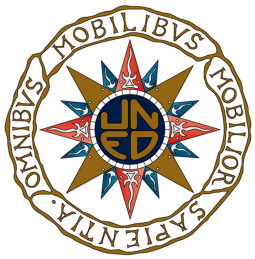


UNIVERSIDAD NACIONAL DE EDUCACIÓN A DISTANCIA

ESCUELA TÉCNICA SUPERIOR DE INGENIERÍA INFORMÁTICA

PROYECTO FIN DE MÁSTER

MÁSTER UNIVERSITARIO EN INGENIERÍA DE SISTEMAS Y DE CONTROL

Modelo de predicción de la radiación solar con aplicación en control automático

Autor:
Rafael Mena Yedra

Directores:
José Sánchez Moreno
José Luis Guzmán Sánchez

Madrid, Septiembre 2014

MÁSTER UNIVERSITARIO EN INGENIERÍA DE SISTEMAS Y DE CONTROL

Título del proyecto:
Modelo de predicción de la radiación solar con aplicación en control automático

Tipo de proyecto:
Proyecto específico propuesto por el alumno

Autor:
Rafael Mena Yedra

Directores:
José Sánchez Moreno
José Luis Guzmán Sánchez

Agradecimientos

Quisiera agradecer a todas las personas que a lo largo de mi carrera de vida me han ido enseñando y empujando para ir hacia delante desde ya pequeño, comenzando por mi madre. A todos mis profesores, compañeros, amigos y entorno porque de todo se va aprendiendo. En especial a los buenos profesores durante mi etapa en la Universidad de Almería, quienes conseguían con su manera de transmitir que mantuviera y creciera mi motivación e ilusión por lo que hacía. Y muy especialmente a Paco Rodríguez y José Luis Guzmán por brindarme la oportunidad de conocer y adentrarme en este mundo de la investigación y darme el apoyo necesario, por permitir que pudiera trabajar y desarrollarme aportando motivación y siempre aprendiendo, porque nunca hay que dejar de hacerlo.

A todos ellos, gracias

Resumen

En este trabajo se ha pretendido abordar el problema de la predicción de la radiación solar tanto a corto plazo (horas) como a medio plazo (días) desde un punto de vista de la inteligencia artificial y el análisis de datos, y con la intención de dar solución a un problema común en el campo del control automático como es la predicción de las perturbaciones. En especial, la radiación solar es un factor crítico y limitante para muchos sistemas basados en este recurso.

En el caso de los fotobiorreactores de microalgas, conocer esta variable y su dinámica es fundamental para plantear estrategias avanzadas de control predictivo que optimicen los recursos y aumenten la productividad. A partir de una base de datos recolectada de una planta real de fotobiorreactores, se plantea una combinación de estrategias basadas en el análisis de datos y el aprendizaje automático para el desarrollo de modelos predictivos a corto y medio plazo de la radiación solar.

Más concretamente, los modelos se basan en las máquinas de vectores de soporte (SVM) para la tarea de aprendizaje automático y predicción, así como en la teoría de inmersión para el análisis de series temporales de datos no lineales y la reconstrucción del espacio de estado utilizando optimización global mediante algoritmos genéticos (GA) para el ajuste de parámetros. Así mismo, se ha planteado la utilización de algoritmos de clustering para la generación de perfiles diarios de radiación solar como base del planteamiento de predicción a medio plazo. Al tratarse de modelos basados en datos y aprendizaje automático, también se ha empleado otra variable que juega un papel directo en la radiación solar terrestre: la nubosidad, para ello se ha planteado una estrategia para caracterizar el conjunto de datos histórico así como crear el soporte para poder emplear predicciones externas de nubosidad de fuentes externas cuando se utilice el modelo en tiempo real.

Los resultados obtenidos de la predicción a corto plazo son favorables con un CV-RMSE en la predicción a 1 hora que oscila entre el 6 % y el 24 % para días de diferente estación y perfil de nubosidad, y utilizando datos no incluidos en la calibración del modelo. Para el caso de la predicción a corto plazo de 12 horas se han obtenido resultados con un ligero incremento en el error, pero similar en el sentido de que captan bien la dinámica de la evolución de la radiación solar. El error (CV-RMSE) para las pruebas de predicción a 12 horas oscila entre el 11 % y el 35 %. Por otro lado los resultados de la predicción a medio plazo son muy prometedores al conseguir predecir adecuadamente el perfil de los siguientes días, haciendo pruebas hasta 14 días, mediante la propuesta de obtener perfiles de radiación. El CV-RMSE para estos dias oscila entre el 7 % y el 35 %.

Palabras clave: *Predicción de radiación solar, series temporales no lineales, máquinas de vectores de soporte, clustering, algoritmos genéticos, control automático.*

Abstract

In this paper it has been tried to address the problem of prediction of solar radiation in the short term (hours) and medium term (days) from the point of view of the artificial intelligence and data analysis, and with the intention to solve a common problem in the field of automatic control as the disturbances prediction. In particular, solar radiation is a critical limiting factor for many systems based on this resource.

For microalgae photobioreactors, knowing the value and evolution of this variable is basic for implementing advanced predictive control strategies which optimize resources and increase productivity. A combination of strategies based on data analysis and machine learning are developed to obtain predictive models of the solar radiation in the short and medium term through the use of a database collected from a real plant of photobioreactors.

More specifically, the models are based on support vector machines (SVM) for the task of machine learning, as well as immersion theory for the analysis of nonlinear time series data and the reconstruction of the state space using global optimization through genetic algorithms (GA) for finding the optimal parameters. It also has been proposed the use of clustering algorithms to generate daily profiles of solar radiation as a basis for the prediction approach in the medium term. Since these models are data-driven and they are based on machine learning, it has been also employed another variable that plays a direct role in the terrestrial solar radiation: a cloudiness index. It has been proposed a strategy to characterize the data set historical and create the platform to use cloudiness predictions from external sources when the model is used in real time.

The results of short-term prediction are favorable with a CV-RMSE in 1-hour prediction ranging from 6 to 24% for days of different season and with different profile of cloudiness. Data used for tests were not included in the calibration of the model. In the case of short-term prediction of 12 hours, results were obtained with a slight increase in the error, but similar in the sense that it recognizes appropriately the dynamics of the solar radiation evolution. The error (CV-RMSE) obtained in the 12-hours prediction tests ranges from 11% to 35%.

Furthermore, results of medium-term prediction are promising to adequately predict the profile of the next few days, with testing up to 14 days, by proposing the generation and use of profiles of radiation. The CV-RMSE for these days ranges from 7% and 35%.

Keywords: Solar radiation prediction, nonlinear time-series, support vector machines, clustering, genetics algorithms, automatic control.

Nomenclatura

AAE	Agencia Andaluza de la Energía
AEMET	Agencia Estatal de Meteorología
AMV	Atmospheric Motion Vector fields Campos de vectores de movimiento atmosférico
ANN	Artificial Neural Network Red neuronal artificial
ARIMA	Auto-Regressive Integrated Moving Average Modelo autorregresivo integrado de media móvil
DFT	Discrete Fourier Transform Transformada discreta de Fourier
GA	Genetic Algorithms Algoritmos genéticos
kNN	k-Nearest Neighbors algorithm Algoritmo de los k-vecinos más cercanos
LS-SVM	Least-Squares Support Vector Machines Máquinas de vectores de soporte mediante mínimos cuadrados
MAPE	Mean Absolute Percentage Error Error medio absoluto porcentual
MAE	Mean Absolute Error Error Absoluto Medio
MSE	Mean Squared Error Error Cuadrático Medio
MPC	Model Predictive Control Control predictivo basado en modelo
NMBE	Normalised Mean Bias Error Error Medio Normalizado del BIAS

NWP	Numerical Weather Prediction Predicción meteorológica numérica
RMSE	Root-Mean-Square Error Raíz cuadrada del error cuadrático medio
R^2	Coefficient of determination Coeficiente de determinación
SCADA	Supervisory Control and Data Acquisition Supervisión, control y adquisición de datos
SVM	Support Vector Machines Máquinas de soporte vectorial
SVR	Support Vector Regression Máquinas de soporte vectorial para regresión

Índice general

Índice de figuras

Índice de tablas

1 INTRODUCCIÓN

1.1. Motivación para el desarrollo del proyecto

Los sistemas basados en energías renovables y mitigación del efecto invernadero son una de las principales preocupaciones de este siglo. Se están haciendo grandes esfuerzos en todo el mundo tratando de buscar recursos limpios y nuevas tecnologías para hacer frente a estas cuestiones [1] y durante los últimos años, muchas contribuciones han surgido en busca de diversificar el suministro de energía mediante el desarrollo de tecnologías energéticas más limpias y más eficientes para aumentar sustancialmente la cuota renovable (solar, eólica, olas, biomasa,...) del suministro de energía mundial. Por lo tanto, desde finales de 2004 hasta el 2009, la capacidad mundial de energía renovable creció a tasas del 10-60 % anual para muchas tecnologías [2].

La biotecnología es una de las áreas emergentes que puede contribuir ampliamente a los dos desafíos mencionados anteriormente: nuevos recursos renovables y la reducción de la contaminación industrial, así como producir productos de alto valor haciendo uso de las células microbianas, vegetales y animales, así como componentes de éstas. En particular, las microalgas se consideran con un gran potencial para la producción de biocombustibles en el futuro. La biomasa procedente de las microalgas puede alcanzar hasta el 80 % del peso seco bajo ciertas condiciones de estrés, se pueden cultivar con una gran productividad por área a diferencia de otros cultivos, algunas cepas tienen un alto contenido de lípidos, se requiere un consumo bajo de agua y es posible su producción en terrenos áridos [3, 4]. Al mismo tiempo, las microalgas ayudan a la mitigación del CO_2 mediante la absorción de éste. Estas ventajas colocan a las microalgas en una buena posición para la producción de energía renovable a gran escala. Un claro ejemplo de esto es que varias empresas petroleras, como Exxon, BP, Chevron, Shell, Neste Oil, Repsol YPF y Acciona (que participa en el marco del proyecto MACROBIO) están invirtiendo en la investigación de microalgas con fines energéticos [4].

Actualmente, la mayoría de las aplicaciones de microalgas se realizan a pequeña escala y se centran en la obtención vitaminas, cosméticos y alimentación, con sólo alrededor de 10000 toneladas de biomasa seca por año [3]. Por lo tanto, se requiere un cultivo de microalgas a gran escala para aumentar la producción de biomasa con el fin de ser utilizado como una fuente de energía renovable. Pero como cualquier sistema biológico, este tipo de procesos son inmensamente complejos, debido a la presencia de organismos vivos y la alta complejidad de las reacciones metabólicas en las que los microorganismos están involucrados. Por lo tanto, desde un punto

de vista de ingeniería, los sistemas de producción en biotecnología no podrían ser diseñados y controlados como otros procesos en muchos otros tipos de industria. Con el fin de aprovechar los beneficios del diseño y control de procesos basado en modelo ya experimentado en la mayoría de otras industrias, se deben hacer esfuerzos considerables en el modelado, simulación y control. Los procesos basados en microalgas implican nuevos retos para el modelado y el control dado que aparte de las clásicas características no lineales y complejas que caracterizan la mayoría de procesos biotecnológicos, el comportamiento no estacionario permanente, la presencia de perturbaciones junto con una fuerte realimentación desde el nivel de la población hasta el nivel celular a través de la atenuación de la luz hacen estos procesos aún más desafiantes [3]. Por otra parte, las microalgas se cultivan por lo general en condiciones al aire libre, y por lo tanto estos organismos crecen en condiciones cambiantes permanentes siendo sometidas a variaciones de luz y temperatura constantemente.

Por tanto también se debe hacer frente a otra serie de desafíos involucrados en el proceso y que hacen factible el objetivo global mencionado dentro del marco general de mejorar la eficiencia, productividad y optimización de los procesos de microalgas a gran escala mediante un adecuado modelado y técnicas avanzadas de control basadas en modelo o eventos para obtener una biomasa de alta calidad, mitigar el CO_2 y reducir la energía al mismo tiempo creando un entorno casi óptimo para el crecimiento, multiplicación y producción de biomasa.

Entre ellos, se requiere una mejor predicción de la productividad para una población de microalgas sometido a las variaciones de alta frecuencia de la luz. Sólo existen unos pocos estudios sobre la optimización de la productividad de la biomasa frente a la luz fluctuante [3]. La razón principal es la falta de modelos adecuados y estrategias de control que describen la fotoadaptación de las microalgas a una luz fluctuante, y que debe ser tenido en cuenta para mejorar la optimización del fotobiorreactor [5,6].

Por otro lado, existen muchos sistemas en los que la estimación de las perturbaciones es necesaria para mejorar el rendimiento general del sistema de control. Algunos ejemplos se pueden encontrar en los sistemas de control que utilizan la radiación solar como fuente de energía principal, tal como en sistemas de energía renovables basados en la energía solar [7] o cultivos en invernadero [8]. Este es también el caso de los fotobiorreactores, donde la luz solar se utiliza a menudo como la fuente de energía para el cultivo de microalgas [9], y este enfoque tiene numerosas ventajas. En primer lugar, la luz solar es gratuita, mientras que las fuentes de luz artificial son muy caras. En segundo lugar, la energía solar contiene todo el espectro de energía de la luz, y puede proporcionar una longitud de onda de absorción adecuada tanto para el crecimiento celular como para la producción de biomasa [10]. Sin embargo, el rendimiento de estos fotobiorreactores abiertos al aire libre es por lo general difícil de controlar, debido a los ciclos de día y noche y la variación diurna de la intensidad de la luz. Para explotar eficazmente el potencial comercial de las algas, es necesario un uso barato, duradero, fiable y altamente eficiente de la fuente de luz [11].

En definitiva, el problema de control relativo a la producción de biomasa de mi-

croalgas en fotobiorreactores a gran escala está compuesta por diferentes niveles de control, ya que, dependiendo del uso final de la biomasa resultante, diferentes objetivos de control tendrán que cumplirse durante el proceso de producción. Por lo tanto la existencia de diferentes objetivos (productividad, problemas económicos, ambientales, aspectos de calidad, etc) generan un problema de control jerárquico multi-escala que puede ser abordado mediante diferentes técnicas de control, incluyendo el control predictivo basado en modelo (MPC) no lineal multi-escala, así como enfoques de control y muestreo basados en eventos [12]. Por lo tanto, es necesario desarrollar modelos, estimadores y predictores, de las variables de concentración de biomasa, las variables ambientales y el resto de las variables de proceso asociados.

El resultado del proyecto será validado a través de diferentes instalaciones industriales de fotobiorreactores, disponibles en la Estación Experimental de la Fundación Cajamar situada en Almería, en el sudeste de España. Se utilizarán fotobiorreactores tubulares y fotobiorreactores verticales planos.

1.2. Objetivos del proyecto

El objetivo principal de este trabajo es diseñar e implementar un modelo para la predicción a corto plazo (1-12 horas) y a medio plazo (1-14 días) de la radiación solar y su aplicación en una planta real con diferentes tipos de fotobiorreactores de microalgas.

Cualquiera que sea el fotobiorreactor usado, la producción fotosintética de las algas siempre está acompañada por la producción de oxígeno y la absorción de dióxido de carbono. Este hecho provoca alteraciones en el medio de cultivo y que el pH cambie constantemente. Los niveles de oxígeno por encima del nivel de saturación del aire pueden inhibir la fotosíntesis en muchas especies de algas, incluso si la concentración de dióxido de carbono se mantiene en niveles elevados. Además, los niveles elevados de oxígeno combinado con altos niveles de irradiación pueden conducir a una grave foto-oxidación [13]. Por esta razón, es preciso disponer de modelos de predicción de las variables ambientales, y especialmente la radiación solar, para optimizar el rendimiento del proceso y con fines de alimentación directa y de compensación al controlador MPC.

De entre las múltiples técnicas de predicción existentes [14], se ha optado por utilizar máquinas de soporte vectorial (SVM). Las SVM se han introducido hace relativamente poco como técnica de aprendizaje automático con un fuerte trasfondo estadístico [**?**, 15], y precisamente por ello es una técnica más robusta en diferentes ámbitos especialmente cuando existen datos con ruido presente, e incluyen aspectos y técnicas de aprendizaje automático, estadística, análisis matemático y optimización convexa. Las SVMs se han aplicado con éxito en tareas de clasificación [16–18], regresión [19, 20] y predicción [21–23], siendo capaces de modelar sistemas multivariable y no lineales como es el caso de la radiación solar [24]. Además, son capaces

de manejar conjuntos de datos de entrenamiento pequeños y a menudo alcanzan una mayor precisión de clasificación que otros métodos tradicionales [25]. El principio subyacente que beneficia a las SVMs es el proceso de aprendizaje que se conoce como minimización del riesgo estructural, de esta forma las SVMs minimizan el error en aquellas observaciones que van a predecirse sin hacer suposiciones previas sobre la distribución de probabilidad de los datos, mientras que otras técnicas normalmente asumen que la distribución de los datos se conoce a priori lo cual no ocurre en los datos de radiación solar que suelen presentarse como series temporales y no hay suposiciones estadísticas fiables que se puedan emplear. También hay otra característica atractiva de las SVMs y tiene que ver con lo que muchos autores denominan sobreajuste o sobreentrenamiento (*overfitting*) [26], otros lo refieren como compensación sesgo-varianza [27] o control de capacidad [28] y que consiste en que el modelo aprende muy bien el conjunto de entrenamiento pero pierde la capacidad de generalización en nuevas observaciones, en este caso las SVMs son capaces de alcanzar un balance entre la precisión obtenida en un conjunto finito de datos de entrenamiento y la habilidad para generalizar ante nuevas observaciones. Aunque no todo son beneficios, también hay desafíos y el mayor quizás tenga que ver con la elección de algunos parámetros y que como se verá más adelante se intentará hacer frente a ello utilizando diferentes propuestas.

Por tanto, para el desarrollo del trabajo y lograr el objetivo final de disponer de una herramienta para la predicción de la radiación solar será necesario resolver una serie de subobjetivos:

- Adquisición de los datos y su preprocesamiento para tratar los datos con el objetivo de mejorar la calidad de los datos al eliminar valores anómalos, eliminar el ruido de la señal, recuperar huecos en los datos o realizar una normalización de los valores. Posteriormente se realizará la división del conjunto de datos disponible en un conjunto para el entrenamiento del modelo y otro de validación.

- Desarrollo del marco teórico de las SVMs sobre el que se asienta el modelo propuesto. Estudio de su aplicación en la predicción a corto plazo de la radiación solar así como la propuesta de optimización mediante algoritmos genéticos (GA) para superar alguna de las dificultades inherentes de la técnica y la aplicación.

- Propuesta y desarrollo de un modelo basado en SVMs para la predicción a medio plazo de la radiación solar y utilizando técnicas de *clustering* o agrupamiento como apoyo para la obtención de diferentes perfiles diarios de radiación.

