

RESUMEN

La energía es una de las claves esenciales que marcan el nivel de desarrollo de un país o región, erigiéndose como uno de los retos actuales la búsqueda de soluciones que permitan satisfacer la cobertura de la demanda eléctrica con los mayores niveles posibles de independencia energética y sostenibilidad económica. Ante el posible agotamiento de los recursos energéticos fósiles, la energía eólica se presenta como una alternativa renovable, sólida y madura que podría dar cobertura, al menos es una parte importante a la demanda creciente en los sistemas eléctricos. Dichos escenarios de máxima penetración de energía eólica sólo son posibles si se disponen de medios para cuantificar el potencial eólico de los proyectos singulares, ya sea con fines relacionados al prediseño de las instalaciones o para el desarrollo de predicciones que ayuden en su gestión a tiempo real, paliando una de las principales dificultades de dicha tecnología, la intermitencia del recurso eólico.

En esta tesis, se evalúa la influencia de determinados parámetros relacionados con aspectos meteorológicos y técnicos, con el propósito de mejorar las estimaciones de potencia eólica cuando se emplean técnicas de Machine Learning (ML) para la predicción. Asimismo, se evalúa la eficiencia de tres técnicas ML, Support Vector Regression (SVR), Random Forest (RF) y Artificial Neural Networks (ANN), en la búsqueda de aquella alternativa que pueda considerarse como más eficiente en la resolución de problemas relacionados con el ajuste a largo (metodología MCP). Entre las soluciones propuestas, se han planteado el uso de métodos Feature Selection (FS) para la búsqueda automática y exhaustiva de las variables que aportan significancia a los modelos, y un test no paramétrico de permutación pareado con el que se evalúa de una manera más objetiva la fiabilidad de los modelos cuando se comparan entre ellos.

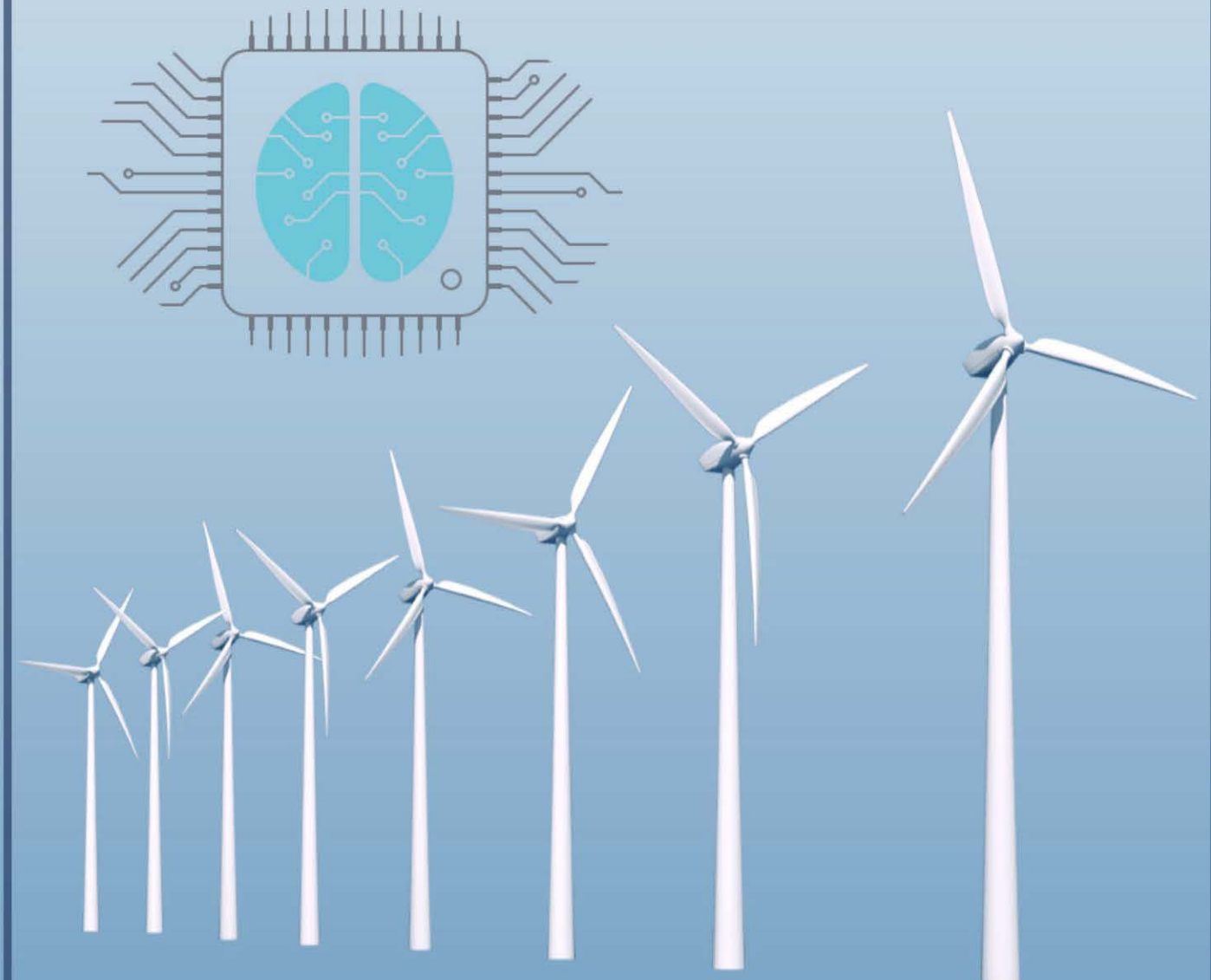
De entre los principales resultados obtenidos cabe destacar que, frente a la tendencia de sólo considerar las velocidades y direcciones como variables de partida de estos modelos, la inclusión de la densidad del aire no sólo mejora los resultados sino que lo hace de forma significativa tanto cuando el objetivo es la estimación de las densidades de potencia (WPD) o la potencia de una turbina eólica (WTPO). De las tres técnicas ML evaluadas, SVR y RF presentan una mejora significativa a la tradicional arquitectura ANN, aspectos que han sido discutidos y publicados en el marco de la presente tesis doctoral.



ANÁLISIS DE LA INFLUENCIA DE PARÁMETROS METEOROLÓGICOS Y Y FUNCIONALES EN LA ESTIMACIÓN DE LA POTENCIA EÓLICA MEDIANTE EL EMPLEO DE TÉCNICAS DE MACHINE LEARNING

Santiago Díaz Ruano

Las Palmas de Gran Canaria, Julio 2018





UNIVERSIDAD DE LAS PALMAS DE GRAN CANARIA
ESCUELA DE INGENIERÍAS INDUSTRIALES Y CIVILES
Doctorado en Ingenierías Química, Mecánica y de Fabricación

TESIS DOCTORAL

**ANÁLISIS DE LA INFLUENCIA DE PARÁMETROS METEOROLÓGICOS Y FUNCIONALES
EN LA ESTIMACIÓN DE LA POTENCIA EÓLICA MEDIANTE EL EMPLEO DE
TÉCNICAS DE MACHINE LEARNING**

Santiago Díaz Ruano

Directores

Dr. D. José Antonio Carta González

Dr.D. José María Matías Fernández

LAS PALMAS DE GRAN CANARIA, JULIO 2018

LA DRA. DÑA. MARÍA DOLORES MARRERO ALEMÁN, SECRETARIA DEL PROGRAMA DE DOCTORADO EN INGENIERÍAS QUÍMICA, MECÁNICA Y DE FABRICACIÓN DE LA UNIVERSIDAD DE LAS PALMAS DE GRAN CANARIA,

CERTIFICA,

Que la Comisión Académica de programa de Doctorado tomó el acuerdo de dar el consentimiento para su tramitación a la tesis doctoral titulada *Análisis de la influencia de parámetros meteorológicos y funcionales en la estimación de la potencia eólica mediante el empleo de técnicas de machine Learning*, presentada por el doctorando D. Santiago Díaz Ruano y dirigida por los doctores D. José Antonio Carta González y D. José María Matías Fernández.

Y para que así conste, a efectos de lo previsto en el reglamento de Estudios de Doctorado de esta Universidad, firmo la presente en Las Palmas de Gran Canaria, 26 Agosto de 2018.

Fdo: Dra. Dña. María Dolores Marrero Alemán

RESUMEN

La energía es una de las claves esenciales que marcan el nivel de desarrollo de un país o región, erigiéndose como uno de los retos actuales la búsqueda de soluciones que permitan satisfacer la cobertura de la demanda eléctrica con los mayores niveles posibles de independencia energética y sostenibilidad económica. Ante el posible agotamiento de los recursos energéticos fósiles, la energía eólica se presenta como una alternativa renovable, sólida y madura que podría dar cobertura, al menos es una parte importante a la demanda creciente en los sistemas eléctricos. Dichos escenarios de máxima penetración de energía eólica sólo son posibles si se disponen de medios para cuantificar el potencial eólico de los proyectos singulares, ya sea con fines relacionados con el prediseño de las instalaciones o para el desarrollo de predicciones que ayuden en su gestión a tiempo real, paliando una de las principales dificultades de dicha tecnología, la intermitencia del recurso eólico.

En esta tesis, se evalúa la influencia de determinados parámetros relacionados con aspectos meteorológicos y técnicos, con el propósito de mejorar las estimaciones de potencia eólica cuando se emplean técnicas de Machine Learning (ML) para la predicción. Asimismo, se evalúa la eficiencia de tres técnicas ML, Support Vector Regression (SVR), Random Forest (RF) y Artificial Neural Networks (ANN), en la búsqueda de aquella alternativa que pueda considerarse como más eficiente en la resolución de problemas relacionados con el ajuste a largo (metodología MCP) y a corto plazo (predicción). Entre las soluciones propuestas, se han planteado el uso de métodos Feature Selection (FS) para la búsqueda automática y exhaustiva de las variables que aportan significancia a los modelos, y un test no paramétrico de permutación pareado con el que se evalúa de una manera más objetiva la fiabilidad de los modelos cuando se comparan entre ellos.

De entre los principales resultados obtenidos cabe destacar que, frente a la tendencia de sólo considerar las velocidades y direcciones como variables de partida de estos modelos, la inclusión de la densidad del aire no sólo mejora los resultados sino que lo hace de forma significativa tanto cuando el objetivo es la estimación de las densidades de potencia (WPD) o la potencia de una turbina eólica (WTPO). De las tres técnicas ML evaluadas, SVR y RF presentan una mejora significativa a la tradicional arquitectura ANN, aspectos que han sido discutidos y publicados en el marco de la presente tesis doctoral.

AGRADECIMIENTOS

El desarrollo de una tesis doctoral es una carrera de fondo, bien lo saben todos aquellos que han pasado por esta fase antes que yo. No obstante, para mí ha sido un verdadero placer dedicar estos últimos tres años a su desarrollo y me anima a continuar con mi actividad investigadora de forma activa ya que ahora cuento con conocimientos con los que antes no disponía y, en este trayecto, he coincidido con un gran número de personas quienes me han ayudado a desarrollarme de manera personal y profesional. En las próximas líneas quiero expresar mi más sincera gratitud a todas estas personas.

En primer lugar quiero dar las gracias a José Antonio Carta y José María Matías, mis dos directores de tesis. José Antonio Carta ha sido mi gran apoyo en la realización de este trabajo, persona por la que siento una inmensa gratitud y admiración. Hemos trabajado mano a mano durante estos años, periodos en los que incluso me ha dedicado su tiempo libre porque yo debía adaptar el desarrollo de la tesis con mi trabajo, horas que, a pesar de suponer algún que otro dolor de cabeza para los dos causados por los modelos que tratábamos de implementar juntos, se pasaban volando y siempre conseguimos encausar gracias a su enorme experiencia, sabiduría y lo que más admiro de él, su persistencia, vital para nuestra profesión de ingenieros, nunca escatimando esfuerzos cuando, a pesar de lograrse una solución, ésta podía mejorarse. Nadie mejor que tú podría haberme introducido en la actividad investigadora y estoy seguro que continuaremos desarrollando trabajos juntos. Gracias también a José María Matías por asesorarme con sus conocimientos en materia estadística. Admiro su capacidad de análisis y razonamiento. Gracias por tu ayuda y por, a pesar de la distancia, siempre mostrarte dispuesto a enseñarme y cooperar en los trabajos que hemos realizado juntos. También quiero dar las gracias a Alejandro Yáñez, mi tutor de tesis, él me dio las claves para realizar búsquedas efectivas de referencias bibliográficas, enseñándome incluso a manejar gestores bibliográficos los cuales han sido de mucha utilidad para la realización de estos trabajos. Gracias también por ayudarme con todos los trámites administrativos que han supuesto la tesis doctoral.

También me siento muy afortunado de haber formado parte del Instituto Tecnológico de Canarias (ITC), donde inicié mi etapa laboral y aún continúo adquiriendo conocimientos técnicos en materia de energías renovables y computación gracias al gran equipo humano que lo forma. Obviamente, si hoy presento mi tesis doctoral en materia de energía eólica es porque ellos me han impulsado en esta dirección. Agradecer en especial a Salvador Suárez, para mí una referencia al ser una de las personas más trabajadoras y con mayores conocimientos en materia de energías renovables de cuantas conozco. Gracias por el gran apoyo que me ha brindado desde el comienzo de mi andadura en esta empresa. A Daniel Henríquez, otro gran apoyo, referencia y magnífico jefe, quien me ha enseñado a desenvolverme en proyectos de cooperación internacional, manejando como nadie los conocimientos tecnológicos y con el quien me encanta entablar conversaciones técnicas porque siempre salen buenas ideas de ellas, muchas de las cuales hemos llevado a la práctica juntos en estos dos últimos años y las que quedan por venir. También me gustaría agradecer especialmente a Fernando Castellano, gran ingeniero y mejor amigo, fuistes la primera persona que me dio una oportunidad laboral, me introdujistes en el mundo de las energías renovables y ha sido un placer trabajar contigo durante estos años. Asimismo, quisiera agradecer a Tomas Cambreleng, una de las personas que conozco con mayores y mejores conocimientos sobre energía eólica y quien me

introdujo sin ninguna duda en esta línea de investigación, a mis compañeros de equipo de la sección de microredes aisladas, Rayco Parra y Deullanela Díaz, es para mí un placer trabajar y continuar aprendiendo con ustedes todos los días. También quiero expresar mi gratitud a Rafael Nebot, quien me ha aportado grandes conocimientos en el ámbito computacional y me ha ayudado a consolidar conocimientos en el campo de la predicción. Siempre ha estado ahí cuando me han surgido dudas y siempre ha sido capaz de resolvérmelas, razón por la que lo considero a día de hoy como una referencia de lo máximo a lo que podría aspirar en mi vida profesional. Gracias a Cataisa Santana, Dunia Mentado, Delia Cabrera, Elías Medina, Jesús de León, Celia Bueno y Ramón García por haberos interesado en todo momento sobre los trabajos que iba haciendo en mi tesis doctoral, ayudando con vuestra experiencia y apoyándome moralmente en los momentos más duros durante su desarrollo. Estos agradecimientos los hago extensibles al resto de compañeros del Departamento de Energías Renovables y al Departamento de Computación científica y tecnológica, gracias a todos porque haber estado ahí siempre.

Por último, pero no menos importante, es más, yo diría que mis principales puntales tanto durante la realización de mi tesis doctoral como de cualquier proyecto que he emprendido hasta el momento, quiero dar las gracias a mis padres Santiago Díaz e Ignacia Ruano, a quienes no sólo quiero sino que idolatro, quienes me han enseñado a ser todo lo que soy y quienes me han inculcado la cultura del esfuerzo ya que, en esta vida, todo siempre sale a base de constancia. Incluyo en este grupo a mis hermanas Tita, María Pino y Vanessa Díaz con las que siempre he estado muy unido y quienes sólo con mirarme son capaces de dar con las palabras clave para sacar siempre lo mejor de mí, a mis sobrinos Yoel Bautista, Yanira López, Airam Bautista y Carla Acosta, todos encantadores y con una capacidad innata de sacarme una sonrisa sólo con mirarme con independencia del estado de ánimo que pudiera tener, y a mis cuñados Santiago López, Juan Acosta y José Bautista, los cuales son también parte importante de mi núcleo familiar.

A todos ustedes, gracias por vuestra ayuda.

Santiago Díaz,
Julio 2018

Índice general

1. Introducción general.....	9
1.1. Introducción	9
1.2. Objetivos de la tesis.....	13
1.3. Metodología de la tesis.....	14
1.4. Estructura de la tesis doctoral.....	14
2. Introducción al estudio del recurso eólico	17
2.1. Introducción	17
2.2. Características generales del recurso eólico.....	17
2.3. Medición del viento	22
2.3.1. Parámetros básicos de medida.....	22
2.3.2. Registro de datos y dispositivos de almacenamiento	25
2.3.3. Duración de campaña de medida	27
2.4. Representación de los datos de viento en el análisis	28
2.4.1. Distribuciones temporales.....	28
2.4.2. Distribuciones de frecuencia	30
2.5. Extrapolación de los datos en altura	32
2.6. Principios de conversión de la energía eólica	34
2.6.1. Energía eólica disponible.....	34
2.6.2. Energía eólica aprovechable.....	35
2.6.3. Estimación de la energía producida	38
2.7. Estimación de la densidad del aire	39
2.8. Formulación usada para la estimación de la potencia eólica a través de datos medidos en las estaciones meteorológicas.....	44
2.9. Necesidad del estudio del potencial a largo plazo	47
2.10. Empleo de técnicas MCP para el estudio del recurso eólico a largo plazo.....	49
2.11. Requisitos de los datos de partida para la estimación del recurso eólico a largo plazo empleando métodos MCP.....	53
2.11.1. Correlación entre los datos de la estación candidata y la estación de referencia.....	53
2.11.2. Amplitud mínima de los datos de la estación candidata.....	54
2.11.3. Estabilidad climática a largo plazo	55
2.11.4. Adecuación de los protocolos de medida	56
2.11.5. Alternativas al empleo de estaciones meteorológicas de superficie.....	57

2.11.6. Influencia de la dirección del viento en la caracterización del modelo	58
2.12. Potencial de las técnicas de Machine Learning para el estudio del recurso eólico	59
3. Técnicas de Machine Learning para el estudio del potencial eólico a largo plazo	63
3.1. Introducción.....	63
3.2. Enfoque de regresión para la resolución de problemas MCP.....	64
3.3. Least Squares Multiple Linear Regression	65
3.4. Artificial Neural Networks (ANN)	66
3.5. Support Vector Regression (SVR)	69
3.6. Random Forest (RF)	74
3.7. Implementación de las técnicas seleccionadas.....	77
3.7.1. Support Vector Machine	78
3.7.2. Random Forest.....	92
3.7.3. Artificial Neural Networks	100
4. Nuevos métodos de estimación de la potencia eólica mediante técnicas de Machine Learning	109
4.1. Evaluación del recurso eólico en Canarias. Muestra de datos.....	109
4.1.1. Introducción	109
4.1.2. Medición meteorológica de vientos en Canarias	110
4.1.3. Estaciones anemométricas consideradas en los estudios.....	112
4.1.4. Análisis inicial del recurso mediante la validación de las muestras de datos	120
4.1.5. Análisis de correlaciones lineales y selección de estaciones que actúan como referencias y objetivos	123
4.2. Análisis comparativo entre varios modelos MCP que usando la técnica SVR estiman la densidad de potencia.....	126
4.2.1. Introducción	126
4.2.2. Antecedentes.....	126
4.2.3. Objetivo del estudio desarrollado.....	129
4.2.4. Muestras de datos usadas para el desarrollo del estudio.....	130
4.2.5. Descripción de las técnicas y modelos matemáticos usados para la simulación	131
4.2.6. Metodología	134
4.2.7. Análisis de resultados.....	141
4.2.8. Conclusiones del estudio.....	152

4.3. Evaluación de la fiabilidad de cinco modelos MCP para estimar las WTPO a largo plazo usando tres técnicas de Machine Learning	153
4.3.1. Introducción.....	153
4.3.2. Antecedentes.....	154
4.3.3. Objetivo del estudio desarrollado.....	156
4.3.4. Muestra de datos usadas para el desarrollo del estudio	157
4.3.5. Descripción de las técnicas y modelos matemáticos usados para la simulación	158
4.3.6. Metodología.....	162
4.3.7. Análisis de resultados.....	169
4.3.8. Conclusiones	183
5. Conclusiones y líneas futuras de investigación	185
5.1. Introducción	185
5.2. Contribución de la tesis al conocimiento de la industria eólica.....	186
5.3. Conclusiones	188
5.4. Líneas futuras de investigación	189
Referencias bibliográficas.....	193
Anexos. Publicaciones, congresos y valoración externa del trabajo realizado en la tesis	207
Introducción	207
A. Publicaciones.....	208
B. Congresos.....	210
C. Valoración externa del trabajo de investigación.....	222

Índice de Figuras

Figura 1 Media horaria de datos de velocidad del viento a 40 metros [Lanzarote]	19
Figura 2 Zonas de interés para ejecución de parques eólicos en presencia de obstáculos [46]	20
Figura 3 Efecto de un obstáculo sobre el perfil vertical de la velocidad del viento	22
Figura 4 Anemómetros tipo (cazoleta y hélice)	24
Figura 5 Armario de dispositivo de registro y almacenamiento y Datalogger	27
Figura 6 Velocidades medias del viento mensuales y anuales	28
Figura 7 Desviaciones típicas de la velocidad del viento mensual.....	29
Figura 8 Evaluación de la intensidad de la turbulencia para una posición determinada.....	30
Figura 9 Rosa de los vientos	30
Figura 10 Distribución de velocidades medias del viento [Lanzarote]	31
Figura 11 Energía extraída del tubo de corriente de aire en la turbina eólica	36
Figura 12 Representación límite de Betz.....	37
Figura 13 Densidad teórica de energía eólica extraíble con una turbina.....	38
Figura 14 Curvas de potencia utilizadas en los análisis de WTPO (Capítulo 9). Modelos: WT-1 (Derecha) y WT-2 (Izquierda).....	45
Figura 15 Sensibilidad de la rentabilidad frente a la producción eólica de un P.E de 2 MW	48
Figura 16 Arquitectura MCP	50
Figura 17 Validación cruzada	52
Figura 18 Distribución de velocidades medias del viento por sectores de dirección	58
Figura 19 Entrenamiento y validación del modelo de regresión (Fase 1).....	77
Figura 20 Comparativa de resultados del método Wrapper Approach implementado para SVR	86
Figura 21 Proceso de selección de parámetros mediante el proceso de validación cruzada	87
Figura 22 Ajuste de correlación R^2 obtenido tras ejecutar el modelo con 5 estaciones de referencia	90
Figura 23 Posiciones de las estaciones anemométricas [254]	115
Figura 24 Histogramas de velocidad y densidad de potencia, rosas de viento y perfiles medios mensuales y diarios para las estaciones WS-1 a WS-5 [268]	118
Figura 25 Histogramas de velocidad y densidad de potencia, rosas de viento y perfiles medios mensuales y diarios para las estaciones WS-6 a WS-10 [268]	119
Figura 26 Control de calidad y detección de errores WS-4.....	122
Figura 27 Esquema general de la metodología empleada para el desarrollo del estudio	135
Figura 28 Errores MAE obtenidos cuando se estiman las densidades de potencia eólica en cada modelo y estación objetivo seleccionada.....	142
Figura 29 Errores MARE obtenidos cuando se estiman las densidades de potencia eólica en cada modelo y estación objetivo seleccionada.....	142
Figura 30 Coeficientes R^2 obtenidos cuando se estiman las densidades de potencia eólica en cada modelo y estación objetivo seleccionada	143
Figura 31 Curvas de potencia utilizadas en los análisis de WTPO. Modelos: WT-1 (Derecha) y WT- 2 (Izquierda).....	162

Figura 32 Procedimiento llevado a cabo para el análisis comparativo desarrollado en este estudio.....	164
Figura 33 Valores medios y desviaciones estándar de las métricas MAE, MARE y R^2 obtenidos al aplicar 10 Folds Cross Validation cuando se estiman las WTPO con cinco modelos que usan tres técnicas ML para la estación WS-3	171
Figura 34 Valores medios y desviaciones estándar de las métricas MAE, MARE y R^2 obtenidos al aplicar 10 Folds Cross Validation cuando se estiman las WTPO con cinco modelos que usan tres técnicas ML para la estación WS-4	172
Figura 35 Valores medios y desviaciones estándar de las métricas MAE, MARE y R^2 obtenidos al aplicar 10 Folds Cross Validation cuando se estiman las WTPO con cinco modelos que usan tres técnicas ML para la estación WS-5	173
Figura 36 Frecuencias de la densidad del aire media y desviación estándar para el rango de velocidades de viento en operación en la estación WS-5.....	181
Figura 37 Comparativa entre WTPOs (datos observados, M1 y M3) medias mensuales y horarias del conjunto de testeo en corto plazo (2014) y evolución horaria para un periodo de 4 días seleccionado de manera arbitraria.	182

Índice de Tablas

Tabla 1 Valores tipo de longitud de rugosidad Z_0 en metros	21
Tabla 2 Coeficientes de las Ecuaciones 2.19, 2.20 y 2.21.....	40
Tabla 3 Coeficientes de las Ecuaciones 2.25, 2.26 y 2.27.....	42
Tabla 4 Coeficientes de las Ecuaciones 2.28 a 2.31	42
Tabla 5 Parámetros de cálculo de la función ksvm (Kernlab).....	83
Tabla 6 Librerías Random Forest recomendadas por la industria	93
Tabla 7 Librerías Artificial Neural Networks recomendadas por la industria.....	100
Tabla 8 Medida del recurso eólico en los aeropuertos canarios	110
Tabla 9 Buenas prácticas en la instalación de torres meteorológicas	113
Tabla 10 Estaciones anemométricas empleadas [254]	117
Tabla 11 Valores tipos de test de rango [182]	121
Tabla 12 Valores tipos de test relacional [182]	121
Tabla 13 Disponibilidad de los datos	123
Tabla 14 Análisis de correlaciones lineales entre velocidades del viento[254].....	124
Tabla 15 Análisis de correlaciones lineales entre densidades de potencia eólica [254].....	124
Tabla 16 Análisis de correlaciones lineales entre estaciones de referencia y objetivo (Velocidad del viento).....	125
Tabla 17 Análisis de correlaciones lineales entre estaciones de referencia y objetivo (WPD)	125
Tabla 18 Modelos considerados para la estimación de las densidades de potencia	133
Tabla 19 Coeficientes de correlación lineal entre velocidades del viento de las estaciones anemométricas evaluadas.....	137
Tabla 20 Coeficientes de correlación lineal entre densidades de potencia de las estaciones anemométricas evaluadas.....	137
Tabla 21 Inputs finalmente seleccionados	145
Tabla 22 Hiperparámetros de los modelos implementados	146
Tabla 23 Análisis de significancia estadística cuando la métrica analizada es MAE. Test pareado	147
Tabla 24 Análisis de significancia estadística cuando la métrica analizada es MARE. Test pareado	148
Tabla 25 Análisis de significancia estadística cuando la métrica analizada es R^2 . Test pareado	149
Tabla 26 Coeficientes de correlación lineal entre resultados de las métricas en cada Folds. WS-5 ...	152
Tabla 27 Coeficientes de correlación lineales entre velocidades del viento de las estaciones evaluadas.....	165
Tabla 28 Features no seleccionados en el procedimiento Wrapper	176
Tabla 29 Hiperparámetros	177
Tabla 30 Factores ξ seleccionados	177
Tabla 31 Tiempos medios de cómputo	177
Tabla 32 Análisis de diferencias estadísticamente significativas para la métrica MAE (kW). P-valores	178
Tabla 33 Análisis de diferencias estadísticamente significativas para la métrica MARE. P-valores	178
Tabla 34 Análisis de diferencias estadísticamente significativas para la métrica R^2 (%). P-valores	179
Tabla 35 Análisis de diferencias estadísticamente significativas entre M1 y M3. P-valores.....	180

CAPÍTULO

1

Introducción general

1.1. Introducción

En la sociedad actual la energía es una de las claves fundamentales de la situación económica y política, demostrándose la existencia de altas correlaciones entre el consumo energético y el nivel de desarrollo de un país [1-3]. Teniendo esto presente, las distintas instituciones gubernamentales impulsan el avance del sector energético mediante la apuesta por aquellas tecnologías capaces de satisfacer la cobertura de la demanda con los mayores niveles posibles de independencia energética y sostenibilidad económica. Ante el posible agotamiento de los recursos energéticos fósiles, la energía eólica se erige como una alternativa renovable, sólida y madura que puede dar respuesta, al menos en una parte importante, a la demanda creciente garantizando una competitividad económica adecuada. En este escenario, es **requisito indispensable el estudio detallado del potencial eólico**, logrando definir la cantidad de energía que podría ser generada durante la vida útil de un parque eólico y con ello la rentabilidad económica de dicha inversión. Este tipo de estudios es también vital para la fase de diseño puesto que permite determinar la clase de viento de la región (que define el tipo de aerogenerador a instalar) y para la operación del parque eólico dado que posibilita la gestión a tiempo real de las instalaciones. En este sentido, se distinguen entre estimaciones a corto y largo plazo.

Las estimaciones del recurso eólico a corto plazo están enfocadas fundamentalmente a la gestión en tiempo real de la generación eólica y la búsqueda de fluctuaciones del recurso que puedan derivar en pérdidas de estabilidad en el sistema eléctrico, ejecutándose pronósticos para un periodo temporal normalmente inferior a seis horas [4,5], si bien otros [6] emplean dicha metodología para predecir el comportamiento del recurso eólico hasta horizontes de 72 horas. Esta información de partida es de vital importancia para asegurar la correcta gestión del sistema eléctrico, tanto es así que la propia normativa española regula [7] la obligatoriedad de que los titulares de parques eólicos suministren las previsiones de producción cuando la potencia instalada supera los 0.5 MW [8]. En el caso particular de las Islas Canarias, los sistemas eléctricos insulares con redes pequeñas y débiles no interconectadas en su mayoría originan una importante restricción al uso de energías renovables, no obstante, existen grandes expectativas en el empleo de sistemas de almacenamiento energético y de gestión de demanda, los cuales deben inexorablemente apoyarse sobre técnicas de estimación del recurso eólico a corto plazo para la gestión continua del sistema.

De otra parte, las estimaciones a largo plazo están íntimamente relacionadas con aspectos económicos. Un proyecto de estas características suele tener un periodo de explotación de hasta 25 años, estando su rentabilidad muy condicionada con las circunstancias locales del recurso eólico. Comúnmente antes de la instalación de un parque eólico se suele realizar una campaña de medición meteorológica que en raras ocasiones supera 2 años, discerniendo aspectos fundamentales del diseño como la velocidad y dirección del viento o la intensidad de turbulencia del emplazamiento. Sin embargo, autores como Hiester et al. [9] concluyen que en la propia estimación de la velocidad media del viento la incertidumbre no se reduce a grados aceptables cuando la duración de la campaña de medida es inferior a 10 años, e incluso hay quienes apuntan a un plazo comprendido entre 20 o 30 años [10,11]. Por todo ello, si bien lo ideal sería disponer de series temporales de datos meteorológicos para periodos comprendidos entre 10 y 30 años, este hito es inalcanzable puesto que supondría el aplazamiento de la decisión de inversión durante un largo periodo de tiempo [11]. Este aspecto adquiere mayor importancia hoy día donde a pesar de existir regímenes retributivos específicos como los marcados para los parques eólicos en modalidad de vertido a red con el Real Decreto 413/2014 [8], éste se supedita a que los niveles de producción sean superiores a los marcados por el Ministerio en cada semiperiodo regulatorio y sistema eléctrico.

En respuesta a este problema se han desarrollado distintos métodos de predicción meteorológica a largo plazo que pueden ser clasificados en función de los procedimientos empleados en modelos de predicción física, técnicas de dinámica de fluido computacional CFD o modelos estadísticos de extrapolación temporal MCP.

Los modelos de predicción física toman como punto de partida un mapa eólico generado con un sistema NWP (Numerical Weather Prediction) el cual se alimenta de datos de reanálisis¹,

¹ Bases de datos de reanálisis: Están compuestas por una red tridimensional de datos meteorológicos históricos generados por el NCEP (US National Center for Environmental Prediction) y el NCAR (National Center for Atmospheric Research), el cual integra las condiciones meteorológicas a macroescala recogidos durante un periodo de 15 años con intervalos de cada 6 horas a distintas alturas.

radiosondeos, estaciones de superficie y datos geofísicos. Cada una de las fuentes de datos comentadas presenta deficiencias en relación con el grado de detalle, de ahí se desprende la necesidad del sistema NWP, el cual simula los fenómenos físicos fundamentales que rigen el comportamiento de la atmósfera, principalmente el principio de conservación de la masa y las leyes generales de la dinámica de fluidos, realizando simulaciones sucesivas desde un mallado amplio de 30 km hasta uno menor de 1.2 km, de forma que los resultados de este proceso iterativo son cada vez más fiables. Posteriormente estos resultados son refinados a Microescala, donde se añaden las influencias locales de la topografía de la región e incluso la rugosidad² de la superficie a una resolución que puede llegar hasta los 20 metros. Como resultado de este proceso complejo, partiendo de datos históricos se consigue predecir las condiciones medias del recurso eólico local a largo plazo, estimaciones que posteriormente deben ser corregidas con los datos obtenidos a partir de la estación anemométrica ubicada en el emplazamiento.

Por su parte, las técnicas de dinámica de fluido computacional CFD están ganando cada vez mayor importancia principalmente en terrenos complejos donde los modelos de predicción física presentan limitaciones [12]. Estos modelos simulan mediante métodos numéricos el comportamiento real de los fluidos en su interacción con la superficie, siendo posible obtener mapas de intensidad de turbulencia ambiental y de inclinación de la velocidad del viento en el emplazamiento, los cuales permiten una mejor comprensión de los efectos locales cuando se añaden a éstos los datos obtenidos en la fase de medición meteorológica. El gran avance propuesto con este método radica en que no solo se tienen en cuenta los efectos locales de un punto determinado (coordenada de ubicación de la torre meteorológica), sino que se hace extensivo en un mallado de hasta 2 x 2 km.

Ya por último, los modelos estadísticos de extrapolación temporal, también conocidos como MCP (Medir – Correlacionar – Predecir), toman como datos de partida al menos dos series temporales de datos meteorológicos [13,14]: estación candidata, que constituyen los datos medidos directamente en el emplazamiento durante 1 o 2 años, y observatorio climatológico de referencia, que incorporan datos medidos en una estación meteorológica ajena con un periodo de medida a largo plazo comprendido entre 15 y 20 años. Adicionalmente, este método requiere que parte de las series temporales sean coincidentes en un periodo de tiempo [13]. Tras desarrollar una correlación entre ambas, se consigue extrapolar los datos del observatorio de referencia a la estación candidata, pudiéndose ignorar influencias locales como las derivadas de la topografía de las regiones, siendo ésta una de las ventajas de su empleo tal como argumenta Khadem et al. [15].

Una variante del enfoque anterior es el uso de métodos Machine Learning basados en la inteligencia artificial y el empleo de técnicas estadísticas avanzadas con las cuales se entrena a un algoritmo para que ante una entrada determinada genere una salida previsible lo más próxima a la realidad. Tal como señala Foley et al. [5], la ventaja de este método radica en la estructura paralela y en el

² Rugosidad: Este fenómeno se evalúa mediante el parámetro conocido como longitud de rugosidad (metros), cuyo sentido físico se corresponde con la altura desde el suelo a la que la velocidad media del viento es cero y que depende de aspectos como las condiciones de la superficie terrestre, la existencia de obstáculos tales como la vegetación y los edificios o la porosidad del suelo.

aprendizaje, pudiéndose emplear más de dos estaciones de referencia en el análisis o incluso otros datos que puedan influir en la decisión final. En este ámbito autores como Carta et al. [16] han demostrado que las estimaciones empleando métodos Machine Learning, en este caso redes neuronales artificiales (Artificial Neural Network – ANN), arrojan errores menores con respecto a las mediciones reales que el empleo de dos algoritmos estándar de cálculo MCP.

Independientemente del modelo que se utilice en el proceso de estimación, el objetivo último que se pretende conseguir es que la correlación entre el resultado obtenido y el dato real medido sea la mayor posible. En los últimos años ha existido un interés creciente por el empleo de técnicas Machine Learning para el estudio del potencial eólico, pasando entre otras, por el uso de técnicas ANN [11,17-20], redes Bayesianas [16], árboles de decisión [21], máquinas de vector soporte (Support Vector Machines – SVM) [4,22,23] y modelos híbridos [24,25].

De todas las técnicas Machine Learning existentes, no existían al inicio de la tesis referencias en artículos divulgados que dieran a conocer cuáles eran las que mayor grado de predictibilidad manifestaban en la estimación de la potencia eólica. Asimismo, la estimación de **potencia juega un papel secundario, sólo calculándose de modo indirecto tras la aplicación de estas técnicas para estimar la velocidad del viento [16,17,26].** Asimismo, los modelos de estimación (tanto en corto como a largo plazo) dan todo el peso de la predicción a la variable de velocidad del viento [27-29], y en los últimos años a la dirección [17,24,30,31], **no considerándose significativas otras variables tales como la densidad del aire o incluso otras características que podrían tener un impacto directo como la influencia que tiene sobre la estimación el tipo de control del aerogenerador simulado (pitch – regulated o stall – regulated).**

Por otra parte, en otros estudios comparativos de métodos de estimación, tanto para el recurso eólico como para la predicción de potencia, autores han optado por el análisis a través del error detectado entre el resultado del modelo y el dato real medido [11,18,28,32-36]. No obstante, una comparación fundamentada y robusta, más acorde con el método científico, debería conseguirse con el uso de **métodos de significación estadística** que consideran la distribución de probabilidad de los resultados obtenidos en lugar de sólo unos valores puntuales. Este procedimiento confronta una hipótesis nula, en la que se supone que las técnicas de evaluación del recurso eólico son igual de efectivas, con una hipótesis alternativa donde se asume distinta efectividad, decidiéndose finalmente por una u otra en función de la compatibilidad estadística de los datos con la distribución de probabilidad subyacente en la hipótesis nula. En todo caso, la comparación final debería complementar el análisis de efectividad realizado por estas técnicas, con otros aspectos como la complejidad del modelo o los requisitos computacionales necesarios.

Ya para concluir conviene mencionar que a la vista del potencial de mercado existente en el estudio del potencial eólico, muchas empresas han diseñado paquetes informáticos tales como WAsP [37], WindPro [38], OpenWind [39] o WindSim [40]. Sin embargo, independientemente de los modelos utilizados en estos programas, **ha sido una práctica muy extendida el uso de una única estación de referencia en el proceso de estimación e incluso cuando se insertan varias estaciones de referencia estos sistemas tienden a tratarlas de manera aislada.** Autores como Carta et al. [16] han

demostrado que independientemente del grado de correlación existente entre varias estaciones de referencia, la fidelidad del resultado mejora notoriamente con respecto a cuando sólo se utiliza una serie temporal de datos meteorológicos a largo plazo. Ésta es otra de las ventajas fundamentales existentes en el empleo de métodos Machine Learning para dar solución a este tipo de problemas.

1.2. Objetivos de la tesis

Como se desprende de la introducción, muchos artículos y libros de gran impacto en materia de evaluación del recurso eólico [9,12,41,42] concluyen que las técnicas Machine Learning son una alternativa potente para la estimación del potencial eólico tanto a corto como a largo plazo, demostrándose incluso una mayor precisión de los resultados obtenidos en comparación con otras técnicas clásicas tales como los algoritmos de regresión lineal MCP en el caso del estudio a largo plazo. A pesar de lo anterior, no se disponía al inicio de la tesis doctoral de ninguna referencia donde se divulgaran resultados de estudios comparativos rigurosos entre las distintas técnicas Machine Learning existentes aplicadas a la estimación de la potencia eólica. Por todo ello, los grandes objetivos de este trabajo son los siguientes:

1. **Se realiza una comparación entre tres técnicas fundamentales, ANN** (que tal como se evaluará es la opción más utilizada para este tipo de aplicaciones en la industria eólica) [11,13,17,43-45], **SVR y RF** (las cuales, según el estudio realizado, son las que mejor representan el estado del arte en campos del conocimiento afines a éste). Además, este estudio se desarrolla no sólo teniendo en cuenta la mejora de la fiabilidad de los modelos, sino otros aspectos de diversa índole como la complejidad computacional o los tiempos de ejecución de los modelos.
2. En contra de la metodología habitual, **los modelos desarrollados en esta tesis se formulan con el objetivo principal de estimar la potencia eólica** y no la velocidad del viento. En este sentido, se estudian diferentes formas funcionales hasta dar con aquella que provea los mejores resultados posibles en cualquiera de sus formas, ya sea a través de las densidades de potencia eólica (Wind Power Densities – WPD) o la potencia del parque eólico (Wind Turbine Power Output – WTPO).
3. Se evalúa **la influencia de un conjunto de variables de características meteorológicas diferentes a las comúnmente utilizadas para este tipo de estudios** (velocidad y dirección del viento), en concreto la temperatura ambiente, la presión atmosférica, la humedad relativa o, en su forma agregada, la densidad del aire, variable que a pesar de considerarse en la fórmula teórica de la potencia aportada por un aerogenerador, suele ser eliminada para este tipo de estudios justificándose por su baja influencia en la estimación de potencia.
4. Por otra parte, en el análisis de los distintos modelos formulados para la estimación de la potencia eólica (WTPO), **se tendrá en cuenta la forma funcional de las expresiones matemáticas que representan las dos principales familias de sistemas de control** existentes en el mercado actual, en concreto, **el sistema pitch – regulated y stall – regulated**.

5. Con el objetivo de dar el mayor rigor posible al estudio desarrollado, para el análisis de la influencia de las variables meteorológicas ensayadas **se utilizará una técnica Feature Selection con la que se buscará de forma automática y exhaustiva aquellas variables que aportan mayor significado a los modelos desarrollados.**
6. Asimismo, **la comparativa entre los diferentes modelos de estimación de potencia evaluados, se realizará utilizando un método de significación estadística**, en concreto un test pareado de permutación no paramétrico el cual será descrito en este documento más adelante.

1.3. Metodología de la tesis

Con el propósito de conseguir los objetivos marcados en esta tesis doctoral se ha establecido una metodología basada en las siguientes fases:

1. Recopilación y análisis de la bibliografía relacionada con el tema objeto a estudio.
2. Estudio del estado del arte en técnicas Machine Learning para el análisis del recurso eólico. Como resultado de esta investigación se seleccionaron de forma motivada aquellas que presentan mayor interés para la estimación de la potencia eólica.
3. Selección de las series temporales de datos de meteorológicos y técnicos de las estaciones meteorológicas y parque eólicos que sirven de referencia para el desarrollo de los estudios.
4. Implementación de las rutinas necesarias para generar cada uno de los modelos seleccionados en la fase 2.
5. Diseño de los diferentes modelos candidatos para la estimación de potencia, incluyendo su forma funcional y sus variables, así como su entrenamiento con los datos disponibles utilizando los algoritmos de aprendizaje de Machine Learning seleccionados.
6. Comparación de los resultados obtenidos en cada uno de los modelos analizados ya sea con métricas estándar como con un test estadístico no paramétrico de permutación pareado.
7. Presentación de las conclusiones del estudio.
8. Exposición y divulgación de los resultados de investigación en revistas indexadas y congresos científicos de reconocido prestigio internacional.
9. Identificación de líneas futuras de investigación.

1.4. Estructura de la tesis doctoral

La tesis doctoral ha sido estructurada en **5 Capítulos** acorde que los objetivos y la metodología descrita, estos son:

CAPITULO	DESCRIPCIÓN
Capítulo 1: Introducción general	Se argumentan las razones que motivan el estudio de la influencia de parámetros meteorológicos y funcionales en la estimación de la potencia eólica mediante técnicas de Machine Learning.
Capítulo 2: Introducción al estudio del recurso eólico y conceptos básicos.	Se realiza una introducción al estudio del recurso eólico presentando una serie de conceptos básicos necesarios para la comprensión de los trabajos que se realizan.
Capítulo 3: Técnicas de Machine Learning para el estudio a largo plazo.	Con este análisis se concretan las técnicas de Machine Learning que presentan mayor interés para el análisis desarrollado en este trabajo. En concreto, se han seleccionado las técnicas SVR, RF y ANN las cuales son aptas para la estimación del recurso eólico a largo plazo con múltiples estaciones de referencia y variables.
Capítulo 4: Nuevos métodos de estimación de la potencia eólica mediante técnicas de Machine Learning.	<p>Este apartado centra el núcleo de la tesis, habiéndose estructurado en tres apartados fundamentales:</p> <p>4.1 Datos de partida y evaluación del recurso eólico en Canarias: Se presentan los datos de partida que serán utilizados para los análisis desarrollados en la tesis doctoral. También se describen las condiciones singulares del recurso eólico de Canarias con vistas a entender los resultados esperables en las simulaciones.</p> <p>4.2 Análisis comparativo entre varios modelos MCP que usando la técnica SVR estiman la densidad de potencia: Usando los datos medidos en las estaciones anemométricas expuestas en el Apartado 4.1 se evalúa la técnica SVR como opción en la estimación de las densidades de potencia a largo plazo. En este análisis se estima paralelamente la forma funcional en la que las variables de partida se introducen en el modelo.</p> <p>4.3 Evaluación de la fiabilidad de cinco modelos MCP para estimar las WTPO a largo plazo usando tres técnicas de Machine Learning: En esta ocasión las tres técnicas seleccionadas se emplean para estimar las WTPOs a largo plazo partiendo de las conclusiones manifestadas en el apartado 4.2 y los mismos datos que para el estudio anterior</p>

CAPITULO	DESCRIPCIÓN
Capítulo 5: Conclusiones y líneas futuras de investigación.	Se plantean las conclusiones a las que se llega mediante el desarrollo de la presente tesis doctoral y las líneas futuras de investigación que se derivan de la misma.
Referencias bibliográficas.	Se relacionan las referencias bibliográficas utilizadas en el trabajo.
Anexos. Publicaciones, congresos y valoración externa del trabajo realizado en la tesis	En este apartado se adjuntan los artículos de revistas indexadas y proceedings publicados fruto de los estudios desarrollados en la tesis.

Introducción al estudio del recurso eólico

2.1. Introducción

En este capítulo se realiza una introducción general al estudio del recurso eólico, la caracterización del potencial energético y la estimación de la potencia eólica con el propósito de establecer una serie de conceptos necesarios necesarios para su entendimiento. Asimismo, se esbozan las características generales de un modelo MCP y los requisitos necesarios para su ejecución, poniendo el acento sobre el estudio con más de una estación de referencia.

2.2. Características generales del recurso eólico

La atmósfera se divide en una serie de capas horizontales basándose dicha clasificación en la diferencia de temperatura existente ente capas, no obstante desde el punto de vista de la energía eólica sólo interesa el viento que se produce en la parte más baja de la troposfera a unos pocos centenares de metros.

En términos generales el viento es consecuencia directa de la radiación solar que recibe la tierra, la cual produce diferencias de temperatura por zonas y derivan en **gradientes horizontales de presión**.

Para compensar estos desequilibrios se producen desplazamientos de aire desde las zonas de alta presión hasta las de baja tensión, siendo los vientos más fuertes cuanto mayor sea el gradiente de presión. La fuerza de presión por unidad de masa puede expresarse mediante la Ecuación 2.1 [46,47].

$$-\frac{1}{\rho} = \frac{\partial p}{\partial n} \quad (2.1)$$

siendo ρ la densidad del aire y $\partial p / \partial n$ el gradiente horizontal de presión. Como consecuencia, cuanto más juntas se encuentren las isobaras mayor será la viento.

Estos gradientes de presión causados por las diferencias de temperatura también se producen en altura ya que comúnmente cuanto mayor es la altura, menor es la temperatura por aspectos relacionados con la transparencia del aire que fomenta el calentamiento terrestre [46,48]. En este sentido puede definirse tres estados de **estabilidad atmosférica**, estos son:

1. **Atmósfera neutra:** Este estado atmosférico se produce cuando una burbuja de aire que se eleva de forma adiabática se encuentra a la misma temperatura que el aire en la capa superior, por lo que las diferencias de densidad entre el aire que se encontraba en esa capa y el aire que ha subido es nula y por tanto no se producirán nuevos movimientos.
2. **Atmósfera estable:** Por el contrario si la burbuja de aire que se eleva de forma adiabática se encuentra a menor temperatura que el aire existente en la capa superior, esta burbuja tenderá a descender nuevamente a su posición inicial. Bajo este fenómeno se puede apreciar una clara estratificación en capas horizontales ya que no se producen mezclas entre ellas.
3. **Atmósfera inestable:** Es justo el fenómeno contrario en el que el aire que se eleva esta a mayor temperatura que el existente en una capa superior por lo que la burbuja continuaría subiendo a otras capas y el hueco liberado por este es ocupado por aire más denso. Este proceso de mezcla motiva mayores velocidades del viento en superficie que cuando la atmósfera es estable.

Este aspecto es de gran importancia cuando se analiza la variación diaria del recurso eólico en superficie. En las horas nocturnas la atmósfera tiende a ser estable y por tanto no existen grandes transferencias de cantidad de movimiento entre las capas inferiores y superiores, sin embargo en las horas de sol existe mayor potencial para que se produzcan atmósferas inestables que generan vientos en superficie. De la misma forma cuando la atmósfera es inestable, el aire tiene menos resiliencia para subir y superar obstáculos con facilidad, mientras que con atmósferas estables las capas superiores ejercen un efecto tapón que obliga al aire a buscar caminos diferentes sorteando obstáculos.

En la Figura 1 puede observarse la media horaria de datos de velocidad del viento recogidos durante un periodo equivalente a dos años en la región Este de la isla de Lanzarote a 40 metros de altura.

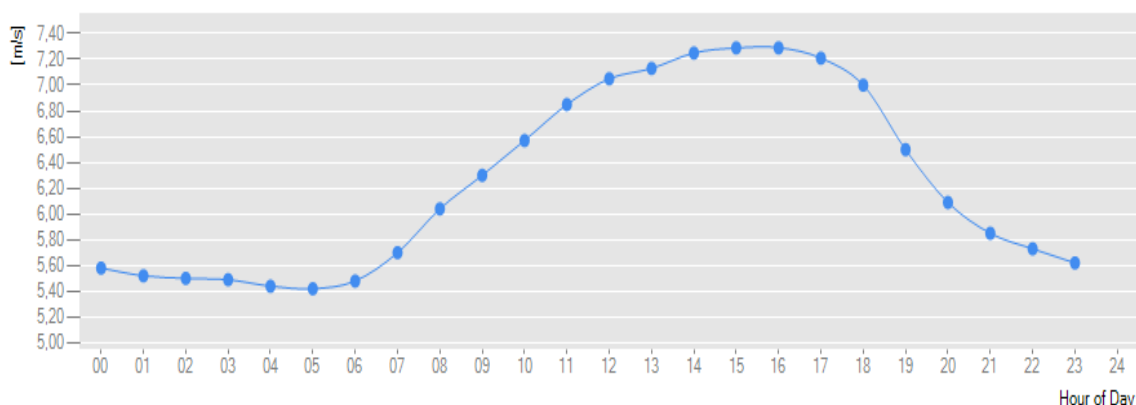


Figura 1 Media horaria de datos de velocidad del viento a 40 metros [Lanzarote]

Así pues se observa como en las horas de puesta de sol la velocidad del viento tiende a disminuir mientras que en el resto del día el viento puede llegar a aumentar hasta 1.5 m/s en valor promedio.

En lo que respecta a las variaciones de la velocidad del viento en periodos a largo plazo por este fenómeno, autores como Palutikoff, Guo y Halliday [49] han concluido que si bien las variaciones interanuales suelen ser escasas, éstas se producen según los datos históricos recabados en diferentes localizaciones a nivel mundial. En este sentido se ha debatido cuáles podrían ser las claves de ello señalándose determinados aspectos como el cambio climático o las consecuencias derivadas por la actividad humana. Estos cambios como consecuencia de las diferencias térmicas también se han producido de un año a otro, según apuntan estos autores, por fenómenos naturales como los ciclones e incluso las erupciones volcánicas.

A los efectos del gradiente de presión y la diferencia de temperatura hay que añadir otros fenómenos como las **fuerzas de Coriolis**, siendo ésta una fuerza ficticia que aparece como consecuencia de que las masas de aire se muevan sobre la tierra cuando ésta se encuentra a su vez en rotación [50,51]. Como consecuencia, si bien bajo condiciones naturales el viento es perpendicular a las isobaras, en algunos casos sobre todo en la atmósfera libre lejos de las capas superficiales, el viento sopla en paralelo a las isobaras. Este fenómeno tiene a atenuarse cuanto menor es la altura y por tanto menor es la velocidad del viento, pero en tanto se reduce la fuerza de Coriolis aparecen otros efectos de gran importancia como la fricción con el suelo.

Los efectos de fricción con el suelo dependen muy directamente de las **características locales del emplazamiento a estudio**, existiendo efectos locales debidos a fenómenos térmicos y efectos locales ocasionados por aspectos orográficos.

En relación con los efectos locales de origen térmico éstos se producen cuando existe un cambio de paisaje drástico, como entre mar y tierra, ciudad y campo o incluso entre ladera y valle. A priori no se considera que tengan el suficiente contenido energético como para motivar la instalación de un parque eólico de alta potencia, sin embargo si es cierto que en ubicaciones de gran potencial estos fenómenos tienden a maximizar la obtención de energía en determinados periodos horarios, aspecto que se tiene en cuenta en el Micrositting de las instalaciones. Se citan a continuación algunos ejemplos muy comunes en las Islas Canarias:

- **Brisas:** En el caso de regiones costeras en las horas de sol la tierra se calienta más rápidamente que el mar y por tanto el aire situado sobre tierra tiende a subir ocasionando formaciones de viento que soplan fuerte hacia tierra. Por su parte por la noche el flujo se invierte ya que el aire sobre el mar es más cálido.
- **Vientos en valles entre montañas:** En este caso los flujos del viento son distintos con dependencia de la hora del día. Durante las primeras horas de la mañana las laderas del valle tienden a calentarse y comienzan a aparecer en ellas flujos ascendentes de aire los cuales son máximos al mediodía. Por la tarde ya no se producen vientos en las laderas y el viento sólo circula valle arriba hasta que tras la puesta del sol se vuelven a producir vientos de ladera descendentes.

En lo relativo a los rasgos orográficos, ciertas formaciones naturales y artificiales generan efectos que según su posición y características pueden suponer un aumento del recurso eólico en un emplazamiento, como por ejemplo en las zonas altas de las cadenas montañosas donde a los efectos del aumento de velocidad del viento por altura se suman las aceleraciones causadas por la orografía.

También pueden ser zonas de alto potencial las ubicaciones que se encuentran en valles entre montañas, donde dependiendo de la dirección del viento se puede producir una canalización del aire como consecuencia de que los cambios energéticos necesarios para acelerar el flujo del aire alrededor de un obstáculo son menores que los requeridos para superar dicho obstáculo por altura (energía potencial), generando un efecto túnel entre cadenas montañosas (Figura 2).

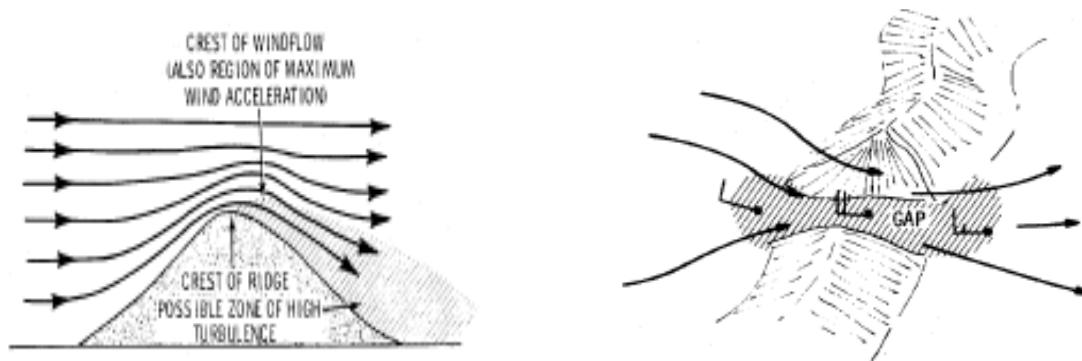


Figura 2 Zonas de interés para ejecución de parques eólicos en presencia de obstáculos [46]

En cualquier caso no todas las cadenas montañosas son aptas para la ejecución de parques eólicos. A las dificultades técnicas, ambientales y normativas hay que añadir el problema de las turbulencias, y es que en zonas demasiado escarpadas o accidentadas se suelen producir vórtices donde el viento cambia de forma rápida sus condiciones de presión y velocidad, lo que afecta de manera muy negativa a los aerogeneradores, anulando cualquier ventaja existente de mayor velocidad de viento.

Entre los estudios desarrollados a nivel mundial suele ser una práctica común la recomendación de no instalar aerogeneradores a una distancia inferior a 10 veces la altura del obstáculo en la dirección predominante del viento para evitar cualquier interferencia de ésta [52].

A efectos prácticos para añadir en los análisis del potencial eólico las condiciones derivadas de la complejidad orográfica ha sido establecido un parámetro conocido como **longitud de rugosidad** (Ecuación 2.2), cuyo sentido físico se corresponde con la altura a la que la velocidad media del viento es cero y que depende de aspectos como la dispersión de los elementos de rugosidad y sus características superficiales.

$$Z_0 = 0.5 \cdot \frac{h_S}{A_H} \quad (2.2)$$

Donde h es la altura y S es la sección de cara al viento de los elementos de rugosidad y A_H es el área horizontal media del obstáculo. En términos generales el valor Z_0 tiende a ser pequeño cuando A_H es mucho más grande que S , por ejemplo en regiones libres de obstáculos de gran altura, mientras que cuando los dos parámetros anteriores son del mismo orden Z_0 tiende a aumentar, como ocurre en ciudades donde existen obstáculos de gran altura y muy cercanos entre sí. En este sentido se han tipificado unas tablas de valores en función del uso o el medio propiamente dicho (Tabla 1).

VALORES TIPO DE LONGITUD DE RUGOSIDAD	
Descripción del tipo de terreno	Longitud de rugosidad Z_0
Superficies muy lisas tales como hielo	0.00001 m
Mares en calma	0.0002 m
Mar revuelto	0.0005 m
Superficies nevadas	0.003 m
Césped de hierba	0.008 m
Pastos	0.01 m
Campos cultivables en barbecho	0.03 m
Campos de cosecha	0.05 m
Zonas con pocos árboles	0.1 m
Zonas arboladas, setos o edificios dispersos	0.25 m
Bosques o zonas muy arboladas	0.5 m
Suburbios	1.5 m
Centro de ciudades con presencia de edificios altos	3.0 m

Tabla 1 Valores tipo de longitud de rugosidad Z_0 en metros

Suele ser una práctica habitual el empleo de estas tablas las cuales se equiparan con los usos del suelo a través de herramientas de información geográfica (GIS), adoptando un valor estimativo de longitud de rugosidad por emplazamiento. Este dato es además muy útil para la ejecución de extrapolaciones de datos en altura tal como se describirá en apartados posteriores.

La suma de la longitud de rugosidad y la altura media de los obstáculos en las inmediaciones definen la capa límite donde la velocidad media del viento tiende a ser nula. Además dependiendo de su posición existen regiones donde por la complejidad del terreno se crean depresiones que invierten el sentido del perfil de viento como se expone en la Figura 3.

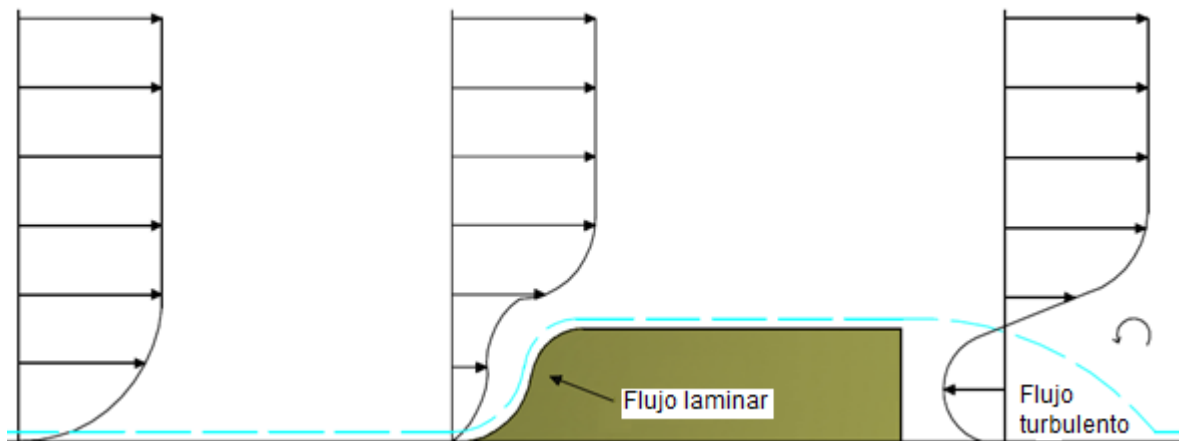


Figura 3 Efecto de un obstáculo sobre el perfil vertical de la velocidad del viento

Autores como Tieleman [53] sugieren que altas longitudes de rugosidad son indicativas de la existencia de flujos turbulentos presentándose frecuentemente esta situación en terrenos de orografía compleja, por lo que el parámetro de longitud de rugosidad se toma como valor cuantitativo de la complejidad topográfica de la ubicación.

Otros autores como Schlez [54] han determinado en sus investigaciones que en determinadas localizaciones una distorsión del flujo causada por las circunstancias locales del terreno pueden derivar en una reducción de la intensidad de turbulencia. Como ejemplo señala que las aceleraciones del flujo originadas en lo alto de las colinas generan un cambio del espectro turbulento hacia bajas frecuencias, mientras que en zonas a nivel del suelo, donde el flujo no muestra cambio, éste se mantiene imperturbable.

2.3. Medición del viento

Todos los aspectos tratados en el apartado anterior señalan la existencia de patrones en el recurso eólico que pueden ser considerados a efectos de la estimación del potencial energético de una posición, sin embargo, como se ha podido comprobar existe un cierto comportamiento caótico derivado de las circunstancias locales que difícilmente puede ser evaluado sin el desarrollo de una campaña de medición meteorológica de vientos en las posiciones a estudio. En este apartado se analizan los aspectos fundamentales de la medida del viento.

2.3.1. Parámetros básicos de medida

En la realización de una campaña de medición meteorológica tipo con fines energéticos fundamentalmente suele recabarse información acerca de la velocidad y la dirección del viento. De modo común con dichas variables podría conocerse la producción aproximada de un aerogenerador ubicado en la misma posición donde se sitúan los sensores. Una estimación más certera podría conocerse considerando la variación de la potencia producida en función de la densidad del aire,

para lo cual es necesaria la medida de la temperatura del aire ambiente, la humedad relativa y la presión atmosférica en la misma posición. Se describen a continuación dichas variables con mayor detalle:

1. Velocidad del viento:

Es quizás el dato más importante ya que permite la cuantificación del potencial energético de un emplazamiento. Cuando las torres meteorológicas se instalan con fines energéticos, es conveniente que la altura de medida sean coincidente con la altura de buje del aerogenerador a instalar, sin embargo, esto no es siempre posible por el aumento de los costes y la dificultad de instalación, razón por la que normalmente la altura de medida máxima suele establecerse entre los 40 y los 60 metros. Asimismo, en muchas ocasiones al inicio del proyecto no se conoce el modelo de aerogenerador a instalar y por ende la estación anemométrica difícilmente coincide con la altura de buje.

Para resolver en cierta medida el problema anterior suele ser recurrente la instalación de sensores en al menos otra altura inferior (comúnmente a 10, 20 metros) permitiendo que por métodos matemáticos pueda ser determinada de manera aproximada el perfil de vientos de la zona a estudio. Esta altura inferior también es relevante puesto que determina la velocidad del viento próxima al extremo inferior de la pala, lo cual es interesante para establecer el perfil de cizallamiento del viento con fines energéticos y estructurales.

A día de hoy quizá la altura de medida de velocidad de viento más estándar es 10 metros puesto que dichas medidas parten de las bases recomendadas por la Agencia Estatal de Meteorología si bien dichas recomendaciones no son comúnmente alabadas con fines energéticos. En cierta medida si bien aporta muchos datos interesantes, este tipo de medidas tienen un mayor riesgo de sufrir apantallamientos por edificios o incluso vegetación que las medidas recabadas por estaciones anemométricas diseñadas explícitamente para los estudios de diseño de un parque eólico.

Para medir la velocidad del viento se utilizan los anemómetros los cuales pueden ser clasificados [55,56] en tres categorías básicas, anemómetros rotacionales (cazoletas), anemómetros de presión (Dines) y anemómetros especiales como los LIDAR (anemómetro láser) y los SODAR (anemómetro de efecto Doppler).

De todos ellos quizás los más ampliamente utilizados son los anemómetros de cazoleta por ser de fácil instalación y muy económicos. En estos el viento incide en el lado cóncavo de las cazoletas produciendo un movimiento rotacional que es traducido a una señal eléctrica registrando la velocidad del fluido. Su principal problema se deriva de su principio de funcionamiento, y es que se aceleran muy rápido pero tardan en pararse, lo que comúnmente genera una pequeña sobreestimación. De la misma forma suele plantear problemas por el movimiento que unido a su operación en ambientes agresivos puede generar errores de medida.

Para detectar los problemas derivados de los anemómetros anteriores suele ser usual la instalación de medidas redundantes con anemómetros de otra tecnología como los anemómetros de hélice.



Figura 4 Anemómetros tipo (cazoleta y hélice)

2. Dirección del viento:

También es de importancia definir la dirección predominante del viento puesto que permite definir la posición exacta donde se debería ubicar una turbina en función de los obstáculos que la rodean. Hoy en día este aspecto es tenido en cuenta incluso a efectos de tramitación del parque eólico ya que la normativa [57] restringe las proximidades entre parques eólicos en función de la dirección predominante del viento.

Para determinar la dirección del viento se utilizan veletas las cuales no son más que un dispositivo montado sobre un eje vertical fijo y estable el cual genera una señal eléctrica en función de su posición. Para que las medidas obtenidas sean exactas la veleta debe estar perfectamente orientada hacia el norte en el momento de su instalación, valor que tomará de referencia a la hora de indicar los grados de dirección del viento.

Sabiendo que prácticamente no hay variaciones significativas entre direcciones de viento a diferentes alturas, sólo suele instalarse una veleta en lo alto de la torre, aunque las torres más precisas incorporan este instrumento en cada punto donde se haya medido la velocidad del viento.

3. Temperatura del aire:

La temperatura del aire es un importante descriptor del ambiente de operación en proximidades a la altura de buje. Es también importante porque permite estimar la densidad del aire en el emplazamiento, siendo su cálculo fundamental para determinar la potencia que podría ser generada por una turbina eólica.

Para la medida de la temperatura del viento se suelen instalar termómetros al nivel del suelo (a los tres metros de altura), sin embargo, en algunos casos los termómetros se instalan a diferentes alturas. Tal como señalan autores como, Carta et al. [58] a pesar de que comúnmente para el cálculo de la energía generada por una turbina eólica en una posición se toma un único valor de densidad, se demuestra que ésta varía con la altura de forma semejante a la velocidad, resultando estimaciones más acertadas cuando se dispone de un perfil de densidades con la altura.

4. Humedad relativa y Presión atmosférica:

Este tipo de sensores se han instalado con menor frecuencia en estaciones anemométricas con fines energéticos, si bien actualmente se considera una fuente de información indispensable para conocer las condiciones locales del emplazamiento en lo que a la evolución del recurso eólico se refiere.

Por lo general el higrómetro suele situarse en una caseta próxima a la estación anemométrica junto con el sensor de temperatura inferior de la torre meteorológica (normalmente a 2 – 3 metros). Por su parte, el barómetro se instala a la misma altura en a que se ubica el sensor de velocidad del viento (comúnmente en la posición más próxima a la altura de buje). En términos generales, existen detractores de la instalación de barómetros ya que ciertos estudios indican que es difícil medir con precisión la presión atmosférica debido a que en ambientes muy ventosos se inducen dinámicas de flujo de viento que podrían afectar al instrumento de medida [55]. Aun considerando este aspecto, no podría estimarse la densidad del aire sin datos de presión tal como se justifica con la expresión citada en la Ecuación 2.3 la cual es considerada como la versión más simplista en la estimación de la densidad del aire.

$$\rho = \frac{P}{R \cdot T} \quad (2.3)$$

donde:

ρ = Densidad del aire (kg m^{-3}).

P= Presión atmosférica (Pa).

T= Temperatura ($^{\circ}\text{C}+273$).

R= Constante específica del aire ($287 \text{ J kg}^{-1}\text{K}^{-1}$).

Se presenta en el Apartado 2.7, una versión de la formulación de la densidad del aire más exacta y apta para un rango mayor temperaturas y presiones atmosféricas.

Básicamente con temperaturas más bajas la densidad del aire es mayor, lo que se traduce en una mayor fluidez de las moléculas del aire sobre la pala y, por tanto, en un aumento de la potencia obtenida por el aerogenerador.

Como valor aproximado de la presión atmosférica podría tomarse como referencia los datos estimados de este parámetro a través de un modelo WRF. No obstante, este tipo de estimaciones también presenta claras deficiencias dado que provienen de modelos matemáticos simplificados con vista a la reducción de las cargas computacionales.

2.3.2. Registro de datos y dispositivos de almacenamiento

Las guías técnicas sobre campañas de medición meteorológica [55] establecen que los parámetros anteriormente citados sean recabados con resoluciones inferiores a 10 segundos y grabados como medias, desviaciones estándar, valores máximos y mínimos en una base de datos donde cada valor sea relacionado con la fecha y hora en la que dicha medida hubiera sido recabada. Este procedimiento permite que a posteriori los datos puedan ser traducidos a frecuencias de entre 10 minutos o 1 hora tal como establece los estándares internacionales para estudios relacionados con

finés energéticos o a intervalos minutales si por el contrario dichos datos pretendieran ser utilizados para predicciones a corto plazo o estudios relacionados con la estabilidad del sistema eléctrico, evitando que la resolución no sea nunca un problema en el análisis.

Adicionalmente con la desviación estándar es posible determinar aspectos relevantes del diseño como la intensidad de turbulencia del emplazamiento. Para que la desviación estándar sea adecuada debe emplearse intervalos de medida de entre uno o dos segundos. Estos datos se graban en una memoria RAM hasta que se obtiene la desviación estándar de los 10 minutos, momento en el que se graba el valor resultante y se borra la RAM para volver a procesar la información.

Ya por último en relación con los valores máximos y mínimos se adoptan los mismos procedimientos de cálculo que para la estimación de la desviación estándar. Estos valores presentan información relevante a efectos del análisis del potencial eólico del emplazamiento y las condiciones extremas que podrían producirse en dicha ubicación.

En el año 1981 Hiester et al. [9] realizó una clasificación de los sistemas de adquisición y de almacenamiento de datos en función de su operación, destacándose cuatro grupos fundamentales:

- **Clase I (Sin almacenamiento):** En este caso no se registran los datos y, por tanto, sus aplicaciones sólo son de interés para medidas de corta duración en la que un observador anota manualmente los resultados que se fueran produciendo.
- **Clase II (Almacenamiento selectivo):** En esta situación a pesar de haberse efectuado medidas para un intervalo amplio de tiempo, estos datos se desprecian sólo grabándose la información para un periodo dado como los valores medios diarios.
- **Clase III (Almacenamiento con registros procesados):** Los datos pasan por un microprocesador que genera estadísticas de interés en función de los datos recabados por los sensores. En este caso sí existe un almacenamiento amplio para el análisis de la información.
- **Clase IV (Almacenamiento con registros procesados y brutos):** Estos sistemas son los más completos ya que procesa la información con registros estadísticos y además almacena los datos brutos por lo que haciendo uso de programas de tratamientos de datos pueden obtenerse información más detallada.

Lógicamente hoy día los sistemas más usados son los de clase IV por ser los más flexibles y a los que más partido se saca con fines relacionados con estudios energéticos en las campañas de medición meteorológica. Para la alimentación de estos sistemas se suelen emplear baterías alimentadas o bien por energía eléctrica procedente de la red o por sistemas renovables como placas solares fotovoltaicas debido a su bajo consumo. En lo que al acceso a los datos se refiere, en el estado actual de la técnica se puede recuperar los datos a través de sistemas remotos que conectan con los dispositivos de almacenamiento. Estos equipos suelen instalarse en un armario a pie de la torre meteorológica.

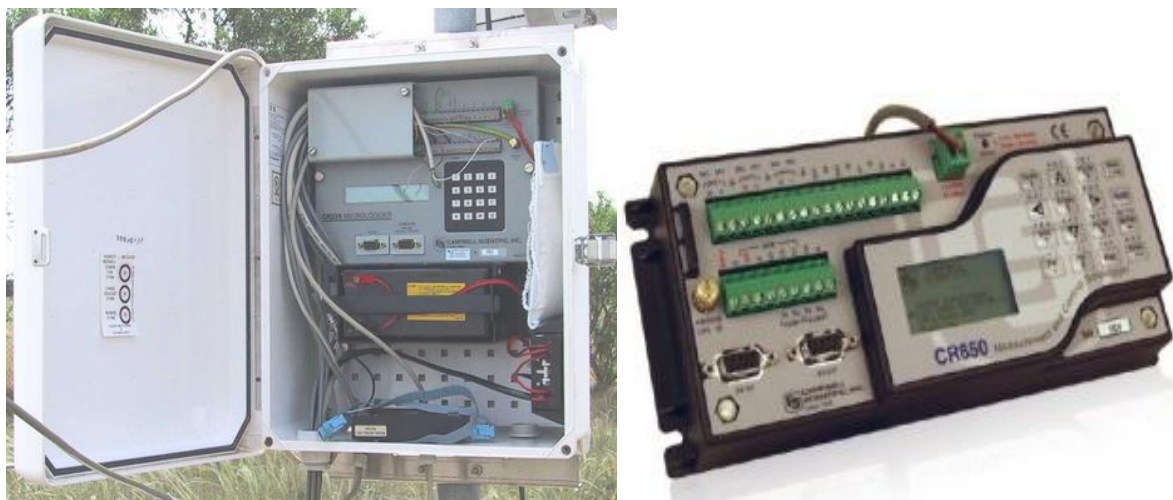


Figura 5 Armario de dispositivo de registro y almacenamiento y Datalogger

En términos generales, las unidades de medida comúnmente empleadas son los m s^{-1} para velocidades del viento, los grados para la dirección del viento, los grados centígrados para temperaturas, valores porcentuales para humedades relativas y hPa para presión atmosférica. En el caso de la dirección del viento, los grados son contados a partir del norte magnético. En algunas ocasiones, principalmente en datos recabados a través de bases de datos de reanálisis, también es posible encontrarse con unidades como km h^{-1} (caso AEMET) o nudos (caso de predicciones proporcionadas por NOAA) en la medida de la velocidad del viento, debiéndose informar de esta configuración en el propio fichero de salida del sistema de registro.

2.3.3. Duración de campaña de medida

Si bien conviene que la toma de datos se realice durante el periodo de tiempo más amplio posible, suele ser habitual la ejecución de campañas de medición meteorológica durante 1 o 2 años ya que en la práctica los principales distribuidores de equipos de medida ofrecen la posibilidad de ejecutar una campaña de medición meteorológica cediendo los equipos en régimen de alquiler hasta la finalización de la misma. Todo ello deriva en que cuanto más tiempo se encuentre instalada la torre meteorológica, mayor será el coste de la instalación. A modo orientativo, llevar a cabo una campaña de medición meteorológica de estas características supondría una inversión de en torno a 45.000 € para medidas a dos alturas (40 y 20 metros), incluyendo los trabajos de generación de informes trimestrales y anuales.

En la elección de los instrumentos de medida suele ser relevante aspectos como el coste, la fiabilidad, la sensibilidad de la medida, la resolución, y la exactitud. Estos parámetros suelen ser necesarios para estimar el error de los modelos matemáticos ejecutados en procesos posteriores de tratamiento de la información. En algunas ocasiones el tipo de sistema de almacenamiento puede limitar la resolución de la medida independientemente de que la resolución pueda ser elevada.

2.4. Representación de los datos de viento en el análisis

De la campaña de medición meteorológica desarrollada se obtiene una cantidad ingente de datos del potencial eólico, siendo necesaria su evaluación mediante métodos estadísticos con el objetivo de evaluar el régimen local del recurso eólico en condiciones normales e identificar anomalías. Normalmente los parámetros estadísticos más utilizados son los que reflejan la posición central y marginal de las variables de interés (media, mediana, y demás cuantiles) y las que reflejan su variabilidad (desviación típica, máximos, mínimos, etc). A estas se unen las distribuciones de rumbo de dirección del viento conocidas como rosas de los vientos y las distribuciones de frecuencia de vientos. En este apartado se realiza un corto repaso de ellas:

2.4.1. Distribuciones temporales

Estos métodos estadísticos son muy útiles puesto que permite el reconocimiento de los valores típicos y las características generales del emplazamiento a estudio. Además, permiten su representación gráfica lo que facilita el entendimiento de los datos y el reconocimiento de patrones de variación:

- **Valores medios:** Este tipo de estadística puede ser desarrollada para distintas frecuencias temporales siendo interesante con el fin de averiguar aspectos como cuáles son los meses en los que se registran mayores velocidades de viento independientemente de que durante un corto espacio de tiempo se hubiera producido una desviación del comportamiento típico estacional. Además en muchos casos es el parámetro más indicativo del potencial eólico de una región en comparación con otra.

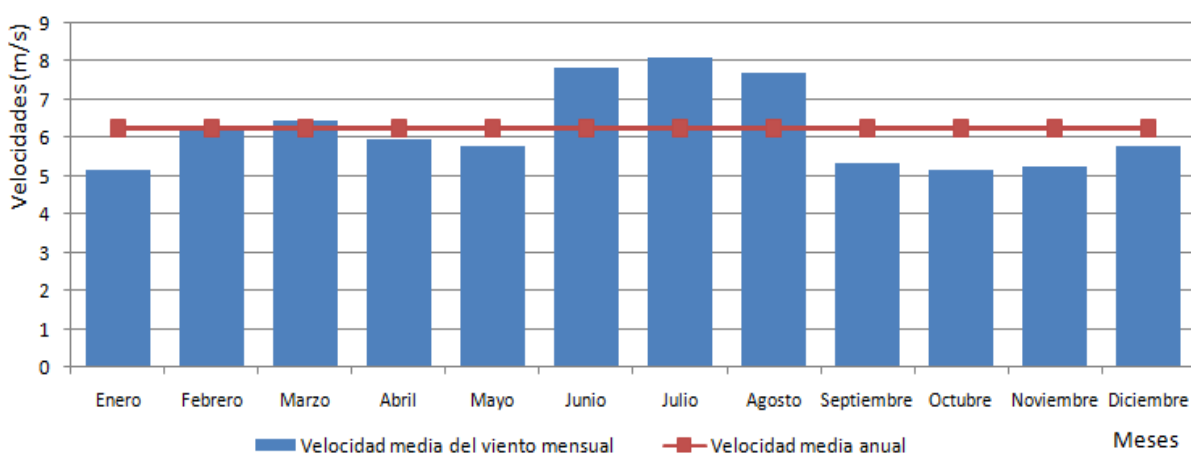


Figura 6 Velocidades medias del viento mensuales y anuales

El valor medio también puede ser interesante para el estudio de las direcciones del viento, sin embargo, su utilidad hoy en día pierde importancia a favor de la distribución de frecuencias de la rosa de los vientos. También es aplicado para variables tales como la temperatura, la presión atmosférica o la humedad relativa.

- **Desviación típica estacional de la velocidad y la dirección del viento:** Por su parte, con la desviación típica se determina el grado de variación de la variable a analizar con respecto a su valor medio, lo que en definitiva se traduce en un indicador de la dispersión o incertidumbre de la variable.

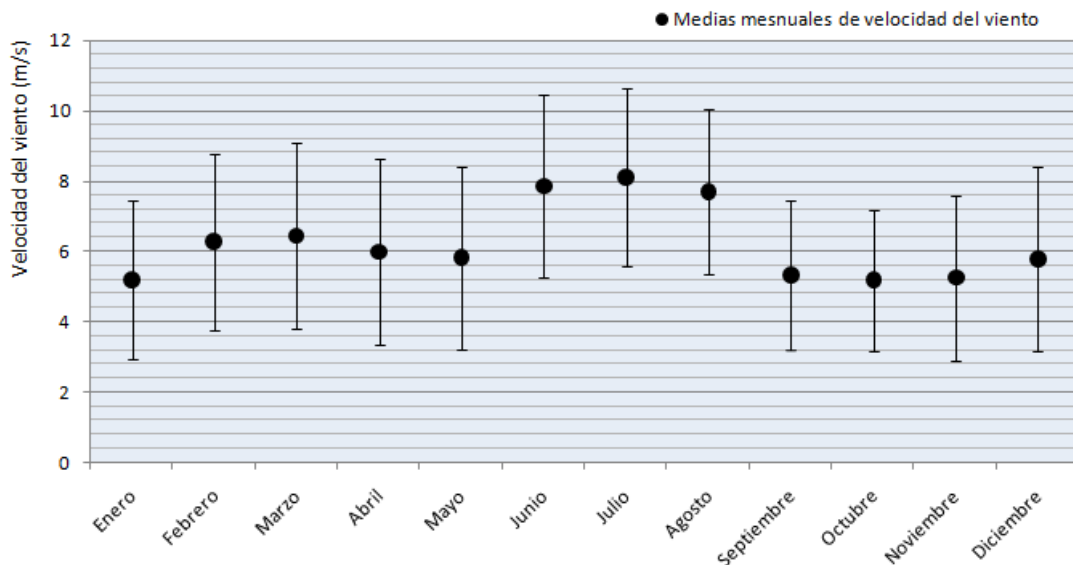


Figura 7 Desviaciones típicas de la velocidad del viento mensual

- **Valores máximos y mínimos de la velocidad del viento:** También son de interés los valores máximos y mínimos de la velocidad del viento ya que permiten el diseño de la instalación frente a condiciones extremas (tiempos máximos de parada por escasos del recurso eólico, esfuerzos máximos soportados en la estructura por acción del viento,...).
- **Intensidad de turbulencia:** La intensidad de la turbulencia describe la rapidez con la que cambia la velocidad del viento en una posición determinada por causas relacionadas con los efectos locales y las condiciones atmosféricas [59]. Para su cálculo es necesario determinar la desviación típica de la velocidad del viento con frecuencias de muestreo de 1 a 2 segundos y dividirla entre la velocidad media del viento en periodos comprendidos entre 10 minutos y una hora.

