que se cumplirá cuando la curva de medias (línea roja del gráfico) coincida aproximadamente con el eje de abscisas, que es el valor cero de los residuos.
- La hipótesis de linealidad, cuando no se observe ningún patrón que sea indicio de requerir un modelo con efecto no lineal.

Además, también permite estudiar la presencia de *outliers*, es decir, observaciones para las cuales la respuesta y_i, es muy diferente al valor predicho por el modelo, $\hat{y}_i$. Debe valorarse si es necesario eliminar dichos datos de la base de datos, pues aunque suele llevar a una mejora más o menos significativa en la bondad de ajuste del modelo, los *outliers* pueden indicar deficiencias en el funcionamiento del modelo, y no simplemente errores en la recolección de los datos.

- **Hipótesis de varianza constante u homocedasticidad**

La estabilidad de la varianza residual (σ^2) se estudia mediante un gráfico llamado *Scale-location*. Para que se cumpla la hipótesis de varianza constante (situación conocida como homocedasticidad), no se debe observar ninguna tendencia en los datos de dicho gráfico. Es decir, que la línea del gráfico que indica la variación de la varianza (línea roja en el gráfico), debe permanecer constante (paralela al eje de abscisas).

- **Estudio de los puntos influyentes**

Existen observaciones, con valores inusuales de las covariables x_i, que puedan tener una influencia exagerada en el modelo. Estas observaciones se denominan "puntos de alto leverage", y son puntos que de forma individual, son capaces de afectar al comportamiento del modelo (p.e., por influir en la inclinación de la pendiente).

Estos puntos se estudian mediante el gráfico *Residuals vs. Leverage*, en el cual viene recogida la distancia de Cook, herramienta estadística que mide el efecto de eliminar una observación dada.

En el análisis de los modelos candidatos seleccionados, se considerará el efecto sobre las estimaciones asociado a la eliminación de datos con distancia de Cook mayor que 0.5.

- **Hipótesis de distribución normal**

Se estudia la hipótesis de distribución normal de los residuos mediante dos tipos de gráficos, gráfico de probabilidad normal e histograma de residuos.

El gráfico de probabilidad normal *(Normal Q-Q)* es un gráfico de distribución de los residuos estandarizados. Para considerar que la distribución de los residuos es normal, los puntos del gráfico deben estar, bien sobre la diagonal, bien lo más cercanos posible a ella. Además, si los residuos siguen una distribución normal, el 68% de los residuos estandarizados deberían quedar en el intervalo entre (-1,+1), y el 95% de ellos entre (-2,+2).

Cuando el tamaño muestral es grande, la distribución de los residuos también se puede estudiar mediante un histograma. En este caso, para considerar que la distribución de los residuos es normal, el histograma debe tener una apariencia similar a la campana de Gauss.

VALIDACIÓN

Se llama error de validación (*test error)* a la desviación típica del residuo que resulta de usar un método estadístico para predecir la respuesta en una nueva observación (es decir, una medida que no se ha usado para llevar a cabo el ajuste del método). Este error de validación se puede calcular fácilmente, en caso de disponer de un nuevo conjunto de observaciones (o conjunto *test*), simplemente aplicando a dicho conjunto el método estadístico desarrollado. Es una medida importante ya que nos muestra la capacidad predictiva real del modelo. En contraste, el error de entrenamiento es la media del error que resulta de aplicar el método estadístico a las observaciones usadas en su propio ajuste. Normalmente, difiere del error de validación y tiende a infraestimarlo (James et al., 2013).

Ya que en este caso no ha sido disponible validar el modelo en un nuevo conjunto de datos, se ha validado mediante un método de remuestreo llamado validación cruzada. Este método permite "apartar" del proceso de ajuste del modelo un subconjunto de las observaciones en estudio, para aplicar finalmente el método estadístico a ese subconjunto, hallando así el error de validación. El método más utilizado es la validación cruzada para *k* subconjuntos (*k-fold-cross-validation).*

Esta validación se lleva a cabo en dos pasos: la división de la base de datos disponible y el reajuste del modelo para dicha base de datos.