- Desarrollo de un módulo para la obtención de datos de predicción de la nubosidad provenientes de fuentes meteorológicas externas para incorporarlo al modelo para la predicción a corto y medio plazo.

- Validación de los resultados.

1.3. Contexto

Este trabajo se ha realizado en el marco del proyecto “Modelado, simulación, control y optimización de fotobiorreactores” (CICYT DPI 2011-27818-C02-01/02) financiado por el Ministerio de Ciencia e Innovación.

En este proyecto, se va a desarrollar un modelo predictivo de la radiación solar en el ámbito de las instalaciones de producción de microalgas de la “Estación Experimental Las Palmerillas”, propiedad de la Fundación Cajamar (+36°47′35,7426”, −2°43′12,0606” - Almería, España), y cuya actividad está estructurada en torno a cuatro líneas de trabajo fundamentales: tecnología de invernaderos, fruticultura subtropical mediterránea, biotecnología y sostenibilidad. En concreto, dentro de las instalaciones de este centro tecnológico, el modelo predictivo será aplicado en dos tipos de fotobiorreactores: tubulares y verticales planos (Figura 1.1).

Tradicionalmente, las microalgas han sido cultivadas en fotobiorreactores abiertos como los llamados *open raceways* por la simplicidad y el bajo coste de este diseño, sin embargo este tipo de fotobiorreactores solo permiten un control reducido de las condiciones de operación y además los cultivos se pueden contaminar fácilmente. Por otra parte, existen fotobiorreactores cerrados de tipo tubular en los que el cultivo permanece reproducible y libre de contaminación [29] permitiendo la elaboración de productos de alto valor derivados de las cepas de microalgas que no pueden ser mantenidas en un entorno abierto. Pero, independientemente del tipo de fotobiorreactor, la radiación solar constituye un factor limitante para el crecimiento y desarrollo de las microalgas, y poder conocer de forma anticipada el valor y dinámica de esta variable clave aportará considerables ventajas al control y optimización del proceso.

Más concretamente, existen dos fotobiorreactores verticales planos al aire libre y diez fotobiorreactores de tipo tubular situados dentro de un invernadero bajo una cubierta de plástico transparente. Todos los fotobiorreactores cuentan con un sistema automatizado para las entradas al sistema (agua para la refrigeración, aire, CO_2, medio de dilución con nutrientes y la frecuencia de la bomba de recirculación para los fotobiorreactores tubulares) así como para la lectura y realimentación del caudal de estas variables y una red de sensores de pH, oxígeno disuelto y temperatura para cada uno de los fotobiorreactores. Todos estos datos son transmitidos a dos ordenadores donde se realiza el control de los fotobiorreactores mediante un sistema software desarrollado para la supervisión, control y adquisición de datos (SCADA) y cuyos valores de radiación solar son tomadas por un sensor (Figura 1.1).

1.4. Principales resultados

Los resultados obtenidos en el modelo de predicción a corto plazo son muy favorables. En concreto, para la predicción a una hora los resultados oscilan entre el 6 % y el 24 % de CV-RMSE para el peor caso teniendo encuentra la diversidad de condiciones

Figura 1.1: Imágenes de un fotobiorreactor abierto tipo *raceway* (arriba izquierda), otro cerrado tipo tubular (arriba derecha), sensor de radiación (abajo izquierda) y captura del sistema SCADA (abajo derecha)

y que son días no presentados en el entrenamiento del modelo. Además el modelo reconoce con bastante buen criterio la dinámica de la radiación solar, lo cual es muy útil en tareas de control. Para el caso de la predicción a corto plazo de 12 horas se han obtenido resultados con un ligero incremento en el error, pero similar en el sentido de que captan bien la dinámica de la evolución de la radiación solar. El error (CV-RMSE) para las pruebas de predicción a 12 horas oscila entre el 11 % y el 35 %.

En la pruebas realizadas con el modelo a medio plazo se ha hecho una predicción de hasta 14 días estimando de manera iterativa el perfil del siguiente día, lo cual se ha conseguido con éxito, y además se comprueba el perfil asignado con los valores reales de radiación solar de ese dia para comprobar que las dinámicas son efectivamente similares así como otras características de la señal: valor máximo, tendencia. En general, los errores de predicción oscilan entre el 7 % y el 35 %, con un peor caso del 66 %.

1.5. Estructura de la memoria del proyecto

En el Capítulo 2 se hace una revisión bibliográfica sobre el estado del arte de la predicción de la radiación solar. A continuación, en el Capítulo 3 se explica el conjunto de datos utilizado y se detalla el proceso de tratamiento para su utilización en el

desarrollo de los modelos. En el Capítulo 4, se describe el marco teórico sobre el que se asienta el modelo de predicción incluyendo técnicas y desarrollo. Los resultados se presentan en el Capítulo 5, tanto para el modelo a corto plazo como a medio plazo. Finalmente, en el Capítulo 6 se exponen algunas conclusiones del trabajo realizado, así como posibles líneas de trabajo e investigación futuras.

2 ESTADO DEL ARTE

Existen diferentes métodos para la predicción de la radiación solar [30] que pueden clasificarse en dos grupos principales:

- Por un lado, los **modelos físicos** están basados en ecuaciones matemáticas que describen la física y la dinámica de la atmósfera. Estas ecuaciones no tienen una única solución debido a su no-linealidad por lo que se obtienen soluciones aproximadas, y por ello son conocidos también como modelos numéricos de predicción meteorológica (NWP). Los errores de estos modelos varían significativamente y dependen del clima y la dinámica de la radiación solar en el lugar de estudio. Existen trabajos previos utilizando estos modelos numéricos sólo para una resolución temporal horaria y diaria [31, 32], y en [33] se puede ver una comparación de los resultados de diferentes modelos NWP incluyendo localidades de España. La media de los errores de predicción por hora varían entre el 20,8 % y el 31,7 % para el primer día, el 21,3 % y el 36,8 % para el segundo día y el 22,4 % y el 40,9 % para el tercer día de la predicción en términos de la raíz cuadrada del error cuadrático medio relativo (CV-RMSE).
- Por otro lado, los **modelos estadísticos** se basan en establecer relaciones entre observaciones pasadas y valores futuros para predecir la radiación solar. Dependiendo de la información que utilicen, se pueden diferenciar dos grupos:
    - Los *modelos clásicos*: La denominación de clásico refleja la importancia de estos modelos en el periodo en el que no había disponible información de modelos NWP y actualmente se utilizan para la predicción a corto y medio-largo plazo, donde la información de NWP no está disponible. Estos modelos son menos complejos que los NWP porque necesitan menos información y menos tiempo computacional para realizar las predicciones. Utilizando una base de datos radiométrica para entrenar a los modelos y obteniendo mediciones terrestres en tiempo real de la ubicación bajo estudio, es posible hacer predicciones de radiación solar. En esta categoría se engloban tanto los enfoques estadísticos clásicos [34–37] como el modelo autorregresivo integrado de media móvil (ARIMA), así como los modelos basados en métodos de inteligencia artificial [38] como redes neuronales artificiales (ANN), inferencia Bayesiana, cadenas de Márkov o el algoritmo de los vecinos más cercanos (kNN) [39–45].
    - *Downscaling estadístico*: Son otra importante aplicación de los modelos estadísticos con el fin de mejorar la salida producida por los modelos

NWP cuando alguna de las variables o localización no está representada con la suficiente precisión y se emplea principalmente para horizontes de tiempo diario u horario. Algunos ejemplos muestran que los errores (CV-RMSE) de predicción para la radiación solar diaria oscila entre el 18 % y el 25 % [46,47], mientras que para la radiación solar horaria el error oscila entre el 30 % y el 40 % [48].

Otra forma de utilizar modelos estadísticos es en combinación con otros instrumentos empleados para estimar la radiación solar como los satélites meteorológicos [49–52] o las cámaras todo-cielo [53–56] para convertir la información a partir del procesamiento de las imágenes en modelos determinísticos que estiman la radiación solar basándose en el fuerte impacto de la nubosidad en la radiación terrestre. Por lo tanto, la descripción del desarrollo temporal de la nubosidad es esencial para la predicción de la radiación solar. Los métodos para predecir la radiación solar a partir de imágenes de satélite o cámaras todo-cielo se basan en campos de vectores de movimiento atmosférico (AMV), ANNs [52] o análisis de series temporales. Un ejemplo de trabajo con imágenes de satélite [50] muestra un error de predicción (CV-RMSE) del 10 % para un horizonte de predicción de 30 minutos, mientras que si el horizonte asciende hasta 6 horas, el error es del 25 %.

La elección del método para la predicción de radiación solar depende principalmente del horizonte de predicción, que puede variar de unos pocos segundos o minutos, unas pocas horas o unos cuantos días, así como del tipo de precisión requerida y del coste computacional que está dispuesto asumirse en función de los recursos disponibles. En la Tabla 2.1 puede verse una tabla comparativa de diferentes trabajos de predicción de la radiación solar.

Autor/es	**Año**	**País**	**Tipo de Modelo**	**Horizonte de predicción**	**Variables de Entrada**	**Variables de Salida**	**Resultados**
Lorenz *et ál.* [32]	2009	Alemania	Físico	1 a 3 días	Parámetros físicos de la nubosidad	Radiación solar	CV-RMSE$_1$ = 36 % CV-RMSE$_3$ = 46,3 %
Lorenz *et ál.* [33]	2009	España	Físico	1 a 3 días	Parámetros físicos de la nubosidad	Radiación solar	CV-RMSE$_1$ = 31,7 % CV-RMSE$_3$ = 40,9 %
Hammer *et ál.* [49]	1999	Alemania	Estadístico	30 minutos a 2 horas	Nubosidad en movimiento (Imágenes de satélite)	Radiación solar	CV-RMSE = 19 %
Hammer *et ál.* [50]	2001	Alemania	Estadístico	30 minutos a 6 horas	Nubosidad en movimiento (Imágenes de satélite)	Radiación solar	CV-RMSE$_{30}$ = 10 % CV-RMSE$_6$ = 25 %
Chow *et ál.* [53]	2011	San Diego, EEUU	Estadístico	30 segundos	Imágenes de satélite	Nubosidad	$\overline{e}$ = 4 %
Marquez y Coimbra [55]	2012	California, EEUU	Estadístico	Hasta 15 minutos	Imágenes de satélite	Radiación solar	CV-RMSE = 40 %
Safi *et ál.* [57]	2002	Marruecos	Estadístico	1 día	Radiación solar	Radiación solar	NMBE = 2,28 %
Reikard [35]	2008	EEUU	Estadístico (ARIMA)	1 a 4 horas	Radiación solar pasada	Radiación solar	MAPE$_1$ = 26,42 % MAPE$_4$ = 38,23 %
Mellit *et ál.* [36]	2010	Arabia Saudita	Estadístico (Modelo adaptativo)	1 hora	Radiación solar, temperatura del aire, humedad relativa y horas de sol	Radiación solar (global y las componentes difusa y directa)	CV-RMSE = 2,18 %

Tabla 2.1: Comparación de trabajos de predicción de radiación solar.

Autor/es	**Año**	**País**	**Tipo de Modelo**	**Horizonte de predicción**	**Variables de Entrada**	**Variables de Salida**	**Resultados**
Mellit *et ál.* [36]	2010	Arabia Saudita	Estadístico (Modelo adaptativo)	1 hora	Radiación solar, temperatura del aire y humedad relativa	Radiación solar (global y las componentes difusa y directa)	CV-RMSE = 9,87 %
Sfetsos y Coonick [39]	1999	Grecia	Inteligencia artificial (ANNs)	Por hora	Radiación solar pasada, velocidad y dirección del viento, temperatura, presión atmosférica	Radiación solar	CV-RMSE = 31 %
Kemmoku y Nakagawa [58]	1999	Japón	Inteligencia artificial (ANNs)	1 día	Datos de presión atmosférica, radiación solar, temperatura e índice de nubosidad	Radiación solar	MAPE = 20 %
Cao y Lin [40]	2008	Shanghai, China	Inteligencia artificial (ANNs y Wavelets)	1 hora	Radiación solar pasada y futura mediante modelo ASHRAE, índice de nubosidad, día y hora	Radiación solar	CV-RMSE = 4,76 %
Crispim *et ál.* [41]	2008	Portugal	Inteligencia artificial (ANNs)	Hasta 30 minutos	Radiación solar pasada e índices de nubosidad	Radiación solar	CV-RMSE = 42,35 %
Paoli *et ál.* [42]	2010	Francia	Inteligencia artificial (ANNs)	1 día	Radiación solar pasada e índice de nubosidad	Radiación solar	CV-RMSE = 21 %
Martín *et ál.* [45]	2010	España	Inteligencia artificial (ANNs)	1 a 3 días	Radiación solar pasada e índice de nubosidad	Radiación solar acumulada (dos valores diarios)	CV-$RMSE_1 = 20{,}58\,\%$ CV-$RMSE_3 = 30{,}39\,\%$

Tabla 2.1: Comparación de trabajos de predicción de radiación solar.

Autor/es	**Año**	**País**	**Tipo de Modelo**	**Horizonte de predicción**	**Variables de Entrada**	**Variables de Salida**	**Resultados**
Marquez y Coimbra [43]	2011	California, EEUU	Inteligencia artificial (ANNs)	1 día	Índice de nubosidad, probabilidad de precipitación, temperatura mínima y el coseno del ángulo cenital solar	Radiación solar y componente directa	CV-RMSE = 17,7 %
Ekici [59]	2014	Turquía	Inteligencia artificial (SVMs)	1 día	Día del año, temperatura media diaria, temperatura máxima diaria, horas de sol y radiación solar pasada	Radiación solar (acumulada)	CV-RMSE = 9,46 %
Ramedani *et ál.* [60]	2014	Irán	Inteligencia artificial (SVMs)	1 día	Día del año, temperatura máxima y mínima, horas de sol,	Radiación solar (acumulada)	RMSE = 3,3 R^2 = 88,9 %

Tabla 2.1: Comparación de trabajos de predicción de radiación solar.

3 ADQUISICIÓN Y TRATAMIENTO DE LOS DATOS

3.1. Adquisición de los datos

El conjunto de datos históricos se ha obtenido del sistema SCADA encargado de la adquisición de datos, control y supervisión de los fotobiorreactores de microalgas en la "Estación Experimental Las Palmerillas", propiedad de la Fundación Cajamar (Almería, España) y la Universidad de Almería. El sistema registra un total de 46 variables (valores de los sensores de pH, oxígeno disuelto, temperaturas, lectura y escritura de actuadores) con un tiempo de muestreo de 1 segundo, aunque se ha tomado sólo los correspondientes a la radiación solar ($W \cdot m^{-2}$) puesto que los modelos propuestos sólo utilizarán como datos de entrada la radiación solar pasada y una variable generada a partir de la misma (índice de nubosidad, Subsección 3.2.4). Concretamente, se han tomado los datos de radiación solar comprendidos entre el 18/10/2013 y el 11/08/2014 con lo que se dispone casi de un período anual entero para tener una representación completa las característica de estacionalidad y dinámica de radiación, aunque dicho conjunto de datos presentaban algunas anomalías y discontinuidades en el tiempo que se tratarán mediante un preprocesamiento de los datos (Sección 3.2). Este conjunto de datos servirá tanto para entrenar los diferentes modelos de predicción propuestos como para realizar la validación de los mismos comprobando su capacidad de generalización, y realizando una adecuada división de los datos en subconjuntos disjuntos.

3.2. Tratamiento de los datos

Tras la obtención de los datos, el siguiente paso es examinarlos y realizar un preprocesamiento sobre ellos con el objetivo de eliminar valores erróneos, reducir el ruido de las señales, recuperar datos a través de la interpolación y aplicar una normalización.

El objetivo final de este preprocesamiento es facilitar la convergencia del modelo e incrementar la capacidad de generalización [61, 62].

3.2.1. Filtrado de datos

3.2.1.1. Falsas medidas

Para comenzar a filtrar los datos, el primer paso es eliminar las falsas medidas. Para ello se descartan aquellos valores que están fuera del rango de lectura de cada señal respectivamente, pues se entiende que han sido marcados como erróneos por algún fallo en el sistema. También se descartan las lecturas duplicadas en el tiempo.

3.2.1.2. Eliminación de ruido

El siguiente paso es eliminar el ruido de las señales y suavizar la tendencia de los datos. Existen múltiples técnicas para el filtrado y suavizado de datos [63]. De entre ellas, finalmente se ha optado por el filtro de Savitzky-Golay [64, 65] por su buen rendimiento; además de preservar características de la distribución inicial tales como los máximos y mínimos relativos, así como el ancho de los picos, que normalmente desaparecen con otras técnicas de promediado. El resultado completo del proceso de filtrado para un día concreto puede verse en la Figura 3.1.

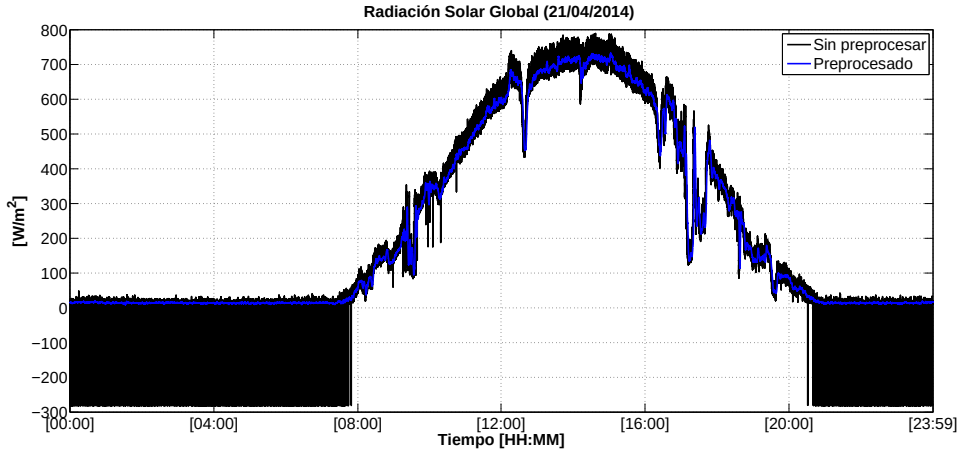


Figura 3.1: Resultado de la señal generada tras realizar el filtrado (azul) a partir de la señal original (negro) de radiación solar

3.2.2. Interpolación de datos

Al tratarse de un problema de predicción basado en series temporales, lo ideal sería que hubiera continuidad a lo largo del tiempo en los datos. Sin embargo, es difícil conseguir que esto sea así debido a las interrupciones en los sistemas de adquisición de datos.

Por ello, se plantea una solución para recuperar información perdida a través de la interpolación. Si bien hay que ser cauto con el tamaño del espacio de tiempo

a interpolar ya que se trata de un problema no-lineal y muy dinámico ya que la situación de las nubes puede cambiar en cuestión de minutos.. Por eso se ha optado por un tamaño máximo de interpolación de 15 minutos (equivalente a 900 puntos de datos por el tiempo de muestreo en segundos).

