EVALUACIÓN DE TÉCNICAS DE REGRESIÓN BASADAS EN EL APRENDIZAJE AUTOMÁTICO PARA EL RELLENO DE LAGUNAS ESPACIALES DE LA BASE DE DATOS DE HUMEDAD DEL SUELO DE CCI EN LA PENÍNSULA IBÉRICA

L. Almendra-Martín[1]*, J. Martínez-Fernández[1], Á. González-Zamora[1], P. Benito-Verdugo[1] y M. Piles[2]

[1]Instituto Hispano Luso de Investigaciones Agrarias (CIALE), Universidad de Salamanca. C/ Duero, 12. 37185. Villamayor. Salamanca. lauraalmendra@usal.es
[2]Laboratorio de Procesado de Imágenes. Parc Científic Universitat de València C/ Catedrático José Beltrán, 2. 46980. Paterna. València.

RESUMEN. La disponibilidad de series de humedad del suelo (SM) largas y continuas en espacio y tempo es fundamental para un gran número de aplicaciones. En este trabajo se analiza la idoneidad de diferentes técnicas de regresión basadas en el aprendizaje automático, aplicadas al dominio espacial, para rellenar las lagunas del producto de SM de la *Climate Change Initiative* (CCI) en la Península Ibérica entre 1991-2019. Los algoritmos evaluados han sido: *Support Vector Machine* (SVM), *Random Forest* (RF) y *Gaussian Processes* (GP). Las variables que se incorporaron como *inputs* en cada modelo fueron la textura del suelo, las coordenadas geográficas y la SM del día precedente. Además, se comparó la precisión de cada método con los parámetros recomendados por defecto por la implementación del software y una vez optimizados. Los tres métodos han demostrado ser suficientemente adecuados para rellenar las lagunas de CCI, siendo el GP el que proporcionaba mayor precisión.

ABSTRACT. The existence of long and continuous soil moisture (SM) series in space and time is fundamental for a variety of applications. This paper analyses the suitability of different machine learning regression techniques to fill the gaps in the SM product of the Climate Change Initiative (CCI) over the Iberian Peninsula in the period 1991-2019. The evaluated algorithms have been Support Vector Machine (SVM), Random Forest (RF) and Gaussian Processes (GP). The methods were applied in the spatial domain, and the variables incorporated as inputs in each model were the soil texture, the geographical coordinates and the SM of the preceding day. In addition, the precision of each method was compared using the default parameters of the software and once optimized. All three methods have proven to be adequate to fill the gaps in CCI, being the GP the most accurate.

1.- Introducción

La humedad del suelo superficial (SM) es una variable climática esencial que ejerce control sobre el intercambio de agua y energía entre el suelo y la atmósfera. Su monitorización es fundamental para un buen número de disciplinas como la agricultura, la meteorología, la gestión de recursos hídricos o los riesgos naturales (Almendra-Martín et al., 2021a; González-Zamora et al., 2021; Martínez-Fernández et al., 2021). Es por esto que, el interés por conseguir series de datos espacialmente continuas de esta variable ha crecido en las últimas décadas. En la actualidad, existe una variedad de bases de datos que proporcionan estimaciones de la humedad del suelo con diferentes orígenes, aproximaciones (observaciones, modelos o reanálisis) y características (Beck et al., 2020). Sin embargo, no hay tantas series de esta variable que sean de larga duración.

La *Climate Change Initiative* (CCI) es una iniciativa de la *European Space Agency* (ESA) que tiene como objetivo aprovechar todo el potencial de las series satelitales generadas durante los últimos 40 años para producir productos observacionales de un gran número de variables esenciales climáticas. Entre estas variables se encuentra SM (Dorigo et al., 2017; Gruber et al. 2017, 2019). Su serie se obtiene combinando todas las estimaciones obtenidas con satélites activos o pasivos de microondas (1-10 GHz) disponibles desde el año 1978, y constituye la base de datos de humedad del suelo satelital más larga disponible, y por tanto "a priori" ideal para estudios climáticos o de tendencias. Sin embargo, una de las principales limitaciones de este producto es la existencia de lagunas de datos provocadas por diferentes factores (contaminación por interferencia de radiofrecuencia, tiempo de revisita de los satélites, presencia de hielo, dificultad de recuperación de SM en zonas de costa y de montaña, etc.) (Cui et al., 2019).