1. <u>División de la base de datos disponible</u>

En primer lugar, se divide la base de datos en *k* partes. En general, la base de datos se divide en 5 o 10 partes o conjuntos de datos (*folds*), que incluyan igual número de observaciones. El número de conjuntos dependerá del número de observaciones de que disponga la base de datos en estudio. Esta división de observaciones entre los *k* conjuntos puede hacerse de forma:

- Consecutiva (*consecutive*), tomando como conjunto a una sucesión de observaciones (Figura 8).

- Aleatoria (*random*), es decir, asignando las observaciones a los conjuntos de forma aleatoria (Figura 9).

2. <u>Reajuste del modelo</u>

Una vez se han dividido las observaciones en *k* subconjuntos, se lleva a cabo el reajuste del modelo, en *k* iteraciones. Así, en la primera iteración, el primer subconjunto (*fold*) se usa como *datos de validación* (es decir, datos que se dejan apartados del proceso de ajuste del modelo) y el resto de datos se usan como *datos de entrenamiento*, en los cuales se ajusta el modelo. El hecho de separar un cierto número de observaciones del proceso de ajuste, permite validar el modelo posteriormente, es decir, utilizar dichas observaciones como si de un conjunto de observaciones "nuevas" se tratara.

Este proceso se repite *k* veces, y en cada una los datos de prueba y de entrenamiento cambian. Al final del proceso se obtienen *k* estimaciones del error de validación, en forma de desviación típica residual. La estimación del proceso de validación cruzada para el modelo estudiado, se calcula promediando los *k* valores obtenidos de la desviación típica residual (James et al., 2013).

La validación cruzada se llevará a cabo en R mediante la función *cvFit* del paquete estadístico *cvTools* (Alfons, 2012).

ANEXO III. ESTUDIOS ADICIONALES

DESCOMPOSICIÓN DE SERIES TEMPORALES EN R

La Figura A. 43 muestra el estudio de la descomposición de la serie temporal en estudio (BD.verano) en R. En ella se puede observar fácilmente la existencia de ciclo y estacionalidad (*seasonal*), aunque no se aprecia ningún patrón en la tendencia (*trend*).

Por tanto, se concluye que la serie de datos reales en estudio no posee una tendencia, probablemente debido a que el periodo en estudio no es lo suficientemente largo como para apreciarla.