Para las variables continuas se ha utilizado un método de interpolación lineal por su sencillez y el rendimiento satisfactorio obtenido. El método funciona de una forma muy simple; dados dos puntos (x_0, y_0) y (x_1, y_1), la interpolación lineal es la línea recta entre estos puntos (Figura 3.2). Para un valor x en el intervalo (x_0, x_1), el valor de y a lo largo de la línea recta viene dado por la ecuación

$$\frac{y - y_0}{x - x_0} = \frac{y_1 - y_0}{x_1 - x_0} \tag{3.1}$$

de la que despejando y, que es el valor desconocido en el punto x, se obtiene

$$y = y_0 + (x - x_0)\frac{y_1 - y_0}{x_1 - x_0} = y_0 + \frac{(x - x_0)y_1 - (x - x_0)y_0}{x_1 - x_0} \tag{3.2}$$

que es la fórmula para la interpolación lineal en el intervalo (x_0, x_1).

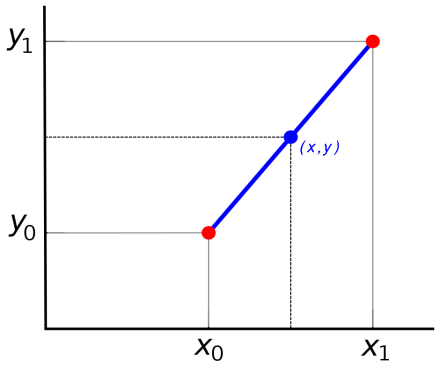


Figura 3.2: Interpolación lineal en (x_0, x_1)

3.2.3. Normalización

La última operación de preprocesamiento consiste en un tratamiento de normalización para escalar los datos a un determinado intervalo entre 0 y 1. Esta operación permite una convergencia más rápida y reduce el error del modelo. La fórmula que se ha utilizado para normalizar los datos es:

$$y_i = \frac{(y_{max} - y_{min})(x_i - x_{min})}{(x_{max} - x_{min})} + y_{min} \tag{3.3}$$

donde:

$[y_{min}, y_{max}] = [0, 1]$; Intervalo de normalización

$[x_{min}, x_{max}]$; Mínimo y máximo de los datos

x_i; Valor a normalizar

3.2.4. Generación de la característica *Índice de nubosidad*

El estado del cielo y la nubosidad presente tiene una influencia directa sobre la radiación terrestre [66], por ello es importante dar esa información al modelo de forma que pueda capturar esa dinámica. De esta manera, además de la propia radiación solar anterior, se va generar una nueva característica que formará parte del conjunto de entradas del modelo. Esta nueva característica, índice de nubosidad δ se calcula como la relación de la variación entre la radiación solar medida y la radiación solar media teórica en ese instante (Ecuación 3.4). Este valor medio se ha obtenido de una base de datos que proporciona la Agencia Andaluza de la Energía (AAE). En la Figura 3.3, puede verse un ejemplo de un día concreto con nubosidad variable. Otra ventaja de incluir esta nueva característica en el modelo es la posibilidad de incluir predicciones de fuentes externas, como la Agencia Estatal de Meteorología (AEMET) o OpenWeatherMap, cuando se utiliza el modelo en tiempo real aumentando la precisión de las predicciones a medio-largo plazo.

$$\delta = \frac{RadTerrestre}{RadTerrestre\,Media} \tag{3.4}$$

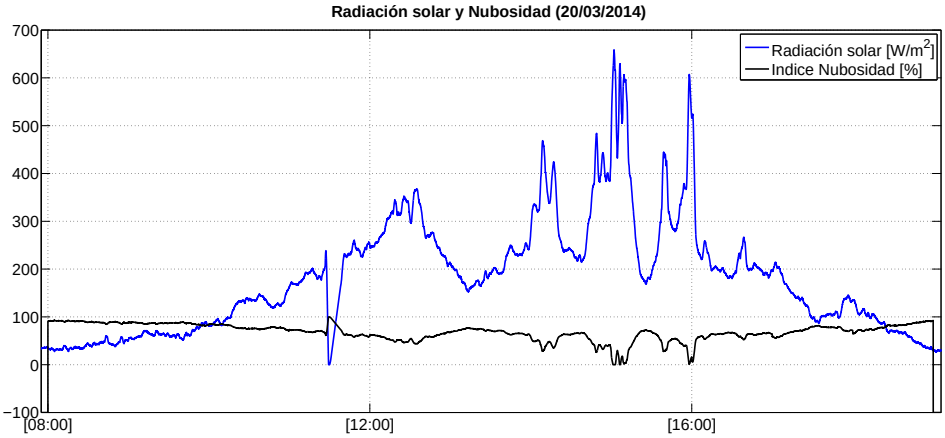


Figura 3.3: Radiación solar de un día con nubosidad variable (azul) y su correspondiente índice de nubosidad generado (negro)

3.3. Preanálisis de los datos

Antes de realizar la división de los datos históricos en subconjuntos para el entrenamiento y validación de los modelos, sería interesante realizar un breve análisis preliminar para conocer cuantitativamente las características del problema observando las variables involucradas. En la Tabla 3.1 se pueden observar las variables del modelo que son simplemente la radiación solar pasada y el índice de nubosidad generado, de esta forma se sigue el principio de parsimonia [67]. Como se puede comprobar en la Tabla 3.2, se puede hacer una distinción entre las estaciones de otoño e invierno por un lado y las de primavera y verano por otro en función de la similaridad de los valores del pico máximo, la media y la desviación típica de los valores de radiación solar que son más altos para las estaciones de primavera y verano. Así mismo sucede con los datos del índice de nubosidad aunque curiosamente la media del índice de nubosidad para el período de verano es mayor que para el resto, hay que tener en cuenta que el índice de nubosidad es un valor relativo sobre la media de radiación obtenida de los datos de la AAE y que tanto el sensor de radiación como las instalaciones de fotobiorreactores se encuentra bajo una cubierta de plástico en un invernadero reduciendo el máximo de radiación en un coeficiente determinado. Si los datos obtenidos con el modelo predictivo no fueran satisfactorios quizás mereciera la pena hacer modelos más específicos basados en distintas condiciones.

Nombre de la variable	**Unidad**	**Rango de lectura**
Radiación solar	$W \cdot m^{-2}$	$[0, 900]$
Índice de nubosidad	%	-1 o $[0, 1]$

Tabla 3.1: Variables de entrada para el modelo predictivo

3.4. División de los datos

3.4.1. Predicción a corto plazo

Para la predicción a corto plazo (minutos a horas) se dispone de un conjunto de bloques de datos contiguos (Tabla 3.3) resultado del proceso de interpolación. Aún así, a causa de la discontinuidad en ellos, se ha optado por realizar una división manual de los datos. Para el conjunto de datos de entrenamiento se han utilizado bloques contiguos correspondientes a las distintas estaciones (Tabla 3.5), intentando utilizar una cantidad de datos similar para cada estación con el objetivo de conseguir

	Radiación solar $[W \cdot m^{-2}]$			
	Otoño	**Invierno**	**Primavera**	**Verano**
Máximo	883,62	828,22	900	900
Mínimo	0	0	0	0
Media	262,71	296,26	371,84	372,07
Desviación típica	167,41	187,81	226,65	203,90

	Índice de nubosidad			
	Otoño	**Invierno**	**Primavera**	**Verano**
Máximo	0,9536	1	1	1
Mínimo	0	0	0	0
Media	0,4173	0,4394	0,3992	0,4767
Desviación típica	0,3231	0,3188	0,3398	0,3258

Tabla 3.2: Valores descriptivos por estación: máximo, mínimo, media y desviación típica

una representatividad lo más completa posible dada la limitación de los datos. Este conjunto de datos servirá para ajustar los parámetros del modelo.

Además, para evitar el sobreentrenamiento y aumentar la capacidad de generalización de la red neuronal, se ha utilizado un conjunto de datos de validación heterogéneo (Tabla 3.6). De esta forma se persigue que el modelo resultante no sea el que mejor se comporte sobre el conjunto de datos de entrenamiento (Tabla 3.5), sino aquel que mejor capacidad para generalizar tenga y minimice la función de error en el conjunto de datos de validación. Este conjunto de datos se utilizará para evaluar la bondad del modelo entrenado con el objetivo de que al mismo tiempo se prueba su capacidad de generalización. Por último, en la Figura 3.4 puede verse una representación gráfica en el tiempo de la distribución de los datos según su finalidad, aquellos que están marcados sin usar no se utilizarán en el entrenamiento ni validación del modelo, pero sí como pruebas independientes.

3.4.2. Predicción a medio plazo

En el caso de la predicción a medio plazo (días) se dispone de un conjunto más limitado de datos puesto que se han tomado únicamente aquellos días completos sin huecos en los datos (Tabla 3.7), esto es porque el modelo toma como entrada el perfil de radiación solar de los días anteriores y en base a eso genera la predicción para los días siguientes como se explicará en las secciones siguientes, donde se explicará que aunque esto pueda ser una limitación también se propone el procedimiento para generar el perfil a partir de días incompletos. Sin embargo, en este caso se ha optado por dejar únicamente días completos para aumentar la precisión del modelo con

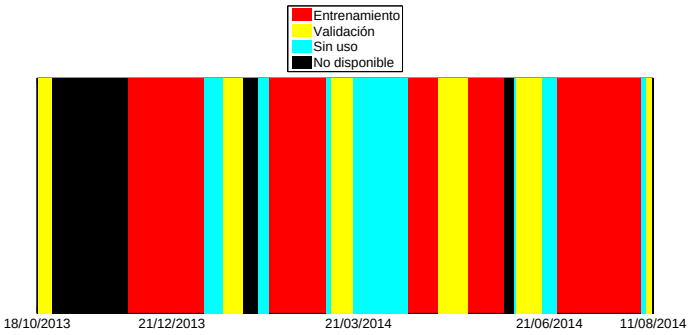


Figura 3.4: Representación gráfica en el tiempo de la distribución de los datos según su utilización en el modelo de predicción a corto plazo

datos de calidad. Dada la limitación de los datos (Tabla 3.7), se ha optado por emplear tanto para el entrenamiento y la validación del modelo el mismo conjunto (Tabla 3.9) dado que el modelo utiliza días anteriores como entrada y éstos deben ser consecutivos, por lo que al ser bloques de mayor tamaño se puede tomar un número mayor de entradas pasadas.

Fecha inicial bloque	**Fecha final bloque**	**Tamaño [s]**
18/10/2013 00:00:00	24/10/2013 10:18:37	555518
30/11/2013 00:00:00	05/01/2014 12:28:58	3155339
06/01/2014 00:00:00	14/01/2014 10:59:32	730773
14/01/2014 11:53:35	24/01/2014 23:59:59	907585
01/02/2014 00:00:00	05/02/2014 14:21:16	397277
06/02/2014 00:00:00	05/03/2014 08:47:10	2364431
05/03/2014 09:14:30	05/03/2014 13:03:35	13746
05/03/2014 13:26:09	07/03/2014 08:17:10	154262
08/03/2014 00:00:00	18/03/2014 08:22:13	894134
18/03/2014 09:09:10	19/03/2014 08:32:42	84213
19/03/2014 08:59:26	21/03/2014 10:34:05	178480
21/03/2014 10:53:08	26/03/2014 12:18:02	437095
26/03/2014 12:56:51	26/03/2014 21:33:41	31011
26/03/2014 21:51:11	26/03/2014 23:55:23	7453
27/03/2014 00:58:40	30/03/2014 01:59:59	262880
30/03/2014 03:00:00	03/04/2014 10:05:20	371121
03/04/2014 10:34:40	07/04/2014 10:27:56	345197
07/04/2014 10:54:33	13/04/2014 19:26:22	549110
13/04/2014 20:23:13	28/04/2014 14:49:27	1275975
29/04/2014 00:01:43	12/05/2014 12:40:20	1168718
12/05/2014 13:17:55	12/05/2014 13:39:27	1293
12/05/2014 14:40:01	30/05/2014 23:59:59	1588799
04/06/2014 00:00:00	04/06/2014 14:58:48	53929
04/06/2014 17:15:59	17/06/2014 06:26:04	1084206
17/06/2014 08:27:34	17/06/2014 08:57:32	1799
17/06/2014 09:19:51	19/06/2014 15:23:39	194629
19/06/2014 17:38:33	22/06/2014 07:38:49	223217
23/06/2014 07:46:06	24/06/2014 00:50:39	61474
24/06/2014 08:39:42	04/08/2014 14:28:48	3563347
04/08/2014 15:00:29	06/08/2014 14:02:14	169306
06/08/2014 14:35:40	09/08/2014 17:57:55	271336
11/08/2014 07:13:12	11/08/2014 10:46:50	12819

Tabla 3.4: Bloques de datos contiguos

Otoño		
Fecha inicial	**Fecha final**	**Tamaño [s]**
30/11/2013 00:00:00	05/01/2014 12:28:58	3155339
Invierno		
Fecha inicial	**Fecha final**	**Tamaño [s]**
06/02/2014 00:00:00	05/03/2014 08:47:10	2364431
Primavera		
Fecha inicial	**Fecha final**	**Tamaño [s]**
13/04/2014 20:23:13	28/04/2014 14:49:27	1275975
12/05/2014 14:40:01	30/05/2014 23:59:59	1588799
Verano		
Fecha inicial	**Fecha final**	**Tamaño [s]**
24/06/2014 08:39:42	04/08/2014 14:28:48	3563347

Tabla 3.5: Conjunto de datos para entrenamiento del modelo de predicción a corto plazo

Otoño		
Fecha inicial	**Fecha final**	**Tamaño [s]**
18/10/2013 00:00:00	24/10/2013 10:18:37	555518
Invierno		
Fecha inicial	**Fecha final**	**Tamaño [s]**
14/01/2014 11:53:35	24/01/2014 23:59:59	907585
08/03/2014 00:00:00	18/03/2014 08:22:13	894134
Primavera		
Fecha inicial	**Fecha final**	**Tamaño [s]**
29/04/2014 00:01:43	12/05/2014 12:40:20	1168718
04/06/2014 17:15:59	17/06/2014 06:26:04	1084206
Verano		
Fecha inicial	**Fecha final**	**Tamaño [s]**
06/08/2014 14:35:40	09/08/2014 17:57:55	271336

Tabla 3.6: Conjuntos de datos para validación del modelo de predicción a corto plazo

Fecha inicial bloque	Fecha final bloque	Número de días
18/10/2013	23/10/2013	6
30/11/2013	04/01/2014	36
06/01/2014	13/01/2014	8
15/01/2014	23/01/2014	9
01/02/2014	04/02/2014	4
06/02/2014	04/03/2014	27
06/03/2014	06/03/2014	1
08/03/2014	17/03/2014	10
20/03/2014	20/03/2014	1
22/03/2014	25/03/2014	4
28/03/2014	29/03/2014	2
31/03/2014	02/04/2014	3
04/04/2014	06/04/2014	3
08/04/2014	12/04/2014	5
14/04/2014	27/04/2014	14
30/04/2014	11/05/2014	12
13/05/2014	29/05/2014	17
05/06/2014	16/06/2014	12
18/06/2014	18/06/2014	1
20/06/2014	21/06/2014	2
25/06/2014	03/08/2014	40
05/08/2014	05/08/2014	1
07/08/2014	08/08/2014	2

Tabla 3.8: Bloques de días completos

Fecha inicial	Fecha final	Número de días
30/11/2013	04/01/2014	36
06/02/2014	04/03/2014	27
25/06/2014	03/08/2014	40

Tabla 3.9: Conjuntos de datos para validación del modelo de predicción a corto plazo

4 MATERIAL Y MÉTODOS

4.1. Máquinas de vectores de soporte - Support Vector Machines (SVM)

Las SVM son una técnica de aprendizaje supervisado que se basan en la teoría del aprendizaje estadístico [15, 68] y derivan de la hipótesis de minimización del riesgo estructural para minimizar tanto el riesgo estructural como el intervalo de confianza de la máquina de aprendizaje consiguiendo una buena capacidad de generalización y robustez frente a datos con ruido presente. Se han aplicado con éxito en tareas de clasificación [16–18], regresión [19, 20, 69] y predicción [21–23], demostrando ser un algoritmo robusto y eficiente para estas tareas. Las SVMs no hacen suposiciones previas sobre la distribución de probabilidad de los datos como otras técnicas que normalmente asumen que la distribución de los datos se conoce a priori lo que no ocurre en los datos de radiación solar que suelen presentarse como series temporales y no hay suposiciones estadísticas fiables que se puedan emplear. También son capaces de modelar sistemas multivariable y no lineales como es el caso de la radiación solar [24]. Y otra característica que hace a las SVMs una técnica interesante para la predicción de radiación solar es que hacen un balance entre la precisión obtenida en un conjunto finito de datos de entrenamiento y la habilidad para generalizar ante nuevas observaciones, lo que constituye una característica muy importante frente al problema del sobreajuste del modelo (*overfitting*).

4.1.1. Clasificación

Las SVMs fueron desarrolladas inicialmente para problemas de clasificación linealmente separables [68, 70]. El objetivo preliminar de clasificación SVM es establecer límites de decisión en el espacio característico para separar los puntos de datos en diferentes clases excluyentes. Su intención es crear un hiperplano de separación óptima entre dos clases para minimizar el error de generalización y con ello maximizar el margen. Si cualesquiera dos clases son separables de entre el infinito número de clasificadores lineales, SVM determina que hiperplano minimiza el error de generalización (es decir, el error en los patrones de prueba) y por el contrario si las dos clases no son separables, SVM trata de buscar que hiperplano maximiza el margen y, al mismo tiempo, minimiza una medida proporcional al número de errores de clasificación. Por lo tanto, el hiperplano seleccionado tendrá el margen máximo entre las dos

clases, donde el margen se define como la suma de la distancia entre el hiperplano de separación y los puntos más cercanos a cada lado de dos clases (Figura 4.1). En la Figura 4.1 se puede observar un plano bidimensional en el que están representados los objetos linealmente separables de dos clases $\{+, o\}$. El objetivo es encontrar un clasificador que las separe perfectamente y puede haber muchas maneras para hacerlo, pero SVM trata de encontrar un único hiperplano que produce el máximo margen posible, es decir maximizando la distancia entre el hiperplano y los puntos de datos más cercanos de cada clase. De esta manera, los objetos de la clase + se encuentran tras el hiperplano H2 mientras que los de la clase *o* tras el hiperplano H1, y los objetos de cada clase que se sitúan exactamente sobre estos hiperplanos H1 y H2 se denominan vectores de soporte (*support vectors*), mientras que la distancia entre estos hiperplanos H1 y H2 (δ) es el máximo margen obtenido. La mayoría de los datos de entrenamiento importantes son vectores de soporte puesto que definen el hiperplano y tienen influencia directa en la ubicación óptima de la superficie de decisión.