La importancia de la intensidad de turbulencia radica en que para determinadas ubicaciones donde dicho indicador es alto, se ejercen cargas dinámicas estructurales sobre los aerogeneradores que pueden provocar el colapso si éstos no están preparados para ello. De la misma forma, también se generarían variaciones en la potencia suministrada ya que los sistemas de control actuarían limitando la producción del aerogenerador para garantizar que la vida útil de la máquina no se vea mermada por trabajar en condiciones anormales. Tal como se representa en la Figura 8, los estándares internacionales señalan tres subcategorías IEC de aerogeneradores comerciales en función de la turbulencia, debiéndose elegir para cada posición la opción más adecuada. Para ello la curva de velocidades medias en la ubicación deberá estar por debajo de la curva estándar correspondiente.

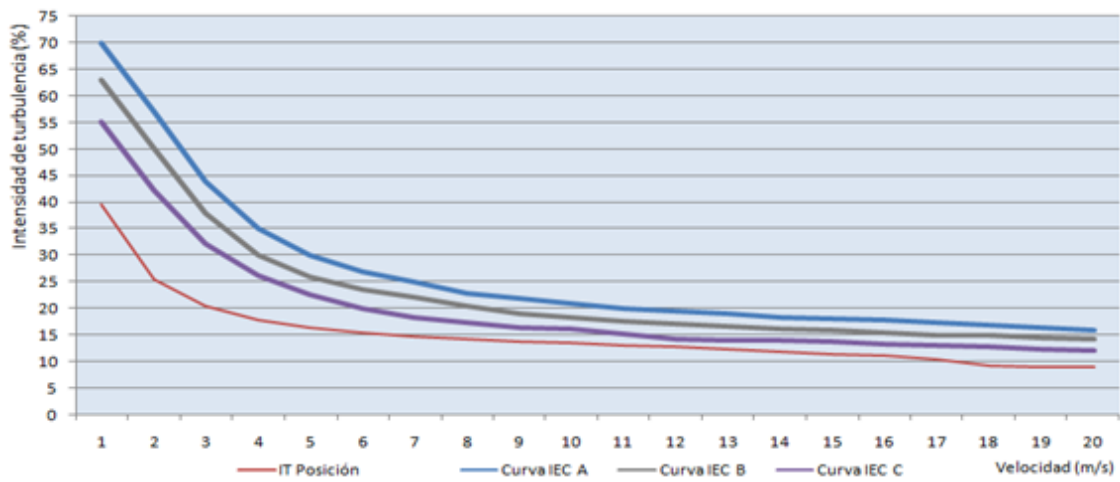


Figura 8 Evaluación de la intensidad de la turbulencia para una posición determinada

2.4.2. Distribuciones de frecuencia

Las técnicas utilizadas para representar las distribuciones de frecuencia suelen ser distintas según se trate de variables velocidad o de variables dirección del viento:

- **Rosa de vientos:** Las características direccionales del viento son de gran importancia para el diseño de una instalación eólica, sobre todo en terrenos con características irregulares y cuando existen agrupaciones de aerogeneradores. Comúnmente, para su representación se suele utilizar una distribución de direcciones del viento llamada “rosa de vientos” en la que se clasifica el conjunto de datos en una serie de sectores de dirección que completan los 360°. Su sentido físico se corresponde con el porcentaje de tiempo en el que el viento proviene de una determinada orientación. El número de sectores en los que se divide la rosa de los vientos depende de la precisión establecida en el análisis, sin embargo, suele ser común utilizar entre 12, 16, 24 o 36 sectores.

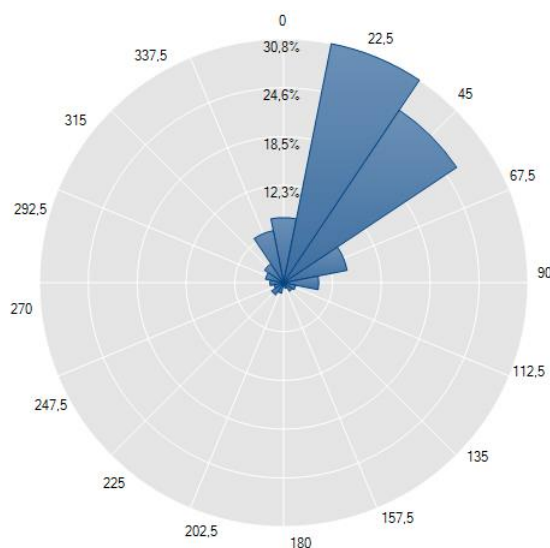


Figura 9 Rosa de los vientos

- **Distribuciones analíticas de velocidad del viento:** Su utilidad se deriva sobre todo de aquéllos análisis en los que se pretenden evaluar el potencial eólico de varios emplazamientos cuando sólo se disponen de datos para una posición determinada, ya que no sólo da información del valor más probable sino que informa de todos los valores posibles en función de la frecuencia.

Actualmente el modelo de distribución más ampliamente utilizado para el estudio del recurso eólico es la distribución de Weibull [60-62] mostrando un comportamiento de precisión aceptable y dependiente de dos parámetros que fijan la forma de la curva, pudiéndose ésta adaptar a la mayoría de los emplazamientos debido a su simplicidad y flexibilidad [63]. La función de densidad de Weibull viene dada por la Ecuación 2.4.

$$f(v, k, A) = \left(\frac{k}{A}\right) * \left(\frac{v}{A}\right)^{k-1} * e^{-\left(\frac{v}{A}\right)^k} \quad (2.4)$$

donde,

K= Parámetro de forma sin dimensión $((\sigma/V)^{-1.086})$.

A= Parámetro de escala en m/s: $\frac{\bar{v}}{\Gamma \cdot (1 + \frac{1}{k})}$ donde Γ es la función Gamma de Euler.

v= Velocidad del viento.

A priori un valor de k entre 2.5 y 3 supone una variación media anual pequeña durante el periodo anual, sin embargo, si su valor se establece entre 1.2 y 1.5 se obtendría una variación estacional de la velocidad del viento elevada. Si el valor de k fuera 2, la distribución de Weibull coincidiría con la distribución de Rayleigh que también ha sido usada para este tipo de estudios puesto que resulta un valor bastante típico en muchas localizaciones a nivel mundial.

Para demostrar el grado de acople de los valores de velocidad media del viento con la distribución de Weibull se adjunta la Figura 10 donde en función de los datos recabados se estima su distribución. Estos datos proceden de una localización ubicada en el Este de Lanzarote.

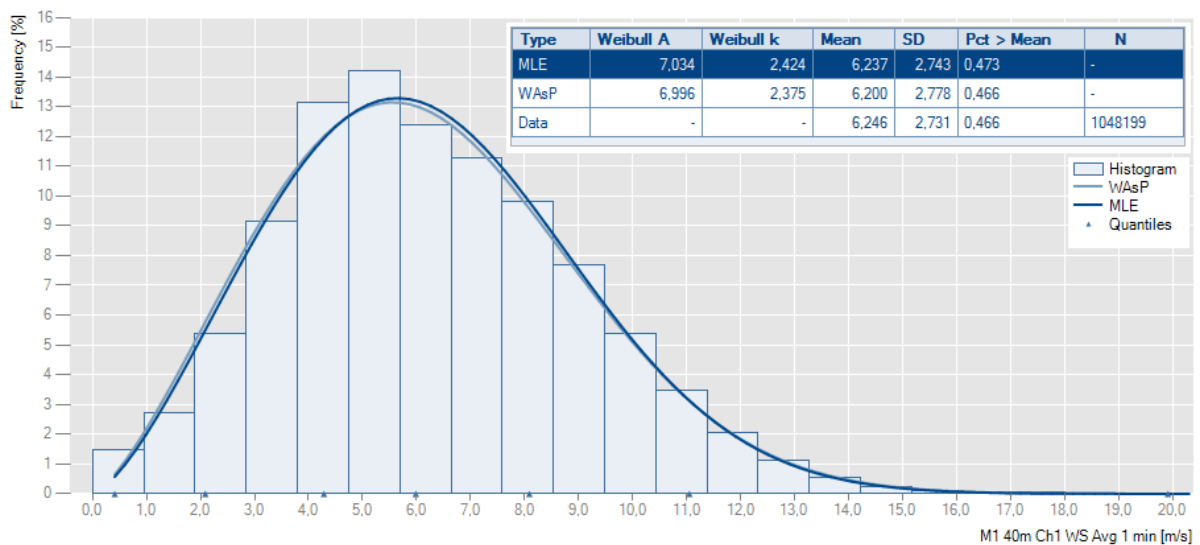


Figura 10 Distribución de velocidades medias del viento [Lanzarote]

También han sido utilizados otros modelos de distribución para la velocidad del viento como la anteriormente mencionada distribución de Rayleigh (R), tanto de forma aislada [64,65] como de forma cooperativa con la distribución de Weibull [66], la distribución Gaussiana inversa (IG) [67], el método de los mínimos cuadrados (LSM) [68], el método de la máxima verosimilitud (MLM), la distribución gamma generalizada a tres parámetros (GG) [69], la función de densidad de probabilidad con parámetro de máxima entropía (ME) o el método del momento (MM).

En el año 2008 Carta et al. [63] desarrollan una revisión de éstos métodos estableciéndose las siguientes conclusiones:

- La distribución IG presenta bajos valores de correlación, lo que contradice las afirmaciones establecidas por otros autores [70,71].
- La distribución de Rayleigh no es adecuada para la representación de una alta gama de regímenes de viento como se sugiere en algunas referencias de literatura científica [46].
- Los métodos LSM y MLM producen un buen ajuste a los histogramas de frecuencia de velocidad del viento, sin embargo, para la estimación de los parámetros característicos es necesario el empleo de métodos numéricos avanzados con la excepción de aquellos casos en los que se puede linealizar con funciones tipo Weibull o Rayleigh.
- Por su parte el método MM no presenta por lo general dificultades en la parametrización pudiendo ser buenos puntos de partida para la estimación de parámetros MLM o LSM.

Lógicamente, si se disponen de series temporales de datos medidos en la posición a estudio y la altura adecuada la solución más acertada es el empleo directo de estos datos. Este aspecto se tratará con mayor detalle en apartados posteriores del presente capítulo.

2.5. Extrapolación de los datos en altura

Se han desarrollado modelos con los cuales los datos de viento de una posición a una altura determinada pueden ser traducidos a diferentes alturas pudiendo completar el perfil de viento característico de la posición a estudio. De todos ellos, quizá los más ampliamente utilizados en la industria son las leyes de distribución logarítmica Ecuación 2.5 y de distribución potencial Ecuación 2.6 [72]:

$$v(z) = v(z_r) \cdot \frac{\ln(z/z_0)}{\ln(z_r/z_0)} \quad (2.5)$$

donde,

z: Altura objetivo.

z_r: Altura de referencia.

z₀: Longitud de rugosidad (m).

$V(z)$: Velocidad media del viento a la altura deseada.

$V(z_r)$: Velocidad media de viento a la altura de referencia.

$$v(Z) = v_x \cdot \left(\frac{Z}{Z_x}\right)^\alpha \quad (2.6)$$

donde,

z : Altura objetivo.

z_x : Altura objetivo.

$v(z)$: Velocidad media del viento a la altura deseada.

v_x : Velocidad media de viento a la altura de referencia.

α : Exponente de variación vertical.

Según los análisis estadísticos de predicción de errores desarrollados por varias universidades de prestigio a nivel mundial [73], no existen diferencias significativas entre ambos métodos para alturas comprendidas entre los 40 y los 100 metros. No obstante, a partir de los 100 metros se suele aconsejar el uso de la ley de distribución potencial al reducirse considerablemente el error. Lógicamente si existen medidas de velocidad del viento a varias alturas, es posible obtener un valor de longitud de rugosidad más exacto, razón por la cual se suelen realizar medidas en al menos dos alturas para una misma posición.

Existen estudios que relacionan el modelo matemático de extrapolación de la altura con el régimen atmosférico [46], concluyéndose que esta expresión es válida únicamente para regímenes atmosféricos neutros. Empíricamente se ha demostrado que este régimen se alcanza cuando las velocidades de viento son superiores a 10 m s^{-1} . Si bien existe la posibilidad de desarrollar un método matemático que corrige el estudio en función de la estabilidad, éste no suele ser empleado debido a su complejidad, puesto que es necesario determinar el gradiente de temperatura y esto exige contar a su vez con medidas a distintas alturas de la temperatura encareciendo los costes de las instalaciones meteorológicas.

Tanto la ley de distribución de potencia como la ley de distribución logarítmica se basan en valores instantáneos y simultáneos para las referencias utilizadas, sin embargo autores como Justus and Mikhail [74] han sugerido que quizá sería más conveniente tratar con un rango amplio de velocidades de viento como si de una distribución de velocidades de viento se tratase. En sus estudios demostraron que existía una equivalencia entre la ley de distribución potencial y los resultados que se obtendría con la ley de distribución de Weibull conforme a la Ecuación 2.7.

$$C_2 = C_1 \cdot \left(\frac{h_2}{h_1}\right)^n \quad k_2 = k_1 \cdot \frac{1 - 0.0881 \ln\left(\frac{h_1}{h_r}\right)}{1 - 0.0881 \ln\left(\frac{h_2}{h_r}\right)} \quad (2.7)$$

donde,

C_2 , C_1 , k_1 y k_2 : Son los coeficientes característicos de distribución de Weibull.

h_2 y h_1 : Son las alturas de medida de la extrapolación.

n : Exponente empírico obtenido a través de la siguiente expresión.

$$n = \frac{0.37 - 0.0881 \ln(C1)}{1 - 0.0881 \ln\left(\frac{h1}{hr}\right)} \quad (2.8)$$

Giovanni Gualtieri et al. [75] desarrollan una comparación entre los métodos clásicos de extrapolación Smedman – Högström y Högström, Panofsky y Dutton y los métodos basados en la distribución Weibull, entre ellos Justus y Mikhail y Spera y Richards, concluyendo que existe una clara ventaja de los métodos basados en Weibull cuando el objetivo es determinar el coeficiente de cizalladura del viento.

2.6. Principios de conversión de la energía eólica

Tal como indican algunas publicaciones [46] las masas de aire desplazadas en las capas bajas de la atmósfera contienen una cantidad de energía que con dependencia de las características locales del emplazamiento y las variaciones temporales del mismo permitirían su aprovechamiento en condiciones adecuadas de eficiencia energética y rentabilidad económica. Para la caracterización del potencial eólico del viento es necesario distinguir entre la energía eólica disponible y la energía eólica aprovechable con fines energéticos, siendo las diferencias entre ambas el tipo de sistema de conversión energética empleado.

2.6.1. Energía eólica disponible

Se define a la energía eólica disponible como la potencia mecánica disponible en las masas de aire en movimiento, la cual es proporcional a la densidad del aire, la sección expuesta en la corriente de aire y el cubo de la velocidad del fluido de acuerdo con la Ecuación 2.9.

$$PD = \frac{1}{2n} \sum_{i=1}^n \rho \cdot V^3 \cdot A \quad (2.9)$$

donde:

PD= Potencia disponible (W).

ρ = Densidad del aire (kg/m³).

V= Velocidad del viento medida (m/s).

n= Número de muestras de velocidad media cargados (intervalos).

A= Área expuesta a la corriente del aire (m²).

Por su parte, un dato indicativo de la potencia eólica disponible en un emplazamiento es una modificación de la Ecuación 2.9 en la que se obtiene la potencia por unidad de área de barrido. De este modo no es necesario definir un aerogenerador tipo a efectos de calcular la potencia en un emplazamiento Ecuación 2.10.

$$WPD = \frac{1}{2n} \sum_{i=1}^n \rho \cdot V^3 \quad (2.10)$$

En relación con la densidad algunos autores señalan que la variación de este parámetro con respecto a la altura puede ser de hasta un 7% [76] de ahí la importancia de su estimación mediante datos para el emplazamiento a estudio. Este aspecto será evaluado con mayor detalle en la presente tesis doctorar. En cualquier caso, es posible estimar de modo aproximado la potencia media disponible sólo con datos de la velocidad media del viento a través de la siguiente ecuación:

$$\frac{\rightarrow}{v^3} = \int_0^{\infty} V^3 f(v) dv \quad (2.11)$$

$$Ed = \left(\frac{Pd}{A} \right) \frac{Nh}{1000} \quad (2.12)$$

donde:

Ed = Potencia media disponible (kWh m⁻²).

Pd= Potencia disponible (W).

ρ = Densidad del aire (kg m⁻³).

V= Velocidad del viento medida (m s⁻¹).

Nh= Número de horas contenidas en la base de datos.

A= Área expuesta a la corriente del aire (m²).

De la misma forma se define el factor de irregularidad (*Ke*) como la relación existente entre la energía eólica disponible calculada y la que se obtendría a través del cubo de la velocidad media del viento de la muestra de datos según Ecuación 2.13.

$$Ke = \frac{v^3}{(\bar{v})^3} \quad (2.13)$$

2.6.2. Energía eólica aprovechable

Tal como definen Burton et al. [49], aunque la fuerza cinética es extraída del flujo del aire, un cambio fortuito en la velocidad de paso no es posible ni deseable debido a las enormes aceleraciones y fuerzas que serían requeridas en la turbina.

A medida que se aproxima el flujo del aire a la turbina la velocidad se ralentiza con respecto al flujo del aire no influenciado por la máquina, de la misma forma el tubo de corriente se expande como consecuencia de la ralentización producida ya que el aire eleva su presión estática para compensar la disminución de la energía cinética. Por otra parte, cuando el aire pasa a través del disco rotor, hay una caída en la presión estática de tal manera que tras el paso por la turbina la presión es inferior a la atmosférica. Esta región de flujo se llama estela y debe ser evitada en la consecución de turbinas en una misma fila [49].

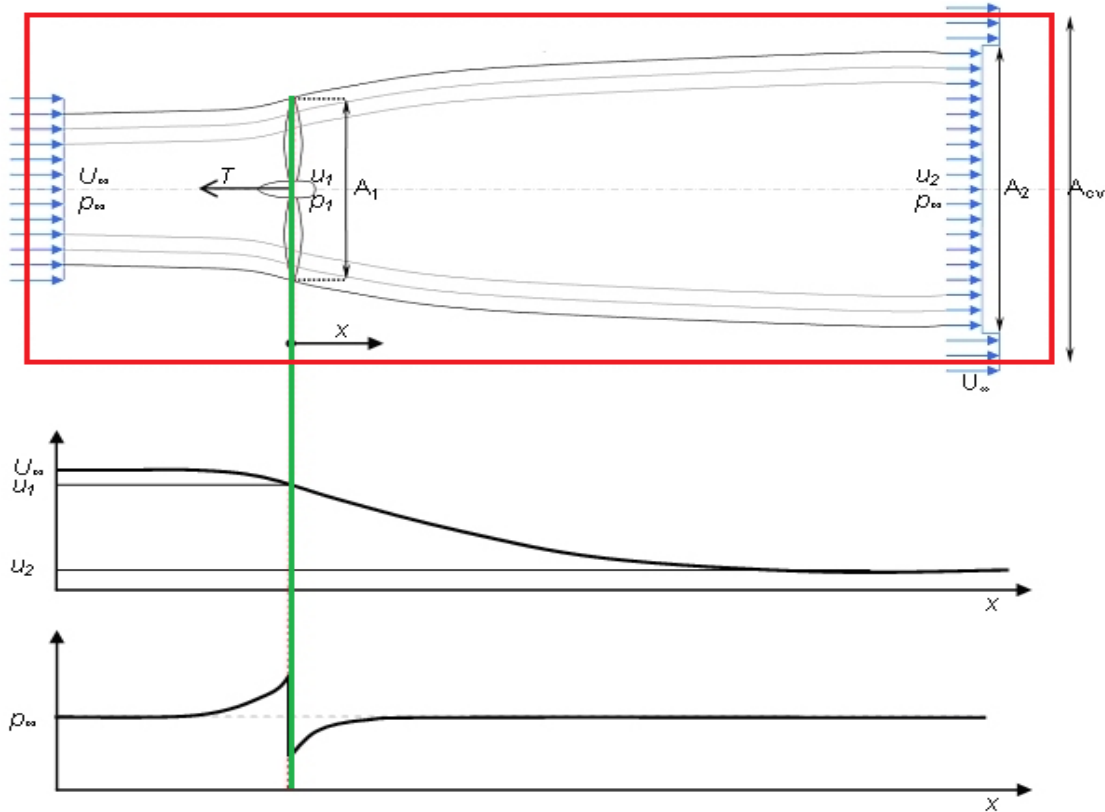


Figura 11 Energía extraída del tubo de corriente de aire en la turbina eólica

Para la modelización de la energía aprovechable se aplican las siguientes expresiones básicas de la mecánica de fluidos:

1. **Principio de continuidad:** Debe cumplirse que la masa se conserve a la entrada y a la salida de la turbina. Para ello se aplica la Ecuación 2.14.

$$\rho \cdot A_1 \cdot v_1 = \rho \cdot A_2 \cdot v_2 \quad (2.14)$$

donde,

ρ = Densidad del aire (kg m^{-3}).

v = Velocidad del viento medida (m s^{-1}).

A = Área expuesta a la corriente del aire (m^2).

2. **Principio de la cantidad de movimiento:** Por otra parte, la fuerza ejercida por el generador eólico sobre el aire tiene un movimiento que se dirige hacia delante y es igual al valor absoluto designado por el teorema de Euler Ecuación 2.15.

$$F = \rho \cdot Q \cdot (v_1 - v_2) = \rho \cdot A \cdot v \cdot (v_1 - v_2) \quad (2.15)$$

3. **Ecuación de Bernouilli:** Esta ecuación se aplica entre la sección de aire libre y la parte anterior al rotor y posteriormente entre la sección trasera del rotor y el área de estela obteniéndose lo siguiente Ecuación 2.16.

$$\frac{1}{2} \cdot \rho \cdot v_1^2 + \rho_0 + \rho \cdot g \cdot h_1 = \frac{1}{2} \cdot \rho \cdot v^2 + \rho^+ + \rho \cdot g \cdot h \quad (2.16)$$

Con todo ello, se define un coeficiente de potencia C_p que expresa la potencia mecánica que puede ser extraída del rotor y que se obtiene en función de cada máquina bajo unas mismas condiciones de velocidad del viento Ecuación 2.17.

$$C_p = \frac{P}{P_0} = \frac{\frac{1}{4} \cdot \rho \cdot A \cdot (v_1^2 - v_2^2) + (v_1 + v_2)}{\frac{1}{2} \cdot \rho \cdot A \cdot v_1^3} \quad C_p = \frac{P}{P_0} = \frac{1}{2} \left| 1 - \left(\frac{v_2}{v_1} \right)^2 \right| \left| 1 + \frac{v_2}{v_1} \right| \quad (2.17)$$

Por todo ello el coeficiente de potencia depende de la relación de velocidades del aire existentes entre la entrada y la salida del rotor de la turbina. En la Figura 12 se representa la variación del coeficiente de potencia en función de la relación de velocidades (v_2/v_1) entre el valor mínimo (0) y el valor máximo (1).

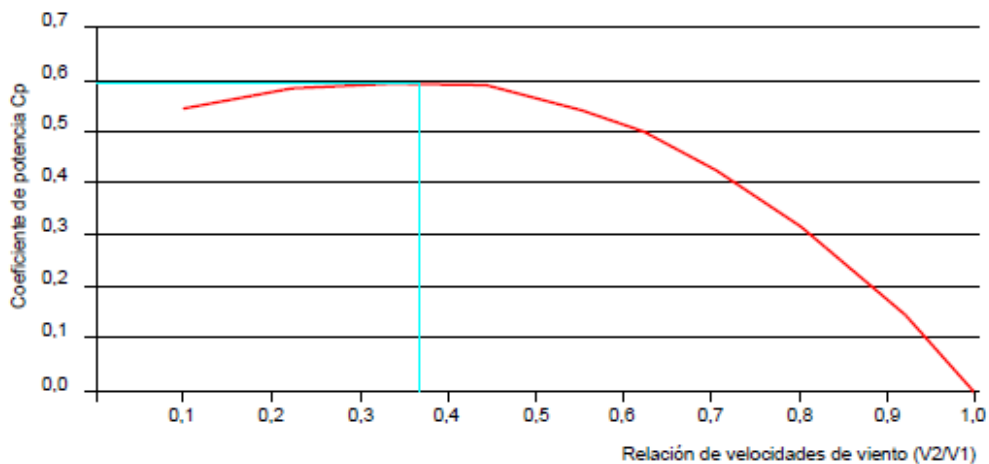


Figura 12 Representación límite de Betz.

El valor máximo del coeficiente de potencia se establece por tanto en $16/27$ (0.593), valor conocido como *límite de Betz*. En la práctica, los modelos actuales de aerogeneradores poseen un coeficiente de potencia de en torno a 0.4 por aspectos relaciones a los límites físicos derivados del rozamiento, la turbulencia y otros fenómenos como la rafagosidad o el cizallamiento [47].

Se define la Ecuación 2.18 para la cuantificación de la potencia eólica aprovechable:

$$P_a = C_p \cdot \frac{1}{2} \cdot \rho \cdot A \cdot v^3 \quad (2.18)$$

En la Figura 13 se representa la energía que teóricamente sería aprovechable en comparación con la que sería perdida debido a límites físicos, esta vez en función de la densidad de potencia [77].

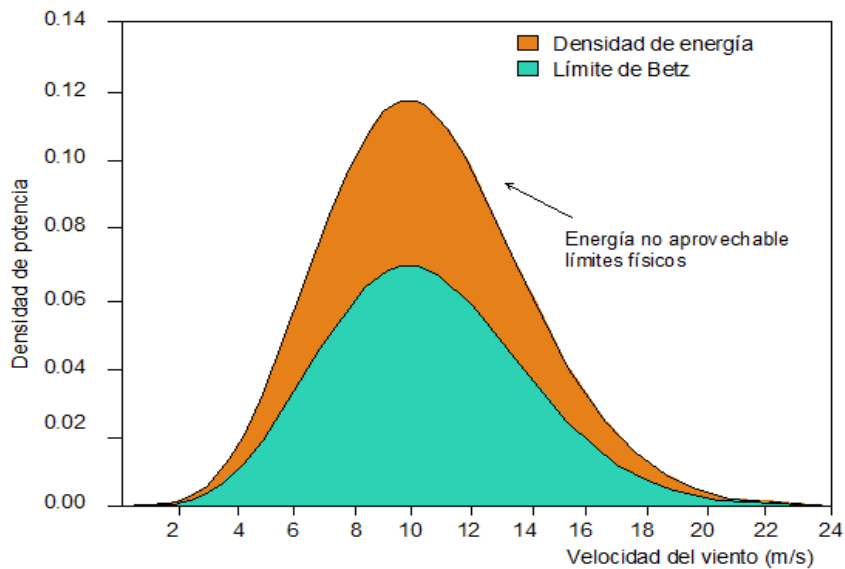


Figura 13 Densidad teórica de energía eólica extraíble con una turbina

2.6.3. Estimación de la energía producida

En la estimación de la energía producida por un parque eólico conviene distinguir entre los métodos de estimación estáticos, los semiestáticos o los cuasidinámicos [46,47]:

- **Métodos estáticos:** Se parte de curvas de distribuciones probabilísticas del viento y de la curva de potencia del aerogenerador proporcionada por el fabricante, desarrollándose un análisis que puede ser analítico o gráfico. Con la primera curva se determina la probabilidad de que se produzca en un emplazamiento una velocidad de viento determinada, mientras que con la segunda se desarrolla una equivalencia entre esa velocidad de entrada y la potencia instantánea que sería generada por un modelo de aerogenerador determinado. Como resultado se logra una curva de probabilidad de potencia cuya integral da como resultado la energía producida por la instalación.

La principal ventaja de este método es su simplicidad, sin embargo, los datos de partida siempre se encontrarían sesgados ya que se parten de curvas que para ambos casos han sido idealizadas, ignorándose efectos estacionarios y otras las pérdidas derivadas de la operación.

- **Métodos semiestáticos:** En este caso se va un poco más allá tomándose series temporales de datos de velocidad del viento para ser comparados con la curva de potencia del aerogenerador tipo a instalar. Así pues, si bien la curva de potencia aún seguiría estando idealizada, existiría mayor información sobre aspectos como la producción generada teniendo en cuenta los tiempos de arranque o parada y otros fenómenos como el comportamiento del aerogenerador frente a los cambios de orientación.
- **Métodos cuasidinámicos:** Ya por último con los métodos cuasidinámicos se toman como datos de partida series temporales de datos de viento en el emplazamiento y un modelo numérico de

simulación del aerogenerador de tal forma que se evalúa de manera muy exacta el procedimiento de operación del aerogenerador y la energía obtenida. Esta es la opción más completa y es comúnmente usada para simulaciones en las que se quiere averiguar las estrategias de control más adecuadas en cada caso así como los criterios de operación y de estabilidad de la red eléctrica en general.

2.7. Estimación de la densidad del aire

Conforme a lo descrito en el Apartado 2.6, uno de los parámetros fundamentales en la estimación de la potencia eólica es la densidad del aire, la cual puede ser estimada por diversos procedimientos de acuerdo con la literatura científica especializada en meteorología [78-80]. De todas ellas, quizá la expresión más precisa es la recomendada por el Comité Internacional de Pesas y Medidas recogida en el procedimiento CIPM-2007 [78], la cual constituye una versión revisada de la CIPM-81/91 [79]. En esta versión, la incertidumbre relativa de la densidad del aire ρ debido a la fórmula de la CIPM-2007 [78] es de aproximadamente 22×10^{-6} . No obstante, las ecuaciones que definen la presión de saturación de vapor (P_{sv}), el enhancement factor (ef) y el factor de compresibilidad (Z) solo son válidas para rangos limitados de la presión barométrica (p) y de la temperatura del aire (t_a). Concretamente, los rangos recomendados de dichos parámetros en los cuales la ecuación de la CIPM-2007 puede ser utilizada son: $600 \text{ hPa} \leq p \leq 1.100 \text{ hPa}$ y $15^\circ\text{C} \leq t_a \leq 27^\circ\text{C}$. Asimismo, se han propuesto fórmulas simplificadas para el cálculo de la densidad del aire [80] que son menos sofisticadas que la fórmula CIPM-2007 y que proporcionan resultados con menor grado de precisión. Sin embargo, dichas fórmulas o presentan restricciones en el rango de presiones, temperaturas y humedades relativas de utilización que no podrían ser utilizadas en ciertos instantes teniendo en cuenta los datos con los que se desarrollan los trabajos de la presente tesis doctoral (ver valores medios de la velocidad y la presión de los datos recopilados y expuestos en el Capítulo 4) o no se especifica (en la bibliografía consultada) los rangos recomendados de las mencionadas variables.

Como puede observarse en la Tabla 10, en las estaciones involucradas en el estudio las presiones atmosféricas se encuentran dentro del rango de utilización recomendado por la CIPM-2007, sin embargo, se han registrado temperaturas que se encuentran fuera del intervalo recomendado señalado por la mencionada norma. Aunque el porcentaje de datos medios horarios de temperaturas que no se encuentran en el rango indicado por la CIPM-2007 es bajo (inferior al 7% para todas las series temporales), para estimar las densidades del aire en el intervalo de temperatura comprendidos entre $7.3^\circ\text{C} \leq t_a \leq 37.4^\circ\text{C}$ se ha recurrido al uso de expresiones que han sido propuestas para un intervalo más amplio de dicha variable.

En este contexto, la densidad de la mezcla de aire (moist air) ρ (kg m^{-3}), es evaluada empleando una ecuación de estado, Ecuación 2.19 [78].

$$\rho = \frac{p M_a}{Z R (t_a + 273.15)} \left[1 - \psi_w \left(1 - \frac{M_v}{M_a} \right) \right] \quad (2.19)$$

En Ecuación 2.19 p es la presión atmosférica en Pascales, M_a es la masa molar del aire seco ($28.96546 \times 10^{-3} \text{ kg mol}^{-1}$), M_v es la masa molar del vapor de agua ($18.01528 \times 10^{-3} \text{ kg mol}^{-1}$), R es la constante universal de masa molar del gas ($8.314472 \text{ J K}^{-1}\text{mol}^{-1}$), t_a es la temperatura ambiente en grados Celsius, Z es un parámetro no dimensional que determina el factor de compresibilidad del aire y ψ_w es la fracción molar del vapor de agua, Ecuación 2.20.

$$\psi_w = ef \frac{H}{100} \frac{p_{sv}}{p} \quad (2.20)$$

En Ecuación 2.20 H es la humedad relativa del aire (%). Para estimar la presión de saturación sobre agua líquida, p_{sv} , se ha utilizado la fórmula propuesta por Hyland y Wexler 1983 [81], Ecuación 2.21, la cual cubre el rango de temperatura de 0°C a 200°C.

$$p_{sv} = \text{Exp} \left[\frac{g_1}{t_a + 273.15} + \sum_{i=2}^5 g_i (t_a + 273.15)^{i-2} + g_6 \ln(t_a + 273.15) \right] \quad (2.21)$$

En Ecuación 2.21, p_{sv} viene dada en unidades de Pascales cuando las unidades de t_a están en grados Celsius y se utilizan las constantes g_i recogidas en la Tabla 2.

Presión de saturación del vapor		Compresibilidad isotermica		Volumen molar de agua líquida saturada	
g_1	$-5.8002206 \cdot 10^3$	ic_1	$5.088496 \cdot 10$	mv_1	-2403.360201
g_2	1.3914993	ic_2	$6.163813 \cdot 10^{-1}$	mv_2	-1.40758895
g_3	$-4.8640239 \cdot 10^{-2}$	ic_3	$1.459187 \cdot 10^{-3}$	mv_3	0.1068287657
g_4	$4.1764768 \cdot 10^{-5}$	ic_4	$2.008438 \cdot 10^{-5}$	mv_4	$-2.914492351 \cdot 10^{-4}$
g_5	$-1.4452093 \cdot 10^{-8}$	ic_5	$-5.847727 \cdot 10^{-8}$	mv_5	$3.73497936 \cdot 10^{-6}$
g_6	6.5459673	ic_6	$4.104110 \cdot 10^{-10}$	mv_6	$-2.1203787 \cdot 10^{-10}$
		ic_7	$1.967348 \cdot 10^{-2}$		

Tabla 2 Coeficientes de las Ecuaciones 2.19, 2.20 y 2.21

Para calcular el factor de compresibilidad Z (adimensional) en el intervalo considerado de temperaturas del aire, se ha utilizado la siguiente ecuación de estado de mezcla virial [82], Ecuación 2.22.

$$Z = \frac{p\bar{v}}{R(t_a + 273.15)} = 1 + \frac{B_m}{\bar{v}} + \frac{C_m}{\bar{v}^2} \quad (2.22)$$

En Ecuación 2.22 \bar{v} es el volumen molar de la mezcla (molar mixture volumen), B_m y C_m se conocen como coeficientes molares viriales de la mezcla (segundo y tercer coeficiente molar virial de la mezcla, respectivamente) y se han calculado utilizando las Ecuaciones 2.23 y 2.24 [82,83].

$$B_m = (1-\psi_w)^2 B_{aa} + 2(1-\psi_w)\psi_w B_{aw} + \psi_w^2 B_{ww} \quad (2.23)$$

$$C_m = (1-\psi_w)^3 C_{aaa} + 3(1-\psi_w)^2 \psi_w C_{aaw} + 3(1-\psi_w)\psi_w^2 C_{aww} + \psi_w^3 C_{www} \quad (2.24)$$

En la Ecuación 2.23, B_{aa} ($\text{m}^3 \text{mol}^{-1}$) es el segundo coeficiente virial aire – aire y se estima mediante la Ecuación 2.25, B_{aw} ($\text{m}^3 \text{mol}^{-1}$) es el segundo coeficiente virial cruzado y su valor se determina con la Ecuación 2.26, B_{ww} ($\text{m}^3 \text{mol}^{-1}$) es el segundo coeficiente virial agua – agua que se calcula con la Ecuación 2.27.

En la Ecuación 2.24, C_{aaa} ($\text{m}^6 \text{mol}^{-2}$) es el tercer coeficiente virial del aire y se estima mediante la Ecuación 2.28, C_{aaw} ($\text{m}^6 \text{mol}^{-2}$) es el tercer coeficiente virial aire – arie – agua y se calcula mediante Ecuación 2.29, C_{aww} ($\text{m}^6 \text{mol}^{-2}$) es el tercer coeficiente virial aire – agua – agua y se determina mediante la Ecuación 2.30, C_{www} ($\text{m}^6 \text{mol}^{-2}$) es el tercer coeficiente virial del agua que se estima mediante la Ecuación 2.31.

$$B_{aa} = \sum_{i=1}^7 [aa_i (t_a + 273.15)^{i-1}] \quad (\text{Range: } -100^\circ\text{C to } 380^\circ\text{C}) \quad (2.25)$$

$$B_{aw} = \frac{1}{10^6} \sum_{i=1}^3 [aw1_i \left(\frac{t_a + 273.15}{100}\right)^{aw2_i}] \quad (\text{Range: } -123.15^\circ\text{C to } 1726.85^\circ\text{C}) \quad (2.26)$$

$$B_{ww} = \frac{1}{10^6} (t_a + 273.15) \left[ww_1 + ww_2 e^{\frac{ww_3}{t_a + 273.15}} \right] \quad (\text{Range: } -100^\circ\text{C to } 200^\circ\text{C}) \quad (2.27)$$

$$C_{aaa} = \frac{1}{10^{12}} \sum_{i=1}^7 [aaa_i (t_a + 273.15)^{i-1}] \quad (\text{Range: } -100^\circ\text{C to } 380^\circ\text{C}) \quad (2.28)$$

$$C_{aaw} = \frac{1}{10^{10}} \sum_{i=1}^5 [aaw_i (t_a + 273.15)^{1-i}] \quad (\text{Range: } -100^\circ\text{C to } 200^\circ\text{C}) \quad (2.29)$$

$$C_{aww} = \frac{-1}{10^{-6}} \text{Exp} \left[\sum_{i=1}^4 aww_i (t_a + 273.15)^{1-i} \right] \quad (\text{Range: } -100^\circ\text{C to } 200^\circ\text{C}) \quad (2.30)$$

$$C_{www} = \frac{1}{10^{10}} (t_a + 273.15)^2 \left[www_1 + www_2 e^{\frac{www_3}{t_a + 273.15}} + \sum_{i=3}^4 www_{i+1} \left[e^{\frac{www_6}{t_a + 273.15}} \right]^{i-2} \right] \quad (\text{Range: } -100^\circ\text{C to } 200^\circ\text{C}) \quad (2.31)$$

Los coeficientes aa_i [83], aw_i [84], aaa_i [83], aaw_i [83,85], aww_i , [83,85] ww_i [85] y www_i [85] de las Ecuaciones 2.25 a 2.31, se han obtenido de las referencias citadas y se muestran en las Tabla 3 - 4.

i	aa	aw1	aw2	ww
1	31.831763	$0.665687 \cdot 10^2$	-0.237	0.05820
2	-719.51195	$-0.238834 \cdot 10^3$	-1.048	0.012234
3	-6538137	$-0.176755 \cdot 10^3$	-3.183	1734.29
4	$1.5929828 \cdot 10^9$	-	-	-
5	$-2.5588842 \cdot 10^{11}$	-	-	-
6	$2.2300382 \cdot 10^{13}$	-	-	-
7	$-8.2793465 \cdot 10^{14}$	-	-	-

Tabla 3 Coeficientes de las Ecuaciones 2.25, 2.26 y 2.27

i	aaa	aaw	aww	www
1	1297.5378	$0.482737 \cdot 10^3$	$-0.1072887 \cdot 10^2$	$7.528231 \cdot 10^{-14}$
2	46021.328	$0.105678 \cdot 10^6$	$0.347804 \cdot 10^4$	$-2.3179 \cdot 10^{-6}$
3	40813154.0	$-0.656394 \cdot 10^8$	$-0.383383 \cdot 10^6$	3645.09
4	$-3.302391 \cdot 10^9$	$0.294442 \cdot 10^{11}$	$0.334060 \cdot 10^8$	$1.424471 \cdot 10^{-15}$
5	$2.2964785 \cdot 10^{11}$	$-0.319317 \cdot 10^{13}$	-	$1.497567 \cdot 10^{-16}$
6	$-5.3683467 \cdot 10^{12}$	-	-	1734.29
7	$-2.1183915 \cdot 10^{14}$	-	-	-

Tabla 4 Coeficientes de las Ecuaciones 2.28 a 2.31

El volumen de la mezcla molar \bar{v} ($\text{m}^3 \text{mol}^{-1}$) puede ser obtenido como solución de la Ecuación 2.22. Para evaluar el enhancement factor ef , se utiliza la Ecuación 2.32 [37-38].

$$\begin{aligned}
 \ln(ef) = & \left[\frac{(1 + \kappa p_{sv})(p - p_{sv}) - \kappa \left(\frac{p^2 - p_{sv}^2}{2} \right)}{R(t_a + 273.15)} \right] v_{vs} + \ln \left[1 - \beta_H (1 - \psi_w) p \right] \\
 & + \left[\frac{(1 - \psi_w)^2 p}{R(t_a + 273.15)} \right] B_{aa} - 2 \left[\frac{(1 - \psi_w)^2 p}{R(t_a + 273.15)} \right] B_{aw} - \left[\frac{p - p_{sv} - (1 - \psi_w)^2 p}{R(t_a + 273.15)} \right] B_{ww} \\
 & + \left[\frac{(1 - \psi_w)^3 p^2}{R^2(t_a + 273.15)^2} \right] C_{aaa} + \left[\frac{3(1 - \psi_w)^2 [1 - 2(1 - \psi_w)] p^2}{2R^2(t_a + 273.15)^2} \right] C_{aaw} \\
 & - \left[\frac{3(1 - \psi_w)^2 \psi_w p^2}{R^2(t_a + 273.15)^2} \right] C_{aww} - \left[\frac{(3 - 2\psi_w) \psi_w^2 p^2 - p_{sv}^2}{2R^2(t_a + 273.15)^2} \right] C_{www} \\
 & - \left[\frac{(1 - \psi_w)^2 (-2 + 3\psi_w) \psi_w p^2}{R^2(t_a + 273.15)^2} \right] B_{aa} B_{ww} - \left[\frac{2(1 - \psi_w)^3 (-1 + 3\psi_w) p^2}{R^2(t_a + 273.15)^2} \right] B_{aa} B_{aw}
 \end{aligned} \tag{2.32}$$

$$\begin{aligned}
& + \left[\frac{6(1-\psi_w)^2 \psi_w^2 p^2}{R^2(t_a + 273.15)^2} \right] B_{ww} B_{aw} - \left[\frac{3(1-\psi_w)^4 p^2}{2R^2(t_a + 273.15)^2} \right] B_{aa}^2 \\
& - \left[\frac{2(1-\psi_w)^2 \psi_w (-2 + 3\psi_w) p^2}{R^2(t_a + 273.15)^2} \right] B_{aw}^2 - \left[\frac{p_{sv}^2 - (4 - 3\psi_w) \psi_w^3 p^2}{2R^2(t_a + 273.15)^2} \right] B_{ww}^2
\end{aligned}$$

En la Ecuación 2.32 κ es la compresibilidad isotérmica del agua líquida saturada (Pa^{-1}), β_H es la constante de la ley Henry y v_{sw} ($\text{m}^3 \text{mol}^{-1}$) es el volumen de masa de agua líquida saturada.

Kell [86] ha propuesto una expresión, Ecuación 2.33, para obtener los valores de κ en el rango 0°C a 150°C . Los valores de las constantes ic_i se muestran en la Tabla 2.

$$\kappa = \frac{1}{10^{11}} \left[\sum_{i=1}^7 \frac{ic_i t^{i-1}}{1 + ic_7} \right] \quad (2.33)$$

v_{sw} se ha estimado mediante la Ecuación 2.34 [42]. Los valores de las constantes ic_i se muestran en la Tabla 2.

$$v_{sw} = \frac{1}{10^6} \left[\sum_{i=1}^6 mv_i (t_a + 273.15)^{i-1} \right]^{-1} \left[-61692.295 + 291.8088(t_a + 273.15) \right] \quad (2.34)$$

Si se considera que el aire se compone solo de oxígeno (fracción molar del oxígeno en el aire, $x_{O_2} = 0.22$) y nitrógeno (fracción molar del nitrógeno en el aire $x_{N_2} = 0.78$), la constante de Henry β_H puede ser estimada mediante la Ecuación 2.35 [83].

$$\beta_H = \frac{0.0001}{101325} \left[\frac{x_{O_2}}{\beta_{O_2}} + \frac{x_{N_2}}{\beta_{N_2}} \right] \quad (2.35)$$

Nelson y Sauer [83] han propuesto ecuaciones para representar las variaciones de β_{O_2} y de β_{N_2} en función de la temperatura, las cuales han sido utilizadas para las estimaciones de densidad del aire realizadas en la presente tesis doctoral. El enhancement factor ef se calcula de forma iterativa mediante la Ecuación 2.32, ya que ψ_w , Ecuación 2.20, es función de ef .

Tal y como se describe con mayor detalle en el Capítulo 4 de esta tesis, para la estimación de la densidad del aire se han usado datos de temperatura y humedad relativa ubicados a una altura de 2 metros sobre el nivel del suelo para la misma posición en la que se recaban los datos de recurso eólico en cada estación anemométrica. Sin embargo, según se especifica en la norma IEC 61400-12-1 [87], los sensores para la medida de la temperatura del aire y la humedad deben estar montados, si se utilizan con fines relacionados a la estimación de potencia eólica, a una distancia menor de 10 m

de la altura del buje de la turbina eólica para representar la temperatura del aire en el centro del rotor de la misma. Asimismo, se indica en la IEC 61400-12-1 [87] que el sensor de presión de aire debe montarse en el mástil meteorológico cerca de la altura del buje para representar la presión del aire en el centro del rotor de la turbina eólica.

Dado que los datos registrados de temperatura así como de humedad relativa no cumplirían con los mencionados requisitos, en los trabajos realizados en la presente tesis doctoral se ha recurrido al procedimiento que se describe a continuación:

1. Como el sensor de presión no estaría montado cerca de la altura del buje de los aerogeneradores, las mediciones de la presión de aire se corregirían a la altura del buje de las mismas, siguiendo las especificaciones de la norma ISO 2533 [88], tal como se indica en la norma IEC 61400-12-1 [87].
2. No se considera la humedad relativa en esta expresión ya que la IEC 61400-12-1 [87] sólo recomienda que se considere dicha variable en el cálculo de la densidad del aire en el caso de altas temperaturas y las temperaturas medias de las estaciones meteorológicas consideradas como objetivos son inferiores a los 30°C.
3. En el caso de la temperatura se considera que esta disminuye linealmente con la altura con un gradiente $\beta = -6.5 \text{ K km}^{-1}$. Con los anteriores puntos de partida, la densidad del aire a la altura de buje del aerogenerador se ha estimado a través de la Ecuación 2.36.

$$\rho(z)_i = \rho(z_r)_i \left[\frac{T(z_r) + \beta \cdot [z - z_r]}{T(z_r)} \right]^{-\frac{g}{\beta \cdot R}} \quad (2.36)$$

En la Ecuación 2.36 g es la aceleración de la gravedad y R es la constante del aire seco. Los valores utilizados para dichos parámetros han sido de 9.80617 m s^{-2} y $287.053 \text{ m K}^{-1} \text{ s}^{-2}$, respectivamente. Las unidades de la temperatura del aire a la altura de referencia, $T(z_r)$, vienen dadas en la Ecuación 2.36 en grados Kelvin.

2.8. Formulación usada para la estimación de la potencia eólica a través de datos medidos en las estaciones meteorológicas.

Los fabricantes de aerogeneradores, siguiendo procedimientos normalizados, tales como los recogidos en IEC 61400-12-1 [87], suelen proporcionar las curvas de potencia de los modelos de aerogenerador ofertados estableciendo una equivalencia entre la velocidad del viento a altura del buje (z) y la potencia que sería producida en ese instante en formato discretizado asumiendo unos valores concretos de densidad del aire orientativos entre los que siempre se encuentra el valor estándar de 1.225 kg m^{-3} (ρ_0). Dependiendo de la estrategia de control, la curva de potencia del aerogenerador puede tener una forma concreta, distinguiéndose fundamentalmente dos tipos

principales, aerogeneradores con control tipo stall – regulated y aerogeneradores con control active power (pitch – regulated).

Tal y como se describe con mayor detalle en el Capítulo 9, una de las principales incógnitas que se pretende dilucidar en la presente tesis doctoral es la influencia de la densidad del aire sobre la estimación de la potencia producida por un aerogenerador en estudios relacionados con la estimación a largo plazo. En este sentido, dependiendo del tipo de aerogenerador seleccionado las conclusiones obtenidas podrían diferir, razón por la cual se ha decidido realizar los ensayos propuestos teniendo en cuenta distintos tipos de aerogenerador en función del sistema de control utilizado. Tras realizar una búsqueda de modelos comerciales se ha decidido realizar los análisis usando los coeficientes de potencia eléctrica (C_p)³ y las curvas de potencia de los dos modelos de aerogenerador expuestos en la Figura 14, correspondiéndose la imagen de la izquierda (WT-2) con el modelo que implementa el sistema de control tipo Stall – regulated [89] y la imagen de la derecha (WT-1) con el aerogenerador que implementa el sistema active power [90].

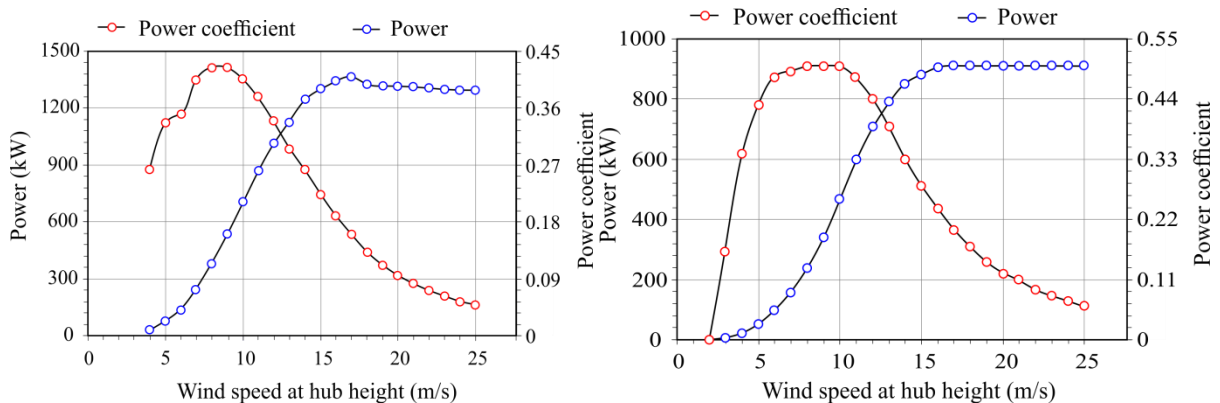


Figura 14 Curvas de potencia utilizadas en los análisis de WTPO (Capítulo 9). Modelos: WT-1 (Derecha) y WT-2 (Izquierda)

Es importante mencionar que, a pesar de la búsqueda exhaustiva llevada a cabo, no se han encontrado modelos que cuadraran a la perfección en cuanto a su potencia nominal de salida, razón por la cual se ha apostado por dos modelos que aunque la potencia nominal es distinta (WT-1 de 900 kW y WT-2 de 1300 kW), la diferencia entre las potencias de dichos modelos es la menor de cuantas opciones fueron valoradas.

Como ha sido anticipado en el Apartado 2.6.3, en la literatura científica se han propuesto diversos modelos de curvas de potencia de turbinas eólicas que representan de forma continua (desde la velocidad de arranque a la velocidad de parada) la relación entre la velocidad del viento y la potencia producida por las mismas [91,92]. Sin embargo, dichos modelos empíricos son entrenados utilizando

³ Representa la cantidad de energía cinética del viento convertida en potencia eléctrica por un aerogenerador. Es decir, tiene en cuenta el coeficiente de potencia del rotor, el rendimiento de los sistemas mecánicos y el rendimiento eléctrico del generador eléctrico [236].

datos de velocidad y potencia generada por un aerogenerador que han sido registrados en un lugar específico donde dicha turbina ha sido instalada. Así pues, dichos modelos empíricos se proponen en su origen para reducir la discrepancia existentes [91] entre la curva de potencia proporcionada por los fabricantes (para una densidad del aire específica, que típicamente es de 1.225 kg m^{-3}) y las curvas de potencia reales del lugar donde se encuentra instalado y en operación el aerogenerador. A priori, este tipo de ajustes de modelos empíricos (ya sean paramétricos o no paramétricos) se consideran herramientas destinadas a la predicción de la energía eólica. Teniendo en cuenta que el propósito de los estudios desarrollados en esta tesis doctoral es la estimación a largo plazo, en este caso concreto, de la potencia producida de un aerogenerador, no es posible utilizar este tipo de modelos de curvas de potencia empíricos ya que no se disponen de datos reales de producción en las mismas instalaciones (notar que este tipo de estudios se llevan a cabo con antelación a la instalación del aerogenerador o parque eólico). Sin embargo, como señalan Lydia et al. [92] las curvas de potencia proporcionadas por los fabricantes pueden ayudar a estimar el potencial de la energía eólica en un sitio candidato. Además, dado que en la estimación de la potencia producida se tendrá en cuenta la variación de la densidad del aire, se considera que se lograría una cierta adaptación de la curva de potencia proporcionada por el fabricante a las características particulares del lugar objetivo.

Como ha sido mencionado anteriormente, las curvas de potencia proporcionadas por los fabricantes suelen estar discretizadas por valores enteros de velocidad del viento. Con el propósito de estimar las relaciones entre las velocidades del viento y las potencias generadas por los aerogeneradores, entre los puntos de estas curvas de potencia se utilizará el método de interpolación con funciones spline cubicas [93].

Como se ha expuesto en la Figura 14, las curvas de potencia utilizadas para los análisis realizados en la presente tesis doctoral sólo exponen los coeficientes de potencia y las potencias cuando la densidad del aire es coincidente con el valor estándar según la norma ISO ($\rho_0 = 1.225 \text{ kg m}^{-3}$). En este sentido, es necesaria la estimación de las potencias generadas por las mismas cuando las densidades del aire difieran de dicho valor concreto. Para realizar dicha estimación, se distingue entre los aerogeneradores con sistema de control stall – regulated y los aerogeneradores pitch – regulated.

En el caso de aerogeneradores con sistema de control stall – regulated, la potencia horaria generada ($WTPO_i$) a partir de la velocidad horaria del viento $v(z)_i$ y la densidad del aire $\rho(z)_i$ (ambas referidas a la altura de buje (z) en el sitio objetivo), se calculan por medio de la Ecuación 2.37 [43,87,94].

$$WTPO_i(v(z)_i, \rho(z)_i) = \left[P(v(z)_i, \rho_0) \right] \left[\frac{\rho(z)_i}{\rho_0} \right] \quad (2.37)$$

En el caso de los aerogeneradores con control de tipo pitch – regulated, las potencias horarias generadas $WTPO_i$, se estiman modificando la velocidad del viento de acuerdo con la Ecuación 2.38. Así pues, el valor representado con la variable $v(z)_i$ en la Ecuación 2.37 se obtiene mediante la Ecuación 2.38, la cual escala la velocidad en función de la diferencia de ésta con respecto a la obtenida con el valor estándar de densidad del aire según la norma ISO.

$$v^*(z)_i = v(z)_i \left[\frac{\rho(z)_i}{\rho_0} \right]^\gamma \quad (2.38)$$

Por tanto, las potencias de generadas se estimarían haciendo uso de la Ecuación 2.39.

$$WTPO_i(v(z)_i, \rho(z)_i) = \left[P(v^*(z)_i, \rho_0) \right] \quad (2.39)$$

La IEC 61400-12-1 [87] utiliza un exponente γ para la Ecuación 2.37 de 1/3. Sin embargo, en los trabajos realizados en esta tesis se ha apostado por los valores propuestos por Svenningsen [95], los cuales han sido a su vez adoptados por el cálculo realizado mediante el software WindPro [96]. Svenningsen [95], defiende que con los valores de γ que propone se mejora el método de corrección de la curva de potencia que se indica en la IEC 61400-12-1 [87]. Para ello, el autor ha llevado a cabo la calibración utilizando una serie de densidades del aire para curvas de potencia específicas proporcionadas por varios fabricantes. Según Svenningsen [95], en el caso de bajas densidades del aire el método propuesto por la IEC61400-12 [87] puede producir una sobreestimación comprendida entre el 4 y el 5% de la producción de energía anual. Sin embargo, señala el autor que dicho error generalmente se reduce a <1% si se utiliza la corrección que éste propone. En este contexto, en este trabajo se ha utilizado un valor de γ de 1/3 para velocidades del viento menores o iguales a 8 m s⁻¹ y 2/3 para velocidades mayores o iguales a 12 m s⁻¹. Para velocidades del viento superiores a 8 m s⁻¹ e inferiores a 12 m s⁻¹ el exponente γ ha sido escalonado suavemente desde 1/3 hasta 2/3.

2.9. Necesidad del estudio del potencial a largo plazo

El valor estándar de inversión en un parque eólico de nueva construcción varía entre 1402000 y 1450000 €/MW [97], siendo una de las principales inquietudes del promotor el conocimiento de la rentabilidad económica esperada. En este sentido, en los estudios previos a su construcción se suele desarrollar un profundo análisis económico, determinando el beneficio total generado a través de la estimación de los flujos netos de caja anuales para el periodo de explotación de las instalaciones.

Para reducir la incertidumbre de la inversión y apuntalar la toma de decisión del promotor, junto con el estudio de viabilidad económica se ejecuta un análisis de sensibilidad con el cual se demuestra la influencia de cada variable o dato de partida en la rentabilidad del proyecto. Comúnmente, **las variables que mayor influencia tienen en la rentabilidad económica del proyecto son el régimen retributivo y las circunstancias locales del potencial eólico en la región.**

Para ejemplificar la influencia del recurso eólico en la rentabilidad económica de un proyecto, en la Figura 15 se adjuntan los resultados obtenidos en una simulación económica de un parque eólico tipo de 2 MW que fuera instalado en la Isla de Gran Canaria y al que le fuera aplicable el régimen retributivo específico marcado por la legislación vigente. Asimismo, en este análisis económico se ha considerado que el parque eólico estaría en explotación durante un periodo de 25 años, habiendo

ascendido la inversión hasta los 2900000 € (1450000 €/MW) y siendo los costes de explotación de aproximadamente 32 €/kWh actualizables anualmente en función del IPC.

Como se puede observar para unas mismas condiciones de partida, si se varía la producción del parque eólico entre las 1600 y 4600 horas teóricas equivalente de producción anual, valores mínimos y máximos registrados en el archipiélago durante los últimos 30 años según el Anuario Energético de Canarias 2015 [98], **la rentabilidad del proyecto pasa de apenas recuperarse la inversión a generar un beneficio próximo a la cantidad invertida en el momento inicial.**

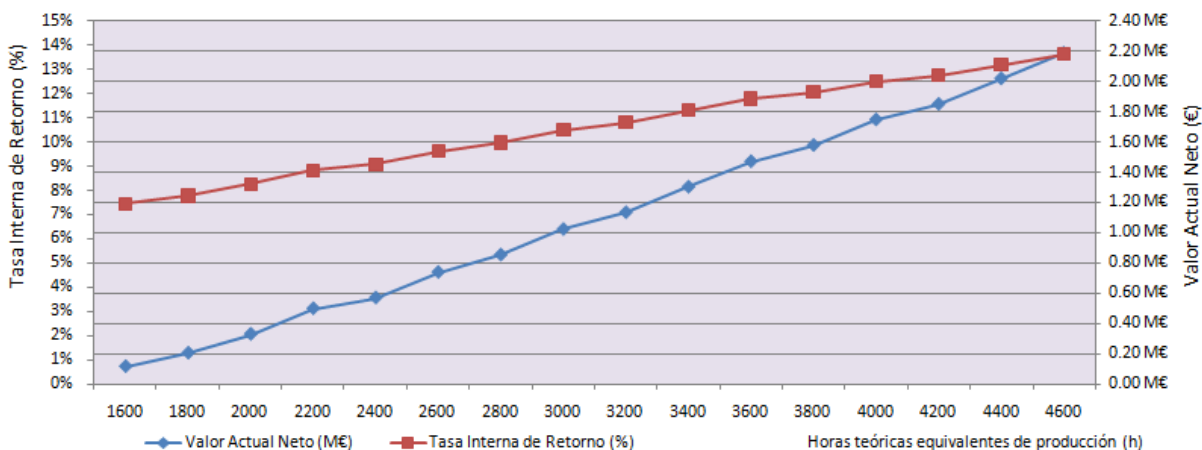


Figura 15 Sensibilidad de la rentabilidad frente a la producción eólica de un P.E de 2 MW

Siendo conscientes de la importancia del estudio del recurso eólico para la apuesta por esta tecnología, como ha sido anticipado en la introducción, antes de la construcción de un parque eólico se desarrolla una campaña de medición meteorológica con la cual se pretende definir las circunstancias locales y los regímenes de viento existentes en las inmediaciones a los aerogeneradores, discerniendo variables fundamentales del diseño como la velocidad y la dirección del viento o las pérdidas energéticas originadas por efectos como la turbulencia o el ángulo de ataque de la pala del aerogenerador. Frecuentemente esta toma de datos se realiza para un periodo comprendido entre 1 o 2 años, tiempo durante el que se redacta el proyecto de construcción y se tramitan las autorizaciones administrativas. Sin embargo, ¿Es correcto la estudio del recurso eólico durante un periodo inferior a 2 años cuando el proyecto tendría un periodo de explotación de 25 años? A esta pregunta han intentado responder diversos expertos de la materia a nivel mundial.

Según Hiester y Pennel [9] uno de los aspectos fundamentales para la caracterización del recurso eólico en un emplazamiento es el reconocimiento de patrones de variación interanual en las lecturas de velocidad media del viento. En este sentido acuerdan que difícilmente la precisión es aceptable cuando se parten de series temporales con un periodo de toma de datos inferior a 10 años en la ubicación candidata. Otros autores como Baker et al. [10] son aún más contundentes. En este caso tomaron datos de estaciones meteorológicas ubicadas en diversas regiones de Estados Unidos, y realizaron un estudio de la variación interanual y la estacionalidad de la velocidad media del viento, verificando patrones de variación del 10% con una probabilidad de ocurrencia del 90%. Como resultado de ese estudio pudieron afirmar que para el desarrollo de un análisis de estas

características lo adecuado sería contar con un periodo de medida meteorológica de entre 20 y 30 años. Como es de suponer, este objetivo es prácticamente inalcanzable puesto que supondría dilatar la decisión de inversión durante un largo periodo de tiempo e incrementar desorbitadamente los costos derivados de la ejecución de la campaña de medición meteorológica.

Una alternativa viable es el empleo de métodos de predicción meteorológica a largo plazo, que tal como se adelantó en el Capítulo 1 pueden clasificarse fundamentalmente en tres grupos básicos, modelos de predicción física, técnicas de dinámica de fluido computacional (CFD) y modelos estadísticos de extrapolación temporal. El objetivo último que se pretende conseguir con estos modelos es que el resultado obtenido sea lo más similar posible a la media de velocidades de viento obtenida durante el periodo de explotación.

A efectos de la presente tesis doctoral son de especial interés los modelos estadísticos de extrapolación temporal (MCP). Una variable de este enfoque que ha tenido una evolución creciente en los últimos años es el empleo de técnicas de Machine Learning basadas fundamentalmente en la minería de datos y el uso de métodos estadísticos avanzados siendo su fuerte el desarrollo de estimaciones usando múltiples estaciones de referencia climatológica.

A continuación se describirá en qué consiste y cuáles son los principios básicos de los métodos MCP. Posteriormente, se definirán las hipótesis de partida comúnmente aceptadas según el estado del arte y los requisitos que deben cumplir los datos para su implementación en un modelo MCP. Finalmente, ya fijándose el objetivo sobre el método, se valorará el potencial existente en el empleo de técnicas de Machine Learning para el desarrollo de modelos MCP.

2.10. Empleo de técnicas MCP para el estudio del recurso eólico a largo plazo

Los métodos MCP parten del concepto de que cuando sólo se disponen de datos de viento a corto plazo en la ubicación a estudio puede desarrollarse una estimación a largo plazo usándose para ello técnicas de correlación estadísticas con los datos recabados en otras estaciones meteorológicas cercanas donde la toma de datos se hubiera llevado a cabo durante periodos superiores a 10 años.

En términos generales para el desarrollo de estos estudios el dato más importante es la velocidad del viento, la cual debe ser obtenida con una frecuencia de muestreo homogénea para toda la serie temporal. Por otra parte los datos eólicos de velocidad se suelen agrupar por sectores dependiendo de la dirección del viento, ejecutándose el modelo para cada uno de los sectores de viento en los que se halla dividido el conjunto de datos de la serie temporal.

En la Figura 16 se representa de manera simplificada la topología básica del estudio eólico empleando técnicas MCP. Para el caso representado en el diagrama se disponen de datos de velocidad y dirección del viento en dos estaciones de referencia (Reference site 1 y Reference site 2), datos que según lo citado anteriormente deberían tener una longitud mínima de 10 años. De otra

parte también se disponen de los datos recabados a corto plazo en la ubicación a estudio (sección en color rojo de la columna "Target site").

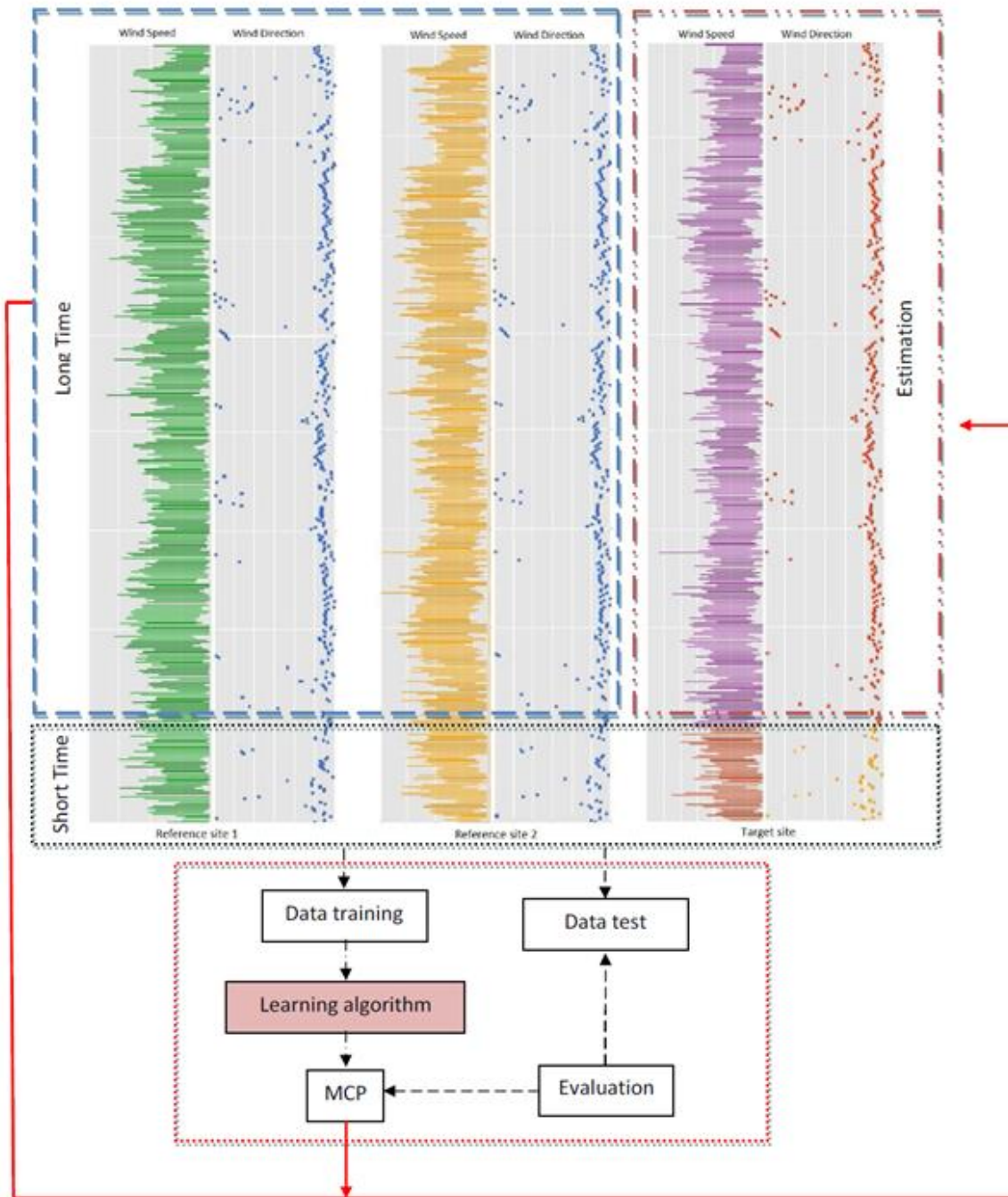


Figura 16 Arquitectura MCP

Los datos recabados en la ubicación a estudio deben ser coincidentes en tiempo con una sección de los datos de las estaciones meteorológicas de referencia, esta condición es indispensable puesto que permite verificar la correlación existente entre las distintas estaciones empleadas para la modelización matemática. La correlación entre los datos de viento no implica en sí misma una

relación de causalidad, sin embargo con este estudio se comprueba si el aumento o disminución del recurso eólico en alguna de las estaciones de referencia utilizadas para un momento determinado va acompañado sistemáticamente de comportamientos homólogos en la estación candidata. A esta etapa se la conoce como fase de entrenamiento (Data training).

En el proceso de entrenamiento comúnmente se utiliza una técnica de evaluación conocida como validación cruzada (Figura 17). Con este fin, los datos a corto plazo de todas las estaciones son divididas en partes o subconjuntos de datos de aproximadamente igual tamaño, dejando fuera una parte (denominada de validación) para evaluar la capacidad predictiva del modelo estimado y utilizando el resto (denominada más comúnmente de entrenamiento) en la estimación. Cada una de las partes está formada por un conjunto de datos tamaño similar ordenados de forma aleatoria para evitar posibles asociaciones espúreas, no obstante, para una misma hora los datos de todas las estaciones implicadas se mueven a la par, manteniéndose siempre las correlaciones existentes entre ellas. Una vez estimado el modelo de correlación con las parte de entrenamiento se utiliza la parte de validación para determinar los errores de predicción del modelo estimado. Este proceso se repite tantas veces como el número de subconjuntos en los que se ha dividido los datos, obteniéndose otros tantos subconjuntos de errores de predicción (uno por cada subconjunto de datos que se va dejando fuera del entrenamiento) y que se corresponden con otros tantos modelos estimados diferentes. Finalmente el error asignado al modelo será el promedio de todos los errores de predicción obtenidos con los diferentes subconjuntos de validación. El número de subconjuntos en que se divide el conjunto de datos depende del tamaño de la muestra disponible. En general, se pretende un compromiso entre el tamaño de los subconjuntos y el número de éstos. Si la muestra es pequeña, suelen utilizarse subconjuntos de tamaño uno, dando lugar a lo que se conoce como validación cruzada "*Leave one out*" (LOO). En este trabajo, hemos elegido dividir la muestra en 10 subconjuntos, habiendo comprobado previamente sus buenos resultados.

Como resultado del proceso anterior, cada modelo candidato de la familia de modelos considerada produce un error medio general de predicción (en los subconjuntos de validación). A continuación, se selecciona el modelo específico que presenta el menor error medio general de predicción. Dicho modelo se entrena posteriormente con la totalidad de los datos del período a corto plazo para aprovechar al máximo la información disponible. Finalmente, este modelo es el que se utiliza para la predicción a largo plazo en la ubicación candidata utilizando los datos a largo plazo de las estaciones de referencia (sección en color violeta en la columna Target site).

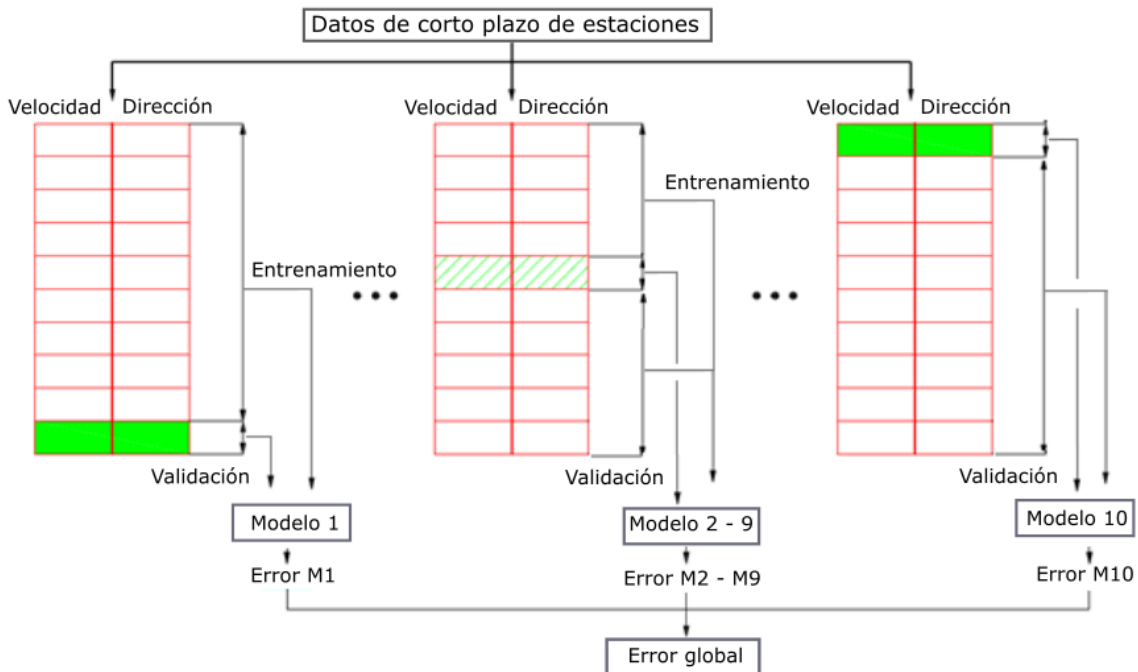


Figura 17 Validación cruzada

Uno de los objetivos fundamentales de los estudios desarrollados en esta tesis doctoral es la **búsqueda de los modelos de aprendizaje estadísticos que mejor comportamiento tenga a la estimación de la potencia eólica. Dicha evaluación ha sido centrada en el horizonte a largo plazo.**

Si bien a priori la mejor opción sería aquella en la que el error es el menor posible, a efectos prácticos debe considerarse otros aspectos como la complejidad del modelo o los tiempos de computación necesarios para su desarrollo. En respuesta a este asunto han sido utilizados **métodos de significación estadística** en los que no únicamente se tienen en cuenta los valores puntuales del error, sino que además se consideran las distribuciones de probabilidad de los resultados obtenidos.

A modo de ejemplo puede desarrollarse este método para comparar dos técnicas en función del error generado. Por una parte se define una “hipótesis nula” en la que se supone que las dos técnicas de evaluación del recurso eólico son igualmente efectivas, mientras que por otra se especifica una hipótesis alternativa donde se asume distinta efectividad, decidiéndose finalmente por una u otra en función de la compatibilidad estadística de los datos con la distribución de probabilidad existente en la hipótesis nula. En otras palabras, si las diferencias de error entre las dos alternativas evaluadas son significativas se adopta la técnica como menor error global. Sin embargo, **si los errores tienen similar efectividad se tomará como solución más adecuada la que menor tiempo de computación tenga**, independientemente de que el error generado en la otra técnica pueda ser inferior.

La mayor parte de los estudios desarrollados a nivel internacional en relación con el uso de métodos MCP han usado para el análisis una única estación meteorológica de referencia, implementando algoritmos que por lo general establecen relaciones lineales entre los datos de viento de las estaciones implicadas (algoritmos de regresión lineal). En contrapartida, tal como se analizará detalladamente en el estudio del estado del arte (Capítulo 3), diversos autores han apuntado a la

existencia de ventajas por la reducción de incertidumbre cuando se utilizan varias estaciones de referencia en lugar de una sola. Sin embargo, se deben afrontar problemas como la sobre especificación, la mayor complejidad computacional y el aumento de los tiempos requeridos para la ejecución de los modelos de extrapolación temporal.

Para la modelización de sistemas con múltiples estaciones de referencia comúnmente se suelen utilizar técnicas de aprendizaje automático las cuales permiten la ejecución del modelo con grandes cantidades de datos. Quizás de las técnicas utilizadas hasta el momento la que mayor aceptación ha tenido en la comunidad científica han sido la Red Neuronal Artificial. No obstante, ésta no es la única técnica existente ya que han sido publicados resultados de investigación con otras técnicas como las máquinas de vector soporte, las redes bayesianas o los árboles de decisión.

2.11. Requisitos de los datos de partida para la estimación del recurso eólico a largo plazo empleando métodos MCP

Como en cualquier modelo matemático, la calidad de los resultados es directamente dependiente de los datos con los que se alimenta el modelo MCP. En este sentido, una amplia relación de autores han valorado cómo afecta a los resultados del modelo cambios en aspectos fundamentales como la correlación mínima necesaria entre la estación candidata y la estación de referencia o la amplitud mínima de las series temporales y la frecuencia de muestreo de los datos. Estos y otros aspectos serán tratados en este apartado relacionándose las hipótesis de partida comúnmente aceptadas según el estado de la técnica actual.

2.11.1. Correlación entre los datos de la estación candidata y la estación de referencia

A la hora de ejecutar un modelo de extrapolación temporal es necesario cuestionarse si los datos tomados de partida como estación de referencia son representativos de las condiciones de viento que pudieran generarse en la ubicación candidata. En este sentido diversos estudios han tratado de concretar la mínima correlación que debería existir entre la estación de referencia y la estación candidata tomando de base aspectos como la escala de tiempo de las series temporales o la distancia existente entre dichas estaciones meteorológicas.

En el año 2000 Ayotte et al. [99] publica un análisis por el cual tomando datos de varias estaciones meteorológicas distanciadas entre ellas regularmente entre 1 y 100 km valida la evolución de la correlación entre estaciones conforme al tiempo y el espacio. Con respecto al tiempo determina que estacionalmente se producen cambios considerables en las velocidades medias del viento que logran estabilizarse únicamente después de promediar sus valores en periodos anuales. De igual forma comprueba que para su caso particular el error tiende a estabilizarse con medias anuales de velocidad del viento superiores a 5 años si bien reconoce que lo ideal sería contar con series mucho

más largas de hasta 20 años. Por otra parte detecta una gran relación de dependencia entre la distancia y la frecuencia de los datos de partida de las estaciones meteorológicas implicadas puesto que a medida que aumentaba la distancia desarrollaba distintas simulaciones con frecuencias temporales mayores verificando que el error descendía gradualmente hasta un cierto momento en el que alcanzaba el mínimo posible. En cualquier caso apunta que los resultados obtenidos obedecen a las características locales de los emplazamientos analizados y que para otras ubicaciones la relación de dependencia entre distancia y frecuencia puede ser diferente.

A conclusiones semejantes apuntó Nielsen et al. [100] en el año 2012 cuando expuso que previo a la aplicación de un método MCP debe considerarse la escala de tiempo usada para la realización de las medias de velocidad del viento, existiendo una relación de dependencia entre el tiempo y la distancia. Así pues si para un estudio de estas características se adoptaran medias de velocidad de viento con periodos de muestreo inferiores a 10 minutos cuando las estaciones se encuentran considerablemente distanciadas, es probable que la correlación existente entre esas estaciones sea baja. Conforme a lo anterior ciertas publicaciones [13,101,102] han sugerido que cuando en el análisis de la relación existente entre las series temporales de la estación de referencia y candidata arrojan bajos coeficientes de correlación, quizás sea conveniente valorar nuevamente su semejanza cambiando la frecuencia temporal de los datos de velocidad y dirección del viento.

2.11.2. Amplitud mínima de los datos de la estación candidata

Otro aspecto de gran importancia concordante con lo anterior que ha sido valorado por diversos autores en relación con la implementación de modelos de extrapolación temporal es la amplitud mínima de los datos meteorológicos de la estación candidata, los cuales deben ser a la par coincidentes en tiempo con una parte de los datos de la estación de referencia para su correlación y el entrenamiento de los algoritmos involucrados.