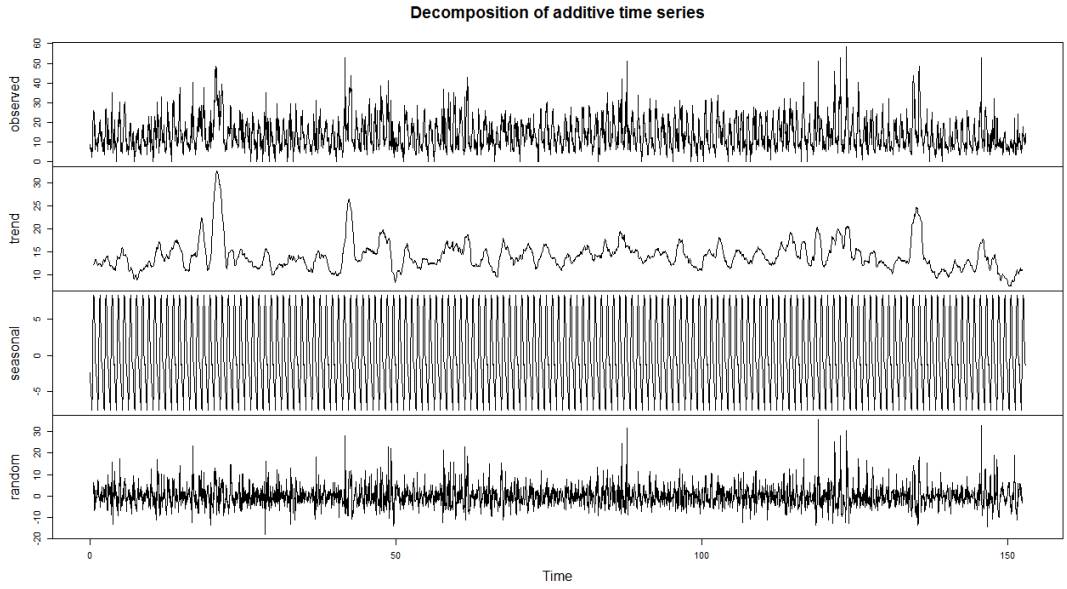


Figura A. 43. Descomposición de la BD.max2, como serie de datos temporal. En esta descomposición se puede observar una estacionalidad, pero no una tendencia clara.

ESTUDIO DE LOS DATOS DE TEMPERATURA REAL

- **Estudio de la relación instantánea**

Se ha realizado un estudio paralelo al desarrollo del trabajo, con el fin de averiguar si existe alguna relación entre los datos de velocidad real y temperatura, pues vista la estacionalidad diaria del régimen de vientos, se cree que también existirá dicha relación lineal entre ambas covariables, ya que el régimen de temperaturas también ofrece diferencias notables entre sus medidas registradas a lo largo del día. Los datos de temperatura real utilizados se han seleccionado con el mismo procedimiento usado para la selección de los datos de dirección correspondientes a las velocidades de viento de interés (ver descripción de P2 en apartado 2.1.2 de la metodología).

El gráfico de dispersión entre la velocidad real y la temperatura real, recogido en la Figura A. 44, muestra la existencia de una clara relación lineal entre ambas covariables; en términos generales, conforme aumenta la velocidad del viento aumenta la temperatura, y viceversa (línea azul gráfico). Esta situación se cumple bastante bien hasta velocidades de 25-30 km/h, aunque no para

velocidades mayores (datos incluidos en el círculo verde en el gráfico), que seguramente, no sigan un comportamiento "normal", si no que se trate de eventos puntuales.

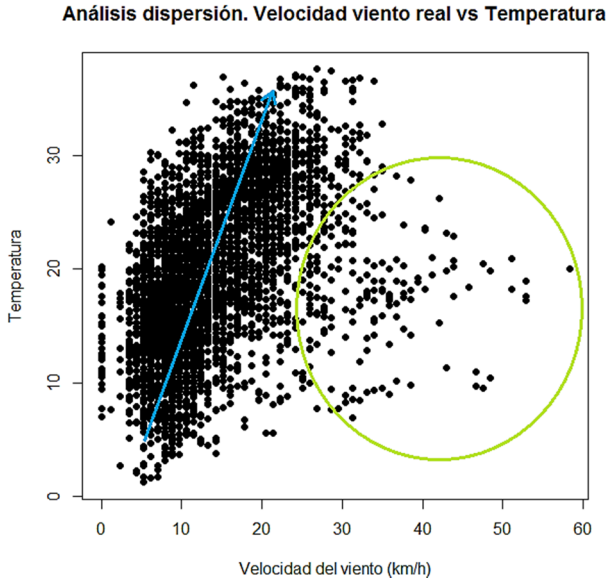


Figura A. 44 Gráfico de dispersión entre las covariables velocidad real de viento y temperatura.

- **Estudios de correlación**

La correlación instantánea entre las covariables temperatura real y velocidad del viento real (de la BD.verano) es de 0.4466, lo cual indica que existe una relación lineal importante entre ambas variables. Además, al analizar el gráfico de correlación cruzada (ver Figura A. 45), se observa que ambas variables siguen un ciclo sinusoidal, donde la tendencia de dicha curva cambia cada 12 horas.