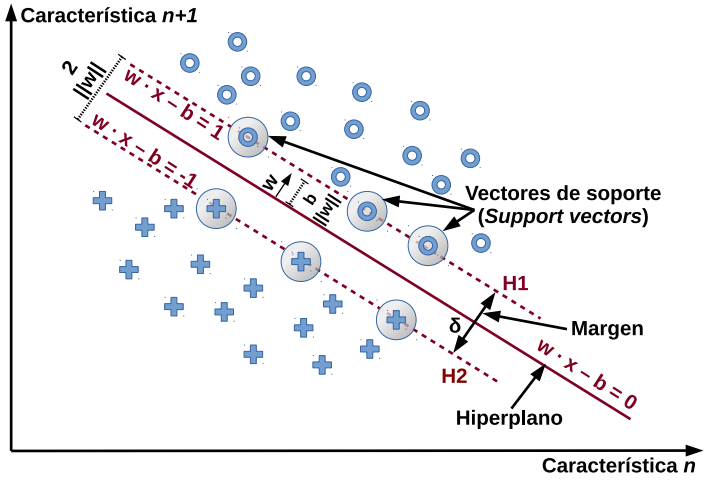


Figura 4.1: Hiperplano de máxima separación

Más formalmente, dado un conjunto de datos de entrenamiento T formado por n puntos de la forma:

$$\mathrm{T} = \{(\mathrm{x_i}, \mathrm{y_i}) | \mathrm{x_i} \in \mathbb{R}^{\mathrm{p}}, \mathrm{y_i} \in \{-1, 1\}\}_{\mathrm{i=1}}^{\mathrm{n}} \tag{4.1}$$

donde y_i indica la clase ($\{-1, 1\}$) a la que pertenece cada punto x_i y cada punto x_i es un vector p-dimensional en $\mathbb{R}$. En este espacio, cualquier hiperplano se puede expresar como un conjunto de puntos x que satisfacen:

$$w \cdot x - b = 0 \tag{4.2}$$

donde $\cdot$ es el producto escalar y w el vector normal, no necesariamente normalizado, a el hiperplano. El parámetro $\frac{b}{\|w\|}$ determina el desplazamiento del hiperplano desde el origen a lo largo del vector w. Si los datos son linealmente separables, se pueden seleccionar dos hiperplanos de forma que separen los puntos de ambas clases sin tener puntos entre ellos y al mismo tiempo maximizando el margen. Estos hiperplanos óptimos se pueden describir por las ecuaciones:

$$\begin{aligned} w \cdot x - b &= 1 \\ w \cdot x - b &= -1 \end{aligned} \tag{4.3}$$

Mediante geometría, se tiene que la distancia entre estos dos hiperplanos es $\frac{2}{\|w\|}$, así que el objetivo del problema es minimizar $\|w\|$ para maximizar el margen entre los hiperplanos. Al mismo tiempo, como se quiere prevenir que haya puntos de datos en dicho margen, se añade la siguiente restricción para cada i:

$$w \cdot x_i - b \geq 1, \quad \text{para los } x_i \text{ de la primera clase} \tag{4.4}$$

o bien:

$$w \cdot x_i - b \leq -1, \quad \text{para los } x_i \text{ de la segunda clase} \tag{4.5}$$

Esto también se puede expresar como:

$$y_i(w \cdot x_i - b) \geq 1, \quad \forall\, 1 \leq i \leq n \tag{4.6}$$

Finalmente, lo que se tiene es un problema de optimización que se puede escribir de la siguiente forma:

$$\begin{aligned} &\text{Minimizar (en } w, b\text{):} \\ &\qquad\qquad \|w\| \\ &\text{sujeto a (para cualquier } i = 1, \ldots, n\text{):} \\ &\qquad\qquad y_i(w \cdot x_i - b) \geq 1 \end{aligned} \tag{4.7}$$

Sin embargo, este problema de optimización es difícil de resolver porque depende de $\|w\|$, la norma de w, lo que implica una raíz cuadrada. Afortunadamente es posible alterar la ecuación sustituyendo $\|w\|$ por $\frac{1}{2}\|w\|^2$ sin cambiar la solución al problema original. Se trata de un problema de optimización cuadrática:

$$\begin{aligned} &\text{Minimizar (en } w, b\text{):} \\ &\qquad \frac{1}{2}\|w\|^2 \\ &\text{sujeto a (para cualquier } i = 1, \ldots, n\text{):} \\ &\qquad y_i(w \cdot x_i - b) \geq 1 \end{aligned} \tag{4.8}$$

Introduciendo los multiplicadores de Lagrange α, el problema de optimización con restricciones se puede expresar como:

$$\begin{aligned} &\text{Minimizar (en } w, b\text{), maximizar (en } \alpha \geq 0\text{):} \\ &\qquad \frac{1}{2}\|w\|^2 - \sum_{i=1}^{n} \alpha_i \left[y_i(w \cdot x_i - b) - 1\right] \end{aligned} \tag{4.9}$$

es decir que se busca un punto de silla . De este modo todos los puntos que se pueden separar como $y_i(w \cdot x_i - b) - 1 > 0$ no importan ya que se debe establecer el correspondiente α_i a cero.

Así, el problema puede resolverse mediante técnicas de programación cuadrática estándar. La condición "estacionaria" de Karush-Kuhn-Tucker [71] implica que la solución puede expresarse como una combinación lineal de los vectores de entrenamiento:

$$w = \sum_{i=1}^{n} \alpha_i y_i x_i \tag{4.10}$$

Sólo unos pocos α_i serán mayores que cero, y los correspondientes x_i serán exactamente los vectores de soporte (support vectors), que se sitúan sobre el margen y satisfacen $y_i(w \cdot x_i - b) = 1$. De aquí se puede derivar que los vectores de soporte también satisfacen:

$$w \cdot x_i - b = \frac{1}{y_i} = y_i \Longleftrightarrow b = w \cdot x_i - y_i \tag{4.11}$$

Lo que permite definir el desplazamiento b. En la práctica, resulta más robusto promediar todos los vectores de soporte N_{SV}:

$$b = \frac{1}{N_{SV}} \sum_{i=1}^{N_{SV}} (w \cdot x_i - y_i) \tag{4.12}$$

Por último, expresar el problema de optimización en su forma dual sin restricciones revela que el hiperplano que maximiza el margen, es decir la tarea de clasificación, es sólo una función de los vectores de soporte, esto es el subconjunto de los datos de entrenamiento que se sitúan en el margen.

Utilizando el hecho de que $\|w\|^2 = w \cdot w$ y sustituyendo $w = \sum_{i=1}^{n} \alpha_i y_i x_i$, se puede ver que el problema dual se reduce al siguiente problema de optimización:

$$\text{Maximizar (en } \alpha_i\text{):}$$

$$\begin{aligned}\tilde{L}(\alpha) &= \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum \alpha_i \alpha_j y_i y_j x_i^T x_j \\ &= \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum \alpha_i \alpha_j y_i y_j k(x_i x_j)\end{aligned}$$

$$\text{sujeto a (para cualquier } i = 1, \ldots, n\text{):}$$

$$\begin{cases} \alpha_i \geq 0, \\ \sum_{i=1}^{n} \alpha_i y_i = 0 \end{cases}$$

Así puede definirse el kernel por $k(x_i x_j) = x_i \cdot x_j$, y w se puede calcular gracias a los términos de α: $w = \sum_i \alpha_i y_i x_i$.

Se puede encontrar una descripción más detallada de los fundamentos y el marco teórico sobre el que se fundamentan las SVM en [68, 72]. Además, existen diferentes variantes de clasificadores SVM: de margen máximo o *hard-margin*, *soft-margin*, k-vecinos más cercanos, de función de base radial... Se puede encontrar una revisión en [73].

4.1.2. Método del Kernel

En los problemas de clasificación reales, lo habitual es que una máquina de aprendizaje lineal no pueda determinar un hiperplano con una separación exacta de los datos por varias limitaciones: existen más de dos variables de entrada, hay curvas no lineales de separación, casos en los que los datos no puedan ser completamente separados o clasificaciones en más de dos categorías. Para hacer frente a estos casos, el espacio de entradas original puede ser mapeado a un nuevo espacio característico de mayor dimensión (espacio de Hilbert) usando funciones no-lineales denominadas funciones de características ϕ (Figura 4.2). A pesar de que el espacio característico es de dimensionalidad alta, no sería factible en la práctica realizar directamente la clasificación. Así que en estos casos, el mapeado no lineal inducido por ϕ se utiliza para el cálculo usando funciones no lineales especiales llamadas *kernels* [74], las más habituales se pueden ver en Ecuación 4.13. El algoritmo resultante es similar, pero el producto escalar $\cdot$ se reemplaza por la función de kernel no-lineal. Esto permite al algoritmo obtener un hiperplano de máximo margen en el nuevo espacio

característico y de esta forma aunque el clasificador sea un hiperplano en el espacio característico de alta dimensionalidad, puede ser no-lineal en el espacio de entrada original Figura 4.3. La función de kernel K está relacionada con la función ϕ por la ecuación $K(x_i, x_j) = \phi(x_i) \cdot \phi(x_j)$. El valor w también está en el nuevo espacio $w = \sum_i \alpha_i y_i \phi(x_i)$. Los productos escalares con w para clasificación se pueden calcular mediante el método del kernel $w \cdot \phi(x) = \sum_i \alpha_i y_i k(x_i, x)$.

$$K(x_i, x_j) = \begin{cases} (x_i \cdot x_j) & \text{Lineal} \\ (x_i \cdot x_j)^d & \text{Polinomial (homogéneo)} \\ (x_i \cdot x_j + 1)^d & \text{Polinomial (no homogéneo)} \\ \exp(-\gamma \parallel x_i - x_j \parallel^2), \quad \gamma > 0 & \text{Función de base radial (RBF)} \\ \tanh(\kappa x_i \cdot x_j + c), \quad \kappa > 0; c > 0 & \text{Sigmoidea} \end{cases} \tag{4.13}$$

Las características de las SVMs y el método del Kernel incluyen: el rendimiento está garantizado puesto que están basados puramente en ejemplos teóricos de aprendizaje estadístico, el espacio de búsqueda tiene un mínimo único, el entrenamiento es extremadamente robusto y eficiente, y la capacidad de generalización permite un equilibrio entre la complejidad del modelo y el error empírico. La arquitectura general de una SVM puede verse en Figura 4.4.

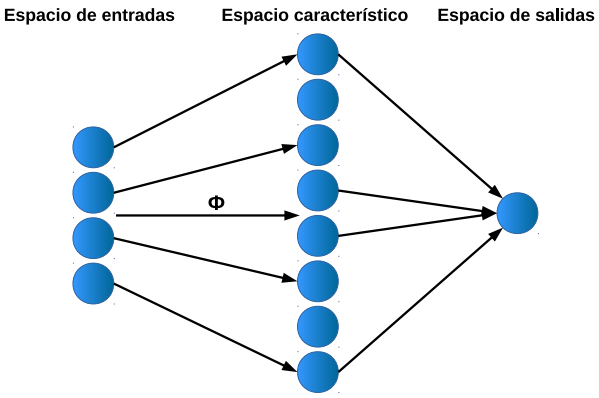


Figura 4.2: Transformación del espacio de entrada al espacio característico de mayor dimensionalidad mediante ϕ

4.1.3. Support Vector Regression (SVR)

Las SVMs no solo se pueden aplicar a problemas de clasificación, sino también para regresión y estimación de funciones. De la misma manera que en el enfoque visto anteriormente para clasificación, también hay un interés en buscar y optimizar los

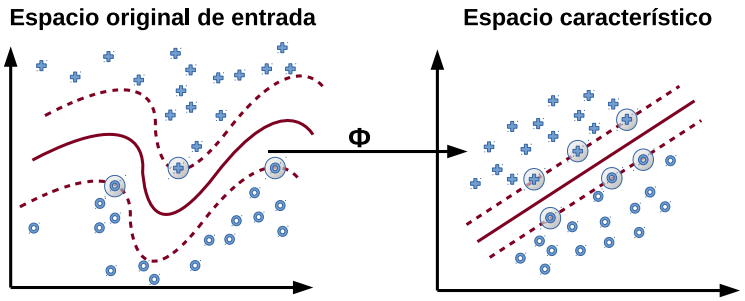


Figura 4.3: Hiperplano en el espacio característico tras la transformación mediante ϕ

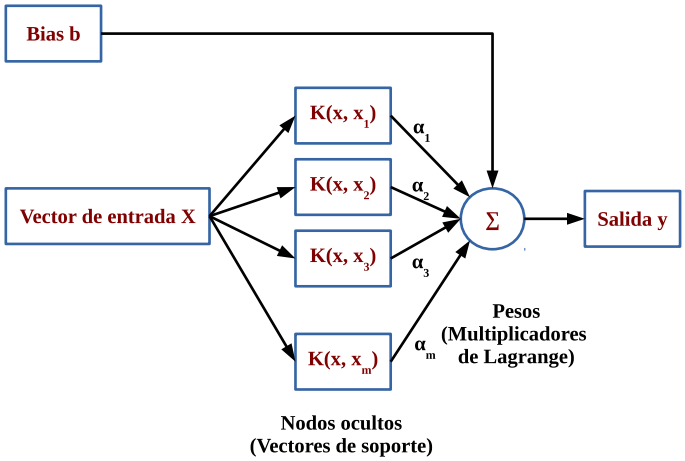


Figura 4.4: Arquitectura de una SVM

límites de generalización dados para regresión, estos se basan en la definición de una función de pérdida que ignora los errores que están situados dentro de cierto umbral del verdadero valor. Este tipo de enfoque se conoce como ε-SVR, en la Figura 4.5 se puede ver un ejemplo de función de regresión con la banda ε, donde las variables ξ_i miden el coste de los errores en los puntos de entrenamiento siendo cero para aquellos que caen dentro de la banda. En SVR, el espacio de entrada primero se mapea en un espacio característico de mayor dimensión utilizando algún mapeado no-lineal fijado Φ y entonces se construye un modelo lineal en ese espacio característico. Esto se puede expresar mediante:

$$f(x) = \sum_{i=1}^{n} w_i \Phi(x_i) + b \tag{4.14}$$

La SVR estándar utiliza una función de pérdida $L_\varepsilon(y, f(x))$ que describe la desvia-

ción de la función estimada y la original, y se define como:

$$L_\varepsilon(y, f(x)) = \begin{cases} 0, & \text{si } |y - f(x)| \leq \varepsilon \\ |y - f(x)| - \varepsilon, & \text{en otro caso} \end{cases} \tag{4.15}$$

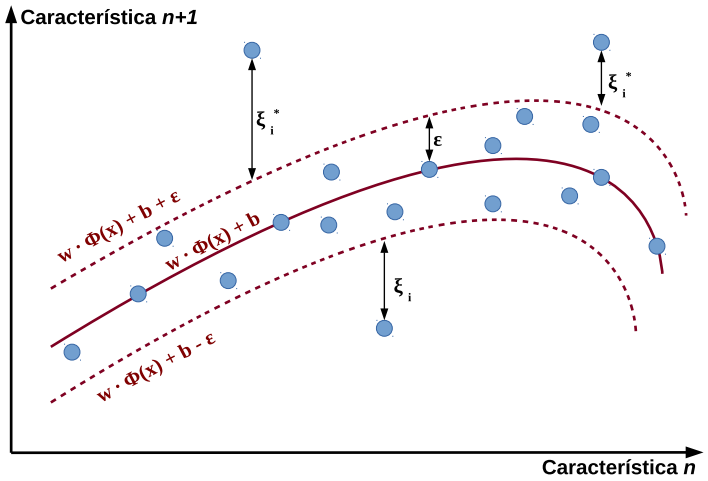


Figura 4.5: Función de regresión no-lineal con la banda ϵ

El problema de optimización para SVR se define como:

$$\begin{aligned} &\text{Minimizar (en } w\text{):} \\ &\qquad \frac{1}{2} \parallel w \parallel^2 + C \sum_{i=1}^{n} (\xi_i + \xi_i^*) \\ &\text{sujeto a (para cualquier } i = 1, \ldots, n\text{):} \\ &\qquad \begin{cases} y_i - f(x_i, w) \leq \varepsilon + \xi_i^* \\ f(x_i, w) - y_i \leq \varepsilon + \xi_i \\ \xi_i, \xi_i^* \geq 0 \end{cases} \end{aligned} \tag{4.16}$$

De esta manera, al mismo tiempo que se reduce la complejidad del modelo por medio de la minimización de $\parallel w \parallel^2$, también se reduce el error introduciendo las variables de holgura no negativas ξ_i, ξ_i^*para medir la desviación de los puntos de entrenamiento fuera de la zona de ε.

Este problema de optimización puede transformarse en su problema dual y la solución viene dada por:

Maximizar (en α_i,α_i^*):

$$-\frac{1}{2}\sum_{i,j=1}^{n}(\alpha_i-\alpha_i^*)(\alpha_i-\alpha_j^*)\left\langle\phi(x_i)\cdot\phi(x_j)\right\rangle$$
$$-\varepsilon\sum_{i=1}^{n}(\alpha_i+\alpha_i^*)+\sum_{i=1}^{n}y_i(\alpha_i-\alpha_i^*)$$

sujeto a (para cualquier $i=1,\ldots,n$):

$$\begin{cases}\sum_{i=1}^{n}\alpha_i-\alpha_i^*=0\\ 0\le\alpha_i\le C,\\ 0\le\alpha_i^*\le C\end{cases}\tag{4.17}$$

Finalmente, aplicando el método del kernel se obtiene la función de decisión del SVR no-lineal, y cuya arquitectura se puede ver representada en Figura 4.4:

$$\sum_{i=1}^{N_{SV}}(\alpha_i-\alpha_i^*)K(x_i,x_j)+b\tag{4.18}$$

La precisión y generalización del modelo SVR depende en buena medida de la correcta configuración de meta-parámetros C, ε y los parámetros del kernel. La elección de un tipo de kernel en particular normalmente depende del conocimiento sobre la aplicación del dominio, pero debería reflejar la distribución de los datos de entrada. El parámetro C determina el equilibrio entre la complejidad del modelo y el grado en que se toleran las desviaciones de error más grandes que ε en la optimización, por ejemplo si C es demasiado grande el objetivo pasa a ser reducir al mínimo el riesgo empírico únicamente y omitiendo la parte de complejidad del modelo en la formulación del problema de optimización. El parámetro ε controla el ancho de la zona umbral usada para ajustar los datos de entrenamiento. El valor de ε puede afectar al número de vectores de soporte usados para construir la función de regresión, cuanto más grande sea habrá menos vectores de soporte seleccionados. Ambos parámetros afectan a la complejidad del modelo aunque de manera diferente.