En el año 2004 Taylor et al. [103] presentan en una conferencia un estudio analítico de las variables que afectan en la estimación a largo plazo del recurso eólico a través de métodos MCP. Para ello realizan diferentes simulaciones en las que varían el periodo temporal de datos coincidentes entre las estaciones candidata y de referencia para rangos de 1, 3, 6, 12, 18 y 24 meses. Con sus resultados demostraron que la incertidumbre se iba reduciendo progresivamente a medida que aumentaba el periodo temporal considerado, pasándose de una desviación estándar relativa de hasta incluso el 12% cuando se tomaban datos de 1 a 3 meses, al 2% cuando los periodos considerados se encontraban entre los 18 y 24 meses. Según los autores, los resultados manifestaban una clara evidencia de que existe una influencia considerable en la variación estacional del recurso eólico a largo plazo. De la misma forma no recomendaban el desarrollo de estudios de estas características con datos para menos de 12 meses en la estación candidata.

Ya en el año 2010 Oliver y Zarling [104] presentan otro estudio de similares características. En éste toman datos de 14 pares de estaciones meteorológicas candidatas y de referencia situadas en Estados Unidos, concluyendo que con datos inferiores a un año se generan errores considerables derivados de la estacionalidad del recurso eólico. Esto les lleva a su vez a darse cuenta de la

importancia que tiene el ser conscientes del momento exacto en el que un promotor debe iniciar la ejecución de una campaña de medición meteorológica, ya que si por cuestiones de tramitación o de cumplimiento de la línea temporal del proyecto desarrolla el modelo con medidas inferiores a un año se produciría un sesgo importante en el estudio por las razones de estacionalidad citadas anteriormente, lo que pondría en entredicho la validez de los estudios desarrollados.

2.11.3. Estabilidad climática a largo plazo

Otro problema en la ejecución de modelos MCP que ha sido objeto a estudio es la conocida como estabilidad climática, condición inherente de los métodos estadísticos empleados los cuales ignoran los efectos del cambio climático en el desarrollo de las estimaciones de recurso eólico, lo que puede significar un importante error en problemas de estimación del recurso eólico a largo plazo.

Algunos autores como Michaelides [105], Emesis [106] o Diandong [107] han apuntado en sus publicaciones que existe una clara relación entre el cambio climático y el viento. Tal como lo describen, las diferencias de temperaturas que se producirán entre el ecuador y los polos provocarían a largo plazo una reducción de la velocidad media del viento que afectaría sobre todo a ubicaciones próximas a latitudes medias. Por el contrario Pryor y Barthelmie [108] publicaron en el año 2009 una revisión de los trabajos de investigación desarrollados hasta ese momento en el estudio de impacto del cambio climático sobre el viento, concluyendo que no existían evidencias claras que suscitasen que las reducciones de potencial eólico originadas en Europa en las últimas décadas fueran como consecuencia del cambio climático, si bien la mayor parte de dichos estudios no eran concluyentes, razón por la que señalaban la necesidad de que fueran llevados a cabo nuevas investigaciones que alabaran fehacientemente dichas proyecciones.

Otros autores han apuntado a condiciones relacionadas con las características locales de los emplazamientos. En el año 2010 Vautard et al. [109] publican un estudio en el que tomando datos de 822 estaciones meteorológicas de superficies ubicadas en diferentes zonas del hemisferio norte, detectan una reducción de la velocidad media del viento de entre el 5% y el 15% para los años comprendidos entre 1979 y 2008. Por otra parte verifican que los mayores cambios se producen en latitudes medias del norte y principalmente para los vientos superficiales más intensos, mientras que en las simulaciones realizadas a nivel del mar no se detectaban cambios significativos. Con todo ello concluyen que una de las principales razones de esta reducción en el recurso eólico se deriva del aumento de la rugosidad superficial, puesto que las regiones más afectadas han sido aquellas en las que el nivel de la biomasa forestal ha incrementado durante los últimos 30 años.

Independientemente de la controversia existente en este tema, a efectos de evaluar su importancia en la estimación del recurso eólico a largo plazo, lo conveniente sería su estudio para un horizonte temporal semejante a la vida útil media de un parque eólico, si bien prácticamente la totalidad de los estudios desarrollados evalúan los cambios para periodos superiores a 50 años e incluso los 100 años [110,111]. Por otra parte tal como señalan Pryor y Barthelmie, los pocos estudios llevados a cabo hasta el momento que han analizado los cambios entre años e incluso décadas del recurso eólico muestran que las variaciones han sido mínimas en sus patrones de comportamiento, y que apenas

han afectado a la explotación de este recurso con fines energéticos [112]. En cualquier caso algunos autores han cuestionado estos resultados puesto que se plantean si realmente los modelos meteorológicos usados hoy día son adecuados para replicar la variabilidad histórica del recurso eólico [108].

2.11.4. Adecuación de los protocolos de medida

También ha sido de interés el análisis de cómo afecta a la estimación del recurso eólico el uso de datos obtenidos a través de campañas de medición meteorológica inadecuadas. Como parece lógico, para llegar a comprender con precisión el recurso eólico existente en un emplazamiento es fundamental el uso de estaciones meteorológicas, siendo la validez de los datos obtenidos dependientes de su ubicación y la correcta instalación de los medios técnicos implicados y el tiempo que dura la campaña de medición meteorológica, aspectos que no deben ser considerados a la ligera.

Carta et al. [13] publican un review en el año 2013 en el que entre otros aspectos evalúan los factores determinantes que influyen en el desarrollo de modelos MCP. En lo relativo a los protocolos de medida señala que a la hora de seleccionar el emplazamiento donde ubicar la estación meteorológica es vital tener en cuenta las afecciones que puedan provocar otros obstáculos en sus inmediaciones, e incluso más importante aún que durante la duración de la campaña de medida no se produzcan alteraciones en el medio que originen nuevas distorsiones. Teniendo en cuenta que la estación candidata se ubica en las inmediaciones del parque eólico a estudio, existirá una relación entre la ubicación de la estación y las pérdidas de potencias que puedan provocarse por la influencia de obstáculos cercanos. Este aspecto fue estudiado por Ruck y Gruber [52] donde haciendo uso de simulaciones con túnel de viento determinan que la distancia mínima entre un aerogenerador y un obstáculo (en este caso singular una colina) debía ser como mínimo de 7.5 veces la altura del obstáculo para que las pérdidas provocadas por éste sean insignificantes. En cualquier caso, los mayores problemas siempre se originan en las estaciones de referencia puesto que las campañas de medida duran muchos años y a los problemas de obstáculos se suma la mayor probabilidad de que el mantenimiento de dichas instalaciones sea inadecuado.

Los autores del review también exponen y analizan las afirmaciones desarrolladas por Probst y Cárdenas [113] en el año 2010 cuando concluyen que necesariamente para la correlación de datos entre estaciones es necesario que las alturas de medida sean coincidentes. Por otra parte si esto no fuera posible la solución más acertada sería ajustar la altura de la estación candidata a la estación de referencia haciendo uso de las leyes de extrapolación vertical (ley de distribución logarítmica y ley de distribución de potencia), y tras haber sido aplicado el método MCP y obtenido la serie temporal a largo plazo, escalarlo de nuevo a la altura de la estación candidata. Partiendo de dichas afirmaciones los autores del review desarrollan diferentes simulaciones en las que variando las alturas entre estaciones analizan la correlación cruzada obtenida para un caso singular situado en la isla de Gran Canaria. Conforme a los análisis desarrollados determinan que las diferencias de correlación cuando se altera la altura de la estación candidata con respecto a la estación de referencia no son significativas, sin embargo sí observan que la peor correlación se obtiene cuando los datos de la estación de referencia son de 10 metros y los de la estación candidata son de 60 metros, lo cual los

autores consideran normal por aspectos relacionados con las turbulencias causadas por la rugosidad superficial a nivel del suelo. Además señalan que no siempre existe disponibilidad de los datos necesarios para el cálculo del perfil vertical de viento a lo que se suma que dichas leyes comúnmente sólo tienen en cuenta los regímenes atmosféricos neutros [73].

2.11.5. Alternativas al empleo de estaciones meteorológicas de superficie

Una alternativa al empleo de estaciones meteorológicas de superficie como estaciones de referencia es el uso de datos de reanálisis. Estas bases de datos están compuestas por una red tridimensional de datos meteorológicos históricos generados por organismos como el NCEP (US National Center for Environmental Prediction) y el NCAR (National Center for Atmospheric Research) [114], las cuales integran las condiciones meteorológicas a macroescala recogidas por sistemas como satélites, globos sondas o estaciones de superficie, durante un periodo de 15 años para una frecuencia de 6 horas a distintas alturas.

Algunos autores como Brower [13,115] han argumentado que la ventaja esencial de este recurso radica en que es de libre acceso y se desarrolla para un mallado regular por lo que se puede seleccionar la más cercana a la posición a estudio. Además no siempre es posible acceder a una estación meteorológica con datos recabados durante más de 10 años, por lo que en ocasiones es la única posibilidad para el desarrollo de estudios de estimación del recurso eólico a largo plazo. Sin embargo Brower es consciente de que estos datos no son lo suficientemente válidos para su uso en métodos MCP. En el mismo coloquio Pinto et al. [13,116] manifiesta que ellos habían hecho un estudio comparativo entre los datos de reanálisis NCEP/NCAR y 20 estaciones de medidas ubicadas a lo largo de Portugal, obteniéndose que los datos cada 6 horas de la fuente de reanálisis tenía una escasa correlación con los datos recogidos en las estaciones locales, sin embargo cuando los datos se comparaban en periodos mensuales o anuales la correlación tendía a mejorar.

En el año 2011 Liléo y Petrik [13,117] exponen los resultados de una investigación en la que usan datos de reanálisis como estaciones de referencia para implementar un modelo MCP. De acuerdo con sus deducciones concluyen que el uso conjunto de referencias procedentes de las bases de datos NCEP/CFR y MERRA permitieron una mejora en la precisión de los resultados del modelo en comparación con el empleo de la base de datos NCEP/NCAR, derivada de la mejor representación de las condiciones climáticas locales con un perfil espacial y temporal de mayor resolución. Para la validación de los resultados usaron los datos recabados por anemómetros ubicados en superficie.

No hay que perder de vista que este tipo de bases de datos son claves para algunos modelos físicos como OpenWind [39]. Conforme al método de cálculo usado en este programa es necesario un mapa eólico el cual se genera a través de un modelo meteorológico de mesoescala NWP (Numerical Weather Prediction) el cual simula de manera iterativa ajustando las condiciones atmosféricas de capa posición geográfica a través de este tipo de datos y reduciendo sucesivamente el mallado utilizado desde los 30 km hasta los 1.2 km. Posteriormente se añaden las influencias locales del emplazamiento como la topografía de la región y la rugosidad existente. De todas formas los desarrolladores son conscientes de que estas simulaciones pueden contener errores y en ningún

caso deben sustituir las mediciones directas en el emplazamiento del parque eólico, en este sentido han añadido un proceso conocido como ajuste al mástil en el que tanto el mapa eólico como los mapas de incertidumbre se corrigen en función de los datos medidos directamente en el emplazamiento [118].

2.11.6. Influencia de la dirección del viento en la caracterización del modelo

También ha sido discutido en diferentes publicaciones los datos de partida con los que debe alimentarse un modelo de estas características. Conforme al análisis del estado del arte prácticamente la totalidad de los modelos de evaluación del recurso eólico evitan trabajar únicamente con datos de velocidad del viento puesto que se ha observado una relación de dependencia entre la velocidad del viento y su dirección angular. Tal como analizan algunos autores [13,17,35] la implementación de esta variable permite la clasificación de las velocidades del viento en función de su dirección, lo que supondría que la mayor parte de las medidas registradas en un mismo sector hayan sido afectadas de igual forma por las características locales del emplazamiento. Esta alternativa ha demostrado una reducción del error considerable en comparación con el empleo de datos de velocidades de viento brutos [17], sin embargo en emplazamientos donde la componente predominante del viento está muy marcada, se generan problemas de falta de datos en otros sectores donde los vientos son esporádicos.

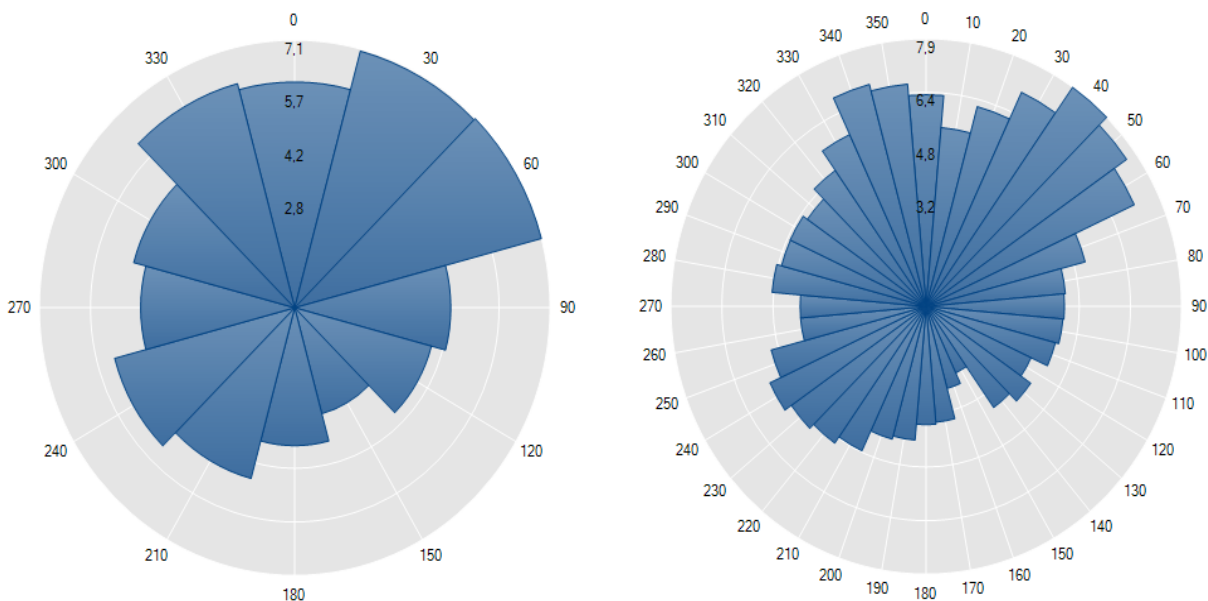


Figura 18 Distribución de velocidades medias del viento por sectores de dirección

Para solucionar el problema de falta de datos por sectores algunos autores como Woods y Watson [119] han propuesto métodos por los cuales varían el número de sectores de dirección del viento en función de la cantidad de datos de velocidad existente en cada sector. En sus conclusiones señalan que el método empleado adquiere mayor potencial en posiciones de orografía compleja obteniéndose menores errores que otros modelos estándar. Otros métodos van más allá proponiéndose soluciones por las cuales se optimizan el número de sectores de dirección del viento y

sus posiciones en función de los coeficientes de correlación obtenidos entre la velocidad de las distintas estaciones [13,120].

A lo anterior hay que añadir que previo al uso de las series temporales suele ser corriente el desarrollo de una fase de filtrado por la cual se descartan del análisis los datos donde existen evidencias de errores o son despreciables a efectos energéticos. Determinados métodos como el presentado a principios de 2015 por Zheng et al. [121] han permiten a través del uso de la minería de datos el cribado de aquellos datos no susceptibles de ser viables para la generación de potencia en parques eólicos o incluso detectar posibles fallos en sensores conforme a las lecturas redundantes existentes por estación meteorológica y la generación del perfil de vientos. En relación con los datos no susceptibles de generar potencia, las propias curvas de potencia aportadas por los fabricantes de aerogeneradores no consideran que se genere energía para vientos con velocidades por debajo de los $3 - 4 \text{ m s}^{-1}$ asumiendo dicho valor como velocidad de arranque [122,123].

Según Llobart et al. [124] para la ejecución de filtrados el método más recurrente es el conocido como estimador de la menor media de los cuadrados (LMS – Siglas en Inglés de Least Mean of Squares) el cual interpreta que todos los datos pueden ser tratados con el mismo modelo y genera problemas cuando se encuentra con valores atípicos ya que un único punto puede destruir el filtrado. En este sentido propone un nuevo método (LMedS – Siglas en Inglés de Least Median Square) semejante al anterior pero en el que se elimina el sumatorio por una mediana dejando sin efecto a los puntos atípicos. Por otra parte propone la eliminación manual de los datos atípicos con una técnica gráfica lo que evita el uso de simulaciones iterativas reduciendo el tiempo de filtrado en un 20% según son conclusiones.

2.12. Potencial de las técnicas de Machine Learning para el estudio del recurso eólico

Según Witten et al. [125] el desarrollo de las tecnologías de la información acaecido durante las últimas tres décadas ha permitido el incremento de la capacidad de almacenamiento de información a través del uso de bases de datos, sin embargo a medida que aumenta la cantidad de datos almacenada disminuye nuestra comprensión sobre la relevancia de estos datos para describir un fenómeno o desarrollar una determinada predicción. En cualquier caso, si bien no somos capaces de identificar el proceso completamente, si somos conscientes de que dicho conjunto de datos no son normalmente arbitrarios y que existen una serie de patrones de comportamiento regulares a lo largo de la serie.

En este marco surge el concepto de minería de datos, procedimiento por el cual a través del uso de técnicas de Machine Learning se fuerza a la búsqueda de patrones estructurales de información en una base de datos, permitiendo la criba de información entre la que se supone que tiene contenido significativo y debe ser tomada a efectos de decisión, de la que no aporta contenido relevante y debe ser aislada del resto.

A su vez las técnicas de Machine Learning se desarrollan a través de **teorías estadísticas de aprendizaje** con las cuales se construyen modelos matemáticos de los que surgen las estimaciones. Hasta finales de los años setenta, la mayor parte de las técnicas estadísticas de aprendizaje utilizadas se basaban en el empleo de métodos lineales, los cuales debido a su simplicidad no son capaces de representar con precisión la relación entre las variables de entrada y salida de un modelo sofisticado a medida que éste fuera cada vez más complejo. Posteriormente, a partir de los años ochenta, los avances en las tecnologías de computación proporcionaron los recursos necesarios para la ejecución de modelos no lineales, técnicas que han demostrado según James et al. [126] un amplio potencial para su implementación práctica en numerosos campos del conocimiento.

No es hasta los años noventa cuando comienza a potenciarse el empleo de técnicas de Machine Learning para propósitos relacionados con las energías renovables. En el año 2001 Kalogirou [127] publica una revisión de los trabajos que habían sido desarrollados hasta ese momento aplicando un tipo de técnica de Machine Learning concreta, la Red Neuronal Artificial (ANN – Siglas en Inglés de Artificial Neural Networks). Esta técnica se basa en el comportamiento de una neurona biológica cuya estructura basada en una entrada, un proceso interno y una salida como respuesta, permite reproducir las relaciones existentes entre un conjunto de variables, una vez sometida a una estrategia de aprendizaje. Según Soteris, la red neuronal había sido utilizada para la modelización de procesos tales como la previsión del recurso, fundamentalmente solar y eólico, la proyección de la demanda y la evaluación de las condiciones económicas asociadas a la energía. De la misma forma basándose en la revisión que desarrolló en su trabajo concluyó que esta metodología presentaba una amplia potencialidad que no debería ser subestimada.

En el campo de estudio del recurso eólico la evolución en el empleo de técnicas de Machine Learning ha sido creciente desde el año 2000 con numerosos artículos publicados de alta relevancia. En el año 2003 Landberg et al. [128] publica una revisión de los diferentes métodos existentes para la estimación del recurso eólico en un emplazamiento, enunciando que una de las ventajas fundamentales en el empleo de técnicas de Machine Learning radica en su capacidad para resolver problemas multivariable donde los métodos estadísticos lineales no logran correlaciones apropiadas. A conclusiones semejantes habían llegado antes Gardner y Dorling [129] en el año 1998 aplicando redes neuronales artificiales al campo de las ciencias atmosféricas y recomendando su uso para tareas de tipo predictivas, clasificatorias o en las que se busca una función objetivo concreta.

También en el año 2000 Addison et al. [130] reflexionan sobre las ventajas del empleo de las técnicas de Machine Learning con respecto a modelos de estimación física, tomando como caso particular el paquete informático de estudio del recurso eólico líder del mercado, WAsP [37]. Según analizan, los modelos de predicción física requieren para su ejecución variables relacionadas con la topografía y la simulación de los fenómenos físicos complejos que originan los cambios de condiciones en el recurso eólico de un emplazamiento determinado. Este problema se agrava en terrenos complejos incrementando considerablemente la incertidumbre generada, hecho por el cual los desarrolladores de WAsP han diferenciado entre orografía simple donde se puede emplear sus modelos básicos basados fundamentalmente en las ecuaciones lineales de Jackson – Hunt, de estimaciones para orografía compleja las cuales ejecutan con técnicas de dinámica de fluido computacional (CFD). Si

bien con esta solución se logra mejorar la calidad de las estimaciones, el modelo CFD requiere el uso de mayores requisitos computacionales incrementando el coste del proyecto [131]. Por otra parte no está del todo claro a partir de qué momento debe considerarse a un terreno como orografía simple o compleja.

Por el contrario las técnicas de Machine Learning, utilizadas como modelo estadístico de extrapolación temporal, independientemente de las condiciones topográficas del terreno, se centran en verificar la validez de los datos históricos de viento de una estación de referencia, correlacionando los datos con los recabados a corto plazo en la posición a estudio. Esta es la principal razón por la que autores como Prasad y Bansal [12] han concluido que el empleo de este tipo de técnicas es más conveniente cuanto mayor la complejidad topográfica, siendo a la par una herramienta con un mayor grado de estandarización.

En el análisis del estado del arte se describirán los principales avances que han sido desarrollados en el campo de la estimación del eólico a largo plazo empleando estaciones de referencia múltiples y empleándose para la mayoría técnicas de Machine Learning tales como las redes neuronales artificiales, las redes bayesianas o las máquinas de vector soporte entre otras.

Técnicas de Machine Learning para el estudio del potencial eólico a largo plazo

3.1. Introducción

Con el análisis desarrollado hasta el momento se puede extraer la conclusión de que los métodos MCP son una alternativa potente para paliar la falta de datos en el estudio del potencial eólico de una ubicación candidata, recurriéndose a técnicas estadísticas con las cuales las series temporales a corto plazo son ajustadas a los datos registrados en una o varias estaciones de referencia posicionadas en cercanías.

De la variedad de métodos MCP publicados hasta el momento para el estudio del recurso eólico, una amplia mayoría se han apoyado en el uso de una única estación climatológica de referencia, implementándose para ello algoritmos de relación lineal entre los datos de la estación candidata y la de referencia [13,49,94,128,132-137]. Sin embargo, diversos autores han apuntado a que existen beneficios cuando se usan varias estaciones de referencia al mismo tiempo, ya que éstas pueden captar diferentes aspectos del recurso eólico en una ubicación objetivo al no dar por hecho que la distribución de direcciones del viento en la estación candidata es semejante a la existente en una única estación de referencia [94,113,130,138]. En cualquier caso, autores como Brower [94] sugieren de la existencia de un número máximo de estaciones de referencia por análisis ya que existe el riesgo de sobre especificación por el cual el modelo pierde flexibilidad y es incapaz de adaptarse a los datos aportados. Algunos autores como Carta et al. [45] han implementado técnicas de selección de parámetros característicos con las que en función del error estimado determinan qué variables y

estaciones son utilizadas para el análisis, reduciéndose el error en el 100% de los casos a costa de un aumento de la capacidad computacional.

Entre las técnicas empleadas hasta el momento para la resolución de problemas de regresión con el enfoque de múltiples estaciones de referencia pueden mencionarse algoritmos de regresión múltiple como el desarrollado por Walls et al. [139], el método J-tris desarrollado por Casella [140], la técnica Multiple Climatic Reduction Technique (MCRT) [141-145] y otras técnicas de minería de datos [125,146] que usan algoritmos de aprendizaje estadístico como las Redes Bayesianas [16] y las Redes Neuronales Artificiales (ANN) [11,17,20,27-29,33,35,147]. De la misma forma, del análisis realizado se extrae la conclusión de que existen otras técnicas de Machine Learning que, sin haber sido utilizadas para el estudio del recurso eólico a largo plazo, si habían sido empleadas con éxito en otros campos del conocimiento afines manifestando su potencialidad, entre ellas las máquinas de vector soporte para regresión (SVR) y los árboles de regresión (RT).

Como ha sido anticipado en el Capítulo 1, el estudio realizado en esta tesis doctoral estará centrado en tres técnicas de Machine Learning fundamentales, ANN, la cual ostenta la condición de referencia en este campo de estudio, y las técnicas SVR y RF, las cuales sí habían sido usadas en otros campos del conocimiento pero no para los fines perseguidos en este trabajo. Partiendo de estas premisas se presentan inicialmente en este capítulo los principios básicos de cada técnica únicamente exponiéndose las variantes que han sido implementadas para el desarrollo del trabajo. Para un estudio detallado de cada técnica, el lector puede dirigirse a [148-150] para ANN, [150,151] para SVR y [150,152] para RF. Adicionalmente, en la segunda sección de este capítulo se expone la metodología que ha sido seguida para la implementación de las técnicas desarrolladas.

3.2. Enfoque de regresión para la resolución de problemas MCP

Por término general, para la resolución de problemas MCP se utiliza un enfoque de regresión múltiple. En este sentido, la función de regresión Ecuación 3.1 debe ser resuelta utilizando las tres técnicas ML seleccionadas o el método Least Squares Multiple Linear Regression, el cual como se explicará en el siguiente subapartado también será usado en este trabajo a modo de referencia en la comparación con métodos lineales.

$$Y_i = f(x_i) + \varepsilon_i, i = 1, \dots, n \quad (3.1)$$

En Ecuación 3.1, para una muestra de datos $T = \{(x_i, y_i), i = 1, \dots, n\}$, $x_i = (x_1, \dots, x_d)^T \in \chi \subset \mathbb{R}^d$ es cada observación de las variables independientes $\mathbf{X} = (X_1, \dots, X_d)^T$ e y_i es la correspondiente observación de la variable dependiente o respuesta Y . Asimismo, ε_i es una variable aleatoria continua que representa el ruido aleatorio que pudiera estar asociado a cada observación de la variable respuesta.

El problema que se pretende resolver radica en estimar la esperanza condicionada $f(\mathbf{x}) = E(Y|\mathbf{X} = \mathbf{x})$. Es decir, el objetivo principal en el problema de regresión es estimar la función $f(x)$ basada en un conjunto de entrenamiento dado T , con el propósito de aproximar y en futuras observaciones de x .

3.3. Least Squares Multiple Linear Regression

La técnica Least Squares MLR es usada en este trabajo a modo de referencia en la comparación con una técnica ML concreta, en este caso la SVR. Dicho estudio ha sido llevado a cabo en el Apartado 4.2, habiéndose escogido dicha alternativa al ser una opción lineal apta para el cálculo cuando se cuenta con varias variables independientes $\mathbf{X} = (X_1, \dots, X_d)^T$.

Partiendo de la Ecuación 3.1, el modelo de regresión lineal multivariante [153] estima la función de regresión como una combinación lineal de las variables independientes:

$$f_{\boldsymbol{\beta}}(\mathbf{x}) = E(Y|\mathbf{X} = \mathbf{x}) = \boldsymbol{\beta}_0 + \sum_{j=1}^d \boldsymbol{\beta}_j x_j = \mathbf{x}^T \boldsymbol{\beta} \quad (3.2)$$

donde $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_d)^T \in \mathbb{R}^{d+1}$, y $\mathbf{X} = (1, X_1, \dots, X_d)^T$. Si se consideran todos los datos de la muestra, el modelo de regresión se escribe como:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \quad (3.3)$$

donde $\mathbf{Y} = (Y_1, \dots, Y_n)^T$, \mathbf{X} es una matriz $n \times (d + 1)$ con primera columna de unos y demás columnas $\mathbf{X}_j = (x_{1j}, \dots, x_{nj})^T, j = 1, \dots, d$.

La estimación del vector de parámetros $\boldsymbol{\beta}$ se realiza por mínimos cuadrados:

$$\min_{\boldsymbol{\beta}} \left\{ \sum_{i=1}^n \ell(y_i, f_{\boldsymbol{\beta}}(\mathbf{x}_i)) \right\} = \min_{\boldsymbol{\beta}} \left\{ \sum_{i=1}^n (y_i - f_{\boldsymbol{\beta}}(\mathbf{x}_i))^2 \right\} \quad (3.4)$$

Donde $\ell(y_i, f(\mathbf{x}_i)) = (y_i - f(\mathbf{x}_i))^2$ es la pérdida cuadrática, que resulta de suponer ruido gaussiano independiente y de media cero en el modelo (Ecuación 3.1). La solución del problema Ecuación 3.4 viene dada por $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}$ con lo que sus predicciones en la muestra son $\hat{\mathbf{Y}} = \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}$.

Este es uno de los métodos usados con mayor frecuencia por su simplicidad, razón por la cual dicha función se encuentra en paquetes estadísticos básicos como Stats [154] para el caso del lenguaje multiplataforma R Statistics (función `lm()`).

3.4. Artificial Neural Networks (ANN)

La variante de red neuronal seleccionada para el desarrollo del trabajo presentado en esta tesis doctoral es conocida como MultiLayer Perceptron (MLP) con conexión hacia delante (feedforward), la cual puede aproximar cualquier función continua, lineal o no lineal, (propiedad de aproximación universal) siempre que posea como mínimo una capa oculta de neuronas [149]. En la mencionada arquitectura sus neuronas se agrupan en tres tipos diferentes de capas:

- a) capa de entrada (input layer), encargada únicamente de recibir las features de entrada y propagarlas al siguiente nivel,
- b) capas ocultas (hidden layers), cuyas neuronas llevan a cabo un procesamiento no lineal de las señales recibidas,
- c) capa de salida (output layer) que es la última capa y proporciona la respuesta de la MLP para cada uno de los vectores de señales de entrada.

La estructura MLP anteriormente mencionada dispondrá en este estudio de tres capas de neuronas (una capa de entrada con d nodos, una capa de neuronas ocultas con m nodos y una capa de salida con un único nodo), la cual ha sido la más frecuentemente utilizada en métodos MCP [13] y la única utilizada en el caso de usar el método MCP para la estimación de potencia en un sitio objetivo [11,13,17,43-45].

La salida de la arquitectura MLP anteriormente especificada vendrá dada por la Ecuación 3.5.

$$f(\mathbf{x}) = g \left(\alpha_0 + \sum_{j=1}^m w_j h_j \left(\beta_j + \sum_{k=1}^d w_{kj} x_k \right) \right) \quad (3.5)$$

En Ecuación 3.5 w_{kj} son los pesos de las conexiones de las d neuronas de entrada con las m neuronas de la capa oculta con bias β_j , w_j son los pesos de las conexiones de las m neuronas ocultas con la neurona de salida con bias α_0 . Por otro lado, $h_j(\cdot)$ y $g_j(\cdot)$ son funciones activación para los nodos de la capa oculta y para el nodo de la capa de salida, respectivamente. Las funciones de activación $h_j(\cdot)$ tomadas en este trabajo tienen forma sigmoideal, mientras que la función $g_j(\cdot)$ se ha tomado lineal, como es usual en la arquitectura MLP.

El algoritmo backpropagation [148,149,155] es utilizado para el aprendizaje de la red, es decir para estimar los parámetros desconocidos de la misma (pesos y bias). Mediante este mecanismo se van adaptando y modificando todos los parámetros de la red con el objetivo de que las entradas produzcan las salidas deseadas. Es decir, con el propósito de minimizar la función error. Para la programación de los modelos MCP basados en ANN utilizados en este trabajo se ha utilizado el paquete nnet [156] del software multiplataforma de licencia libre R Statistics [154]. Se hace un análisis con mayor profundidad acerca de su implementación en el Apartado 3.7.3.

A continuación se citan los trabajos publicados hasta el momento en relación con el empleo de arquitecturas ANN para la estimación del potencial eólico a largo plazo.

En el año 2004 Bechrakis et al. [29] presentan un estudio en el que utilizando una red neuronal artificial desarrollan una extrapolación de mediciones de velocidad del viento de una posición a otro emplazamiento con datos incompletos. A continuación prueban el algoritmo con datos de tres estaciones de medidas ubicadas en el Sureste de Irlanda, obteniéndose una correlación alta que marcaba la viabilidad técnica de este proceso con unos requisitos previos de información reducidos. El problema que presentaba este estudio era los datos tomados de partida para la ejecución del modelo, puesto que a pesar de que el autor era consciente de que una evaluación de estas características requiere del uso de estaciones de referencia con periodos de medida superiores a 10 años, sólo disponía de estaciones con un año de medida, usando para el entrenamiento dos meses de la estación candidata y generando una serie anual.

Posteriormente, en el año 2006 Öztopal [27] utiliza la metodología ANN para la estimación del potencial eólico valiéndose de 10 estaciones de medida ubicadas en el Noroeste de Turquía. En este estudio se detecta la existencia de mayores niveles de correlación en los meses de invierno mientras que en los meses del periodo estival los efectos de la topografía inferían en los resultados. Los resultados fueron posteriormente comparados con las estimaciones desarrolladas mediante un modelo TPCSV (*Trigonometric Point Cumulative Semivariograms*) concluyéndose que es posible evitar algunos problemas típicos de estimación con el uso de la metodología ANN.

Dos años después López et al. [35] presentan un método para la estimación de la velocidad media anual del viento usando una serie temporal de datos medidos en el emplazamiento a estudio y los datos recabados en varias estaciones meteorológicas cercanas con mayor precisión. Para ello emplean la arquitectura ANN con el algoritmo de aprendizaje de regulación Bayesiana disponiéndose de una única capa oculta y 15 neuronas. Entre sus principales conclusiones señalaron que como mínimo es necesario contar con datos de dirección y velocidad del viento para una estación de referencia. Por otra parte, los datos de dirección del viento son vitales cuando la estación candidata se ubica inmersa en un terreno de orografía compleja, reduciéndose el error RMS (*Root Mean Square*) según sus cálculos para este caso particular en un 23%. El error medio obtenido con la simulación se situó sobre el 2% conforme a las series temporales de validación usadas.

Al año siguiente (2010) se publican dos estudios también relativos al uso de modelos ANN. Por un lado Butler [36] sostiene en su artículo que las técnicas de correlación lineales clásicas, si bien han sido un importante paso para el desarrollo de los estudios de recurso eólico, su utilidad se reduce cuanto mayor es la complejidad existente en el medio. En este sentido, propone el uso de arquitecturas ANN con las que definitivamente demuestra una mejora de la precisión cuantificada en una reducción de tres veces el error AEP (*Annual Energy Prediction*). Por otra parte, Fadare [18] utilizando un modelo ANN con una única capa oculta y un algoritmo de aprendizaje Feed-Forward Back Propagation (FFBP) simuló el perfil de vientos de Nigeria bajo distintas configuraciones. Como datos de partida se tomaron 28 estaciones meteorológicas con datos medidos a 10 metros durante 20 años, usando 18 estaciones para el entrenamiento y 10 para la fase de testeo final. Finalmente el

estudio demostró que el modelo implementado tenía una precisión aceptable con un error absoluto porcentual promedio (MAPE – Siglas en Inglés Mean de Absolute Percentage Error) del 8,9%. Los resultados finales fueron traducidos a unos mapas mensuales de velocidades de viento medio por emplazamiento para Nigeria.

Ya en el año 2011 Velázquez et al. [17] verificaron la relación existente entre los datos de partida con los que se alimenta un modelo ANN y la reducción del error en la salida. Entre sus conclusiones destaca que independientemente de la correlación que manifieste el dato de entrada con respecto a la salida, cuantas más estaciones de referencia se utilizan, menores errores se producen si bien aumentan considerablemente los tiempos de ejecución del modelo. De la misma forma el grado de acierto en la estimación aumenta considerablemente cuando además de la velocidad del viento se integra como señal de entrada su dirección angular. En este caso se empleó un modelo MLP con algoritmo de aprendizaje Back Propagation.

En ese mismo año Velázquez et al. [11] utilizaron técnicas ANN estableciendo la velocidad y la dirección del viento a partir de seis estaciones de referencias localizadas en diferentes islas del archipiélago canario con datos recogidos durante 10 años, posteriormente comprobó la energía generada a través de las curvas de potencia de cinco modelos de aerogenerador. Este proceso se ejecutó nuevamente con un modelo MCP lineal, comparándose los resultados a través del error MAPE. Los resultados manifiestan que en todos los casos las tasas de error MAPE del coste específico de la energía son inferiores cuando se utilizan técnicas ANN.

En el año 2013 Deligiorgi et al. [20] desarrollaron un estudio en el cual en su primera parte revisaron los antecedentes teóricos, la formulación matemática así como las ventajas e inconvenientes en el empleo de técnicas ANN. En este estudio la arquitectura implementaba un algoritmo de aprendizaje Back Propagation Levenberg – Marquardt, obteniendo la velocidad media del viento para una región de orografía compleja y comparando los resultados con los obtenidos a través de un modelo clásico MCP, en este caso de regresión múltiple.

En el año 2015 Philippopoulos y Deligiorgi [157] compararon en su estudio dos arquitecturas de redes neuronales artificiales alimentadas hacia adelante con cinco metodologías tradicionales de interpolación espacial (promedio espacial, vecino más cercano, vecino natural, distancia media inversa ponderada y distancia media inversa ponderada al cuadrado) para estimar la velocidad media horaria del viento en zonas costeras de topografía compleja. Sus conclusiones manifiestan una mayor exactitud cuando se utilizan redes neuronales en lugar de los sistemas tradicionales de extrapolación física, si bien se requiere como mínimo un año de datos para el entrenamiento del modelo.

También en el mismo año Ata [158] realiza una revisión de los artículos publicados en los que se utilizan las redes neuronales artificiales con fines relacionados con la estimación del recurso eólico. Para ello clasifica los estudios en función de las aplicaciones desarrolladas distinguiendo la predicción y el control, el estudio del recurso a corto plazo y el análisis a largo plazo. En esta revisión se señalan además las ventajas e inconvenientes en su utilización.

Carta et al. [45] publican en 2015 un estudio comparativo de dos técnicas de selección de variables para el desarrollo de la metodología MCP con redes neuronales artificiales. Con estos recursos se logran reducir los problemas de sobre especificación cuando se cuenta con estaciones de referencia múltiples para el estudio del recurso eólico, efecto negativo que si bien ya había sido detectado por otros autores [94], no había sido abordado aportando una solución alternativa a la utilización de todas las variables de las estaciones de referencia disponibles. Las técnicas de selección consideradas fueron Correlation Feature Selection – Filter Approach (CFS – FA) y Wrapper Approach (WA), ambas herramientas modeladas en el software libre Weka. El estudio se ejecutó con datos de velocidades y direcciones de vientos recabados durante dos años en cinco estaciones meteorológicas, implementándose una red neuronal Perceptron Multicapa y ejecutándose la comparación con las métricas MAE, MAPE y IoA. Sus conclusiones señalan que la técnica WA siempre genera errores inferiores a los obtenidos con CFS – FA, pero que en contrapartida su modelización requiere una alta carga computacional.

3.5. Support Vector Regression (SVR)

Partiendo de la definición general expuesta en el apartado anterior para el caso de redes neuronales artificiales para regresión, esta arquitectura podría ser definida para el caso de una sola capa oculta mediante la Ecuación 3.6 [149].

$$f(\mathbf{x}) = \sum_{j=1}^m \beta_j k(\mathbf{x}; \mathbf{w}_j) + b \quad ; \quad i = 1, \dots, n \quad (3.6)$$

con $\beta_j, b \in \mathbb{R}, \mathbf{w}_j \in \mathbb{R}^d$, y donde k es la función de activación de las unidades de la capa oculta. Dentro de este modelo general, pueden realizarse una segunda distinción según el carácter de la función de transferencia de cada neurona oculta $k(\mathbf{x}; \mathbf{w}_j)$:

1. Las redes de tipo Multilayer Perceptron utilizan una función de transferencia del tipo $k(\mathbf{x}; \mathbf{w}_j) = k(\langle \mathbf{w}_j, \mathbf{x} \rangle) = k(\mathbf{w}_j^T \mathbf{x})$ donde k es una función de activación de tipo sigmoide aplicada sobre una proyección $\mathbf{w}_j^T \mathbf{x}$ de los datos sobre cada vector de pesos \mathbf{w}_j , el cual representa así una dirección relevante para el problema.
2. Las redes de tipo Radial Basis Function networks (RBFs), utilizan una función de transferencia de tipo radial $k(\mathbf{x}; \mathbf{w}_j) = k(\|\mathbf{x} - \mathbf{w}_j\|)$, donde k es una función de activación frecuentemente de carácter gaussiano.

Bajo determinadas condiciones generales [159] ambos tipos de redes neuronales poseen la propiedad de aproximación universal que permite que su arquitectura, Ecuación 3.6, pueda aproximar la verdadera función de regresión, siempre que esta función sea una función continua.

Las redes neuronales anteriores se entrenan mediante algoritmos de optimización diversos adaptados a su arquitectura [149,160], utilizando en general la pérdida cuadrática $\ell(y, f(\mathbf{x})) =$

$(y - f(\mathbf{x}))^2$, la cual, equivale a suponer un ruido gaussiano ε en la Ecuación 3.1 en la estimación de los parámetros mediante la máxima verosimilitud.

En el marco de la formulación anterior, las Support Vector Machines para regresión pueden considerarse redes neuronales de una sola capa, Ecuación 3.6, que poseen una arquitectura del tipo radial, cuyas diferencias con las redes neuronales RBF se centran en la función de pérdida y en el algoritmo de optimización con los que se estiman sus parámetros. Así, en lugar de la pérdida cuadrática, la función de pérdida de las SVR es la denominada Vapnik ε -insensitive loss, Ecuación 3.7.

$$\ell(y, f(\mathbf{x})) = |y - f(\mathbf{x})|_{\varepsilon}^p = \max\{0, (|y - f(\mathbf{x})| - \varepsilon)^p\} \quad (3.6)$$

En Ecuación 3.7 se define un margen o banda de tamaño ε (no confundir con el ruido ε) a cada lado de la función de regresión dentro de la cual la función de pérdida no percibe error alguno. Si $p = 1$ se trata de una relajación de la pérdida absoluta y si $p = 2$ es una relajación de la pérdida cuadrática.

Utilizando esta pérdida, las SVR se estiman optimizando la siguiente función objetivo que resulta de utilizar la función de pérdida en un marco de regularización [161], es decir, penalizando el tamaño de los parámetros Ecuación 3.7.

$$\min_{\omega} \left\{ \sum_{i=1}^n |y_i - f(x_i; \omega)|_{\varepsilon}^p + \gamma \|\omega\|^2 \right\} \equiv \min_{\omega} \left\{ \|\omega\|^2 + C \sum_{i=1}^n |y_i - f(x_i; \omega)|_{\varepsilon}^p \right\} \quad (3.7)$$

donde denotamos por ω todos los parámetros de la SVR en la Ecuación 3.6 (los coeficientes β_i , los vectores de pesos w_i , y el término independiente b y la solución por $f(x; \omega)$). La constante $\gamma = 1/C$ es una constante regularizadora que pondera cuanta importancia se asigna al ajuste a los datos y cuanta a que la función no sea excesivamente compleja con el fin de evitar el sobreajuste (nótese que cuanto menor es el tamaño de los parámetros menos compleja es la función implementada por la red neuronal). Así, un mayor valor de C equivale a penalizar en mayor medida los errores, y un valor menor a primar la simplicidad de nuestra estimación de la función de regresión ante el riesgo de sobreajuste.

La descripción anterior no responde a la descripción de las SVR que suele encontrarse en la literatura emulando su desarrollo histórico, es decir, partiendo de las Support Vector Machines para el problema de clasificación. Esa descripción tradicional ayuda a ver el problema de regresión como un problema de clasificación con dos clases (los puntos que se sitúan por encima y por debajo de la curva de regresión) cuya frontera óptima sería la función de regresión, pero contribuye a ocultar el verdadero lugar que ocupan estos modelos entre las técnicas de regresión, y que hemos intentado perfilar anteriormente.

En todo caso, la formulación estándar se obtiene a partir de lo ya descrito sin más que tener en cuenta que el problema de optimización expresado por la Ecuación 3.7 equivale al siguiente problema de optimización (el factor $\frac{1}{2}$ que afecta a la norma de los parámetros no afecta a la solución), Ecuación 3.8.

$$\min_{\omega} \left\{ \frac{1}{2} \|\omega\|^2 + C \sum_{i=1}^n (\xi_i + \xi'_i) \right\} \quad (3.8)$$

$$\begin{cases} f(\mathbf{x}_i; \omega) - y_i \leq \varepsilon + \xi_i \\ y_i - f(\mathbf{x}_i; \omega) + b \leq \varepsilon + \xi'_i \\ \xi_i, \xi'_i \geq 0, i=1, \dots, n \end{cases}$$

donde se han tratado los errores superiores a ε que afectan a la ε -insensitive loss como variables slack ξ_i, ξ'_i a minimizar en la función objetivo.

Si en el marco del problema, Ecuación 3.8, se considera un modelo de tipo lineal $f(\mathbf{x}; \omega) = \langle \mathbf{w}, \mathbf{x} \rangle + b = \mathbf{w}^T \mathbf{x} + b = \sum_{j=1}^d \mathbf{w}_j \mathbf{x}_j + b$ (que bajo el paradigma conexionista puede verse como una red neuronal sin capa oculta), el problema de optimización Ecuación 3.8 es un problema cuadrático con restricciones lineales que posee solución única que puede obtenerse más fácilmente mediante su formulación dual [150,151]. El vector solución w resulta una combinación lineal de un subconjunto S de puntos de la muestra que se denominan vectores soporte, es decir, $\mathbf{w} = \sum_{\mathbf{x}_i \in S} \beta_i \mathbf{x}_i$ con lo que la estimación de la regresión resulta, Ecuación 3.9.

$$f(\mathbf{x}; \omega) = \langle \mathbf{w}, \mathbf{x} \rangle + b = \sum_{\mathbf{x}_i \in S} \beta_i \langle \mathbf{x}_i, \mathbf{x} \rangle + b \quad (3.9)$$

Por tanto, los vectores soporte $\mathbf{x}_i \in S$ son los puntos de la muestra que resultan relevantes para la estimación de la función de regresión pues solamente en ellos se apoya su forma funcional. Como resultado, esta forma funcional tiende a poseer una gran sparsity.

Sin embargo, un modelo lineal no sería adecuado para muchos problemas en los que la regresión es una función compleja. El método propuesto (*kernel trick*) para conseguir no linealidad en estos casos [162] consiste en realizar una transformación previa ϕ del espacio de features o input space χ en un nuevo espacio $\phi(\chi)$ normalmente de mayor dimensionalidad en el que se aplica el modelo lineal anterior. Así, sustituyendo en la Ecuación 3.9 los puntos de la muestra \mathbf{x}_i por los puntos transformados $\phi(\mathbf{x}_i)$, se obtiene, Ecuación 3.10.

$$f(\mathbf{x}; \omega) = \sum_{i=1}^m \beta_i \langle \phi(\mathbf{x}), \phi(\mathbf{x}_i) \rangle = \sum_{i=1}^m \beta_i k(\mathbf{x}, \mathbf{x}_i) \quad (3.10)$$

donde k es una función definida positiva que define el producto interno $k(\mathbf{x}, \mathbf{x}') = \langle \phi(\mathbf{x}), \phi(\mathbf{x}') \rangle$ en el espacio de entrada transformado $\phi(\chi)$. La elección de esta función k (denominada kernel porque este planteamiento se realiza en un Reproducing Kernel Hilbert Space (RKHS) [151] evita tener que especificar a priori el tipo de transformación ϕ a utilizar pero, además, evita la carga computacional asociada al cálculo del producto interno entre los features transformados.

En consecuencia, las propiedades analíticas de la solución descansan en las propiedades analíticas del kernel k . En particular, si el kernel elegido es no lineal se dota de carácter no lineal a la función implementada por la SVR.

Por supuesto, cada kernel diferente determina un modelo diferente de SVM. Sin embargo, el modelo final siempre se caracteriza por incorporar en su solución los vectores de la muestra que resultan relevantes en la regresión.

En todo caso, las redes SVR con arquitectura indicada en Ecuación 3.10 con núcleo k , verificando determinadas condiciones generales poseen la propiedad de aproximación universal [163]. Los criterios que guían la selección del kernel son diversos y dependen del problema en cuestión: si no se dispone de información a priori sobre el grado de complejidad de la función de regresión, se elige como kernel una función lo suficientemente flexible como para representar diferentes grados de complejidad, y se selecciona posteriormente el grado óptimo de complejidad en función de los datos.

La problemática de la selección óptima del kernel [161] no forma parte de nuestros objetivos por lo que, en este trabajo se seleccionó el kernel gaussiano debido a sus buenas propiedades ampliamente contrastadas en multitud de ámbitos de aplicación, Ecuación 3.11.

$$k(\mathbf{x}, \mathbf{x}_i) = \exp\left[-\frac{1}{2\sigma^2}\|\mathbf{x} - \mathbf{x}_i\|^2\right] \quad (3.11)$$

donde el parámetro σ regula el grado de complejidad de la solución —a mayor valor de σ mayor suavidad (menor complejidad) de la función de regresión— y puede adaptarse a los datos.

En comparación con las redes neuronales presentadas al inicio de este apartado, las SVR presentan los siguientes ventajas:

- 1) son el resultado de un problema de optimización cuadrática con restricciones lineales que posee solución única—a diferencia de la mayor inestabilidad de los problemas de optimización no lineal de las redes neuronales que poseen muchos óptimos locales y cuyos algoritmos de optimización son más inestables y dependientes de los puntos iniciales, o que de manera expresa persiguen soluciones subóptimas (como es el caso de algunos algoritmos para redes RBF),
- 2) la ya mencionada sparsity de su forma funcional,
- 3) la interpretabilidad de la solución en términos de los vectores soporte y, como consecuencia,
- 4) la posibilidad de, para muestras de gran tamaño, particionar la muestra realizando un entrenamiento por partes utilizando los diferentes subconjuntos muestrales.

En contrapartida, las SVR requieren una importante labor de selección del modelo, es decir, de selección de los denominados hiperparámetros: ϵ de la pérdida ϵ -insensitive (Ecuación 3.6), el parámetro regularizador C de la Ecuación 3.7 y el parámetro σ que determina la complejidad del kernel gaussiano.

Para la estimación de los modelos SVR utilizados en este trabajo se ha utilizado nuevamente el software multiplataforma de licencia libre R [154]. Asimismo, como base de partida en la implementación de la SVR, se utilizó el paquete Kernlab [164] el cual implementa una extensa gama de métodos de kernelización y máquinas de vector soporte para la resolución de problemas de clasificación y regresión. Su implementación ha sido descrita en el Apartado 3.7.1.

Si bien para el estudio del recurso eólico a corto plazo la técnica SVR si ha sido usada [4,22,25,165-171], para el estudio a largo plazo no se había utilizado hasta las publicaciones producidas fruto del trabajo realizado en esta tesis.

De todas las referencias citadas en el párrafo anterior para el estudio del recurso eólico a corto plazo, el artículo publicado por Peng et al. [169] merece una mención aparte dada la similitud de su aplicación con respecto a un modelo a largo plazo. En este caso se usaba como referencia una estación con datos de viento para diez meses y se calculaba la homónima en otra posición donde únicamente existían datos para dos meses. En su estudio se aplicaron tres técnicas distintas, un método de correlación lineal, un método físico y el algoritmo SVM. Según sus conclusiones el método más robusto fue el SVM ya que obtuvo el error MAPE más bajo de las tres simulaciones (5%) y a diferencia del método físico podía utilizarse indistintamente en localizaciones de orografía suave y compleja.

Por otra parte, Mohandes et al. [171] realizó una comparación entre el algoritmo SVM y una red neuronal MLP con el propósito de obtener una predicción de la velocidad media del viento a través de datos recabados en una estación durante 12 años. Para ambos casos la totalidad de los datos de partida fueron divididos en tres secciones, entrenamiento, validación y testeo, verificándose su adecuación conforme a la media del error al cuadrado (MSE). Según sus conclusiones el modelo SVM manifiesta mejores resultados que el MLP para sistemas de orden comprendido entre 1 y 11.

En el análisis del estado del arte también se han encontrado otros artículos en los que el algoritmo SVR se hibrida con otras técnicas para el estudio del recurso eólico a largo plazo. Yu et al. publica en el año 2013 [172] un estudio en el que se presenta un nuevo algoritmo hibridado llamado GMCM (Gaussian Mixture Copula Model) – GPR (Gaussian Process Regression). En este caso el GMCM se usaba para el entrenamiento a partir de la serie histórica de velocidades del viento, mientras que el algoritmo GPR permitía la construcción del modelo de regresión con el que se estiman las velocidades medias del viento. En este caso parten de datos históricos recabados durante 14 años en tres parques eólicos ubicados en los Estados Unidos en localizaciones con diferentes patrones climatológicos y geográficos con el objetivo de caracterizar la “multiestacionalidad” del viento en las regiones para el desarrollo de estimaciones a largo plazo más precisas, comparando los resultados obtenidos con otras técnicas hibridadas, GMCM – ARIMA y GMCM – SVR. De sus conclusiones los autores concluyen que la técnica más precisa con una notoria reducción del error MAPE y RMSE es la GMCM – GPR.

3.6. Random Forest (RF)

Los random forests son técnicas de Machine Learning encuadradas bajo lo que se conoce como Ensemble Learning [173], cuya predicción resulta de la combinación de las predicciones de otras técnicas. La motivación de estas técnicas es la mejora de la capacidad de predicción de las técnicas individuales que se combinan⁴, bien sea consiguiendo la reducción de su sesgo (aplicable a modelos de baja complejidad como es el caso por ejemplo de árboles dicotómicos, o a modelos entrenados con subconjuntos de datos o de features, que por sí solos no podrían resolver el problema satisfactoriamente), o bien persiguiendo la reducción de su varianza en los casos de técnicas inestables (como es el caso por ejemplo de árboles con poca poda y otras técnicas de gran complejidad).

La técnica de RF aplicada a regresión consiste en una combinación de árboles de regresión que utilizan tanto el enfoque de bootstrap aggregation (bagging) como de aleatoriedad [155].

Un árbol de regresión consiste en un conjunto de condiciones que se organizan en una estructura jerárquica (compuesta de nodos y ramas) con el propósito de predecir una variable objetivo continua Y a partir de un conjunto de variables predictoras X . Un árbol de regresión contiene un nudo raíz, nodos interiores y nodos hojas (o nodos terminales) que se conectan mediante ramas. En el nodo raíz y en los nodos interiores del árbol se especifican condiciones exhaustivas y excluyentes a las features predictoras. El número de niveles posibles que una variable predictora puede tomar establece el número de ramas descendentes de un nodo. En cada nodo hoja R_i de los M que contiene el árbol se especifica un nivel predictivo de la variable objetivo Y .

Una vez establecidos los nodos hojas, un árbol de regresión puede considerarse como una especie de modelo aditivo que puede expresarse [155] en la forma indicada en Ecuación 3.12.

$$f(\mathbf{x}) = \sum_{i=1}^M \kappa_i I(\mathbf{x} \in R_i) \quad (3.12)$$

Donde $I(\cdot)$ es una función que devuelve el valor 1 si el argumento es cierto y 0 en caso contrario; R_1, \dots, R_M representan a los nodos hojas, de tal manera que $R = \bigcup_{i=1}^M R_i$ y $R = \bigcap_{i=1}^M \emptyset$; κ_i son constantes en cada región, ya que para cada observación que cae en el interior de un nodo hoja R_i se establece la misma predicción, la cual es simplemente la media K_i de los valores objetivo y_j pertenecientes a dicho nodo hoja, Ecuación 3.13.

⁴ Bajo la denominación genérica de ensemble learning, se incluyen diversas técnicas que alcanzan ese objetivo con diversas estrategias de combinación: Boosting, bagging, stacking y optimal linear combinations, bayesian model averaging (BMA), mixtures of experts, etc.

$$\kappa_i = \bar{y}_{R_i} = \frac{1}{n_i} \sum_{x_j \in R_i} y_j \quad (3.13)$$

En Ecuación 3.13 n_i es el número total de observaciones de la variable objetivo Y pertenecientes al nodo hoja R_i .

El error de predicción del árbol puede ser evaluado calculando *Residual Sum of Squares (RSS)*, Ecuación 3.14.

$$RSS = \sum_{i=1}^M \sum_{x_j \in R_i} (y_j - \bar{y}_{R_i})^2 \quad (3.14)$$

En el proceso de construcción de un árbol de regresión se suele tomar [155] como estrategia de división de cada uno de los nodos (raíz e internos) aquella división que proporciona la mayor reducción del error evaluado con Ecuación 3.14. En este contexto, puede emplearse un greedy algorithm que, mediante un proceso iterativo conocido como división binaria recursiva, proporciona para cada nodo el punto de corte. Dicho punto de corte permite distribuir las observaciones en dos regiones (dos nodos hijos) cada una de las cuales se conecta con el nodo padre con una rama. El proceso iterativo finaliza cuando cada nodo terminal contiene menos de un número "nodesize" de observaciones establecido por el usuario u otra norma de stop.

La técnica de RF recurre al *bootstrapping para generar L pseudo-training sets* a partir de la muestra T de n datos indicada en el Apartado 3.2 con los que ajustar separadamente L árboles de regresión $f_k(x)$ y después promediarlos, Ecuación 3.15.

$$f_{RF}(x) = \frac{1}{L} \sum_{k=1}^L f_k(x) \quad (3.15)$$

Los árboles no son sometidos a podado⁵ (*pruning*), por lo que tienen varianza alta y relativamente baja bias. Con el *bagging se puede lograr* reducir la varianza del promedio de los árboles de regresión. Sin embargo, para lograr una reducción significativa de la varianza se precisa que la correlación entre cualquier par de los L árboles de regresión sea baja.

Para lograr la reducción de dichas correlaciones en el proceso de construcción de los L árboles la técnica de RF lleva a cabo una selección aleatoria de $mtry < d$ predictores antes de evaluar cada división del árbol, donde d es el número de predictores totales. Por tanto, $f_k(x)$ en Ecuación 3.1 representa al k -ésimo árbol del RF, el cual está caracterizado por las variables de división, puntos de corte de cada nodo y valores del nodo hoja.

⁵Proceso de eliminación de condiciones de las ramas del árbol para huir del sobreajuste y obtener modelos más generales [155].

Para la programación de los modelos MCP basados en RF utilizados en este trabajo se ha utilizado el paquete randomForest [174] del software multiplataforma de licencia libre R Statistics [154]. La descripción de su implementación se presenta en el Apartado 3.7.2.

La totalidad de los artículos publicados hasta el momento sobre árboles de regresión en aplicaciones relacionadas con la materia eólica se han basado en el estudio del recurso eólico con fines predictivos a muy corto plazo. En el año 2009 Mori y Umezawa [21] proponen en una conferencia un nuevo método para la construcción de reglas de datos y la predicción del recurso eólico con fines relacionados con la estabilidad, implementándose un algoritmo híbrido conocido como NBTree el cual se basaba en el árbol de decisión C4.5 y las ecuaciones de Naive Bayes (NB). Los resultados indicaron que las variables más influyentes en la predicción de la potencia dependen de las estaciones, así existe una mayor relación de dependencia de la presión atmosférica en los meses del periodo estival y de la humedad relativa en los meses de invierno.

También relacionado con el estudio del recurso eólico con fines predictivos a muy corto plazo, en el año 2015 Troncoso et al [31] publican un estudio comparativo en el que presentan ocho métodos de árboles de regresión híbridos⁶ con modelos lineales como el método de los mínimos al cuadrado y modelos no lineales como k – Nearest Neighbors o LWLR (Locally Weighted Linear Regression). Estos modelos fueron testeados con datos de ocho parques eólicos ubicados en España y los resultados fueron comparados con diferentes métricas de error, entre ellas MSE, RMSE, MAE, RMAE y R^2 , obteniéndose que las predicciones más exactas se obtenían con los modelos basados en algoritmos k – Nearest Neighbors. Posteriormente las técnicas de árbol de decisión más exactas fueron comparadas a su vez con otros métodos implementados en la industria, Perceptrón multicapa, Máquina de vector soporte, ELM (Extreme Learning Machines), GMDH (Group Method of Data Handling), CART (Classification and Regression Trees) y CHAIT (Chi – Squared Automatic Interaction Detection). En sus conclusiones manifiestan que las técnicas de árboles de regresión obtenían resultados con errores incluso inferiores a los generados con MLP y SVM, siendo a la par de las técnicas con menores tiempos de computación.

⁶Técnica que permite la combinación de las regresiones obtenidas por varios modelos con el objetivo de mejorar las predicciones obtenidas y la robustez de una solución ante fallos de una de las técnicas ensambladas.

3.7. Implementación de las técnicas seleccionadas

En la metodología MCP para el estudio del recurso eólico a largo plazo se puede distinguir dos fases fundamentales de acuerdo a la descripción presentada en el Apartado 2.8, éstas son:

- **Fase 1:** Entrenamiento y validación,
- **Fase 2:** Correlación a largo plazo con el modelo previamente validado.

Los estudios que se realizan en la presente tesis se centran en la primera fase en la que se genera el modelo de regresión a través de las técnicas mencionadas.

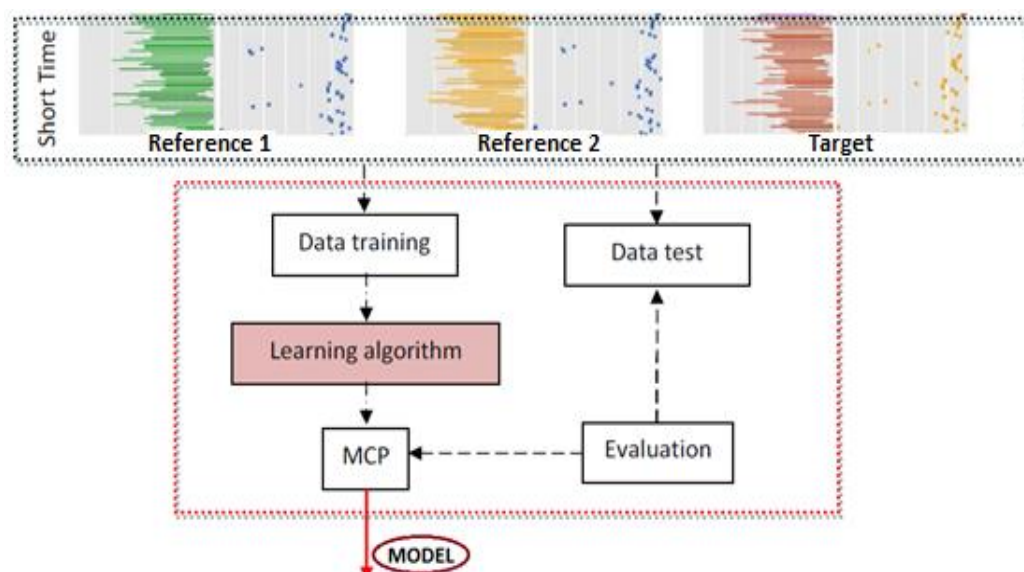


Figura 19 Entrenamiento y validación del modelo de regresión (Fase 1)

Si bien algunos autores han recurrido a mantener al margen de la modelización parte de la serie temporal de datos meteorológicos recabados en la estación candidata, para que una vez ejecutado el modelo MCP pueda ser testeada su fiabilidad [13,175], en este caso se va a optar por utilizar en la fase 1 todos los datos disponibles que hubieran sido recabados en el corto plazo en paralelo con las estaciones de referencia (un año de datos), mejorando previsiblemente las correlaciones obtenidas [11,16]. La fiabilidad se medirá por tanto en función del error obtenido tomando como referencia los datos de validación seleccionados con el proceso de validación cruzada.

Para el análisis comparativo de las técnicas Machine Learning (Apartado 4.3) los datos de partida serán siempre los mismos independientemente de la técnica utilizada en cada momento.

La descripción de cada una de las soluciones implementadas comenzará con la técnica Support Vector Regression dado que ésta fue la primera alternativa modelada desde el comienzo de la tesis doctoral, utilizándose para todos los estudios desarrollados en la tesis, esto incluye la estimación de densidades de potencia WPD, las potencias teóricas WTPO y las velocidades del viento a largo plazo. Estos estudios son discutidos en los Apartados 4.2 – 4.3 y en el extended abstract expuesto en el Anexo. Seguidamente, se describe la programación de las técnicas RF y ANN. Dichos modelos se

utilizan para el estudio desarrollado en el Apartado 4.3, el cual tienen como objetivo la estimación a largo plazo de las WTPO y para el extended abstract donde se estima la velocidad del viento a largo plazo con las tres técnicas ML.

3.7.1. Support Vector Machine

Como fue anticipado en el Apartado 3.5, para el estudio con la técnica SVR se ha empleado el software R Statistics [154], lenguaje interpretado de alto nivel ampliamente utilizado y reconocido por la industria gracias a su potencial y su capacidad para añadir a los recursos básicos del programa una extensa librería de paquetes que extiende el potencial matemático del software [176].

De entre los paquetes disponibles en la librería CRAN tiene especial interés en este ámbito de estudio el paquete Kernlab [164], el cual implementa una extensa gama de métodos de kernelización y máquinas de vector soporte para la resolución de problemas de clasificación y regresión de múltiples tipologías, herramientas que han sido la referencia en el desarrollo de máquinas de vector soporte durante la última década alabadas por los premios que han ido consiguiendo en múltiples estudios relacionados fundamentalmente con la predicción [177].

El algoritmo SVR ha sido editado en un único archivo R Script (*.R) el cual ha sido dividido internamente en cinco secciones, definición de datos de partida, búsqueda de parámetros óptimos (C , ϵ y σ), búsqueda de variables predictoras significativas con un método Feature Selection, entrenamiento y testeo de la SVR y generación de resultados. En los siguientes subapartados se detalla cada una de las secciones de la rutina y se presentan los métodos matemáticos utilizados de acuerdo con los fundamentos descritos a lo largo de este capítulo.

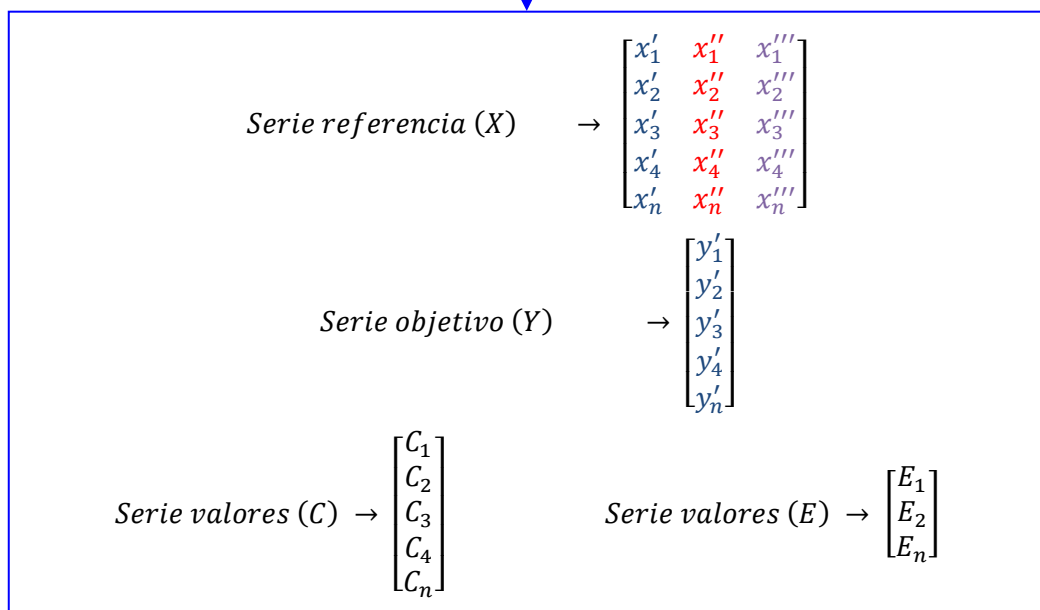
3.7.1.1 Definición de los datos de partida

Una de las potencialidades del lenguaje utilizado es su capacidad para la lectura y carga automática de las variables empleadas en el proceso matemático a través del uso de archivos en formato *.csv (delimitados por comas). Éste es el cometido principal de la primera sección.

En la rutina empleada inicialmente se opta por eliminar todas las variables almacenadas en el espacio de trabajo de R, evitando que para cada nueva simulación el programa acceda a las variables que hubieran sido almacenadas en procesos anteriores y, muy en especial, las variables conflictivas, entendiéndose éstas como los objetos acumulados con el mismo nombre en más de dos simulaciones. Seguidamente se establece la ruta al directorio de trabajo definido por el usuario, donde se cargarán los ficheros *.csv que almacenan las variables necesarias para la ejecución del modelo. Así pues, dichos archivos deben encontrarse en el directorio de trabajo que hubiera sido enrutado y además, los nombres deben ser idénticos y siempre seguidos de su extensión (*.csv). Por otra parte, las posiciones decimales estarán marcadas por puntos y nunca por comas para evitar conflictos en la lectura de estos ficheros dado que las comas suelen usarse para la separación entre columnas. Cuatro son los archivos que serán necesarios para la ejecución del algoritmo SVR:

- **Variable X:** Se refiere a los set de datos de las estaciones de referencia empleadas para la modelización (variables independientes de la Ecuación 3.1). En este sentido, será una matriz dependiendo del número de variables empleadas en el procedimiento MCP, presentándose en columnas cada una de las variables por estación de referencia utilizadas en el proceso. La resolución de los datos será horaria, existiendo una fila por cada hora en la que existen datos para todas las estaciones de referencia consideradas.
- **Variable Y:** Hace referencia al conjunto de datos de la variable dependiente en la Ecuación 3.1, el cual estará compuesto por un único vector con tantos valores como horas existan en el set de datos.
- **Valores Épsilon:** Este fichero *.csv presenta en una única columna todos los valores Épsilon que se deben ser testeados durante la ejecución del algoritmo.
- **Valores C:** Sigue el mismo principio que la variable épsilon, siendo necesario organizar un vector con cuantos valores C se pretenda testear.

<i>Estación objetivo</i>	$\rightarrow Y' = y'_1, y'_2, y'_3, y'_4, y'_n$
<i>Estación de referencia 1</i>	$\rightarrow X' = x'_1, x'_2, x'_3, x'_4, x'_n$
<i>Estación de referencia 2</i>	$\rightarrow X'' = x''_1, x''_2, x''_3, x''_4, x''_n$
<i>Estación de referencia 3</i>	$\rightarrow X''' = x'''_1, x'''_2, x'''_3, x'''_4, x'''_n$
<i>Serie de parámetros C</i>	$\rightarrow C = C_1, C_2, C_3, C_4, C_n$
<i>Serie de parámetros Epsilon</i>	$\rightarrow E = E_1, E_2, E_n$



Las variables X e Y deben estar correctamente emparejadas, lo que significa que la fila 1 de la variable Y debe corresponderse en día y hora con las medidas recabadas en las estaciones de referencia (variable X). Para realizar este proceso se ha utilizado la función merge() en la etapa de preproceso, la cual genera una única matriz con las dos variables en función del índice, que en este caso es la fecha y hora. Ya en el código de la SVR se eliminan todas las filas donde existe al menos un valor NaN. Este proceso se lleva a cabo en las variables X e Y.

Ya por último, teniendo en cuenta que el proceso 10-Folds Cross Validation va a ser aplicado en este caso, se comprueba que el número de filas de las variables X e Y es divisible por 10, y si no lo fueran, se eliminan filas hasta que se cumpla dicha condición.

```
##### DEFINICION DE DATOS DE PARTIDA (MODELO) #####
rm(list=ls())
remove(list = conflicts(detail=TRUE)$globalEnv)
getwd()
setwd("/home/superuser/Modelos personales/sdruano/MCP/Pitch/40m/RF/Model1/ws5")

# Se cargan las variables de entrada desde archivo csv almacenado en directorio:
X=data.matrix(read.csv("model1x.csv",header=TRUE,sep=";",dec="."))
Y=data.matrix(read.csv("model1y.csv",header=TRUE,sep=";",dec="."))
eps=read.csv("value_e.csv",header=TRUE,sep=";",dec=".")
Vc=read.csv("value_c.csv",header=TRUE,sep=";",dec=".")
Folds=10

#Se cargan las librerias que son utilizadas en este algortimo:
library("caret", lib.loc=~/.anaconda3/lib/R/library")
library("kernlab", lib.loc=~/.anaconda3/lib/R/library")
library("BioPhysConnector", lib.loc=~/.anaconda3/lib/R/library")

##### DETERMINA LA UBICACION DE LOS DATOS PERDIDOS Y SUSTITUYE POR NA #####

# Los valores de 0 son sustituidos por un numero pequeno para evitar error en MAPE:
X[, seq(1,dim(X)[2],2)][X[, seq(1,dim(X)[2],2)] == 0] <- NA
Y[, seq(1,dim(Y)[2],2)][Y[, seq(1,dim(Y)[2],2)] == 0] <- NA

# Se comprueba que las variables de partida no contienen datos perdidos:
completos = complete.cases(X)
X = X[completos, ]
Y = Y[completos, ]
completos2 = complete.cases(Y)
Y = Y[completos2]
X = X[completos2, ]

# Luego se comprueba que la serie tiene un numero de filas divisible entre 10 (10-folds):
n=dim(X)[1]
if(n%%10!=0){
  exacto=n%%10
  vector=c(1:exacto)
  borrar=n-vector
  X=X[-borrar,]
  Y=Y[-borrar]
}

# Guarda la serie corregida por si es necesaria su revision:
write.table(data.frame(X),file="X_corregida.csv")
write.table(data.frame(Y),file="Y_corregida.csv")
```

Puede observarse en el código presentado, que además de las variables se han cargado los paquetes R que son utilizados para el desarrollo de los modelos.

3.7.1.2 Búsqueda de parámetros óptimos C, Epsilon y σ

Para la búsqueda de parámetros óptimos, con independencia de la técnica de Machine Learning seleccionada, se han propuesto en la industria diferentes metodologías, entre ellas las técnicas Random Search, Grid Search, Particle Swarm Optimization, CMA-ES y Wrapper [178].

Asumiendo los principios expuestos en el Apartado 3.5, los parámetros característicos que deben ser estimados durante el desarrollo de la SVR son C , ϵ y σ , esta última asumiendo que la función de Kernel empleada para proyectar los datos en un espacio de mayores dimensiones es la Gaussian Radial Basic Function Ecuación 3.11.

Para la estimación del hiperparámetro óptimo σ se emplea la función sigest() perteneciente a la librería kernlab, la cual emplea un método Heurístico en su estimación. De acuerdo con [176] dicha estimación está basada sobre un valor cuantil de $\|x - x'\|^2$ comprendido entre 0.1 y 0.9, procedimiento que ha demostrado su utilidad para generar buenas aproximaciones del susodicho valor óptimo.

Por otra parte, en la estimación de los hiperparámetros óptimos C y ϵ se ha decidido aplicar la metodología Grid Search Ecuación 3.16, método indirecto en el que se testea en base al error Mean Square Error (Ecuación 3.17) generado diferentes valores predefinidos (Variables C y ϵ) ordenados en ejes cartesianos.

$$\begin{matrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \epsilon_4 \\ \epsilon_5 \\ \vdots \\ \epsilon_N \end{matrix} \quad \begin{matrix} C_1 \\ C_2 \\ C_3 \\ \vdots \\ C_N \end{matrix} \quad \rightarrow \quad \text{Error MSE} = \begin{matrix} & \text{Epsilon} \\ \begin{matrix} E_{11} & E_{12} & E_{13} & E_{14} & E_{15} & E_{1N} \\ E_{21} & E_{22} & E_{23} & E_{24} & E_{25} & E_{2N} \\ E_{31} & E_{32} & E_{33} & E_{34} & E_{35} & E_{3N} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ E_{N1} & E_{N2} & E_{N3} & E_{N4} & E_{N5} & E_{NN} \end{matrix} & C \end{matrix} \quad (3.16)$$

Conforme a la Ecuación 3.16 se selecciona en cada caso un valor diferente de C y ϵ ejecutando la máquina de vector soporte con dichos valores. Posteriormente, el resultado del error MSE es almacenado en una matriz con un número de columnas igual al número de valores ϵ testeados y un número de filas coincidente con la longitud del vector C . Una vez desarrollado dicho proceso se busca el valor mínimo de la matriz, cuyos índices determinan la posición de los parámetros característicos óptimos del caso evaluado.

$$MSE = \frac{1}{n} \sum_{i=1}^n (\hat{e}_i - o_i)^2 \quad (3.17)$$

En la Ecuación 3.16 o_i son los datos observados, \hat{e}_i los datos estimados con el modelo SVR y n el número de observaciones del conjunto de datos evaluados.

Para el desarrollo de esta rutina es indispensable el uso de la función ksvm() perteneciente a la librería kernlab. El cambio de los parámetros C y ϵ en cada una de las iteraciones descritas en el párrafo anterior se realiza con bucles cíclicos donde se integra la función ksvm(). De este proceso sólo se almacena el error obtenido en cada simulación. Se muestra a continuación la arquitectura empleada en este proceso:

```

optimos_svr <- function(X,Y,vc=vc,eps=eps,Folds=Folds){
  require(kernlab)
  vc=t(vc)
  eps=t(eps)
  nc = dim(vc)[2]
  ne = dim(eps)[2]
  n = dim(X)[1]

  #Comprobacion de variables de entrada:
  stopifnot(n > 1, dim(Y)[1] == n)
  stopifnot(Folds > 0, Folds==trunc(Folds))
  stopifnot((nc >= 1) || (ne >= 1))

  #Creando matriz de almacenamiento de errores MSE optimo:
  MSE = matrix(0,nrow=nc,ncol=ne)
  colnames(MSE, do.NULL = FALSE, prefix = "epsilon")
  rownames(MSE, do.NULL = FALSE, prefix = "c")

  #Creando vector de busqueda de epsilon minimo (bucle for):
  MSEeps = matrix(0,nc)
  jeps = matrix(0,nc)

  #Calculo de parametros por metodo iterativo:
  for (i in 1:nc) {
    for (j in 1:ne) {
      grid=ksvm(X, Y, type= "eps-svr", kernel="rbfdot", kpar="automatic", C=vc[i],
                epsilon = eps[j], cross=Folds)
      MSE[i,j] = cross(grid)
    }
    jeps[i] = which.min(MSE[i,])
    MSEeps[i] = min(MSE[i,])
  }
  #Determinacion de los parametros optimos y las variables de salida:
  ic = which.min(MSEeps)
  Coptimo = vc[ic]
  epsoptimo = eps[jeps[ic]]
  ijoptimo = c(ic,jeps[ic])
  MSEoptimo = MSE[ic,jeps[ic]]
  return(list(Coptimo=Coptimo,epsoptimo=epsoptimo,MSE=MSE,ijoptimo=ijoptimo))
}

```

La función `ksvm()` se considera el núcleo fundamental de la rutina desarrollada, la cual es apta para la resolución de distintas tipologías de problemas de clasificación y regresión. En la Tabla 5 se resume las tipologías y parámetros disponibles en la función utilizada [179].

Para el caso estudiado se ha decidido optar por la máquina de vector soporte para regresión “Eps-SVR” la cual responde a la resolución del problema de optimización presentado en la Ecuación 3.8.

En el transcurso de búsqueda de parámetros óptimos interesa que el proceso de validación cruzada sea ejecutado automáticamente por la máquina de vector soporte, ya que el único parámetro de salida que interesa es el error *MSE* obtenido como resultado de las medias de los errores en cada pliegue. Esta es la principal razón por la que se ha configurado la función `ksvm()` para que realice este proceso incluyendo la variable `Folds` que se ha asociado al número de conjuntos a ejecutar (10 en este caso).

PARÁMETROS DE CÁLCULO DE LA FUNCIÓN KSVM (KERNLAB)		
Valor	Función	Descripción
Type	C-svc	Clasificación C
	nu-svc	Clasificación nu.
	C-bsvc	Máquina de vector soporte para clasificación de límites constantes.
	Spoc-svc	Crammer. Singer native multi-class.
	Kbb-svc	Weston, Watkins native multi-class.
	One-svc	Novelty detection.
	Eps-svr	Epsilon Regression.
	Nu-svr	Nu regression.
	Eps-bsvr	Máquina de vector soporte para regresión de límites constantes.
Kernel	rbfdot	Gaussian Radial Basic Function Kernel
	Polydot	Polynomial Kernel
	Vanilladot	Linear Kernel
	Tanhdot	Hyperbolic tangent Kernel
	Laplacedot	Laplace Kernel
	besseldot	Bessel Kernel
	Anovadot	Anova RBF Kernel
	Splinedot	Spline Kernel
	Stringdot	String Kernel
Kpar	Valor de rbfdot	Configurado en automático o valor dato para las funciones de Kernel Laplace o RBF.
	Valor polydot	Grado, escala o compensación de la función polinomial
	Valor tanhdot	Escala o compensación para la tangente hiperbólica de la función tanhdot.
	Valor besseldot	Valor sigma, orden o grado de la función Bessel.
	Valor anovadot	Valor sigma o grado de la función ANOVA.
	Valor stringdot	Valores de longitud, lambda (decay factor) y normalizada de la función stringdot.
C	C	Parámetro de la ecuación característica. Término de regularización de la formulación de Lagrange.
Epsilon	Epsilon	Definición de las bandas de la función de pérdida para las máquinas de vector soporte de regresión.
Cross	Cross	Número de pliegues del proceso de validación cruzada cuyo valor debe ser superior a $k > 0$.

Tabla 5 Parámetros de cálculo de la función ksvm (Kernlab)

Tras desarrollar los cálculos mencionados en esta sección, la rutina sale de la función sólo almacenando los valores óptimos de C , ϵ y σ con los que se procederá a entrenar el modelo. De

la misma forma, para asegurar que los cálculos han sido desarrollados de manera conveniente, se ha configurado el algoritmo para que muestre el mejor valor calculado de la métrica MSE, así como la matriz de errores obtenida con el proceso Grid Search.