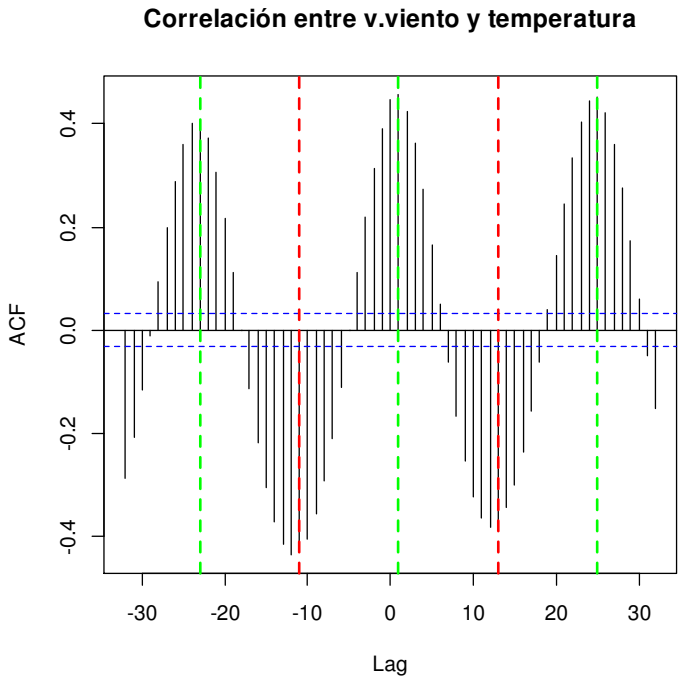


Figura A. 45 Correlograma entre covariables de temperatura y velocidad real del viento, para el periodo en estudio.

Los resultados obtenidos mediante el análisis exploratorio de los datos reales de temperatura, hacen pensar que el ciclo diario existente en la velocidad del viento, puede estar altamente influenciado por la variación diaria de la temperatura.

Sin embargo, no se ha podido disponer de datos de predicción de temperatura y por tanto, no se ha podido completar con ellos el modelo estadístico de predicción desarrollado.

ESTUDIO DEL PRONÓSTICO DE LOS MODELOS SELECCIONADOS. UMBRAL DE 20 KM/H

Ya que el interés del trabajo radica en las velocidades de viento más elevadas (superiores a 30 km/h), se ha incluido en este apartado el estudio realizado para el pronóstico de los modelos *ms1* y *ms2* finalmente seleccionados, para el umbral de velocidades superiores a 20 km/h, ya que como se ha dicho, tienen un menor interés para este trabajo.

Se ha llevado a cabo el mismo proceso realizado en el apartado 4.2.5. En primer lugar, se han llevado a cabo tablas de contingencia para comparar el comportamiento del modelo forma instantánea, tomando como umbral v>20 km/h. En este caso, el modelo *ms1* ofrece un pronóstico instantáneo acertado en un 51.5% de los casos para *ms1,* para el intervalo entre 20 y 30 km/h, y en un 51.1% de los casos para *ms2* (ver recuadros verdes en Tabla A. 46).

ms1	**Intervalos velocidad (km/h)**				
P(v>20 km/h)	**(0,10]**	**(10,20]**	**(20,30]**	**(30,40]**	**40**
(0, 0.25]	1325	1089	156	19	3
(0.25,0.50]	23	321	166	13	4
(0.50,0.75]	8	102	191	31	2
(0.75,1]	1	17	88	53	18
ms1					
(0, 0.25]	1318	1096	145	13	1
(0.25,0.50]	30	319	185	15	5
(0.50,0.75]	7	94	184	40	2
(0.75,1]	2	20	87	48	19

Tabla A. 46 Como Tabla 13, para probabilidad de superar los 20 km/h

En segundo lugar, se han construido tablas que incluyen el número de observaciones/día, y la capacidad de los modelos para pronosticar el umbral en estudio, en al menos una ocasión a lo largo del día en que se han dado las observaciones. Para el umbral de velocidad de 20 km/h, ambos modelos pronostican falsas alarmas (1 en el caso de *ms1* y 2 en el caso de *ms2*). Respecto al porcentaje de aciertos en días con 1-2 observaciones, éste supone un 26.47% para ambos modelos. El porcentaje de aciertos para días con más de 3 observaciones (y hasta 19), es de 78.4% para *ms1* y 82% para *ms2* (ver Tabla A. 47). Por tanto, para el umbral de 20 km/h, es el modelo *ms2* el que funciona ligeramente mejor.