Existen otras técnicas de SVR como el SVM de mínimos cuadrados (Least-squares SVM: LS-SVM) propuesta en [75]. Es una técnica de aprendizaje automático ampliamente aplicable tanto para clasificación como regresión. La solución de LS-SVM deriva de las ecuaciones lineales de Karush-Kuhn-Tucker en lugar del problema de programación cuadrática del tradicional SVM. Una desventaja es que utilizar una función de pérdida cuadrática sin ninguna regularización tiende a estimaciones menos robustas. Para evitarlo, se ha propuesto un LS-SVM ponderado donde los pesos se asignan a los datos en un método de entrenamiento de dos pasos.

También las nu-SVM (ν-SVM) se basan en que el margen flexible tiene que estar en el rango $[0,1]$ [76], donde la SVM emplea hiperplanos de separación no homogéneos,

es decir, no incluye el origen. El parámetro ν no rige el equilibrio entre el error de entrenamiento y el de generalización, sino que ahora tiene un doble rol: es el límite superior de la fracción de errores del margen y además es el límite inferior de la fracción de vectores de soporte.

4.2. Reconstrucción del espacio de estados

En el análisis de series temporales no lineales, como es el caso de la radiación solar [24] y con el objetivo de realizar una tarea de predicción, es necesario introducir brevemente conceptos de la teoría de inmersión y la reconstrucción del espacio de estados [77].

El estado de un sistema dinámico determinista es la información necesaria para determinar la evolución del sistema en el tiempo. En tiempo discreto, esta evolución se puede describir mediante el siguiente sistema de ecuaciones en diferencias:

$$x(n+1) = F[x(n)] \tag{4.19}$$

donde $x(n) \in \mathbb{R}^d$ es el estado del sistema en el instante n, y F es una función vectorial no lineal. Una serie temporal es un conjunto de mediciones ordenadas en el tiempo $\{x(n)\}, n = 1, \ldots, N$ de una magnitud escalar observada en la salida del sistema. Esta cantidad observable se define en términos del estado $x(n)$ del sistema subyacente de la siguiente manera:

$$x(n) = h[x(n)] + \varepsilon(t) \tag{4.20}$$

donde h es una función escalar no lineal, ε es una variable aleatoria que representa incertidumbre en el modelado y/o mediciones de ruido. Se suele asumir que $\varepsilon(t)$ se obtiene de un proceso de ruido blanco Gaussiano. Se puede deducir a partir de la Ecuación 4.20 que las observaciones $\{x(n)\}$ se pueden ver como una proyección del espacio de estados multivariable del sistema en el espacio unidimensional. Las ecuaciones Ecuación 4.19 y Ecuación 4.20 conjuntamente describen el comportamiento del espacio de estado del sistema dinámico.

Con el fin de realizar la predicción, es necesario reconstruir (estimar) tan bien como sea posible el espacio de estado del sistema usando la información proporcionada por $\{x(n)\}$ únicamente. En [78] se demuestra que, bajo condiciones muy generales, el estado de un sistema dinámico determinista puede ser reconstruido de manera precisa por una ventana de tiempo finito que se desplaza a lo largo de las series temporales observadas de la siguiente manera:

$$\hat{x}(n) = [x(n), x(n-\tau), \ldots, x(n-(d_E-1)\tau)] \tag{4.21}$$

donde $x(n)$ es la muestra de datos de la serie temporal en el instante n, d_E es la dimensión de inmersión y τ es el retardo de inmersión. La Ecuación 4.21 implementa el teorema de inmersión con retardos [79]. Según este teorema, una colección de valores pasados en un espacio vectorial de dimensión d_E deben proporcionar suficiente información para reconstruir los estados de un sistema dinámico observable. Al hacer esto, lo que se trata de conseguir es desplegar la proyección de nuevo a un espacio de estado multivariable cuyas propiedades topológicas son equivalentes a las del espacio de estado que genera realmente la serie temporal observable, siempre que la dimensión de inmersión d_E sea suficientemente grande.

El teorema de inmersión también proporciona un marco teórico para la predicción de series temporales no lineales, donde la relación de predicción entre el actual estado $x(n)$ y el siguiente valor de la serie temporal viene dada por la siguiente ecuación:

$$x(n+1) = g[x(n)] \tag{4.22}$$

Una vez elegidos la dimensión d_E y el retardo τ de inmersión, resta aproximar la función de mapeado g cuya aproximación se hará mediante SVR:

$$\hat{x}(n+1) = \hat{g}[x(n)] \tag{4.23}$$

donde $\hat{x}(n+1)$ es una estimación de $x(n+1)$ y $\hat{g}$ es la correspondiente aproximación de g. El error de estimación, $e(n+1) = x(n+1) - \hat{x}(n+1)$, suele utilizarse para evaluar la calidad de la aproximación.

De esta forma, el vector de entrada, denotado por $x(n)$, se define por las coordenadas de inmersión de la Ecuación 4.21, por lo que está compuesto de d_E valores reales de las series temporales observadas, separados unos de otros por τ instantes de tiempo:

$$x(n) = [x(n), x(n-\tau), \ldots, x(n-(d_E-1)\tau)] \tag{4.24}$$

4.3. Optimización mediante algoritmos genéticos

Los algoritmos genéticos (GA) pertenecen a la categoría de métodos de búsqueda estocástica, pero a diferencia de otros métodos (temple simulado, aceptación de umbral...) que operan sobre una única solución, los GA operan sobre una población de soluciones por lo que es una técnica muy recurrida de optimización global. Los aspectos más importantes de los GA son: la definición de la función objetivo, la definición e implementación de la representación de las soluciones y la definición e implementación de los operadores genéticos [80, 81].

Para usar GA, primero se deben codificar las soluciones del problema de forma que pueda almacenarse en el ordenador y donde cada solución se considera un genoma o cromosoma. El GA crea una población de genomas a los que aplica operaciones de cruzamiento y mutación para generar nuevos individuos. Los individuos que se seleccionan como candidatos suelen ser los mejores individuos, aunque existen diferentes criterios de selección. La función objetivo es la que determina como de bueno es cada individuo solución. Existen diferentes representaciones para los genomas como cadenas de bits, árboles, listas... Aunque lo importante es que sea una representación completa de la solución al problema que trata de optimizarse. El operador de cruce realiza la combinación entre dos individuos, los padres, para producir uno o más individuos nuevos con el propósito de extender los mejores individuos de la anterior generación a la siguiente, los individuos seleccionados para el cruce son elegidos en base a un operador de selección que no elige solamente los mejores individuos porque entonces el algoritmo convergería rápidamente a esa solución sino que selecciona otros no tan bueno pero con la esperanza de que aporten algo útil en el cruce genético. El operador de mutación tiene como objetivo introducir una cierta cantidad de aleatoriedad en la búsqueda con la intención de explorar nuevas soluciones a los que el operador de cruce por sí solo no encontraría. El operador de reemplazo es útil cuando se tienen varias poblaciones al mismo tiempo durante el GA.

Los GA tienen características que los hacen una técnica muy valiosa para problemas de optimización: no se quedan atrapados en mínimos locales (optimización global), pueden tratar problemas discretos y continuos, y además se pueden paralelizar fácilmente, sin embargo eso no resta el hecho de que sea una técnica costosa computacionalmente. En general, los GA obtienen mejor rendimiento que los métodos de búsqueda basados en gradiente si el espacio de búsqueda tiene muchos óptimos locales.

4.4. Clustering

El clustering o agrupamiento se considera el problema más importante del aprendizaje no supervisado, en el que se trata estructurar un conjunto de datos no etiquetados. Es decir, cada cluster es una colección de objetos que son similares entre sí y no con el resto de objetos que pertenecen a otros clusters tal como se puede ver un ejemplo en la Figura 4.6, donde el criterio de similaridad es la distancia. De modo que el objetivo del clustering es determinar el agrupamiento intrínseco que existe en un conjunto de datos no etiquetados, pero decidir cual es un buen agrupamiento es una tarea difícil y depende de la información y configuración de parámetros suministrados al algoritmo utilizado. Existen numerosos algoritmos [82, 83] de clustering que pueden dividirse en los jerárquicos, tanto aglomerativos como divisivos, y los no jerárquicos en los que el número de cluster se determina de antemano y las observaciones se van asignando a los cluster en función de su cercanía.

A menudo, el algoritmo de clustering más apropiado para un problema particular se

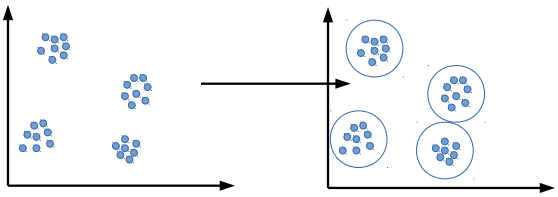


Figura 4.6: Proceso de clustering sobre un conjunto de datos dado

necesita elegir de forma experimental. En este caso, se ha optado por utilizar el algoritmo de clustering Quality Threshold Clustering (QT-Clustering). Este algoritmo fue inventado inicialmente para el clustering en genética [84]. Aunque es más costoso computacionalmente que el algoritmo K-Means por ejemplo, tiene la ventaja de que no requiere especificar el número de cluster y es predecible al garantizar la misma partición en diferentes ejecuciones. El parámetro de entrada es un diámetro umbral o *quality threshold* que viene a determinar el diámetro máximo de los cluster, es decir la máxima distancia permisible entre el centro y cualquier otro punto del cluster siendo el centro aquel punto con mínima distancia al resto. El algoritmo se divide en dos fases: en primer lugar empieza con un conjunto global que incluye todos los puntos y cada punto sirve de base para formar un cluster candidato. De esta forma, a cada cluster candidato de cada punto se van uniendo los puntos más cercanos uno a uno hasta que el diámetro del cluster sobrepase el umbral establecido, por tanto al final habrá un cluster candidato por cada punto de como máximo el diámetro umbral y en el que puede haber cluster que compartan puntos. El segundo paso es seleccionar uno de los candidatos para convertirlo en un QT-cluster real, éste es aquel que tiene mayor número de puntos y estos ya no se tienen en cuenta en el momento de volver al paso anterior de generación de cluster candidatos. El proceso completo se repite hasta que todo punto está asignado a un cluster real.

4.5. Modelo de predicción a corto plazo

La predicción de la radiación solar a corto plazo trata con problemas de predicción con horizontes de tiempo del orden de minutos u horas sin llegar a alcanzar más de un día. Para ello, el modelo se va a basar en las SVM y más concretamente en ε-SVR (Subsección 4.1.3). El problema de regresión está basado en el análisis de las series temporales no lineales con el objetivo de reconstruir el espacio de estados (Sección 4.2), con lo cual se debe buscar los parámetros τ y d para resolver la Ecuación 4.21. Esto se va a realizar empleando algoritmos genéticos (Sección 4.3) para realizar una búsqueda global en la que también se incluirá como parámetro el tipo de función kernel utilizado por el modelo (Ecuación 4.13). En el problema de optimización, el objetivo es reducir el error de predicción por lo que la función objetivo tratará de minimizar el error cuadrático medio (MSE), pero también se

incluye un factor para ponderar esta puntuación en base al máximo error obtenido con la intención de desmerecer aquellos modelos menos fiables:

$$f(\hat{Y}, Y) = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2 + \left(1 - \frac{1}{e^{\text{máx}(|\hat{y}_i - y_i|)}}\right), \quad \forall i = 1, \ldots, n \tag{4.25}$$

donde el primer término corresponde al MSE y el segundo al factor de ponderación del error máximo. Así mismo, hay que tener en cuenta de que el conjunto de datos utilizado para el entrenamiento de los modelos (Tabla 3.5) es distinto al que se usa para la puntuación y validación de cada uno (Tabla 3.6), de esta forma se persigue obtener modelos con mayor capacidad de generalización y disminuir el sobreajuste. La variable índice de nubosidad (Subsección 3.2.4) también forma parte del conjunto de datos aunque no como entrada con retardos.

Los parámetros utilizados para la optimización mediante GA se puede ver en la Tabla 4.1. Mientras que los resultados de las mejores combinaciones de parámetros obtenidas de la optimización genética se pueden ver en la Tabla 4.2 para el modelo a una hora y en la Tabla 4.3 para el modelo a 12 horas, ambos en orden descendiente de mejor a peor solución obtenida.

4.6. Modelo de predicción a medio plazo

Para el modelo de predicción a medio plazo se ha utilizado un planteamiento diferente. En este caso el horizonte de predicción es de uno a varios días y por ello el enfoque que se le ha dado es distinto al anterior que está basado en la regresión de la serie temporal. El enfoque que se propone se basa en la utilización de un proceso de clustering (Sección 4.4) para la obtención de un conjunto de perfiles diarios de radiación solar que sean representativos del conjunto de datos utilizados y por consiguiente de la zona local en la que se realiza la calibración de este modelo. De esta forma lo que se pretende es obtener una estimación del perfil de radiación desde un día hasta varios días más adelante y esto en base al perfil de los días anteriores así como al índice de nubosidad del día. Además, el proceso de clustering lleva consigo un preproceso de los datos para reducir la dimensionalidad de los datos ya que cada día contiene $24\,\text{horas} \cdot 60\,\text{minutos} \cdot 60\,\text{segundos} = 86400$ puntos de datos, para ello primeramente se realiza un remuestreo cada 60 puntos y posteriormente se realiza la transformada discreta de Fourier (DFT) con lo que se genera un nuevo vector de características formado por los l primeros coeficientes de la DFT de la señal original puesto que se pasa de trabajar de un espacio de dimensión m (en el dominio del tiempo) a un espacio de dimensión inferior l (en el dominio de la frecuencia). Una característica importante de la DFT es que preserva la distancia euclídea entre dos señales tanto en el dominio del tiempo como en el de la frecuencia. Dado una

Parámetro	**Valor**
Parámetros a optimizar	$[\tau, d, K]$
Valores mínimos de los parámetros	$[1, 1, 0]$
Valores máximos de los parámetros	$[120, 30, 3]$
Función objetivo	$f(\hat{Y}, Y) = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2 + \left(1 - \frac{1}{e^{\text{máx}(\lvert\hat{y}_i - y_i\rvert)}}\right)$
Función para escalar puntuación	Escala según *ranking* del individuo
Número de generaciones	5
Individuos por generación	40
Función de creación de población inicial	Uniforme
Élite con garantía de supervivencia	$0{,}05 \cdot \text{mín}(\text{máx}(10 \cdot \text{numVariables}, 40), 100)$
Índice de cruzamiento (aparte de élite)	0,8
Función de selección de padres	Estocástica uniforme según puntuación
Función de cruzamiento	Combinación dispersa
Función de mutación	Gaussiana
Intervalo de migración	20
Índice de migración	0,2
τ: retardo de inmersión d: dimensión de inmersión (Sección 4.2) K: Tipo de kernel (Ecuación 4.13)	

Tabla 4.1: Parámetros de configuración para la optimización con algoritmos genéticos (GA) del modelo de predicción a corto plazo

señal $\vec{x} = [x_0, x_1, \ldots, x_{m-1}]$, la DFT es una secuencia $\vec{X}$ de números complejos X_f, $f = 0, 1, \ldots, m-1$:

$$X_f = \frac{1}{\sqrt{m}} \sum_{i=0}^{m-1} x_i \cdot \exp(-j\frac{2\pi}{m}\imath f), \quad \forall f = 0, 1, \ldots, m-1 \tag{4.26}$$

El resultado puede verse en la Figura 4.7, donde se ven los 12 perfiles diarios (cluster) obtenidos. Los cuales servirán para estimar la predicción a medio plazo de los días siguientes. También hay que mencionar que uno de los perfiles es nulo, de lo que se puede extraer que el sensor no es perfecto y un preprocesado más minucioso sería recomendable.

El siguiente paso es construir y entrenar el modelo predictivo utilizando los conjuntos de datos descritos en la Tabla 3.9. Faltaría estimar el número de días pasados que se utilizan para predecir el perfil del día siguiente y para ello se va realizar una construcción iterativa de modelos utilizando un número diferente de días pasados utilizados en la estimación del siguiente perfil. Como se puede observar en la Tabla 4.4, a partir de 22 días se obtiene una precisión del 100 % en los datos de entrenamiento con lo que puede empezar por utilizar dicho valor $d = 22$.