En cuanto a los valores C y ϵ preseleccionados para la ejecución del proceso Grid Search (vectores de datos de partida) se ha tenido en cuenta las recomendaciones sugeridas por otros autores. Así pues, el valor de C se define en función del modelo ejecutado, adoptando un valor próximo a la magnitud del rango de valores de la variable objetivo. Algo semejante sucede con el parámetro ϵ , donde se aconseja que el valor seleccionado sea tal que permita obtener un número de vectores soportes de aproximadamente el 50% del número de muestras totales [146,180-182]. La definición de estos parámetros marca en cierta medida los tiempos de computación y por ende la complejidad del modelo, siendo necesario en todo caso realizar pruebas con las que se valoran la evolución del error. En este punto también hay que tener presente que el algoritmo utilizado aplica de manera automática un proceso de Feature Scaling, lo cual condiciona en buena medida la elección del valor óptimo de C que tiende a ser menor a medida que aumenta la magnitud de la variable y que obliga al usuario a tener un control importante del proceso que modela.

3.7.1.3 Feature Selection aplicado para la técnica SVR

Uno de los problemas generales de los métodos de Machine Learning es el riesgo de sobreajuste, el cual ha sido tratado con anterioridad en este capítulo y que marca la pérdida de flexibilidad del modelo cuando existe sobre especificación. En la industria del Machine Learning se conoce como Feature Selection a la familia de aplicaciones que tienen por objetivo solucionar este problema mediante la selección del número de variables predictoras adecuadas para el procedimiento modelado, reduciendo el número de referencias utilizadas al mínimo necesario mediante la eliminación de las variables redundantes y las que no aportan significado alguno tomando como base la serie objetivo.

Comúnmente, los métodos Feature Selection (FS) son presentados en tres categorías en función de cómo combina el algoritmo de selección con el propio modelo de aprendizaje, Filter Method, Wrapper Method y Embedded Method [45,183]. Para los objetivos planteados en esta tesis doctoral se ha seleccionado un método Wrapper al ser la opción más exhaustiva, en concreto el algoritmo *Recursive Feature Elimination* (RFE), procedimiento supervisado que determina el subconjunto de variables óptimas tras introducir en un algoritmo de aprendizaje distintas combinaciones de variables predictoras. En cada simulación se puntúa la significancia de cada variable a través de la medida del error, y se descartan aquellas que empeoran los resultados con respecto a la situación de máxima bondad [184].

Para la ejecución del algoritmo RFE se han utilizado los recursos disponibles en el paquete Caret (*Classification and Regression Training*) [185] del software R Statistics. En el algoritmo RFE, el modelo de selección puede ser distinto al modelo de aprendizaje, sin embargo, otros autores han concluido que la fiabilidad de los resultados en este caso es menor ya que las pautas de búsqueda de los dos modelos pueden seguir distintas estrategias, lo que se traduce en que los óptimos del primer modelo

pueden no coincidir con los del segundo [45]. Siguiendo la línea referida, en este algoritmo se usa la misma función SVR para la selección y el aprendizaje, estando ésta inmersa a su vez en un mecanismo propio de validación cruzada 10 – Folds:

```
#####IMPLEMENTACION DEL METODO WRAPPER (FEATURE SELECTION)#####
#Se utiliza el método Feature selection del paquete caret. De entre las opciones
#disponibles se opta por Recursive Feature Selection y en concreto la función
#"svmRadial" del paquete de entrenamiento. El proceso se integra en validación cruzada
#10 folds tomando parámetros óptimos del proceso anterior.

require(caret)
ctrl<-rfeControl(functions = caretFuncs, method="cv", number=Folds)
tune_grid<-data.frame(C=Coptimo,sigma=svr_train_I@kernelF@kpar$sigma)
Features<-rfe(X, as.numeric(Y), rfeControl=ctrl,
              sizes = c(seq(1,30,1)), method='svmRadial',tuneGrid=tune_grid,
              trControl=trainControl(method = 'cv', number=Folds))

Vopt<-Features$optVariables
xoptimos<-X[,c(Vopt)]

#Guarda la evolución del error en función del número de entradas consideradas:
nombre=paste("Feature_Selection.jpg", sep="")
jpeg(nombre)
plot(Features)
dev.off()
```

En cuanto a la técnica de selección de subconjuntos de muestreo, el algoritmo RFE implementa el método *Backward Elimination*, el cual comienza estimando el error RMSE (Ecuación 3.18) con todos los predictores existentes, clasificándose las características en función de su importancia ($S_1 > S_2 > S_N$). A continuación, mientras los predictores de mayor importancia son retenidos, el resto de variables son permutadas de manera secuencial valorando en cada caso la exactitud del resultado y reajustando la clasificación de los predictores en función del error ($RMSE_1 < RMSE_2 < RMSE_N$). Cuando la mejor combinación de variables es determinada, se eliminan todos los predictores que producen un error mayor que el obtenido con la opción óptima [184,186]. Tras aplicar el procedimiento Feature Selection se obtiene un conjunto de variables óptimas S_i el cual es usado para el entrenamiento y el testeo de la SVR. Se consigue con todo ello una reducción del riesgo de sobreajuste a costa de un aumento de los tiempos y los requerimientos computacionales en la fase de entrenamiento.

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (o_i - \hat{e}_i)^2} \quad (3.18)$$

En la Ecuación 3.18 o_i son los datos observados, \hat{e}_i los datos estimados con el modelo SVR y n el número de observaciones del conjunto de datos evaluados.

En el caso concreto mostrado en el código anterior, el algoritmo parte de los valores de C y $\acute{Epsilon}$ seleccionados en el proceso anterior, sin embargo, una alternativa de mayor exactitud es la supresión de dichos valores óptimos forzando a que esta función busque los valores en cada una de las iteraciones ejecutadas. Este método genera unas estimaciones más precisas, no obstante, los tiempos de computación aumentan hasta un 800%.

En la última etapa de esta sección, el algoritmo RFE grafica la evolución del error en función del número de variables seleccionadas. Se muestran en la Figura 20 dos ejemplos de los resultados generados para dos modelos independientes.

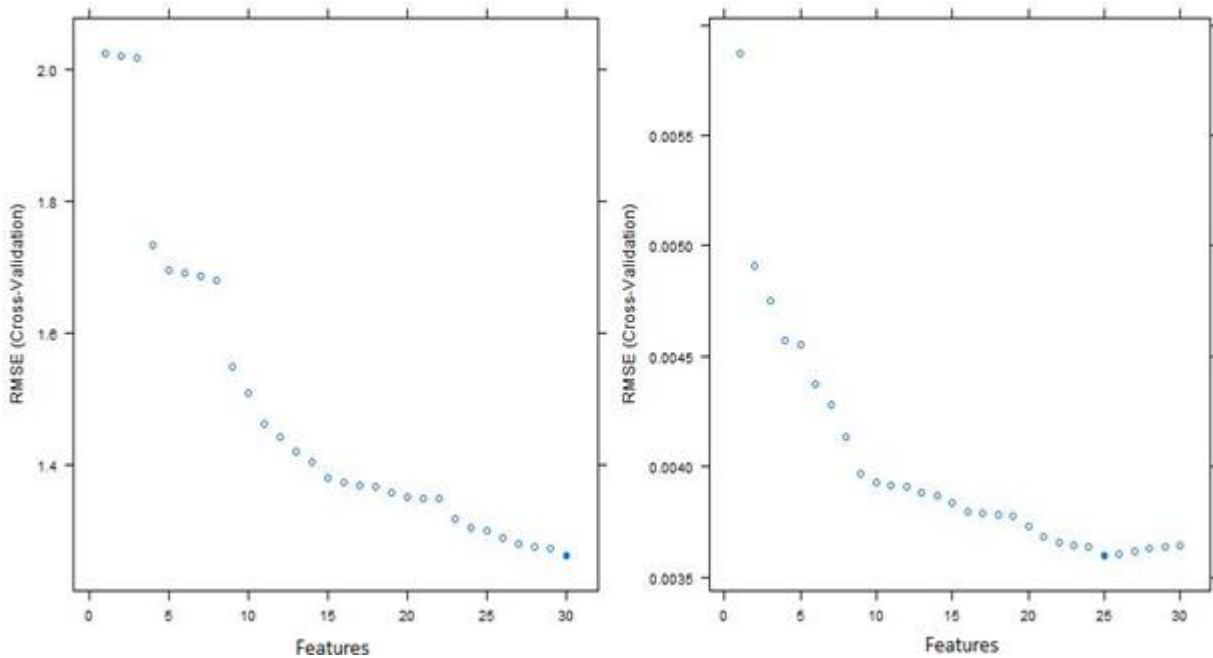


Figura 20 Comparativa de resultados del método Wrapper Approach implementado para SVR

En la Figura de la izquierda se observa que el mejor error RMSE se obtuvo en aquella situación en la que se seleccionaban todas las variables predictoras, en contra del modelo representado en la Figura de la derecha para la cual se eliminó 5 predictores.

3.7.1.4 Entrenamiento y testeo de la técnica SVR

Una vez seleccionados los parámetros y las variables predictoras óptimas, se realiza el entrenamiento del modelo. Este proceso de recálculo, si bien en un principio parece no ser necesario, es útil para analizar los resultados obtenidos en cada uno de los 10 conjuntos de validación cruzada.

```

#Division para validacion cruzada:
folds = sample(cut(seq(1,nrow(X)),breaks=Folds,labels=FALSE))
for (i in 1:Folds) {
  testIndexes = which(folds==i,arr.ind=TRUE)
  xtest = X[testIndexes, ]
  ytest = Y[testIndexes]
  xtrain = X[-testIndexes, ]
  ytrain = Y[-testIndexes]
  ntest = length(ytest)

  #Entrenamiento del modelo y almacenamiento de resultados por iteración:
  svr_train=ksvm(xtrain,ytrain,type= "eps-svr", kernel="rbfdot", kpar="automatic",
                C=Coptimo, epsilon=epsoptimo)

  MSE_train[i]=error(svr_train)
  ytrainv = fitted(svr_train)
  ypredict_train[1:length(ytrainv),i]=ytrainv
  supportv[i]=nsv(svr_train)
}

```

En este proceso, inicialmente se procede a crear un vector de valores índices comprendidos entre 1 y 10 (Folds) cuya longitud coincide con el número de filas de la variable X. Seguidamente, cada uno de los índices generados en el vector folds se asocia a las variables X e Y dependiendo de su posición. Este proceso se realiza mediante un bucle con el que se asocia cada índice i a cada valor de las variables X e Y de forma aparejada como se muestra en Figura 21 para un conjunto simple de 20 datos por variable.

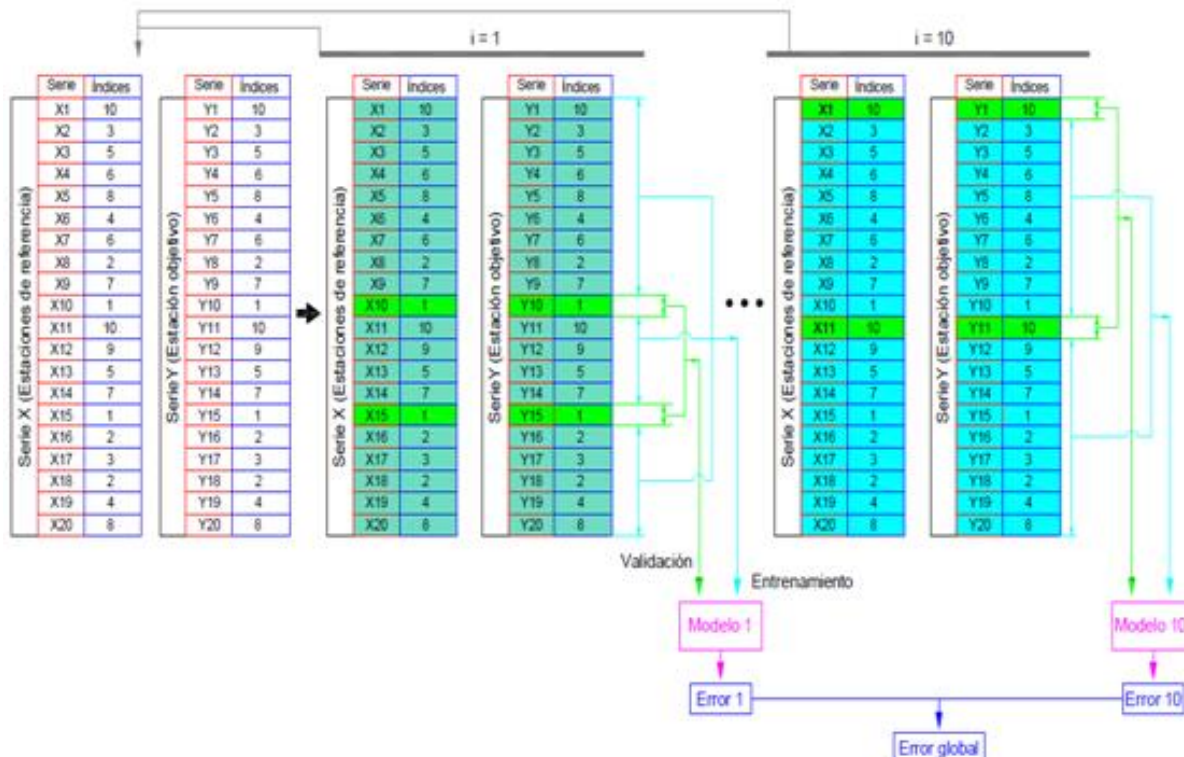


Figura 21 Proceso de selección de parámetros mediante el proceso de validación cruzada

Tras haberse procedido al reparto entre los conjuntos de entrenamiento y validación, se ejecuta nuevamente la función `ksvm()`, esta vez usando sólo las variables por estaciones de referencia que

superaron la prueba Feature Selection y con los valores óptimos de C , ϵ y σ . Del mismo modo, ahora se deshabilita la opción de que se ejecute de manera automática el método de validación cruzada ya que, como se ha explicado, ahora se realiza de manera manual para acceder a la máxima información disponible de cómo se ejecuta el proceso. Tras ejecutar el modelo se devuelve en el espacio de trabajo los siguientes resultados por cada pliegue de validación cruzada:

- Valor del error MSE obtenido en la fase de entrenamiento.
- Vector con los valores que conforman la curva Y de entrenamiento.
- Número de vectores soporte totales.

De acuerdo con los principios básicos de la metodología SVR, sólo los valores x_i con coeficientes α_i ó α_i^* distintos de cero son tenidos en cuenta en el modelo SVR, los cuales se definen como vectores soportes y son los responsables del trazado de la función de regresión estimada. Para determinar el número de vectores soporte sólo se tienen en cuenta aquellos que se encuentran ubicados justo en la frontera de la banda de radio ϵ . Éstos satisfacen la condición complementaria $0 < \beta < C$ [150].

Siguiendo con el código, posteriormente se recurre a la función `predict()`, también perteneciente al paquete `kernlab`, para generar la estimación de valores Y con el conjunto X de validación.

```
#Testeo con predict y almacenamiento de resultados:
ytestv = predict(svr_train,xtest)
ytestv = ifelse(ytestv<0,0.4,ytestv)
MSE_test[i] = sum((ytestv-ytest)^2)/ntest
MAPE_test[i] = mean(abs((ytestv-ytest)/ytest))
MAE_test[i] = mean(abs(ytestv-ytest))
ypredict_test[,i] = ytestv
```

Este proceso sólo es posible si previamente se ha ejecutado el modelo `ksvm()` y se disponen de todos los datos necesarios para obtener la función de regresión. El resultado de esta expresión se almacena en un vector denominado `ytestv` según el código anteriormente expuesto y que obedece a la Ecuación 3.1.

A continuación se compara la serie y y estimada (\hat{y}_i) con el vector y de datos observados (y_i), obteniéndose el error de testeo. Para el cálculo de este error se han considerado distintas métricas puesto que, dependiendo de la simulación realizada, los resultados obtenidos con cada una de ellas podrían diferir.

En términos generales, para los estudios llevados a cabo en esta tesis doctoral fueron seleccionadas las métricas Mean Absolute Error (MAE) según la Ecuación 3.19, Mean Absolute Relative Error (MARE) Ecuación 3.20 y el coeficiente de determinación R^2 según Ecuación 3.21.

$$MAE = \frac{1}{n} \sum_{i=1}^n |\hat{e}_i - o_i| \quad (3.19)$$

$$MARE = \frac{1}{n} \sum_{i=1}^n \left| \frac{o_i - \hat{e}_i}{o_i} \right| \quad (3.20)$$

$$R^2 = 1 - \frac{\sum_{i=1}^n (o_i - \hat{e}_i)^2}{\sum_{i=1}^n (o_i - \bar{o})^2} \quad (3.21)$$

En las Ecuación 3.19, Ecuación 3.20 y Ecuación 3.21 o_i son los datos observados, \bar{o} la media de los valores observados, \hat{e}_i los datos estimados con el modelo MCP y n el número de observaciones del conjunto de datos evaluados.

La lectura de los índices de error (Ecuación 3.19 y Ecuación 3.20) definen el grado de exactitud que se obtiene con el modelo desarrollado, siendo su estudio necesario para la comparación de las distintas técnicas de Machine Learning o simplemente la idoneidad del método empleado en la estimación del potencial eólico a largo plazo. Adicionalmente, se calcula el coeficientes R^2 (Ecuación 3.21), el cual define un valor en tantos por unidad que miden la calidad del modelo para estimar nuevas proyecciones en Y , o lo que es lo mismo, la capacidad para predicción de datos de potencia a largo plazo en la estación objetivo [187].

```
#Calculo del coeficiente de determinación R2:
correlacion = matrix(0,ntest)
for (j in 1:ntest){
  correlacion[j]=(ytest[j]-ytestv[j])^2

#Generacion grafica correlacion test:
fitter=lm(ytestv~ytest)
p=fitted.values(fitter)
nombre=paste("Correlacion_", i, ".jpg", sep="")
jpeg(nombre)
plot(ytest,ytestv,type="p",xlab="Serie validación objetivo",ylab="Estimación sv)
points(ytest,ytest,type="c",col="green")
points(ytest,p,type="l",col="blue")
title(main=list("Coeficiente de correlacion"))
legend("topleft",legend=c("Data","Y=Predict","Fit"),lty=c(4,1,2),col=c(2,1,3))
dev.off()
}
SSE = sum(correlacion)
SST = sum((ytest-mean(ytest))^2)
rsquared[i] = 1-(SSE/SST)
```

Todos los procesos mencionados desde el inicio de este apartado se desarrollan por separado para cada uno de los pliegues definidos. Teniendo en cuenta que los resultados obtenidos son valores o vectores, éstos se almacenan en vectores o matrices según correspondan para cada uno de los casos. Para finalizar se representa en la Figura 22 el ajuste del modelo a través de la comparación del conjunto de validación (dato real) con la predicción.

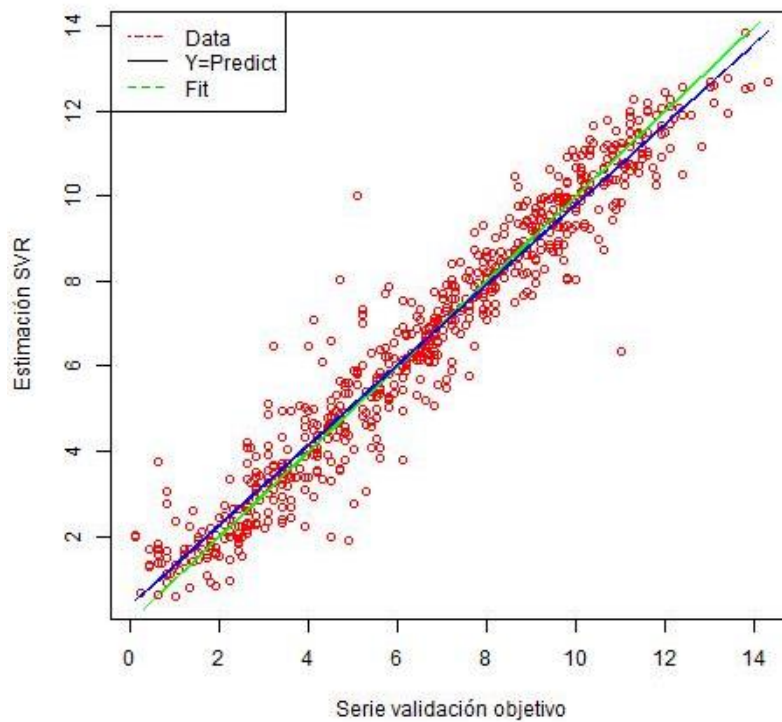


Figura 22 Ajuste de correlación R^2 obtenido tras ejecutar el modelo con 5 estaciones de referencia

Además, del algoritmo creado se obtienen los siguientes resultados generales los cuales se envían al espacio de trabajo:

- Matriz de valores de error MSE obtenidos durante la búsqueda de valores óptimos C y Épsilon.
- Vector de valores de error MSE en la fase de testeo. Cada valor representa el resultado de cada uno de los 10 pliegues.
- Vector de valores de error MAE en la fase de testeo. Cada valor representa el resultado de cada pliegue.
- Vector de valores de error MAPE en la fase de testeo. Cada valor representa el resultado de cada pliegue.
- Vector de valores de R^2 en la fase de testeo. Cada valor representa el resultado de cada pliegue.
- Matriz de estimaciones Y generadas para cada uno de los 10 pliegues.
- Vector definido con el número de vectores soporte en cada pliegue.

3.7.1.5 Generación de resultados en el algoritmo SVR

Para finalizar, el algoritmo envía a la ruta especificada por el usuario los resultados del modelo los cuales se presentan en formato *.csv, facilitando el análisis posterior y evitando que por acciones de simulaciones consecutivas se pierdan los análisis realizados en los modelos anteriores.


```
##### EXPORTACION DE LOS RESULTADOS A CSV #####

#Plotea los resultados en formato csv para el directorio de trabajo definido en
#la seccion inicial de la rutina.

write.table(data.frame(MSE_test_I),file="MSE_test_Model1.csv")
write.table(data.frame(MSE_test),file="MSE_test_W_Model1.csv")
write.table(data.frame(MAPE_test_I),file="MAPE_test_Model1.csv")
write.table(data.frame(MAPE_test),file="MAPE_test_W_Model1.csv")
write.table(data.frame(MAE_test_I),file="MAE_test_Model1.csv")
write.table(data.frame(MAE_test),file="MAE_test_W_Model1.csv")
write.table(data.frame(MSE_train_I),file="MSE_train_Model1.csv")
write.table(data.frame(MSE_train),file="MSE_train_W_Model1.csv")
write.table(data.frame(rsquared_I),file="RSquare_Model1.csv")
write.table(data.frame(rsquared),file="RSquare_W_Model1.csv")
write.table(data.frame(IoA_I),file="IoA_Model1.csv")
write.table(data.frame(IoA),file="IoA_W_Model1.csv")
write.table(data.frame(ypredict_test_I),file="Predict_test_Model1.csv")
write.table(data.frame(ypredict_test),file="Predict_test_W_Model1.csv")
write.table(data.frame(ypredict_train_I),file="Predict_train_Model1.csv")
write.table(data.frame(ypredict_train),file="Predict_train_W_Model1.csv")
write.table(data.frame(MSE),file="Busqueda_paropt_Model1.csv")
write.table(data.frame(MSEW),file="Busqueda_paropt_W_Model1.csv")
write.table(data.frame(vopt),file="Parametros_optimos_Model1.csv")
write.table(data.frame(MSE_test_pot),file="Potencia_MSE_test_W_Model1.csv")
write.table(data.frame(MAPE_test_pot),file="Potencia_MAPE_test_W_Model1.csv")
write.table(data.frame(SMAPE_test_pot),file="Potencia_SMAPE_test_W_Model1.csv")
write.table(data.frame(MAE_test_pot),file="Potencia_MAE_test_W_Model1.csv")
write.table(data.frame(rsquared_pot),file="Potencia_RSquare_W_Model1.csv")
write.table(data.frame(IoA_pot),file="Potencia_IoA_W_Model1.csv")
write.table(data.frame(potenciapredict_test),file="Potencia_Predict_test_W_Model1.csv")
save.image("C:/Users/Dim/Desktop/Articulo densidades SDR/Modelo 1/Resultados.RData")
```

3.7.2. Random Forest

En concordancia con lo descrito en el Capítulo 1 y de modo semejante a como se ha procedido con la técnica Support Vector Regression en el Apartado 3.7.1, para el desarrollo de los estudios que se presentan en este trabajo, se parte de implementaciones de los distintos algoritmos en código abierto los cuales son posteriormente adaptados al estudio del potencial eólico a largo plazo. Así pues, la comprensión de los fundamentos teóricos de la técnica analizada se comete con vistas a comprender los resultados obtenidos en la práctica y no para el diseño de nuevas herramientas.

Para el caso singular del algoritmo Random Forest, su simplicidad estructural y su semejanza a la técnica de árboles de regresión ha motivado que actualmente existan múltiples librerías y paquetes informáticos que implementen dicha solución. En general, estas librerías desarrollan el algoritmo original propuesto por Breiman [188], incluyéndose la técnica Bagging propuesta por el mismo autor en el año 1996 [189]. No obstante, algunos paquetes informáticos como OpenCV ofrecen otras alternativas como el algoritmo Extremely Randomized Trees propuesto por Geurts et al. [190] o la variante Random Subspaces desarrollada por Ho en el año 1998 [191].

La mayoría de los algoritmos Random Forest fueron programados de inicio en lenguajes de programación de bajo nivel, comúnmente C/C++ y Java. Posteriormente, éstos fueron traducidos y en ocasiones mejorados con otros lenguajes de programación de alto nivel tales como R Statistics, Matlab o Python.

De acuerdo con el estudio del estado del arte, en la Tabla 6 se exponen las principales librerías de código abierto que implementan la técnica Random Forest, mencionándose en la misma línea otros datos relevantes como el criterio de parada considerado en cada modelo, el programa de desarrollo original, las variantes de algoritmo implementadas y su capacidad para desarrollar tareas de clasificación y/o regresión en Multihilo.

Debe hacerse constar que la relación expuesta en dicha tabla no contempla todas las opciones existentes en la industria. No obstante, si se puede afirmar que son las de mayor relevancia medida en términos de trabajos publicados durante la última década.

Conforme a la información sintetizada en la tabla anterior, todas las librerías a excepción de H2O y Weka permiten la resolución de problemas de regresión. En lo referente al criterio de parada, los distintos algoritmos suelen apostar por dos estrategias fundamentales, estos son: i) Número mínimo de muestras requeridas por nodo (por ejemplo nodesize en R Statistics), ii) Definición de la variable upper bound (para librerías como Scikit-Learn). Asimismo, para medir la calidad de la estimación los distintos paquetes suelen implementar distintas opciones, las cuales determinan la entropía del modelo ejecutado, siendo las más comunes Gini y MSE.

Desde el punto de vista computacional, algunas de las librerías mencionadas tienen la potencialidad de ejecutar varios procesos al mismo tiempo aprovechando sus capacidades de análisis Multihilo, lo cual podría considerarse una ventaja y en ocasiones un criterio aceptable de decisión. Además, por

tiempos computaciones, algunos estudios señalan que las mejores opciones son aquellas que están programadas en lenguajes de alto nivel.

LIBRERÍAS RANDOM FOREST RECOMENDADAS POR LA INDUSTRIA					
Librería	Clase	Algoritmo	Criterio de parada	Programa	Multihilo
Weka	Clasificación	Random Forest y Bagging	<i>depth</i>	Weka, Java	Si
randomForest	Clasificación y regresión	Random Forest y Bagging	<i>nodesize, maxnodes</i>	R Statistics	No
Scikit - Learn	Clasificación y regresión	Random Forest, Bagging, Extremely Randomized Trees, Random Subspaces	<i>max_depth, min_samples_split, min_samples_leaf, max_leaf_nodes</i>	Python	Si
OpenCV	Clasificación y regresión	Random Forest, Bagging, Extremely Randomized Trees	<i>max_depth, min_samples_count, forest_accuracy</i>	C/C++	Si
OK ₃	Clasificación y regresión	Random Forest, Bagging, Extremely Randomized Trees	<i>varmin, nmin, maxnbsplits</i>	C	No
Orange	Clasificación y regresión	Random Forest y Bagging	<i>worst_acceptable, min_susbet, min_instances, max_depth, max_majority</i>	Phyton	No
H2O	Clasificación	Random Forest y Bagging	<i>max_depth</i>	Java	Si

Tabla 6 Librerías Random Forest recomendadas por la industria

Para los estudios desarrollados en la presente tesis doctoral se ha optado por la librería *randomForest*, motivándose la elección en los siguientes aspectos:

1. La librería *randomForest* está desarrollada bajo las directrices del lenguaje de programación de alto nivel R Statistics, lo que supone una ventaja pues permite su adaptación a las necesidades derivadas del estudio estadístico requerido para en estos trabajos.
2. Además de las funcionalidades propias del paquete *randomForest* hay que considerar las relativas a otros paquetes del mismo entorno, mejorando el análisis estadístico desarrollado, la interfaz y los modos de presentación de los resultados obtenidos.
3. El hecho de que no se cuente con función Multihilo no supone un problema pues se cuenta con la suficiente capacidad computacional para el procesamiento de los datos en los distintos análisis desarrollados. Adicionalmente, si fuera necesario el desarrollo de procesos de estas características, se puede anidar distintos procesos paralelos a través de un bucle inmerso en el procedimiento de validación cruzada.

4. La citada librería permite la formulación del algoritmo sólo definiendo dos parámetros característicos, manteniendo la coherencia con los trabajos hasta el momento desarrollados en la tesis doctoral con la técnica SVR.
5. El paquete randomForest está alabado por múltiples publicaciones desarrolladas en la última década.
6. Se ha apostado por utilizar un mismo entorno, en este caso el programa R Statistics, para el desarrollo de todas las técnicas de Machine Learning evaluadas en la presente tesis doctoral.

En el diseño de los modelos de estimación a largo plazo de la presente tesis doctoral se ha considerado clave que, independientemente de la técnica de aprendizaje estadístico implementada, la estructura general de los modelos desarrollados mantengan una serie de condiciones comunes para garantizar que en la comparación de los resultados, las diferencias no sean debidas a aspectos que no tengan que ver con la propia técnica de Machine Learning empleada en cada caso. Así pues, partiendo del análisis realizado en el Apartado 3.7.1 para Support Vector Regression, en el diseño del modelo de estimación para la técnica Random Forest se estipulan las siguientes condiciones de partida:

1. Los ensayos que se realizan en la tesis doctoral se centran en la fase de entrenamiento y validación de los modelos. Como se analizará en el Capítulo 4, sólo se cuentan con datos meteorológicos recabados para el año 2014 con frecuencia horaria, lo cual se considera suficiente para el entrenamiento de los modelos de estimación a largo plazo y su validación.
2. La fiabilidad del modelo se medirá en función de los índices de error y correlación obtenidos en la fase de entrenamiento (mismas métricas que han sido usadas en el caso de la técnica SVR) al aplicar el procedimiento conocido como 10 – Folds Cross Validation.
3. El modelo debe estar preparado para estimar usando múltiples variables explicatorias entre las que destacarían la velocidad y dirección del viento, así como la densidad del aire o las potencias teóricas generables en cada una de las 9 estaciones usadas como referencia en cada análisis.
4. Los problemas relacionados con el sobreajuste serán controlados en su caso mediante una técnica de Feature Selection aplicada con anterioridad a la ejecución de la metodología MCP. En este caso, se usará al igual que para el caso de SVR la técnica Wrapper Approach implementada con el paquete Caret de R Statistics.
5. Para la búsqueda de los hiperparámetros característicos óptimos *mtree* y *nodesize* se utiliza, al igual que para el caso de SVR, la técnica Grid Search. Así mismo en el caso del hiperparámetro *mtry*, se recurre al método heurístico implementado en el propio paquete randomForest.

De acuerdo con las conclusiones expuestas en el apartado anterior, el software que será empleado para la modelización será R Statistics, recurriéndose en este caso al paquete randomForest.

El algoritmo RF es desarrollado usando una estrategia de factorización semejante a la que se ha empleado en el caso de la técnica SVR, dividiéndose el script en cinco secciones principales, definición de datos de partida, búsqueda de parámetros óptimos (*mtree*, *mtry* y *nodesize*), búsqueda de variables predictoras significativas a través del método Wrapper, entrenamiento y testeo de la RF y generación de resultados. En los siguientes subapartados se realiza una descripción de la estructura general del modelo RF diseñado. En este sentido, sólo se ahondará en aquellas secciones que presentan cambios fundamentales con respecto a la arquitectura diseñada y ya descrita en el Apartado 3.7.1 para la técnica SVR.

3.7.2.1 Definición de los datos de partida

Como fue descrito en el Apartado 3.7.1 para el caso particular de SVR, en la implementación del modelo RF también se ha aprovechado la capacidad de R para la carga automática de las variables de partida desde archivos en formato *.csv.

```
# Se cargan las variables de entrada desde archivo csv almacenado en directorio:
X=data.matrix(read.csv("model1x.csv",header=TRUE,sep=";",dec="."))
Y=data.matrix(read.csv("model1y.csv",header=TRUE,sep=";",dec="."))
potencia_ref=data.matrix(read.csv("potencia1.csv",header=TRUE,sep=";",dec="."))
pot_aero=data.matrix(read.csv("potencia_aero4.csv",header=TRUE,sep=";",dec="."))
folds=data.matrix(read.csv("folds.csv",header=TRUE,sep=";",dec="."))
ntree=read.csv("ntree.csv",header=TRUE,sep=";",dec=".")
nodesize=read.csv("nodesize.csv",header=TRUE,sep=";",dec=".")
Folds=10
Y = ifelse(Y>40,Y/1000,Y)

#Se cargan las librerías que son utilizadas en este algoritmo:
library("caret", lib.loc=~ /anaconda3/lib/R/library")
library("randomForest", lib.loc=~ /anaconda3/lib/R/library")
library("bootstrap", lib.loc=~ /anaconda3/lib/R/library")
library("BioPhysConnector", lib.loc=~ /anaconda3/lib/R/library")

##### DETERMINA LA UBICACION DE LOS DATOS PERDIDOS Y SUSTITUYE POR NA #####

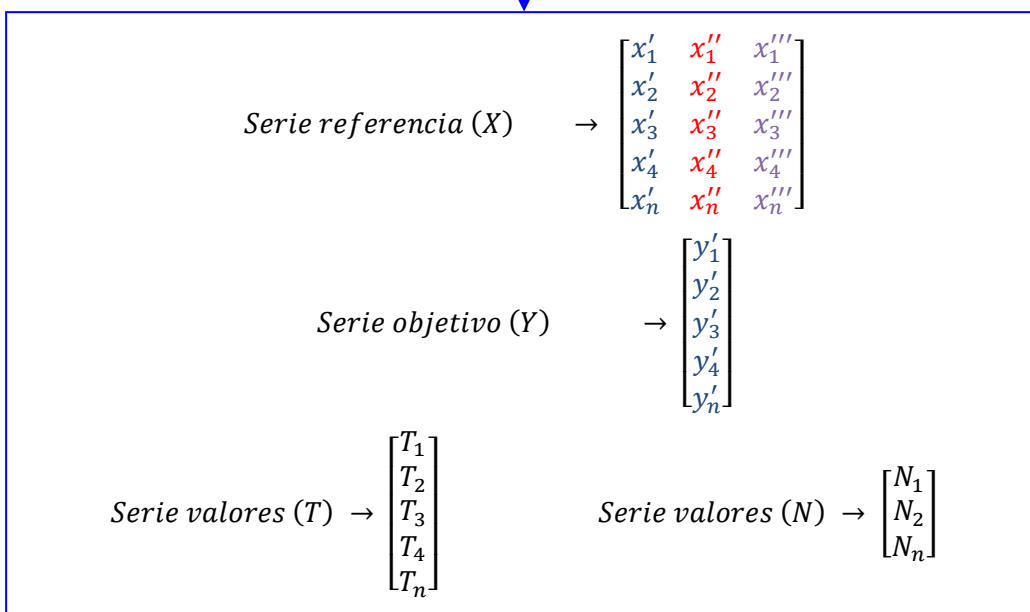
# Los valores de 0 son sustituidos por un numero pequeno para evitar error en MAPE:
X[, seq(1,dim(X)[2],2)][X[, seq(1,dim(X)[2],2)] == 0] <- NA
Y[, seq(1,dim(Y)[2],2)][Y[, seq(1,dim(Y)[2],2)] == 0] <- NA

# Se comprueba que las variables de partida no contienen datos perdidos:
completos = complete.cases(X)
X = X[completos, ]
Y = Y[completos, ]
completos2 = complete.cases(Y)
Y = Y[completos2]
X = X[completos2, ]

# Luego se comprueba que la serie tiene un numero de filas divisible entre 10 (10-folds):
n=dim(X)[1]
if(n%%10!=0){
  exacto=n%%10
  vector=c(1:exacto)
  borrar=n-vector
  X=X[-borrar,]
  Y=Y[-borrar]
}

# Guarda la serie corregida por si es necesaria su revision:
write.table(data.frame(X),file="X_corregida.csv")
write.table(data.frame(Y),file="Y_corregida.csv")
```

<i>Estación objetivo</i>	$\rightarrow Y' = y'_1, y'_2, y'_3, y'_4, y'_n$
<i>Estación de referencia 1</i>	$\rightarrow X' = x'_1, x'_2, x'_3, x'_4, x'_n$
<i>Estación de referencia 2</i>	$\rightarrow X'' = x''_1, x''_2, x''_3, x''_4, x''_n$
<i>Estación de referencia 3</i>	$\rightarrow X''' = x'''_1, x'''_2, x'''_3, x'''_4, x'''_n$
<i>Serie de parámetros mtree</i>	$\rightarrow T = T_1, T_2, T_3, T_4, T_n$
<i>Serie de parámetros nodesize</i>	$\rightarrow N = N_1, N_2, N_n$



Como para el caso de SVR, el algoritmo carga desde el directorio especificado cuatro archivos independientes, estos son:

- **Variable X:** Matriz de idénticas características al explicado en el Apartado 3.7.1.1.
- **Variable Y:** Vector de idénticas características al explicado en el Apartado 3.7.1.1.
- **Valores mtree:** En este caso se dispondría de una única columna en la que se especificaría por filas el número de árboles de regresión considerados en cada análisis. De acuerdo con las recomendaciones estipuladas en la literatura científica, se establecen valores de entre 500 y 3000 árboles con pasos cada 500.
- **Valores nodesize:** Sigue una metodología semejante a la estipulada para *mtree*, ingresándose en cada una de las filas del archivo *csv los distintos valores *nodesize* que se quisieran testear. En este caso se prueban valores de 1, 3, 4, 5, 6, 8, 10, 13, 15 y 20.

Las variables X e Y deben ser interpretadas como matrices para el correcto funcionamiento del modelo.

3.7.2.2 Búsqueda de parámetros óptimos `nodesize`, `mtry` y `mtry`

Para la selección de los hiperparámetros óptimos se recurren a dos métodos diferenciados, el método Grid Search para definir las variables de número árboles de regresión y `nodesize` y al método `tuneRF` para seleccionar el valor óptimo de la variable `mtry`.

```

optimos_rf <- function(Xoptimos,Y,ntree=ntree,nodesize=nodesize,Folds=Folds)
  require(randomForest)
  require(bootstrap)
  ntree=t(ntree)
  nodesize=t(nodesize)
  nt = dim(ntree)[2]
  nno = dim(nodesize)[2]
  n = dim(Xoptimos)[1]

  #Comprobacion de variables de entrada:
  stopifnot(n > 1, dim(Y)[1] == n)
  stopifnot(Folds > 0, Folds==trunc(Folds))
  stopifnot((nt >= 1) || (nno >= 1))

  #Creando matriz de almacenamiento de errores MSE optimo:
  MSE = matrix(0,nrow=nt,ncol=nno)
  colnames(MSE, do.NULL = FALSE, prefix = "nodesize")
  rownames(MSE, do.NULL = FALSE, prefix = "ntree")

  #Creando Vector de busqueda de nodesize (bucl e for):
  mtry_opt=matrix(0,nt)
  MSEnode = matrix(0,nt)
  jnode = matrix(0,nt)

  # Funciones de entrenamiento inicial. Necesarias para la aplicacion del Search Grid:
  trainRF = function(Xoptimos,Y)
  {rf = randomForest(Xoptimos,Y,ntree=hi,mtry=hk,nodesize=hj)}
  predictRF = function(modelo,Xoptimos)
  {predict(modelo,as.matrix(Xoptimos))}

  #Calculo de parametros por metodo iterativo:
  for(i in 1:nt){
    hi=ntree[i]
    mtry=tuneRF(Xoptimos,Y,ntreeTry=hi,doBest=TRUE)
    hk=mtry$mtry
    mtry_opt[i]=hk
    for(j in 1:nno){
      hj=nodesize[j]
      iteracion = crossval(Xoptimos,Y,trainRF,predictRF,ngroup=Folds)
      MSE[i,j] = sum((Y-iteracion$cv.fit)^2)/n
    }
    jnode[i] = which.min(MSE[i,])
    MSEnode[i] = min(MSE[i,])
  }

  #Determinacion de los parametros optimos y las variables de salida:
  itree = which.min(MSEnode)
  mtree_optimo = ntree[itree]
  mtry_optimo = mtry_opt[itree]
  nodesize_optimo = nodesize[jnode[itree]]
  ijoptimo = c(itree,jnode[itree])
  MSEoptimo = MSE[itree,jnode[itree]]
  return(list(mtree_optimo=mtree_optimo,nodesize_optimo=nodesize_optimo,MSE=MSE,
            ijoptimo=ijoptimo,MSEoptimo=MSEoptimo,mtry_optimo=mtry_optimo))

```

Como ya ocurría en el caso de las SVR para el hiperparámetro σ , en cada iteración del Grid Search el parámetro `mtry` debe ser ajustado dado que depende del número de árboles de regresión.

Por otra parte, dado que la función `RandomForest` no dispone de ninguna opción embebida que posibilite el desarrollo del procedimiento de validación cruzada tal como ocurría con la técnica SVR, se ha recurrido a la función `crossval()` del paquete `bootstrap` la cual posibilita la ejecución de dicho procedimiento de validación cuando la técnica evaluada tiene función `predict()` (método por el cual

un algoritmo concreto puede ser ejecutado para obtener los valores Y cuando se le especifica una matriz de datos de partida X .

3.7.2.3 Feature Selection aplicado para la técnica Random Forest

Como ha sido anticipado al inicio del presente apartado, para Random Forest, el método Feature Selection utilizado es el mismo que el descrito en el Apartado 3.7.1.3 para SVR, en concreto el algoritmo *Recursive Feature Elimination* (RFE), procedimiento supervisado que establece el subconjunto de variables óptimas tras evaluar en un algoritmo de aprendizaje distintas combinaciones de variables de referencia. Así pues, en cada simulación se puntuaría la significancia de cada variable a través de la medida del error, y se descartarían aquellas que empeoran los resultados con respecto a la situación de máxima bondad [184].

```
##### SELECCION DE VARIABLES (WRAPPER APPROACH - CARET) #####
#Se opta por Recursive Feature Selection y en concreto la funcion "rfFuncs"
#El proceso se integra en validacion cruzada 10 folds
Features<-rfe(X, as.numeric(Y), sizes = c(seq(1,30,1)),
             rfeControl=rfeControl(functions=rfFuncs,method = 'cv', number=Fold))
Vopt<-Features$optVariables
Xoptimos<-X[,c(Vopt)]

#Guarda la evolucion del error en funcion del numero de entradas consideradas:
nombre=paste("Feature_selection.jpg", sep="")
jpeg(nombre)
plot(Features)
dev.off()
```

El método ha sido implementado a través del paquete Caret [185] del software R Statistics, donde el modelo de selección es idéntico al modelo de aprendizaje puesto que dicho paquete recurre a la librería *randomForest* en el proceso de selección si se especifica que el método Feature Selection debe ser desarrollado mediante la técnica de aprendizaje RF. Además, la selección y el aprendizaje está inmerso en un mecanismo propio de validación cruzada 10 – Folds tal como se observa en la función anteriormente expuesta.

3.7.2.4 Entrenamiento y testeo de la técnica RF

Tras la selección de los hiperparámetros óptimos (*ntree*, *nodesize* y *mtry*) y el conjunto de variables significativas del problema de regresión tratado, se lleva a cabo el proceso de entrenamiento para dar como resultado el algoritmo que describe la relación entre las variables explicativas y la respuesta esperada.

Teniendo en cuenta que el objetivo de este estudio es la validación del método, el proceso 10 – Folds Cross Validation vuelve a ejecutarse esta vez de forma disgregada para calcular las métricas de error y correlación en cada uno de los subconjuntos de la matriz usada para el entrenamiento y la validación. El resultado final de cada métrica se obtiene con la media aritmética de las métricas obtenidas para cada uno de los Folds del problema.


```
##### CALCULO CON METODO RANDOM FOREST Y VARIABLES OPTIMAS #####

train_RF_wrapper = function(Xoptimos,Y,mtree_optimo,nodesize_optimo,mtry_optimo,Folds)
  require(randomForest)
  n = dim(Xoptimos)[1]
  stopifnot(n> 1, dim(Y)[1] == n)
  MSE_test = matrix(0,Folds)
  MAPE_test = matrix(0,Folds)
  MAE_test = matrix(0,Folds)
  ypredict_test = matrix(0,(n/Folds),Folds)
  rsquared = matrix(0,Folds)
  IoA = matrix(0,Folds)
  MSE_test_pot = matrix(0,Folds)
  MAPE_test_pot = matrix(0,Folds)
  SMAPE_test_pot = matrix(0,Folds)
  MAE_test_pot = matrix(0,Folds)
  potenciapredict_test = matrix(0,(n/Folds),Folds)
  rsquared_pot = matrix(0,Folds)
  IoA_pot = matrix(0,Folds)
  #Potencias en aerogenerador:
  MSE_test_Aerog = matrix(0,Folds)
  MAPE_test_Aerog = matrix(0,Folds)
  SMAPE_test_Aerog = matrix(0,Folds)
  MAE_test_Aerog = matrix(0,Folds)
  Aerog_predict_test = matrix(0,(n/Folds),Folds)
  rsquared_Aerog = matrix(0,Folds)
  IoA_Aerog = matrix(0,Folds)

  for (i in 1:Folds) {
    testIndexes = which(folds==i,arr.ind=FALSE)
    xtest = Xoptimos[testIndexes, ]
    ytest = Y[testIndexes]
    xtrain = Xoptimos[-testIndexes, ]
    ytrain = Y[-testIndexes]
    ptest = potencia_ref[testIndexes]
    pot_test = pot_aero[testIndexes]
    ntest = length(ytest)

    #Entrenamiento del modelo y almacenamiento de resultados por iteracion:
    RF_train=randomForest(xtrain,ytrain,mtree=mtree_optimo,mtry=mtry_optimo,
                          nodesize=nodesize_optimo)

    #Testeo con predict y almacenamiento de resultados:
    ytestv = predict(RF_train,xtest)
    ytestv = ifelse(ytestv<0,0.4,ytestv)
    MSE_test[i] = sum((ytestv-ytest)^2)/ntest
    MAPE_test[i] = mean(abs((ytestv-ytest)/ytest))
    MAE_test[i] = mean(abs(ytestv-ytest))
    ypredict_test[,i] = ytestv

    #Calculo del coeficiente de determinacion R2:
    correlacion = matrix(0,ntest)
    for (j in 1:ntest){
      correlacion[j]=(ytest[j]-ytestv[j])^2
    }
  }
}
```

3.7.2.5 Generación de resultados en el algoritmo RF

Esta sección es idéntica la expuesta en el Apartado 3.7.1.5 para el caso de SVR, planteándose los gráficos y guardándose en formato *.csv los principales resultados del análisis entre los que destacan las métricas, los parámetros óptimos finalmente seleccionados así como las estimaciones obtenidas para cada valor objetivo y en cada uno de los folds del proceso de validación cruzada. Tanto los gráficos como los archivos *.csv son guardados de manera automática en el directorio de trabajo.

3.7.3. Artificial Neural Networks

Dado que el fin perseguido en este trabajo en la adaptación de la técnica ANN al estudio del potencial eólico a largo plazo, se asume la misma filosofía que ha sido adoptada en los Apartados 3.7.1 – 3.7.2 para las técnicas Support Vector Regression y Random Forest. Así pues, para la programación de las redes neuronales artificiales se adopta la estrategia de que el algoritmo o núcleo del modelo provenga de una librería consolidada que cuente con el suficiente respaldo para considerar que la versión desarrollada no tiene defectos de diseño y, por tanto, cualquier disminución de la capacidad de aprendizaje es como consecuencia de la estructura del método y no tanto por fallos de programación.

La red neuronal artificial es la técnica de Machine Learning que más ampliamente ha sido usada en multitud de campos del conocimiento, razón por la cual existen numerosas opciones en cuanto a alternativas disponibles en la industria para la implementación de esta estrategia de aprendizaje. Se expone en la Tabla 7 las principales librerías disponibles de acuerdo con el estudio del estado del arte desarrollado en este trabajo.

LIBRERÍAS ARTIFICIAL NEURAL NETWORKS RECOMENDADAS POR LA INDUSTRIA				
Librería	Métodos	Software	Clase	Lenguaje
Neural designer	MLP, RBF.	Protegido	Clasif/Regr	-
GMDH Shell	Regresión multivariable.	Protegido	Clasif/Regr	-
Neuroph	Varios (librería).	Protegido	Clasif/Regr	Java
Darknet	MLP, RBF.	Libre	Clasif/Regr	C/CUDA
DeepLearningKit	Deep Learning, Convolutional Neural Networks (CNN).	Libre	Clasif/Regr	C
Tflearn	Deep learning, CNN, Long-Short term memory networks (LSTM), BiRNN, Residual network, Generative network.	Libre	Clasif/Regr	Python
ConvNetJS	MLP, RBF, Deep Learning.	Libre	Clasif/Regr	Java
NeuroSolution	MLP, RBF.	Protegido	Clasif/Regr	-
Torch	MLP, RBF.	Libre	Clasif/Regr	C/CUDA
Keras	MLP, RBF, CNN, Recurrent Neural Networks (RNN), Deep Learning.	Libre	Clasif/Regr	Python/R
nNet	Arquitecturas FeedForward y Multinomial Log-Linear Models.	Libre	Regresión	R
NeuralNet	Arquitecturas Feed-Forward con backpropagation, Resilient BackPropagation (RBP).	Libre	Clasif/Regr	R
H2O	Arquitecturas Feed-Forward, Deep learning.	Libre	Clasif/Regr	Python, R, Java

Tabla 7 Librerías Artificial Neural Networks recomendadas por la industria

LIBRERÍAS ARTIFICIAL NEURAL NETWORKS RECOMENDADAS POR LA INDUSTRIA				
Sklearn	Arquitecturas Feed-Forward con backpropagation, Resilient BackPropagation (RBP).	Libre	Clasif/Regr	Python
Net	Arquitecturas Feed-forward.	Protegido	Clasif/Regr	Matlab
Weka Neural Network	MLP.	Libre	Clasif/Regr	Java
NVIDIA Digits	Deep Learning.	Protegido	Clasif	-
Stuttgart Neural Network Simulator	Interface con las grandes familias de redes neuronales en R.	Libre	Clasif/Regr	R
DeepPy	Arquitecturas Feed-forward.	Libre	Clasif/Regr	Python
MLPNeuralNet	Forward propagation.	Libre	Clasif/Regr	Matlab, Python, R
Synaptic	MLP, Multilayer, Long-Short term memory networks (LSTM).	Libre	Clasif/Regr	Java
DNNGraph	Deep Learning.	Protegido	Clasif	-
NeuralN	Varios (librerías).	Protegido	Clasif/Regr	C
AForge.Neuro	Arquitecturas Feed-forward.	Libre	Clasif/Regr	C
NeuralTalk2	-	Protegido	Clasif	-
Cuda-convnet2	Neural networks, Deep Learning.	Libre	Clasif	CUDA, Python
Knet	Neural networks, Deep Learning.	Libre	Clasif/Regr	Julia
DN2A	Neural networks.	Libre	Clasif	Java
Neon	Deep learning, Convnet NN, LSTM, RNN.	Libre	Clasif/Regr	Python
HNN	Arquitecturas Feed-forward.	Libre	Clasif/Regr	C
Lasagne	MLP, RBF, CNN, RNN, LSTM.	Libre	Clasif/Regr	Python
Gobrain	Feed-Forwards, Elman Recurrent Neural Networks (ERNN).	Libre	Clasif	Python
LambdaNet	Arquitecturas Feed-forward.	Libre	Clasif	Haskell
RustNN	MLP con Backpropagation.	Libre	-	-
Mocha	Deep Learning, CNN.	Libre	Clasif/Regr	Julia, C
Deeplearn-rs	Deep Learning, LSTM.	Libre	Class	-

Tabla 7 Librerías Artificial Neural Networks recomendadas por la industria (continuación)

Si bien en esta tabla no se contemplan todas las opciones existentes en la industria, si se expone la mayor parte de las librerías que podrían utilizarse para los fines perseguidos en este trabajo. Son de especial interés aquellas opciones de código abierto programadas en lenguajes de alto nivel tales como Python, R Statistic o Matlab y que a la vez tenga el potencial de ejecución en multihilo. Asimismo, únicamente se podrían utilizar las alternativas válidas para la resolución de problemas de regresión y en concreto, en base a las conclusiones extraídas del Apartado 3.4, aquellas en las que se hubieran programado el método MLP.

De acuerdo con los principios expresados en el párrafo anterior se selecciona la librería *Nnet* [156]. Se argumentan las razones que han motivada la elección a continuación:

1. La librería *Nnet* es la principal opción elegida por los desarrolladores para la programación de las redes neuronales artificiales en R Statistics cuando el algoritmo de cálculo es MLP y el objetivo es la resolución de problemas de regresión.
2. Al utilizarse R Statistics se garantiza que las tres técnicas de Machine Learning comparadas en este trabajo sean evaluadas en el mismo entorno. Por consiguiente, los métodos de pre-tratamiento y post-procesado son idénticos en las tres alternativas y, se garantiza con ello que en la comparación llevada a cabo las posibles diferencias existentes no son como consecuencia de aspectos ajenos al propio algoritmo de aprendizaje.
3. La librería *Nnet* tiene la categoría de código abierto, por consiguiente, tiene la potencialidad de que dicho algoritmo pueda ser adaptado a las condiciones particulares del estudio llevado a cabo en este trabajo. Por ejemplo, a diferencia de otras alternativas, en esta se pueden implementar métodos de validación diferentes (en este caso 10 Folds Cross Validation).
4. Entre los argumentos que deben ser especificados por el usuario para llevar a cabo la estrategia de aprendizaje, pueden proporcionarse los siguientes criterios relacionados con el Early-Stopping y la precisión del resultado:
 - a. *Wts*: Vector donde se especifica los pesos iniciales del proceso de entrenamiento. Si no se especifica, toma como método la alternativa random.
 - b. *Size*: Número de unidades de la capa oculta.
 - c. *Subset*: Argumento en el que se puede especificar la lista de elementos que deben ser considerados en el proceso de entrenamiento.
 - d. *Maxit*: Número máximo de iteraciones.
 - e. *Abstol* y *restol*: Parada si el error de entrenamiento alcanza el valor especificado por el usuario.
 - f. *MaxWts*: Número máximo permitido de pesos.
5. A pesar de que a dicho paquete se accede mediante el lenguaje de programación R Statistics, las funciones básicas de este algoritmo se ejecutan con base C++, lo que reduce considerablemente los tiempos de procesados.

La red neuronal generada a través del proceso de entrenamiento puede ser usada directamente en otras herramientas estadísticas de R Statistics dado que puede ser convocado con la función *predict()*. Un ejemplo de esta característica es que la red neuronal generada con *Nnet* puede ser usada directamente en el algoritmo Shapley Permutation implementado en el paquete Sensibility.

Se describe en este apartado los principios básicos seguidos para la programación de la técnica de redes neuronales artificiales. De acuerdo con los aspectos tratados en este capítulo, el núcleo de aprendizaje ha sido desarrollado con la librería *Nnet* de R Statistics la cual implementa el algoritmo MLP con función de activación sigmoide y método de entrenamiento backpropagation.

Como para las otras dos técnicas ML evaluadas en este trabajo, el análisis comparativo se centrará en las fases de entrenamiento y posterior validación usando la metodología 10 – Folds Cross Validation. Los posibles efectos relacionados con el sobreajuste debido al número y las características de los inputs se controlan con la aplicación de la técnica *Recursive Feature Elimination* (RFE).

Se detallan en los siguientes subapartados las características particulares del algoritmo desarrollado para el caso de redes neuronales artificiales.

3.7.3.1 Definición de datos de partida

Para el desarrollo de las simulaciones, los datos de partida son extraídos de ficheros *.csv en el mismo formato que para el resto de técnicas evaluadas. Tras su precarga en el entorno de R tanto los inputs como los outputs del proceso de entrenamiento son sometidos a una fase de pretratamiento.

```
# Se cargan las variables de entrada desde archivo csv almacenado en directorio:
X=data.matrix(read.csv("model9x.csv",header=TRUE,sep=";",dec="."))
Y=data.matrix(read.csv("model9y.csv",header=TRUE,sep=";",dec="."))
folds=data.matrix(read.csv("folds.csv",header=TRUE,sep=";",dec="."))
Folds=10
Neurons=c(10,20,25) # Number of units in the hidden layer
Epochs=c(1000) # Maximum number of iterations
Stop=0.001 # Stop if the fit criterion falls below

#Se cargan las librerías que son utilizadas en este algoritmo:
library("BioPhysConnector", lib.loc=~anaconda3/lib/R/library")
library("nnet", lib.loc=~anaconda3/lib/R/library")
library("bootstrap", lib.loc=~anaconda3/lib/R/library")

# Los valores de 0 son sustituidos por un número pequeño para evitar error en MAPE:
X[, seq(1,dim(X)[2],2)][X[, seq(1,dim(X)[2],2)] == 0] <- NA
Y[, seq(1,dim(Y)[2],2)][Y[, seq(1,dim(Y)[2],2)] == 0] <- 5

# Se comprueba que las variables de partida no contienen datos perdidos:
completos = complete.cases(X)
X = X[completos,]
Y = Y[completos,]
completos2 = complete.cases(Y)
Y = Y[completos2,]
X = X[completos2,]

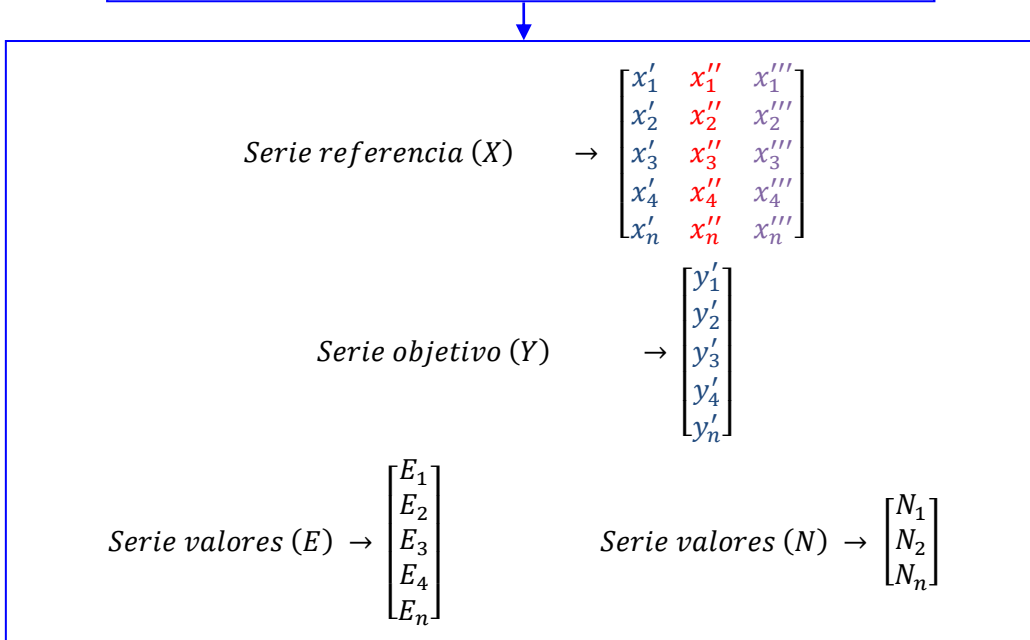
# Luego se comprueba que la serie tiene un número de filas divisible entre 10 (10-folds):
n=dim(X)[1]
if(n%%10!=0){
  exacto=n%%10
  vector=c(1:exacto)
  borrar=n-vector
  X=X[-borrar,]
  Y=Y[-borrar,]
}
# Guarda la serie corregida por si es necesaria su revision:
write.table(data.frame(X),file="X_corregida.csv")
write.table(data.frame(Y),file="Y_corregida.csv")
```

Por lo general, los valores iguales a cero han tenido que eliminarse del proceso de entrenamiento dado que su consideración causaría indeterminaciones en el proceso de validación cuando se plantea la métrica MAPE. Asimismo, se han eliminado los registros *NaN*. A pesar de que estos valores *NaN* han sido eliminados en el proceso de pretratamiento, la función *Nnet* permite que se le especifique a dicho algoritmo la acción de debe realizarse de manera automática cuando un valor *NaN* es descubierto tanto en el set de variables explicativas como en la respuesta.

Como ya ocurría en las técnicas SVR y RF, además de los inputs y el target, debe especificarse al modelo los propios hiperparámetros a tener en cuenta en el proceso de aprendizaje, en este caso, el

número máximo de neuronas en la capa oculta, el número máximo de iteraciones permitidas y la tolerancia máxima del error en la fase de entrenamiento (criterios que condicionan el Early-Stopping de la red neuronal).

<i>Estación objetivo</i>	$\rightarrow Y' = y'_1, y'_2, y'_3, y'_4, y'_n$
<i>Estación de referencia 1</i>	$\rightarrow X' = x'_1, x'_2, x'_3, x'_4, x'_n$
<i>Estación de referencia 2</i>	$\rightarrow X'' = x''_1, x''_2, x''_3, x''_4, x''_n$
<i>Estación de referencia 3</i>	$\rightarrow X''' = x'''_1, x'''_2, x'''_3, x'''_4, x'''_n$
<i>Serie de parámetros epochs</i>	$\rightarrow E = E_1, E_2, E_3, E_4, E_n$
<i>Serie de parámetros neurons</i>	$\rightarrow N = N_1, N_2, N_n$



3.7.3.2 Búsqueda de parámetros óptimos en ANN

Para la selección del número óptimo de neuronas en la capa oculta y el número máximo de iteraciones de la red neuronal se recurre nuevamente al método Grid Search definido con detalle en el Apartado 3.7.1. Los valores iniciales de estos dos vectores fueron establecidos asumiendo que el rango de valores testeados fuera lo suficientemente amplio para reconocer aquella combinación que más se acercara a la respuesta óptima del sistema, todo ello teniendo en cuenta para cada caso el número inputs considerados y la forma de la señal objetivo.

```

recurso_MLP <- function(Xoptimos,Y,Folds=Folds,Neurons=Neurons,Epochs=Epochs,Stop=Stop){
  require(nnet)
  require(bootstrap)
  # Sanity-check inputs
  nx = dim(Xoptimos)[1]
  ny = dim(Y)[2]
  nN = length(Neurons)
  nE = length(Epochs)
  stopifnot(nx > 1, nx == ny)
  erroritera = matrix(0,nrow=nN,ncol=nE)
  errorneuronsbests = matrix(0,nN)
  jneuronsbests = matrix(0,nN)
  colnames(erroritera) = Epochs
  rownames(erroritera) = Neurons

  # Funciones de entrenamiento inicial. Necesarias para la aplicacion del Search Grid:
  trainMLP = function(Xoptimos,Y)
  {mlp = nnet(Xoptimos,Y,size=hh,linout=TRUE,maxit=epoch,trace=FALSE,absto=Stop)}
  predictMLP = function(modelo,Xoptimos)
  {predict(modelo,as.matrix(Xoptimos))}

  for(i in 1:nN){
    hh=Neurons[i]
    for(j in 1:nE){
      epoch=Epochs[j]
      iteracion = crossval(Xoptimos,Y,trainMLP,predictMLP,ngroup=Folds)
      aa = sum((Y - iteracion$cv.fit)^2)/nx
      erroritera[i,j] = sum((Y-iteracion$cv.fit)^2)/nx
    }
    jneuronsbests[i] = which.min(erroritera[i,])
    errorneuronsbests[i] = min(erroritera[i,])
  }
  ihbest = which.min(errorneuronsbests)
  hopt = Neurons[ihbest]
  epochopt = Epochs[jneuronsbests[ihbest]]
  ijopt = c(ihbest,jneuronsbests[ihbest])
  errorcvopt = erroritera[ihbest,jneuronsbests[ihbest]]
  mlp = nnet(Xoptimos,Y,size=hopt,linout=TRUE,maxit=epochopt,trace=FALSE,absto=Stop)
  return(list(mlp=mlp,hopt=hopt,epochopt=epochopt,errorcvopt=errorcvopt))
}

prediccionMLP <- function(modelo,xoptimos)
{predict(modelo,as.matrix(xoptimos))}

```

Como puede extraerse del análisis en detalle del código formulado, la combinación óptima de neuronas en la capa oculta y máximo número de iteraciones que generan el mejor resultado se obtiene con el análisis de aquella alternativa que arroja menor error MSE.

3.7.3.3 Feature Selection aplicado para la técnica ANN

El procedimiento RFE ha sido descrito con detalle en el Apartado 3.7.1.3, usándose como para SVR y RF la función *Backward Elimination*. En este caso, se mantienen los mismos parámetros de simulación que los supuestos para las otras dos técnicas de Machine Learning pero la función de entrenamiento y testeo utilizada en la permutación es en este caso *Nnet*. Adicionalmente, en este método se le especifican los hiperparámetros óptimos calculados en el proceso anterior con la técnica Grid Search.


```
##### SELECCION DE VARIABLES (WRAPPER APPROACH - CARET) #####
#Se opta por Recursive Feature Selection y en concreto la funcion "rfFuncs"
#El proceso se integra en validacion cruzada 10 folds
require(caret)
ctrl<-rfeControl(functions = caretFuncs, method="cv", number=Fold)
my_grid=expand.grid(.size=Neurons, .decay= c(0))
FeatureS<-rfe(X, as.numeric(Y), rfeControl=ctrl,sizes = c(seq(1,30,1)), method='nnet',
              tuneGrid = my_grid,linout=TRUE, trace=FALSE)

Vopt<-Features$optVariables
Xoptimos<-X[,c(Vopt)]

#Guarda la evolucion del error en funci?n del n?mero de entradas consideradas:
nombre=paste("Feature_Selection.jpg", sep="")
jpeg(nombre)
plot(Features)
dev.off()
```

3.7.3.4 Entrenamiento y testeo de la técnica ANN

```
for(i in 1:Folds){
  test.rows = which(folds==i,arr.ind=FALSE)
  xtrain = Xoptimos[-test.rows,]
  ytrain = Y[-test.rows]
  xtest = Xoptimos[test.rows,]
  ytest = Y[test.rows]
  ntest = length(ytest)
  ytrain=log(ytrain)

# Las siguientes funciones llaman a los procesos anteriores:
mlp = recurso_MLP(xtrain,ytrain,Folds=Folds,Neurons=Neurons,Epochs=Epochs,Stop=Stop)
yprediccion = prediccionMLP(mlp$mlp,xtest)
yprediccion = exp(yprediccion)

# Calculo de metricas de error para cada uno de los Folds:
MSE_test[i] = sum((yprediccion-ytest)^2)/ntest
Power_observer = ytest
Power_estimated = yprediccion
completos3 = complete.cases(Power_observer)
Power_observer = Power_observer[completos3]
Power_estimated = Power_estimated[completos3]
completos4 = complete.cases(Power_estimated)
Power_observer = Power_observer[completos4]
Power_estimated = Power_estimated[completos4]
MAPE_test[i] = mean(abs((Power_estimated-Power_observer)/Power_observer))
SMAPE_test[i] = 2*mean(abs(Power_estimated-Power_observer)/abs(Power_estimated+Power_observer))
MAE_test[i] = mean(abs(Power_estimated-Power_observer))
ypredict_test[,i] = yprediccion

# Calculo de coeficiente de determinacion R2:
correlacion = matrix(0,ntest)
for (j in 1:ntest){
  correlacion[j]=(ytest[j]-yprediccion[j])^2

  #La grafica se genera dentro de este bucle para que salga una por cada Fold:
  fitter=lm(yprediccion~ytest)
  p=fitted.values(fitter)
  nombre=paste("Correlacion_", i, ".jpg", sep="")
  jpeg(nombre)
  plot(ytest,yprediccion,type="p",xlab="Serie validacion objetivo",ylab="Estimacion MLP",col="red")
  points(ytest,ytest,type="c",col="green")
  points(ytest,p,type="l",col="blue")
  title(main=list("Coeficiente de correlacion"))
  legend("topleft",legend=c("Data","Y=Predict","Fit"),lty=c(4,1,2),col=c(2,1,3))
  dev.off()
}
SSE = sum(correlacion)
SST = sum((ytest-mean(ytest))^2)
Rsquared[i] = 1-(SSE/SST)
```

Tras finalizar la ejecución del algoritmo RFE ya se conocen los parámetros óptimos (número de neuronas de la capa oculta e iteraciones máximas del problema) y el subconjunto de variables explicativas que aseguran la mejor respuesta de la red neuronal según el conjunto de datos de

entrenamiento disponible. Por consiguiente, en esta etapa se realiza la validación pormenorizada de la capacidad predictiva del modelo generado.

3.7.3.5 Generación de resultados en el algoritmo ANN

Como en el resto de casos, en la sección final del código se generan los resultados en formato *.csv permitiendo el análisis de los resultados. En este caso, además de los ficheros estáticos se guarda las variables de entorno producidas durante la modelización, asegurando que los resultados puedan ser consultados en cualquier momento con el máximo nivel de detalle posible.

Nuevos métodos de estimación de la potencia eólica mediante técnicas de Machine Learning

4.1. Evaluación del recurso eólico en Canarias. Muestra de datos

4.1.1. Introducción

Las técnicas de Machine Learning utilizadas para la estimación tanto del recurso como la potencia eólica presentan la característica común de tratarse de métodos de aprendizaje supervisado. Así pues, la estimación sólo es posible si se cuenta con al menos datos de velocidad del viento recabados en una posición donde los coeficientes de correlación con respecto a la situación objetivo superan el 80%. Adicionalmente, en los estudios desarrollados en esta tesis doctoral se valora la capacidad de las técnicas de Machine Learning seleccionadas para estimar con múltiples estaciones de referencia y múltiples variables, razón por la cual interesa contar con el mayor número posible de series temporales de datos climatológicos relacionados y aptos para ser considerados como referencias en la metodología de aprendizaje.

Por otra parte, aunque el estudio se centra en la estimación de la potencia eólica, la eficiencia de la metodología MCP es medida usando distintas variables objetivo, fundamentalmente la velocidad del viento (V), la densidad de potencia eólica (WPD) y la potencias eólicas (teórica o real) a la salida del aerogenerador ($WTPO$) de acuerdo con los principios de cálculo expuestos en el Capítulo 2. Para todas ellas es vital que se disponga de datos recabados de velocidad y dirección del viento como variables explanatorias [13]. Por otra parte, en los casos en los que se estima las WPD y las $WTPO$ es necesario contar con datos estimados de densidad del aire, para lo cual las estaciones seleccionadas como objetivo y referencias deberían contar con datos de presión atmosférica, humedad relativa y temperatura ambiente. En el proceso comparativo que se desarrolla en este trabajo se toman como muestra de datos las medidas recabadas en estaciones anemométricas ubicadas en diferentes localizaciones de las Islas Canarias.

En el Apartado 4.1 se describen inicialmente las campañas de medición del viento que han sido históricamente desarrolladas en Canarias. Posteriormente, se presentan los datos que han sido seleccionados para la realización de los trabajos y los procesos de filtrado que han sido llevados a cabo para asegurar la mayor validez posible en los datos tomados de partida en el análisis. Finalmente, conforme a un análisis de correlaciones lineales, se decidirá el conjunto de estaciones que serán usadas en los diferentes trabajos llevados a cabo como objetivos y referencias del análisis.

4.1.2. Medición meteorológica de vientos en Canarias

Las primeras campañas de medida del recurso eólico en Canarias fueron realizadas en los aeropuertos, teniendo como fin la planificación de las rutas y las maniobras de aproximación de la aviación a las terminales. La serie de datos de viento más antigua de Canarias procede del aeropuerto de Los Rodeos (Tenerife Norte) donde se comenzó a registrar datos en el año 1931. Se presentan en la siguiente tabla las coordenadas geográficas y los periodos de medición disponibles conforme a la información publicada en el portal web NOAA (National Oceanic Atmospheric Administration) [192]. Todas estas estaciones tienen una altura de 10 metros (Tabla 8).

MEDIDA DEL RECURSO EÓLICO EN AEROPUERTOS CANARIOS						
Código NOAA	Nombre	Abscisa	Norte	Cota	Comienzo medida	Registro disponible
n/a	Santa Cruz	377.606	3.147.689	36.0 m	03/01/1931	84 años
GCCR	Lanzarote	635.951	3.202.804	14.3 m	10/03/1950	65 años
GCLP	Gran Canaria	461.925	3.089.730	23.8 m	16/03/1950	65 años
GCFV	Fuerteventura	611.228	3.147.910	25.3 m	27/03/1950	65 años
GCLA	La Palma	230.551	3.169.768	32.6 m	01/01/1960	55 años
GCHI	El Hierro	215.591	3.080.055	31.4 m	13/03/1973	42 años
GCTS	Tenerife Sur	345.392	3.103.185	63.7 m	01/07/1980	35 años
n/a	La Gomera	282.055	3.102.840	219 m	29/05/2004	11 años

Tabla 8 Medida del recurso eólico en los aeropuertos canarios

Como se puede comprobar con la descarga de los datos accesibles en el portal web, no es hasta bien entrados los años 90 cuando los registros de viento comienzan a mostrar una resolución regular

adecuada para su utilización con fines energéticos (datos medidos cada minuto y promediados con frecuencia horaria), aspecto derivado del proceso de lectura y registro de datos y de los criterios de operación de la aviación y su evolución histórica.

En el periodo comprendido entre 1985 y 1990 comienzan a desarrollarse campañas de medición meteorológicas en el archipiélago con fines relacionados con la energía, habiéndose iniciado estas acciones por los promotores que pretendían conocer la viabilidad técnico económica en la instalación de los parques eólicos. Estas estaciones anemométricas se ubicaron en los puntos donde a priori se asumía mayor potencial para la instalación de parques eólicos de acuerdo con los regímenes de vientos locales. Por otra parte, las campañas de medición meteorológicas no se desarrollaban para periodos superiores a los 2 años [11], momento en el cual se desmantelaban las instalaciones. En algunos casos se optó por proseguir con la medida a través de los registros recabados en los propios aerogeneradores, sin embargo, estos datos no tenían utilidad por su escasa fiabilidad. También existe constancia de que las primeras torres anemométricas instaladas disponían de anemómetros de baja calidad, por lo que su fiabilidad ha quedado en algunos casos en entredicho [193].

La primera campaña para la caracterización del recurso eólico en el ámbito insular data del año 1987, cuando el Departamento de Ingeniería Mecánica (DIM) de la Universidad de Las Palmas de Gran Canaria (ULPGC) inició una campaña de medición a través de 10 estaciones anemométricas ubicadas en Fuerteventura que fueron financiadas por el Cabildo de Fuerteventura y por la Consejería de Industria y Energía del Gobierno de Canarias. En este caso, partiendo de que los fondos financieros eran limitados, se optó por mantener algunas de estas estaciones en una ubicación fija y movilizar las restantes por la geografía insular una vez se dispusiera de un registro aceptable [194,195].

El ensayo realizado en la isla de Fuerteventura motivó a que se emprendieran acciones en el año 1989 para desarrollar campañas de idénticas características en el resto de las islas, dando como resultado el primer Mapa Eólico de Canarias. Este convenio entre la ULPGC y la Dirección de Energía del Gobierno de Canarias contó en su primera fase con las estaciones empleadas para Fuerteventura y con 11 estaciones más desarrollándose un proceso semejante al comentado en el párrafo anterior. Tres años después comenzaba la segunda fase en la que a las 11 estaciones de la primera fase se añadían 8 estaciones financiadas por la ULPGC. Un año después, fruto de un acuerdo de colaboración con UNELCO, se incorporaron al estudio los registros de otras 12 estaciones propiedad de la empresa eléctrica [193].

Ya en el año 1992 se funda el Instituto Tecnológico de Canarias, S.A (ITC) [196] ente adscrito a la Consejería de Economía, Industria, Comercio y Conocimiento del Gobierno de Canarias y que ha venido continuando con la instalación de estaciones anemométricas en zonas aptas para el explotación eólica, ya sea por fines vinculados a la caracterización del recurso o como estudio de viabilidad para la ejecución de nuevos parques eólicos en la isla. Las estaciones promovidas con fines de caracterización planteaban por lo general alturas de 10 metros mientras que para la ejecución de estudios de viabilidad se optaba por instalar torres con alturas coincidentes con la altura de buje de los aerogeneradores a instalar (entre 40 y 90 metros) y con al menos dos posiciones de medida en vertical.

En el año 2005 se encarga a la empresa AWS Truewind un estudio del recurso eólico en las islas empleando para ello el modelo de predicción física MesoMap que integra el modelo MASS de simulación de la atmósfera y el modelo simplificado de flujo de viento WindMap, el cual añade los efectos locales de la orografía y la rugosidad [197]. Los resultados de este estudio conformaron el nuevo Mapa Eólico de Canarias, empleándose bases de dato de reanálisis, en concreto NCEP y NCAR, mediciones de radiosondas y las campañas de medición de vientos realizadas por el ITC desde sus comienzos hasta esa fecha [197]. Fruto de estos trabajos se publicó la aplicación Web “Recurso Eólico de Canarias”, aplicación de libre acceso que tuvo un papel fundamental en los concurso de asignación de potencia eólica iniciados en el 2007 y que permitía concretar una estimación inicial de la producción obtenida por un parque en una posición determinada [198]. Años después el IDAE publica un Mapa Eólico de idénticas características y procedimiento pero en este caso se extendía a toda España [199].

Con los años se ha comprobado que tanto el Mapa Eólico de Canarias desarrollado por el ITC como el publicado por el IDAE presentan limitaciones inherentes a la metodología empleada y la escases de información de partida, produciéndose una sobreestimación de los vientos bajos y una subestimación de los altos debido a los efectos de suavizado. No obstante, se ha comprobado que es una herramienta idónea para los procesos preliminares del proyecto ya que permite localizar y diferenciar las ubicaciones con mayor potencial en el ámbito geográfico [199].

Además de las citadas campañas de medición meteorológicas hay que tener en cuenta aquellas que han sido desarrolladas por la Agencia Estatal de Meteorología (AEMET). Actualmente, esta es la red de estaciones activa de mayor importancia en Canarias con unas 90 estaciones ubicadas en distintos puntos del archipiélago con fines no relacionados a la energía pero que pueden tener utilidad para la aplicación de técnicas MCP. Todas estas estaciones están estandarizadas a una altura de 10 metros.

4.1.3. Estaciones anemométricas consideradas en los estudios

De acuerdo con el análisis realizado en la fase inicial de la tesis doctoral, en Canarias existirían datos recabados en aproximadamente 160 posiciones diferentes ubicadas en las distintas islas del archipiélago canario si se tienen en cuenta las campañas meteorológicas realizadas hasta el momento por la Agencia Estatal de Meteorología (AEMET) y el Instituto Tecnológico de Canarias, S.A. No se consideran en esta estimación, aquellas campañas de medición meteorológicas desarrolladas por particulares así como promotores eólicos ubicados en el archipiélago. De acuerdo con los fines del presente estudio, se realiza la selección de las estaciones anemométricas conforme a los criterios argumentados a continuación:

1. Asegurar el mantenimiento de las condiciones climatológicas y la estacionalidad del recurso eólico. Sólo se emplearan las estaciones anemométricas con **registros continuos anuales** que comiencen en el mes de Enero y finalicen en Diciembre, con independencia de que existan posibles fallos de registro puntuales u outliers.

2. Garantizar la simultaneidad entre series. Las estaciones que formen parte del mismo análisis deben mantener una coherencia en fechas, **garantizando que los datos fueran recabados en el mismo periodo temporal** tanto para las estaciones de referencia como para la serie objetivo.
3. Priorizar las estaciones de mayor fiabilidad. También se considera como aspecto clave la **fiabilidad de las muestras de datos**, priorizando en el análisis las estaciones de las que se tiene constancia que los protocolos de medida aplicados han sido los correctos de acuerdo con los estándares establecidos en el Anexo G de la norma EN 61400-12-1:2007. Se resume en la Tabla 9 los principales aspectos recogidos en esta norma. Dichas recomendaciones permiten reducir la incertidumbre hasta el 1.5 – 2.0%, lo cual se supone una fiabilidad aceptable incluso para el desarrollo de protocolos de certificación de curvas de potencia de aerogeneradores.
4. Usar estaciones que dispongan de **datos medidos de velocidad, dirección del viento, temperatura del aire, presión atmosférica y humedad relativa**.
5. Garantizar que la frecuencia de los datos sea regular. En este caso, **se considera suficiente si durante el periodo de registro se cuenta con valores medios de cada variable en cada hora**.