ms1	**Nº horas/día con observaciones >20 km/h**		
Intervalos P(v>20 km/h)	**0**	**1-2**	**3-19**
(0, 0.25]	3	4	2
(0.25,0.50]	6	19	22
(0.50,0.75]	1	8	45
(0.75,1]	0	1	42
ms2			
(0, 0.25]	1	2	0
(0.25,0.50]	7	21	20
(0.50,0.75]	2	7	46
(0.75,1]	0	2	45

Tabla A. 47 Pronóstico de los modelos *ms1* y *ms2* para número de días con un número determinado de observaciones > 20 km/h

SCRIPT DE R QUE PERMITE OBTENER PREDICCIONES A 24 HORAS, APLICANDO EL MODELO ÓPTIMO DESARROLLADO (*MS1*)

IMPORTANTE: ANTES DE EMPEZAR, HAY QUE INSTALAR LOS PAQUETES *lubridate*, *chron* y #### *modeest*

```
setwd('C:/')      # Aquí es donde se encuentra el archivo Excel al cual introducir predicciones de
                  # AEMET, para el día de interés (entre las 3 am y la 1 am). El archivo se debe
                  # llamar prediccion y tener formato .csv. Los decimales de la velocidad del viento
                  # vendrán especificados con puntos. Fecha y hora se especificarán en la misma
                  # columna fecha.hora, predicción de la velocidad del viento vendrá en la
                  # columna p.v.viento y predicción de la dirección del viento en la columna
                  # p.dir.viento

prevision.hoy <- read.csv("prediccion.csv", header=T, sep=';')
prevision.hoy <- as.data.frame(prevision.hoy)
attach(prevision.hoy)

# 1. Se separa “hora” para que el programa la pueda leer

    auxiliar.fecha<-strptime(fecha.hora, "%d/%m/%Y %H:%M")
    require(lubridate) # devuelve mes como número
    require(chron)
    require(modeest)
    prevision.hoy$año<-year(auxiliar.fecha)
    prevision.hoy$mes<-month(auxiliar.fecha)
    prevision.hoy$dia<-day(auxiliar.fecha)
    prevision.hoy$hora<-hour(auxiliar.fecha)

# 2. Se crea un data.frame que contenga todas las covariables necesarias que el modelo necesita

    # 1. Se crean covariables post y lag, de v.viento y de p.v.viento
```

```
# Para v.viento
    attach(prevision.hoy)
    auxiliar.nombre.variable <- 'p.v.viento'
    auxiliar.variable.interes <- prevision.hoy[[auxiliar.nombre.variable ]]

    for(i.lag in 1:4) {
        auxiliar.nombre.variable.interes <- paste(auxiliar.nombre.variable, 'lag', i.lag,
        sep='.', collapse='')
        auxiliar.variable<-c(rep(NA,i.lag),
        auxiliar.variable.interes[1:(length(auxiliar.variable.interes)-i.lag)])
        prevision.hoy <- cbind(prevision.hoy, auxiliar.variable)
        names(prevision.hoy)[length(names(prevision.hoy))] <-
        auxiliar.nombre.variable.interes

    }

    for(i.post in 1:4) {
        auxiliar.nombre.variable.interes <- paste(auxiliar.nombre.variable, 'post', i.post,
        sep='.', collapse='')
        auxiliar.variable <-
        c(auxiliar.variable.interes[(1+i.post):length(auxiliar.variable.interes)],rep(NA,
        i.post)) # movemos los datos
        prevision.hoy <- cbind(prevision.hoy, auxiliar.variable)
        names(prevision.hoy)[length(names(prevision.hoy))] <-
        auxiliar.nombre.variable.interes
    }

# Para p.v.viento
    auxiliar.nombre.variable <- 'p.dir.viento'