τ	**d**	**Kernel**	ε	**no. SV**	**Puntuación**	**MSE**	**ErrorMáx**	$\mathbf{R^2}$
108	15	RBF	0,004836	9645	0,31288	0,0015184	0,37304	0,97414
109	29	RBF	0,004990	9303	0,31349	0,0015615	0,37386	0,97385
116	17	RBF	0,004613	9579	0,3199	0,0015908	0,38319	0,97337
118	16	RBF	0,004678	9594	0,32037	0,0015911	0,38387	0,97328
75	23	RBF	0,004619	9620	0,32456	0,0015588	0,39009	0,97365

Tabla 4.2: Resultados de la optimización de parámetros para el modelo de predicción a 1 hora

τ	**d**	**Kernel**	ε	**no. SV**	**Puntuación**	**MSE**	**ErrorMáx**	$\mathbf{R^2}$
113	21	RBF	0,007182	9305	0,20998	0,0016288	0,23363	0,97256
109	24	RBF	0,007019	9263	0,22035	0,0017039	0,24673	0,97177
110	21	RBF	0,007031	9326	0,22139	0,0016729	0,2481	0,97178
113	25	RBF	0,007107	9218	0,2287	0,00171594	0,25746	0,97178
111	24	RBF	0,007245	9247	0,22953	0,0017198	0,25852	0,97142

Tabla 4.3: Resultados de la optimización de parámetros para el modelo de predicción a 12 horas

Días pasados	**Precisión**
1	67 % (67/100)
2	65,97 % (64/97)
3	67,02 % (63/94)
4	70,32 % (64/91)
5	71,59 % (63/88)
6	72,94 % (62/85)
7	79,26 % (65/82)
8	78,48 % (62/79)
9	77,63 % (59/76)
10	78,08 % (57/73)
11	81,42 % (57/70)
12	83,58 % (56/67)

Días pasados	**Precisión**
13	82,81 % (53/64)
14	85,24 % (52/61)
15	93,10 % (54/58)
16	94,54 % (52/55)
17	94,23 % (49/52)
18	93,87 % (46/49)
19	95,65 % (44/46)
20	95,34 % (41/43)
21	95 % (38/40)
22	100 % (37/37)
23	100 % (34/34)
24	100 % (31/31)

Tabla 4.4: Resultados del entrenamiento del modelo de predicción a medio plazo

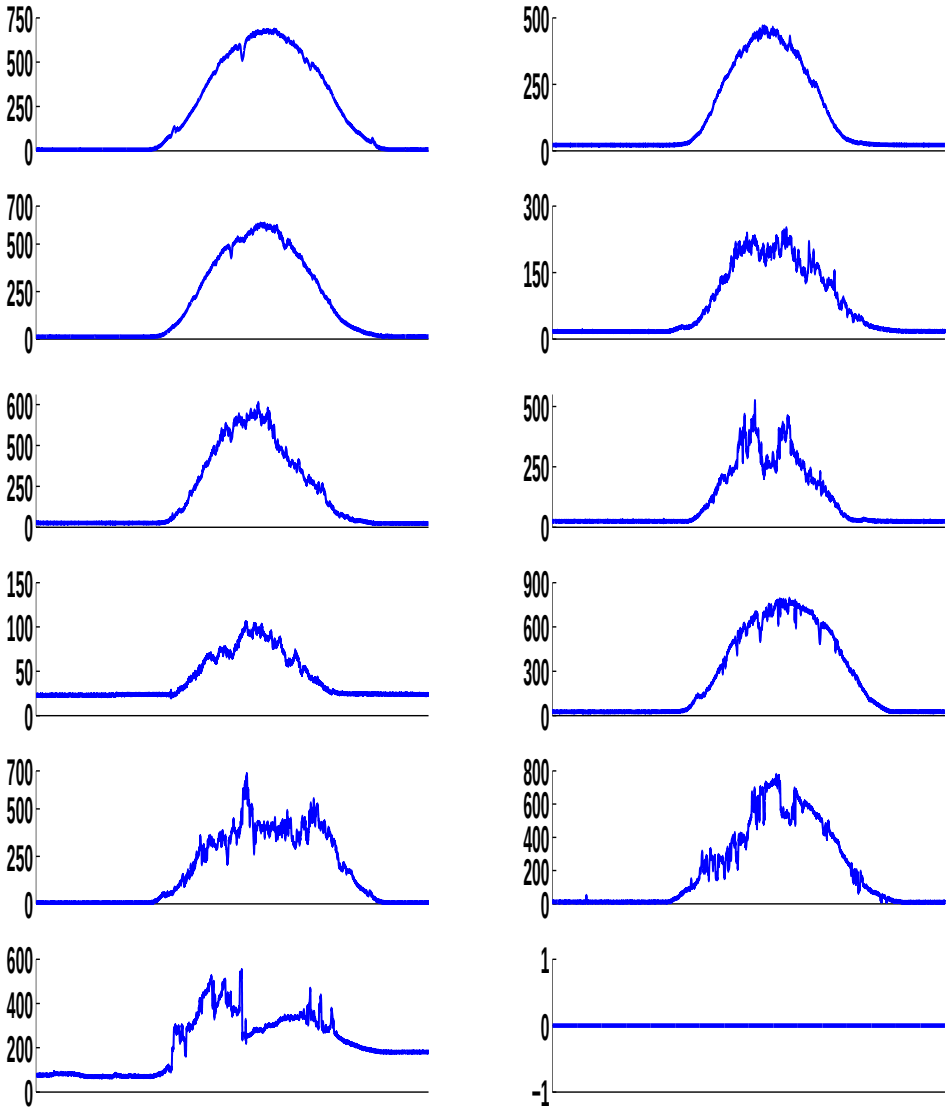


Figura 4.7: Perfiles diarios de radiación solar obtenidos mediante el proceso de clustering

5 RESULTADOS

En este capítulo se presentan los resultados de predicción de la radiación solar. Para evaluar la bonanza del modelo de estimación se ha utilizado la medida estadística del Coeficiente de Variación de la Raíz Cuadrada del Error Cuadrático Medio (CV-RMSE), que es una medida que no depende de la escala y cuya fórmula es:

$$CV - RMSE = \frac{RMSE}{\bar{Y}} \tag{5.1}$$

donde RMSE es la raíz cuadrada del error cuadrático medio:

$$RMSE = \sqrt{MSE} \tag{5.2}$$

y el error cuadrático medio (MSE) se calcula de la siguiente manera:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(\hat{Y}_i - Y_i)^2 \tag{5.3}$$

siendo $\hat{Y}$ el vector con las n predicciones y Y el vector con los valores reales.

A modo de complementar y para conocer la bondad de los resultados en su escala, también se ha utilizado el Error Absoluto Medio (MAE):

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|\hat{Y}_i - Y_i| \tag{5.4}$$

5.1. Predicción a corto plazo

La predicción a corto plazo es inferior a 1 día. En concreto se ha utilizado un modelo para la predicción a una hora y otro modelo para la predicción a 12 horas, esto es que en el instante i, el modelo produce la predicción para el instante $\hat{x}(i+n)$. Como batería de pruebas se ha utilizado 4 días con diferentes condiciones de radiación, de

Test	Día	Estación	Perfil de radiación
A	20/01/2014	Invierno	Soleado
B	10/01/2014	Invierno	Nublado
C	12/06/2014	Verano	Soleado
D	10/06/2014	Verano	Nublado

Tabla 5.1: Batería de pruebas para el modelo de predicción a corto plazo

diferente estación y que no se han empleado en el entrenamiento o validación del modelo (Tabla 5.1).

En el caso del modelo de predicción a una hora ($\hat{x}(i+1\,\text{hora})$) se ha utilizado la mejor combinación de parámetros obtenida en el proceso de entrenamiento y optimización (Tabla 4.2) con $\tau = 108$, $d = 15$ y función de kernel RBF (función de base radial). Como se ve en la tabla Tabla 5.2, los resultados de predicción a 1 hora oscilan entre un error del $[6 - 24]\,\%$ lo cual índica un buen índice de predicción si además se tiene en cuenta que son días totalmente nuevos para el modelo y que el día con peor resultado ($24\,\%$) es un día nublado de verano, y probablemente habrá pocos dias con este perfil en el conjunto de datos. Además, observando las gráficas de predicción (Figura 5.1), se ve como el modelo predice con buen criterio la dinámica de la radiación solar.

Test	CV-RMSE [%]	MAE $[W \cdot m^{-2}]$	Error Máximo $[W \cdot m^{-2}]$
A	17,36	14,940	105,901
B	17,56	7,579	56,408
C	6,96	10,251	78,591
D	24,45	18,742	192,119

Tabla 5.2: Resultados descriptivos de la predicción a corto plazo de la radiación solar (1 hora)

Para el modelo de predicción a 12 horas ($\hat{x}(i + 12\,\text{horas})$) también se ha utilizado la mejor combinación de parámetros obtenida en el proceso de entrenamiento y optimización (Tabla 4.3) con $\tau = 113$, $d = 21$ y función de kernel RBF (función de base radial). El error CV-RMSE oscila entre $[11 - 35]\,\%$, que es mayor que para el caso anterior de la predicción a 1 hora, aunque el modelo sigue prediciendo la dinámica de la evolución de la radiación solar, ofreciendo un buen rendimiento.

Test	CV-RMSE [%]	MAE $[W \cdot m^{-2}]$	Error Máximo $[W \cdot m^{-2}]$
A	35,52	28,448	185,247
B	20,31	9,639	42,797
C	11,78	14,966	129,890
D	16,34	15,838	75,082

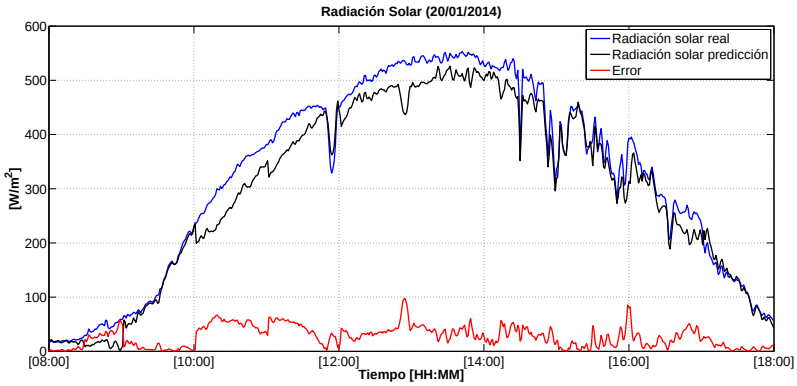
Tabla 5.3: Resultados descriptivos de la predicción a corto plazo de la radiación solar (12 horas)

5.2. Predicción a medio plazo

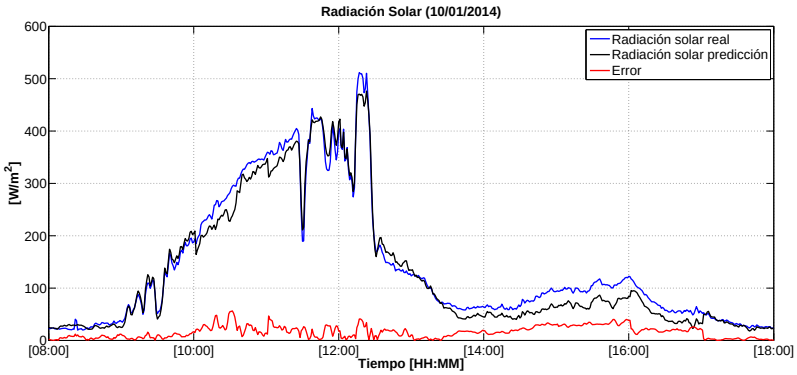
Para la predicción a medio plazo, se ha utilizado un conjunto de datos perteneciente a verano de 2014 y que ha sido empleado en el entrenamiento del modelo debido a la limitación en la cantidad de datos para este tipo de predicción (Tabla 3.9). Concretamente, se ha realizado una predicción hasta 14 días realimentando la salida del modelo para predecir el siguiente día y utilizando como entradas el perfil de radiación y el índice de nubosidad de los $d = 22$ días anteriores, es decir que en el instante i el modelo toma la información de los 22 días pasados y construye de forma iterativa la predicción de los 14 siguientes $\hat{x}(i+14)$ añadiendo la propia salida del modelo como entrada en cada iteración. Los resultados se pueden observar en la Tabla 5.4, donde se ve que para todos los días el modelo realiza una predicción correcta del perfil de radiación con el que han sido etiquetados en el proceso de clustering los 14 días y que en general se obtiene un CV-RMSE que ronda el $[10 - 30]$ % salvo para el día número 10 en el que el error es del 66,22 %, lo que quiere decir que hay una buena sintonía en la generación y asignación de los perfiles diarios de radiación solar. Mirando las gráficas (Figura 5.3-Figura 5.6), primero se observa que los perfiles de radiación solar que el modelo ha estimado para cada día adopta la dinámica de cada día en cuanto a valores máximos, tendencias y oscilaciones, aunque es complicado acertar con exactitud puesto que se trabaja sobre una base reducida de perfiles. Además se ve como algunos como el día 11 presenta un perfil atípico. En definitiva, los resultados son aceptables, pero una mayor cantidad y calidad de datos así como un preprocesamiento más minucioso puede conducir a una mejora considerable de este enfoque de predicción.

Día de predicción	**CV-RMSE** [%]	**MAE** $[W \cdot m^{-2}]$	**Error Máximo** $[W \cdot m^{-2}]$	**Predicción perfil**	**Perfil real**
1	26,01	61,062	253,173	3	3
2	28,77	59,465	458,771	3	3
3	12,63	33,481	412,614	1	1
4	8,06	21,968	143,777	1	1
5	10,19	26,580	189,385	1	1
6	16,29	27,383	580,257	1	1
7	7,89	22,439	138,621	1	1
8	19,40	45,230	196,834	3	3
9	19,35	47,184	151,254	3	3
10	66,22	72,535	416,271	4	4
11	35,53	90,595	270,429	11	11
12	0,0003	0,002	0,0005	8	8
13	23,35	50,370	305,592	3	3
14	23,07	50,550	580,350	3	3

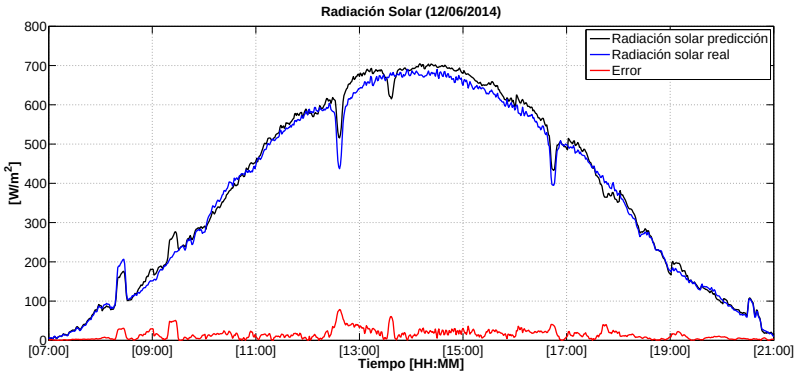
Tabla 5.4: Resultados descriptivos de la predicción a medio plazo de la radiación solar (1 a 14 días)



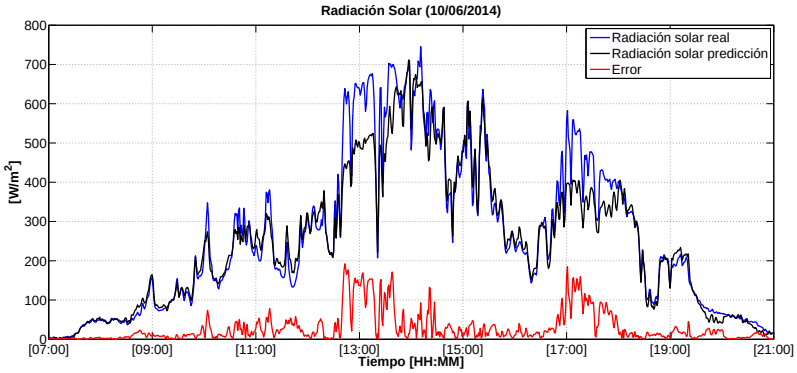
(a) Día de invierno soleado: 20/01/2014



(b) Día de invierno nublado: 10/01/2014

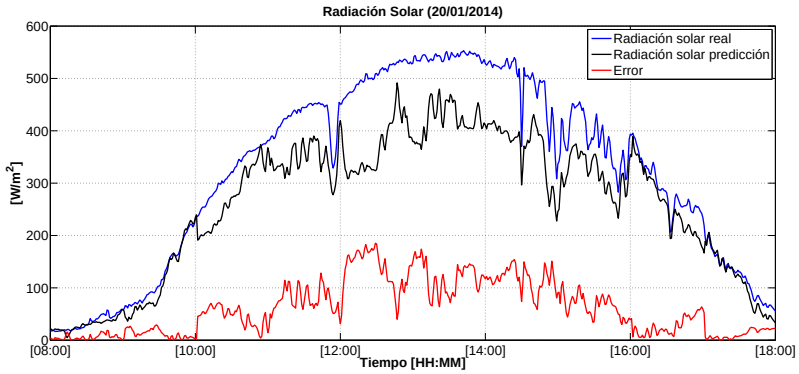


(c) Día de verano soleado: 12/06/2014

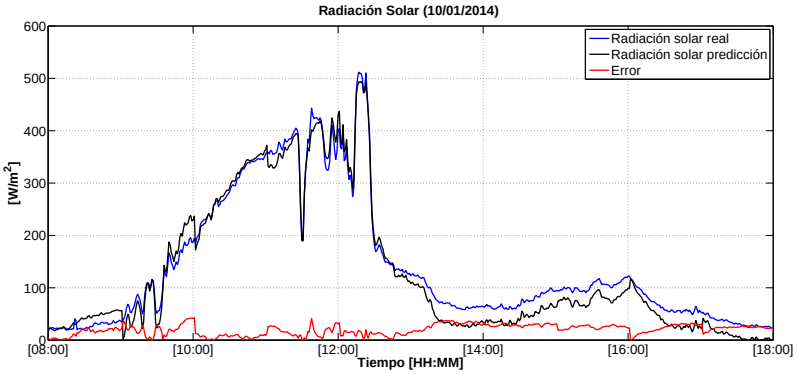


(d) Día de verano nublado: 10/06/2014

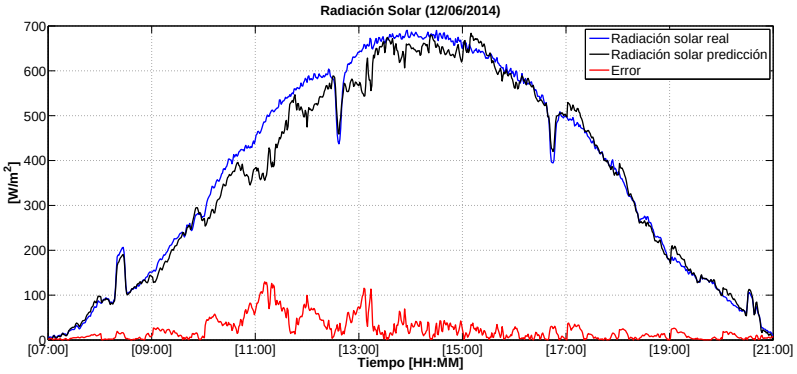
Figura 5.1: Predicción a corto plazo (1 hora)