BUENAS PRÁCTICAS EN LA INSTALACIÓN DE TORRES METEOROLÓGICAS	
Longitud de los vientos (“boom length”).	La longitud de los brazos donde se instala el anemómetro debe ser mayor a 6 veces el diámetro de la torre para geometrías tubulares y de 5 veces para geometrías triangulares. Para geometrías cuadradas debe ser calculada con la formulación expuesta en la norma EN 61.400 – 12 – 1: 2007.
Dirección de los brazos.	Se recomienda instalar los brazos a 45° con respecto a la dirección predominante del viento para torres tubulares y en perpendicular a la misma dirección para torres triangulares. La instalación de los brazos en cualquier otra posición motivaría la aplicación de los coeficientes de correctores expuestos en el Anexo G de la norma EN 61.400 – 12 – 1: 2007.
Separación del anemómetro con respecto al viento.	El anemómetro debe separarse entre 15 y 25 veces el diámetro del viento.
Distancia entre instrumentos.	Superior a 1.5 metros.
Sensores de presión y temperatura	Deben instalarse cercanos a la posición de toma de datos de los anemómetros, siendo conveniente que existan medidas a dos alturas. El termómetro debe contar con un protector de radiación y el sensor de presión tiene que ubicarse en una caja de intemperie bien ventilada.

Tabla 9 Buenas prácticas en la instalación de torres meteorológicas

BUENAS PRÁCTICAS EN LA INSTALACIÓN DE TORRES METEOROLÓGICAS	
Anemómetros	<p>Considerar los siguientes aspectos:</p> <ul style="list-style-type: none"> ▪ Los sensores deben estar montados en un tubo redondo vertical del mismo diámetro al utilizado para la calibración, conduciendo el cable por su interior y no existiendo ningún elemento que pueda perturbar el flujo en una longitud mínima de 0.75 metros. ▪ Su desviación con respecto a la proyección vertical no debería ser menor de 2°. ▪ El mástil que sostiene el anemómetro deberá estar introducido en el cuerpo de la torre para una longitud no inferior a 1:5 veces su tamaño. ▪ Ningún aparato de medida puede estar a menos de 1.5 metros sobre la cruceta de soporte de la torre. ▪ Conviene instalar un segundo anemómetro por altura de medida que ejerza la función de control, determinando problemas del primario e incluso tomando sus medidas de referencia en caso de que se detectaran fallos. Para su instalación existen dos opciones: <ul style="list-style-type: none"> - Montaje vertical: Entre 1.5 y 2.5 metros por debajo anemómetro principal y dentro del 10% de la altura de buje - Montaje paralelo: Anemómetros principal y de control montados a la misma altura por medio de una cruceta soporte que distancie los sensores entre 1.5 y 2.5 metros. El tubo soporte vertical se distanciaría de la torre entre 15 y 25 veces el diámetro del tubo. ▪ Conviene instalar sensores en al menos dos alturas para definir el perfil de viento vertical.
Veletas	<p>Las recomendaciones son las siguientes:</p> <ul style="list-style-type: none"> ▪ La veleta debe tener su punto de referencia orientado al norte verdadero y no al norte magnético. En cualquier caso, en la ficha de la estación se debe indicar el criterio que se haya seleccionado. ▪ Se respetarán las distancias de separación citadas anteriormente con respecto a los anemómetros. ▪ La veleta se instalará sobre la propia cruceta soporte cuando esta exista alejada de otros elementos que pudieran afectar a la medida.
Calibración de los sensores	<p>La recomendación es que los sensores hayan sido calibrados en un periodo que va desde 1 a 3 años en función del medio donde se encuentre dicho equipo.</p>
Cables de conexión.	<p>Los cables deben transcurrir por el interior de los tubos evitando la obstaculización del flujo y la producción de estelas que puedan suponer un aumento de las incertidumbres.</p>

Tabla 9 Buenas prácticas en la instalación de torres meteorológicas (continuación)

Se exponen en la Figura 23 las estaciones anemométricas seleccionadas de acuerdo con los criterios anteriores para las investigaciones llevadas a cabo en la presente tesis doctoral.

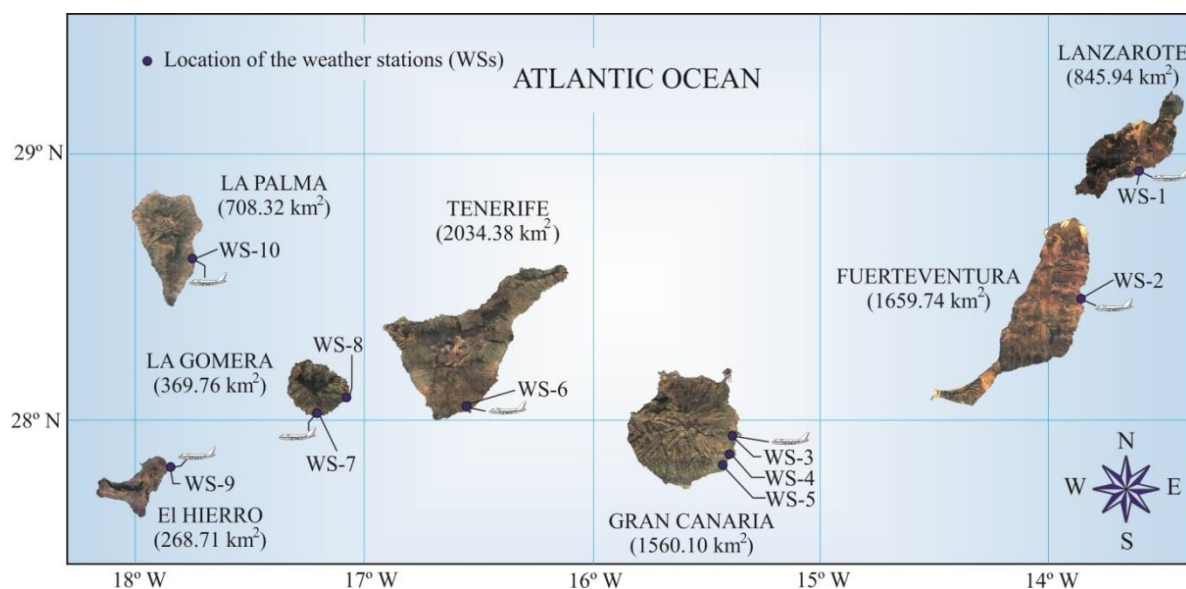


Figura 23 Posiciones de las estaciones anemométricas [254]

Se dispone de datos climatológicos medidos durante 2014 cada hora en 10 estaciones meteorológicas (WS) instaladas en las Islas Canarias (España). Los datos capturados en las estaciones WS-1, WS-2, WS-3, WS-6, WS-7, WS-8, WS-9, WS-10 fueron concedidos exclusivamente para los fines de investigación de la presente tesis doctoral por la Agencia Estatal de Meteorología (AEMET). Por su parte, los datos de las estaciones WS-4 y WS-5 fueron facilitados con el mismo fin por el Instituto Tecnológico de Canarias, S.A.

De los datos disponibles en cada estación meteorológica se utilizan las medidas de velocidad y dirección del viento, así como los datos de temperatura, presión atmosférica y humedad relativa del aire, estos últimos para el cálculo de la densidad del aire en aquellas investigaciones relacionadas con la estimación de la potencia eólica.

Los anemómetros y las veletas de las estaciones propiedad de la AEMET están localizados a 10 metros de altura con respecto al nivel del suelo. En el caso de las estaciones propiedad del ITC, las medidas han sido recabadas a distintas alturas, sin embargo, se ha optado por aquellas alturas donde la medida se sitúa en la posición más próxima a 10 metros de cuantas opciones estuvieran disponibles en ambas posiciones para asegurar la mayor compatibilidad estadística en comparación con los datos recabados por la AEMET. Por todo ello, en el caso de la estación WS-4 la altura seleccionada de instalación del anemómetro y la veleta es 20 metros mientras que en WS-5 dicha altura asciende hasta 10 metros.

En lo que respecta a los termómetros y los higrómetros éstos fueron instalados a 2 metros de altura sobre el nivel del suelo para todas las posiciones mientras que los barómetros se encuentran

posicionados en la misma altura en la que se ubican los sensores de medida de velocidad y dirección del viento.

En la Tabla 10 se muestran los códigos asignados a cada estación meteorológica, las alturas sobre el nivel del suelo a las que han sido captadas dichas medidas. Dicha tabla resume adicionalmente los valores medios, desviaciones estándar, valores máximos y mínimos de las velocidades del viento, presiones atmosféricas, humedades relativas y temperaturas medidas durante el año 2014 con frecuencia horaria. De la misma forma, se exponen los mismos valores estadísticos para el caso de la densidad del aire calculada por medio de los datos anteriores y el procedimiento de cálculo CIPM 2007 [78].

Por otra parte, en las Figura 24 – 25 se presentan los histogramas de frecuencia para velocidades del viento y densidades de potencia eólica, así como las rosas de viento y los valores medios mensuales y diarios graficados para cada una de las estaciones anemométricas consideradas.

De acuerdo con la información presentada, una de las principales características de los datos meteorológicos empleados es su ubicación, ya que dichos datos proceden de estaciones localizadas al nivel del mar para todos los casos y donde la altitud es inferior a 65 metros salvo para la estación WS-7 (estación ubicada en el Aeropuerto de La Gomera). Por consiguiente, los cambios en términos de presión atmosférica, humedad relativa y temperatura ambiente son mínimos y, por tanto, las estimaciones de densidad del aire arrojan diferencias poco significativas entre estaciones tanto en términos medios como en desviaciones típicas.

Por otra parte, se puede comprobar que las velocidades medias del viento son por lo general superiores a 6 m s^{-1} en todas las posiciones, registrándose además velocidades de hasta 12 m s^{-1} con frecuencias elevadas, por lo cual desde el punto de vista del recurso eólico, la mayoría de las posiciones serían aptas para la propuesta de casos a estudio pudiéndose simular el comportamiento con aerogeneradores reales. Asimismo, para ninguno de los casos las velocidades máximas superan los 25 m s^{-1} , aspecto indicativo de que durante el periodo de muestreo no se produjeron fenómenos meteorológicos anormales conforme a las condiciones del recurso eólico de Canarias.

En lo que respecta a las direcciones del viento, destaca la presencia de un comportamiento marcado por la predominancia de los vientos Alisios, cuyas componentes principales provienen de direcciones Norte – Noreste para todas las posiciones salvo la estación WS-7. En dichas posiciones existen periodos donde se producen algunos cambios en la distribución de frecuencias del viento proviniendo estos fundamentalmente del Sur - Suroeste. Si se tiene en cuenta la posición relativa de cada estación en las islas, este fenómeno se debe inequívocamente al efecto de las borrascas sobre el archipiélago. En el caso de la estación WS-7, la rosa de vientos obtenida es coherente con la posición geográfica en la que ésta se encuentra. Si se analiza la variación mensual de los valores de velocidad del viento medio se puede comprobar que los meses del periodo estival son para todas las posiciones salvo la estación WS-7 donde se producen los mayores regímenes de viento (influencia de los vientos alisios en Canarias) mientras que en los meses de Septiembre y Octubre se produce una caída en los valores de velocidad media mensual.

ESTACIONES ANEMOMÉTRICAS EMPLEADAS																				
Estaciones	Altitud (m)	Velocidad del viento (m/s)			Presión atmosférica (hPa)				Humedad relativa (%)				Temperatura (Celsius)				Densidad del aire (kg/m ³)			
		Media	SD	Max	Media	SD	Max	Min	Media	SD	Max	Min	Media	SD	Max	Min	Media	SD	Max	Min
WS-1	14	6.07	3.26	16.39	1015.86	4.27	1.027	993	68.60	12.38	98.0	14.0	21.11	3.87	33.4	10.3	1.183	0.022	1.247	1.126
WS-2	25	6.07	2.72	18.33	1014.95	4.39	1.026	992	70.12	10.14	95.0	18.0	20.90	3.29	31.9	11.6	1.181	0.019	1.243	1.131
WS-3	24	7.73	3.63	18.90	1013.12	4.00	1.024	993	67.66	9.95	100.0	12.0	21.27	3.26	35.0	11.1	1.179	0.019	1.236	1.133
WS-4	7	8.67	5.07	25.00	1010.62	5.30	1.025	990	72.37	7.67	96.0	17.6	20.35	2.42	31.8	10.6	1.197	0.016	1.245	1.153
WS-5	6	8.47	3.57	19.54	1019.92	4.14	1.031	988	68.72	8.92	92.3	21.1	22.10	4.72	37.4	7.3	1.182	0.028	1.264	1.086
WS-6	64	5.35	3.21	18.61	1009.83	3.94	1.020	989	66.23	11.44	97.0	11.0	20.77	3.44	35.4	11.2	1.178	0.019	1.243	1.125
WS-7	219	2.47	2.07	21.11	993.39	3.76	1.004	978	70.67	11.69	99.0	12.0	19.71	3.08	32.0	10.7	1.163	0.018	1.213	1.118
WS-8	15	3.52	1.82	10.83	1016.65	3.97	1.028	999	72.27	8.71	97.0	17.0	21.13	2.96	31.3	12.3	1.183	0.018	1.239	1.139
WS-9	32	6.65	2.63	19.17	1015.80	3.89	1.027	998	71.82	9.02	100.0	8.0	21.23	2.69	32.5	12.9	1.181	0.016	1.224	1.142
WS-10	33	4.97	2.88	18.61	1014.59	4.34	1.027	993	69.42	8.67	99.0	24.0	20.68	2.92	34.0	12.3	1.183	0.018	1.236	1.137

Tabla 10 Estaciones anemométricas empleadas [254]

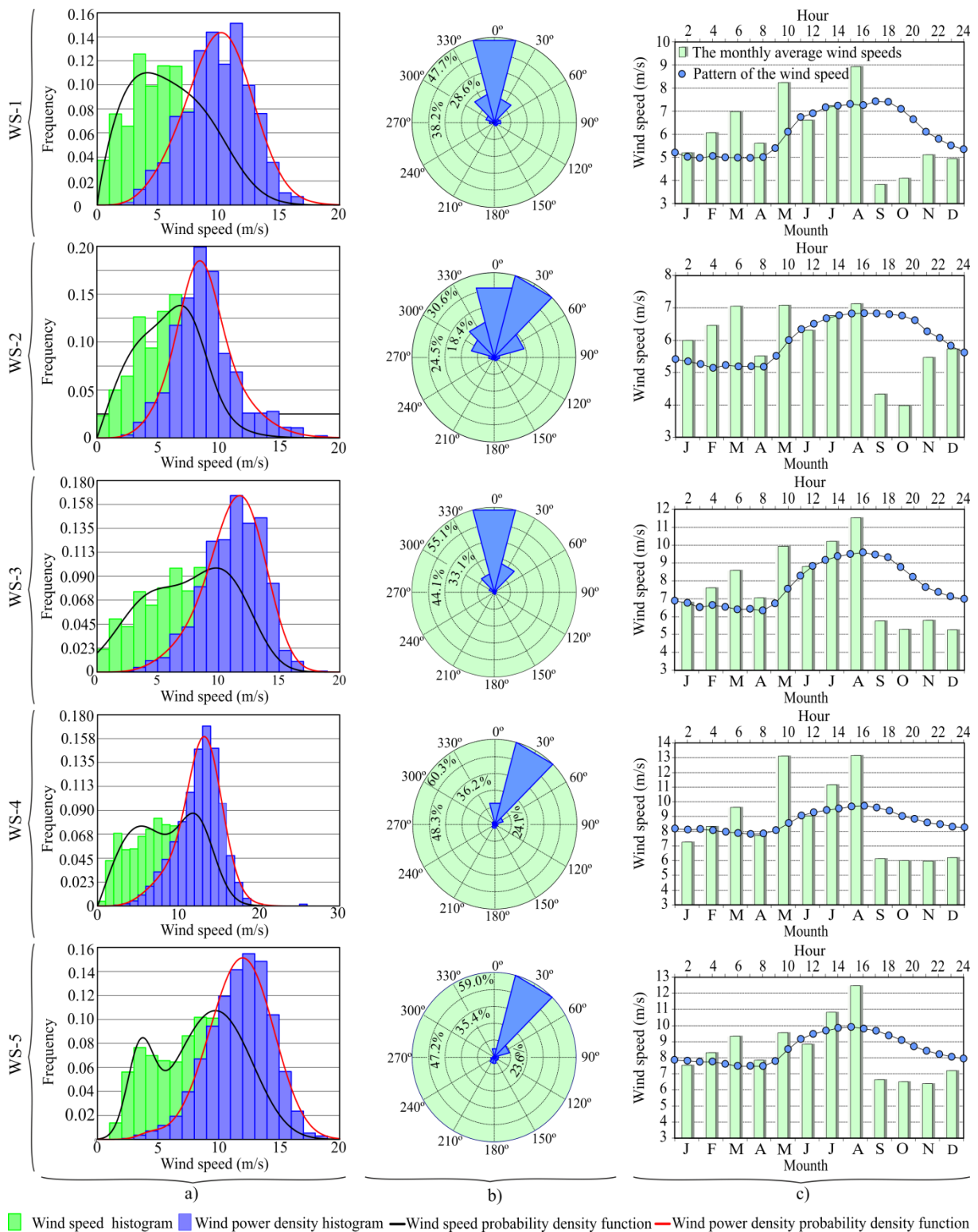


Figura 24 Histogramas de velocidad y densidad de potencia, rosas de viento y perfiles medios mensuales y diarios para las estaciones WS-1 a WS-5 [268]

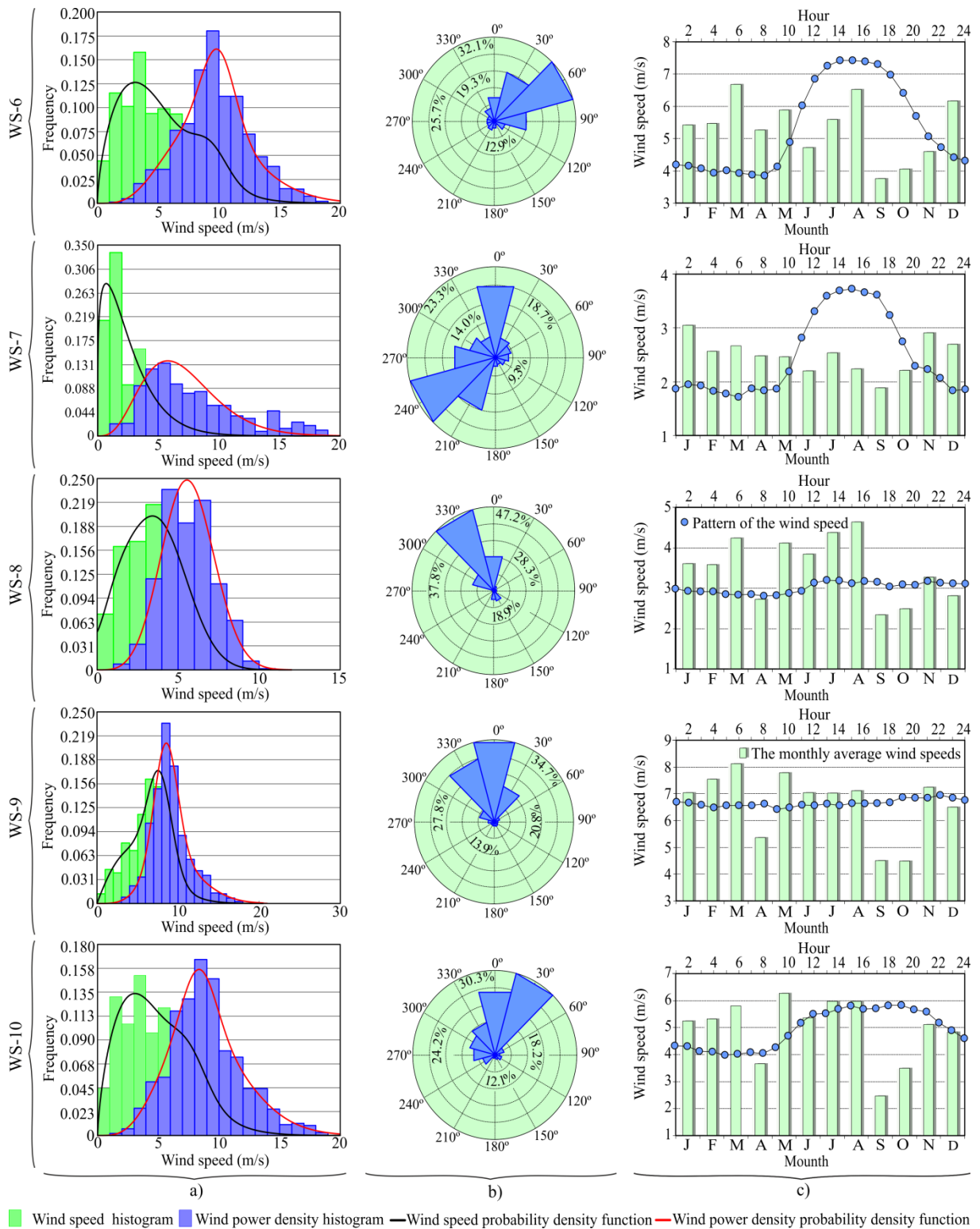


Figura 25 Histogramas de velocidad y densidad de potencia, rosas de viento y perfiles medios mensuales y diarios para las estaciones WS-6 a WS-10 [268]

En lo referente a la variación diaria del recurso eólico, destaca que por cuestiones relacionadas con la transparencia del aire, en los periodos diurnos se produce un calentamiento de la superficie terrestre que no se produce en horario nocturno. Estas diferencias de temperatura generan gradientes de presión que modifican las condiciones de estabilidad atmosférica entre la noche y el día. Así pues, en las horas nocturnas, la atmósfera tiende a ser estable y, por tanto, no existen grandes transferencias de energía entre las capas inferiores y superiores de la misma. Sin embargo, en las horas de sol, existe un mayor potencial para que se produzcan atmósferas inestables que de por sí generan vientos en superficie. Asimismo, cuando la atmósfera es inestable, el aire tiene menos dificultad para subir y superar obstáculos con facilidad, mientras que con atmósferas estables las capas superiores ejercen un “efecto tapón” que obliga al aire a buscar caminos diferentes, sorteando los posibles obstáculos que pudieran existir en la superficie terrestre [46,48].

Analizando lo sucedido en estas estaciones meteorológicas, se observa que el comportamiento es coherente con la teoría expuesta, existiendo un incremento progresivo del recurso eólico durante el periodo diurno, el cual toma su valor máximo en las horas de mayor incidencia solar en las islas (13:00 – 15:00). Seguidamente, se produce una rampa descendente hasta estabilizarse a partir de las 23:00. La variación del recurso eólico entre la noche y el día depende de la estación analizada, no obstante, puede comprobarse que en términos medios el aumento de la velocidad del viento es de entre $2 - 2.5 \text{ m s}^{-1}$.

4.1.4. Análisis inicial del recurso mediante la validación de las muestras de datos

Para todas las estaciones anemométricas seleccionadas los datos han sido recabados de sus registradores y traducidos a formato *.csv, archivos que pueden ser interpretados por todas las herramientas de análisis de datos empleadas para el desarrollo de la tesis doctoral. Posteriormente, se lleva a cabo un proceso de **control de calidad y validación de las muestras de datos** con anterioridad a su utilización para la correlación a largo plazo.

El control de calidad y validación de los datos de partida tiene como objetivo asegurar la mayor precisión posible en los resultados finales, garantizando que los procesos de correlación nunca sean ejecutados cuando existen contrariedades en las series utilizadas.

En la industria eólica se han desarrollado distintas herramientas informáticas que tienen por objetivo detectar de forma automática mediante técnicas estadísticas básicas aquellos datos que presentan irregularidades, para seguidamente decidir si estos datos son válidos o deben ser eliminados. Se distinguen dos fases en este proceso de validación [94,200]:

1. **Control de calidad:** Esta fase que puede ser automática tiene por objetivo la detección de datos sospechosos sin tomar aún ninguna decisión en cuanto al cribado de la información. Inicialmente se realiza un chequeo general de la estructura de los datos verificando que la secuencia de tiempo es continua y que no existen problemas obvios como secciones de la serie en la que los valores de velocidad del viento son cero o devuelven un valor del tipo “NaN”.

Seguidamente se realiza un control de los parámetros de medida por medio de los test de rango, tendencias y relaciones:

- a. **Test de rango:** Con este test se comprueban que los datos de viento son normales y que por tanto se encuentran dentro de los límites típicos en función de sus valores medios, máximos, mínimos y de desviación estándar [39,94].

VALORES TIPOS DE TEST DE RANGO		
Parámetros	Límite inferior	Límite superior
Velocidad media del viento horizontal	0 m s ⁻¹	30 m s ⁻¹
Desviación estándar de la velocidad del viento horizontal	0 m s ⁻¹	3 m s ⁻¹
Valor máximo de ráfagas	0 m s ⁻¹	35 m s ⁻¹
Velocidad media del viento vertical (Terreno simple – complejo)	0 m s ⁻¹	(2 – 4 m s ⁻¹)
Desviación estándar de la velocidad del viento vertical	0 m s ⁻¹	(1 – 2 m s ⁻¹)
Valor máximo de ráfagas (Terreno simple – complejo)	0 m s ⁻¹	(3 – 6 m s ⁻¹)
Valor medio de la dirección del viento	0°	360°
Desviación estándar de la dirección del viento	3°	75°

Tabla 11 Valores tipos de test de rango [182]

- b. **Test de tendencias:** Se analiza las tasas de cambio de las variables con el paso del tiempo. Este test no es aplicable a la dirección del viento puesto que existen ocasiones en los que si son posibles los cambios de dirección de forma abrupta. Se consideran fuera de lo normal cambios horarios de la velocidad del viento por encima de 5 m s⁻¹ [94].
- c. **Test relacional:** Sólo se realiza cuando existen medidas recabas en estaciones próximas o en la propia torre a diferente altura. El objetivo es detectar situaciones improbables en las que se observan diferencias por encima de lo normal o cambios de tendencias entre registros [39,94].

VALORES TIPOS DE TEST RELACIONAL	
Parámetros	Criterio de validación
Máxima rafagosidad de la velocidad del viento en relación a la velocidad media.	Max. Rafaga ≤ 2.5 x Avg.
60/40 m. Diferencia de velocidades medias del viento.	≤ 3 m s ⁻¹
60/40 m. Máxima diferencia de velocidades de viento.	≤ 5 m s ⁻¹
60/25 m. Diferencia de velocidades medias del viento.	≤ 5 m s ⁻¹
60/25 m. Máxima diferencia de velocidades de viento.	≤ 8 m s ⁻¹
Misma altura. Diferencia de velocidades medias del viento.	≤ 0.5 m s ⁻¹
Misma altura. Máxima diferencia de velocidades de viento.	≤ 3.0 m s ⁻¹
60/25 m. Diferencias de direcciones medias.	≤ 20°
60/25 m. Cizalladura del viento medio.	-0.05 < α ^b < 0.45

Tabla 12 Valores tipos de test relacional [182]

En la Figura 26 se ejemplifica los aspectos mencionados sobre la detección de posibles errores, en este caso para las series de la estación WS-4, estación anemométrica ubicada en la isla de Gran Canaria.

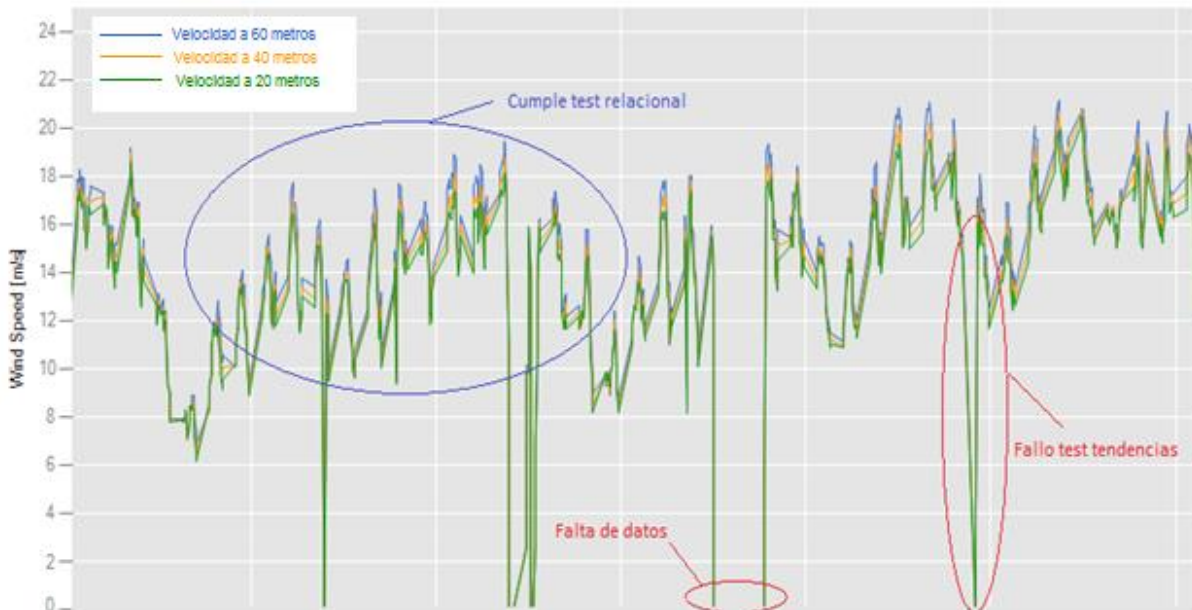


Figura 26 Control de calidad y detección de errores WS-4

Este proceso se realiza con todas las series temporales utilizando para ello el software libre WindData Explorer [201] del fabricante de aerogeneradores Vestas, donde a través de la aplicación de filtros se detectan los posibles errores automáticamente.

- 2. Verificación:** Una vez detectados los datos sospechosos se procede con el análisis propiamente dicho, discerniendo la estrategia de actuación sobre los datos señalados. Algunos valores inusuales que a priori podrían clasificarse como datos erróneos pueden ser debidos a condiciones meteorológicas y por tanto no deberían eliminarse del análisis. Por el contrario un aumento progresivo de las tasas de error podrían indicar la degradación del sensor o problemas en el Data logger y por tanto deberían ser eliminados. La decisión final se realizará de manera manual estudiando cada caso particular en concreto.

Una vez conocidos los datos erróneos hay que decidir cómo proceder con ellos. Algunos autores [175,202] han propuesto sistemas basados en el ajuste a la regresión por el cual los valores erróneos son sustituidos por el valor de la regresión en ese mismo momento, manteniendo la tendencia del recurso durante este periodo temporal. En cualquier caso, esta estrategia supone una manipulación previa de los datos brutos que no es aconsejable para estudios MCP ya que puede derivar en problemas de sobreajuste.

Tras haber aplicado el análisis expuesto en el presente apartado, se resume en la Tabla 13 la disponibilidad total por estación y variable.

DISPONIBILIDAD DE LOS DATOS					
WS	Velocidad del viento	Dirección del viento	Presión atmosférica	Humedad relativa	Temperatura ambiente
WS-1	99.7%	99.7%	94.6%	99.5%	99.5%
WS-2	94.0%	94.0%	86.1%	94.0%	94.1%
WS-3	99,9%	99.9%	9,5%	95.1%	97.9%
WS-4	95.3%	95.3%	95.3%	95.3%	95.3%
WS-5	97.3%	97.3%	94.3%	94.3%	94.3%
WS-6	99.8%	99.8%	96.8%	99.3%	99.3%
WS-7	98.6%	98.6%	99.5%	98.0%	96.7%
WS-8	92.6%	92.6%	90.7%	90.7%	90.3%
WS-9	99.8%	99.8%	94.4%	99.7%	100%
WS-10	99.9%	99.9%	95.0%	99.9%	99.9%

Tabla 13 Disponibilidad de los datos

Así pues, se concluye que la disponibilidad media de los datos empleados supera el 90% para todas las estaciones y variables. La variable con mayores indisponibilidades es la presión atmosférica para todas las estaciones, siendo dicha disponibilidad coincidente con la que existiría para la densidad del aire ya que en el cálculo de esta última son necesarias las alturas simultáneas de presión, humedad relativa y temperatura.

Para los estudios desarrollados en la presente tesis doctoral, **se opta por eliminar en todas las series temporales aquellas horas con indisponibilidad en cualquiera de las variables y estaciones consideradas**, sólo desarrollando los procesos de entrenamiento y testeo con los datos originales en los que se ha comprobado un comportamiento consistente.

4.1.5. Análisis de correlaciones lineales y selección de estaciones que actúan como referencias y objetivos

Uno de los aspectos indicativos de la idoneidad de los datos disponibles para la estimación a largo plazo con múltiples estaciones de referencia y variables es el coeficiente de correlación lineal existente entre las estaciones implicadas en la estimación. Teniendo en cuenta el carácter experimental de los trabajos desarrollados, no se define de antemano una posición concreta en la que se pretenda estimar las características a largo plazo de la potencia eólica. En este sentido, las estaciones objetivo y sus referencias son definidas mediante un análisis de correlaciones lineales asegurando que inexorablemente se cumplan los criterios fijados para la aplicación de la metodología MCP [13] tal como se resume en el Apartado 2.9 de este trabajo.

Se presenta en la Tabla 14 un análisis de correlaciones lineales practicado mediante el método de Pearson tomando como referencias las velocidades del viento horarias de las 10 estaciones seleccionadas. Dicho análisis de correlaciones lineales también es realizado para las densidades de potencia eólica (*WPD*) en la Tabla 15 dado que esta variable también será estimada a largo plazo empleando la metodología MCP.

ANÁLISIS DE CORRELACIONES LINEALES ENTRE VELOCIDADES DEL VIENTO										
WS	WS-1	WS-2	WS-3	WS-4	WS-5	WS-6	WS-7	WS-8	WS-9	WS-10
WS-1	1.00	0.73	0.74	0.67	0.69	0.54	0.26	0.56	0.52	0.56
WS-2	0.73	1.00	0.68	0.58	0.63	0.55	0.30	0.54	0.58	0.57
WS-3	0.74	0.68	1.00	0.82	0.83	0.53	0.23	0.60	0.51	0.58
WS-4	0.67	0.58	0.82	1.00	0.83	0.51	0.18	0.55	0.45	0.50
WS-5	0.69	0.63	0.83	0.83	1.00	0.63	0.21	0.55	0.42	0.50
WS-6	0.54	0.55	0.53	0.51	0.63	1.00	0.39	0.36	0.31	0.40
WS-7	0.26	0.30	0.23	0.18	0.21	0.39	1.00	0.23	0.24	0.34
WS-8	0.56	0.54	0.60	0.55	0.55	0.36	0.23	1.00	0.56	0.55
WS-9	0.52	0.58	0.51	0.45	0.42	0.31	0.24	0.56	1.00	0.56
WS-10	0.56	0.57	0.58	0.50	0.50	0.40	0.34	0.55	0.56	1.00

Tabla 14 Análisis de correlaciones lineales entre velocidades del viento [254]

ANÁLISIS DE CORRELACIONES LINEALES ENTRE DENSIDADES DEL POTENCIA EÓLICA										
WS	WS-1	WS-2	WS-3	WS-4	WS-5	WS-6	WS-7	WS-8	WS-9	WS-10
WS-1	1.00	0.60	0.69	0.61	0.62	0.48	0.16	0.49	0.40	0.37
WS-2	0.60	1.00	0.56	0.44	0.49	0.51	0.36	0.43	0.53	0.43
WS-3	0.69	0.56	1.00	0.77	0.82	0.52	0.18	0.51	0.33	0.38
WS-4	0.61	0.44	0.77	1.00	0.78	0.45	0.12	0.47	0.28	0.27
WS-5	0.62	0.49	0.82	0.78	1.00	0.58	0.10	0.49	0.26	0.27
WS-6	0.48	0.51	0.52	0.45	0.58	1.00	0.37	0.37	0.29	0.23
WS-7	0.16	0.36	0.18	0.12	0.10	0.37	1.00	0.17	0.30	0.28
WS-8	0.49	0.43	0.51	0.47	0.49	0.37	0.17	1.00	0.41	0.39
WS-9	0.40	0.53	0.33	0.28	0.26	0.29	0.30	0.41	1.00	0.44
WS-10	0.37	0.43	0.38	0.27	0.27	0.23	0.28	0.39	0.44	1.00

Tabla 15 Análisis de correlaciones lineales entre densidades de potencia eólica [254]

De acuerdo con la Tabla 14 los mayores coeficientes de correlación se producen entre las estaciones WS-3, WS-4 y WS-5, estaciones ubicadas en la isla de Gran Canaria, donde debido a su proximidad geográfica, todas las estaciones se encontrarían afectadas por las mismas condiciones meteorológicas. A este respecto, los coeficientes de correlación entre las estaciones WS-3, WS-4 y WS-5 superan el valor de 0.8 y, por tanto, podrían ser catalogadas como “buenas” (rango de 0.8 – 0.9) en concordancia con la literatura especializada en materia de ajuste a largo plazo MCP [13].

En el caso de las densidades de potencia eólica (Tabla 15) las correlaciones entre estaciones son ligeramente inferiores a las registradas para velocidades del viento. Así pues, si se utilizaran las reglas generales recomendadas para velocidades en los métodos MCP [13], los coeficientes de correlación

entre densidades de potencia eólica podrían catalogarse como “buenos” (0.8 – 0.9) entre las estaciones WS-3 y WS-5 y “moderados” (0.7 – 0.8) entre las estaciones WS-4 y WS-5.

Conforme a lo anterior y de acuerdo con el estudio del estado del arte realizado, cuando los métodos MCP se han apoyado en la información proporcionada por una única estación de referencia, se han desechado como referencia aquellas estaciones en las que el coeficiente de correlación era inferior a 0.70 [13], dado que se ha considerado que los datos de referencia presentarían una elevada incertidumbre al tratar de estimar las condiciones eólicas a largo en la ubicación objetivo. Sin embargo, en los modelos MCP que permiten estimar dichas circunstancias usando varias estaciones de referencia, no debe descartarse a priori el empleo de ninguna de las estaciones de referencia aunque sus coeficientes de correlación sean bajos, siempre y cuando al menos una de estas referencias manifestara una correlación alta [17]. En este último caso, los posibles problemas derivados del efecto de sobreajuste podrían ser eliminados usando métodos Feature Selection antes de la ejecución de las técnicas de aprendizaje estadístico.

De acuerdo con los principios manifestados en el párrafo anterior, las condiciones mínimas para la aplicación de la metodología MCP **sólo podrían ser alcanzadas si el estudio tratara de averiguar las condiciones del potencial eólico a largo plazo en las estaciones WS-3, WS-4 y WS-5** dado que si se seleccionara alguna de estas posiciones como objetivo y el resto de estaciones fueran referencias, para cada uno de los tres análisis (uno por estación) **al menos una de las referencia presentarían correlaciones “buenas” con respecto al objetivo como se demuestra en la Tabla 16 para el caso de velocidades del viento y en la Tabla 17 para el caso de densidades de potencia.**

ANÁLISIS DE CORRELACIONES LINEALES ENTRE ESTACIONES DE REFERENCIA Y OBJETIVO (VELOCIDAD)										
Objetivo	Referencias									
	WS-5	WS-4	WS-3	WS-1	WS-2	WS-8	WS-10	WS-6	WS-9	WS-7
WS-3	0.83	0.82	-	0.74	0.68	0.60	0.58	0.53	0.51	0.23
WS-4	0.83	-	0.82	0.67	0.58	0.55	0.50	0.51	0.45	0.18
WS-5	-	0.83	0.83	0.69	0.63	0.55	0.50	0.63	0.42	0.21

Tabla 16 Análisis de correlaciones lineales entre estaciones de referencia y objetivo (Velocidad del viento)

ANÁLISIS DE CORRELACIONES LINEALES ENTRE ESTACIONES DE REFERENCIA Y OBJETIVO (WPD)										
Objetivo	Referencias									
	WS-5	WS-4	WS-3	WS-1	WS-2	WS-6	WS-8	WS-10	WS-9	WS-7
WS-3	0.82	0.77	-	0.69	0.56	0.52	0.51	0.38	0.33	0.18
WS-4	0.78	-	0.77	0.61	0.44	0.45	0.47	0.27	0.28	0.12
WS-5	-	0.78	0.82	0.62	0.49	0.58	0.49	0.27	0.26	0.10

Tabla 17 Análisis de correlaciones lineales entre estaciones de referencia y objetivo (WPD)

4.2. Análisis comparativo entre varios modelos MCP que usando la técnica SVR estiman la densidad de potencia

4.2.1. Introducción

Las densidades de potencia eólica (Wind Power Density – WPD) medias a largo plazo constituye un indicador relevante del viento como fuente de potencia que se suele incluir en los mapas de recurso eólico regionales como información previa útil en la identificación de sitios atractivos para la instalación de proyectos eólicos. En el Apartado 4.2 se comparan ocho modelos de MCP propuestos para estimar las WPDs en un sitio objetivo. Siete de dichos modelos utilizan la técnica Support Vector Regression (SVR) mientras que el octavo emplea la técnica Multiple Linear Regression (MLR), la cual sirve de base en la comparación con respecto al resto de modelos diseñados. Además, se ha utilizado una técnica Wrapper con validación cruzada de 10 folds para seleccionar el conjunto óptimo de variables de entrada a los modelos SVRs y MLR.

De los ocho modelos, algunos se han entrenado para estimar directamente los valores medios horarios de las WPDs en el sitio objetivo, mientras que en otros casos el entrenamiento se desarrolla adoptando como variable respuesta los parámetros de las que aquellas dependen. Las variables explicativas consideradas son distintas combinaciones de la media horaria de velocidad y dirección del viento así como las densidades del aire recabadas durante 2014 en 10 estaciones anemométricas instaladas en distintas posiciones del archipiélago canario de acuerdo con la información presentada en el Apartado 4.1.

Fruto del estudio desarrollado en este apartado ha sido publicado un artículo [203] en la revista indexada con índice JRC Q1 Energy Conversion and Management.

4.2.2. Antecedentes

Con el objetivo de llevar cabo una predicción retrospectiva de las características del viento a largo plazo en lugares donde solo se dispone de medidas registradas en un corto plazo de tiempo ha sido propuesta en la literatura científica una amplia colección de los denominados métodos MCP [13], habiéndose expuesto en el Capítulo 2 de este documento las bases consideradas para este análisis.

Los métodos más frecuentemente propuestos y utilizados hasta el momento en la industria eólica se han apoyado en la información proporcionada por una sola estación de referencia. Sin embargo, en la bibliografía científica relacionada con las energías renovables puede observarse un creciente número de propuestas de métodos que se apoyan en varias estaciones de referencia para estimar las características del viento a largo plazo en un sitio objetivo [13,45]. Los defensores de estas propuestas han argumentado que el uso de varias estaciones de referencia permite capturar los detalles intrínsecos en el recurso eólico de la posición objetivo que de otra forma se perderían si solamente se utilizara como referencia una única estación. En este contexto, se han propuesto diversas técnicas como métodos MCP [16,26,45,204], aunque la técnica de minería de datos más frecuentemente empleada ha sido las redes neuronales artificiales [11,13,17,27-29,35,45,157].

Hay que señalar que generalmente las únicas variables de las estaciones de referencia que han alimentado a los modelos MCP han sido las velocidades del viento que han sido registradas en las mismas [27-29]. Sin embargo, algunos autores [17,35,157] destacan la importancia de usar las velocidades y direcciones del viento como referencias, especialmente en zonas con orografía compleja. En la inmensa mayoría de los casos, la variable que se ha pretendido estimar a largo plazo en el sitio objetivo ha sido exclusivamente la velocidad del viento. Ello se justifica, entre otras razones [11,203], por el hecho de que con las velocidades del viento estimadas a largo plazo en una determinada zona se puede calcular la velocidad media del viento a largo plazo, la cual constituye uno de los tipos de información relevante que se suele presentar en los mapas de recurso eólico [94,132,137,205].

Llegados a este punto, se manifiesta que a pesar de la búsqueda exhaustiva realizada en el marco de esta tesis doctoral, no se ha encontrado propuestas de modelos MCP que hayan sido entrenados y evaluados con el propósito principal de estimar las WPDs a largo plazo en un sitio objetivo. Sin embargo, la media anual de las densidades de potencia es una medida relevante del viento como fuente de potencia y frecuentemente es utilizada como indicador del potencial eólico existente en una región [137,205-207]. Es decir, proporciona información sobre la cantidad de energía eólica que está disponible en un potencial sitio para ser convertida por una turbina eólica.

De hecho, dada la importancia de este indicador, también se elaboran mapas de recurso eólico donde se indica la mencionada densidad de potencia media anual [94,132,137,205] en términos de clases de potencia eólica, donde cada clase representa un rango de medias anuales de densidades de potencia eólica. Por tanto, también resulta de sumo interés disponer de modelos MCP que permitan estimar las WPDs a largo plazo en un sitio objetivo con el propósito de computar a partir de ellas la media anual de WPDs. En este sentido, los métodos MCP que utilizan información meteorológica (series de datos de densidad del aire, estimados éstos a partir de series de datos de temperatura, presión y humedad, así como series de datos de velocidad y dirección del viento) registrada en varias estaciones de referencia puede ser de utilidad para estimar las WPDs en un sitio objetivo donde sólo se cuentan con datos meteorológicos recabados a corto plazo fruto de campañas de medición meteorológicas.

Si la densidad del aire ρ y la velocidad del viento v son variables dependientes que se encuentran distribuidas de acuerdo con una distribución bivariable y continua de probabilidad, $f_{\rho v}(\rho, v)$, se obtiene $E[WPD]$ de acuerdo con la Ecuación 4.1 [58].

$$E[WPD] = \frac{1}{2} E[\rho v^3] = \frac{1}{2} \int_0^{\infty} \int_0^{\infty} \rho v^3 f_{\rho v}(\rho, v) d\rho dv \quad (4.1)$$

En la Ecuación 4.1, si ρ se expresa en kg m^{-3} y v se en m s^{-1} , la $E[WPD]$, o media de la potencia por unidad de área perpendicular a la dirección desde la que el viento sopla, se obtiene en Wm^{-2} . En el caso de disponer en el periodo de consideración de series de n medidas de velocidades del viento v_i y de sus densidades del aire asociadas ρ_i , la medida de densidad de potencia de la muestra (\overline{WPD}) puede ser expresada como se indica en la Ecuación 4.2 [94,207].

$$\overline{WPD} = \frac{1}{2n} \sum_{i=1}^n \rho_i v_i^3 \quad (4.2)$$

La precisión de Ecuación 4.2 depende del intervalo de registro de datos, ya que dicha ecuación ignora las variaciones de la densidad del aire y fundamentalmente de la velocidad del viento dentro de cada intervalo de registro [94,207]. Por tanto, como señala Brower [94], la verdadera media del periodo en consideración es generalmente algo mayor que la computada con Ecuación 4.2, aunque la diferencia no suele ser generalmente importante para la evaluación del recurso eólico.

Hay que indicar que en algunas referencias [14,17,18,208,209] una vez estimadas mediante el empleo de un método MCP las velocidades del viento a largo plazo se han estimado a partir de ellas las E[WPD]. Es decir, en los mencionados artículos los autores no han considerado la variabilidad de la densidad del aire y no han entrenado y evaluado los modelos MCP con el objetivo principal de minimizar el error en la predicción de las WPDs.

Para ello, parten de la hipótesis de que ρ y v son variables aleatorias estadísticamente independientes y, por tanto, la función de probabilidad conjunta $f_{\rho v}(\rho, v)$, de la Ecuación 4.1, es igual al producto de sus probabilidades marginales $f_{\rho}(\rho)$ e $f_v(v)$ [58], Ecuación 4.3.

$$E[WPD] = \frac{1}{2} \int_0^{\infty} \rho f_{\rho}(\rho) d\rho \int_0^{\infty} v^3 f_v(v) dv = \frac{1}{2} E[\rho] E[v^3] \quad (4.3)$$

La cual puede ser estimada con la Ecuación 4.4:

$$E[\widehat{WPD}] = \frac{1}{2} \bar{\rho} \left(\frac{1}{n} \sum_{i=1}^n v_i^3 \right) = \frac{1}{2} \bar{\rho} \overline{v^3} \quad (4.4)$$

La mayoría de los autores [14,17,208,209] usan la Ecuación 4.4 asumiendo que la densidad del aire es constante en el tiempo y utilizan el valor estándar de 1.225 kg m^{-3} , el cual corresponde a condiciones de atmósfera estándar (Aire completamente seco, presión y temperatura medias a nivel del mar de 1013.25 hPa y 15°C, respectivamente). En este sentido, habría que matizar que la densidad del aire es una variable continua que depende de la presión, temperatura y humedad [58], aunque su variabilidad tiene una influencia muy inferior a la variabilidad de la velocidad del viento a la hora de estimar las WPDs. Sin embargo, en estaciones ubicadas en lugares con altitud elevada, la utilización de una densidad media del aire de 1.225 kg m^{-3} conduce a la sobrestimación de la E[WPD]. Con respecto al empleo de la densidad media del aire en Ecuación 4.4 Koepl [207] recomienda que se utilice la densidad media del aire ponderada, Ecuación 4.5.

$$\widehat{E[\rho]} = \bar{\rho} = \frac{\sum_{i=1}^n \rho_i v_i^3}{\sum_{i=1}^n v_i^3} \quad (4.5)$$

También habría que señalar que algunos autores [18] han computado las medias mensuales en sitios objetivo haciendo uso de la Ecuación 4.6, que utiliza las velocidades medias mensuales del viento estimadas, \bar{v}_m .

$$E[\widehat{WPD}] = \frac{1}{2} \bar{\rho} \left(\frac{1}{n} \sum_{i=1}^n v_i \right)^3 = \frac{1}{2} \bar{\rho} (\bar{v}_m)^3 \quad (4.6)$$

Sin embargo, como señalan Brower [94] y Jain [132], dicha estimación no es adecuada, ya que se ignora la distribución estadística de la velocidad del viento. Las velocidades del viento superiores a la velocidad media contribuyen mucho más a $E[WPD]$ que las velocidades del viento inferiores a la velocidad media, debido al exponente cúbico.

4.2.3. Objetivo del estudio desarrollado

Como ha sido indicado en la descripción de los antecedentes de la materia objeto de este trabajo, el objetivo principal (único objetivo en la inmensa mayoría de los casos), del amplio abanico de métodos MCP que han sido propuestos en la literatura científica relacionada con las energías renovables, ha sido estimar las velocidades del viento a largo plazo en un sitio objetivo donde sólo se disponen de datos recabados a corto plazo [203].

Asimismo, ha sido subrayado que de acuerdo con la bibliografía consultada se puede concluir que, en aquellos casos [14,17,58,208,209] en que se han estimado las WPDs mediante métodos MCP se ha realizado de forma marginal. Es decir, el objetivo principal de los métodos MCP fue estimar con precisión las velocidades del viento y una vez estimadas estas se utilizaron para estimar las WPDs. Por tanto, no se ha analizado si dichos modelos, que han sido entrenados minimizando el error en la predicción de las velocidades del viento, también minimizan el error en la predicción de las WPDs, o por el contrario, se requiere utilizar modelos MCP especialmente entrenados para minimizar el error en la predicción de las WPDs. Además, en dichos trabajos se ha asumido que la densidad del aire es constante en el tiempo y se ha utilizado el valor típico de 1.225 kg m^{-3} , correspondiente a condiciones de atmósfera estándar.

Ante este escenario, uno de los objetivos del estudio desarrollado en este apartado, que supone una aportación original y que ha permitido la primera publicación en el marco de la presente tesis doctoral (véase Anexo), es que se proponen por primera vez diferentes modelos MCP, que, utilizando múltiples estaciones de referencia, tengan como propósito principal estimar las WPDs a largo plazo en un sitio objetivo. Y ello, considerando no solo la variabilidad de las velocidades del viento en las estaciones de referencia, sino también de las direcciones del viento y de las densidades del aire (proponiendo una formulación válida para una humedad relativa del aire entre 0% y 100% y amplios rangos de presión barométrica y de temperatura del aire), así como la forma funcional en que dichas variables participan en los modelos MCP propuestos. Se persigue analizar si existe diferencia estadísticamente significativa (5% de significación) entre los errores de estimación de las WPDs producidos por los modelos MCP aquí propuestos y los obtenidos por los modelos utilizados hasta ahora (en los que la estimación de las WPDs no se realizó directamente). Con el ensayo de hipótesis

estadística aplicado se pretende fundamentar estadísticamente las conclusiones alcanzadas al comparar las métricas MAE (Mean Absolute Error), MARE (Mean Absolute Relative Error) y R^2 (Coefficient of determination) obtenidas con los modelos MCP analizados con el objetivo de evitar tratar diferencias puramente circunstanciales originadas por la aleatoriedad de las muestras, como si fuesen debidas a mecanismos estructurales dignos de mención.

En base a la investigación realizada utilizando datos meteorológicos medios horarios (temperatura, presión y humedad relativa del aire y velocidad y dirección del viento) de diez estaciones meteorológicas (WSs) instaladas en el archipiélago canario se trata de aportar a los usuarios y a los diseñadores de los métodos MCP, de información relevante a la hora de estimar las densidades de potencia eólica en un sitio objetivo.

Como ha sido señalado en el apartado anterior, cuando se ha hecho uso de información recopilada simultáneamente en varias estaciones de referencia, las técnicas de minería de datos más frecuentemente utilizadas como método MCP han sido las ANNs [11,13,17,27-29,35,45,157]. Sin embargo, en este trabajo se han utilizado las técnicas de máquinas de vector soporte, más concretamente las máquinas de vector soporte para regresión (SVR), como métodos MCP en los que se seleccionan las variables de entrada mediante una técnica Wrapper y los parámetros característicos de dichos modelos se estiman usando un método heurístico y Grid Search con validación-cruzada de 10-Folds.

La elección de la SVR está motivada por ser una de las técnicas que mejor representan el estado del arte de Machine Learning (ML) debido a su comprobada capacidad de predicción de primera clase en muy diferentes escenarios, mostrándose con frecuencia superior a la ANNs Multi-Layer Perceptron (MLP) [171,210-212]. Esta capacidad predictiva está fundamentada no sólo en su propiedad de aproximación universal a cualquier función continua (presente también en las ANNs) sino también en un algoritmo de entrenamiento más eficaz y estable que proporciona siempre una solución única al problema de estimación y una mayor “Sparsity” en dicha solución.

En este contexto, siete de los ocho modelos MCP propuestos utilizaran las SVR como método de predicción. Sin embargo, el octavo modelo MCP propuesto para la estimación de la WPDs de un sitio objetivo empleará la técnica de Multiple Linear Regression (MLR) [146] con objeto de servir de referencia comparativa del rendimiento del resto de modelos analizados.

Este apartado se estructura de la siguiente forma: En el siguiente subapartado se describen los materiales utilizados, incluyendo la muestra de datos, el método de estimación de la densidad del aire empleado y los modelos SVR y MLR usados. A continuación, se detalla la metodología seguida para realizar el trabajo de comparación y se presentan y analizan los resultados obtenidos. Por último, se presentan las conclusiones más importantes del trabajo realizado.

4.2.4. Muestras de datos usadas para el desarrollo del estudio

Como ha sido detallado en el Apartado 4.1, los datos meteorológicos utilizados en este artículo proceden de 10 estaciones climatológicas (WS) ubicadas en las siete islas mayores que constituyen el

Archipiélago Canario (España), Figura 23. Se dispone de medidas medias horarias de velocidad (captadas con anemómetros rotativos de cazoletas) y dirección del viento (captados mediante veletas), así como de temperatura, presión atmosférica y humedad relativa del aire, registrada durante el año 2014. La longitud de dichas series de datos permite que los modelos, en su aprendizaje, puedan obtener conocimiento del patrón de las variaciones estacionales de los datos meteorológicos de las estaciones objetivo y de referencia representativos del corto plazo (periodo en el que las series temporales de datos de las estaciones de referencia coinciden en longitud y fecha con la serie temporal de datos registrados en el sitio considerado como objetivo), tal como se recomienda en las hipótesis en las que se sustentan los métodos MCP [13]. Todas las series de datos han sido captadas a 10 metros sobre el nivel del suelo, salvo en la estación WS-4 donde los sensores han sido ubicados a 20 metros de altura.

En la Tabla 10 se han mostrado los códigos asignados a cada estación meteorológica, sus altitudes o elevaciones con respecto al nivel del mar y las medias anuales, las desviaciones estándar, los valores máximos y mínimos de las series registradas durante el año 2014.

En la Figura 24 - 25 también se representan las rosas de los vientos correspondientes al año 2014 en las diez WSs consideradas en este estudio.

4.2.5. Descripción de las técnicas y modelos matemáticos usados para la simulación

4.2.5.1 Cálculo de la densidad del aire

En este trabajo las densidades del aire han sido estimadas con los datos de temperatura, humedad relativa y presión atmosférica capturados en cada una de las 10 estaciones que servirán para el desarrollo de este estudio.

La formulación empleada para el cálculo de la densidad del aire en este estudio ha sido presentada en el Apartado 2.7.

4.2.5.2 Técnicas utilizadas para la estimación de las densidades de potencia

Tal como se ha indicado en el Capítulo 3, en este trabajo se utiliza un enfoque de regresión para construir los distintos modelos MCP en la estimación de las WPDs en un sitio objetivo. Para estimar la función de regresión en cada uno de los modelos comparados, se utiliza la técnica SVR, con la excepción de un modelo lineal empleado como referencia comparativa, que se estima mediante Least Squares MLR.

En el Apartado 3.3 de la presente tesis doctoral podrá encontrar una descripción de la técnica Least Squares Multiple Linear Regression, la cual ha sido utilizada en la presente tesis doctoral a modo de referencia en la comparativa de modelos desarrollada en este apartado. Podrá encontrar una descripción con mayor detalle en [153].

Por otra parte, como para la técnica anterior, en el Apartado 3.5 de la presente tesis doctoral se ha realizado un análisis de la técnica SVR. Para la estimación de los modelos SVR utilizados en este trabajo se ha utilizado el software multiplataforma de licencia libre R que posee una amplia librería de paquetes producto de la contribución desinteresada de multitud de equipos de investigación estadística de primera línea a nivel mundial. Como base de partida en la implementación de la SVR, se utilizó el paquete Kernlab [164]. Una descripción con mayor detalle puede ser encontrada en [150,211,213].

4.2.5.3 Modelos evaluados para la estimación de densidades de potencia con MCP

Se sintetizan en la Tabla 18 las formulaciones de los 8 modelos para la estimación, mediante métodos MCP, de las densidades de potencia eólica a largo plazo en un sitio objetivo, los cuales se evalúan y comparan en el presente trabajo. En la Tabla 18 el subíndice t indica el momento t o instante evaluado, el símbolo “hat” indica estimación, y las variables con subíndice $j \in \{1, \dots, d\}$ se refieren a la estación de referencia j -ésima y las que no tienen dicho subíndice se refieren al sitio objetivo, Asimismo, $f_A(B_1, \dots, B_p)$ representa los métodos de regresión utilizados donde se obtiene una estimación \hat{A} de la variable A con los features B_1, \dots, B_p . Los modelos del 1 al 7 estiman las variables del sitio objetivo a partir de la técnica SVR. Adicionalmente, en el modelo 8 se formula una alternativa construida a través de la técnica de Multiple Linear Regression (MLR) que supondría la estrategia de menor complejidad estructural que permitiría la resolución del problema formulado. Hay que señalar que únicamente los modelos 4, 6, 7 y 8 son los únicos modelos que utilizan la variable WPD o su logaritmo base 10 como variable objetivo, ya que los restantes modelos se entrenan tratando de minimizar el error en la predicción de otras variables de interés, las cuales, al estar relacionadas con las WPDs se utilizan posteriormente para estimar a estas últimas. Sin embargo, hay que matizar que únicamente los modelos 4 y 7 estiman de forma directa las WPDs ya que los modelos 6 y 8 se entrenan tratando de minimizar el error en la predicción de los logaritmos base 10 de las WPDs y para estimar las WPDs se precisa calcular el inverso de dichos logaritmos.

El modelo 1 (M1) recoge el procedimiento más frecuentemente utilizado para estimar de forma marginal las WPDs. Estos modelos se entrenan mediante una técnica SVR, que utilizando como variables de entrada las velocidades y direcciones del viento de los sitios de referencia estiman las velocidades del viento del sitio objetivo. Posteriormente, estas se elevan al cubo y se multiplican por $1/2$ y por el valor estándar de la densidad del aire ($\rho_0 = 1.225 \text{ kg m}^{-3}$), para estimar las WPDs.

Models	Formulation
Modelo 1	$\widehat{WPD}_t = \frac{1}{2} \cdot \rho_0 \cdot \widehat{V}_t^3 = \frac{1}{2} \cdot 1.225 \cdot [f_V(V_{1t}, \dots, V_{dt}, D_{1t}, \dots, D_{dt})]^3$
Modelo 2	$\widehat{WPD}_t = \frac{1}{2} \cdot \bar{\rho} \cdot \widehat{V}_t^3 = \frac{1}{2} \cdot \bar{\rho} \cdot [f_V(V_{1t}, \dots, V_{dt}, D_{1t}, \dots, D_{dt})]^3$
Modelo 3	$\widehat{WPD}_t = \frac{1}{2} \cdot \bar{\rho} \cdot \widehat{V}_t^3 = \frac{1}{2} \cdot \bar{\rho} \cdot f_{V^3}(V_{1t}^3, \dots, V_{dt}^3, D_{1t}, \dots, D_{dt})$
Modelo 4	$\widehat{WPD}_t = f_P(WPD_{1t}, \dots, WPD_{dt}, D_{1t}, \dots, D_{dt})$
Modelo 5	$\widehat{WPD}_t = \frac{1}{2} \cdot f_\rho(\rho_{1t}, \dots, \rho_{dt}, D_{1t}, \dots, D_{dt}) \cdot [f_V(V_{1t}, \dots, V_{dt}, D_{1t}, \dots, D_{dt})]^3$
Modelo 6	$\widehat{WPD}_t = 10^{\log \widehat{WPD}_t}$, with: $\log \widehat{WPD}_t = f_{\log WPD}(\log WPD_{1t}, \dots, \log WPD_{dt}, D_{1t}, \dots, D_{dt})$
Modelo 7	$\widehat{WPD}_t = f_{WPD}(V_{1t}, \dots, V_{dt}, D_{1t}, \dots, D_{dt}, \rho_{1t}, \dots, \rho_{dt})$
Modelo 8	$\widehat{WPD}_t = 10^{\log \widehat{WPD}_t}$, with: $\log \widehat{WPD}_t = \sum_{j=1}^d (a_j \cdot \log \rho_{jt} + b_j \cdot \log V_{jt} + c_j \cdot D_{jt}) + b$

Tabla 18 Modelos considerados para la estimación de las densidades de potencia

El modelo 2 (M2) únicamente difiere del M1 en que se sustituye la densidad del aire ρ_0 por la densidad media del aire ($\bar{\rho}$) calculada en el sitio objetivo durante el periodo de disposición de series de datos de temperatura, presión y humedad relativa del aire.

El modelo 3 (M3) difiere del M2 en que en lugar de estimar las velocidades del viento estima los cubos de las velocidades. En el caso de este modelo las SVR se alimentan de los valores de las velocidades al cubo y de las direcciones del viento de los sitios de referencia. Una vez estimados los cubos de la velocidades en el sitio objetivo se multiplican por 1/2 y por la densidad media del aire ($\bar{\rho}$) calculada en el sitio objetivo para estimar las densidades de potencia eólica.

El modelo 4 (M4) estima directamente las densidades de potencia eólica del sitio objetivo. En este caso los modelos SVR se nutren de las densidades de potencia eólica y de las direcciones del viento de las estaciones de referencia.

El Modelo 5 (M5) emplea otros dos submodelos para estimar de forma desagregada las dos componentes de las densidades de potencia eólica en el sitio objetivo: la velocidad del viento y la densidad del aire en dicho sitio. Uno de ellos estima las velocidades del viento por medio de la técnica SVR, utilizando como variables de entrada las velocidades y las direcciones del viento de las estaciones de referencia. El otro estima las densidades del aire haciendo uso de la técnica SVR y empleando como variables de entrada las densidades del aire y las direcciones del viento de las estaciones de referencia. La estimación indirecta de las densidades de potencia eólica se lleva a cabo elevando al cubo las velocidades del viento estimadas y multiplicándolas por las correspondientes densidades del aire estimadas y por 1/2.

El modelo 6 (M6) difiere del M4 en un cambio de escala que se realiza en las variables densidad de potencia eólica de los sitios de referencia y objetivo. Es decir, los modelos SVR, utilizando como variables de entrada las direcciones del viento y los logaritmos base 10 de las densidades de potencia eólica de las de las estaciones de referencia estiman directamente los logaritmos base 10 de las densidades de potencia eólica del sitio objetivo. Las densidades de potencia eólica se estiman posteriormente calculando el inverso del logaritmo base 10 de las variables estimadas.

El modelo 7 (M7) al igual que el M4 estima directamente las densidades de potencia eólica del sitio objetivo. Sin embargo, difiere de aquel en las variables que alimentan a los modelos SVR. En este caso, los modelos SVR utilizan como variables de entrada las velocidades y direcciones las direcciones del viento y las densidades del aire de las estaciones de referencia.

El modelo 8 (M8), de forma similar al modelo 6, estima directamente los logaritmos base diez de las WPDs y, por tanto, para estimar las WPDs se precisa calcular el inverso de dichos logaritmos. Sin embargo, M8 utiliza la técnica de MLR en lugar de la SVM y las variables predictoras son los logaritmos base diez de las velocidades y direcciones del viento y de las densidades del aire de las estaciones de referencia. Este modelo lineal se incluye como referencia en la comparación, pero también está inspirado en el carácter lineal del logaritmo de la WPD (Ecuación 4.7), por lo que, debería producir buenos resultados.

$$WPD = \frac{1}{2} \cdot \rho \cdot V^3 \Leftrightarrow \log(WPD) = \log\left(\frac{1}{2}\right) + \log(\rho) + 3 \cdot \log(V) \quad (4.7)$$

4.2.6. Metodología

4.2.6.1 Preámbulo

En la Figura 27 se muestra un esquema de la metodología que, siguiendo las bases de los procedimientos MCP, ha sido utilizada en este trabajo para llevar a cabo la construcción y comparación de los diferentes modelos de estimación de WPDs propuestos.

En la parte superior izquierda de la Figura 27 se muestran las variables (features) disponibles en los sitios de referencia (WS-1, WS-2,..., WS-10), cuyas series de valores registrados en el periodo de corto plazo pueden ser usados como datos de entrada en los ocho diferentes modelos analizados. Es decir, las series de datos de velocidad del viento (V), de dirección del viento (D), de densidad del aire (ρ) y de densidad de potencia eólica (WPD). Cada valor medio horario de las series de densidad del aire (ρ_i) se ha determinado haciendo uso de la Ecuación 2.19 y de las series de datos de temperatura, presión atmosférica y humedad relativa del aire mencionadas en el Apartado 4.2.4, siguiendo el procedimiento señalado en el Apartado 2.7. Cada valor medio horario de las series de densidad de potencia eólica (WPD_i) se ha determinado, con la Ecuación 4.8, haciendo uso de los valores medios horarios de las series de densidad del aire (ρ_i) y de los valores medios horarios de las series de velocidades del viento (v_i).

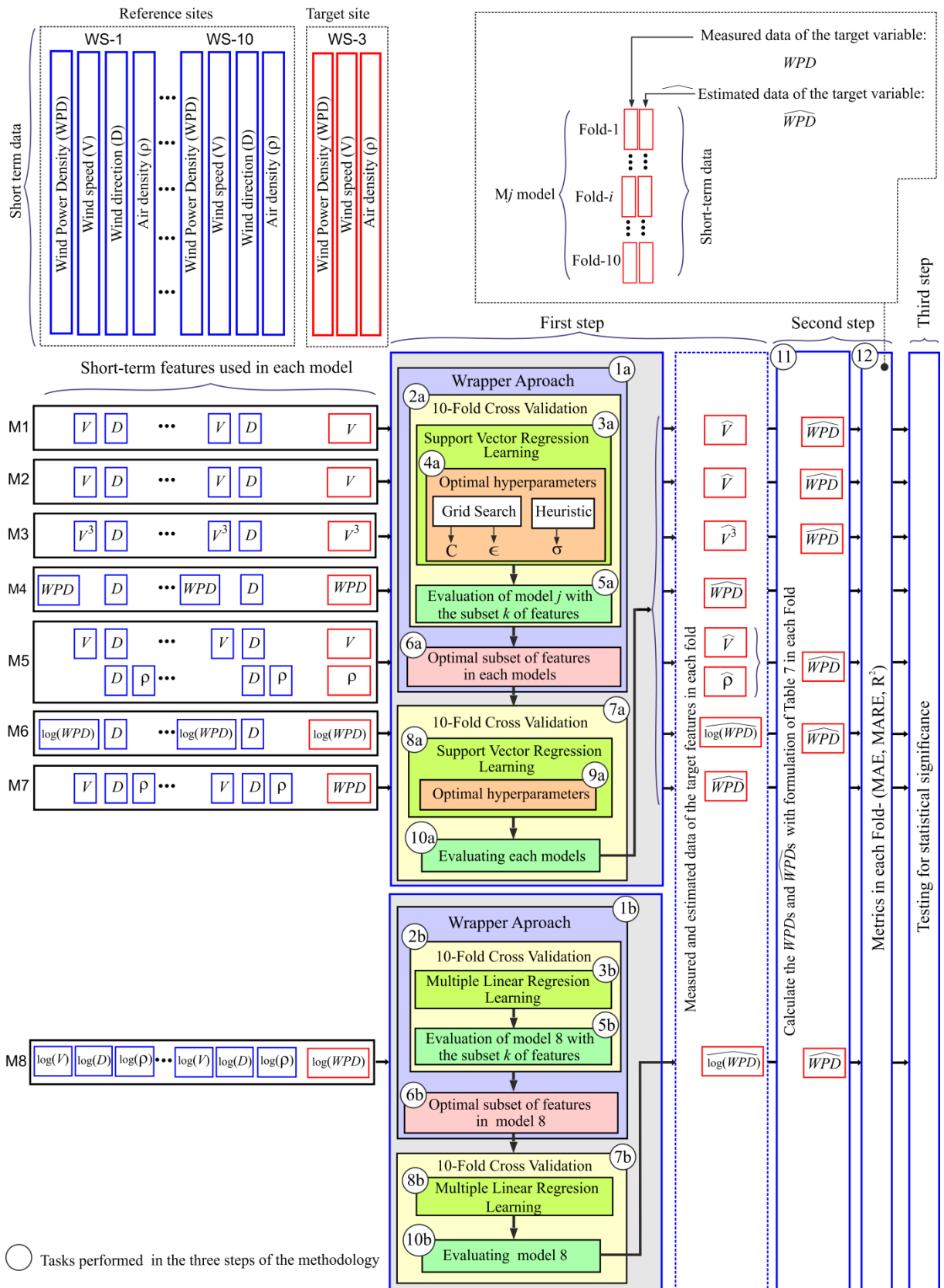


Figura 27 Esquema general de la metodología empleada para el desarrollo del estudio

$$WPD_i = \frac{1}{2} \rho_i v_i^3 \quad (4.8)$$

Asimismo, se muestran en dicha zona de la Figura 27 las variables del sitio objetivo que pueden ser utilizadas en el aprendizaje supervisado de los modelos analizados. Es decir, la velocidad del viento (V), la densidad del aire (ρ) y la densidad de potencia eólica (WPD) del sitio objetivo.

Con el propósito de facilitar la interpretación del mencionado esquema (Figura 27), este se ha particularizado para el caso en que el sitio objetivo sea WS-3 y las series de datos meteorológicos registrados a corto plazo en las estaciones de referencia disponibles sean los nueve restantes indicadas en la Tabla 10. Sin embargo, en el trabajo realizado se han analizado también los casos en los que las estaciones WS-4 y WS-5 han ejercido como sitio objetivo y las restantes nueve WSs señaladas en la Tabla 10 se han utilizado como estaciones de referencia disponibles. El motivo por el que se han seleccionado únicamente las estaciones WS-3, WS-4 y WS-5 como sitios objetivos se ha fundamentado en una de las hipótesis básicas de los métodos MCP [13], que señala la necesidad de que los datos meteorológicos de las estaciones de referencia sean representativos del clima del lugar objetivo. Esencialmente, ello suele concretarse en la literatura científica relacionada con las energías renovables en la existencia de un alto coeficiente de correlación de Pearson entre las velocidades de viento de la estación objetivo y de referencia.

En este contexto, las tres mencionadas WSs son las únicas que presentan coeficientes de correlación entre los datos de viento de la estación objetivo y de al menos una estación de referencia superiores a 0.8 (Tabla 20), los cuales son considerados en la mencionada literatura como "buenos" (0.8-0.9) [13]. Los coeficientes de correlación registrados entre los datos de vientos de las tres WSs antes mencionadas cumplen con la recomendación señalada en [135], donde se indica que dichos coeficientes no deberían ser menores del 70%.

Cuando los métodos MCP se han apoyado en la información proporcionada por una única estación de referencia esta ha sido desechada [13] cuando el coeficiente de correlación ha sido bajo, ya que se considera que los datos de referencia presentan una sustancial incertidumbre al tratar de estimar las condiciones eólicas a largo plazo en el sitio objetivo. Sin embargo, en los modelos MCP que permiten utilizar información registrada en varias estaciones de referencia, no debe descartarse a priori el empleo de ninguna de las estaciones de referencia disponibles, aunque los coeficientes de correlación no sean altos, si se dispone de al menos de una estación de referencia con un buen coeficiente de correlación [17]. En este último caso se suelen utilizar métodos Feature Selection con los cuales se resuelven los posibles problemas derivados del sobreajuste [45].

WS	WS-1	WS-2	WS-3	WS-4	WS-5	WS-6	WS-7	WS-8	WS-9	WS-10
WS-1	1.00	0.73	0.74	0.67	0.69	0.54	0.26	0.56	0.52	0.56
WS-2	0.73	1.00	0.68	0.58	0.63	0.55	0.30	0.54	0.58	0.57
WS-3	0.74	0.68	1.00	0.82	0.83	0.53	0.23	0.60	0.51	0.58
WS-4	0.67	0.58	0.82	1.00	0.83	0.51	0.18	0.55	0.45	0.50
WS-5	0.69	0.63	0.83	0.83	1.00	0.63	0.21	0.55	0.42	0.50
WS-6	0.54	0.55	0.53	0.51	0.63	1.00	0.39	0.36	0.31	0.40
WS-7	0.26	0.30	0.23	0.18	0.21	0.39	1.00	0.23	0.24	0.34
WS-8	0.56	0.54	0.60	0.55	0.55	0.36	0.23	1.00	0.56	0.55
WS-9	0.52	0.58	0.51	0.45	0.42	0.31	0.24	0.56	1.00	0.56
WS-10	0.56	0.57	0.58	0.50	0.50	0.40	0.34	0.55	0.56	1.00

Tabla 19 Coeficientes de correlación lineal entre velocidades del viento de las estaciones anemométricas evaluadas

Hay que señalar que los coeficientes de correlación registrados entre las WPDs (Tabla 20) de las tres estaciones seleccionadas son algo inferiores a los registrados para las velocidades (si se utilizasen las reglas generales recomendadas para las velocidades en los métodos MCP [13], los coeficientes de correlación entre densidades de potencia eólica de las estaciones WS-3 y WS-5 podrían ser catalogados como "buenos" (0.8 a 0.9), sin embargo, los registrados entre las estaciones WS-3 y WS-4 y entre las estaciones WS-4 y WS-5 se catalogarían como "moderados" (0.7-0.8).

Tal como se ha indicado en el Apartado 6.4 las series de datos meteorológicos registrados durante el año 2014 se han considerado representativos del corto plazo. Es decir, serán los datos que se utilizarán para llevar a cabo la construcción de los modelos cuya comparación es objetivo de este trabajo.

WS	WS-1	WS-2	WS-3	WS-4	WS-5	WS-6	WS-7	WS-8	WS-9	WS-10
WS-1	1.00	0.60	0.69	0.61	0.62	0.48	0.16	0.49	0.40	0.37
WS-2	0.60	1.00	0.56	0.44	0.49	0.51	0.36	0.43	0.53	0.43
WS-3	0.69	0.56	1.00	0.77	0.82	0.52	0.18	0.51	0.33	0.38
WS-4	0.61	0.44	0.77	1.00	0.78	0.45	0.12	0.47	0.28	0.27
WS-5	0.62	0.49	0.82	0.78	1.00	0.58	0.10	0.49	0.26	0.27
WS-6	0.48	0.51	0.52	0.45	0.58	1.00	0.37	0.37	0.29	0.23
WS-7	0.16	0.36	0.18	0.12	0.10	0.37	1.00	0.17	0.30	0.28
WS-8	0.49	0.43	0.51	0.47	0.49	0.37	0.17	1.00	0.41	0.39
WS-9	0.40	0.53	0.33	0.28	0.26	0.29	0.30	0.41	1.00	0.44
WS-10	0.37	0.43	0.38	0.27	0.27	0.23	0.28	0.39	0.44	1.00

Tabla 20 Coeficientes de correlación lineal entre densidades de potencia de las estaciones anemométricas evaluadas

4.2.6.2 Procedimiento

Como puede observarse en Figura 27, el procedimiento seguido consta de tres pasos. En el primer paso del procedimiento se llevan a cabo diez tareas, las cuales se señalizan con números correlativos encerrados en un círculo. En las seis primeras tareas se lleva a cabo la selección del subconjunto óptimo de variables de entrada de los ocho modelos para cada una de las tres WSs objetivo (WS-3, WS-4 y WS-5) consideradas, se eligen los ocho modelos generados para cada una de las mencionadas estaciones objetivo que en el proceso de evaluación (tarea 10) han proporcionado menores errores de muestra (utilizando los datos que se poseen de la evidencia) y se obtienen las predicciones producidas por dichos modelos (señalado en un recuadro con línea de puntos y rayas en Figura 27).

El segundo paso consta de las tareas 11 y 12. En la tarea 11, utilizando los datos estimados se lleva a cabo el cálculo de las predichas \overline{WPDs} , con la formulación indicada en la Tabla 18 para aquellos modelos que no han sido entrenados para estimar directamente las densidades de potencia eólica del sitio objetivo (M1, M2, M3, M5, M6 y M8). En la tarea 12 se calculan los errores (métricas MAE, MARE y R^2) cometidos al estimar las WPDs con todos los modelos analizados.

En el Tercer y último paso del procedimiento seguido se lleva a cabo un ensayo de hipótesis estadística. Con el mismo se persigue analizar si existe diferencia estadísticamente significativa (5% de significación) entre los errores de estimación de las WPDs producidos por los ocho modelos MCP analizados. En los sub-aparatos siguientes se detallan los tres mencionados pasos.

4.2.6.2.1 Primer paso

En la parte inferior izquierda de la Figura 27 se muestran las variables de las estaciones de referencia disponibles que pueden alimentar a cada uno de los ocho modelos propuestos. Dado que el número de estaciones de referencia disponibles se eleva en el caso de estudio a nueve y que el número de variables que cada estación puede aportar a cada modelo oscila entre dos y tres, el número total de predictores disponibles para ser utilizados en los modelos es de entre 18 y 27. En este contexto, con el propósito de minimizar el riesgo de sobreajuste de los modelos MCP, la metodología propuesta utiliza una técnica wrapper (tareas "1a" y "1b") para la selección de features de cada modelo. Concretamente, se utiliza el algoritmo Recursive Feature Elimination (RFE) [184] implementado en el paquete Caret [185] del software R Statistics.

Esta técnica, conocida como "Backward Elimination" determina, basándose en la medida de la raíz del error cuadrático medio (root mean-square error) (RMSE) Ecuación 3.18, el subconjunto de variables de entrada óptimas de cada modelo tras analizar distintas combinaciones de las features predictoras. Cuando la mejor combinación de variables es determinada, se eliminan todos los predictores que producen un error mayor al obtenido con el subconjunto óptimo [184,186].

En este caso, en la Ecuación 3.18 o_i son los datos observados, \hat{e}_i los datos estimados con los modelos SVR y MLR y n el número de observaciones del conjunto de datos evaluados.

El algoritmo que se ha implementado en la técnica wrapper utilizada para llevar a cabo la selección de variables óptimas ha sido el mismo que el utilizado en los modelos MCP para llevar a cabo la predicción de las variables objetivo. De esta forma, el algoritmo puede aprovechar al máximo las variables seleccionadas [45]. Por tanto, en este trabajo se ha utilizado la técnica MLR como algoritmo integrado en el wrapper y en el modelo MCP, en el caso del M8 (señalado con las siglas "3b" y "8b" encerradas en círculos en Figura 27) y la técnica SVR como algoritmo integrado en el wrapper y en el modelo MCP en los restantes modelos (señalado con las siglas "3a" y "8a" encerradas en círculos en la Figura 27).

Con el objetivo de reducir la dependencia de los errores obtenidos al evaluar los modelos del modo en el que se lleva a cabo la partición entre los datos de entrenamiento utilizados para el aprendizaje y los datos de ensayo usados para evaluar los modelos se ha utilizado la técnica estadística de validación cruzada con 10 folds (tareas "2a" y "2b"). Dicha técnica ha sido utilizada con anterioridad en la evaluación de métodos MCP [16,45,206] y ha sido ampliamente aceptada en la comunidad de la minería de datos [125]. Para aplicar la validación cruzada los datos se ordenan de forma aleatoria y posteriormente se dividen en 10 subconjuntos disjuntos de similar tamaño [45]. Uno de los folds se utiliza como datos de ensayo y el resto (9) como datos de entrenamiento. El procedimiento de validación cruzada se repite durante 10 iteraciones, tomando cada vez un subconjunto diferente como datos de ensayo. En cada una de las 10 iteraciones se determina el error cometido. El error de la validación cruzada se obtiene como la media aritmética de los 10 errores. Otra ventaja de la validación cruzada es que la varianza de los 10 errores de muestra permite estimar la variabilidad del método de aprendizaje con respecto a la evidencia. En este punto, debemos señalar que en los ocho modelos MCP ensayados se emplearon las mismas particiones aleatorias de validación cruzada. Por tanto, todos los modelos compartieron la misma unidad experimental, minimizándose así la varianza de la diferencia entre las métricas medias obtenidas por los ocho modelos.