    auxiliar.variable.interes <- prevision.hoy[[auxiliar.nombre.variable ]]

    for(i.lag in 1:4) {
        auxiliar.nombre.variable.interes <- paste(auxiliar.nombre.variable, 'lag', i.lag,
        sep='.', collapse='')
        auxiliar.variable <-c(rep(NA,i.lag),
        auxiliar.variable.interes[1:(length(auxiliar.variable.interes)-i.lag)])
        prevision.hoy <- cbind(prevision.hoy, auxiliar.variable)
        names(prevision.hoy)[length(names(prevision.hoy))]<-
        auxiliar.nombre.variable.interes
    }

    for(i.post in 1:4) {
        auxiliar.nombre.variable.interes <- paste(auxiliar.nombre.variable, 'post', i.post,
        sep='.', collapse='')
        auxiliar.variable <-
        c(auxiliar.variable.interes[(1+i.post):length(auxiliar.variable.interes)],rep(NA,
        i.post))  # movemos los datos
```

```
        prevision.hoy <- cbind(prevision.hoy, auxiliar.variable)
        names(prevision.hoy)[length(names(prevision.hoy))]<-
        auxiliar.nombre.variable.interes
    }

    prevision.hoy # comprobamos que se han añadido las nuevas covariables

# Se cambia el el nombre a las covariables anteriores, para facilitar su manejo

  # velocidad
    detach()
    attach(prevision.hoy)
    x <- p.v.viento
    x1 <- p.v.viento.lag.1
    x2 <- p.v.viento.lag.2
    x3 <- p.v.viento.lag.3
    x4 <- p.v.viento.lag.4
    x.p1 <- p.v.viento.post.1
    x.p2 <- p.v.viento.post.2
    x.p3 <- p.v.viento.post.3
    x.p4 <- p.v.viento.post.4

  # dirección
    dx <- p.dir.viento
    d1 <- p.dir.viento.lag.1
    d2 <- p.dir.viento.lag.2
    d3 <- p.dir.viento.lag.3
    d4 <- p.dir.viento.lag.4
    d.p1 <- p.dir.viento.post.1
    d.p2 <- p.dir.viento.post.2
    d.p3 <- p.dir.viento.post.3
    d.p4 <- p.dir.viento.post.4

    # Se crean los armónicos de dirección de viento
      cosdv <- cos(dx/360 * 2 * pi)
      sendv <- sin(dx/360 * 2 * pi)
      cosd1 <- cos(d1/360 * 2 * pi)
      send1 <- sin(d1/360 * 2 * pi)
      cosd2 <- cos(d2/360 * 2 * pi)
      send2 <- sin(d2/360 * 2 * pi)
      cosd3 <- cos(d3/360 * 2 * pi)
      send3 <- sin(d3/360 * 2 * pi)
      cosd4 <- cos(d4/360 * 2 * pi)
      send4 <- sin(d4/360 * 2 * pi)
      cosd.p1 <- cos(d.p1/360 * 2 * pi)
      send.p1 <- sin(d.p1/360 * 2 * pi)
```

```
        cosd.p2 <- cos(d.p2/360 * 2 * pi)
        send.p2 <- sin(d.p2/360 * 2 * pi)
        cosd.p3 <- cos(d.p3/360 * 2 * pi)
        send.p3 <- sin(d.p3/360 * 2 * pi)
        cosd.p4 <- cos(d.p4/360 * 2 * pi)
        send.p4 <- sin(d.p4/360 * 2 * pi)

prediccion.hoy <- cbind(mes, dia, hora, x,x1,x2,x3,x4,x.p1,x.p2,x.p3,x.p4,
                        sendv, send1, send2, send3, send4, send.p1, send.p2,
                        send.p3, send.p4, cosdv, cosd1, cosd2, cosd3, cosd4,
                        cosd.p1, cosd.p2, cosd.p3, cosd.p4)

summary(prediccion.hoy)       # Se comprueba que la BD se ha creado correctamente
prediccion.hoy <- as.data.frame(prediccion.hoy)

detach()
attach(prediccion.hoy)

# Se crean armónicos de estacionalidad y ciclos

    # Estacionalidad
        dias.transcurridos <- 121