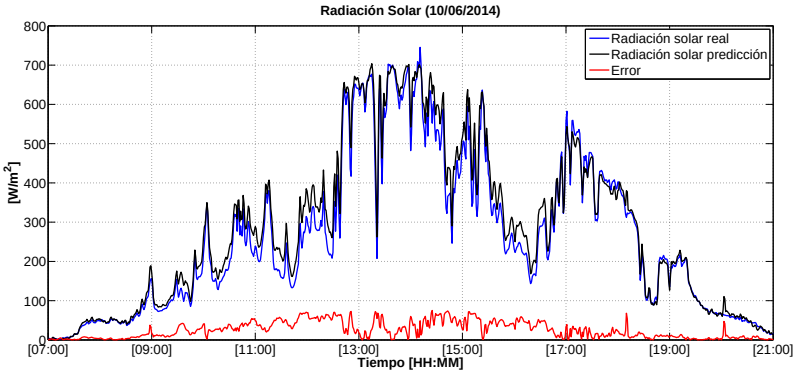
(a) Día de invierno soleado: 20/01/2014



(b) Día de invierno nublado: 10/01/2014

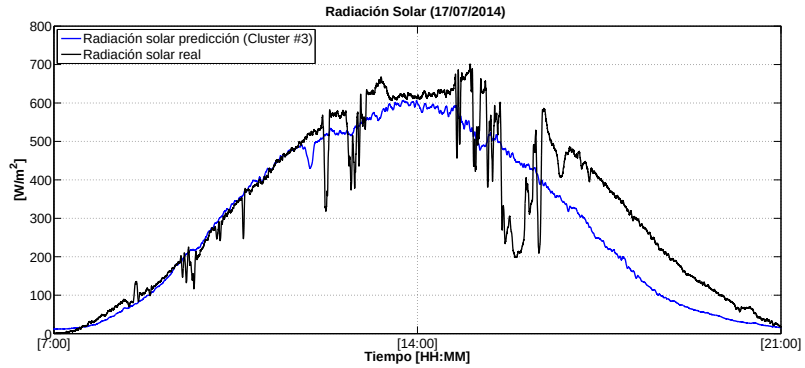


(c) Día de verano soleado: 12/06/2014

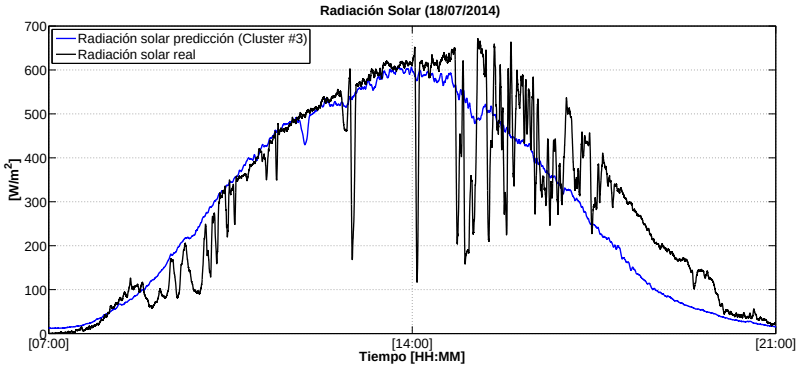


(d) Día de verano nublado: 10/06/2014

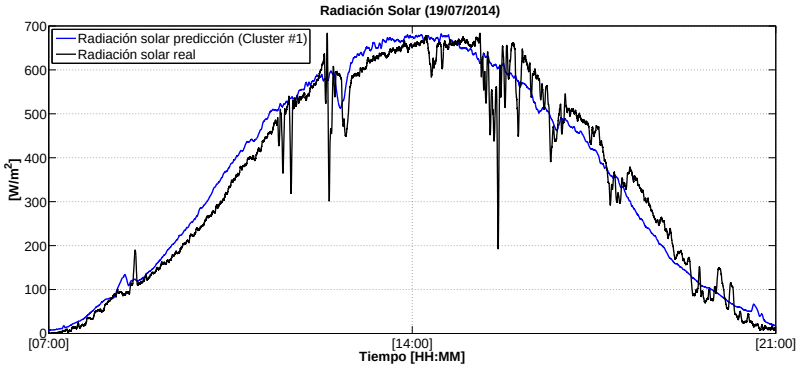
Figura 5.2: Predicción a corto plazo (12 horas)



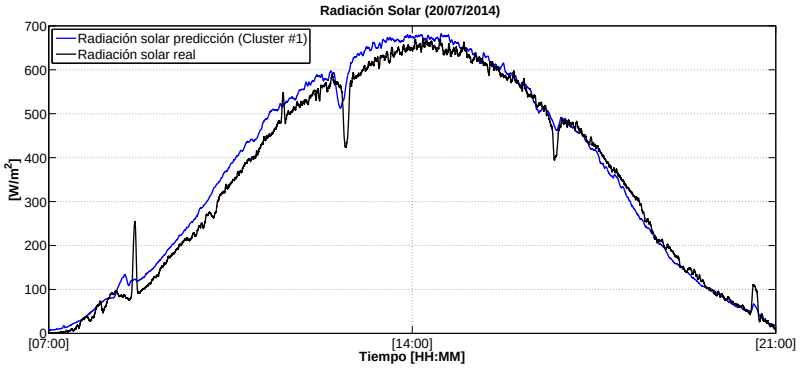
(a) Día de predicción 1: 17/07/2014



(b) Día de predicción 2: 18/07/2014

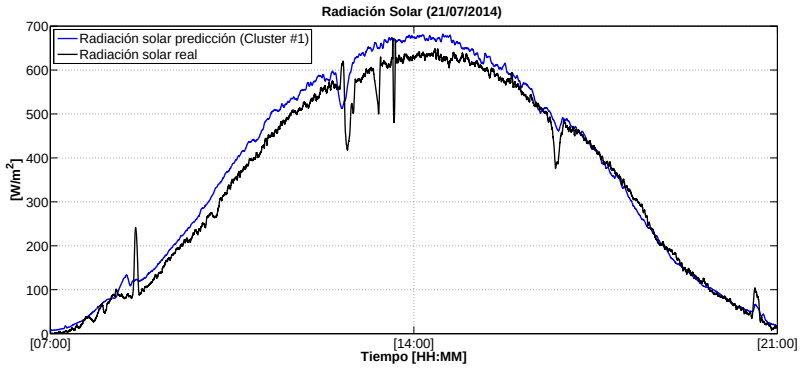


(c) Día de predicción 3: 19/07/2014

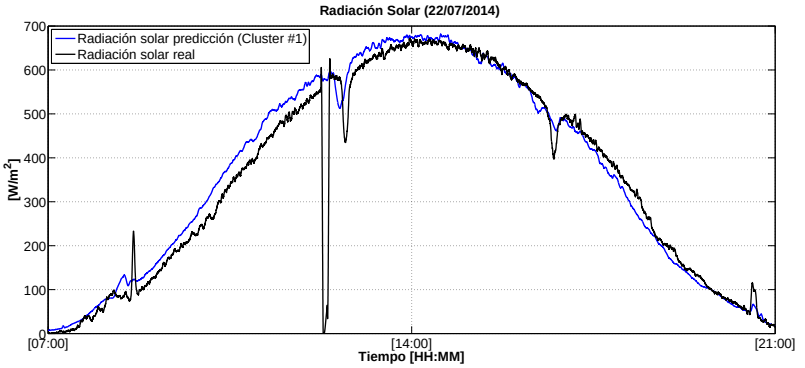


(d) Día de predicción 4: 20/07/2014

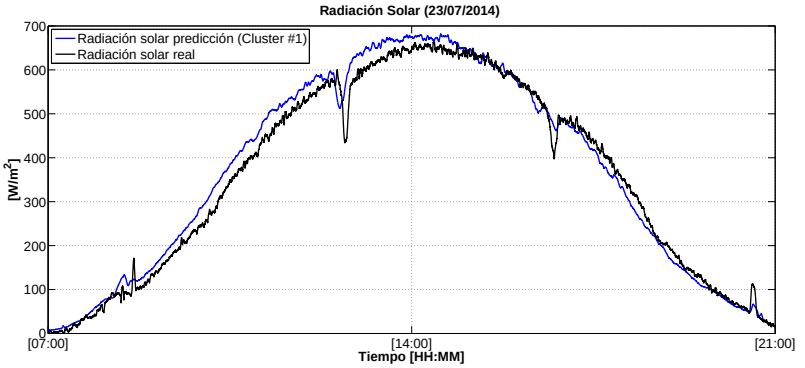
Figura 5.3: Predicción a medio plazo de perfiles de radiación solar (1 a 4 días)



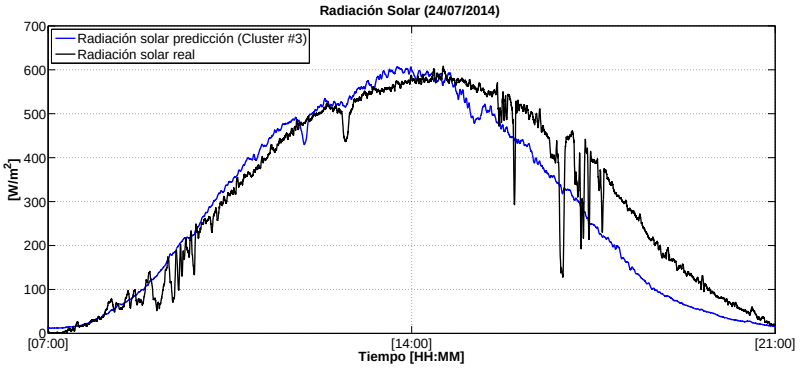
(a) Día de predicción 5: 21/07/2014



(b) Día de predicción 6: 22/07/2014

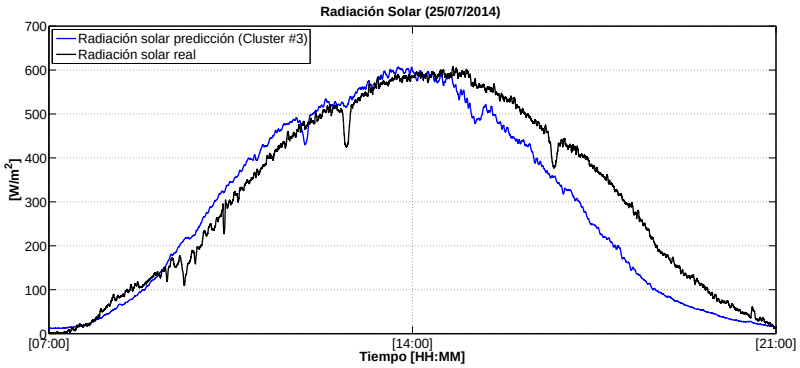


(c) Día de predicción 7: 23/07/2014

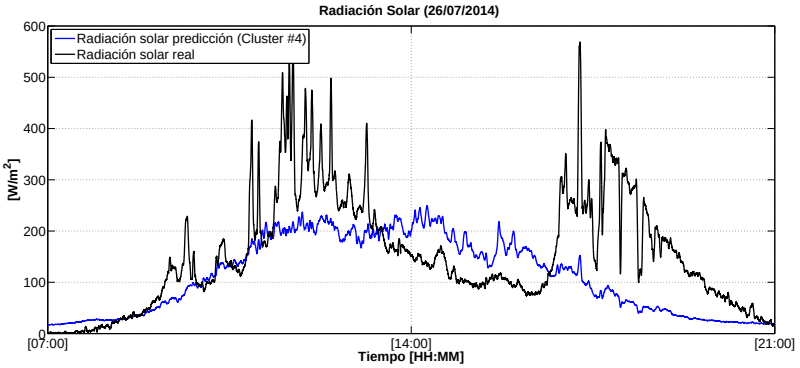


(d) Día de predicción 8: 24/07/2014

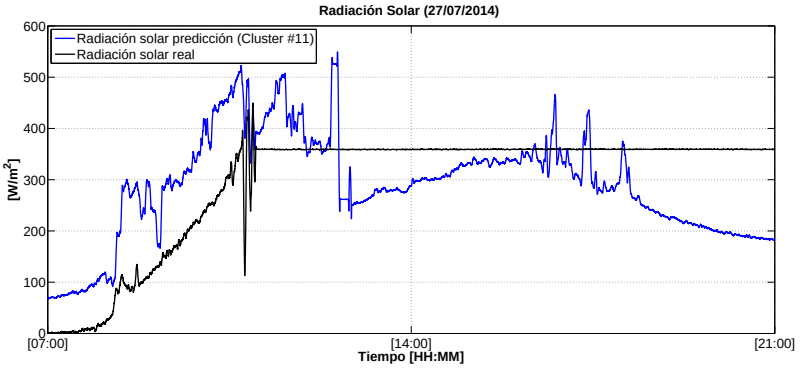
Figura 5.4: Predicción a medio plazo de perfiles de radiación solar (5 a 8 días)



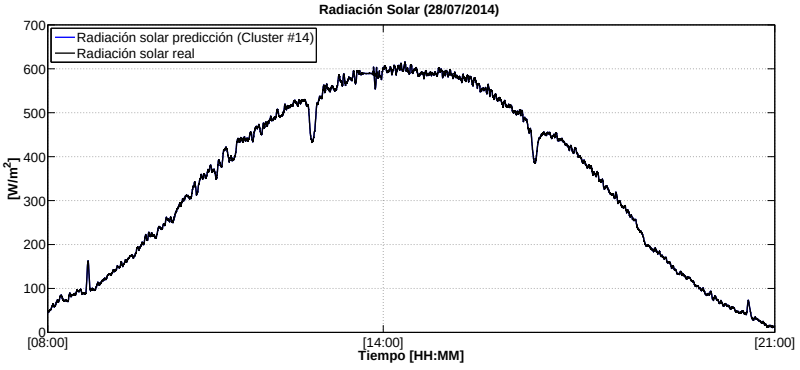
(a) Día de predicción 9: 25/07/2014



(b) Día de predicción 10: 26/07/2014

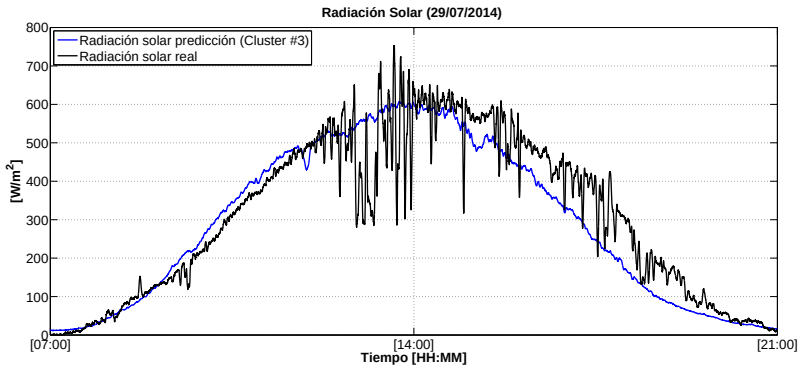


(c) Día de predicción 11: 27/07/2014

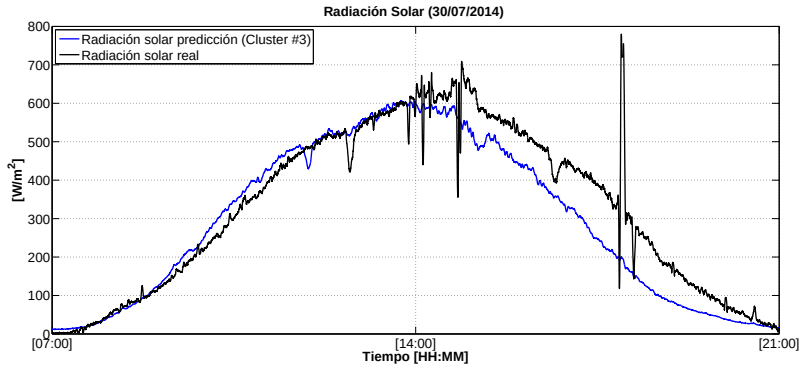


(d) Día de predicción 12: 28/07/2014

Figura 5.5: Predicción a medio plazo de perfiles de radiación solar (9 a 12 días)



(a) Día de predicción 13: 29/07/2014



(b) Día de predicción 14: 30/07/2014

Figura 5.6: Predicción a medio plazo de perfiles de radiación solar (13 a 14 días)

6 CONCLUSIONES Y FUTUROS TRABAJOS

En este trabajo se ha pretendido abordar el problema de la predicción de la radiación solar tanto a corto plazo (horas) como a medio plazo (días) desde un punto de vista de la inteligencia artificial y el análisis de datos, y con la intención de dar solución a un problema común en el campo del control automático como es la predicción de las perturbaciones. En especial, la radiación solar es un factor crítico y limitante para muchos sistemas basados en este recurso. En el caso concreto de los fotobiorreactores de microalgas, conocer esta variable y su dinámica es fundamental para plantear estrategias avanzadas de control predictivo que optimicen los recursos y aumenten la productividad. A partir de una base de datos recolectada de una planta real de fotobiorreactores, se plantea una combinación de estrategias basadas en el análisis de datos y el aprendizaje automático para el desarrollo de modelos predictivos a corto y medio plazo de la radiación solar, en concreto los modelos se basan en las máquinas de vectores de soporte (SVM) para la tarea de aprendizaje automático y predicción, así como en la teoría de inmersión para el análisis de series temporales de datos no lineales y la reconstrucción del espacio de estado utilizando optimización global mediante algoritmos genéticos (GA) para el ajuste de parámetros. Así mismo, se ha planteado la utilización de algoritmos de clustering para la generación de perfiles diarios de radiación solar como base del planteamiento de predicción a medio plazo. Al tratarse de modelos basados en datos y aprendizaje automático, también se ha empleado otra variable que juega un papel directo en la radiación solar terrestre: la nubosidad, para ello se ha planteado una estrategia para caracterizar el conjunto de datos histórico así como crear el soporte para poder emplear predicciones externas de nubosidad de fuentes externas cuando se utilice el modelo en tiempo real.

Los resultados obtenidos en el modelo de predicción a corto plazo son muy favorables. En concreto, para la predicción a una hora los resultados oscilan entre el 6 % y el 24 % de CV-RMSE para el peor caso teniendo encuentra la diversidad de condiciones y que son días no presentados en el entrenamiento del modelo. Además el modelo reconoce con bastante buen criterio la dinámica de la radiación solar, lo cual es muy útil en tareas de control. Para el caso de la predicción a corto plazo de 12 horas se han obtenido resultados con un ligero incremento en el error, pero similar en el sentido de que captan bien la dinámica de la evolución de la radiación solar. El error (CV-RMSE) para las pruebas de predicción a 12 horas oscila entre el 11 % y el 35 %.

El problema de la predicción a medio plazo se ha abordado con otro tipo de enfoque.

En lugar de construir un modelo cuya finalidad sea predecir mediante regresión los valores futuros, se plantea una estrategia para obtener una serie de perfiles diarios de radiación solar que sean representativos del conjunto de datos utilizado y por consiguiente de la localidad en la que se calibra y se utilizará el modelo. Los resultados de este proceso ya son útiles de por sí como base para el análisis de datos y evaluación de estrategias de control, pero además la utilización de estos perfiles de radiación junto con el valor de nubosidad permite al modelo realizar predicciones de los perfiles de radiación de hasta 14 días posteriores de forma recursiva y con un buen índice de precisión en la clasificación y ajuste del perfil. En concreto, en la pruebas realizadas se ha hecho una predicción de hasta 14 días estimando de manera iterativa el perfil del siguiente día, lo cual se ha conseguido con éxito, y además se comprueba el perfil asignado con los valores reales de radiación solar de ese día para comprobar que las dinámicas son efectivamente similares así como otras características de la señal: valor máximo, tendencia. En general, los errores de predicción oscilan entre el 7 % y el 35 %, con un peor caso del 66 % para un día concreto con un perfil de radiación muy oscilatorio y con ruido presente. Esto quiere decir que hay una buena sintonía en la generación y asignación de los perfiles diarios de radiación solar. De nuevo, la utilización de la variable de nubosidad permite utilizar el modelo en tiempo real utilizando información de fuentes externa de meteorología. Este tipo de planteamiento y los resultados obtenidos son muy prometedores, sin embargo se debe hacer especial hincapié y es una consideración para trabajos futuros en la cantidad y calidad de los datos utilizados así como en realizar un preprocesamiento de los mismos aún más minucioso para evitar algunas anomalías presentes y que se han ido descubriendo gracias a este planteamiento.