La ϵ -SVR con Gaussian kernel constituye la técnica de Machine Learning utilizada para la estimación de las WPDs en el presente estudio para los modelos 1 al 7. En este sentido, deben establecerse los hiperparámetros característicos C , ϵ y σ en coherencia con la situación simulada en cada modelo de aprendizaje ejecutado (tareas "4a" y "9a"). Con respecto a los parámetros característicos C , ϵ se utiliza el método Grid Search (Figura 27), el cual testea distintas combinaciones de estos parámetros hasta dar con la alternativa que menores errores produce en los datos dejados fuera en cada iteración de validación cruzada, habiéndose introducido al inicio de la rutina dos vectores con distintos valores de C , ϵ , para los cuales se han seguido las recomendaciones establecidas en diversas publicaciones de la comunidad científica [146,154,181,182]. Sin embargo, el parámetro σ ha sido seleccionado utilizando un método heurístico implementado en la función `Sigest` () del paquete `kernlab` [164] (Figura 27). La función `Sigest`, partiendo de los datos de la muestra de entrenamiento, estima el rango de valores del parámetro σ que podría proporcionar buenos resultados cuando se utiliza con SVR. Dicha función `Sigest` () estima un valor comprendido entre los cuantiles 0.1 y 0.9 de la distancia euclidiana cuadrática de los datos de partida en el espacio de trabajo $\|x - x_i\|^2$. Según las referencias [164], cualquier valor comprendido entre estos dos límites producirá buenos resultados.

Al finalizar la última tarea del primer paso (tarea "10a" y "10b"), se obtiene con el subconjunto óptimo de variables de cada modelo (tareas "6a" y "6b") y los hiperparámetros seleccionados en cada uno de ellos (tarea "9a") los valores predichos (\widehat{V} , \widehat{V}^3 , \widehat{WPD} , $\widehat{\rho}$ y $\log(\widehat{WPD})$) por cada modelo (M1, M2, M3, M4, M6, M7 y M8) y submodelo (dos submodelos pertenecientes al modelo M5) para cada posición objetivo (WS-3, WS-4 y WS-5), los cuales son utilizados en el segundo paso de la metodología.

4.2.6.2.2 Segundo paso

El segundo paso del procedimiento seguido consta de las tareas 11 y 12 anteriormente indicadas (Figura 27). En la tarea 11 las predicciones realizadas en el Primer Paso y cuyo objetivo no era estimar directamente las \widehat{WPD} (M1, M2, M3, M5, M6 y M8), utilizan las expresiones expuestas en la Tabla 5 para estimar las \widehat{WPD} en los grupos de validación cruzada (10-folds) dejados sucesivamente fuera, mientras que para los modelos en los que se ha estimado directamente las \widehat{WPD} (M4 y M7) se pasa directamente a la tarea 12.

Las métricas utilizadas en la tarea 12 para calcular los errores cometidos al estimar las WPDs con todos los modelos analizados han sido el Mean Absolute Error (MAE) Ecuación 3.19, el Mean Absolute Relative Error (MARE) Ecuación 3.20 y el coeficiente de determinación R^2 Ecuación 3.21. Dichas métricas se aplican sobre los datos estimados que han sido obtenidos en la tarea 11 (Figura 27) y los valores medidos de las variable WPD en los target site (WS-3, WS-4 y WS-5) que han sido asignados a los diez subconjuntos definidos en el proceso. Por tanto, se obtienen diez errores para cada una de las métricas, pudiéndose calcular el error medio de los mismos y la desviación estándar.

4.2.6.2.3 Tercer paso

Con alguna excepción [45], los estudios comparativos de modelos MCP que han sido llevados a cabo hasta el momento, de acuerdo con la información recopilada, han recurrido a la evaluación de hipótesis basada en precisión [11,28,33-36]. Es decir, se considera como mejor modelo aquel que proporciona el menor error. Sin embargo, en este tercer paso del procedimiento metodológico seguido, tal como ha sido mencionado anteriormente, se pretende analizar si existe diferencia estadísticamente significativa (5% de significación) entre los errores de estimación de las WPDs producidos por los diferentes modelos. Así pues, aunque se siga tomando como mejor modelo aquel que proporciona menor error, dicha precisión se considera una variable aleatoria y, por tanto, se le aplica un test de hipótesis a sus medias poblacionales a la hora de decretar la mejor fiabilidad de los modelos comparados. Para ello se lleva a cabo un ensayo de hipótesis estadística. Es decir, se pretende fundamentar estadísticamente las conclusiones alcanzadas al comparar las métricas MAE, MARE y R^2 obtenidas con los modelos MCP analizados con el objetivo de evitar tratar diferencias puramente circunstanciales originadas por la aleatoriedad de las muestras, como si fuesen debidas a mecanismos estructurales dignos de mención.

Con este procedimiento se confronta una hipótesis nula (H_0), en la que se considera que no existe diferencia significativa entre los errores medios generados por dos modelos "i" y "j" de estimación de

las WPD, con una hipótesis alternativa (H_1) unilateral, en la que se acepta que la métrica media (μ_i) del modelo "i" es significativamente mayor que la métrica media (μ_j) de otro modelo "j", Ecuación 4.9.

$$H_0: \mu_i \leq \mu_j ; H_1: \mu_i > \mu_j \quad (4.9)$$

Dado que la hipótesis de normalidad es rechazada por los datos (de acuerdo con el test de normalidad de Anderson–Darling [214] usado con un nivel de significancia estadística de $\alpha=0.05$) y que el número de datos disponibles es pequeño (los diez errores de los diez Folds) se ha utilizado un test no paramétrico de permutación para datos apareados [215,216]. Se ha utilizado un test de datos apareados dado que, en los ocho modelos MCP ensayados los errores se han calculado bajo similares condiciones de ensayo. Es decir, los modelos que se comparan comparten muestras de algunos features, las muestras de las variables objetivo son idénticas y se han realizado las mismas particiones aleatorias de validación cruzada. Por tanto, todos los modelos compartieron la misma unidad experimental, minimizándose así la varianza de la diferencia entre las métricas medias obtenidas por los ocho modelos, pudiéndose obtener así, en general, un mejor contraste.

El test no paramétrico de permutación se puede usar para muestras de tamaño pequeño y bajo la hipótesis nula no precisa realizar supuestos distribucionales con respecto a la distribución muestral de la estadística de ensayo.

Para llevar a cabo las comparaciones múltiples para poder explorar las diferencias estadísticamente discernibles entre los errores medios de los modelos, cada par de modelos son ordenados en función de la magnitud de la métrica a comparar, siendo el modelo i el de mayor valor y el modelo j el de menor. Bajo estas circunstancias, con la métrica R^2 , si la hipótesis nula es desestimada ($p\text{-value} \leq \alpha$), el modelo i se consideraría mejor que el modelo j . Asimismo, con las métricas MAE y MARE, si se rechaza la hipótesis nula, el modelo j sería considerado mejor que el modelo i . Además, antes de tomar cualquier decisión, los p – valores obtenidos en cada comparación del test de permutación han sido ajustados a través del procedimiento propuesto por Benjamini y Hochberg (BH) [217], reduciéndose el riesgo de que se produzcan falsos positivos (False Discovery Rate).

4.2.7. Análisis de resultados

Las Figura 28-29 sintetizan los resultados generales obtenidos en la etapa 12 del tercer paso del procedimiento metodológico seguido. En dichas figuras se muestran los valores medios y las desviaciones estándar de las métricas MAE, MARE y R^2 obtenidos al estimar las WPDs de cada una de las tres estaciones climatológicas consideradas como targets (WS-3, WS-4 y WS-5).

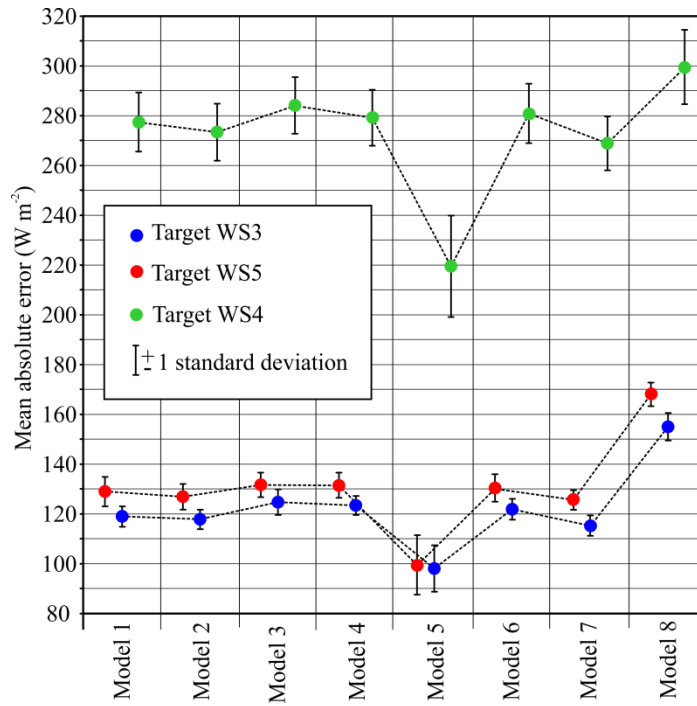


Figura 28 Errores MAE obtenidos cuando se estiman las densidades de potencia eólica en cada modelo y estación objetivo seleccionada

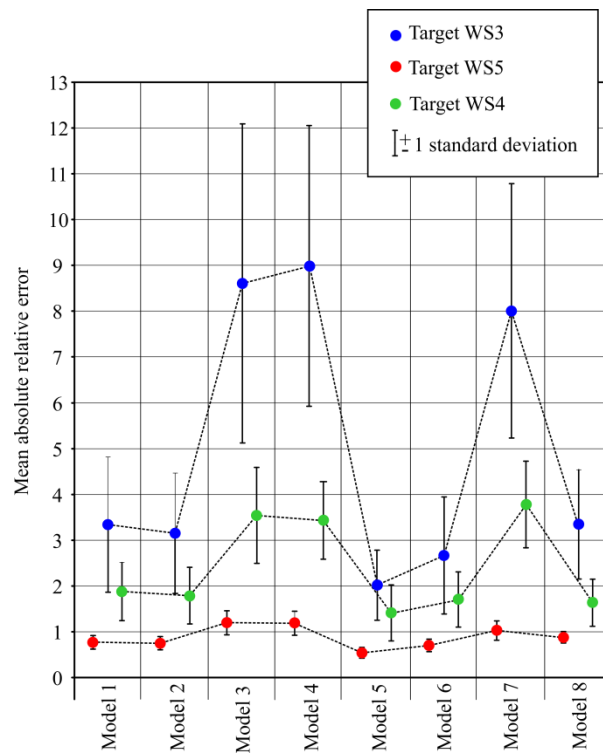


Figura 29 Errores MARE obtenidos cuando se estiman las densidades de potencia eólica en cada modelo y estación objetivo seleccionada

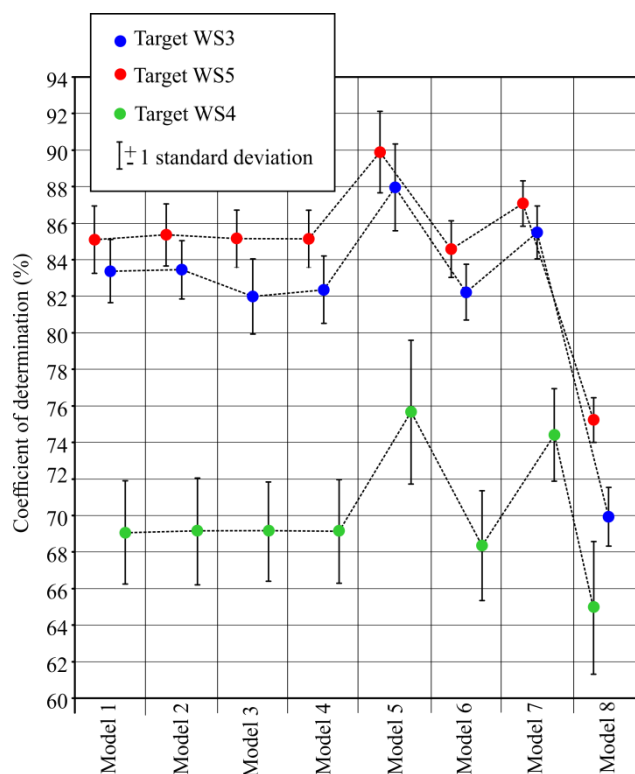


Figura 30 Coeficientes R^2 obtenidos cuando se estiman las densidades de potencia eólica en cada modelo y estación objetivo seleccionada

Conforme a la información presentada en la Tabla 21, el uso del algoritmo Recursive Feature Elimination no ha supuesto una mejora trascendental de la capacidad de predicción de los modelos. De los 24 modelos ejecutados en este análisis comparativo (considerando las tres estaciones objetivo), sólo se eliminaron variables de 8 modelos concretos, estableciéndose la mejora del RSME en valores por lo general inferiores al 1%. Sin embargo, se ha comprobado que los tiempos de computación aumentaron en un 800% con respecto a la situación en la que no se utiliza el método Wrapper y se toman todas las estaciones de referencia disponibles. A pesar de lo descrito, su utilización se considera aconsejable cuando se sospecha que la posible redundancia entre los predictores considerados pudiera introducir demasiado ruido perjudicando el proceso de estimación. Por tanto, puede concluirse que la capacidad de regulación del método SVR ha sido suficiente en la mayoría de los casos para resolver los posibles problemas de sobreajuste en las simulaciones.

Se puede observar en la Tabla 21 que en todos los casos analizados en los que intervienen como features de entrada las velocidades del viento estas son las que presentan mayor relevancia para los modelos (Tabla 21), frente al resto de tipo de features potencialmente disponibles para ser utilizadas. En el caso de los modelos 4 y 6 son las densidades de potencia eólica las features que presentan mayor relevancia para dichos modelos. Los modelos 7 y 8, que en la práctica son los dos únicos casos en los que se consideran features de velocidad y densidad de potencia al mismo tiempo, siempre han antepuesto los registros de velocidad frente a los de densidad de potencia.

En la Tabla 21 se presenta un resumen de los aspectos de mayor importancia en relación con el método wrapper aplicado. En la misma se muestra el número de features seleccionadas por el algoritmo frente al número total de features disponibles en cada modelo, los errores RSME que se han obtenido en el proceso de evaluación (tarea 5 en Figura 27) utilizando todas las features disponibles y los RMSE obtenidos en el mencionado proceso utilizando el conjunto de features seleccionadas al aplicar el método wrapper, la relación de las cinco features de mayor influencia y las features descartadas en cada caso por el método wrapper. En la Tabla 22 se muestran los hiperparámetros óptimos establecidos con la técnica Grid Search y el método heurístico para cada uno de los modelos ejecutados basados en la técnica SVR.

El orden de la relevancia de las features seleccionados está en función de la magnitud de su coeficiente de correlación con la serie objetivo. En este sentido, la feature de mayor relevancia es aquella que presenta el más alto coeficiente de correlación con respecto al target analizado, ya sea en términos de velocidad (Tabla 19) o de densidad de potencia eólica (Tabla 20).

En las Tabla 23 – 25 se muestran los resultados del análisis de significancia estadística realizado para cada una de las métricas utilizadas en este estudio. Las tres tablas toman una estructura semejante, presentándose de forma matricial los p-valores obtenidos con el test de permutación no paramétrico (para conjuntos pareados) y ajustados según el método BH en la comparativa entre pares de modelos.

Estación Objetivo	SVR Modelo		Variables selected/All Features	RSME		5 Features más relevantes					Inputs eliminados					
				Todos Features	Wrapper	1	2	3	4	5	1	2	3	4	5	
WS3	Modelo 1		18/18	1.298		WS5.V	WS4.V	WS1.V	WS2.V	WS8.V	No					
	Modelo 2		18/18	1.298		WS5.V	WS4.V	WS1.V	WS2.V	WS8.V	No					
	Modelo 3		13/18	334.4	332.7	WS5.V	WS4.V	WS1.V	WS2.S	WS6.V	WS7.V	WS8.D	WS7.D	WS1.D	WS9.D	
	Modelo 4		18/18	196.1		WS5.P	WS4.P	WS1.P	WS2.P	WS10.P	No					
	Modelo 5		18/18	1.298		WS5.V	WS4.V	WS1.V	WS2.V	WS8.V	No					
	Modelo 5		ρ	16/18	0.00375	0.00315	WS2.ρ	WS8.ρ	WS7.ρ	WS4.ρ	WS6.ρ	WS6.D	WS7.D	-	-	-
	Modelo 6		18/18	0.409		WS4.P	WS5.V	WS1.P	WS2.P	WS8.P	No					
	Modelo 7		26/27	178.3	177.7	WS5.V	WS4.V	WS1.V	WS2.V	WS8.V	WS7.D	-	-	-	-	
Modelo 8		25/27	0.4562	0.4561	WS4.V	WS5.V	WS2.V	WS9.V	WS5.V	WS2.ρ	WS6.D	-	-	-		
WS4	Modelo 1		18/18	2.191		WS5.V	WS3.V	WS1.V	WS2.V	WS5.D	No					
	Modelo 2		18/18	2.191		WS5.V	WS3.V	WS1.V	WS2.V	WS5.D	No					
	Modelo 3		18/18	898.9		WS5.V	WS3.V	WS1.V	WS8.V	WS6.V	No					
	Modelo 4		18/18	538.4		WS5.P	WS3.P	WS1.P	WS8.P	WS2.P	No					
	Modelo 5		18/18	2.191		WS5.V	WS3.V	WS1.V	WS2.V	WS5.D	No					
	Modelo 5		ρ	17/18	0.00373	0.00369	WS3.ρ	WS8.ρ	WS2.ρ	WS7.ρ	WS9.ρ	WS7.D	-	-	-	-
	Modelo 6		18/18	0.395		WS5.P	WS3.P	WS1.P	WS2.P	WS5.10.D	No					
	Modelo 7		27/27	491.0		WS5.V	WS3.V	WS1.V	WS2.V	WS8.V	No					
Modelo 8		23/27	0.4119	0.4118	WS5.V	WS3.V	WS6.ρ	WS10.V	WS10.ρ	WS1.S	WS1.ρ	WS2.D	WS6.D	-		
WS5	Modelo 1		18/18	1.254		WS4.V	WS3.V	WS1.V	WS6.V	WS2.V	No					
	Modelo 2		18/18	1.254		WS4.V	WS3.V	WS1.V	WS6.V	WS2.V	No					
	Modelo 3		18/18	353.2		WS3.V	WS4.V	WS1.V	WS6.V	WS8.V	No					
	Modelo 4		18/18	208.5		WS3.P	WS4.P	WS1.P	WS6.P	WS2.P	No					
	Modelo 5		18/18	1.254		WS4.V	WS3.V	WS1.V	WS6.V	WS2.V	No					
	Modelo 5		ρ	18/18	0.0064		WS3.ρ	WS1.ρ	WS7.ρ	WS6.ρ	WS8.ρ	No				
	Modelo 6		18/18	0.275		WS4.P	WS3.P	WS1.P	WS2.P	WS6.P	No					
	Modelo 7		26/27	193.8	192.9	WS4.V	WS3.V	WS1.V	WS6.V	WS2.V	WS9.D	-	-	-	-	
Modelo 8		25/27	0.3156	0.3155	WS4.V	WS3.V	WS6.V	WS6.D	WS1.V	WS1.ρ	WS7.ρ	-	-	-		

Tabla 21 Inputs finalmente seleccionados

Modelo/Objetivo	Parámetro característico (C)			Parámetro Intensive Loss Function (ε)			Parámetro característico de la función Kernel Gaussian Radial Basic Function (σ)		
	WS3	WS4	WS5	WS3	WS4	WS5	WS3	WS4	WS5
Modelo 1	7	9	7	0.05	0.10	0.10	0.0296	0.0378	0.0379
Modelo 2	3	9	7	0.05	0.10	0.10	0.0303	0.0378	0.0379
Modelo 3	3	8	3	0.10	0.10	0.10	0.0540	0.0456	0.0449
Modelo 4	3	3	5	0.08	0.08	0.08	0.0326	0.0455	0.0454
Modelo 5	10	10	8	0.01	0.01	0.01	0.0420	0.0484	0.0430
Modelo 6	3	5	5	0.08	0.08	0.08	0.0343	0.0396	0.0406
Modelo 7	5	10	10	0.10	0.10	0.10	0.0192	0.0267	0.0286

Tabla 22 Hiperparámetros de los modelos implementados

A título de ejemplo, en la Tabla 26 se muestran los coeficientes de correlación lineal existentes entre los errores (10 errores obtenidos en los 10 folds) de los modelos que se comparan, para el caso del sitio objetivo WS-5. Puede observarse que existe dependencia positiva entre los distintos modelos. Por tanto, se corrobora el adecuado uso del test pareado de permutación no paramétrico.

En la Tabla 23-25, los modelos "i" representan las columnas y los modelos "j" las filas, Ecuación 4.9. Si la lectura de las tablas se realiza por filas, las celdas con los p-valores escritos en negrita corresponden a aquellos casos donde la hipótesis nula ($H_0: \mu_i = \mu_j$) es rechazada a favor de la hipótesis alternativa ($H_1: \mu_i > \mu_j$). Es decir, el modelo "j" significativamente mejor será aquel que tenga un mayor número de celdas con valores escritos en negrita. Ya que dicho modelo "j" presentará una métrica MAE (Tabla 23) o MARE (Tabla 24) significativamente menor (5% nivel de significancia) que la correspondiente al modelo i con el que se compara. En la Tabla 25, los modelos "i" representan las filas y los modelos "j" las columnas. En este caso, el significativamente mejor que el modelo "i" será aquel que tenga un mayor número de celdas de la fila con valores escritos en negrita, ya que dicho modelo "i" presentará una métrica R^2 significativamente mayor que las correspondientes a los modelos "j".

De la observación de las Figura 28 – 30 se deduce que es el modelo 5, integrado por los submodelos que estiman la densidad del aire y la velocidad del viento, el que en la evaluación de hipótesis basada en precisión ha presentado las mejores métricas MAE, MARE y R^2 , independientemente de la estación objetivo considerada. Además, como se muestra en las Tablas 10-12, las métricas obtenidas al evaluar dicho modelo 5 son, desde el punto de vista estadístico, significativamente (5% nivel de significancia) mejores que las proporcionadas por los restantes modelos analizados en el 100% de los casos cuando se utiliza la métrica MAE, el 95% con la métrica R^2 y el 76% con la métrica MARE. Podríamos atribuir estos resultados al hecho de que estructuralmente el modelo 5 estima las velocidades y las densidades de manera desagregada a través de dos submodelos MCP, lo que le otorga mayor capacidad para reducir los comportamientos redundantes existentes entre las distintas variables ingresadas en el algoritmo de aprendizaje (velocidad y dirección del viento y densidad del aire). El modelo 7 podría considerarse en principio el método más potente en términos de

estructura, prediciéndose directamente la densidad de potencia eólica y utilizando como referencias las velocidades y direcciones del viento y las densidades del aire.

WS3	i	Modelo1	Modelo2	Modelo3	Modelo4	Modelo5	Modelo6	Modelo7	Modelo8
j	Media	118.99	117.84	124.72	123.46	98.12	121.91	115.28	155.04
Modelo1	118.99	-	0.268000	0.009579	0.017780	0.000200	0.079009	0.035000	0.000000
Modelo2	117.84	0.268000	-	0.002965	0.003889	0.000200	0.024933	0.094617	0.000000
Modelo3	124.72	0.009579	0.002965	-	0.268000	0.000000	0.112896	0.000000	0.000000
Modelo4	123.46	0.017780	0.003889	0.268000	-	0.000000	0.213985	0.000560	0.000000
Modelo5	98.12	0.000200	0.000200	0.000000	0.000000	-	0.000000	0.000200	0.000000
Modelo6	121.91	0.079009	0.024933	0.112896	0.213985	0.000000	-	0.001575	0.000000
Modelo7	115.28	0.035000	0.094617	0.000000	0.000560	0.000200	0.001575	-	0.000000
Modelo8	155.04	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	-
WS4	i	Modelo1	Modelo2	Modelo3	Modelo4	Modelo5	Modelo6	Modelo7	Modelo8
j	Media	277.33	273.22	284.04	279.12	219.56	280.77	268.89	299.36
Modelo1	277.33	-	0.257133	0.144060	0.368770	0.000000	0.287538	0.087422	0.001680
Modelo2	273.22	0.257133	-	0.038150	0.178933	0.000000	0.129095	0.248713	0.001556
Modelo3	284.04	0.144060	0.038150	-	0.209745	0.000000	0.287538	0.007700	0.017800
Modelo4	279.12	0.368770	0.178933	0.209745	-	0.000000	0.000000	0.041341	0.002291
Modelo5	219.56	0.000000	0.000000	0.000000	0.000000	-	0.000000	0.000000	0.000000
Modelo6	280.77	0.287538	0.129095	0.287538	0.000000	0.000000	-	0.020347	0.009046
Modelo7	268.89	0.087422	0.248713	0.007700	0.041341	0.000000	0.020347	-	0.000000
Modelo8	299.36	0.001680	0.001556	0.017800	0.002291	0.000000	0.009046	0.000000	-
WS5	i	Modelo1	Modelo2	Modelo3	Modelo4	Modelo5	Modelo6	Modelo7	Modelo8
j	Media	129.09	127.02	131.71	131.53	99.44	130.49	125.65	168.10
Modelo1	129.09	-	0.248835	0.198000	0.205418	0.000000	0.329056	0.104484	0.000000
Modelo2	127.02	0.248835	-	0.042165	0.054133	0.000000	0.121100	0.299717	0.000000
Modelo3	131.71	0.198000	0.042165	-	0.464600	0.000000	0.333631	0.007800	0.000000
Modelo4	131.53	0.205418	0.054133	0.464600	-	0.000215	0.342015	0.009520	0.000000
Modelo5	99.44	0.000000	0.000000	0.000000	0.000215	-	0.000215	0.000215	0.000000
Modelo6	130.49	0.329056	0.121100	0.333631	0.342015	0.000215	-	0.034650	0.000000
Modelo7	125.65	0.104484	0.299717	0.007800	0.009520	0.000215	0.034650	-	0.000000
Modelo8	168.10	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	-

*Si la lectura es por filas (j) es el mejor modelo si la celda está resaltada.

Tabla 23 Análisis de significancia estadística cuando la métrica analizada es MAE. Test pareado

Los resultados obtenidos apuntan a que tantas variables no han sido gestionadas tan eficientemente por los modelos SVR ya que se emplean kernels homocedásticos y, por tanto, en este caso se asumiría la misma variabilidad para todas los features y todo el rango de cada una de ellas,

independientemente de las posibles diferencias de concentración en sus distintas zonas, lo que reduce en cierta medida la capacidad de aprendizaje del modelo, si bien sería la segunda mejor opción conforme a los resultados presentados en el artículo.

WS3	i	Modelo1	Modelo2	Modelo3	Modelo4	Modelo5	Modelo6	Modelo7	Modelo8
j	Media	3.34	3.15	8.60	8.99	2.02	2.67	8.01	3.35
Modelo1	3.34	-	0.407556	0.000000	0.000000	0.009553	0.204800	0.000000	0.496000
Modelo2	3.15	0.407556	-	0.000000	0.000000	0.019133	0.265109	0.000000	0.407556
Modelo3	8.60	0.000000	0.000000	-	0.407556	0.000000	0.000000	0.398533	0.000000
Modelo4	8.99	0.000000	0.000000	0.407556	-	0.000000	0.000373	0.282070	0.000000
Modelo5	2.02	0.009553	0.019133	0.000000	0.000000	-	0.138084	0.000000	0.008225
Modelo6	2.67	0.204800	0.265109	0.000000	0.000373	0.138084	-	0.000000	0.169680
Modelo7	8.01	0.000000	0.000000	0.398533	0.282070	0.000000	0.000000	-	0.000000
Modelo8	3.35	0.496000	0.407556	0.000000	0.000000	0.008225	0.169680	0.000000	-
WS4	i	Modelo1	Modelo2	Modelo3	Modelo4	Modelo5	Modelo6	Modelo7	Modelo8
j	Media	1.87	1.79	3.55	3.43	1.42	1.70	3.78	1.64
Modelo1	1.87	-	0.403719	0.000400	0.000000	0.102900	0.343117	0.000000	0.273600
Modelo2	1.79	0.403719	-	0.000000	0.000560	0.168494	0.403719	0.000400	0.343117
Modelo3	3.55	0.000400	0.000000	-	0.403719	0.000000	0.000000	0.343117	0.000000
Modelo4	3.43	0.000000	0.000560	0.403719	-	0.000000	0.000233	0.273600	0.000000
Modelo5	1.42	0.102900	0.168494	0.000000	0.000000	-	0.242511	0.000000	0.273600
Modelo6	1.70	0.343117	0.403719	0.000000	0.000233	0.242511	-	0.000233	0.405200
Modelo7	3.78	0.000000	0.000400	0.343117	0.273600	0.000000	0.000233	-	0.000000
Modelo8	1.64	0.273600	0.343117	0.000000	0.000000	0.273600	0.405200	0.000000	-
WS5	i	Modelo1	Modelo2	Modelo3	Modelo4	Modelo5	Modelo6	Modelo7	Modelo8
j	Media	0.77	0.75	1.20	1.19	0.54	0.70	1.03	0.88
Modelo1	0.77	-	0.268000	0.000280	0.000000	0.001508	0.145824	0.005600	0.062491
Modelo2	0.75	0.268000	-	0.000000	0.000000	0.001600	0.217862	0.002625	0.032480
Modelo3	1.20	0.000280	0.000000	-	0.456500	0.000000	0.000000	0.078278	0.000933
Modelo4	1.19	0.000000	0.000000	0.456500	-	0.000000	0.000000	0.087267	0.002613
Modelo5	0.54	0.001508	0.001600	0.000000	0.000000	-	0.007811	0.000000	0.000000
Modelo6	0.70	0.145824	0.217862	0.000000	0.000000	0.007811	-	0.000509	0.007000
Modelo7	1.03	0.005600	0.002625	0.078278	0.087267	0.000000	0.000509	-	0.034800
Modelo8	0.88	0.062491	0.032480	0.000933	0.002613	0.000000	0.007000	0.034800	-

*Si la lectura es por filas (j) es el mejor modelo si la celda está resaltada.

Tabla 24 Análisis de significancia estadística cuando la métrica analizada es MARE. Test pareado

WS3	j	Modelo1	Modelo2	Modelo3	Modelo4	Modelo5	Modelo6	Modelo7	Modelo8
i	Media	83.38%	83.46%	81.99%	82.35%	87.96%	82.23%	85.50%	69.93%
Modelo1	83.38%	-	0.451700	0.079467	0.129150	0.000200	0.084000	0.007700	0.000000
Modelo2	83.46%	0.451700	-	0.066021	0.100678	0.000200	0.068180	0.009059	0.000000
Modelo3	81.99%	0.079467	0.066021	-	0.381584	0.000000	0.419677	0.000373	0.000000
Modelo4	82.35%	0.129150	0.100678	0.381584	-	0.000000	0.446963	0.000000	0.000000
Modelo5	87.96%	0.000200	0.000200	0.000000	0.000000	-	0.000000	0.009489	0.000000
Modelo6	82.23%	0.084000	0.068180	0.419677	0.446963	0.000000	-	0.000000	0.000000
Modelo7	85.50%	0.007700	0.009059	0.000373	0.000000	0.009489	0.000000	-	0.000000
Modelo8	69.93%	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	-
WS4	j	Modelo1	Modelo2	Modelo3	Modelo4	Modelo5	Modelo6	Modelo7	Modelo8
i	Media	69.06%	69.13%	69.12%	69.13%	75.65%	68.34%	74.41%	64.96%
Modelo1	69.06%	-	0.500000	0.500000	0.500000	0.000933	0.383473	0.001273	0.010827
Modelo2	69.13%	0.500000	-	0.500000	0.500000	0.001273	0.369067	0.000933	0.010554
Modelo3	69.12%	0.500000	0.500000	-	0.500000	0.000933	0.369067	0.001244	0.011375
Modelo4	69.13%	0.500000	0.500000	0.500000	-	0.001633	0.369067	0.001244	0.010800
Modelo5	75.65%	0.000933	0.001273	0.000933	0.001633	-	0.001200	0.330244	0.000000
Modelo6	68.34%	0.383473	0.369067	0.369067	0.369067	0.001200	-	0.000000	0.029812
Modelo7	74.41%	0.001273	0.000933	0.001244	0.001244	0.330244	0.000000	-	0.000000
Modelo8	64.96%	0.010827	0.010554	0.011375	0.010800	0.000000	0.029812	0.000000	-
WS5	j	Modelo1	Modelo2	Modelo3	Modelo4	Modelo5	Modelo6	Modelo7	Modelo8
i	Media	85.10%	85.35%	85.14%	85.14%	89.89%	84.58%	87.07%	75.23%
Modelo1	85.10%	-	0.440272	0.500370	0.500370	0.000255	0.316018	0.011694	0.000000
Modelo2	85.35%	0.440272	-	0.440272	0.440272	0.000255	0.218695	0.015400	0.000000
Modelo3	85.14%	0.500370	0.440272	-	0.515400	0.000000	0.280933	0.005600	0.000000
Modelo4	85.14%	0.500370	0.440272	0.515400	-	0.000467	0.280933	0.004853	0.000000
Modelo5	89.89%	0.000255	0.000255	0.000000	0.000467	-	0.000255	0.003600	0.000000
Modelo6	84.58%	0.316018	0.218695	0.280933	0.280933	0.000255	-	0.002369	0.000000
Modelo7	87.07%	0.011694	0.015400	0.005600	0.004853	0.003600	0.002369	-	0.000000
Modelo8	75.23%	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	-

*Si la lectura es por filas (i) es el mejor modelo si la celda está resaltada

Tabla 25 Análisis de significancia estadística cuando la métrica analizada es R². Test pareado

Llegados a este punto debemos señalar que el método 5 propuesto permite determinar las funciones de la densidad de potencia eólica que han sido propuestas en la literatura científica relacionada con las energías renovables. Dado que el modelo 5 estima a partir de sus dos submodelos, las series de densidades del aire y de velocidades del viento, permite, además de estimar las WPDs, construir las distribuciones de probabilidad de las densidades de potencia $f_{WPD}[\rho, v]$ que consideran que ρ y v son

variables aleatorias estadísticamente dependientes y por tanto, tienen en cuenta sus distribuciones conjuntas [58] $f_{\rho v}(\rho, v)$ Ecuación 4.10.

$$f_{WPD}[\rho, v] = \frac{\rho v^3 f_{\rho v}(v, \rho)}{\int_0^{\infty} \int_0^{\infty} \rho v^3 f_{\rho v}(\rho, v) d\rho dv} \quad (4.10)$$

El modelo 5 también permite, dado que uno de sus submodelos predice la velocidad del viento, construir las distribuciones de probabilidad de densidad de potencia siguiendo el procedimiento más frecuentemente utilizado en la bibliografía científica [218]. Es decir, formular la distribución considerando que ρ y v son variables aleatorias estadísticamente dependientes. Por tanto, sus distribuciones conjuntas $f_{\rho v}(\rho, v)$ son igual al producto de sus probabilidades marginales. Además, se suele asumir que la densidad del aire es constante. Por tanto, la distribución de probabilidad de las densidades de potencia depende solo de la velocidad del aire $f_{WPD}[\rho, v]$. Sin embargo, el modelo 7, dado que estima únicamente WPDs no permitiría construir las distribuciones de probabilidad de las densidades de potencia con los procedimientos propuestos [58,218].

Los resultados recogidos en las Figura 28 – 30 y en las Tabla 23 - 25, ponen de manifiesto que los modelos que han sido entrenados minimizando el error en la predicción de las velocidades del viento no logran minimizar también el error en la predicción de las WPDs. Es decir, se requiere utilizar modelos MCP especialmente entrenados para minimizar el error en la predicción de las WPDs o modelos en los que además de predecirse la velocidad contemple la predicción de la densidad del aire para determinar seguidamente las WPDs.

Por ello, la variación de la densidad del aire es importante que sea considerada incluso en regiones a nivel del mar donde en términos medios se cumplen los estándares establecidos por la ISA (International Standard Atmosphere) para la definición del valor de densidad del aire 1.225 kg m^{-3} .

Como puede observarse en las Tabla 23 - 25, no existen diferencias estadísticamente significativas entre las métricas de los modelos 1 y 2. El motivo de ello es que todas las posiciones evaluadas en este estudio se encuentran a nivel del mar. Sin embargo, si se hubiesen utilizado datos de estaciones ubicadas a gran altitud donde existiría una mayor diferencia entre el valor estándar de (1.225 kg m^{-3}) y la media de densidades del aire del emplazamiento ($\bar{\rho}$), se estima que la diferencia entre los modelos 1 y 2 podría ser apreciable. Asimismo, se estima que en esta circunstancia la diferencia entre las métricas del modelo 1 y el modelo 5 se incrementaría.

Los modelos 3 y 4 han sido los que peores métricas han generado, independientemente de la estación objetivo involucrada, cuando se emplea la técnica SVR. No han existido diferencias estadísticamente significativas entre dichos modelos para ninguna de las métricas empleadas. En el caso concreto del modelo 4, el uso directo de las WPDs como features de referencia elimina las influencias que sobre la estación objetivo tienen las features de velocidad. Se desprende de los resultados obtenidos que los mejores modelos son siempre aquellos que de una forma u otra incorporan como predictores las velocidades en las estaciones de referencia y esto se debe a que en

determinadas ocasiones, diferentes combinaciones de valores de velocidades en las estaciones de referencia producen valores similares de densidad de potencia eólica en la estación objetivo, y por ende, cuando se rechazan las velocidades como predictores a favor de las densidades, se pierden algunas interrelaciones relevantes entre referencias. Por su parte, los resultados del modelo 3 sugieren que no conviene despojar a los modelos de Machine Learning de uno de sus puntos fuertes, su capacidad para adoptar la forma funcional más adecuada en cada momento, adaptándose a los datos de partida y extrayendo las distintas relaciones existentes entre estaciones de referencia y objetivo. Cuando se impone la restricción adicional de considerar como muestra de datos las velocidades elevadas al cubo, las estimaciones generadas con estos modelos tienden a empobrecerse y no describen con fiabilidad el comportamiento del sistema que se pretende modelar, obteniéndose una respuesta contraria a la que se obtiene cuando se linealiza las variables de entrada a través de la escala logarítmica.

El modelo 6, que afectos prácticos es la versión linealizada del modelo 4, sigue unas pautas de comportamiento muy parecidas a las descritas en su versión primitiva (modelo 4), si bien el logaritmo ha ayudado a estrechar sensiblemente el rango de valores, facilitando en cierta medida la tarea de aprendizaje del modelo. En el caso particular de la técnica SVR, este estrechamiento del rango de valores con la escala logarítmica tiende a linealizar las relaciones entre los predictores y la variable respuesta, y por tanto, simplifica su forma aunque dicha linealización no se consiga, mejorando las lecturas de error y correlación. Analizando lo sucedido en este estudio, se observa que el modelo 6 produjo un valor de MAE ligeramente inferior al obtenido con el modelo 4 para todas las estaciones objetivo, y estas diferencias también se observaron en el resto de métricas. A pesar de todo ello, las lecturas de las métricas de error no se consideran significativas desde el punto de vista estadístico a excepción de cuando el análisis se realiza tomando como referencia la métrica MARE.

El modelo 8 es el que peores estimaciones ha generado, condicionado por su alto grado de linealidad, siendo el extremo contrario del modelo 7 en términos de cantidad de estructura. De acuerdo con lo descrito, parece obvio que la relación entre la WPD de la estación objetivo y las velocidad y densidades de las estaciones de referencia no son las mismas relaciones lineales que se encuentran implícitas en la definición de la densidad de potencia según la Ecuación 4.8, sólo habiéndose captado parte de estas relaciones con los modelos no lineales.

Si se presta atención al análisis estadístico, para todos los casos comparados, la técnica SVR logra una mejora significativa de las estimaciones. A modo de referencia, el progreso puede ser cuantificado con los indicadores MAE e R^2 . Si se compara el modelo 5 con el modelo 8, se consigue una reducción del error MAE de entre el 27 – 41%, y un incremento del coeficiente de determinación de entre el 16 – 26% con dependencia de la estación objetivo evaluada.

Metric MAE								
Modelo	Modelo1	Modelo2	Modelo3	Modelo4	Modelo5	Modelo6	Modelo7	Modelo8
Modelo1	1.00	0.99	0.88	0.86	0.56	0.94	0.90	0.30
Modelo2	0.99	1.00	0.94	0.93	0.59	0.97	0.91	0.42
Modelo3	0.88	0.94	1.00	0.99	0.58	0.96	0.83	0.64
Modelo4	0.86	0.93	0.99	1.00	0.54	0.94	0.87	0.69
Modelo5	0.56	0.59	0.58	0.54	1.00	0.65	0.33	0.23
Modelo6	0.94	0.97	0.96	0.94	0.66	1.00	0.86	0.44
Modelo7	0.90	0.91	0.83	0.87	0.33	0.86	1.00	0.42
Modelo8	0.30	0.42	0.64	0.69	0.23	0.44	0.42	1.00
Metric MARE								
Modelo	Modelo1	Modelo2	Modelo3	Modelo4	Modelo5	Modelo6	Modelo7	Modelo8
Modelo1	1.00	1.00	0.85	0.81	0.72	0.97	0.92	0.96
Modelo2	1.00	1.00	0.86	0.82	0.73	0.96	0.93	0.96
Modelo3	0.85	0.86	1.00	0.99	0.87	0.71	0.95	0.80
Modelo4	0.81	0.82	0.99	1.00	0.89	0.66	0.95	0.77
Modelo5	0.72	0.73	0.87	0.89	1.00	0.66	0.87	0.79
Modelo6	0.98	0.96	0.71	0.66	0.66	1.00	0.84	0.97
Modelo7	0.92	0.93	0.95	0.95	0.87	0.84	1.00	0.90
Modelo8	0.96	0.96	0.80	0.77	0.79	0.97	0.90	1.00
Metric R ²								
Modelo	Modelo1	Modelo2	Modelo3	Modelo4	Modelo5	Modelo6	Modelo7	Modelo8
Modelo1	1.00	0.99	0.91	0.91	0.60	0.90	0.98	0.50
Modelo2	0.99	1.00	0.96	0.96	0.63	0.95	0.98	0.55
Modelo3	0.91	0.96	1.00	1.00	0.66	0.96	0.92	0.65
Modelo4	0.91	0.96	1.00	1.00	0.63	0.96	0.93	0.65
Modelo5	0.60	0.63	0.66	0.63	1.00	0.62	0.49	0.46
Modelo6	0.90	0.95	0.96	0.96	0.62	1.00	0.92	0.46
Modelo7	0.98	0.98	0.92	0.93	0.49	0.92	1.00	0.50
Modelo8	0.50	0.55	0.65	0.65	0.46	0.46	0.50	1.00

Tabla 26 Coeficientes de correlación lineal entre resultados de las métricas en cada Folds. WS-5

4.2.8. Conclusiones del estudio

En el estudio presentado en este apartado se han propuesto por primera vez diferentes modelos MCP, que, utilizando múltiples estaciones de referencia, tienen como propósito principal estimar las WPDs a largo plazo en un sitio objetivo. Y ello, considerando no solo la variabilidad de las velocidades del viento en las estaciones de referencia, sino también de la direcciones del viento y de las densidades del aire, así como la forma funcional en que dichas variables participan en los modelos MCP propuestos.

De los resultados alcanzados se concluye que los modelos MCP hasta ahora propuestos que han sido entrenados minimizando el error en la predicción de las velocidades del viento, no logran minimizar también el error en la predicción de las WPDs. Es decir, que se requiere utilizar modelos MCP especialmente entrenados para minimizar el error en la predicción de las WPDs (con features de velocidad, dirección y densidad del aire), o modelos en los que además de predecirse la velocidad contemplen la predicción de la densidad del aire. Por ello, es importante considerar la variación

anual de la densidad del aire incluso en regiones a nivel del mar donde en términos medios se cumplen los estándares establecidos por la ISA (densidad estándar próxima a 1.225 kg m^{-3}).

Asimismo, se concluye que de los ocho modelos comparados en el estudio realizado, el modelo 5 es el que ha logrado las mejores métricas MAE, MARE y R^2 , independientemente de la estación objetivo considerada. Las métricas obtenidas al evaluar dicho modelo 5 son desde el punto de vista estadístico significativamente (5% de nivel de significancia) mejores que las proporcionadas por los restantes modelos analizados en el 100% de los casos cuando se utiliza la métrica MAE, el 95% con la métrica R^2 y el 76% con la métrica MARE. Estos resultados se atribuyen al hecho de que estructuralmente el modelo 5 estima las velocidades del viento y las densidades del aire de manera desagregada a través de dos submodelos MCP, lo que le otorga mayor capacidad para adaptarse a la complejidad y especificidades de cada una de dichas variables y gestionar mejor los comportamientos redundantes existentes entre las distintas variables ingresadas en el algoritmo de aprendizaje.

El modelo 7 constituye el método más potente en términos de estructura a la hora de predecir directamente la densidad de potencia eólica, utilizando como features de referencias las velocidades y direcciones del viento y las densidades del aire. Sin embargo, conforme a los resultados obtenidos en este trabajo constituye la segunda mejor opción entre los ocho métodos analizados.

Conviene mencionar que el método Wrapper ha supuesto un aumento importante de los tiempos de computación. Sin embargo, para el caso de los datos considerados en este paper, los resultados muestran que no se ha producido una mejora considerable de las estimaciones con respecto a la situación de referencia donde se utilizan todas las variables predictoras, siendo la propia capacidad de regulación de la técnica SVR suficiente para gestionar la mayoría de los problemas de sobreajuste. No obstante, se recomienda su uso cuando se sospecha que por el número de variables poco significativas o que, siendo significativas, éstas serían sustituidas por otras covariables ya presentes, podría estar introduciéndose excesivo ruido en la estimación del modelo sin beneficio aparente.

4.3. Evaluación de la fiabilidad de cinco modelos MCP para estimar las WTPO a largo plazo usando tres técnicas de Machine Learning

4.3.1. Introducción

Para estimar la potencia eólica a largo plazo conocidas como Wind Turbine Power Outputs (WTPO) en un sitio objetivo donde solo se dispone de datos meteorológicos medidos en un corto plazo de tiempo han sido utilizados diversos modelos basados en métodos MCP. Los modelos MCP usados hasta ahora comparten el postulado de que es poco relevante la influencia de la variación de la densidad del aire, asumen el valor estándar de 1.225 kg m^{-3} y solo tienen en cuenta aerogeneradores con control blade pitch.

En el estudio realizado en este apartado se evalúa el comportamiento de los modelos usados hasta el momento y otros propuestos en este trabajo, los cuales incorporan la densidad del aire al modelo MCP como una covariable adicional en la estimación de la WTPO a largo plazo y consideran tanto los aerogeneradores con control blade pitch como aerogeneradores con control stall-regulated. Las ventajas de la incorporación de dicha covariable se evalúan utilizando diferentes formas funcionales y diferentes algoritmos de Machine Learning para implementarlas, en concreto las técnicas Artificial Neural Networks, Support Vector Regression y Random Forest.

Como para el estudio desarrollado en el Apartado 4.2, las variables explicativas consideradas en este trabajo proceden de las 10 estaciones anemométricas descritas en el Apartado 4.1, habiéndose usado los datos de velocidad, dirección del viento, humedad relativa, presión y temperatura para 2014 en alturas comprendidas entre 10 y 20 metros.

Fruto del estudio desarrollado en este apartado ha sido publicado un artículo [219] en la revista indexada con índice JRC Q1 Applied Energy.

4.3.2. Antecedentes

A la hora de tomar la decisión de instalar o no una determinada turbina eólica (WT) en un sitio objetivo resulta de interés conocer las Wind Turbine Power Outputs (WTPOs), estimadas haciendo uso de la curva de potencia de dicha WT y de las características del régimen de vientos y de la densidad del aire del lugar donde la misma se pretende instalar. Las WTPO medias a largo plazo en un sitio objetivo permiten pronosticar si dicha cantidad de energía será suficiente para satisfacer una demanda dada. Además, dicho parámetro tiene una fuerte influencia la efectividad del aerogenerador [11]. Como la curva de potencia de una WT depende de la densidad del aire y de la velocidad del viento a la altura de su buje, la WTPO dependerá también de dichas variables [43].

Dado que en muchas ocasiones solo se dispone en el sitio objetivo de medidas registradas a corto plazo de los datos meteorológicos de los que depende la WTPO, en la literatura científica relacionada con las energías renovables se han propuestos diversos modelos para estimar las WTPOs [11,13,16,17,26,44] a largo plazo. Para ello, dichos modelos suelen utilizar los denominados métodos MCP, que se apoyan en la información meteorológica registrada durante largo plazo en estaciones vecinas que sirven de referencia [13].

Respecto al número de estaciones de referencia utilizadas hay que señalar que la mayoría de los modelos propuestos hasta ahora (algunos de ellos implementados en los programas de la industria de la energía eólica) se han apoyado en una única estación de referencia [13]. Sin embargo, con el propósito de capturar la mayor información posible del comportamiento eólico de la zona donde se encuentra el sitio objetivo, existe una creciente tendencia en la propuesta de modelos que utilizan múltiples estaciones de referencia [11,13,16,17,26,44]. Dichos modelos utilizan técnicas de Machine Learning (ML) que permiten no solo la utilización simultánea de múltiples estaciones de referencia sino que son capaces de captar relaciones no lineales entre las features [44,45,203] involucradas.

Llegado a este punto, hay que señalar que en la gran mayoría de los casos, la técnica de ML utilizada ha sido entrenada de forma supervisada de tal manera que se minimice el error en la estimación de las velocidades del viento en el sitio objetivo. Una vez estimadas dichas velocidades, estas se usan para estimar las WTPOs [11,16,17,26,44]. En este contexto, Velázquez et al. [17] han utilizado redes neuronales artificiales (ANNs) que tienen como features de entrada las velocidades y direcciones del viento de múltiples estaciones de referencia para estimar las velocidades del viento a largo plazo en un sitio objetivo. Dichas velocidades estimadas se emplean posteriormente para estimar las potencias generadas por diversas WT con blade pitch control de diferentes potencias nominales, haciendo uso de las curvas de potencia de las mismas. Con el mismo tipo de variables de entrada y de salida y el mismo objetivo, Carta et al., [16] han usado Bayesian Networks. Sin embargo, Zhang et al. [6] han utilizado un método MCP híbrido. Las técnicas de ML utilizadas por estos autores han sido ANN y Support Vector Regression (SVR), cada una de las cuales tiene como entrada la velocidad del viento de una única estación de referencia y como salida la velocidad del viento de la estación objetivo. Las direcciones del viento no constituyeron variables de entrada en dichas técnicas, sino que los autores han utilizado el método binning de dirección del viento ampliamente utilizado en los métodos MCP que se apoyan en una única estación de referencia [13]. Por tanto, las técnicas de ML (ANN y SVR) se han aplicado a los datos de velocidades del viento asignados a cada uno de los sectores de dirección que han definido para cada una de las estaciones de referencia consideradas.

En dichos estudios [16,17,26] los modelos MCP no han sido entrenados y evaluados con el propósito principal de estimar las WTPOs, sino que estas se han estimado de forma marginal. Es decir, los modelos MCP no han sido especialmente entrenados para minimizar el error en la predicción de las WTPOs, sino entrenados para minimizar el error en la predicción de las velocidades del viento. Además, en dichos trabajos se ha asumido que la densidad del aire de la estación objetivo es constante en el tiempo y se toma el valor típico de 1.225 kg m^{-3} , correspondiente a condiciones de atmósfera estándar (aire completamente seco, presión y temperatura medias a nivel del mar de 1013.25 hPa y 15°C , respectivamente). Las curvas de potencia de las WTs utilizadas en dichos estudios cuentan con blade pitch control y han sido las proporcionadas por los fabricantes para dicha específica densidad del aire. Sin embargo, en un estudio realizado por Jung y Kwon [44] se indica que una ANN entrenada con una medida de error convencional puede subestimar significativamente la producción anual de energía. En este contexto, estos autores [44] con el objetivo de mejorar la precisión de la energía producida por una WT calculada con las velocidades estimadas del viento a largo plazo utilizan ANNs con funciones de error ponderadas. Sin embargo, dichos autores no consideran en su estudio la variación de la densidad del aire, argumentando que su influencia es secundaria para la producción de energía de una WT y que su magnitud es mucho más pequeña que la variación de la velocidad del viento.

Además, señalan que la producción de energía es proporcional a la densidad del aire mientras que es proporcional al cubo de la velocidad del viento. Sin embargo, como señala Hau [43] las diferencias entre las densidades del aire registradas en sitios ubicados en alturas que se diferencian entre sí en unos pocos cientos de metros puede ser notoria, de tal manera que la influencia de la densidad del aire en el rendimiento de una WT no debería ser despreciada. También Liu y Liu [220] indican que en

algunas áreas interiores de China las altitudes de las ubicaciones de los parques eólicos son generalmente elevadas y, por tanto, la densidad del aire difiere de la densidad estándar del aire. Dichos autores señalan que, aunque la velocidad del viento es el principal factor que afecta a la potencia de salida del aerogenerador, los efectos de la densidad del aire deben ser considerados en la estimación de la potencia en la etapa de diseño de un parque eólico. Además, el cambio en el rango de temperaturas entre verano e invierno influye en las variaciones de las densidades del aire, que deberían tenerse en cuenta, especialmente en los países calientes [43]. Sin embargo, hay diferencias importantes en el nivel de influencia de la variación de la densidad del aire en la curva de potencia de una WT dependiendo del tipo de regulación de la misma. En el caso de stall-regulated WTs la potencia producida se ve modificada por la densidad del aire en todo el rango de velocidades de funcionamiento [87]. Es decir, desde la velocidad de arranque a la velocidad nominal y de la velocidad nominal a la velocidad de parada. Por tanto, un cálculo de la producción de energía de la WT sin tener en cuenta la variación de la densidad del aire podría conducir a una variación no despreciable de la cantidad de potencia esperada. Sin embargo, en las WTs con control de paso de pala (es decir, aerogeneradores con control activo de potencia), la influencia de la variación de la densidad del aire es menos severa. Ello es debido a que en este tipo de WTs la variación en la densidad del aire no tiene influencia en la variación de la potencia eléctrica generada por las mismas cuando estas trabajan en el rango de carga completa [87]. Sin embargo, en el intervalo de carga parcial, la curva de potencia generada varía proporcionalmente a la densidad del aire como lo hace la curva de potencia de las WTs de paso fijo (stall-regulated WTs).

4.3.3. Objetivo del estudio desarrollado

Del análisis bibliográfico realizado se desprende que cuando se han utilizado MCP métodos, basados en información proporcionada por múltiples estaciones de referencia, para estimar las WTPOs a largo plazo en un sitio objetivo se han utilizado técnicas de ML que generalmente se entrenan para estimar las velocidades del viento y a partir de ellas se estiman las WTPOs. Sin embargo, los modelos propuestos desprecian la variación de la densidad del aire, asumen que esta es constante de valor 1.255 kg m^{-3} , correspondiente a condiciones de atmosfera estándar y consideran únicamente WTs con blade pitch control.

En este contexto, uno de los objetivos del estudio desarrollado en este apartado, que supone una aportación original, es que por primera vez se evalúa el comportamiento de cinco modelos propuestos para estimar las WTPO a largo plazo en un sitio objetivo donde solo se dispone de datos meteorológicos medidos en un corto plazo de tiempo. Dichos modelos se diferencian entre sí en la manera funcional en la que los features de densidad del aire, velocidad del viento y dirección intervienen y por la técnica de ML utilizada en los mismos para realizar la regresión.

Dada las diferentes influencias que la densidad del aire tiene en las curvas de potencia de las WTs dependiendo del tipo de regulación de las mismas, en este estudio los modelos contemplan tanto las WTs con blade pitch control como stall-regulated WTs. De esta forma se garantiza que los resultados aquí obtenidos serían consistentes con la extensa gama de modelos de aerogenerador disponibles en

el catálogo comercial existente en la actualidad en contra del proceso comúnmente utilizado en el que sólo se simulan aerogeneradores con control blade pitch.

Las técnicas de ML seleccionadas han sido ANN con arquitectura Multi-Layer perceptron (MLP), SVR y Random Forest (RF). La elección de la ANN con arquitectura MLP ha sido consecuencia de que la misma, según se desprende de análisis bibliográfico realizado [11,13,17,43-45] y de los aspectos expuestos en el Capítulos 3, ha sido la técnica más frecuentemente utilizada como método MCP para estimar las WTPOs a largo plazo en un sitio objetivo. La utilización de la SVR está motivada por ser una de las técnicas que mejor representan el estado del arte de ML debido a su comprobada capacidad de predicción de primera clase en muy diferentes escenarios, mostrándose con frecuencia superior a la ANNs con arquitectura MLP [211]. Esta capacidad predictiva está fundamentada, como señala Díaz et al. [203] (Apartado 4.2), no sólo en su propiedad de aproximación universal a cualquier función continua (presente también en las ANNs) sino también en un algoritmo de entrenamiento más eficaz y estable que proporciona siempre una solución única al problema de estimación y una mayor “Sparsity” en dicha solución.

Finalmente, la técnica de RF, propuesta por Breiman [188], resulta de una combinación de bagging, subspace sampling y regression trees [221] y se ha seleccionado en este trabajo por ser una técnica de ML emergente que ha crecido en popularidad en los últimos años, dado que su estrategia de operación ha resultado ser muy buena en comparación con muchas otras técnicas, incluyendo las SVRs y las ANNs, además de ser robusta contra el sobreajuste [188]. Aunque hasta el momento la técnica de RF no ha sido implementada en modelos basados en métodos MCP, está siendo usada con éxito en diversos estudios relacionados con la energía [222-224]. En este contexto, Ma y Cheng [222] han investigado la influencia de 171 features posiblemente relacionadas con el uso de la energía en el sector residencial usando la técnica RF. Los autores comparan RF con métodos lineales típicos tales como Multiple Linear Regression y Lasso, concluyendo que los resultados obtenidos con el modelo RF poseen menor error cuadrático medio.

Ibrahim y Khatib [223] han presentado un modelo híbrido para la predicción horaria global de radiación solar usando la técnica RF y el algoritmo firefly. Según los autores, los resultados del modelo propuesto son comparados con métodos típicos ANN y muestran mejores resultados que las técnicas base utilizada.

La técnica RF ha sido también propuesta por Lahouar y Slama [224] para construir un modelo de predicción eólica a una hora vista. Según los autores, los resultados muestran una interesante mejora de la predicción así como una reducción importante de los errores comparados con métodos ANN.

4.3.4. Muestra de datos usadas para el desarrollo del estudio

Se dispone de medidas medias horarias de velocidad y dirección del viento, así como de temperatura, presión atmosférica y humedad relativa del aire, registradas durante el año 2014 en 10 estaciones climatológicas (WSs) instaladas en las Islas Canarias (España), Figura 23.

En la Tabla 10 se muestran los códigos asignados a cada estación meteorológica, sus altitudes y alturas sobre el nivel del suelo en las que las series de datos de viento han sido captadas. Los termómetros e higrómetros fueron instalados a 2 metros sobre el nivel del suelo y los barómetros se posicionaron en la misma situación en la que se encuentran los sensores de viento. También se recoge en la Tabla 10 las medias anuales, las desviaciones estándar, los valores máximos y mínimos de las series de velocidades del viento registradas durante el año 2014, así como los valores correspondientes a las densidades del aire. Las densidades del aire han sido obtenidas a partir de las series de datos de temperatura, presión atmosférica y humedad relativa del aire registradas durante el año 2014, haciendo uso de una ecuación de estado y de las fórmulas y procedimiento usado por Díaz et al. [203]. Asimismo, se indica en el Apartado 4.1 la fuente de procedencia de los datos meteorológicos utilizados.

Estos datos meteorológicos son recabados en origen con sensores de alta precisión que recaban la medida con una frecuencia inferior a 10 segundos, procediéndose al cálculo de los valores medios horarios por variable en el propio equipo de adquisición de datos. Dicho procedimiento es aplicado de modo semejante en todas las estaciones anemométricas usadas para la redacción del artículo en cuestión.

En la Figura 24 – 25 se representan los histogramas de frecuencia de las velocidades del viento, los histogramas de frecuencia de las densidades de potencia eólica (calculadas estas en el supuesto generalmente realizado en la literatura científica [218] de que las densidades del aire fuesen constantes) y las funciones de densidad de probabilidad (PDF) de ambas variables. En lugar de usar las distribuciones mezcla de Weibull (dos parámetros) se ha usado la mezcla singular truncada Normal–Weibull, dado que esta aporta muy buenos resultados tanto para distribuciones de frecuencia unimodal como bimodal en la región evaluada de acuerdo con los errores relativos obtenidos en base a la media anual de WPD [62,63]. También se representan en estas ilustraciones las rosas de los vientos y las medias mensuales y horarias de las velocidades del viento para las 10 estaciones en el mismo periodo.

4.3.5. Descripción de las técnicas y modelos matemáticos usados para la simulación

4.3.5.1 Cálculo de las curvas de potencia de las turbinas eólicas

Como ha sido descrito en el Apartado 2.8, los fabricantes de aerogeneradores, siguiendo procedimientos normalizados, tales como los recogidos en IEC 61400-12-1 [87], suelen proporcionar las curvas de potencia de los modelos de aerogenerador ofertados estableciendo una equivalencia entre la velocidad del viento a altura del buje (z) y la potencia que sería producida en ese instante en formato discretizado asumiendo unos valores concretos de densidad del aire orientativos entre los que siempre se encuentra el valor estándar de 1.225 kg m^{-3} (ρ_0). Dependiendo de la estrategia de control, la curva de potencia del aerogenerador puede tener una forma concreta, distinguiéndose fundamentalmente dos tipos principales, aerogeneradores con control tipo stall – regulated y aerogeneradores con control active power (pitch – regulated).

Dado que el propósito de este estudio es evaluar modelos propuestos para estimar las WTPO a largo plazo en un sitio objetivo donde solo se dispone de datos meteorológicos medidos en un corto plazo de tiempo, no es posible utilizar los modelos de curvas de potencia empíricos. En su lugar serán utilizados los modelos generales aplicables a las dos grandes familias de aerogeneradores (stall – regulated y pitch – regulated, formulaciones las cuales han sido expuestas en el Apartado 2.8.

Por otra parte, con el propósito de estimar las relaciones entre velocidades del viento y potencias generadas por las WT entre los puntos de las curvas de potencia proporcionadas por los fabricantes en este estudio se ha utilizado una interpolación con funciones spline cúbicas [93].

4.3.5.2 Técnicas de Machine Learning

En este estudio para construir los distintos modelos MCP formulados para llevar a cabo la estimación de las WTPOs en un sitio objetivo se utiliza un enfoque de regresión múltiple. La función de regresión en cada uno de los modelos evaluados, es estimada utilizando tres técnicas de ML, la técnica ANN, la técnica SVR y la técnica RF.

Se ha expuesto en el Capítulo 3 una descripción de las técnicas de Machine Learning usadas en este estudio. En cualquier caso, para un análisis detallado puede dirigirse a [148-150] para ANN, [150,151] para SVR y [150,152] para RF.

Como ha sido justificado en el Apartado 3.4, para la construcción de las redes neuronales artificiales utilizadas en este trabajo se ha recurrido a la metodología Multi-layer Perceptron (MLP) con función de activación sigmoidea. Asimismo, el algoritmo backpropagation [148,149] es utilizado para el aprendizaje de la red, es decir para estimar los parámetros desconocidos de la misma (pesos y bias). Además, para la programación de los modelos MCP basados en ANN utilizados en este trabajo se ha utilizado el paquete nnet [156] del software multiplataforma de licencia libre R Statistics [154].

4.3.5.2.1 Support Vector Regression

Por otra parte, como ya ocurría en el estudio desarrollado en el Apartado 4.2, para la resolución del problema SVR ha sido seleccionada la técnica denominada ϵ -SVR [151]. Para la programación de los modelos MCP basados en SVR utilizados en este trabajo se ha utilizado el paquete Kernlab [164] del software multiplataforma de licencia libre R Statistics.

Finalmente, para la programación de los modelos MCP basados en RF utilizados en este trabajo se ha utilizado el paquete randomForest [174] del software multiplataforma de licencia libre R Statistics, habiéndose presentado en la parte final del mencionado capítulo una descripción de la estructura del código definida como caso general para la resolución de problemas de regresión.

4.3.5.3 Modelos aplicados para estimar las WTPOs con métodos MCP

Los modelos analizados en este trabajo se diferencian entre sí en la forma funcional en la que los features de densidad del aire, velocidad del viento y dirección intervienen y por la técnica de ML utilizada (ANN, SVR y RF) en los mismos para realizar la regresión. Seguidamente se describen los

cinco modelos (M1,...,M5) para estimar las WTPOs, basados en métodos MCP con múltiples estaciones de referencia, que se pretenden evaluar y se muestran las formas funcionales de los mismos.

El modelo M1, representado en Ecuación 4.11, utiliza el método más frecuentemente propuesto en la literatura científica [16,17,26] para estimar a largo plazo las WTPOs en un sitio objetivo utilizando métodos MCP con múltiples estaciones de referencia. Como ya ha sido indicado, las tres técnicas de ML (ANN, SVR, RF), utilizando las velocidades y direcciones del viento de los sitios de referencia como variables de entrada, se entrenan en este modelo buscando hacer mínimo el error en la predicción de la velocidad del viento del sitio objetivo. Una vez estimadas las velocidades del sitio objetivo \widehat{V}_t se estiman las potencias producidas por las turbinas eólicas \widehat{WTPO}_t haciendo uso de la curva de potencia $P(V, \rho_0)$ de la WT que proporciona el fabricante de la misma referida a condiciones estándares ISO, $\rho_0 = 1.225 \text{ kg m}^{-3}$, (Figura 31).

$$\widehat{WTPO}_t = P(\widehat{V}_t, \rho_0) = P(f_v(V_{1f}, \dots, V_{kt}, \dots, V_{pt}, D_{1f}, \dots, D_{kt}, \dots, D_{pt}), \rho_0) \quad (4.11)$$

Tanto en Ecuación 4.11, como en las Ecuación 4.12, Ecuación 4.13, Ecuación 4.14 y Ecuación 4.16, $f_Z(X_1, \dots, X_d)$ representa a una función de regresión con la cual se obtiene una estimación \widehat{Z} de la variable Z con las features X_1, \dots, X_d , donde Z es la variable que se predice en cada caso (la velocidad del aire V , la densidad del aire ρ , etc.). El subíndice t indica el instante evaluado y el subíndice $k \in \{1, \dots, p\}$ se refiere a la estación de referencia k -ésima de las p WSs disponibles.

En el modelo M2, representado en Ecuación 4.12, se estiman las velocidades del viento siguiendo exactamente el mismo procedimiento que el M1. La única diferencia es que a la hora de estimar las potencias producidas, en lugar de utilizar la curva de potencia de la WT que proporciona el fabricante de la misma referida a condiciones estándares ISO, se emplea la curva de potencia corregida (según se indica en el Apartado 2.8) para la densidad media del aire ($\bar{\rho}$) calculada a la altura de buje del aerogenerador en el sitio objetivo durante el periodo de disposición de series de datos de dicha variable. Por tanto, M2 se diferencia de M1 en que pretende tener en cuenta los efectos en la producción de las WTs señaladas por Hau [43] y Liu y Liu [220] cuando la densidad del aire registrada en los sitios objetivos no coinciden con la densidad estándar del aire generalmente utilizada en la literatura científica.

$$\widehat{WTPO}_t = P(\widehat{V}_t, \bar{\rho}_0) = P(f_v(V_{1f}, \dots, V_{kt}, \dots, V_{pt}, D_{1f}, \dots, D_{kt}, \dots, D_{pt}), \bar{\rho}_0) \quad (4.12)$$

El modelo M3, representado en Ecuación 4.13, coincide con los modelos M1 y M2 en el proceso seguido para estimar las velocidades del viento de la estación objetivo. Sin embargo, difiere de ellos en que contempla la estimación (de forma separada de la estimación de la velocidad del viento) de la densidad del aire a largo plazo en las estaciones objetivo. Por tanto, M3 se apoya en dos submodelos que estiman de forma separada las dos mencionadas variables. Las features de entrada a las tres técnicas de ML utilizadas para estimar la densidad del aire son las densidades del aire medias horarias y las direcciones medias horarias del viento de las estaciones de referencia. Una vez

estimadas las velocidades del viento y las densidades del aire del sitio objetivo se estiman las WTPOs haciendo uso de la curva de potencia de la WT que proporciona el fabricante (referida a la densidad del aire estándar) pero corregida (según se indica en el Apartado 2.8) para la densidades medias horarias del aire estimadas en el sitio objetivo.

$$\widehat{WTPO}_t = P(\widehat{V}_t, \widehat{\rho}_t) = P\left(f_v(V_{1f}, \dots, V_{kt}, \dots, V_{pt}, D_{1f}, \dots, D_{kt}, \dots, D_{pt})\right) \quad (4.13)$$

$$f_\rho(\rho_{1f}, \dots, \rho_{kt}, \dots, \rho_{pt}, D_{1f}, \dots, D_{kt}, \dots, D_{pt})$$

El modelo M4, representado en Ecuación 4.14, se diferencia fundamentalmente de los restantes modelos en que las tres técnicas de ML no se entrenan para minimizar el error en la predicción de la velocidad del viento o de la densidad de aire, sino que estas, utilizando las velocidades y direcciones del viento y las densidades del aire de los sitios de referencia como variables de entrada, se entrenan para minimizar el error de la estimación de las WTPOs.

$$\widehat{WTPO}_t = f_{WTPO}(V_{1f}, \dots, V_{kt}, \dots, V_{pt}, D_{1f}, \dots, D_{kt}, \dots, D_{pt}, \rho_{1f}, \dots, \rho_{kt}, \dots, \rho_{pt}) \quad (4.14)$$

El modelo M5, representado en Ecuación 4.16, está basado en la propuesta de Jung y Kwon [44] de utilizar modelos que estimen las velocidades del viento a largo plazo utilizando ANNs con funciones de error ponderadas a la hora de estimar la energía producida por una WT. Según los mencionados autores utilizar modelos como los M1 entrenados en base a una medida de error convencional pueden subestimar significativamente la producción anual de energía. Las velocidades del viento estimadas por el modelo M5 que es evaluado en este trabajo se centra en la ponderación de la función de error convencional aplicada a las ANNs. Es decir, en la ponderación de la medida de la raíz del error cuadrático medio (RMSE), Ecuación 3.18, aplicada a la técnica de ML propuesta por Jung y Kwon [44].

En Ecuación 3.18 o_i son los datos observados, \hat{e}_i las estimaciones logradas con la técnica ANN, n el número de observaciones del conjunto de datos evaluados y c_i son los pesos aplicados. Los pesos c_i son estimados haciendo uso de la curva de potencia de la WT que proporciona el fabricante referida a condiciones estándares ISO, $\rho_0 = 1.225 \text{ kg m}^{-3}$, (Figura 14), haciendo uso de la Ecuación 4.15, tal como proponen Jung y Kwon [44].

$$c_i = c(o_i) = \frac{1 + \xi \cdot Pr \cdot P(v(z)_i = o_i, \rho_0)}{1 + \xi \cdot Pr^2} \quad (4.15)$$

El factor $0 \leq \xi \leq 1$ en Ecuación 4.15 controla el grado de desactivación de las velocidades del viento que no producen toda la potencia nominal Pr . Si $\xi=0$, los errores no son ponderados y el modelo M5 coincide con el modelo M1. Según ξ se incrementa, las velocidades del viento que no dan lugar a que la WT trabaje a su potencia nominal Pr tendrán menos importancia ya que los errores se multiplicarán por pesos c_i decrecientes ($c_i < 1$).

$$\widehat{WTPO}_t = P(\widehat{V}_t, \rho_0) = P(f_v(V_{1f}, \dots, V_{kt}, \dots, V_{pt}, D_{1f}, \dots, D_{kt}, \dots, D_{pt}), \rho_0) \quad (4.16)$$

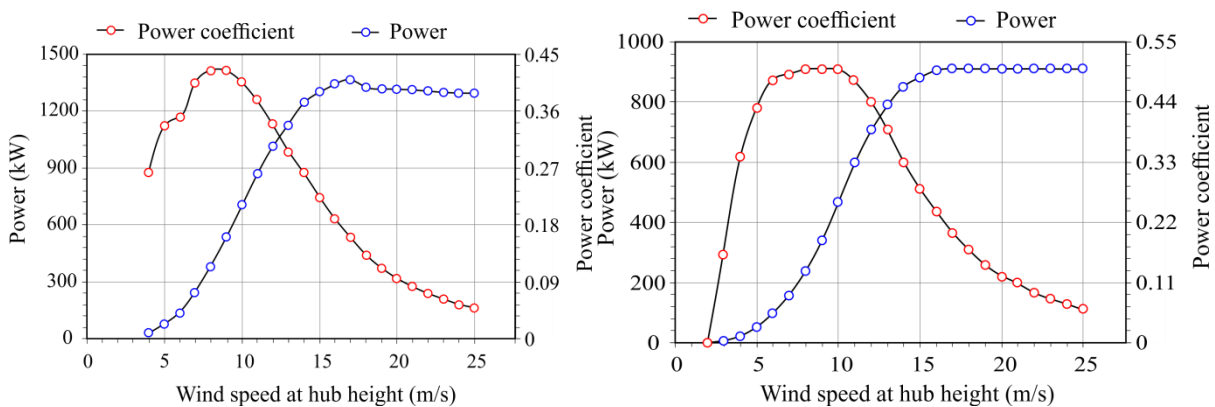


Figura 31 Curvas de potencia utilizadas en los análisis de WTPO. Modelos: WT-1 (Derecha) y WT-2 (Izquierda)

4.3.5.4 Hardware usado para el desarrollo de los cálculos

Para realizar las simulaciones requeridas en este trabajo se ha utilizado un ordenador equipado con dos procesadores Intel® Xeon E5-2630 con frecuencia base de 2.3 GHz y con 20 unidades de procesamiento independientes que permiten hasta 40 hilos de ejecución. Asimismo, dicho ordenador cuenta con 60 GB de memoria RAM y 7 TB de almacenamiento local. El sistema operativo utilizado en este equipo es Ubuntu 16.04 LTS (Long-Term Support).

4.3.6. Metodología

En los subapartados siguientes se relata la metodología utilizada en este trabajo.

4.3.6.1 Preámbulo

En la Figura 32 se esquematiza el procedimiento metodológico utilizado con el propósito de comparar los cinco modelos diseñados para estimar las WTPOs a largo plazo en un sitio objetivo, que han sido descritos en el Apartado 4.3.5.3. Como ha sido señalado en el abstract y la introducción de este paper, dichos modelos están basados en métodos MCP. Por tanto, la metodología esquematizada en la Figura 32 se apoya en el procedimiento normalmente empleado por dichos métodos [13] para llevar a cabo el entrenamiento y ensayo de las funciones que se proponen para establecer, en el periodo común de disposición de datos (denominado short-term)⁷, una relación entre los datos meteorológicos (velocidad, dirección del viento y densidad del aire) de los sitios de referencia y las variables del sitio objetivo. En este trabajo las series temporales de datos meteorológicos representativas del corto plazo han sido registradas durante el año 2014.

⁷Periodo en el cual las series temporales de datos de las variables de los sitios de referencia coinciden en longitud y fecha con la serie temporal de datos disponibles en el sitio considerado como objetivo.

Las series de datos de las variables velocidad del viento (V), dirección del viento (D) y densidad del aire (ρ) que se disponen durante el corto plazo en los sitios de referencia (WS-1, WS-2,..., WS-10), las cuales pueden ser seleccionadas para ser usadas como datos de entrada en los cinco diferentes modelos (M1,...,M5) analizados se indican en la parte superior izquierda de la Figura 32. Asimismo, se indican en dicha posición las series de datos de las variables (V , ρ y $WTPO$) de los sitios objetivo (WS-3, WS-4 y WS-5) que pretenden ser estimadas con las técnicas de ML propuestas (ANN, SVR y RF) para realizar la regresión. Es decir, aquellas variables que los modelos analizados utilizan para su entrenamiento.

En la Figura 32 la estación WS-5 se muestra como estación objetivo y las otras estaciones meteorológicas (WS-1,..., WS-4, WS-6,..., WS-10) como estaciones de referencia. Aunque hay que señalar que también, como se ha indicado anteriormente, se han simulado los casos en los que las estaciones WS-3 y WS-4 han representado a las estaciones objetivo y las nueve estaciones restantes han representado a las estaciones de referencia. La razón para seleccionar estas tres estaciones como estaciones objetivo es la existencia de coeficientes de correlación de Pearson superiores a 0.8 entre las velocidades de viento de la estación objetivo y al menos una de las estaciones de referencia. Dichos coeficientes de correlación, que se encuentran comprendidos entre 0.8 y 0.9 (Tabla 18), son catalogados en la literatura relacionada con los métodos MCP [13] como "buenos" indicadores de que el clima de una estación de referencia es representativo del clima del lugar objetivo, suposición clave que subyace todos los métodos MCP.

Hay que señalar que, antes de alimentar (Figura 32) a los cinco modelos (M1,...,M5) considerados en este estudio, las series de valores medios horarios de las velocidades del viento $v(z_r)_i$ y de las densidades del aire $\rho(z_r)_i$, medidas a una altura de referencia z_r en los sitios objetivos son extrapolados para cuantificar los valores $v(z)_i$ y $\rho(z)_i$ correspondientes a la hub height z de cada WT. Asimismo, para cada una de las WTs consideradas, cada valor medio horario de las series de $WTPO_i$, se ha determinado haciendo uso de la curva de potencia de la WT (Figura 14), de los valores medios horarios extrapolados de las series de densidad del aire, $\rho(z)_i$, y de los valores medios horarios extrapolados de las series de velocidades del viento $v(z)_i$.

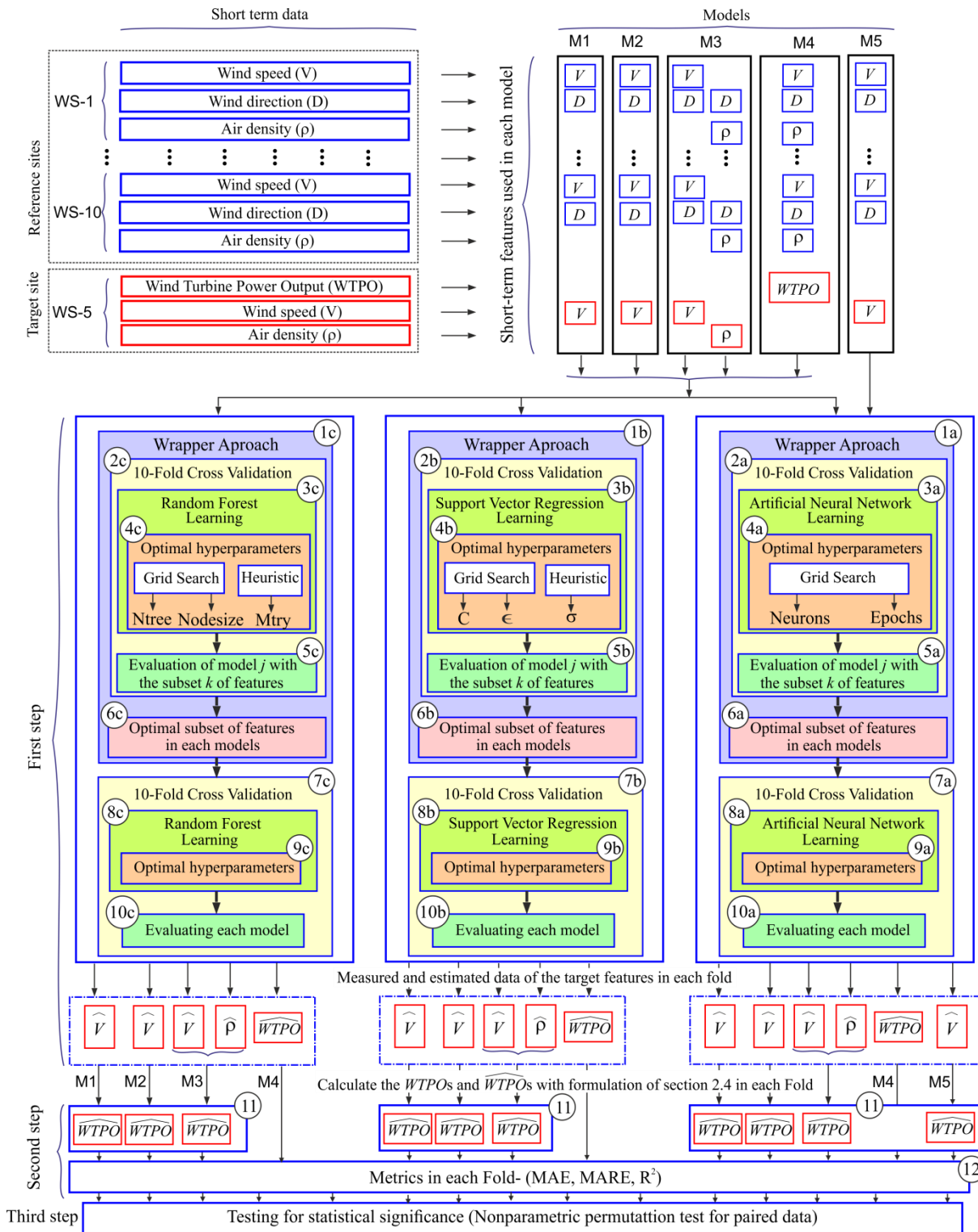


Figura 32 Procedimiento llevado a cabo para el análisis comparativo desarrollado en este estudio

WS	WS-1	WS-2	WS-3	WS-4	WS-5	WS-6	WS-7	WS-8	WS-9	WS-10
WS-1	1.00	0.73	0.74	0.67	0.69	0.54	0.26	0.56	0.52	0.56
WS-2	0.73	1.00	0.68	0.58	0.63	0.55	0.30	0.54	0.58	0.57
WS-3	0.74	0.68	1.00	0.82*	0.83	0.53	0.23	0.60	0.51	0.58
WS-4	0.67	0.58	0.82	1.00	0.83	0.51	0.18	0.55	0.45	0.50
WS-5	0.69	0.63	0.83	0.83	1.00	0.63	0.21	0.55	0.42	0.50
WS-6	0.54	0.55	0.53	0.51	0.63	1.00	0.39	0.36	0.31	0.40
WS-7	0.26	0.30	0.23	0.18	0.21	0.39	1.00	0.23	0.24	0.34
WS-8	0.56	0.54	0.60	0.55	0.55	0.36	0.23	1.00	0.56	0.55
WS-9	0.52	0.58	0.51	0.45	0.42	0.31	0.24	0.56	1.00	0.56
WS-10	0.56	0.57	0.58	0.50	0.50	0.40	0.34	0.55	0.56	1.00

Tabla 27 Coeficientes de correlación lineales entre velocidades del viento de las estaciones evaluadas

*En negrita los coeficientes de correlación de Pearson que son catalogados en la literatura relacionada con los métodos MCP [7] como "buenos" indicadores de que el clima de una estación de referencia es representativo del clima del lugar objetivo

4.3.6.1.1 Estimación del perfil vertical de viento

Se ha utilizado la ley logarítmica para el perfil vertical de la velocidad del viento, la cual es una de las más fundamentales ecuaciones de la metodología aplicada a la capa meteorológica límite [128]. Se describe la metodología de cálculo en el Apartado 2.5 de este documento.