        dia.estudio <- rep(dias.transcurridos + prediccion.hoy$dia + (prediccion.hoy$mes -
        5)*31)
            sen.dia <- sin(dia.estudio *2*pi/365)
            cos.dia <- cos(dia.estudio *2*pi/365)
            sen.dia2 <- sin(dia.estudio *4*pi/365)
            cos.dia2 <- cos(dia.estudio *4*pi/365)

    # Ciclo
            sen.hora <- sin(prediccion.hoy$hora*2*pi/24)
            cos.hora <- cos(prediccion.hoy$hora*2*pi/24)
            sen.hora2 <- sin(prediccion.hoy$hora*4*pi/24)
            cos.hora2 <- cos(prediccion.hoy$hora*4*pi/24)
            sen.hora3 <- sin(prediccion.hoy$hora*8*pi/24)
            cos.hora3 <- cos(prediccion.hoy$hora*8*pi/24)
            sen.hora4 <- sin(prediccion.hoy$hora*16*pi/24)
            cos.hora4 <- cos(prediccion.hoy$hora*16*pi/24)

prediccion.hoy <- cbind(prediccion.hoy, sen.dia, cos.dia, sen2.dia, cos2.dia, sen.hora,
cos.hora, sen2.hora, cos2.hora, sen3.hora, cos3.hora, sen4.hora, cos4.hora)
```

```
# 3. Se aplica el modelo a nuevo conjunto de datos, con función predict()
        # Se carga área de trabajo BD.final.RData (donde está el modelo óptimo ms1
        # guardado). Es importante cargarla en este momento y no antes, ya que de lo
        # contrario la base de datos no se creará correctamente

    prediccion.prueba <- predict.lm(ms1, newdata=prediccion.hoy, se.fit=TRUE, type='response',
    terms=NULL)

# 4. Se calculan los pronósticos, para umbrales de velocidad de 40, 30, 20 y 10 km/h

    horas.utiles <- c(7:21)
    auxiliar.p.40 <- pnorm(sqrt(40) , mean=na.omit(prediccion.prueba$fit),
    sd=prediccion.prueba$residual.scale)
    auxiliar.p.30 <- pnorm(sqrt(30) , mean=na.omit(prediccion.prueba$fit) ,
    sd=prediccion.prueba$residual.scale)
    auxiliar.p.20 <- pnorm(sqrt(20) , mean=na.omit(prediccion.prueba$fit),
    sd=prediccion.prueba$residual.scale)
    auxiliar.p.10 <- pnorm(sqrt(10) , mean=na.omit(prediccion.prueba$fit),
    sd=prediccion.prueba$residual.scale)

plot(horas.utiles, (na.omit(prediccion.prueba$fit))^2, ylim=c(0,60),
    xlab='hora',ylab='v.viento (km/h)', pch=19, main='Gráfico predicciones instantáneas y
pronósticos')
points(prevision.hoy$v.viento[horas.utiles], pch=19, col='blue')
abline(h=c(10, 20,30,40), lty=2)
lines(horas.utiles, (1-auxiliar.p.40)*60,col='red', lwd=2)
abline(h=c(0.25,0.5,0.75)*60,col='red',lty=2)
axis(4,at=60*c(0.25,0.5,0.75,1),labels=c(0.25,0.5,0.75,1),tick=T)
mtext(side=4,text='P (y> umbral)')
lines(horas.utiles,(1-auxiliar.p.30)*60,col='orange',  lwd=2)
lines(horas.utiles,(1-auxiliar.p.20)*60,col='green',  lwd=2)
lines(horas.utiles,(1-auxiliar.p.10)*60,col='yellow',  lwd=2)

legend(7,60,legend=c('umbral = 40 km/h','umbral = 30 km/h','umbral = 20 km/h', 'umbral = 10
km/h'), col=c('red','orange','green','yellow'), lty=c(1,1))

# 4. Por último, se crea el vector con predicciones instantáneas

    hora.estudio <- as.data.frame(horas.utiles)
    predicciones.instantaneas <- (na.omit(prediccion.prueba$fit))^2

    archivo.predicciones.instantaneas <- cbind(hora.estudio, predicciones.instantaneas)
```