Por último, como líneas de trabajo futuras aparte de trabajar sobre una base de datos más elaborada, existe la continuidad natural de aplicar este trabajo a tareas de control automático para servir como solución a los problemas de predicción de perturbaciones. Dentro del área de la predicción de radiación solar se puede plantear la utilización de nuevas variables, o incluso la utilización de imágenes de satélite o cámara todo-cielo para su posterior análisis mediante algoritmos de visión artificial para incorporar información adicional sobre la nubosidad en tiempo real al modelo. Y como extensión del aprendizaje automático también se plantea la posibilidad de desarrollar la metodología en otro contexto o localizaciones, así como la utilización en otros ámbitos o áreas de investigación con el suficiente conocimiento del dominio.

Bibliografía

[1] UNFCCC, "United nations framework convention on climate change," 2010.

[2] REN21, "Renewables 2010 global status report," tech. rep., Paris: REN21 Secretariat. Deutsche Gesellschaft für Technische Zusammenarbeit (GTZ) GmbH, 2010.

[3] O. Bernard, "Hurdles and challenges for modelling and control of microalgae for co2 mitigation and biofuel production," *Journal of Process Control*, vol. 21, no. 10, pp. 1378–1389, 2011.

[4] N.-H. Norsker, M. J. Barbosa, M. H. Vermuë, y R. H. Wijffels, "Microalgal production - a close look at the economics," *Biotechnology Advances*, vol. 29, no. 1, pp. 24–27, 2011.

[5] M. Berenguel, F. Rodrıguez, F. Acién, y J. Garcıa, "Model predictive control of ph in tubular photobioreactors," *Journal of Process Control*, vol. 14, no. 4, pp. 377–387, 2004.

[6] I. Fernández, J. Peña, J. Guzmán, M. Berenguel, y F. Acién, "Modelling and control issues of ph in tubular photobioreactors," en *Proceedings of the 11th IFAC Symposium on Computer Applications in Biotechnology (CAB 2010)*, (Leuven, Belgium.), 2010.

[7] M. Berenguel, M. Arahal, y E. Camacho, "Modelling the free response of a solar plant for predictive control," *Control engineering practice*, vol. 6, no. 10, pp. 1257–1266, 1998.

[8] A. Pawlowski, J. L. Guzmán, F. Rodrıguez, M. Berenguel, y J. E. Normey-Rico, "Predictive control with disturbance forecasting for greenhouse diurnal temperature control," en *Proceedings of the 18th World Congress of IFAC, Milan, Italy*, 2011.

[9] Y. Chisti, "Biodiesel from microalgae," *Biotechnology advances*, vol. 25, no. 3, pp. 294–306, 2007.

[10] C.-Y. Chen, G. D. Saratale, C.-M. Lee, P.-C. Chen, y J.-S. Chang, "Phototrophic hydrogen production in photobioreactors coupled with solar-energy-excited optical fibers," *International Journal of Hydrogen Energy*, vol. 33, no. 23, pp. 6886–6895, 2008.

[11] C.-Y. Chen, K.-L. Yeh, R. Aisyah, D.-J. Lee, y J.-S. Chang, "Cultivation, photobioreactor design and harvesting of microalgae for biodiesel production: a critical review," *Bioresource technology*, vol. 102, no. 1, pp. 71–81, 2011.

[12] A. Ramírez-Arias, F. Rodríguez, J. L. Guzmán, y M. Berenguel, "Multiobjective hierarchical control architecture for greenhouse crop growth," *Automatica*, vol. 48, no. 3, pp. 490–498, 2012.

[13] M. Rebolloso Fuentes, J. García Sánchez, J. Fernández Sevilla, F. Acién Fernández, J. Sánchez Pérez, y E. Molina Grima, "Outdoor continuous culture of porphyridium cruentum in a tubular photobioreactor: quantitative analysis of the daily cyclic variation of culture parameters," *Progress in Industrial Microbiology*, vol. 35, pp. 271–288, 1999.

[14] M. R. Arahal, M. Berenguel Soria, y F. Rodríguez Díaz, *Técnicas de predicción con aplicaciones en Ingeniería*. Sevilla: Universidad de Sevilla. Servicio de publicaciones, 2006.

[15] C. Cortes y V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.

[16] C. Cai, L. Han, Z. L. Ji, X. Chen, y Y. Z. Chen, "Svm-prot: web-based support vector machine software for functional classification of a protein from its primary sequence," *Nucleic acids research*, vol. 31, no. 13, pp. 3692–3697, 2003.

[17] C. Schuldt, I. Laptev, y B. Caputo, "Recognizing human actions: a local svm approach," en *ICPR 2004. Proceedings of the 17th International Conference on Pattern Recognition*, vol. 3, pp. 32–36, IEEE, 2004.

[18] A. Mathur y G. Foody, "Multiclass and binary svm classification: Implications for training and classification users," *Geoscience and Remote Sensing Letters, IEEE*, vol. 5, no. 2, pp. 241–245, 2008.

[19] R. Cogdill y P. Dardenne, "Least-squares support vector machines for chemometrics: an introduction and evaluation," *Journal of Near Infrared Spectroscopy*, vol. 12, no. 2, pp. 93–100, 2004.

[20] L. Xia, J. Meng, R. Xu, B. Yan, y Y. Guo, "Modeling of 3-d vertical interconnect using support vector machine regression," *Microwave and Wireless Components Letters, IEEE*, vol. 16, no. 12, pp. 639–641, 2006.

[21] F. E. Tay y L. Cao, "Application of support vector machines in financial time series forecasting," *Omega*, vol. 29, no. 4, pp. 309–317, 2001.

[22] C. Sivapragasam, S. Liong, y M. Pasha, "Rainfall and runoff forecasting with ssa-svm approach," *Journal of Hydroinformatics*, vol. 3, pp. 141–152, 2001.

[23] X. Yu, S. Liong, y V. Babovic, "Ec-svm approach for real-time hydrologic forecasting.," *Journal of Hydroinformatics*, vol. 6, pp. 209–223, 2004.

[24] Z. Zeng, H. Yang, R. Zhao, y J. Meng, "Nonlinear characteristics of observed solar radiation data," *Solar Energy*, vol. 87, pp. 204–218, 2013.

[25] P. Mantero, G. Moser, y S. B. Serpico, "Partially supervised classification of remote sensing images through svm-based probability density estimation," *Geoscience and Remote Sensing, IEEE Transactions on*, vol. 43, no. 3, pp. 559–570, 2005.

[26] D. C. Montgomery, E. A. Peck, y G. G. Vining, *Introduction to linear regression analysis*, vol. 821. John Wiley & Sons, 2012.

[27] S. Geman, E. Bienenstock, y R. Doursat, "Neural networks and the bias/variance dilemma," *Neural computation*, vol. 4, no. 1, pp. 1–58, 1992.

[28] I. Guyon, V. Vapnik, B. Boser, L. Bottou, y S. Solla, "Capacity control in linear classifiers for pattern recognition," en *Pattern Recognition, 1992. Vol. II. Conference B: Pattern Recognition Methodology and Systems, Proceedings., 11th IAPR International Conference on*, pp. 385–388, IEEE, 1992.

[29] B. Wang, C. Q. Lan, y M. Horsman, "Closed photobioreactors for production of microalgal biomasses," *Biotechnology advances*, vol. 30, no. 4, pp. 904–912, 2012.

[30] C. F. Coimbra, J. Kleissl, y R. Marquez, "Chapter 8-overview of solar-forecasting methods and a metric for accuracy evaluation," en *Solar Energy Forecasting and Resource Assessment* (J. Kleissl, ed.), pp. 171 – 194, Boston: Academic Press, 2013.

[31] D. Renné, "Semi-annual status report: November 2009. task 36: Solar resource knowledge management," tech. rep., IEA. IEA SHC Task 36 Solar Resource Knowledge Management., 2009.

[32] E. Lorenz, J. Hurka, D. Heinemann, y H. G. Beyer, "Irradiance forecasting for the power prediction of grid-connected photovoltaic systems," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 2, no. 1, pp. 2–10, 2009.

[33] E. Lorenz, J. Remund, S. C. Müller, W. Traunmüller, G. Steinmaurer, D. Pozo, J. Antonio, V. L. F. Ruiz-Arias, L. Ramirez, M. G. Romeo, *et al.*, "Benchmarking of different approaches to forecast solar irradiance," en *Proceedings of the 24th European Photovoltaic Solar Energy Conference*, pp. 4199–4208, 2009.

[34] D. S. Wilks, *Statistical methods in the atmospheric sciences*, vol. 100. Academic press, 2011.

[35] G. Reikard, "Predicting solar radiation at high resolutions: A comparison of time series forecasts," *Solar Energy*, vol. 83, no. 3, pp. 342–349, 2009.

[36] A. Mellit, H. Eleuch, M. Benghanem, C. Elaoun, y A. M. Pavan, "An adaptive model for predicting of global, direct and diffuse hourly solar irradiance," *Energy Conversion and Management*, vol. 51, no. 4, pp. 771–782, 2010.

[37] A. Pawlowski, J. L. Guzmán, F. Rodríguez, M. Berenguel, y J. Sanchez, "Application of time-series methods to disturbance estimation in predictive control problems," en *IEEE International Symposium on Industrial Electronics (ISIE)*, pp. 409–414, IEEE, 2010.

[38] A. Mellit, "Artificial intelligence technique for modelling and forecasting of solar radiation data: a review," *International Journal of Artificial intelligence and soft computing*, vol. 1, no. 1, pp. 52–76, 2008.

[39] A. Sfetsos y A. Coonick, "Univariate and multivariate forecasting of hourly solar radiation with artificial intelligence techniques," *Solar Energy*, vol. 68, no. 2, pp. 169–178, 2000.

[40] J. Cao y X. Lin, "Application of the diagonal recurrent wavelet neural network to solar irradiation forecast assisted with fuzzy technique," *Engineering Applications of Artificial Intelligence*, vol. 21, no. 8, pp. 1255–1263, 2008.

[41] E. M. Crispim, P. M. Ferreira, y A. E. Ruano, "Prediction of the solar radiation evolution using computational intelligence techniques and cloudiness indices," *International Journal of Innovative Computing, Information and Control*, vol. 2, p. 2, 2008.

[42] C. Paoli, C. Voyant, M. Muselli, y M.-L. Nivet, "Forecasting of preprocessed daily solar radiation time series using neural networks," *Solar Energy*, vol. 84, no. 12, pp. 2146–2160, 2010.

[43] R. Marquez y C. F. Coimbra, "Forecasting of global and direct solar irradiance using stochastic learning methods, ground experiments and the nws database," *Solar Energy*, vol. 85, no. 5, pp. 746–756, 2011.

[44] H. T. Pedro y C. F. Coimbra, "Assessment of forecasting techniques for solar power production with no exogenous inputs," *Solar Energy*, vol. 86, no. 7, pp. 2017–2028, 2012.

[45] L. Martín, L. F. Zarzalejo, J. Polo, A. Navarro, R. Marchante, y M. Cony, "Prediction of global solar irradiance based on time series analysis: application to solar thermal power plants energy production planning," *Solar Energy*, vol. 84, no. 10, pp. 1772–1781, 2010.

[46] J. Jensenius y G. Cotton, "The development and testing of automated solar energy forecasts based on the model output statistics (mos) technique," en *1st Workshop on terrestrial solar resource forecasting and on use of satellites for terrestrial solar resource assessment, Washington, DC*, 1981.

[47] D. Baker y M. Casper, "Subjective forecasting of received solar radiation," en *Proc. First Workshop on Terrestrial Solar Resource Forecasting and on the Use of Satellites for Terrestrial Solar Resource Assessment*, pp. 8–11, 1981.

[48] R. Perez, K. Moore, S. Wilcox, D. Renné, y A. Zelenka, "Forecasting solar radiation–preliminary evaluation of an approach based upon the national forecast database," *Solar Energy*, vol. 81, no. 6, pp. 809–812, 2007.

[49] A. Hammer, D. Heinemann, E. Lorenz, y B. Lückehe, "Short-term forecasting of solar radiation: a statistical approach using satellite data," *Solar Energy*, vol. 67, no. 1, pp. 139–150, 1999.

[50] A. Hammer, D. Heinemann, C. Hoyer, y E. Lorenz, "Satellite based short-term forecasting of solar irradiance - comparison of methods and error analysis," en *The 2001 EUMETSAT meteorological satellite data users conference*, pp. 677–684, Citeseer, 2001.

[51] A. Hammer, D. Heinemann, C. Hoyer, R. Kuhlemann, E. Lorenz, R. Müller, y H. G. Beyer, "Solar energy assessment using remote sensing technologies," *Remote Sensing of Environment*, vol. 86, no. 3, pp. 423–432, 2003.

[52] R. Marquez, H. T. Pedro, y C. F. Coimbra, "Hybrid solar forecasting method uses satellite imaging and ground telemetry as inputs to anns," *Solar Energy*, vol. 92, pp. 176–188, 2013.

[53] C. W. Chow, B. Urquhart, M. Lave, A. Dominguez, J. Kleissl, J. Shields, y B. Washom, "Intra-hour forecasting with a total sky imager at the uc san diego solar energy testbed," *Solar Energy*, vol. 85, no. 11, pp. 2881–2893, 2011.

[54] R. Marquez, V. Gueorguiev, y C. Coimbra, "Forecasting solar irradiance using sky cover indices," *ASME J. Sol. Energy Eng*, 2012.

[55] R. Marquez y C. F. Coimbra, "Intra-hour dni forecasting based on cloud tracking image analysis," *Solar Energy*, vol. 91, pp. 327–336, 2013.

[56] J. Casa Nova, C. Boaventura, y P. de Moura Oliveira, "Solar irradiation forecast model using time series analysis and sky images," en *Proceedings of 5th Conference of the European Federation for Information Technology in Agriculture, Food and Environment (EFITA/WCCA 2005), Vila Real (Portugal)*, 2005.

[57] S. Safi, A. Zeroual, y M. Hassani, "Prediction of global daily solar radiation using higher order statistics," *Renewable energy*, vol. 27, no. 4, pp. 647–666, 2002.

[58] Y. Kemmoku, S. Orita, S. Nakagawa, y T. Sakakibara, "Daily insolation forecasting using a multi-stage neural network," *Solar Energy*, vol. 66, no. 3, pp. 193 – 199, 1999.

[59] B. B. Ekici, "A least squares support vector machine model for prediction of the next day solar insolation for effective use of pv systems," *Measurement*, vol. 50, no. 0, pp. 255 – 262, 2014.

[60] Z. Ramedani, M. Omid, A. Keyhani, S. Shamshirband, y B. Khoshnevisan, "Potential of radial basis function based support vector regression for global solar radiation prediction," *Renewable and Sustainable Energy Reviews*, vol. 39, no. 0, pp. 1005 – 1011, 2014.

[61] A. Azadeh, M. Sheikhalishahi, M. Tabesh, y A. Negahban, "The effects of preprocessing methods on forecasting improvement of artificial neural networks," *Australian Journal of Basic and Applied Sciences*, vol. 5, no. 6, pp. 570–580, 2011.

[62] C. Wu, K. Chau, y Y. Li, "Predicting monthly streamflow using data-driven models coupled with data-preprocessing techniques," *Water Resources Research*, vol. 45, no. 8, 2009.

[63] W. Wettayaprasit, N. Laosen, y S. Chevakidagarn, "Data filtering technique for neural networks forecasting," en *Proceedings of the 7th WSEAS International*

Conference on Simulation, Modelling and Optimization, pp. 225–230, World Scientific and Engineering Academy and Society (WSEAS), 2007.

[64] A. Savitzky y M. J. E. Golay, "Smoothing and differentiation of data by simplified least squares procedures," *Analytical chemistry*, vol. 36, no. 8, pp. 1627–1639, 1964.

[65] P. O. Persson y G. Strang, "Smoothing by Savitzky-Golay and Legendre filters," *Institute for Mathematics and Its Applications*, vol. 134, pp. 301–316, 2003.

[66] G. Sánchez, A. Serrano, y M. Cancillo, "Effect of cloudiness on solar global, solar diffuse and terrestrial downward radiation at badajoz (southwestern spain)," *Optica pura y aplicada*, vol. 45, no. 1, pp. 33–38, 2012.

[67] P. M. Bentler y A. Mooijaart, "Choice of structural model via parsimony: A rationale based on precision.," *Psychological bulletin*, vol. 106, no. 2, p. 315, 1989.

[68] V. Vapnik, *The nature of statistical learning theory*. springer, 1995.

[69] V. Vapnik, S. E. Golowich, y A. Smola, "Support vector method for function approximation, regression estimation, and signal processing," *Advances in neural information processing systems*, pp. 281–287, 1997.

[70] V. Vapnik, *Statistical learning theory*, vol. 2. Wiley New York, 1998.

[71] R. Fletcher, *Practical methods of optimization*. John Wiley & Sons, 1987.

[72] N. Cristianini y J. Shawe-Taylor, *An introduction to support vector machines and other kernel-based learning methods*. Cambridge university press, 2000.

[73] S. Abe, *Support vector machines for pattern classification*. Springer, 2010.

[74] A. Aizerman, E. M. Braverman, y L. Rozoner, "Theoretical foundations of the potential function method in pattern recognition learning," *Automation and remote control*, vol. 25, pp. 821–837, 1964.

[75] J. A. Suykens, T. Van Gestel, J. De Brabanter, B. De Moor, J. Vandewalle, J. Suykens, y T. Van Gestel, *Least squares support vector machines*, vol. 4. World Scientific, 2002.

[76] B. Schölkopf, A. J. Smola, R. C. Williamson, y P. L. Bartlett, "New support vector algorithms," *Neural computation*, vol. 12, no. 5, pp. 1207–1245, 2000.

[77] H. D. I. Abarbanel, T. W. Frison, y L. S. Tsimring, "Obtaining order in a world of chaos," *Signal Processing Magazine,IEEE*, vol. 15, no. 3, pp. 49–65, 1998.

[78] F. Takens, "Detecting strange attractors in turbulence," en *Dynamical Systems and Turbulence, Warwick 1980* (D. Rand y L.-S. Young, eds.), vol. 898 of *Lecture Notes in Mathematics*, pp. 366–381, Springer Berlin / Heidelberg, 1981.

[79] H. Kantz y T. Schreiber, *Nonlinear time series analysis*, vol. 7. Cambridge University Press, 2004.

[80] J. H. Holland, *Adaptation in natural and artificial systems: An introductory analysis with applications to biology, control, and artificial intelligence.* University of Michigan Press, 1975.

[81] D. E. Goldberg, *Genetic algorithms in search, optimization, and machine learning.* Addison Wesley, 1989.

[82] A. K. Jain y R. C. Dubes, *Algorithms for clustering data.* Prentice-Hall, Inc., 1988.

[83] A. K. Jain, M. N. Murty, y P. J. Flynn, "Data clustering: a review," *ACM computing surveys (CSUR)*, vol. 31, no. 3, pp. 264–323, 1999.

[84] L. J. Heyer, S. Kruglyak, y S. Yooseph, "Exploring expression data: identification and analysis of coexpressed genes," *Genome research*, vol. 9, no. 11, pp. 1106–1115, 1999.