Como valores de la longitud de rugosidad superficial se han asumido en este estudio los valores indicados en Corine Land Cover 2006 [225] de la Agencia europea del medioambiente. Desde la web [225] se descargó un archivo raster, el cual fue abierto con el software de licencia libre Quantum GIS [226]. Dicho archivo tiene una estructura de datos que permite la representación de mapas donde se informa de los usos del suelo y la longitud de rugosidad asociada. La información aportada por este mapa fue cruzada con la información disponible en el Sistema de Información de Ocupación del Suelo de España (Siglas en español: SIOSE) del Ministerio de Fomento del Gobierno de España [227], con el propósito de comprobar que los usos del suelo de los lugares analizados eran los mismos que los indicados en Corine [225]. Las longitudes de rugosidad indicadas en [106] para las áreas donde se ubican las estaciones WS-3, WS-4 y WS-5 son 0.15, 0.21 y 0.1 metros, respectivamente.

Hay que señalar que la Ecuación 2.5 realmente solo es aplicable para proyectar la velocidad del viento en el caso de que se produzcan condiciones térmicas neutras a nivel del suelo [228]. Sin embargo, como indica Brower [94] el efecto de la estabilidad térmica debe ser estimada determinando la Monin-Obukhov stability length [128], la cual debe ser obtenida de datos de temperatura a múltiples alturas y desafortunadamente estos no están disponibles en este trabajo. Por otro lado, Emeis [106] indica que para altas velocidades del viento, las cuales son las más favorables para la conversión de la energía eólica, la estratificación de la capa límite suele ser casi neutra. Además, como señalan Lackner et al. [229], el modelo simplificado es muy comúnmente utilizado en la práctica durante las campañas reales de evaluación de sitios en U.S. y ha sido utilizado

con frecuencia en los artículos publicados en revistas destacadas de energía [230-232] para ajustar la velocidad del viento a la altura de buje del aerogenerador desde la altura de medida de la misma.

4.3.6.1.2 Estimación de la densidad del aire a la altura de buje

Para la estimación de la densidad del aire a la altura del buje se ha usado la formulación expuesta en el Apartado 2.7 de este documento. La aplicación de dicha formulación es necesaria dado que los sensores de temperatura y humedad relativa están montados a 2 metros. Sin embargo, según especifica la IEC 61400-12-1 [87], los sensores para la medida de la temperatura del aire y la humedad deben estar montados, si se utilizan para los fines perseguidos en este trabajo, a una distancia de 10 m de la altura del buje de la WT, lo cual representaría la temperatura del aire en el centro del rotor de la misma. Asimismo, se puntualiza en la IEC 61400-12-1 [87] que el sensor de presión de aire debe montarse en el mástil meteorológico cerca de la altura del buje para representar la presión del aire en el centro del rotor de WT.

También es importante hacer constar que en esta extrapolación no se ha considerado la humedad relativa. Sin embargo, la IEC 61400-12-1 [87] solo recomienda que se considere dicha variable en el cálculo de la densidad del aire en el caso de altas temperaturas y las temperaturas medias de las estaciones meteorológicas consideradas como objetivos son inferiores a los 30°C. Finalmente, en el caso de la temperatura, se ha considerado que esta disminuye linealmente con la altura con un gradiente $\beta = -6.5 \text{ K km}^{-1}$.

4.3.6.1.3 Cálculo de las series de datos WTPO

Para el cálculo de las WTPO ha sido empleada la formulación expuesta en el Apartado 2.8 de la presente tesis doctoral, tanto para las versiones stall – regulated como pitch - regulated.

4.3.6.2 Descripción del procedimiento seguido

La metodología seguida para alcanzar los objetivos propuestos en este trabajo se desarrolla en tres pasos (Figura 32), en cada uno de los cuales se llevan a cabo una serie de tareas. En este contexto, puede observarse en Figura 32 que el primer paso comprende diez tareas, cuyo orden de prelación se indica con un número encerrado en un círculo. En el primer paso los números indicativos de las tareas se acompañan de una letra, con el propósito de señalar las diferencias existentes entre los procesos realizados con cada una de tres técnicas de ML utilizadas (ANN, SVR, RF). En los subapartados siguientes se detallan los tres mencionados pasos.

Conviene señalar que, aunque las series temporales de datos meteorológicos usados para los trabajos aquí desarrollados presentan una muy alta disponibilidad (superior al 98%), existen ocasiones en las que no existen datos para una hora determinada. En este sentido, previo a la ejecución del primer paso, se ha llevado a cabo un proceso de filtrado, eliminando del entrenamiento aquellas horas en las que se detecta celdas en blanco (*NaN*). De esta forma se asegura el funcionamiento del modelo sin fallos de ejecución.

4.3.6.2.1 Primer paso: Entrenamiento de los modelos y predicción de las series objetivo

Como ya ha sido señalado con anterioridad, el número de WSs de referencia disponibles y potencialmente utilizables por los cinco modelos considerados en este estudio es de nueve y el número de variables con que cada WS de referencia puede contribuir es de dos (V y D) o tres (V, D y ρ), Figura 32. Por tanto, el número total de variables que potencialmente pueden constituir variables de entrada de los modelos considerados se encuentra entre dieciocho y veintisiete. Un notable número de variables de entrada puede dar lugar a una sobre-especificación, la cual causa efectos negativos, tales como incremento del error de estimación, sobreajuste, etc. Dichos efectos, pueden afectar a la capacidad de generalización de los modelos cuando estos se alimenten con nuevos datos (no utilizados en el entrenamiento de los modelos) para llevar a cabo predicciones [45].

Con el objetivo de evitar dichos efectos negativos, en este estudio se utiliza una técnica wrapper para la selección de las features de cada modelo, cuyo proceso se representa en la Figura 32 con el número 1 encerrado en un círculo y acompañado con las letras a , b y c , para representar su aplicación cuando se usa la técnica ANN, SVR y RF, respectivamente. El algoritmo utilizado para seleccionar las variables de entrada óptimas de cada modelo se le conoce con el nombre de Recursive Feature Elimination (RFE) [185] (misma técnica empleada en el Capítulo anterior) y está disponible en el paquete Caret [185] del software R Statistics. Cuando la mejor combinación de variables es determinada (tareas $6a$, $6b$ y $6c$), se eliminan todas las features que producen un error mayor al obtenido con el subconjunto óptimo [184,186]. Además, la técnica wrapper se ha ejecutado utilizando la misma técnica de ML ($3a$, $3b$, $3c$) que se emplea para realizar la predicción ($8a$, $8b$, $8c$), como puede observarse en Figura 32. Para evaluar los modelos con el wrapper (tareas $2a, 2b, 2c$) y en los modelos de predicción que se alimentan con subconjunto óptimo de features ($7a, 7b, 7c$) se ha hecho uso de la técnica de validación cruzada con 10 folds (10-Folds Cross Validation), que ha sido empleada con anterioridad en la evaluación de métodos MCP [45,203]. Además, se han usados las mismas particiones aleatorias para que todos los modelos empleasen la misma unidad experimental y lograr de esta manera minimizar la varianza de la diferencia entre las métricas medias obtenidas por los cinco modelos.

Para determinar los hiperparámetros de las tres técnicas de ML empleadas se ha utilizado el método Grid Search [125] (Figura 32), tanto en las tareas 4 (a , b y c) como en las tareas 9 (a , b y c). Sin embargo, en el caso de las técnicas SVR (tareas $4b$ y $9b$) y RF (tareas $4c$ y $9c$) también se ha recurrido al uso de métodos heurísticos para determinar algunos parámetros. Concretamente, en el caso de la técnica SVR el parámetro σ ha sido seleccionado mediante el concurso de la función *Sigest()* del paquete kernlab [164] y en caso de la técnica RF el parámetro *Mtry* ha sido elegido haciendo uso de la *tuneR()* function implementada en el randomForest package [174].

Una vez consumada la tarea 10 (a , b , c) quedan definidos los modelos que para cada una de las WSs objetivo, tecnologías de WTs empleadas y técnicas de ML consideradas han generado los menores errores de muestra (utilizando los datos extrapolados a la altura de los bujes de las wind turbines que se poseen de las estaciones objetivo). Por tanto, una vez completado el paso 1 se dispone para cada uno de los modelos considerados, el subconjunto óptimo de features, los hiperparámetros

seleccionados y (señalado en un recuadro con línea de puntos y rayas, Figura 32) los valores estimados de las variables (\hat{V} , $\hat{\rho}$ y \widehat{WTPO}) en cada uno de los 10 folds.

4.3.6.2.2 Segundo paso: Determinación de las WTPOs y cálculo de las métricas

Este paso entraña dos tareas, que se señalizan en Figura 32 con los números 11 y 12 encerrados en un círculo. En la tarea 11, utilizando el conjunto de predicciones (para cada uno de los 10 folds usados para el ensayo de las validaciones cruzadas) por cada uno de los modelos, a excepción del modelo M4, se estiman los valores \widehat{WTPO} correspondientes mediante las Ecuación 4.11, Ecuación 4.12, Ecuación 4.13 y Ecuación 4.16.

En la tarea 12 se computan, para cada uno de los 10 folds, el Mean Absolute Error (MAE) Ecuación 3.19, el Mean Absolute Relative Error (MARE), Ecuación 3.20, y el coeficiente de determinación (R^2), Ecuación 3.21, entre los valores estimados \widehat{WTPO} con todos y cada uno de los modelos (y técnicas de ML implementadas en los mismos) y los valores calculados WTPOs para las dos WTs consideradas en los tres target site (WS-3, WS-4 y WS-5). Por tanto, para cada caso, se obtienen diez valores de cada una de las métricas. Las métricas que se utilizarán para las evaluaciones se calculan como la media aritmética de las diez métricas obtenidas de los 10 folds.

Conviene tener en cuenta que en el cálculo de las WTPO, existen determinadas ocasiones en las que la producción es nula. Dichos instantes en los que la velocidad y la WTPO son cero no se eliminan del proceso de entrenamiento dado que es una situación plausible en el normal funcionamiento de un parque eólico, posibilitando con ello que los modelos entrenados sean capaces de reconocer dichas situaciones. No obstante, con anterioridad al cálculo de las métricas MAE, MARE y R^2 se eliminan todos aquellos instantes en los que la producción es cero. Este proceso es necesario dado que de no hacerlo se producirían problemas de interminación k/0 (con $k \neq 0$) cuando se estima la métrica MARE.

4.3.6.2.3 Tercer paso: Comparación estadística de la bondad de los modelos

El tercer paso comprende una única tarea que consiste en la realización de un ensayo de hipótesis estadística. Con el mismo se persigue determinar si existe diferencia estadísticamente significativa entre las métricas obtenidas en la estimación de las WTPOs con los modelos evaluados.

Se desea contrastar la hipótesis nula (H_0) de que la métrica media (μ_i) de un modelo i es menor o igual, con un nivel de significación $\alpha = 5\%$, que la métrica media (μ_j) del modelo j al estimar las WTPOs, frente a una hipótesis alternativa (H_1) unilateral, en la que se acepta que la μ_i es significativamente mayor μ_j , Ecuación 4.13.

Dado que las muestras son pequeñas (las diez métricas de los diez folds de las cross-validation) y dado que en los modelos evaluados las métricas se han obtenido bajo similares condiciones de ensayo, se ha utilizado un test de permutación no paramétrico para datos apareados [215,216]. Como ya se discutía en el capítulo anterior, con el propósito de mitigar el riesgo de que se originen falsos positivos, los p – valores que se obtienen en cada par de comparaciones del test de permutación se ajustan siguiendo el procedimiento propuesto por Benjamini y Hochberg (BH) [217].

4.3.7. Análisis de resultados

En las Figura 33 – 35 se recogen los valores medios y las desviaciones estándar de las métricas MAE, MARE y R^2 obtenidas de la validación cruzada al estimar las WTPOs, con los cinco modelos analizados (con las tres técnicas de ML contempladas) y las dos WTs consideradas, en cada una de las tres WSs objetivo (WS-3, WS-4 y WS-5).

En la Tabla 28 se muestran las features de entrada que han sido descartadas del total disponible al aplicar el método wrapper a los distintos modelos considerados en este estudio. Puede observarse que las velocidades del viento y las densidades del aire son las features con menor número de descartes (lo que refleja la relevancia de dichas features) y que dichos descartes han ocurrido al utilizar las técnicas ANN y RF. Los descartes de las features de velocidad del viento han ocurrido cuando estas han presentado muy bajos coeficientes de correlación (Tabla 18) con respecto al target analizado. Las direcciones del viento comprenden el mayor porcentaje de features no seleccionadas. Asimismo, puede observarse en la Tabla 4 que es el modelo M4, cuando se utilizan las técnicas ANN y RF para estimar las WTPOs en el sitio objetivo WS-5, el que mayor número de features ha descartado.

En la Tabla 29 se muestran los hiperparámetros óptimos obtenidos con la técnica Grid Search y los métodos heurísticos (en su caso), indicados en la Figura 32, para cada uno de los modelos ejecutados basados en las tres técnicas de ML empleadas (ANN, SVR, RF). En la Tabla 30 se presentan los valores obtenidos del factor ξ , Ecuación 4.20, en cada uno de los diez folds de la cross-validation (cuando han actuado como subconjunto para estimar la métrica de muestra parcial y los nueve restantes como subconjunto de entrenamiento) de las validaciones cruzadas realizadas. Los valores del factor ξ se fueron variando entre 0 y 1 con incrementos de 0.1. Sin embargo, puede observarse que solo en el 25% de los casos dicho factor alcanzó el valor 0.2, mientras que en el 55% de los casos este no superó el valor 0.1. En el 20% de los casos $\xi=0$, por tanto los errores no fueron ponderados y por tanto, las métricas parciales del modelo M5 coincidieron con las métricas parciales del modelo M1. En la Tabla 31 se indican los tiempos medios de cálculo (en minutos) que han sido requeridos por los diferentes modelos analizados con las tres técnicas de ML empleadas. Puede observarse que los modelos M1, M2 y M3 han requerido los mismos tiempos medios y que el modelo M4 ha sido el más eficiente en términos computacionales, independientemente de las técnicas ML utilizadas para la ejecución del análisis. Sin embargo, el M5 ha precisado más del doble del tiempo que el M1. Puede comprobarse que la técnica RF es la que menor tiempo ha requerido y la técnica SVR la que más tiempo ha precisado, en cada modelo en las que han sido implementadas.

De la observación de las Figura 33 – 35 se concluye que el modelo M3 ha registrado las mejores métricas MAE, MARE y R^2 , independientemente de la estación objetivo, técnica de ML y tipo de WT consideradas. Asimismo, se observa que las técnicas SVR y RF han presentado mejores métricas que la técnica ANN (generalmente utilizada hasta la fecha [11,13,17,44,45]), independientemente del modelo utilizado, sitio objetivo y tipo de WT considerada.

La técnica RF ha proporcionado en el 65.28% de los casos mejores métricas que la técnica SVR. De acuerdo con los valores obtenidos de R^2 , solo cuando se ha hecho uso del modelo M3 con las técnicas SVR y RF se ha conseguido que entre el 92.78% y el 97.51% de la variabilidad total de WPTO sea explicada por el modelo utilizado. Los resultados obtenidos con el modelo M4 reflejan que a pesar de contar con una potente estructura no es capaz de gestionar eficientemente, con ninguna de las tres técnicas de ML utilizadas, la variabilidad de la densidad del aire.

Si se emplease el modelo M4 en lugar del modelo M3 se podría concluir, erróneamente, que la consideración de la variación de la densidad del aire no tiene una influencia relevante en la estimación de las WTPOs, o incluso, que dicha consideración puede tener una influencia negativa en la estimación de las WTPOs, dado que genera, en la mayoría de los casos, peores métricas que los modelos M1, M2 y M5, los cuales no tienen en cuenta la variabilidad de la densidad del aire. Al comparar los dos modelos (M3 y M4) que tienen en cuenta la variabilidad de la densidad del aire se puede deducir que para que se haga patente el efecto relevante de la variabilidad de la densidad del aire a la hora de estimar las WTPOs es importante la forma funcional en la que la densidad del aire, la velocidad del viento y la dirección intervienen en los modelos. Del estudio realizado se desprende que la forma funcional del modelo M3, Ecuación 4.13, es más apropiada que la del modelo 4, Ecuación 4.14.

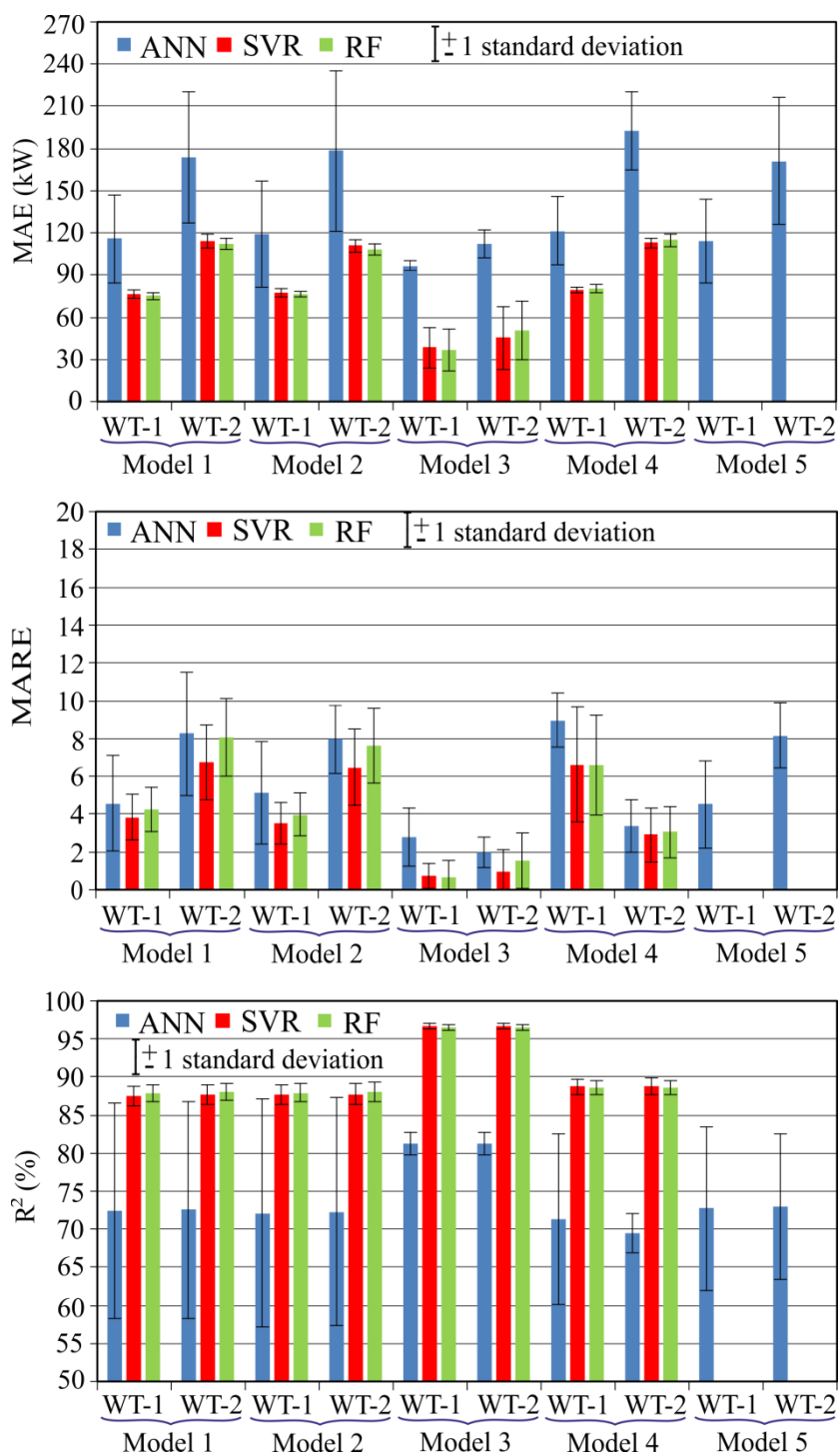


Figura 33 Valores medios y desviaciones estándar de las métricas MAE, MARE y R^2 obtenidos al aplicar 10 Folds Cross Validation cuando se estiman las WTPO con cinco modelos que usan tres técnicas ML para la estación WS-3

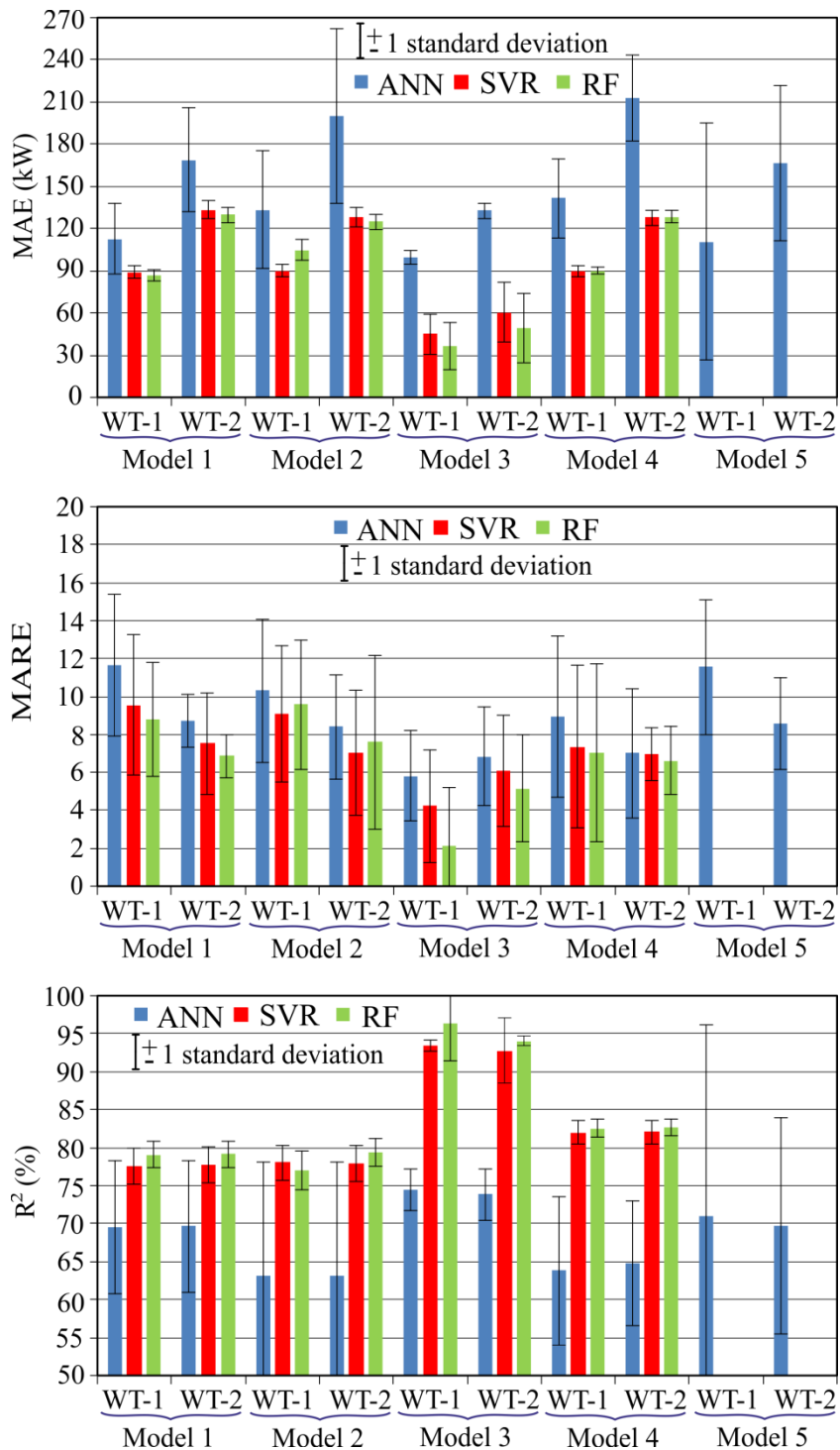


Figura 34 Valores medios y desviaciones estándar de las métricas MAE, MARE y R² obtenidos al aplicar 10 Folds Cross Validation cuando se estiman las WTPO con cinco modelos que usan tres técnicas ML para la estación WS-4

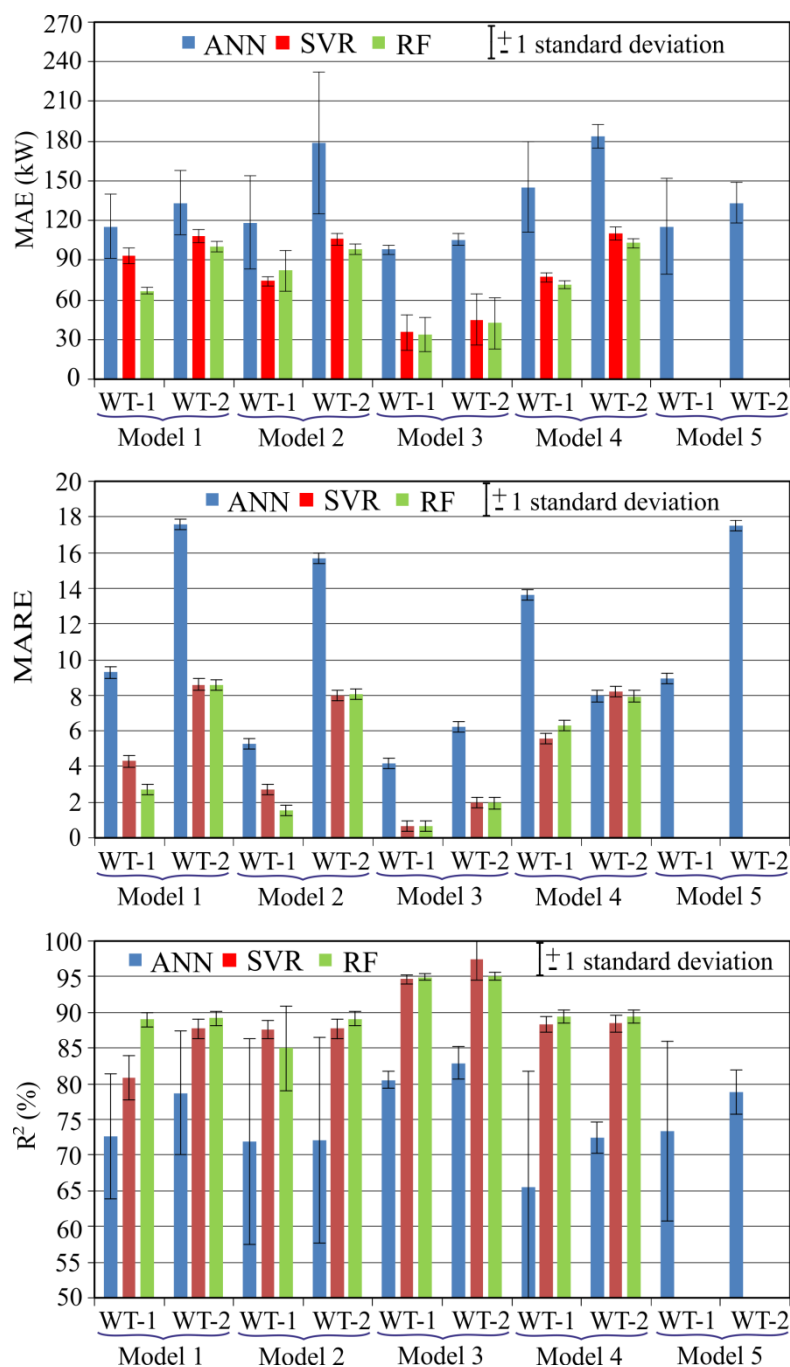


Figura 35 Valores medios y desviaciones estándar de las métricas MAE, MARE y R^2 obtenidos al aplicar 10 Folds Cross Validation cuando se estiman las WTPO con cinco modelos que usan tres técnicas ML para la estación WS-5

De la observación de las Figura 33 - 35 no puede señalarse, en los casos analizados, la existencia de diferencias relevantes entre las métricas obtenidas con los modelos M1 y M2 que permitan señalar a uno de ellos como mejor modelo. Cuando se ha estudiado las WTPOs con la WT-1 el modelo M1 ha presentado mejores métricas que el M2 en el 55.55% de los casos, sin embargo, cuando se ha considerado la WT-2, el modelo M2 ha generado mejores métricas en el 66.66% de los casos.

La no observancia de diferencias notables entre las métricas obtenidas con el modelo M2 y las obtenidas con el modelo M1 puede ser consecuencia de que los tres sitios objetivos considerados en este estudio se encuentran ubicados a nivel del mar y no se producen las diferencias notorias entre las densidades medias del aire de los sitios objetivos (1.194 kg m^{-3} , 1.196 kg m^{-3} y 1.199 kg m^{-3} , en WS-3, WS-4 y WS-5, respectivamente) a la altura del buje de las WTs consideradas y la densidad del aire estándar de 1.225 kg m^{-3} , que según Hau [43] se detectan entre sitios ubicados en alturas que se diferencian entre sí unos cientos de metros.

En el caso de las estaciones objetivo consideradas las frecuencias de las mencionadas densidades medias del aire y de la densidad del aire estándar han sido muy bajas. A título de ejemplo, se muestran en la Figura 36 las frecuencias de dichas densidades (9.3% y 6.2% para la densidad del aire media y la densidad del aire estándar, respectivamente) en función del rango de velocidades de operación de las WTs consideradas en la estación WS-5. Las densidades del aire en WS-5 se encuentran en el intervalo (1.124 kg m^{-3} y 1.263 kg m^{-3}).

De los resultados del estudio realizado se concluye que el modelo M5 (basado en la propuesta de Jung y Kwon [233], pero utilizando una técnica wrapper para la selección de las features y la técnica de 10-Folds Cross Validation para la evaluación), ha generado métricas ligeramente mejores que las proporcionadas por el M1, aunque con requerimientos de tiempos de cálculo muy superiores, como ya ha sido señalado anteriormente.

Con el objetivo de comprobar si existen diferencias significativas entre las técnicas de ML empleadas en el modelo M3, que de acuerdo con las Figura 33 – 35 ha resultado ser el más adecuado para estimar las WTPOs, se ha llevado a cabo un test de hipótesis estadística. En las Tabla 32 – 34 se muestran los resultados alcanzados. La primera columna de dichas tablas indica la estación objetivo considerada (WS-3, WS-4 y WS-5). Para cada estación objetivo existen dos sub-tablas dentro de cada tabla, es decir, una sub-tabla para cada tipo de WT (WT-1 y WT-2).

En cada sub-tabla de la Tabla 32 y Tabla 33 (las cuales recogen los p-valores correspondientes a las métricas MAE y MARE, respectivamente), las técnicas de ML "i" representan las columnas y las técnicas de ML "j" las filas, Ecuación 4.13. Si la lectura de las tablas se realiza por filas, las celdas con los p-valores escritos en negrita corresponden a aquellos casos donde la hipótesis nula ($H_0: \mu_i \leq \mu_j$) es rechazada a favor de la hipótesis alternativa ($H_1: \mu_i > \mu_j$). En la Tabla 34 (que muestra los p-valores asociados a la métrica R^2), los modelos "i" representan las filas y los modelos "j" las columnas. En este caso, la técnica de mejor significancia "i" será aquella que tenga un mayor número de celdas de la fila con valores escritos en negrita, ya que dicha técnica "i" presentará una métrica R^2 significativamente mayor que las correspondientes a las técnicas "j".

Los p-valores indicados en las Tabla 32 y Tabla 34 conducen a rechazar la hipótesis nula y a aceptar la hipótesis alternativa de que las métricas MAE y R^2 obtenidas con las técnicas SVR y RF son significativamente (nivel de significancia 5%) mejores que las obtenidas con la técnica ANN en todos los casos considerados (WSs y WTs objetivos). Se observa en las Tabla 32 – 34 que no se han

presentado diferencias estadísticamente significativas entre las métricas obtenidas con las técnicas SVR y RF.

De acuerdo con los p-valores mostrados en la Tabla 33, cuando se ha considerado la WT-1, en el 100% de los casos, las métricas MARE obtenidas con la técnica RF fueron significativamente mejores que las obtenidas con la técnica ANN. Sin embargo, en el caso de la WT-2 y para la WS-4 no se presentaron diferencias estadísticamente significativas entre las MARE obtenidas con las tres técnicas de ML consideradas. Por tanto, aunque la técnica ANN ha sido generalmente utilizada hasta la fecha [11,13,17,44,45] como técnica de regresión implementada en los métodos MCP, los resultados aquí alcanzados nos conducen a proponer la utilización de las técnicas SVR y RF, ya que se pueden lograr mejores métricas al estimar las WTPOs en un sitio objetivo, con tiempos de cálculo aceptables.

Asimismo, con el propósito de comprobar si las diferencias entre las tres métricas generadas por el M3, que tiene en cuenta la variabilidad de la densidad del aire, y las generadas por el M1, que obvia dicha variabilidad y ha sido el más frecuentemente utilizado en la bibliografía [16,17,26], se llevó a cabo un test de hipótesis estadística. Los resultados, que se muestran en la Tabla 35, los cuales se centran, a título de ejemplo, en las métricas obtenidas con las técnicas SVR, permiten concluir que en todos los casos simulados (con excepción de la métrica MARE en el caso de la WT-2 en la WS-4) las métricas generadas por el M3 son significativamente (nivel de significancia 5%) mejores que las obtenidas por el M1. Este resultado refleja la importancia de la consideración de la variabilidad de la densidad del aire a la hora de estimar las WTPOs en un sitio objetivo.

Como ya ha sido señalado con anterioridad, en las WTs con control de paso de pala la variación en la densidad del aire no tiene influencia en la variación de la potencia eléctrica generada por las mismas cuando estas trabajan en el rango de carga completa [43,87]. Sin embargo, como se observa en Tabla 35 las métricas obtenidas para el caso de la WT-1 con el M3 han resultado significativamente mejores que las obtenidas con M1. Este hecho evidencia que aunque en las WTs con control de paso de pala la influencia de la variación de la densidad del aire es menos severa [43] que en los aerogeneradores stall-regulated, dicha influencia no debe ser a priori despreciada.

Target	Features	Modelo 1 - Modelo 2			Modelo 3						Modelo 4						Modelo 5	
		Velocidad			Velocidad			Densidad del aire			Pitch			Stall			Pitch	Stall
		ANN	SVR	RF	ANN	SVR	RF	ANN	SVR	RF	ANN	SVR	RF	ANN	SVR	RF	ANN	ANN
WS3	Features seleccionados	17/18	18/18	18/18	17/18	18/18	18/18	15/18	17/18	17/18	27/27	27/27	27/27	22/27	27/27	27/27	18/18	17/18
	Features no seleccionados	WS-7 D	-	-	WS-7 D	-	-	WS-9 D	WS-7 D	WS-7 D	-	-	-	WS-1 D	-	-	-	WS-9 D
		-	-	-	-	-	-	WS-10 D	-	-	-	-	-	WS-2 D	-	-	-	-
		-	-	-	-	-	-	WS-8 D	-	-	-	-	-	WS-10 S	-	-	-	-
		-	-	-	-	-	-	-	-	-	-	-	-	WS-8 D	-	-	-	-
-	-	-	-	-	-	-	-	-	-	-	-	WS-6 D	-	-	-	-		
WS4	Features seleccionados	15/18	17/18	18/18	15/18	17/18	18/18	15/18	17/18	17/18	26/27	27/27	26/27	24/27	27/27	26/27	17/18	17/18
	Features no seleccionados	WS-6 D	WS-3 D	-	WS-6 D	WS-3 D	-	WS-9 D	WS-6 D	WS-6 D	WS-6 D	-	WS-7 D	WS-1 D	-	WS-7 D	WS-7 S	WS-6 D
		WS-7 D	-	-	WS-7 D	-	-	WS-10 D	-	-	-	-	-	WS-2 D	-	-	-	-
		WS-8 D	-	-	WS-8 D	-	-	WS-3 D	-	-	-	-	-	WS-7 D	-	-	-	-
WS5	Features seleccionados	17/18	18/18	15/18	17/18	18/18	18/18	15/18	18/18	17/18	25/27	27/27	19/27	20/27	27/27	19/27	18/18	17/18
	Features no seleccionados	WS-7 D	-	WS-7 D	WS-7 D	-	-	WS-4 D	WS-3 D	WS-6 D	WS-2 D	-	WS-1 D	WS-1 D	-	WS-1 D	-	WS-7 D
		-	-	WS-8 D	-	-	-	WS-7 D	-	-	WS-8 p	-	WS-3 D	WS-2 p	-	WS-3 D	-	-
		-	-	WS-10 S	-	-	-	WS-9 D	-	-	-	-	WS-4 p	WS-8 S	-	WS-6 p	-	-
		-	-	-	-	-	-	-	-	-	-	-	WS-6 D	WS-7 D	-	WS-7 D	-	-
		-	-	-	-	-	-	-	-	-	-	-	WS-7 S	WS-7 S	-	WS-7 S	-	-
		-	-	-	-	-	-	-	-	-	-	-	WS-7 D	WS-6 D	-	WS-8 D	-	-
		-	-	-	-	-	-	-	-	-	-	-	WS-8 D	WS-10 D	-	WS-10 D	-	-
-	-	-	-	-	-	-	-	-	-	-	WS-10 D	-	-	WS-10 S	-	-		

Tabla 28 Features no seleccionados en el procedimiento Wrapper

Weather Station	Técnica	Hiper-parametro	Modelo 1 - 2	Modelo 3		Modelo 4		Modelo 5	
			Veloc.	Veloc.	Densid.	WT-1	WT-2	WT-1	WT-2
WS3	ANN	Neurons	30	30	25	40	24	30	30
		Epochs	600	600	600	1000	1000	600	600
	SVR	C	9	9	3	13	13	-	-
		ϵ	0.10	0.10	0.01	0.10	0.10	-	-
		σ	0.040	0.040	0.003	0.027	0.028	-	-
	RF	Ntree	500	500	500	1000	1000	-	-
		Nodesize	4	4	1	5	5	-	-
Mtry		6	6	3	9	9	-	-	
WS4	ANN	Neurons	28	28	25	39	51	40	40
		Epochs	700	700	600	1000	1000	600	600
	SVR	C	13	13	3	13	13	-	-
		ϵ	0.10	0.10	0.01	0.10	0.10	-	-
		σ	0.041	0.041	0.004	0.026	0.026	-	-
	RF	Ntree	800	800	500	1000	1000	-	-
		Nodesize	5	5	1	5	5	-	-
Mtry		6	6	3	8	8	-	-	
WS5	ANN	Neurons	39	39	25	52	30	40	40
		Epochs	600	600	600	1000	1000	600	600
	SVR	C	13	13	3	14	14	-	-
		ϵ	0.10	0.10	0.01	0.10	0.10	-	-
		σ	0.039	0.039	0.003	0.026	0.027	-	-
	RF	Ntree	500	500	500	1000	1000	-	-
		Nodesize	5	5	1	5	5	-	-
Mtry		6	6	3	8	8	-	-	

Tabla 29 Hiperparámetros

Weather Station	Turbina eólica	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Fold 6	Fold 7	Fold 8	Fold 9	Fold 10
WS3	WT-1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1
	WT-2	0.2	0.2	0.1	0.2	0.2	0.2	0.2	0.2	0.1	0.1
WS4	WT-1	0.2	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1
	WT-2	0.2	0.2	0.2	0.2	0.1	0.1	0.2	0.2	0.2	0.1
WS5	WT-1	0.1	0	0	0	0.1	0.1	0	0	0	0.1
	WT-2	0.1	0	0	0	0.1	0.1	0	0	0	0.1

Tabla 30 Factores ξ seleccionados

Técnica ML	Modelo 1	Modelo 2	Modelo 3	Modelo 4	Modelo 5
ANN	1478	1478	1478	1140	3235
SVR	2200	2200	2200	1699	-
RF	1337	1337	1337	1031	-

Tabla 31 Tiempos medios de cómputo

Objetivo	Turbina	ML	WT-1				Turbina	ML	WT-2			
			i	ANN	SVR	RF			i	ANN	SVR	RF
WS-3	WT-1	j	media	96.69	38.23	36.72	WT-2	j	media	112.35	45.28	50.60
		ANN	96.69	-	0.000	0.000		ANN	112.35	-	0.000	0.000
		SVR	38.23	0.000	-	0.275		SVR	45.28	0.000	-	0.274
		RF	36.72	0.000	0.275	-		RF	50.60	0.000	0.274	-
WS-4	WT-1	j	media	99.39	45.01	36.51	WT-2	j	media	132.82	60.52	49.44
		ANN	99.39	-	0.000	0.000		ANN	132.82	-	0.000	0.000
		SVR	45.01	0.000	-	0.236		SVR	60.52	0.000	-	0.273
		RF	36.51	0.000	0.236	-		RF	49.44	0.000	0.273	-
WS-5	WT-1	j	media	97.95	35.24	33.39	WT-2	j	media	105.63	45.15	42.29
		ANN	97.95	-	0.000	0.000		ANN	105.63	-	0.000	0.000
		SVR	35.24	0.000	-	0.275		SVR	45.15	0.000	-	0.276
		RF	33.39	0.000	0.275	-		RF	42.29	0.000	0.276	-

Tabla 32 Análisis de diferencias estadísticamente significativas para la métrica MAE (kW). P-valores

Objetivo	Turbina	ML	WT-1				Turbina	ML	WT-2			
			i	ANN	SVR	RF			i	ANN	SVR	RF
WS-3	WT-1	j	media	2.78	0.74	0.69	WT-2	j	media	1.98	0.92	1.55
		ANN	2.78	-	0.000	0.001		ANN	1.98	-	0.027	0.275
		SVR	0.74	0.000	-	0.406		SVR	0.92	0.027	-	0.273
		RF	0.69	0.001	0.406	-		RF	1.55	0.275	0.273	-
WS-4	WT-1	j	media	5.81	4.21	2.15	WT-2	j	media	6.83	6.07	5.16
		ANN	5.81	-	0.143	0.007		ANN	6.83	-	0.299	0.137
		SVR	4.21	0.143	-	0.115		SVR	6.07	0.299	-	0.277
		RF	2.15	0.007	0.115	-		RF	5.16	0.137	0.277	-
WS-5	WT-1	j	media	4.20	0.64	0.65	WT-2	j	media	6.25	1.98	1.94
		ANN	4.20	-	0.000	0.000		ANN	6.25	-	0.002	0.016
		SVR	0.64	0.000	-	0.471		SVR	1.98	0.002	-	0.471
		RF	0.65	0.000	0.471	-		RF	1.94	0.016	0.471	-

Tabla 33 Análisis de diferencias estadísticamente significativas para la métrica MARE. P-valores

* Para cada posición objetivo y tipo de turbina, si se lee por filas (j) la mejor técnica ML es aquella con más celdas mostradas en negrita.

Objetivo	Turbina	ML	WT-1				Objetivo	ML	WT-2			
			j	ANN	SVR	RF			j	ANN	SVR	RF
WS-3	WT-1	i	media	81.21	96.67	96.56	WT-2	i	media	81.21	96.67	96.58
		ANN	81.21	-	0.000	0.000		ANN	81.21	-	0.000	0.000
		SVR	96.67	0.000	-	0.275		SVR	96.67	0.000	-	0.299
		RF	96.56	0.000	0.275	-		RF	96.58	0.000	0.299	-
WS-4	WT-1	i	media	74.40	93.45	96.36	WT-2	i	media	73.82	92.78	94.00
		ANN	74.40	-	0.000	0.000		ANN	73.82	-	0.000	0.000
		SVR	93.45	0.000	-	0.045		SVR	92.78	0.000	-	0.327
		RF	96.36	0.000	0.045	-		RF	94.00	0.000	0.327	-
WS-5	WT-1	i	media	80.50	94.65	94.97	WT-2	i	media	82.93	97.51	95.15
		ANN	80.50	-	0.000	0.000		ANN	82.93	-	0.000	0.000
		SVR	94.65	0.000	-	0.158		SVR	97.51	0.000	-	0.011
		RF	94.97	0.000	0.158	-		RF	95.15	0.000	0.011	-

Tabla 34 Análisis de diferencias estadísticamente significativas para la métrica R² (%). P-valores

** Para cada posición objetivo y tipo de turbina, si se lee por filas (i) la mejor técnica ML es aquella con más celdas mostradas en negrita.

Es decir, no se debería desestimar que en el intervalo de carga parcial, la curva de potencia de las WT con blade pitch control varía con la densidad del aire como lo hace la curva de potencia de los aerogeneradores stall-regulated. Por tanto, el porcentaje de operación de la WT en dicho rango de carga puede ser un factor a tener en cuenta.

Como se muestra a título de ejemplo en Figura 36 para el caso de la WS-5, el porcentaje de operación a plena carga de las turbinas eólicas consideradas es de un 15.24%, sin embargo el porcentaje de operación de la WT-1 en el intervalo de carga parcial es de un 84.16%. Dichos porcentajes de operación han sido obtenidos haciendo uso de una distribución mixta Normal-Weibull truncada [62,63], cuyos parámetros han sido obtenidos con las velocidades del viento registrados en la WS-5 después de haber sido estos extrapolados a la altura del buje de las turbinas eólicas.

Finalmente, teniendo en cuenta que la técnica 10-Folds Cross Validation ha sido aplicada y que, por tanto, todos los datos de corto plazo usados para el entrenamiento también son usados para el testeo pero nunca en el mismo fold, se presenta a modo de ejemplo en la Figura 37 las WTPOs medias mensuales y horarias observadas, así como las estimaciones que se obtendrían en la estación WS3 con los M1 y M3 cuando la turbina eólica seleccionada es WT1 y la técnica ML usada es SVM. Se puede observar que a modo cualitativo se reafirman las conclusiones obtenidas con el análisis de significancia estadística, existiendo, en términos generales, un mayor ajuste del modelo que

considera la variación horaria de la densidad del aire en contraposición de la metodología más usada hasta el momento por la industria. La última grafica representada en la Figura 37 muestra dicha comparativa entre WTPOs para un periodo concreto seleccionado al azar del conjunto de testeo.

Objetivo	Turbina	Métrica	Modelo	WT-1				Turbina	Métrica	Modelo	WT-2					
				i	M1	M3					i	M1	M3			
WS-3	WT-1	MAE (kW)	j	media	76.17	38.23		WT-2	MAE (kW)	j	media	114.19	45.28			
			M1		76.17	-				0.000	M1		114.19	-	0.000	
			M3		38.23	0.000				-	M3		45.28	0.000	-	
		MARE	j	media	3.84	0.74				MARE	j	media	6.72	0.92		
			M1		3.84	-					0.000	M1		6.72	-	0.000
			M3		0.74	0.000					-	M3		0.92	0.000	-
	R ² (%)	j	media	87.53	96.67		R ² (%)	j	media	87.70	96.67					
		M1		87.53	-			0.000	M1		87.70	-	0.000			
		M3		96.67	0.000			-	M3		96.67	0.000	-			
WS-4	WT-1	MAE (kW)	j	media	88.84	45.01		WT-2	MAE (kW)	j	media	133.22	60.52			
			M1		88.84	-				0.000	M1		133.22	-	0.000	
			M3		45.01	0.000				-	M3		60.52	0.000	-	
		MARE	j	media	9.55	4.21				MARE	j	media	7.51	6.07		
			M1		9.55	-					0.001	M1		7.51	-	0.135
			M3		4.21	0.001					-	M3		6.07	0.135	-
	R ² (%)	j	media	77.61	93.45		R ² (%)	j	media	77.73	92.78					
		M1		77.61	-			0.000	M1		77.73	-	0.001			
		M3		93.45	0.000			-	M3		92.78	0.001	-			
WS-5	WT-1	MAE (kW)	j	media	93.50	35.24		WT-2	MAE (kW)	j	media	108.23	45.15			
			M1		93.50	-				0.000	M1		108.23	-	0.000	
			M3		35.24	0.000				-	M3		45.15	0.000	-	
		MARE	j	media	4.29	0.64				MARE	j	media	8.60	1.98		
			M1		4.29	-					0.000	M1		8.60	-	0.000
			M3		0.64	0.000					-	M3		1.98	0.000	-
	R ² (%)	j	media	80.88	94.65		R ² (%)	j	media	87.75	97.51					
		M1		80.88	-			0.000	M1		87.75	-	0.000			
		M3		94.65	0.000			-	M3		97.51	0.000	-			

Tabla 35 Análisis de diferencias estadísticamente significativas entre M1 y M3. P-valores

* Para cada posición objetivo y tipo de turbina, si se lee por filas (j) la mejor técnica ML es aquella con más celdas mostradas en negrita.

WSP: Percentage for which wind is within a certain speed interval
 WT-1 ($\rho: 1.263 \text{ kg/m}^3$) \circ WT-1 power curve ($\rho: 1.225 \text{ kg/m}^3$)
 WT-1 ($\rho: 1.124 \text{ kg/m}^3$) WT-2 ($\rho: 1.263 \text{ kg/m}^3$)
 \circ WT-2 power curve ($\rho: 1.225 \text{ kg/m}^3$) WT-2 ($\rho: 1.124 \text{ kg/m}^3$)
 — Wind probability density function

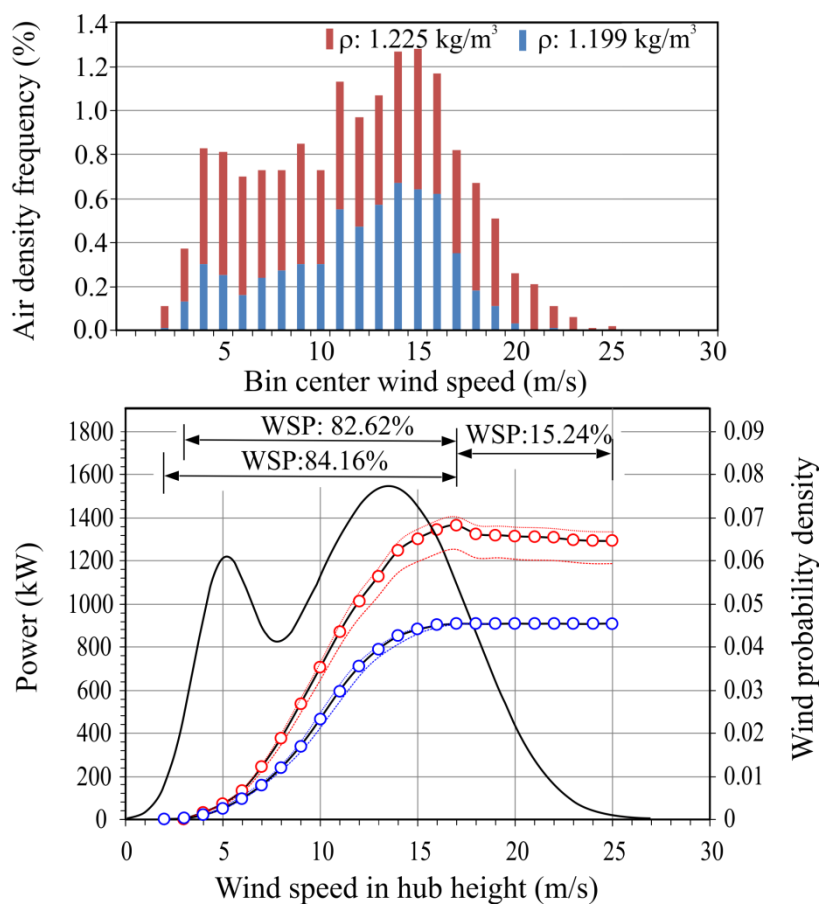


Figura 36 Frecuencias de la densidad del aire media y desviación estándar para el rango de velocidades de viento en operación en la estación WS-5

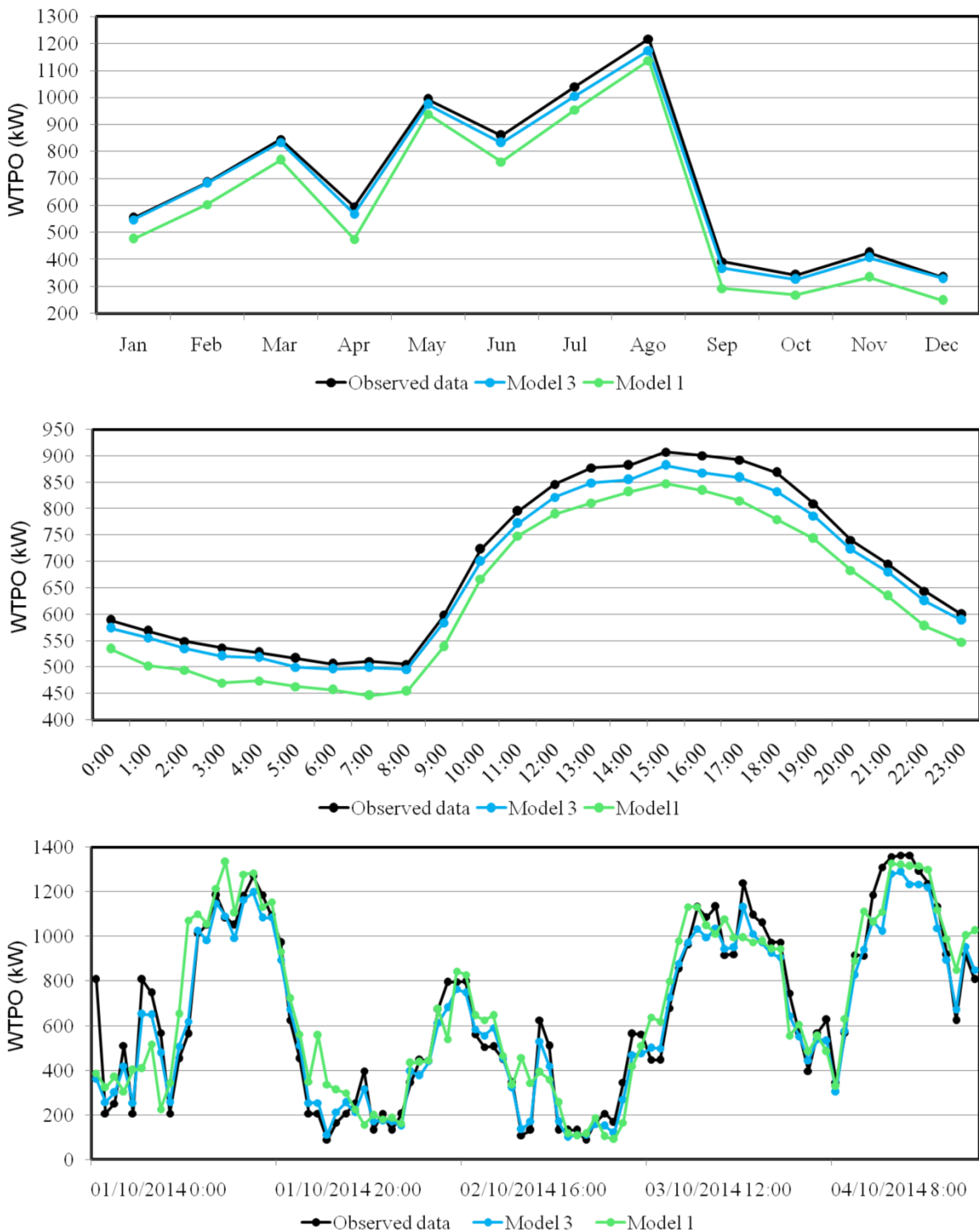


Figura 37 Comparativa entre WTPOs (datos observados, M1 y M3) medias mensuales y horarias del conjunto de testeo en corto plazo (2014) y evolución horaria para un periodo de 4 días seleccionado de manera arbitraria.

4.3.8. Conclusiones

En el trabajo desarrollado en este apartado se ha evaluado el comportamiento de modelos MCP, utilizando múltiples estaciones de referencia, que han sido usados hasta la fecha y modelos propuestos aquí, para estimar las WTPO a largo plazo en un sitio objetivo. Los modelos propuestos hasta la fecha, han utilizado fundamentalmente como técnica de regresión ANN, se han centrado en WT con control de pala pitch y parten de la hipótesis de que es poco relevante la influencia de la variación de la densidad del aire. Dichos modelos han asumido un valor constante de la densidad del aire de 1.225 kg m^{-3} , el cual corresponde a condiciones de atmósfera estándar.

Se han propuesto modelos que tienen en cuenta la variabilidad de la densidad del aire, contemplando el uso tanto de turbinas eólicas con control pitch-regulated como stall-regulated. Además, se ha evaluado el comportamiento en cuatro de los cinco modelos considerando tres técnicas de ML diferentes para llevar a cabo la regresión (ANN, SVR y RF).

Entre las conclusiones extraídas al aplicar los modelos y técnicas de regresión utilizadas en los mismos a las medias horarias de velocidad, dirección del viento y densidad del aire recabada en 2014 para 10 estaciones anemométricas instaladas en el archipiélago canario, pueden señalarse:

- a) Para que se ponga de manifiesto el efecto relevante de la variabilidad de la densidad del aire a la hora de estimar con métodos MCP las WTPOs en un sitio objetivo, tanto si se usan aerogeneradores stall-regulated WT como pitch-regulated, es importante tener en cuenta la forma funcional en la que intervienen las muestras de datos de velocidad, dirección del viento y densidad del aire.
- b) De los cinco modelos MCP evaluados, el modelo M3, que estima la velocidad del viento y la densidad del aire de forma independiente a largo plazo con anterioridad a la predicción de las WTPOs siempre arrojan las mejores métricas MAE, MARE and R^2 , independientemente de la estación objetivo (WS-3, WS-4 y WS-5), la técnica de ML (ANN, SVR, RF) y el tipo de turbina eólica consideradas (WT-1: WT con blade pitch control y WT-2: stall-regulated WT). Además, del análisis de significancia estadística realizado con el propósito de comprobar la existencia de diferencias, usando la técnica SVR, entre las tres métricas generadas por el M3 (que tiene en cuenta la variabilidad de la densidad del aire), y las generadas por el M1 (que obvia dicha variabilidad y ha sido el más frecuentemente utilizado en la bibliografía [16,17,26]), se llegó a la conclusión que en todos los casos simulados (con excepción de la métrica MARE en el caso de la WT-2 en la WS-4) las métricas generadas por el M3 son significativamente (nivel de significancia del 5%) mejores que las obtenidas por el M1. Este resultado refleja la importancia de la consideración de la variabilidad de las densidades del aire a la hora de estimar las WTPOs en un sitio objetivo.
- c) Los modelos que han utilizado SVRs o RFs han proporcionado siempre mejores métricas que los modelos que han usado redes neuronales artificiales, con diferencias estadísticamente significativas (nivel de significancia del 5%) para la mayoría de los casos evaluados.

- d) No se han encontrado diferencias estadísticamente significativas entre las métricas obtenidas con los modelos basados en SVRs y las obtenidas con los modelos basados en RFs.
- e) La no observancia de diferencias notables entre las métricas obtenidas con el modelo M2 y las obtenidas con el modelo M1 puede ser consecuencia de que los tres sitios objetivos considerados en este estudio se encuentran ubicados a nivel del mar y no se producen las diferencias notorias entre las densidades medias del aire de los sitios objetivos (1.194 kg m^{-3} , 1.196 kg m^{-3} y 1.199 kg m^{-3} , en WS-3, WS-4 y WS-5, respectivamente) a la altura del buje de las turbinas eólicas consideradas y la densidad del aire estándar de 1.225 kg m^{-3} .
- f) El modelo M4 se diferencia fundamentalmente de los restantes modelos en que las tres técnicas de ML no se entrenan para minimizar el error en la predicción de la velocidad del viento o de la densidad de aire, sino que se entrenan para minimizar el error en la estimación de las WTPOs. Los resultados obtenidos indican que el modelo M4, a pesar de contar con una potente estructura (que utiliza las velocidades y direcciones del viento y las densidades del aire de los sitios de referencia como variables de entrada y las WTPOs como variable de salida) no es capaz de gestionar eficientemente, con ninguna de las tres técnicas de ML utilizadas, la variabilidad de la densidad del aire. Sin embargo, el modelo M4 ha sido el más eficiente en términos computacionales, independientemente de las técnicas ML utilizadas para la ejecución del análisis.
- g) El modelo M5, basado en la propuesta de Jung y Kwon [233] de utilizar modelos que estimen las velocidades del viento a largo plazo utilizando ANNs con funciones de error ponderadas a la hora de estimar la energía producida por una WT ha sido evaluado con la metodología general aplicada a los restantes modelos aquí considerados. Es decir, se ha utilizado una técnica wrapper para la selección de las features del dicho modelo y la evaluación se ha llevado a cabo con 10-Folds Cross Validation. De los resultados del estudio realizado se concluye que el modelo M5 ha generado métricas ligeramente mejores que las proporcionadas por el M1, sin embargo, ha precisado más del doble del tiempo de computo que el M1.

Conclusiones y líneas futuras de investigación

5.1. Introducción

Tras los estudios desarrollados en la tesis doctoral y de acuerdo con los resultados obtenidos, en este último capítulo se recopilan inicialmente las principales contribuciones que ha supuesto este trabajo al conocimiento de la industria eólica y, en especial, a la aplicación de técnicas de Machine Learning para la estimación de la potencia eólica, la cual ha sido centrada en el estudio a largo plazo tal como se ha analizado en los Apartados 4.2 y 4.3 de este trabajo.

Seguidamente, en el Apartado 5.3 se relacionan las principales conclusiones extraídas del trabajo de investigación, para acabar en el Apartado 5.4 con una relación de las principales líneas de investigación detectadas fruto del trabajo realizado.

5.2. Contribución de la tesis al conocimiento de la industria eólica

Se relaciona en esta sección aquellos aspectos evaluados en la presente tesis doctoral que han supuesto una contribución relevante al conocimiento de la industria eólica.

1. Frente a la tendencia común de sólo considerar en la estimación de la potencia eólica las medidas de velocidad y dirección del viento [14,17,58,208,209] argumentando que la densidad del aire es una variable de escasa influencia en la estimación, se proponen modelos MCP ejecutados con técnicas de Machine Learning donde la densidad del aire es tenida en cuenta alterando las formas funcionales en las que dicha variable se define en el algoritmo de aprendizaje. En este sentido, se proponen alternativas en las que se introducen en los modelos las medidas de temperatura, humedad relativa y presión atmosférica y otras en las que previo al entrenamiento, la densidad del aire es precalculada en base a las tres variables anteriores. También se proponen opciones en las que se ejecutan modelos especializados en la estimación de la velocidad del viento y la densidad del aire a largo plazo de manera desagregada para, posteriormente, calcular la potencia eólica con el objetivo de lograr la mayor capacidad de adaptación posible del modelo desarrollado.

Los resultados obtenidos han sido descritos en los Apartados 4.2 – 4.3. Asimismo, estos resultados han sido publicados para el caso del cálculo de densidades de potencia en la revista *Energy Conversion and Management* (IF: 6.377) con el artículo titulado “*Comparison of several measure-correlate-predict models using support vector regression techniques to estimate wind power densities. A case study*” y en la revista *Applied Energy* (IF: 7.900) con el artículo titulado “*Performance assessment of five MCP models proposed for the estimation of long-term wind turbine power outputs at a target site using three machine learning techniques*”. Las páginas iniciales de estos dos artículos se muestran en el Anexo de la presente tesis doctoral.

2. Los distintos modelos ensayados para la estimación de las potencias teóricas (WTPO) a largo plazo han sido validados para las dos principales familias de sistemas de control existentes en el mercado actual de aerogeneradores, estas son las blade pitch control y las stall-regulated. De esta forma, las conclusiones obtenidas en el análisis realizado en el Apartado 4.3 son de aplicación para cualquier tipo de aerogenerador existente en la gama comercial.

Los resultados obtenidos en este análisis también fueron publicados en el artículo “*Performance assessment of five MCP models proposed for the estimation of long-term wind turbine power outputs at a target site using three machine learning techniques*”.

3. Se comparan tres técnicas de Machine Learning para la estimación de la potencia eólica a largo plazo. Las técnicas comparadas han sido Artificial Neural Networks, la cual ostenta la condición de referencia en la industria [11,13,17,43-45], Support Vector Regression y Random Forest. Estas dos últimas, si bien no habían sido usadas con anterioridad para esta aplicación, si se

consideraba que eran las opciones que mejor representaban el estado del arte en Machine Learning al haberse empleado con éxito en otros campos del conocimiento.

Los resultados de esta comparación han sido expuestos en el Apartado 4.3 de este documento y, adicionalmente, han sido publicados en el artículo *“Performance assessment of five MCP models proposed for the estimation of long-term wind turbine power outputs at a target site using three machine learning techniques”*.

Esta comparación también ha sido desarrollada para cuando el objetivo es la estimación de la velocidad del viento. Los resultados han sido expuestos mediante comunicación oral en una conferencia internacional, en concreto, *“7th International Congress on Energy and Environment Engineering and management (CIEM7) 17-19 July 2017, Spain”*. Asimismo, fue desarrollado un extended abstract el cual fue titulado de la misma forma que la ponencia, *“An empirical assessment of several machine Learning approaches to estimate long-term wind speed conditions”*. Se presenta en el Anexo de la presente tesis doctoral el extender abstract así como el certificado que demuestra la presentación oral desarrollada.

4. Dado que uno de los objetivos básicos de la presente tesis doctoral es la búsqueda de los parámetros que mayor influencia tienen sobre la estimación de la potencia eólica, ha sido usada una técnica Feature Selection, en concreto la Recursive Feature Elimination (RFE) [184,186] la cual se encuentra implementada a través del paquete R Statistic “Caret” [185], y que permite la búsqueda exhaustiva de las variables de partida que aportan significancia al modelo y no produce sobreajustes que perjudican al resultado de la regresión obtenida.

Hasta ese momento, la única referencia del uso de técnicas FS aplicadas a la resolución de problemas MCP fue desarrollada por Carta et al. [45], la cual se aplicó a la técnica de Machine Learning ANN. En este contexto, se propone la adaptación del algoritmo RFE para que además sea un método válido para dar solución a problemas MCP en los que la técnica de regresión sean SVR y RF.

Los resultados han sido descritos en los Apartados 4.2 – 4.3 de este documento y, adicionalmente, han sido publicados en los dos artículos enunciados con anterioridad para las revistas indexadas Energy Conversion and Management y Applied Energy.

5. Los modelos ML que mejores resultados obtuvieron para la comparativa descrita en el punto 2 fueron utilizados para el desarrollo de ajustes a largo plazo empleando la metodología MCP en distintas regiones del archipiélago canario. Con los resultados obtenidos se desarrollaron ajustes de mapas eólicos los cuales fueron inicialmente desarrollados con un modelo Numerical Weather Prediction (NWP) alimentado con datos de reanálisis.

Los resultados obtenidos fueron publicados en el congreso anteriormente descrito mediante la ponencia *“Adjustments of wind maps through data recovered in weather stations. Case study: Canary Island”* a la cual se asiste en calidad de segundo autor.

5.3. Conclusiones

Se relacionan a continuación las conclusiones más relevantes fruto de la investigación desarrollada y en coherencia con los resultados obtenidos en los diferentes estudios realizados:

1. Se demuestra que la densidad del aire no sólo es un parámetro que mejora la capacidad predictiva de los modelos ML entrenados para la estimación de la potencia, sino que además dicha mejora es considerada como estadísticamente significativa, habiéndose aplicado un test no paramétrico de permutación no pareado donde el nivel de significancia ha sido establecido en el 5%. Esta conclusión ha sido validada tanto para cuando el objetivo es la estimación de las densidades de potencia (WPD) o la potencia teórica del parque eólico (WTPO).
2. Adicionalmente se ha demostrado que la importancia de la densidad del aire en la estimación de la potencia eólica es estadísticamente significativa con independencia del tipo de sistema de control considerado (blade pitch control o stall-regulated), lo que reafirma la importancia de la conclusión extraída. Además, ha de tenerse en cuenta que las posiciones evaluadas se encuentran a nivel del mar donde la densidad del aire tiende a mantenerse estable en comparación con regiones de mayor altitud donde la variación es mayor y, por ende, su posible influencia sería aún mayor.
3. De las distintas alternativas evaluadas para la estimación de la potencia, tanto medida en términos de densidad de potencia (WPD) como en potencia teórica (WTPO), los resultados obtenidos demuestran que los mejores resultados se obtienen cuando con anterioridad al cálculo de la potencia a través de la formulación teórica aplicada en cada caso, se ejecutan dos modelos MCP independientes especializados en la estimación de la velocidad del viento y la densidad del aire. Los resultados obtenidos con esta variante producen una mejora estadísticamente significativa en comparación con todas las alternativas evaluadas en este trabajo (hasta 8 propuestas diferentes), entre las que se incluían recomendaciones de otros autores y los modelos genéricos comúnmente empleados.
4. De las tres técnicas ML comparadas SVR y RF han sido las mejores tanto para cuando el objetivo es la estimación de la potencia como la velocidad del viento. Además, no existen diferencias significativas entre estas dos técnicas (asumiendo un nivel de significación del 5%) y siendo una mejor que la otra con dependencia de la situación existente en cada caso evaluado.

Si existen diferencias entre estas dos técnicas en lo relativo a los tiempos computacionales requeridos para su ejecución. La mayor dificultad en la búsqueda de los parámetros óptimos de la técnica SVR, hacen que los tiempos computacionales de esta alternativa sean un 35% superiores a los requeridos para RF.

Por otra parte, de la comparación entre estas dos técnicas con la referencia en la industria (ANN), en concreto la arquitectura Multi-Layer Perceptron (MLP), se concluye que ambas mejoran de forma estadísticamente significativa los resultados obtenidos con la técnica ANN. Esta conclusión se justifica por la mayor capacidad predictiva de las técnicas SVR y RF, las cuales

presentan algoritmos de entrenamiento más eficientes y estables, capaces de proporcionar una solución única al problema en contraste con la técnica ANN la cual su estructura sólo permite la búsqueda de óptimos locales.

5. Se demuestra que la técnica FS es una opción técnicamente viable con la que automatizar la búsqueda de los parámetros óptimos en el entrenamiento de los modelos de aprendizaje tanto para la técnica de referencia (ANN), como para las otras dos propuestas más innovadoras en este campo del conocimiento (SVR y RF).

En general, se concluye que el método RFE es una opción bastante exhaustiva y que elimina del análisis el componente del error humano. No obstante, los tiempos de computación se disparan hasta en un 800%, por lo que se recomienda su utilización sólo en la primera ejecución del modelo en casos en los que el objetivo sea la predicción a corto plazo. Por otra parte, la técnica empleada sólo permite computar hasta dos métricas para el caso de regresión (RMSE y R^2), por lo que una mayor profundidad en el análisis sólo puede ser obtenida si las distintas permutaciones se ejecutan de forma manual o se modifica el algoritmo de entrenamiento utilizado.

6. Se demuestra que la técnica SVR propuesta en este trabajo para el ajuste a largo plazo aplicando la metodología MCP es una opción válida que puede ser utilizada para el reajuste de mapas eólicos en regiones insulares.
7. De acuerdo con los resultados obtenidos en el estudio desarrollado, los mapas eólicos ajustados con este procedimiento manifestaron una mejora considerable de las métricas MAPE y R^2 cuando se comparaban con los resultados obtenidos mediante su versión original, tomando como referencia la producción histórica de parque eólicos situados en las regiones de mayor potencial energético en las Islas Canarias.

5.4. Líneas futuras de investigación

Con la presente tesis doctoral se ha logrado confirmar una serie de hipótesis las cuales han sido definidas en la introducción general (Capítulo 1) y los apartados introductorios y de antecedentes del capítulo de análisis (Capítulo 4). No obstante, a medida que se iba dando solución a algunas hipótesis surgían otras las cuales no han sido resueltas hasta el momento. Dichas hipótesis marcan las futuras líneas de investigación. Se exponen a continuación las líneas detectadas:

1. Como ha sido tratado con anterioridad, una de las contribuciones de este trabajo ha sido la propuesta del algoritmo RFE para la selección de las variables más influyentes en la estimación de la potencia eólica. No obstante, se considera que el análisis estadístico fruto de esta alternativa no es lo suficientemente profundo como para comprender las complejas interrelaciones entre las variables explicatorias con la variable respuesta. En este sentido, se propone el desarrollo de métodos más exhaustivos los cuales incluso integren el uso de test no

paramétricos. Esta línea de trabajo ya se está desarrollando en el grupo de investigación en el marco del proyecto ENERMAC financiado por el programa INTERREG MAC 2014 – 2020 (fondos FEDER).

2. En este estudio se ha demostrado que técnicas más innovadoras como SVR y RF mejoran la predictibilidad en comparación con la que hasta ahora ha sido la referencia ANN. Se propone como línea futura de investigación la evaluación de otras técnicas y, en concreto, métodos que comienzan a despuntar tales como el Deep Learning o alternativas Clustering Regression [125,126,146,150].
3. También plantea interés el estudio en detalle de métodos de hibridación [26,175,223] con los cuales se obtiene una predicción basada en estimaciones generadas con diferentes métodos. En la práctica, esta alternativa es de mucho interés ante situaciones en las que distintos grupos de investigación desarrollan modelos y dichos modelos no reflejan un mismo comportamiento ante el mismo fenómeno y no siendo fácil definir cuál de las estrategias supone una mejor eficacia. En general, este ensamblado se suele desarrollar con métodos de optimización como la programación cuadrática inversa o por la asignación de pesos (la alternativa más usada en energía eólica). Sin embargo, se podría evaluar la búsqueda de mejores estimaciones incluso usando técnicas ML para el ensamblado.
4. De entre las variables explicativas que forman parte de los modelos pueden considerarse otras generadas mediante modelos meteorológicos tales como el Weather Research Forecast (WRF) alimentados con datos de reanálisis, los cuales algunos se ofrecen de manera gratuita. A estos modelos se les conoce como MOS (Model Output Statistic) [234,235] y han probado su relativa eficiencia en distintas aplicaciones. No obstante, sería conveniente su investigación en regiones de orografía compleja como las existentes en Canarias. Este es otra de las líneas de investigación abiertas y que se están desarrollando en el marco del proyecto ENERMAC.
5. Se propone el desarrollo de modelos especialmente entrenados para detectar las rampas de producción a muy corto plazo de tiempo, lo que supondría una mejora considerable de la capacidad de gestión de las instalaciones. En estos análisis la resolución de los datos de los datos debería ser inferior al minuto, si bien sería recomendable el desarrollo de un estudio previo que permita estimar la resolución idónea a utilizar en base a la predicción de aquellas caídas de producción que mayor impacto negativo suponen al sistema eléctrico. Asimismo, entre las señales a introducir en estos modelos se deberían considerarse otro tipo de variables que estuvieran accesibles desde el SCADA del parque eólico.
6. El análisis desarrollado en este trabajo ha estado especialmente centrado en la estimación de potencia a largo plazo donde los tiempos computacionales requeridos para el desarrollo de los modelos no es un problema de especial importancia. No obstante, en el régimen del corto plazo es requerida la búsqueda de soluciones más eficientes que no solo desarrollen las fases de entrenamiento y predicción en etapas diferenciadas sino que los algoritmos sean lo suficientemente rápidos como para ejecutar las predicciones en el menor tiempo posible y sin perder calidad en la estimación. Esto se plantea como otro reto y una línea de investigación de

especial importancia para la operación de estos modelos a tiempo real.

Referencias bibliográficas

- [1] Ang BW. Monitoring changes in economy-wide energy efficiency: From energy–GDP ratio to composite efficiency index. *Energy Policy* 2006; 34:574-82.
- [2] Ayres RU, Turton H, Casten T. Energy efficiency, sustainability and economic growth. *Energy* 2007; 32:634-48.
- [3] Buttel FH. Social structure and energy efficiency: a preliminary cross- national analysis. *Hum Ecol* 1978;6:145-64.
- [4] Kramer O, Gieseke F, Satzger B. Wind energy prediction and monitoring with neural computation. *Neurocomputing* 2013; 109:84-93.
- [5] Aoife M. Foley, Paul G. Leahy, Antonino Marvuglia, Eamon J. McKeogh. Current methods and advances in forecasting of wind power generation. *Renewable Energy* 2011.
- [6] Ramirez-Rosado IJ, Alfredo Fernandez-Jimenez L, Monteiro C, Sousa J, Bessa R. Comparison of two new short-term wind-power forecasting systems. 2009.
- [7] Ministerio de Industria, Energía y Turismo del Gobierno de España. *Procedimientos de Operación de los Sistemas eléctricos No Peninsulares* 2012.
- [8] Ministerio de Industria, Energía y Turismo del Gobierno de España. Real Decreto 413/2014, de 6 de junio, por la que se regula la actividad de producción de energía eléctrica a partir de fuentes de energía renovable, cogeneración y residuos 214.
- [9] Hiester TR PW. *The siting handbook for large wind energy systems*. 1ª Edición ed. New York, 1981.
- [10] Baker RW. Annual and seasonal variations in mean wind speed and wind turbine energy production. *Solar Energy* 1990;45:285; 285,289; 289.

- [11] Velázquez S, Carta JA, Matías JM. Comparison between ANNs and linear MCP algorithms in the long-term estimation of the cost per kWh produced by a wind turbine at a candidate site: A case study in the Canary Islands. *Applied Energy* 2011; 88:3869-81.
- [12] Prasad, R.D., Bansal, R.C. Technologies and methods used in wind resource assessment. In: AF Zobaa RB, editor. *Handbook of Renewable Energy Technology*, Pte. Ltd Singapore: World Scientific Publishing Co.; 2011, p. 69.
- [13] Carta JA, Velázquez S, Cabrera P. A review of measure-correlate-predict (MCP) methods used to estimate long-term wind characteristics at a target site. *Renewable and Sustainable Energy Reviews* 2013; 27:362-400.
- [14] Dinler A. A new low-correlation MCP (measure-correlate-predict) method for wind energy forecasting. *Energy* 2013; 63:152-60.
- [15] Khadem SK, BadgerJ, Ullah SM, Aditya SK, Ghosh HR, Hussain M. The effect of obstacles close to the anemometer mast located on a building on wind flow in the WAsP model. *RETRUD03* 2003.
- [16] Carta JA, Velázquez S, Matías JM. Use of Bayesian networks classifiers for long-term mean wind turbine energy output estimation at a potential wind energy conversion site. *Energy Conversion and Management* 2011; 52:1137-49.
- [17] Velázquez S, Carta JA, Matías JM. Influence of the input layer signals of ANNs on wind power estimation for a target site: A case study. *Renewable and Sustainable Energy Reviews* 2011; 15:1556-66.
- [18] Fadare DA. The application of artificial neural networks to mapping of wind speed profile for energy application in Nigeria. *Applied Energy* 2010; 87:934-42.
- [19] Saavedra-Moreno B, Salcedo-Sanz S, Carro-Calvo L, Gascón-Moreno J, Jiménez-Fernández S, Prieto L. Very fast training neural-computation techniques for real measure-correlate-predict wind operations in wind farms. *J Wind Eng Ind Aerodyn* 2013; 116:49-60.
- [20] Deligiorgi D, Philippopoulos K, Kouroupetroglou G. Artificial neural network based methodologies for the estimation of wind speed. *Green Energy and Technology* 2013; 129:247-66.
- [21] Mori H, Umezawa Y. Application of NB tree to selection of meteorological variables in wind speed prediction. 2009; *Proceedings of the IEEE transmission & distribution conference & exposition, Asia and Pacific*;
- [22] Zhao Y. Wind speed forecasting based on chaotic particle swarm optimization support vector machine. *International journal of applied mathematics & statistics* 2013; 48:347; 347,355; 355.
- [23] Lei M, Shiyan L, Chuanwen J, Hongling L, Yan Z. A review on the forecasting of wind speed and generated power. *Renewable and Sustainable Energy Reviews* 2009; 13:915-20.
- [24] Liu H, Tian H-, Pan D-, Li Y-. Forecasting models for wind speed using wavelet, wavelet packet, time series and Artificial Neural Networks. *Applied Energy* 2013; 107:191-208.
- [25] Wang J, Qin S, Zhou Q, Jiang H. Medium-term wind speeds forecasting utilizing hybrid models for three different sites in Xinjiang, China. *Renewable Energy* 2015; 76:91-101.
- [26] Zhang J, Chowdhury S, Messac A, Hodge B. A hybrid measure-correlate-predict method for long-term wind condition assessment. *Energy Conversion and Management* 2014; 87:697-710.

- [27] Öztopal A. Artificial neural network approach to spatial estimation of wind velocity data. *Energy Conversion and Management* 2006; 47:395-406.
- [28] Bilgili M, Sahin B, Yasar A. Application of artificial neural networks for the wind speed prediction of target station using reference stations data. *Renewable Energy* 2007; 32:2350-60.
- [29] Bechrakis DA, Deane JP, McKeogh EJ. Wind resource assessment of an area using short term data correlated to a long term data set. *Solar Energy* 2004; 76:725-32.
- [30] Wei-Chang Yeh, Yuan-Ming Yeh, Po-Chun Chang, Yun-Chin Ke, Vera Chung. Forecasting wind power in the Mai Liao Wind Farm based on the multi-layer perceptron artificial neural network model with improved simplified swarm optimization. *Electrical power and energy systems* 2014; 55.
- [31] Troncoso A, Salcedo-Sanz S, Casanova-Mateo C, Riquelme JC, Prieto L. Local models-based regression trees for very short-term wind speed prediction. *Renewable Energy* 2015; 81:589-98.
- [32] Schlechtingen M, Santos IF, Achiche S. Using Data-Mining Approaches for Wind Turbine Power Curve Monitoring: A Comparative Study. *Sustainable Energy, IEEE Transactions on* 2013; 4:671-9.
- [33] Velo R, López P, Maseda F. Wind speed estimation using multilayer perceptron. *Energy Conversion and Management* 2014; 81:1-9.
- [34] Azad HB, Mekhilef S, Ganapathy VG. Long-term wind speed forecasting and general pattern recognition using neural networks. *IEEE Transactions on Sustainable Energy* 2014; 5:546-53.
- [35] López P, Velo R, Maseda F. Effect of direction on wind speed estimation in complex terrain using neural networks. *Renewable Energy* 2008; 33:2266-72.
- [36] Butler DA. Intelligent Tools for Wind Resource Correlation. *Wind Eng* 2010; 34:157; 157,166; 166.
- [37] Risø National Laboratory (DTU Wind Energy). WAsP. ; Version 11.
- [38] EMD International A/E. WindPro.
- [39] True Power. OpenWind Enterprise.
- [40] WindSim AS. WindSim.
- [41] Lange M, Focken U. Physical approach to short-term wind power prediction. *Physical Approach to Short-Term Wind Power Prediction* 2006:1-208.
- [42] Mathew S. Wind energy: Fundamentals, resource analysis and economics. *Wind Energy: Fundamentals, Resource Analysis and Economics* 2007:1-246.
- [43] Hau E. Wind Turbines. Fundamentals, Technologies, Application, Economics. 3th ed. 2013th ed. , New York: Springer.
- [44] Jung S, Kwon S. Weighted error functions in artificial neural networks for improved wind energy potential estimation. *Applied Energy* 2013; 111:778-90.
- [45] Carta J.A., Cabrera Pedro, Matías José M, Castellano Fernando. Comparison of feature selection methods using ANNs in MCP-wind speed methods. A case study. *Applied Energy* 2015; 158:490-507.
- [46] Ciemat. Principios de conversión de la energía eólica. Madrid: Editorial Ciemat, 1999.

- [47] Wagner Hermann Josef, Mathur Jyotirmay. Introduction to Wind Energy Systems. Basics, technology and Operation. 2nd ed. : Springer, 2009.
- [48] Bradley EF. The influence of thermal stability and angle of incidence on the acceleration of wind up a slope. J Wind Eng Ind Aerodyn 1983; 15:231-42.
- [49] Burton Tony, Sharpe David, Jenkins Nick, Bossanyi Ervin. Wind Energy Handbook. : Wiley, 2002.
- [50] Pouquet A, Marino R. Geophysical turbulence and the duality of the energy flow across scales. Phys Rev Lett 2013; 111.
- [51] Lu H, Porté-Agel F. Large-eddy simulation of a very large wind farm in a stable atmospheric boundary layer. Physics Fluids 2011; 23.
- [52] Bodo Ruck, Manuel Gruber. Wind power density loss for wind turbines due to upstream hills. Laboratory of Building and Environmental Aerodynamics.
- [53] H. W. Tieleman. Wind characteristics in the surface layer over heterogeneous terrain. J Wind Eng Ind Aerodyn 1992:41-44, 329-340.
- [54] W. Schlez. Voltage fluctuations caused by groups of wind turbines. 2000; Loughborough University.
- [55] AWS Scientific I. Wind resource assessment handbook. National Renewable Energy Laboratory 1997.
- [56] Le Gouriêrês Désiré. Energía eólica. Teoría, concepción y cálculo práctico de las instalaciones. 1983.
- [57] Consejería de Empleo, Industria y Comercio del Gobierno de Canarias. DECRETO 6/2015, de 30 de enero, por el que se aprueba el Reglamento que regula la instalación y explotación de los Parques Eólicos en Canarias. 2015.
- [58] Carta JA, Mentado D. A continuous bivariate model for wind power density and wind turbine energy output estimations. Energy Conversion and Management 2007; 48:420-32.
- [59] Casella L. Performance analysis of the first method for long-term turbulence intensity estimation at potential wind energy sites. Renewable Energy 2015; 74:106-15.
- [60] Carta JA, Ramírez P, Velázquez S. Influence of the level of fit of a density probability function to wind-speed data on the WECS mean power output estimation. Energy Conversion and Management 2008; 49:2647-55.
- [61] Carta JA, Ramírez P, Bueno C. A joint probability density function of wind speed and direction for wind energy analysis. Energy Conversion and Management 2008; 49:1309-20.
- [62] Carta JA, Ramírez P. Use of finite mixture distribution models in the analysis of wind energy in the Canarian Archipelago. Energy Conversion and Management 2007; 48:281-91.
- [63] Carta JA, Ramírez P, Velázquez S. A review of wind speed probability distributions used in wind energy analysis: Case studies in the Canary Islands. Renewable and Sustainable Energy Reviews 2009; 13:933-55.
- [64] Mihelic-Bogdanic A, Budin R. Specific wind energy as a function of mean speed. Renewable Energy 1992; 2:573-6.
- [65] Kaminsky FC. Bivariate probability models for the description of average wind speed at two heights. Solar Energy 1988; 40:49; 49,56; 56.

- [66] Carlin J, Haslett J. Probability distribution of wind power from a dispersed array of wind turbine generators. *J Applied Meteorological* 1982; 21:303-13.
- [67] Bardsley WE. Note on the use of the inverse Gaussian distribution for wind energy applications. *Journal of applied meteorology* (1962) 1980;19 , Sep. 1980, p.1126-1130.
- [68] Canavos G. *Applied probability statistical methods*. 1st ed. 1998.
- [69] van der Auwera L, de Meyer F, Malet LM. Use of the weibull three-parameter model for estimating mean wind power densities. *J Applied Meteorological* 1980; 19:819-25.
- [70] LUNA RE. Estimation of long-term concentrations using a 'universal' wind speed distribution. *J.appl.met.* 1974;13.
- [71] Jaramillo OA, Borja MA. Wind speed analysis in La Ventosa, Mexico: a bimodal probability distribution case. *Renewable Energy* 2004; 29:1613-30.
- [72] Lubitz WD. Power law extrapolation of wind measurements for predicting wind energy production. *Wind Eng* 2009; 33:259-71.
- [73] Berg Jacob, Mann Jakob, Nielsen Morten. *Introduction to micro meteorology for wind energy*. DTU Wind Energy 2013.
- [74] Mikhail AS. Height extrapolation of wind data. *Journal of Solar Energy Engineering, Transactions of the ASME* 1985; 107:10-4.
- [75] Gualtieri G, Secci S. Methods to extrapolate wind resource to the turbine hub height based on power law: A 1-h wind speed vs. Weibull distribution extrapolation comparison. *Renewable Energy* 2012; 43:183-200.
- [76] Heck D., Chickering M., Geiger D. *Learning Bayesian networks: the combination of knowledge and statistical data*. Machine Learning 1995.
- [77] Hansen Martin O. L. *Aerodynamics of Wind Turbines*. : James x James, 2003.
- [78] A. Picard, RS Davis, M Gläser, K Fujii. Revised formula for the density of moist air (CIPM-2007). *IOP Publishing* 2008; 45:149-155.
- [79] Davis RS. Equation for the Determination of the Density of Moist Air (1981/91). *Metrologia* 1992; 29:67-70.
- [80] Mandal G, Kumar A, Sharma D.C, Kumar H. Comparative analysis of different Air Density Equations. *Journal of Metrology Society of India* 2013; 28:51-62.
- [81] Hyland RW, Wexler A. Formulations for the thermodynamic properties of the saturated phases of H₂O from 173.15 K to 473.15 K. *ASHRAE Transactions* 1983; 89:500-519.
- [82] Herrmann S, Kretzschmar HJ, Gatley DP. Thermodynamic properties of real moist air, dry air, steam, water, and ice (RP-1485). *HVAC&R Research* 2009; 15: 961-986.
- [83] Nelson HF, Sauer HJ. Formulation of high-temperature properties for moist air. *HVAC&R Research* 2002; 8: 311-334.
- [84] Harvey AH, Huang PH. First-principles calculation of the air–water second virial coefficient. *International Journal of Thermophysics* 2007; 28:556-565.
- [85] Mago PJ, Sherif SA. Psychrometric charts and property formulations of supersaturated air. *HVAC&R Research* 2005; 11: 147-163.

- [86] Kell GS. Density, thermal expansivity, and compressibility of liquid water from 0° to 150 °C: correlations and tables for atmospheric pressure and saturation reviewed and expressed on 1968 temperature scale. *J Chem Eng Data* 1975; 20: 97–105.
- [87] International Electrotechnical Commission. IEC 61400-12. Wind turbine generator systems – Part 12: Wind turbine power performance testing. 1998.
- [88] Technical Committee 20, Subcommittee 6. ISO 2533. Standard atmosphere. 1975.
- [89] Nordex. Nordex N60. http://www.nordex-online.com/fileadmin/MEDIA/Produktinfos/EN/Nordex_N60_EN.pdf. [Last accessed 15-12-2017].
- [90] Enercon. Catálogo Enercon product overview: http://www.enercon.de/fileadmin/Redakteur/Medien-Portal/broschueren/pdf/en/ENERCON_Produkt_en_06_2015.pdf. Enercon. [Last accessed 08-07-2018].
- [91] Taslimi-Renani E, Modiri-Delshad M, Elias MFM, Rahim NA. Development of an enhanced parametric model for wind turbine power Curve. *Applied Energy* 2016; 177: 544–552.
- [92] Lydia M, Kumar SS, Selvakumar AI, Kumar GEP. A comprehensive review on wind turbine power curve modeling techniques. *Renewable and Sustainable Energy Reviews* 2014; 30:452–460.
- [93] Press WH, Teukolsky SA, Vetterling WT, Flannery BP. *Numerical Recipes*. 3th ed. Cambridge University Press; New York 2007.
- [94] Brower MC, Bailey BH, Beaucage P, Bernadett DW, Doane J, Eberhard MJ et al. *Wind Resource Assessment: A Practical Guide to Developing a Wind Project*. Wind Resource Assessment: A Practical Guide to Developing a Wind Project 2012.
- [95] Svenningsen L. Proposal of improved power curve air-density correction. Paper number: PO.310. Proceedings of the European wind energy conference and exhibition Warsaw, Poland April 2010.:20-3.
- [96] WindPRO. Power curve options. http://www.emd.dk/files/windpro/WindPRO_Power_Curve_Options.pdf. [Last accessed 08-07-2018].
- [97] Ministerio de Industria, Energía y Turismo del Gobierno de España. Orden IET/1459/2014, de 1 de agosto, por la que se aprueban los parámetros retributivos y se establece el mecanismo de asignación del régimen retributivo específico para las nuevas instalaciones de energía eólica y fotovoltaica en los sistemas eléctricos no peninsulares. 2014.
- [98] Consejería de Empleo, Industria y Comercio del Gobierno de Canarias. *Anuario Energético de Canarias 2013*. 2013.
- [99] Ayotte KW, Davy RJ, Coppin PA. A Simple Temporal and Spatial Analysis of Flow in Complex Terrain in the Context of Wind Energy Modelling. *Bound -Layer Meteorol* 2001; 98:275-95.
- [100] Nielsen M. Application of the measure–correlate–predict approach for wind resources assessment. *European Wind Energy Conference. Wind Energy for the New Millennium* 2001:773; 773,776; 776.
- [101] Bowen AJ. *WASP Prediction Errors Due to Site Orography*.
- [102] Landberg L. *Proceedings of the 15th British Wind Energy Association Conference* 1993:119; 119,125; 125.

- [103] Taylor M. Proceedings of the European Wind Energy Conference & Exhibition.
- [104] Oliver A. Proceedings of the AWEA 2010 Windpower Conference and Exhibition 2010.
- [105] Efstathiou E. Michaelides. *Alternative Energy Sources*. 1ª Edición ed. : Springer, 2012.
- [106] Emeis S. *Wind Energy Meteorology* 2013.
- [107] Ren D. Effects of global warming on wind energy availability. *Journal of Renewable & Sustainable Energy* 2010; 2:052301.
- [108] Pryor SC, Barthelmie RJ. Climate change impacts on wind energy: A review. *Renewable and Sustainable Energy Reviews* 2010; 14:430-7.
- [109] Vautard R, Cattiaux J, Yiou P, Thépaut J-, Ciais P. Northern Hemisphere atmospheric stilling partly attributed to an increase in surface roughness. *Nature Geoscience* 2010; 3:756-61.
- [110] Breslow PB, Sailor DJ. Vulnerability of wind power resources to climate change in the continental United States. *Renewable Energy* 2002; 27:585-98.
- [111] Sailor DJ, Smith M, Hart M. Climate change implications for wind power resources in the Northwest United States. *Renewable Energy* 2008; 33:2393-406.
- [112] Christensen JH. Climate change 2007: the physical science basis. Contribution of Working Group I to the fourth assessment report of the intergovernmental panel on climate change 2007.
- [113] Probst O, Cárdenas D. State of the Art and Trends in Wind Resource Assessment. *Energies* 2010; 3:1087-141.
- [114] National Climatic Data Center (NOAA). <http://www.ncdc.noaa.gov/thredds/ncss/grid/oceanwinds6hr/dataset.html>. ; [Last accessed 15-02-2018].
- [115] Brower MC. Proceedings of the European Wind Energy Conference & Exhibition.
- [116] Pinto C. Proceedings of the European Wind Energy Conference & Exhibition.
- [117] Liléo Petrik SO. Proceedings of the European Wind Energy Conference & Exhibition.
- [118] Beaucage Philippe, Brower Michael C. Wind flow model performance. Do more sophisticated models produce more accurate wind resource estimates? 2012: AWS TruePower.
- [119] Woods JC, Watson SJ. A new matrix method of predicting long-term wind roses with MCP. *J Wind Eng Ind Aerodyn* 1997; 66:85-94.
- [120] Bass Jh R. An Improved Measure-correlate-predict Algorithm for the Prediction of the Long Term Wind Climate in Regions of Complex Environment.
- [121] Zheng L, Hu W, Min Y. Raw wind data preprocessing: A data-mining approach. *IEEE Transactions on Sustainable Energy* 2014; 6:11-9.
- [122] Gamesa. <http://www.gamesacorp.com/es/productos-servicios/aerogeneradores/gamesa2mw.html>. [Last accessed 07-01-2018].
- [123] Vestas. http://www.vestas.com/en/products_and_services/turbines/v110-2_0_mw#!power-curve-and-aep. 2015.
- [124] Llombart A, Pueyo C, Fandos JM, Guerrero JJ. Robust data filtering in wind power systems. *European Wind Energy Conference and Exhibition 2006, EWEC 2006* 2006; 2:1611-6.

- [125] Witten IH, Eibe F, Mark AH. Data Mining. Practical Machine Learning Tools and Techniques. 3ª Edición ed. : Morgan Kaufmann, 2011.
- [126] Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani. An Introduction to Statistical Learning. : Springer, 2013.
- [127] Kalogirou SA. Artificial neural networks in renewable energy systems applications: a review. Renewable and Sustainable Energy Reviews 2001; 5:373-401.
- [128] Landberg L, Myllerup L, Rathmann O, Petersen EL, Jørgensen BH, Badger J et al. Wind resource estimation - An overview. Wind Energy 2003;6:261-71.
- [129] Gardner MW, Dorling SR. Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences. Atmos Environ 1998; 32:2627-36.
- [130] Addison JF, Hunter A, Bass J, Rebbeck M. A neural network version of the measure–correlate–predict algorithm for estimating wind energy yield. In. Houston, Texas; 3–8 December 2000. p.917–922.
- [131] WindPRO. WASP CFD in WindPRO 2.0. Energy Calculations with CFD. http://emd.dk/files/windpro2.9/WindPRO_and_WASPCFD.pdf. [Last accessed 12-04-2016]
- [132] Jain P. Wind Energy Engineering 2011.
- [133] Derrick A. Development of the measure-correlate-predict strategy for site assessment. European Community wind energy conference, Bedford: Stephens HS and Associates 1993:p.681-5.
- [134] Derrick A. Development of the Measure-Correlate-Predict Strategy for Site Assessment. 14th British Wind Energy Conference, Nottingham: Mechanical Engineering Publications 1992:p. 259-65.
- [135] Langreder W. Wind resource and site assessment. Wind power generation and wind turbine design 2010; MA: WIT Press:p.49-87.
- [136] Barthelmie RJ PS. Meteorology and wind resource assessment for wind farm development. Wind energy systems 2011; Cambridge: Woodhead Publishing:p.3-27.
- [137] Manwell JF, McGowan JG, Rogers AL. Wind energy explained. 2nd ed. New York: Wiley, 2009.
- [138] Casella L, Lavagnini A, Sorrentino G. Comparison between WASP and a joint probabilistic method for wind climate assessment in a Mediterranean coastal area. Nuovo Cimento della Società Italiana di Fisica C 2007; 30:365-80.
- [139] Walls L, Kline J, Kline Z. Long-term wind speed estimates from short-term data: So many ways to get it wrong! Proceedings of the AWEA Wind Resource Assessment Workshop 2010; Oklahoma City, Oklahoma.
- [140] Casella L. Improving Long-Term Wind Speed Assessment Using Joint Probability Functions Applied to Three Wind Data Sets. Wind Eng 2012; 36:473; 473,483; 483.
- [141] Barros VR EE. On the evaluation of wind power from short wind records. Journal of Climate and Applied Meteorology 1983; 22:1116-1123.
- [142] Barros VR RJ,. Measurement strategies: Use of short observation records for estimating the annual wind variation. Proceedings of The International Colloquium on Wind Energy 1981; BHRA Fluid Engineering, Brighton, UK:pp. 23-28.

- [143] Barros VR EE. Reply. *Journal Climate and Applied Meteorology* 1984; 23:1480-1483.
- [144] Barros VR SI. On extension of climatic series from short records. *Journal Applied Meteorology* 1988; 27:325-335.
- [145] Skibin D. Comment "On evaluation of wind power from short wind records". *Journal Climate and Applied Meteorology* 1984;23:1477-1479.
- [146] Alpaydin E. *Introduction to Machine Learning*. Massachusetts Institute of Technology, 2004.
- [147] Monfared M, Rastegar H, Kojabadi HM. A new strategy for wind speed forecasting using artificial intelligent methods. *Renewable Energy* 2009; 34:845-8.
- [148] Bishop C. *Neural Networks for Pattern Recognition*. : Oxford, 2010.
- [149] Haykin S. *Neural networks. A Comprehensive Foundation*. Prentice Hall, New Jersey, USA. 1999.
- [150] Cristianini N, Shade-Taylor J. *Support Vector Machines and other Kernel-based learning method*. Cambridge, United Kingdom: Cambridge University Press, 2000.
- [151] Smola. A, Schölkopf B. A tutorial on Support Vector Regression. *NeuroCOLT 2 Technical Report Series* 1998.
- [152] Biau G, Scornet E. A random forest guided tour. *Springer* 2016:197-227.
- [153] Draper NR SH. *Applied regression analysis*. 3rd ed. ed. : New York: Wiley, 1998.
- [154] GNU Project. R Statistics;Version 3.3.2: <https://www.r-project.org/>. [Last accessed 18-04-2017]
- [155] Hastie T, Tibshirani R, Friedman J. *The elements of statistical learning. Data mining, inference and prediction*. 2st ed Stanford: Springer 2013.
- [156] Ripley R. Package nnet Version 7.3-12. *Feed-Forward Neural Networks and Multinomial Log-Linear Models*. CRAN R Statistics 2016.
- [157] Philippopoulos K, Deligiorgi D. Application of artificial neural networks for the spatial estimation of wind speed in a coastal region with complex topography. *Renewable Energy* 2012; 38:75-82.
- [158] Rasit A. Artificial Neural Networks applications in wind energy. *Renewable and Sustainable Energy Reviews* 2015:534-562.
- [159] Leshno M, Lin VY, Pinkus A, and Schocken S. Multilayer feedforward networks witha nonpolynomial activation function can approximate any function. *Neural Networks* 6:861–7. 1993.
- [160] Park J SI. Approximation and Radial-Basis-Function Networks. *Neural Computation* 1993;5: 305–316.
- [161] Girosi F, Jones M, Poggio T. Regularization theory and neural networks architectures. *Neural Computation* 1995; 7:219–269.
- [162] Boser B, Guyon I, Vapnik V. A training algorithm for optimal margin classifiers. In: Haussler D (ed) *Proceedings of the Fifth Annual ACM Workshop on Comput Learn Theory*. ACM Press. 1992.
- [163] Hammer B, Gersmann K. A note on the universal approximation capability of Support Vector Machine. *Neural Processing Letters* 2003; 17:43-53.
- [164] Karatzoglou A, Smola A, Hornik K, Zeileis A. Package Kernlab Version 0.9-22. An S4 Package for Kernel Methods in R. *Reference Manual*. *Journal of Statistical Software* 2015; 11:1-20.

- [165] Kusiak A, Zheng H, Song Z. Short-term prediction of wind farm power: A data mining approach. *IEEE Trans Energy Conversion* 2009; 24:125-36.
- [166] Zhou S-, Mao M-, Su J-. Short-term wind speed prediction with support vector machine based on predict error correction. *Xitong Fangzhen Xuebao / Journal of System Simulation* 2012; 24:769-73.
- [167] Wang X, Li H. Multiscale prediction of wind speed and output power for the wind farm. *Journal of Control Theory and Applications* 2012; 10:251-8.
- [168] Sangitab P, Deshmukh SR. Use of support vector machine for wind speed prediction. 2011 International Conference on Power and Energy Systems, ICPS 2011 2011.
- [169] Peng H. Long term wind speed evaluation based on neighboring meteorological masts. *Tàiyáng néng xuébào* 2012; 33:230; 230,235; 235.
- [170] Jia S. A new method for the short-term wind speed forecasting. DRPT 2011 - 2011 4th International Conference on Electric Utility Deregulation and Restructuring and Power Technologies 2011:1320-4.
- [171] Mohandes MA, Halawani TO, Rehman S, Hussain AA. Support vector machines for wind speed prediction. *Renewable Energy* 2004; 29:939-47.
- [172] Yu J, Chen K, Mori J, Rashid MM. A Gaussian mixture copula model based localized Gaussian process regression approach for long-term wind speed prediction. *Energy* 2013; 61:673-86.
- [173] Polikar R. Ensemble Learning. In: Cha Zhang and Yunqian Ma, editors. *Ensemble Machine Learning. Methods and Applications*, New York: Springer 2012:p. 157-175.
- [174] Breiman L, Cutler A, Liaw A, Wiener M. *Package Random Forest (R)*. CRAN R Statistics 2015.
- [175] Wang J, Qin S, Zhou Q, Jiang H. Medium-term wind speeds forecasting utilizing hybrid models for three different sites in Xinjiang, China. *Renewable Energy* 2015; 76:91-101.
- [176] GNU Project. The Comprehensive R Archive Network (CRAN). <http://ftp.cixug.es/CRAN/>. [Last accessed 08-07-2018]
- [177] Chung CC, Jen L. LIBSVM A Library for Support Vector Machine. <https://www.csie.ntu.edu.tw/~cjlin/libsvm/>; [Last accessed 08-07-2018].
- [178] Claesen M, Simm J, Popovic D. Solver Overview. Methods for selecting optimal parameters. <http://optunity.readthedocs.org/en/latest/user/solvers.html>; [Last accessed 08-07-2018].
- [179] Karatzoglou A, Meyer D, Hornik K. Support Vector Machines in R. *Journal of Statistical Software* 2006; 15.
- [180] Hofmann T, Scholkopf B, Smola A. *Kernel Methods in Machine Learning*. 2008.
- [181] Vapnik, Golowich S, Smola A. Support Vector Method for function approximation, regression estimation and signal processing. *Advances in Neural Information Processing Systems* 1998:281-287.
- [182] Sewell M. Support Vector Machines (SVMs). <http://www.svms.org/>; [Last accessed 08-07-2018].
- [183] Kong X, Liu X, Shi R, Lee KY. Wind speed prediction using reduced support vector machines with feature selection. *Neurocomputing* 2015; 169:449-56.

- [184] Kuhn M. Variable Selection Using The caret Package. 2010.
- [185] Kuhn M. Package "caret". Classification and Regression Training. CRAN R Statistics 2015: <https://github.com/topepo/caret/>. [Last accessed 08-07-2018]
- [186] Kuhn M. Building Predictive Models in R Using the caret Package. Journal of Statistical Software 2008;28:Issue 5.
- [187] Steel R. G. D., Torrie J. H. Principles and Procedures of Statistics with Special Reference to the Biological Sciences. 1st ed. ed. : McGraw Hill, 1960.
- [188] Breiman L. Random Forest. Mach Learn 2001; 45:5-32.
- [189] Breiman L. Bias, variance, and arcing classifiers. 1996.
- [190] Geurts P, Ernst D, Wehenkel L. Extremely randomized trees. Mach Learn 2006;63:3.
- [191] Ho T. The random subspace method for constructing decision forests. Pattern Anal Mach Intell 1998; 20:832.
- [192] NOAA. National Oceanic and Atmospheric Administration. <https://gis.ncdc.noaa.gov/map/viewer/#app=cdo&cfg=cdo&theme=hourly&layers=1> [Last accessed 04-12-2016].
- [193] Ramírez P. Modelado estadístico de las características del viento para su evaluación energética. Aplicación a las Islas Canarias. Universidad de Las Palmas de Gran Canaria (ULPGC) 2006.
- [194] Calero R, Carta JA. Action plan for wind energy development in the Canary Islands. Energy Policy 2004; 32:1185-97.
- [195] Carta, J.A., Calero R., Padrón J., Garcia J. Wind potential in the Canarian Archipelago. Proceedings of the 5th European Wind Energy Association Conference and Exhibition 1994; Vol.3:pp.35-41.
- [196] Instituto Tecnológico de Canarias, S.A. <http://www.itccanarias.org/web/>. [Last accessed 08-07-2018].
- [197] Meteosim SL. Estudio del recurso eólico de las Islas Canarias con MesoMap. 2006.
- [198] Instituto Tecnológico de Canarias, S.A. Recurso eólico de Canarias <http://www.itccanarias.org/recursoeolico/>. [Last accessed 05-04-2017].
- [199] Instituto para la Diversificación y el Ahorro de la Energía (IDAE). Atlas eólico de España <http://atlaseolico.idae.es/>. [Last accessed 05-04-2017].
- [200] Measnet. Evaluation of site - specific Wind Conditions. 2009; Version 1.
- [201] Vestas. WindData Explorer; V1.5.6: <http://winddataexplorer.com/>. [Last accessed 10-04-2017]
- [202] Agencia Estatal de Meteorología (AEMET). Climatol http://www.aemet.es/es/idi/clima/registros_climaticos. [Last accessed 05-04-2017].
- [203] Díaz S, Carta JA, Matías JM. Comparison of several measure-correlate-predict models using support vector regression techniques to estimate wind power densities. A case study. Energy Conversion and Management 2017; 140:362-400.
- [204] Patanè D, Benso M, Hernández C., de La Blanca F, López C. Long term wind resource assessment by means of multivariate cross-correlation analysis. Proceedings of the European Wind Energy Conference & Exhibition 2011; 14-17 March, Brussels, Belgium.
- [205] Spera DA. Wind turbine technology: fundamental concepts of wind turbine engineering 1994.

- [206] Carta JA, Velázquez S. A new probabilistic method to estimate the long-term wind speed characteristics at a potential wind energy conversion site. *Energy* 2011; 36:2671-85.
- [207] Koepl GW. Putnam's power from the wind. 2nd ed ed. New York: Van Nostrand Reinhold, 1982.
- [208] Weekes SM, Tomlin AS. Comparison between the bivariate Weibull probability approach and linear regression for assessment of the long-term wind energy resource using MCP. *Renewable Energy* 2014; 68:529-39.
- [209] Weekes SM, Tomlin AS. Data efficient measure-correlate-predict approaches to wind resource assessment for small-scale wind energy. *Renewable Energy* 2014; 63:162-71.
- [210] Drucker H, Burges Chris J.C, Kaufman L, Smola A, Vapnik V. Support Vector Regression Machines. Bell Labs and Monmouth University. *Advances in Neural Information Processing Systems* 1996.
- [211] Matías, J.M., Vaamonde, A., Taboada, J. and González-Manteiga, W. Support vector machines and gradient boosting for graphical estimation of a slate deposit. *Journal of Stochastic Environmental Research and Risk Assessment*, 2004b; 18:309–323.
- [212] Giorgi M, Ficarella A. Error analysis of the short term wind power prediction models. *Applied Energy* 2011; 88:1298-311.
- [213] Deng L. An overview of Deep - Structured Learning for Information Processing. *APSIPA ASC 2011 Xi'an* 2011.
- [214] D'Agostino RB SM. Goodness of fit techniques. 1st ed. New York: Dekker ed. , 1986.
- [215] Good P. Permutation, parametric and bootstrap tests of hypotheses. 3th ed. New York: Springer ed. , 2005.
- [216] Berry KJ, Johnston JE, Mielke PW. A chronicle of permutation statistical methods. 1st ed. New York: Springer ed. , 2014.
- [217] Benjamini Y HY. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J Roy Stat Soc* 1995; 57:289–300.
- [218] Masseran N. Evaluating wind power density models and their statistical properties. *Energy* 2016; 84:533-54.1.
- [219] Díaz S, Carta JA, Matías JM. Performance assessment of five MCP models proposed for the estimation of long-term wind turbine power outputs at a target site using three machine learning techniques. *Applied Energy* 2018; 209:455-477.
- [220] Liu YZ LW. Effect of air density on output power of wind turbine. *Applied Mechanics and Materials* 2013:336-338: 1114-1117.
- [221] Kelleher JD, Namee BM, D'Arcy A. Fundamentals of machine learning for predictive data analytics. London: The MIT Press 2015; 1st ed.
- [222] Ma J CJ. Identifying the influential features on the regional energy use intensity of residential buildings based on Random Forests. *Applied Energy* 2016; 183: 193–201.
- [223] Ibrahim IA KT. A novel hybrid model for hourly global solar radiation prediction using random forests technique and firefly algorithm. *Energy Conversion and Management* 2017;138: 413–425.

- [224] Lahouar A SJ. Hour-ahead wind power forecast based on random forests. *Renewable Energy* 2017;109: 529-541.
- [225] European Environment Agency. CORINE project. <http://www.eea.europa.eu/data-and-maps/data/corine-land-cover-2006-raster-3/>. [Last accessed 08-07-2018].
- [226] Fundación OSGeo. Quantum GIS. <http://www.qgis.org/es/site/>. 2015;v2.1 [Last accessed 08-07-2018].
- [227] Ministerio de Fomento del Gobierno de España. Centro Nacional de Información Geográfica (CNIG). <http://centrodedescargas.cnig.es/CentroDescargas/catalogo.do#selectedSerie>. [Last accessed 08-11-2015].
- [228] Huaquan ZM. *Wind Resource Assessment and Micro-siting*. Science and Engineering. First Edition ed. : Wiley, 2015.
- [229] Lackner MA, Rogers AL, Manwell JF, McGowan JG. A new method for improved hub height mean wind speed estimates using short-term hub height data. *Renewable Energy* 2010;35: 2340-2347.
- [230] Grassi S, Junghans S, Raubal M. Assessment of the wake effect on the energy production of onshore wind farms using GIS. *Applied Energy* 2014; 136: 827–837.
- [231] Fant C, Gunturu B, Schlosser A. Characterizing wind power resource reliability in southern Africa. *Applied Energy* 2016; 161: 565–573.
- [232] D. Carvalho D. Rocha A, Gómez-Gesteira M, Silva Santos C. Sensitivity of the WRF model wind simulation and wind energy production estimates to planetary boundary layer parameterizations for onshore and offshore areas in the Iberian Peninsula. *Applied Energy* 2014;135: 234–246.
- [233] Kwon S. Uncertainty analysis of wind energy potential assessment. *Applied Energy* 2010; 87:856-65.
- [234] National Weather Service (NOAA). Model Output Statistics (MOS) https://www.weather.gov/mdl/mos_home. ; [Last accessed 08-07-2018].
- [235] Castellani F, Burlando M, Taghizadeh S, Astolfi D, Piccioni E. Wind energy forecast in complex sites with a hybrid neural network and CFD based method. *Energy Procedia* 2014; 45:188–97.
- [236] Zhang MH. *Wind resource assessment and micro-siting*. 1st ed. Singapore: Wiley. , 2015.

Anexos. Publicaciones, congresos y valoración externa del trabajo realizado en la tesis

Introducción

Se expone en este apartado los dos artículos publicados durante el desarrollo de la tesis doctoral.

- Díaz S, Carta JA, Matías JM. Comparison of several measure-correlate-predict models using support vector regression techniques to estimate wind power densities. A case study. *Energy Conversion and Management* 2017; 140:362-400.
- Díaz S, Carta JA, Matías JM. Performance assessment of five MCP models proposed for the estimation of long-term wind turbine power outputs at a target site using three machine learning techniques. *Applied Energy* 2018; 209: 455-477.

Lo acompañan los “Abstract” de las dos conferencias desarrolladas durante este periodo, habiéndose presentado en el primero de ellos un “Extended Abstract” que fue publicado en el acta de congreso.

- Díaz S, Carta JA, Castellano F. An empirical assessment of several machine Learning approaches to estimate long-term wind speed conditions. 7th International Congress on Energy and Environment Engineering and management (CIEM7) 17-19 July 2017, Spain.
- Castellano F, Díaz S, Carta JA. Adjustments of wind maps through data recovered in weather stations. Case study: Canary Island. 7th International Congress on Energy and Environment Engineering and management (CIEM7) 17-19 July 2017, Spain.

Finalmente se adjunta la valoración externa del trabajo de investigación desarrollada por Dr. Marie-Laure Nivet, profesora titular de la Universidad de Córsega (Università de Corsica Pasquale Paoli).

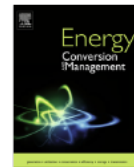
A. Publicaciones

Energy Conversion and Management 140 (2017) 334–354



Contents lists available at ScienceDirect

Energy Conversion and Management

journal homepage: www.elsevier.com/locate/enconman

Comparison of several measure-correlate-predict models using support vector regression techniques to estimate wind power densities. A case study

Santiago Díaz^a, José A. Carta^{a,*}, José M. Matías^b^a Department of Mechanical Engineering, University of Las Palmas de Gran Canaria, Campus de Tafira s/n, 35017 Las Palmas de Gran Canaria, Canary Islands, Spain^b Department of Statistics, University of Vigo, Lagoas Marcosende, 36200 Vigo, Spain

ARTICLE INFO

Article history:

Received 16 November 2016

Received in revised form 22 February 2017

Accepted 23 February 2017

Keywords:

Wind power density
 Measure-correlate-predict
 Support vector regression
 Feature selection
 Statistical significance

ABSTRACT

The long-term annual mean wind power density (WPD) is an important indicator of wind as a power source which is usually included in regional wind resource maps as useful prior information to identify potentially attractive sites for the installation of wind projects. In this paper, a comparison is made of eight proposed Measure-Correlate-Predict (MCP) models to estimate the WPDs at a target site. Seven of these models use the Support Vector Regression (SVR) and the eighth the Multiple Linear Regression (MLR) technique, which serves as a basis to compare the performance of the other models. In addition, a wrapper technique with 10-fold cross-validation has been used to select the optimal set of input features for the SVR and MLR models. Some of the eight models were trained to directly estimate the mean hourly WPDs at a target site. Others, however, were firstly trained to estimate the parameters on which the WPD depends (i.e. wind speed and air density) and then, using these parameters, the target site mean hourly WPDs. The explanatory features considered are different combinations of the mean hourly wind speeds, wind directions and air densities recorded in 2014 at ten weather stations in the Canary Archipelago (Spain).

The conclusions that can be drawn from the study undertaken include the argument that the most accurate method for the long-term estimation of WPDs requires the execution of a specially trained model which considers the variability of the wind speeds of the reference stations, as well as of the wind directions and air densities, and in addition the functional manner in which these variables participate in the proposed MCP models. It is also concluded that it is important to consider the annual variation of air density even in regions at sea level. It is further concluded that, of the eight MCP models under comparison, the one that predicts the WPDs based on two sub-models (which estimate the wind speeds and air densities in an unlinked manner) always provides the best MAE (Mean Absolute Error), MARE (Mean Absolute Relative Error) and R^2 (Coefficient of determination) metrics, with the differences being statistically significant (5% significance) for most of the cases assessed. Additionally, the regulatory capacity of the SVR technique was sufficient to manage most of the overfitting problems, and hence the contribution of the wrapper method was not relevant in our study.

© 2017 Elsevier Ltd. All rights reserved.

1. Introduction

In this section, a background is firstly provided to the problem related to long-term estimation of Wind Power Densities (WPDs) when Measure-Correlate-Predict (MCP) methods are used which are based on information provided by multiple reference weather stations (WSs). Subsequently a description is given of the aim and original contribution of this paper.

1.1. Background

In the scientific literature, an extensive collection of MCP methods [1] have been proposed for hindcasting of the long-term wind characteristics at sites for which only measurements recorded over a short time period are available.

The most commonly proposed and used methods to date in the wind industry have been based on information obtained from a single reference station. However, in the scientific literature concerned with renewable energies a growing number of proposals can be seen for methods which are based on the use of several ref-

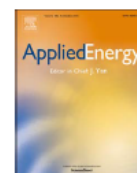
* Corresponding author.

E-mail address: jcarta@dim.ulpgc.es (J.A. Carta).



Contents lists available at ScienceDirect

Applied Energy

journal homepage: www.elsevier.com/locate/apenergy

Performance assessment of five MCP models proposed for the estimation of long-term wind turbine power outputs at a target site using three machine learning techniques

Santiago Díaz^a, José A. Carta^{a,*}, José M. Matías^b

^a Department of Mechanical Engineering, University of Las Palmas de Gran Canaria, Campus de Tafira s/n, 35017 Las Palmas de Gran Canaria, Canary Islands, Spain

^b Department of Statistics, University of Vigo, Lagoas Marcosende, 36200 Vigo, Spain

HIGHLIGHTS

- Five models based on MCP methods to estimate WTPOs at a target site are assessed.
- Models are proposed which take into account air density variability.
- Three machine learning techniques implemented in the models are analysed.
- The models use wind turbines with blade pitch control and stall-regulated wind turbines.
- Statistical hypothesis tests are used to compare the ML techniques in the best model and to compare some of the models.

ARTICLE INFO

Keywords:

Support vector machine
Artificial neural network
Random forest
Wind turbine power curve
Wind turbine power output
Air density

ABSTRACT

Various models based on measure-correlate-predict (MCP) methods have been used to estimate the long-term wind turbine power output (WTPO) at target sites for which only short-term meteorological data are available. The MCP models used to date share the postulate that the influence of air density variation is of little importance, assume the standard value of 1.225 kg m^{-3} and only consider wind turbines (WTs) with blade pitch control.

A performance assessment is undertaken in this paper of the models used to date and of newly proposed models. Our models incorporate air density in the MCP model as an additional covariable in long-term WTPO estimation and consider both WTs with blade pitch control and stall-regulated WTs. The advantages of including this covariable are assessed using different functional forms and different machine learning algorithms for their implementation (Artificial Neural Network, Support Vector Machine for regression and Random Forest).

The models and the regression techniques used in them were applied to the mean hourly wind speeds and directions and air densities recorded in 2014 at ten weather stations in the Canary Archipelago (Spain). Several conclusions were drawn from the results, including most notably: (a) to clearly show the notable effect of air density variability when estimating WTPOs, it is important to consider the functional ways in which the features air density and wind speed and direction intervene, (b) of the five MCP models under comparison, the one that separately estimates wind speeds and air densities to later predict the WTPOs always provided the best mean absolute error, mean absolute relative error and coefficient of determination metrics, independently of the target station and type of WT under consideration, (c) the models which used Support Vector Machines (SVMs) for regression or random forests (RFs) always provided better metrics than those that used artificial neural networks, with the differences being statistically significant (5% significance) for most of the cases assessed, (d) no statistically significant differences were found between the SVM- and RF-based models.

1. Introduction

When making a decision as to whether a particular wind turbine (WT) should be installed at a target site it is of interest to know the wind

turbine power outputs (WTPOs). The WTPOs are estimated using the power curve of the WT and the characteristics of the wind regime and air density at the site where the WT is to be installed (or not). The estimated long-term gross annual mean WTPO at a target site allows a

* Corresponding author.

E-mail address: jcarta@dim.uipgc.es (J.A. Carta).

<https://doi.org/10.1016/j.apenergy.2017.11.007>

Received 8 August 2017; Received in revised form 3 October 2017; Accepted 2 November 2017
0306-2619/© 2017 Elsevier Ltd. All rights reserved.

B. Congresos



7th International Congress
Energy and Environment Engineering and Management

(CIEM7)

17-19 July 2017

Canary Islands, SPAIN

Extending

ABSTRACTS
B O O K

7th International Congress on Energy and Environment Engineering and management (CIEM7)
17-19 July 2017, Canary Islands, SPAIN

Integration of renewable sources of energy in low voltage distribution network. Case small wind turbine	129
A. Argudo Chalán ⁽¹⁾ , F. Déniz ⁽¹⁾ , E. Vega-Fuentes ⁽²⁾ , S. Marrero Marrero ⁽¹⁾ Miguel Martínez Mergarejo ⁽¹⁾	
.....	129
Improving the sustainability of the indigo dyeing by applying an electrochemical reduction process.....	134
V. Buscio ⁽¹⁾ , X. Coma ⁽¹⁾ , V. López-Grimau ⁽¹⁾ , B. Amante ⁽¹⁾ , C. Gutiérrez-Bouzán ⁽¹⁾	
.....	134
Pellets for energy made from pine wood versus grape stalks: a comparative life cycle assessment.....	138
J. Ferreira ⁽¹⁾ , B. Esteves, L. Cruz Lopes, I. Domingos	
.....	138
An empirical assessment of several machine learning approaches to estimate long-term wind speed conditions	142
Santiago Díaz ^{*(1,2)} , José Antonio Carta ⁽¹⁾ , Fernando Castellano ^(1,2)	
.....	142
Analysis of High Integration of Renewable Energy in Low Voltage Distribution Networks	148
S. Hernández-Fragiel ⁽¹⁾ , M. Martínez-Melgarejo ⁽²⁾ , F. Déniz ⁽³⁾ , E. Vega-Fuentes ⁽⁴⁾	
.....	148
Anaerobic Degradation of Synthetic Domestic Wastewater in a Hybrid Reactor.....	153
P. Singh ⁽¹⁾ , A. Singla ⁽²⁾ , P.P.S. Cheema ⁽¹⁾	
.....	153
EVALUATION AND MAPPING OF GROUNDWATER QUALITY STATUS OF LUDHIANA (INDIA)	160
K. Singh ⁽¹⁾ , J. Jyoti ⁽¹⁾ , H.S. Rai ⁽¹⁾	
.....	160
Hydropower for Sustainable and Green Energy in Turkey	165
I. Yuksel ⁽¹⁾ , H. Arman ⁽²⁾ , I. H. Demirel ⁽³⁾	
.....	165
The Role of the Energy Management for Clean Energy in Turkey	169
I. Yuksel ⁽¹⁾ , H. Arman ⁽²⁾ , I. H. Demirel ⁽³⁾	
.....	169
Technical-economic analysis of the use of textile waste for the production of thermal energy	174
L. J. R. Nunes ⁽¹⁾ , R. Godina ⁽²⁾ , J. C. O. Matias ⁽³⁾	
.....	174
Sea floor mapping of coastal ecosystems using very high resolution imagery and OBIA classification.....	179
Eduarne Ibarrola-Ulzurrun ⁽¹⁾ , Javier Marcello ⁽¹⁾ , Consuelo Gonzalo-Martin ⁽²⁾	
.....	179
Energy to Water nexus: trying to understand the differences between urban and rural households consumptions	183
C. Matos ⁽¹⁾⁽²⁾ , D. Faria ⁽¹⁾ , I. Bentes ⁽¹⁾⁽²⁾ , A. Briga-Sá ⁽¹⁾⁽²⁾ , S. Pereira ⁽¹⁾⁽²⁾	
.....	183
Analysis of the deformations of a piston with a solid skirt in elastohydrodynamic lubrication of internal.....	188
Mohamed Benbrik ⁽¹⁾ , Miloud Tahar Abbès ⁽¹⁾ , Patrick Maspeyrot ⁽²⁾ , Ahmed Dekkiche ⁽¹⁾ , Mouna Amara ⁽¹⁾	
.....	188
From water to energy: Methodology to characterize, model and measure the reduction of urban and rural domestic consumptions	193
S. Pereira ⁽¹⁾ , C. Matos ⁽¹⁾ , A. Cunha ⁽³⁾ , F. Pereira ⁽⁴⁾ , I. Bentes ⁽¹⁾ , A. Briga-Sá ⁽¹⁾	
.....	193

7th International Congress on Energy and Environment Engineering and management
(CIEEM7)
17-19 July 2017, Canary Islands, SPAIN

An empirical assessment of several machine learning approaches to estimate long-term wind speed conditions

Santiago Díaz^{*(1,2)}, José Antonio Carta⁽¹⁾, Fernando Castellano^(1,2)

⁽¹⁾ *Department of Mechanical Engineering, University of Las Palmas de Gran Canaria, Campus de Tafira s/n, 35017 Las Palmas de Gran Canaria, Canary Island, Spain.*

**Corresponding author. Tel: 627616515 E-mail address: Santiago.diaz103@alu.ulpgc.es*

⁽²⁾ *Department of Renewable Energies, Canary Island Institute of Technology (ITC), Playa de Pozo Izquierdo s/n, 35119, Santa Lucía, Spain.*

1. Introduction – In the scientific literature, an extensive collection of Measure-Correlate-Predict (MCP) methods have been proposed to recognize long-term wind characteristics at target sites for which only measurements recorded over a short-term are available [1]. These methods allow such characteristics to be determined when short-term target site wind speed measurements are available as well as historical data collected at nearby reference stations (commonly from weather stations located at airports or wind farms in the operational phase) where the sampling periods exceed 15 years and can, therefore, be considered as long-term.

One of the most important advances developed in this methodology in recent years has been the use of Machine Learning (ML) techniques. These statistical learning techniques are not only able to determine non-linear relationships between the features [2], but also allow the use of several reference stations which may be able to capture details of the wind resource at a target site that would otherwise be missed if only one reference station were used. This enables details to be captured of the wind resource at the target site that could not be recognized if only one reference station were used [1-3]. The most commonly used data mining technique has been Artificial Neural Networks (ANNs) [1-3]. However, other techniques such as Support Vector Regression (SVRs) have begun to be explored [4]. Other techniques, including Random Forest (RF) [5], show signs of potential use given the results that have been obtained in similar fields of knowledge [6], though RF has not yet been used for this specific purpose.

The main goal of this study is to demonstrate the validity of SVR and RF methods for long-term wind speed estimation using the MCP methodology and considering multiple reference stations. Their efficiency is demonstrated by means of a comparison with the ANN method, currently considered the reference technique [1-2]. All the techniques compared in this study are formulated with the same starting conditions in relation to the explanatory and response variables. The features considered are the mean hourly wind speeds and directions recorded in 2014 at ten weather stations in the Canary Archipelago (Spain). Likewise, with the aim of reducing the dependency of errors obtained when evaluating these techniques, the 10 Folds Cross Validation technique has been used [1-3, 4, 7] and Grid Search was applied to the selection of the different hyper-parameters necessary for the execution of the ML techniques evaluated [4].

The studies undertaken show that RF not only produces a lower error for all the target sites considered (measured using the metrics Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), R-squared (R^2) and Index of Agreement (IoA)), but also requires approximately 35% less computing times for estimation compared to SVR, its potential competitor. In addition, the results confirm that both the SVR and RF techniques allow direct resolution of this kind of problem in which high dimensionality circumstances occur, obtaining a single solution as opposed to the local optima solutions produced with ANN techniques.

2. Experimental – The mean hourly wind speed and wind direction data recorded over the course of 2014 at 10 weather stations (WS) located in the 7 major islands of the Canary Archipelago (Spain) were used for this study (see **Figure 1**). All the wind speed and direction data series were captured at 10 m above ground level and the length of these data series allowed determination of the pattern of seasonal variations of the meteorological data in accordance with the general recommendations related to MCP methods [1]. These data series were provided by the state meteorological agency (Spanish initials: AEMET) of the Ministry of Agriculture, Food and Environment of the Spanish Government (stations identified as WS-1, WS-2, WS-3, WS-6, WS-7, WS-8, WS-9 and WS-10) and by the Canary Technological Institute (Spanish initials: ITC) (stations identified as WS-4 and WS-5).

7th International Congress on Energy and Environment Engineering and management (CIEM7)
17-19 July 2017, Canary Islands, SPAIN

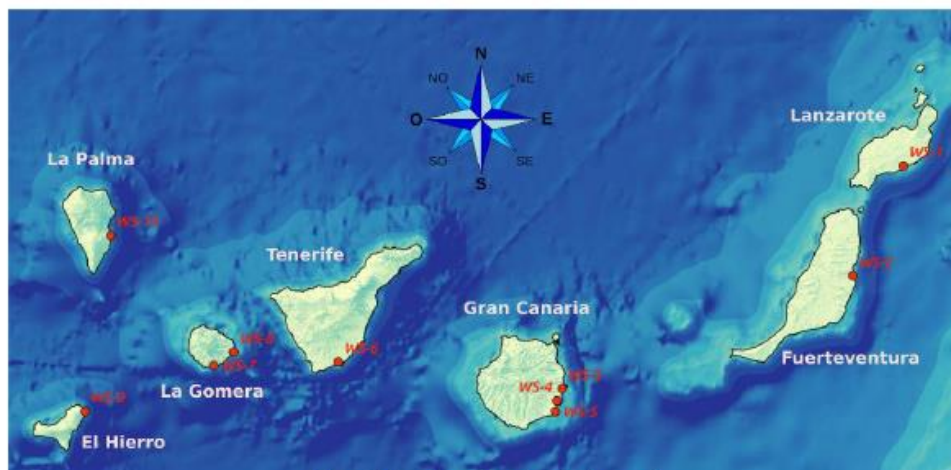


Figure 1. Location of weather stations used.

The regression approaches compared in this study and used to construct the MCP models for target site wind speed estimation were ANN, SVR and RF. The ANN technique assumes a multilayer perceptron (MLP) architecture, which is the most commonly used option to solve this type of problem. It presents a feedforward structure with only one hidden layer composed of multiple neurons (previously selected) and uses a sigmoid activation function. This architecture has been discussed in depth in several papers related to this field of knowledge [2].

The SVR technique uses the formulation ϵ -SVR [8], and has also been widely analyzed by authors including Cristianini and Shawe-Taylor [9] and Hofmann et al. [10]. Likewise, it has previously been used in other studies related to the analysis of long-term wind resources using the MCP methodology [4]. According to its methodology, for the resolution of non-linear problems such as those given in this case, the inputs should be transformed into a hypothetical space where they acquire the condition of linearity [11]. For this process, a positive-definite kernel function is used, which represents the inner product in the feature space [4]. In this work, the Gaussian kernel was chosen since it has been extensively used in the data mining community and has given good results in several fields of knowledge [12].

Implementation of the RF technique was based on the original method proposed by Leo Breiman [13], which is simply an ensemble of regression trees where, for each leaf of each of the regression trees, a methodology known as Classification and Regression Trees (CART) is used [14-15]. In general terms, this method is able to resolve high dimensionality problems with very competitive computational times when compared to other ML techniques. Its formulation and basic principles of operation have been analyzed in great detail by Biau G. and Scornet [15].

Figure 2 shows the general structure of the ML methods that have been implemented and the methodology followed for the development of the comparative analysis carried out in this work. Shown in the top left-hand corner of Figure 2 are the available variables of the reference site (WS-1, ..., WS-10), whose short-term data series of wind velocity (V) and wind direction (D) are used as input data for the three learning approaches that are compared in the present study. The diagram has been particularized for the case in which it is planned to estimate the long-term wind speed conditions for the target site WS-3, using the remaining WSs as references in the training and validation processes of each statistical method. Nevertheless, the same process was also applied considering WS-4 and WS-5 as target sites while the other nine weather stations are still the references. Only the weather stations WS-3, WS-4 and WS-5 were selected as target sites as one of the basic assumptions of the MCP methodology [1] indicates that the meteorological data used as reference data must be representative of the weather conditions of the target site for which the calculations are being made [16]. In this context, Table I shows the linear correlation coefficients between the 10 WSs used in this study. As can be seen, the correlation coefficient only exceeds 0.8 when comparing the time series of WS-3, WS-4 and WS-5. In this way, there is a guarantee that at least one of the references in each model is able to represent the conditions of the target site.

7th International Congress on Energy and Environment Engineering and management (CIEM7)
17-19 July 2017, Canary Islands, SPAIN

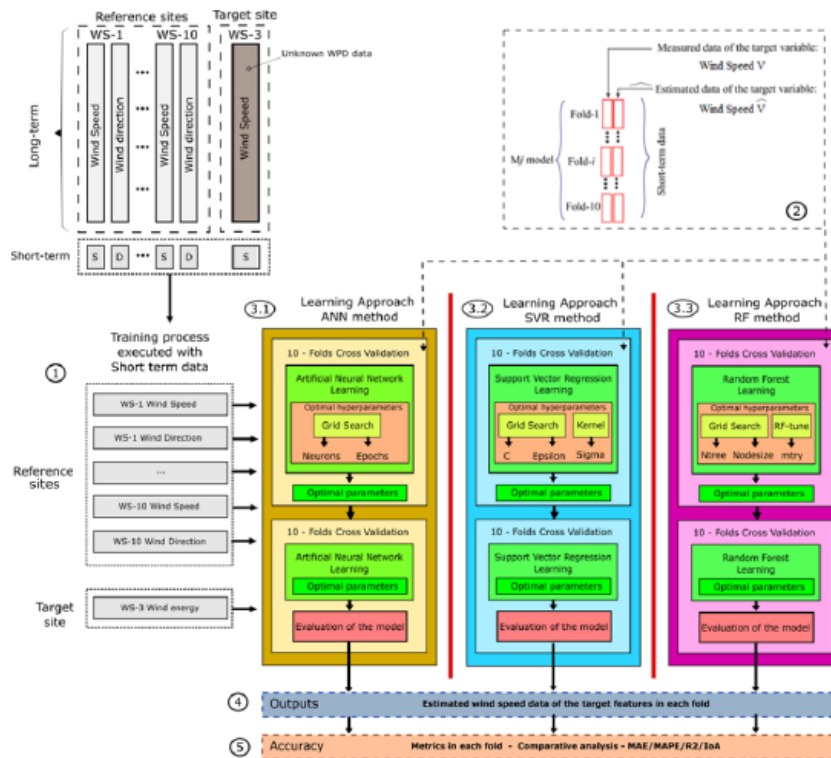


Figure 2. Schematic representation of the methodology developed in this work

The technology used to develop the models is called R Statistics, an open source programming software with an extensive collection of specialized packages created as a result of research carried out by prestigious research teams in multiple fields of knowledge from all parts of the globe. In this case, the mnet [17], Kernlab [18] and randomForest [19] packages were used as starting points in implementation of the ANN, SVR and RF techniques, respectively.

As can be seen in Figure 2, the process can be divided into 5 Steps. In Step 1, the data of the target sites (WS-3, WS-4 or WS-5, respectively) are separated from the remaining WSs which are used as reference stations or explanatory variables. In Step 2, a 10-fold cross-validation procedure is undertaken [2, 4, 7], in which both the target and the reference WSs are subjected to a partition between training and test subsets. In this sense, the data are randomly ordered and divided into 10 folds of equal size using one of the folds as test data and the remaining nine folds as training data. The 10-fold cross-validation process is repeated 10 times until each fold has been considered once as the test subset. Finally, the cross-validation error is obtained as the arithmetic mean of the 10 metric measures calculated.

Step 3 is comprised of two stages as described below. Note that this step is executed separately for each of the tested ML techniques. However, they follow the same methodological procedure outlined in Figure 2. Stage 1 of Step 3 involves the search process for the optimal characteristic hyper-parameters of each of the selected ML techniques. In the case of ANN (Step 3.1), the configured hyper-parameters are the number of neurons in the hidden layer and the maximum number of iterations of the model. Both parameters are defined through the Grid Search method which tests different combinations of these parameters until finding the alternative that produces the smallest error in the data left out in each iteration of the cross-validation process. The Grid Search method is also used to estimate some of

Table I. Linear correlation coefficients between the wind speeds of the Weather Stations

WS	WS-1	WS-2	WS-3	WS-4	WS-5	WS-6	WS-7	WS-8	WS-9	WS-10
WS-1	1.00	0.73	0.74	0.67	0.69	0.54	0.26	0.56	0.52	0.56
WS-2	0.73	1.00	0.68	0.58	0.63	0.55	0.30	0.54	0.58	0.57
WS-3	0.74	0.68	1.00	0.82	0.83	0.53	0.23	0.60	0.51	0.58
WS-4	0.67	0.58	0.82	1.00	0.83	0.51	0.18	0.55	0.45	0.50
WS-5	0.69	0.63	0.83	0.83	1.00	0.63	0.21	0.55	0.42	0.50
WS-6	0.54	0.55	0.53	0.51	0.63	1.00	0.39	0.36	0.31	0.40
WS-7	0.26	0.30	0.23	0.18	0.21	0.39	1.00	0.23	0.24	0.34
WS-8	0.56	0.54	0.60	0.55	0.55	0.36	0.23	1.00	0.56	0.55
WS-9	0.52	0.58	0.51	0.45	0.42	0.31	0.24	0.56	1.00	0.56
WS-10	0.56	0.57	0.58	0.50	0.50	0.40	0.34	0.55	0.56	1.00

of the tested ML techniques. However, they follow the same methodological procedure outlined in Figure 2. Stage 1 of Step 3 involves the search process for the optimal characteristic hyper-parameters of each of the selected ML techniques. In the case of ANN (Step 3.1), the configured hyper-parameters are the number of neurons in the hidden layer and the maximum number of iterations of the model. Both parameters are defined through the Grid Search method which tests different combinations of these parameters until finding the alternative that produces the smallest error in the data left out in each iteration of the cross-validation process. The Grid Search method is also used to estimate some of

7th International Congress on Energy and Environment Engineering and management (CIEM7)
17-19 July 2017, Canary Islands, SPAIN

the SVR hyper-parameters (Step 3.2), in particular the C , ϵ parameters. This technique also requires configuration of the characteristic parameter σ which is required to compute the Gaussian kernel. In this context, the σ parameter is established through the heuristic method implemented in the `sigest()` function in the `kernelab` package [18]. Finally, the definition of three hyper-parameters is required in the case of the RF method (Step 3.3). The number of regression trees and the minimum terminal node size are selected using the Grid Search method, while the number of variables randomly sampled as candidates at each split is defined through another heuristic method implemented in the `randomForest` package, the `tuneRF` function [19]. It should be noted that in this search stage for the optimal hyper-parameters, the authors have followed the recommendations as described in several prestigious publications [2, 15, 20].

Stage 2 of Step 3 involves the actions related to the training process of the learning models. Therefore, the optimal hyper-parameters selected in Stage 1 and the training dataset are used to estimate the long-term target site wind speed according to the patterns recognized with the reference stations. Taking into account use of the 10-fold cross-validation process, at the end of Step 3 there will be 10 independent models (one for each fold) for each ML technique.

In Step 4, each model is fed with the validation subset of the reference stations, estimating a time series of wind speeds for the target site. Based on the time series estimated in Step 4 and the validation subset of the target site (observed wind speed), the accuracy of each method is determined in Step 5. In this case, the metrics Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), R-squared (R^2) and Index of Agreement (IoA) were used. The overall value that defines the performance capacity of each ML technique is obtained by averaging the 10 metrics results of the 10-fold cross-validation process.

The simulations were performed with an Inter® Pentium® CPU B980 PC with base frequency of 2.40 GHz and 4 GB of RAM. The operating system installed in this computer is 64-bit Windows 8.1.

3. Results and Discussion – Shown in **Figure 3** are the general results obtained through the metrics MAE, MAPE, R^2 and IoA when the long-term wind speed conditions at the target sites (WS-3, WS-4 and WS-5) are estimated according to the bases explained in Section 2. **Table II** represents the mean and standard deviations of the metrics for each model and target site.

Firstly, it can be concluded from the results that the SVR and RF techniques show high potential for long-term wind speed estimation at sites where only short-term data are available. In this sense, regardless of the considered target position (WS-3, WS-4 or WS-5) or the metric selected to evaluate the performance of the estimation, SVR and RF gave more accurate estimations than the ANN technique. At the same time, the standard deviations of the error were significantly lower in the SVR and RF techniques if compared to those obtained with the ANN technique. These results support the claim that the SVR and RF statistical models can report an important improvement in estimation of the wind resource at sites where the construction of a wind farm is being considered.

Secondly, the above results confirm that both the SVR and RF techniques allow direct resolution of high dimensionality problems in a more efficient way than the ANN technique [9, 15]. In this context, it can be argued that while the neural networks have many local optima [21], the SVR and RF techniques have more stable optimization algorithms which are able to obtain a unique solution to the formulated problem [9, 15]. Likewise, from a statistical point of view there are fewer problems with the SVR and RF techniques in relation to overfitting and generalization, allowing the detection of outliers which may be overlooked in the training stage.

Of the three evaluated techniques, RF is the most accurate. Even so, the differences with respect to SVR are not significant (**Figure 3**). According to the results presented in **Table II**, the RF technique gives the best MAE, R^2 and IoA metrics for the three WSs considered as target sites. However, when the selected metric is MAPE, the best results are given with the SVR technique, except when the estimation is performed on the WS-5 in which case the RF technique is once again the most accurate alternative.

7th International Congress on Energy and Environment Engineering and management (CIEM7)
17-19 July 2017, Canary Islands, SPAIN

This study demonstrates another important advantage of the RF technique over the selected alternatives, namely its greater facility for model configuration in the learning process. Although configuration of three hyper-parameters is required to execute this model, there is general agreement in the literature about the default values that should be adopted to execute the simulation according to the size and complexity of the problem under study. While there is also general agreement in the literature about the establishment of the characteristic hyper-parameters in the SVR technique, the quality of the estimation is much more dependent on the hyper-parameters that are chosen compared to the RF technique. This aspect makes computing times significantly longer than those required for the RF technique in the learning process.

Regarding computing time, execution of the ANN technique required 13680 seconds on average for the three WSs selected as target sites. In the case of the SVR technique, an average computational time of 22530 seconds was required, and for the RF technique 15120 seconds under the same conditions. It should be noted that these computational times are for model execution without the parallelization strategies that could be applied to the three alternatives under study.

Table II. Summary of results obtained in the comparative analysis. MAE (m/s), MAPE (%), R2 (%) and IoA (%)

TARGET SITE WS-3					
Method	Variable	MAE	MAPE	R2	IoA
ANN	Mean	1.31	33.87%	76.22%	68.68%
	Standard Desv.	0.24	6.32%	7.20%	6.48%
RF	Mean	0.93	23.70%	88.03%	84.66%
	Standard Desv.	0.03	2.94%	0.97%	0.55%
SVR	Mean	0.97	23.06%	87.00%	84.04%
	Standard Desv.	0.03	2.51%	1.16%	0.50%
TARGET SITE WS-4					
Method	Variable	MAE	MAPE	R2	IoA
ANN	Mean	2.05	33.12%	64.75%	63.43%
	Standard Desv.	0.46	9.62%	11.63%	11.39%
RF	Mean	1.30	21.51%	79.53%	83.41%
	Standard Desv.	0.04	0.98%	1.77%	0.54%
SVR	Mean	1.31	19.71%	77.41%	83.24%
	Standard Desv.	0.05	1.04%	2.24%	0.74%
TARGET SITE WS-5					
Method	Variable	MAE	MAPE	R2	IoA
ANN	Mean	1.39	23.80%	73.11%	68.61%
	Standard Desv.	0.26	5.08%	8.88%	8.33%
RF	Mean	0.84	14.05%	89.20%	86.02%
	Standard Desv.	0.02	0.94%	0.68%	0.42%
SVR	Mean	0.89	14.98%	87.57%	85.09%
	Standard Desv.	0.03	1.15%	1.19%	0.66%

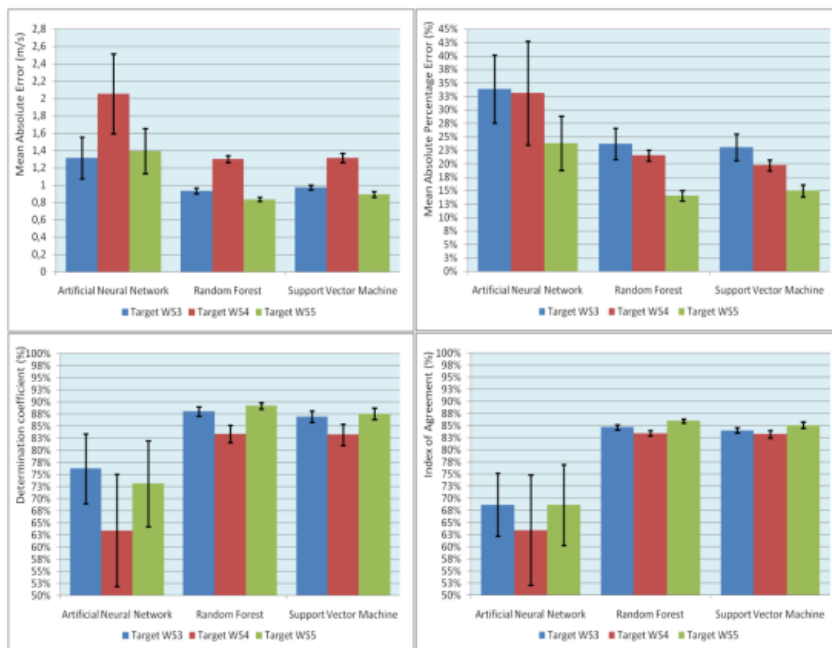


Figure 3. Comparison of MAE (top-left), MAPE (top-right), R2 (bottom-left) and IoA (bottom-right) when estimating the wind speed with each technique and target station.

4. Conclusions – According to the results of the analysis, although the trend nowadays is towards using ANNs in implementation of the MCP methodology, there are alternatives that display greater precision and efficiency in the estimation of the long-term wind speed at a target site.

7th International Congress on Energy and Environment Engineering and management (CIEEM7)
17-19 July 2017, Canary Islands, SPAIN

Random Forest gave the best results for all target sites when measuring the performance with the MAE, R^2 and IoA metrics. Furthermore, when measuring the performance with the MAPE metric, it is unclear which is the best alternative given that the SVR technique was the best option for target sites WS-3 and WS-4 but in the case of WS-5 the best result was obtained with the RF technique. Similarly, the differences between the ANN and the other evaluated techniques are reasonably significant, as shown by the value of the R^2 metric which is at least 10% lower than the values obtained with the SVR and RF techniques for the three target sites.

Finally, the computing times required for the RF method were around 35% lower than those required by its main competitor, the SVR technique. This difference in computing requirements is based on its greater facility to obtain the optimal hyper-parameters for execution of the model.

5. References

- [1] Carta JA, Velázquez S, Cabrera P. A review of measure-correlate-predict (MCP) methods used to estimate long-term wind characteristics at a target site. *Renewable and Sustainable Energy Reviews* 2013; 27:362-400.
- [2] Velázquez S, Carta JA, Matias JM. Comparison between ANNs and linear MCP algorithms in the long-term estimation of the cost per kWh produced by a wind turbine at a candidate site: A case study in the Canary Islands. *Applied Energy* 2011; 88:3869-81.
- [3] Carta JA, Cabrera P., Matias JM, Castellano F. Comparison of feature selection methods using ANNs in MCP-wind speed methods. A case study. *Applied Energy* 2015; 158:490-507.
- [4] Díaz S, Carta J.A, Matias JM. Comparison of several measure-correlate-predict models using support vector regression techniques to estimate wind power densities. A case study. *Energy Conversion and Management*. 2017.
- [5] Breiman L. Random Forest. *Machine Learning* 2001; 45:5-32.
- [6] Lin Y, Kruger U, Zhang J, Wang Q, Lamont L, Chaar LE. Seasonal analysis and prediction of wind energy using Random Forests and ARX model structures. *IEEE Transactions on Control Systems Technology*. Vol. 23, No. 5, pp. 1994-2002, 2015.
- [7] Witten IH, Frank E, Hall MA. *Data Mining. Practical Machine Learning Tools and Techniques*. 3rd ed.: Morgan Kaufmann 2011.
- [8] Vapnik V, Golowich S, Smola A. Support Vector Method for function approximation, regression estimation and signal processing. *Advances in Neural Information Processing Systems* 1998:281-287.
- [9] Cristianini N, Shawe-Taylor J. *An introduction to support vector machines and other kernel-based learning methods*. Cambridge, United Kingdom: Cambridge University Press, 2000.
- [10] Hofmann T, Scholkopf B, Smola AJ. Kernel Methods in Machine Learning. *The Annals of Statistics* 2008; 36:1171-1220.
- [11] Hammer B, Gersmann K. A Note on the Universal Approximation Capability of Support Vector Machines. *Neural Processing Letters* 2003; 17:43-53.
- [12] Schölkopf B, Smola AJ. *Learning with kernels*. MIT Press 2002.
- [13] Breiman L. Random Forest. *Mach Learn* 2001;45:5-32.
- [14] Cutler A, Cutler R, Stevens JR. Chapter 5. Random Forest. In: Zhang C, Yunquian M, editors. *Ensemble Machine Learning: Method and Applications*: Springer; 2012, p. 157.
- [15] Biau G, Scornet E. A random forest guided tour. *Springer* 2016:197-227.
- [16] Langreder W. Wind resource and site assessment. In: Tong W, editor. *Wind power generation and wind turbine design*. MA: WIT Press 2010:49-87.
- [17] Ripley B, Venables W. Package nnet. Feed-Forward Neural Networks and Multinomial Log-Linear Models. 2015. <https://www.r-project.org/> (accessed 20.05.2017).
- [18] Karatzoglou A, Smola A, Hornik K, Zeileis A. Package Kernlab Version 0.9-22. An S4 Package for Kernel Methods in R. *Reference Manual. Journal of Statistical Software* 2015; 11:1-20.
- [19] Breiman L, Cutler A. Breiman and Cutler's Random Forests for Classification and Regression. Package randomForest. Project. R Statistics. 2015. <https://www.r-project.org/> (accessed 20.05.2017).
- [20] Sewell M. Support Vector Machines (SVMs). <http://www.svms.org/> 2015 (accessed 20.05.2017).
- [21] Bishop CM. *Pattern Recognition and Machine Learning*. Springer. 2016

This paper has been co-funded by FEDER funds, INTERREG MAC 2014-2020 programme, within the ENERMACH project (MAC/1.1a/117).



International Congress

Energy and Environment Engineering and Management

17-19 July 2017
CANARY ISLANDS/SPAIN



Eduardo Manuel Cuerda Correa, Chairman of the 7th International Congress of Energy and Environment Engineering and Management (CIIEM7)

CERTIFIES

That SANTIAGO DIAZ RUANO

presented in the aforesaid Conference the oral communication entitled

AN EMPIRICAL ASSESSMENT BETWEEN SEVERAL MACHINE LEARNING APPROACHES TO ESTIMATE LONG-TERM WIND SPEED CONDITIONS.

Las Palmas, 19th July, 2017

CERTIFICATE OF PARTICIPATION





7th International Congress
Energy and Environment Engineering and Management

(CIEM7)

17-19 July 2017

Canary Islands, SPAIN



ABSTRACTS
B O O K

7th International Congress on Energy and Environment Engineering and Management (CIEEM7)
17-19 July 2017, Canary Islands. SPAIN

Carlos Hernández ⁽¹⁾ , Sergio Velázquez ⁽¹⁾ , Julieta Schallenberg ⁽¹⁾	88
Study connective capabilities of solid residues from the waste incineration	89
V. Blahuskova, J. Vlcek, A. Prysycz.....	89
Biosorption of Lead (II) and Nickel (II) from synthetic wastewater using cyanobacterial biomass	90
Upasha Sharma ⁽²⁾ , Sushovan Sen ^{(1)*} , Kalyan Adhikary ⁽²⁾ , Susmita Dutta ⁽¹⁾	90
Degradation of Organic Micropollutants using a Hybrid Bioreactor	91
C. Grandclément ^(1,2,3) , A. Píram ⁽²⁾ , I. Seyssiecq ⁽³⁾ , G. Vanot ⁽¹⁾ , N. Tiliacos ⁽¹⁾ , N. Roche ⁽³⁾ , P. Doumenq ⁽²⁾	91
Adjustment of wind maps through data recovered in weather stations.....	92
Fernando Castellano* ^(1,2) , Santiago Díaz ^(1,2) , José Antonio Carta ⁽¹⁾	92
Treatment of wastewater from an Ecuadorian printing industry with Fenton reactions and organic biofilters	93
D. L. Tinoco ² , J. A. Herrera-Melián ¹ , J. A. Ortega Méndez ¹ , G.F. Torres ² , J. Araña ¹ , E. D. Marin ² and J. M. Doña-Rodríguez ¹	93
Thermal Resources of the Climate of West Polesie, Belarus.....	94
A. Meshyk, M. Sheshka, M. Barushka.....	94
River Flow in Belarusian Polesie and its Climate Change Adaptation	95
A. Volchak, A. Meshyk, An. Vouchak	95
Assessment of Waterpower Potential of Rivers in the Yaselda Catchment.....	96
A. Volchak, An. Vouchak, M. Barushka	96
Management of slurry in Gran Canaria with full-scale Natural Treatment Systems for Wastewater (NTSW). One year’s experience in livestock farms.	97
C.A. Mendieta Pino ⁽¹⁾ , S.O. Pérez Báez ⁽²⁾ , A. Ramos Martín ⁽³⁾ , S. Brito Espino ⁽⁴⁾	97
Methods of Environmental Flow Assessment	98
M. Sheshka	98
The Industrial Emissions Directive implementation in Poland – contaminants identification – case study for research and small tonnage production plant „IChN”	99
A. Paszek, D. Łuczowska, J. Gluzińska.....	99
Characterization of surge phenomenon by the temperature tracking to improve efficiency in power plants turbochargers	100
J. Valencia ¹ , D. Echeverría ¹ , A. Ramos ² and V. Henríquez ²	100

7th International Congress on Energy and Environment Engineering and Management (CIEM7)

17-19 July 2017, Canary Islands. SPAIN

results, such as the removal of diclofenac up to 60%, have been observed. The efficiency of such a prototype still have to be investigated in real environmental operating conditions. Once the optimized conditions found, the scale up of an industrial pilot will allow the validation of such a tertiary process for a wastewater treatment plant with significant micropollutant discharge.

5. References

[1] C. Grandclément, I. Seyssiecq, A. Piram, P. Wong-Wah-Chung, G. Vanot, N. Tiliacos, N. Roche, P. Doumenq, *Water Res.*, **111**, p. 297-317, (2017)

Adjustment of wind maps through data recovered in weather stations. Case study: Canary Islands

Fernando Castellano*^(1,2), Santiago Díaz^(1,2), José Antonio Carta⁽¹⁾

⁽¹⁾ *Department of Mechanical Engineering, University of Las Palmas de Gran Canaria, Campus de Tafira s/n, 35017 Las Palmas de Gran Canaria, Canary Island, Spain.*

⁽²⁾ *Department of Renewables Energies, Canary Island Institute of Technology (ITC), Playa de Pozo Izquierdo s/n, 35119, Santa Lucía, Spain.*

*Corresponding author. Tel: 626459983; e-mail: fcastellano@itccanarias.org

1. Introduction – Conventional large-scale wind resource exploration methods require expensive measurement campaigns and the installation of a large number of meteorological masts in a homogeneous survey network. In addition, conventional models of wind resource forecasting are not accurate in complex terrain areas, as it is the case of Canary Islands. Modern mesoscale and microscale modelling techniques offer a very effective solution to these problems as they effectively combine the use of a sophisticated atmospheric simulation model, capable of reproducing large-scale wind patterns, with a microscale wind model that responds to the characteristics of the terrain and the topography.

In 2006, the Canary Islands Government with the technical advice of ITC developed a Wind Resource Map [1] of each island of the archipelago. The map was based on the MASS meteorological model (Mesoscale Atmospheric Simulation System), developed by the Company AWS-TRUEPOWER [2]. With this work it was possible to have fairly reliable wind data for sites of interest without carrying out specific measurement campaigns.

In 2016, a review of the wind atlas was carried out using an improved mesoscale model. Numerous weather stations of the archipelago were also used for the validation and adjustment of the wind atlas.

2. Experimental – The maps produced according to the methodology described in section 1 have been adjusted with surface met masts located in different areas of the archipelago. To do this, filtering and quality control of the raw data were performed to ensure the results accuracy of the adjustment processes. In a second phase, the met masts wind speed was adjusted to long-term using the MCP (Measure-Correlate-Predict) methodology [3]. Finally, the new wind atlas was calibrated by integrating the long-term tuned stations into the maps using the OpenWind software.

3. Results and Discussion – Original and adjusted maps accuracy has been checked by calculating wind farm production where data of farms are available. **Table I** shows a sample of the results obtained in the comparison between the wind farm production results calculated with the adjusted maps and the original versions.

Metrics	Original map	Adjusted map
MAPE	11,6%	9,5%
RSquared	93,8%	96,0%

Table I. Example of some results obtained

C. Valoración externa del trabajo de investigación



Doctorado en Ingenierías
Química, Mecánica y de Fabricación

E3. REVISIONES PERIÓDICAS DEL TRABAJO DE INVESTIGACIÓN CON EVALUADORES EXTERNOS

Nombre y apellidos del doctorando	Santiago Díaz Ruano	
Nombre y apellidos del director/es de tesis	José Antonio Carta González José María Matías Fernández	
Nombre y apellido de investigador Dr. evaluador externo	Marie-Laure Nivet	
Universidad/centro y lugar del evaluador externo	University of Corsica, France	
Fecha de realización de la evaluación	07/2018	
Lugar/modalidad de la evaluación (marque lo que proceda)	Presencial (Si/No): No Lugar: -	Online(Si/No): Si

Breve informe del investigador evaluador externo sobre el desarrollo de la tesis doctoral:

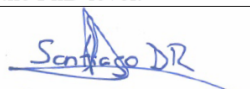
Santiago Diaz Ruano will present his PH.D. work in September 2018. His research concerns the estimation of energy produced by wind farms in long-term conditions. He proposes to use a methodology called Measure-Correlate-Predict (MCP) based on Machine Learning Techniques to estimate Wind Power Speed, Wind Power Density and finally Wind Turbine Power Output of sites where measurements have been recorded over a short period.

His work is based on a clear and rigorous study of the actual scientific literature. He tests and compares different MCP models based on three machine learning techniques which are Artificial Neural Networks (ANN), Support Vector Regression (SVR) and Random Forest (RF). He proposes a methodology to optimise and to compare the models demonstrating his mastery of the subject.

The results of his studies have been presented in July 2017 in the framework of the International Congress 7th International Congress of Energy and Environment Engineering and Management (CIIEM7). Furthermore he published with his co-authors two articles in high-quality peer-reviewed international journals with high impact factor (Energy Conversion and Management (IF : 6.377) and Applied Energy (IF : 7.900)) attesting the interest of the community for his research.

He actively participates to the Horizon 2020 project Technology Innovation for the Local Scale (TILOS) in which he develops short term wind power forecasting models, proving his work can be applied in an operational context.

All these facts show that the serious and consequent work done by Santiago Diaz Ruano is at the PhD level.


SANTIAGO DÍAZ RUANO

Fdo: Estudiante de doctorado

Fdo: Director/es de tesis



Fdo: Investigador externo

CARTA
GONZALEZ JOSE
ANTONIO -
42905050E

Firmado digitalmente por CARTA
GONZALEZ JOSE ANTONIO - 42905050E
Número de reconocimiento (DN): c=ES,
serialNumber=42905050E, sn=CARTA
GONZALEZ, givenName=JOSE ANTONIO,
cn=CARTA GONZALEZ JOSE ANTONIO -
42905050E
Fecha: 2018.07.19 13:23:57 +01'00'

La presente tesis doctoral fue terminada de escribir el día 22 de Julio de
2018 en Las Palmas de Gran Canaria