Existen diferentes trabajos en la literatura que evalúan la aplicabilidad de técnicas de relleno de lagunas a bases de datos de SM (Xing et al., 2017; Xiao et al., 2016; Wang et al., 2012), pero son pocos los trabajos en los que se aplican a la base de datos de SM de la CCI. Por ejemplo, Cui et al. (2019) aplicaron un algoritmo basado en redes neuronales en la Meseta Tibetana para rellenar las lagunas de la base de datos de CCI aplicado al dominio temporal. Llamas et al. (2020) evaluaron técnicas de relleno de datos aplicadas al dominio espacial en una región de Oklahoma, dos basadas en *kriging* y una en modelos lineales. Estos autores obtuvieron buenas aproximaciones de los datos ausentes en las series de SM, sin embargo, su aplicabilidad se basó en los resultados obtenidos en regiones concretas con determinadas características climáticas. Almendra-Martín et al. (2021b) compararon la idoneidad de diferentes métodos aplicados tanto al dominio espacial como temporal en una región más amplia, el sur de Europa. Utilizaron desde metodologías simples y comúnmente usadas como la interpolación, hasta técnicas de regresión más complejas basadas en el

aprendizaje automático como las *Support Vector Machines* (SVM). Los resultados obtenidos demostraron que el algoritmo más complejo, SVM, aplicado al dominio espacial utilizando como *inputs* la textura del suelo, las coordenadas geográficas y la humedad del día anterior, ofrecía las estimaciones más precisas. Los resultados obtenidos por Almendra-Martín et al. (2021b) demostraron la aplicabilidad de técnicas de regresión basadas en aprendizaje automático para rellenar las lagunas de la base de datos de SM de la CCI. Sin embargo, el foco de ese trabajo fue en la elección de la combinación de variables de entrada que más precisión ofrecían en cada modelo, pero no se investigó en profundidad el ajuste (optimización) de los hiperparámetros y su impacto en la precisión obtenida.

En este trabajo se pretende evaluar la idoneidad de diferentes algoritmos de regresión basados en la teoría del aprendizaje automático aplicadas al dominio espacial como son las SVM, los *Random Forest* (RF) y los *Gaussian Processes* (GP) del software MATLAB en la Península Ibérica, utilizando las variables de entrada sugeridas por Almendra-Martín et al. (2021b) y evaluando el impacto del ajuste de los hiperparámetros.

2.- Bases de datos

La serie de SM de la CCI (en adelante CCISM) utilizada en este trabajo fue la versión 5.2 del producto combinado. Este producto fusiona los datos de todos los satélites de microondas, tanto activos como pasivos, que estiman esta variable desde noviembre de 1978 hasta 2020. El resultado es una imagen diaria a una resolución espacial de 0.25° de todo el globo terráqueo (Gruber et al., 2019). En este estudio, se extrajeron las series de los píxeles pertenecientes a la Península Ibérica y se filtraron los datos mediante el indicador de calidad del producto, manteniendo únicamente los píxeles con la calidad más alta. Los sensores de los satélites que componen el producto poseen diferentes características técnicas y cobertura temporal, lo que implica una distribución espaciotemporal de datos muy heterogénea (Gruber et al., 2017). En concreto, para la zona de estudio, en la primera década, con tan solo un satélite pasivo operativo, no se supera el 10% de datos disponibles (Fig. 1a). Mientras que en el año 2003 se observa un punto de inflexión alcanzando aproximadamente el 60% y aumentando progresivamente hasta casi 90% en los últimos años de la serie. Debido a la escasa disponibilidad de datos en los primeros años, este estudio se centra en el periodo 1991-2020, en el que se dispone de al menos un 25% de datos anuales. En este periodo el porcentaje medio de datos disponibles es de un 60%, siendo ligeramente menor en algunas zonas costeras, de montaña o bosque (Fig. 1b).

Los datos de textura de suelo se obtuvieron del producto SoilGrids del ISRIC-*World Soil Information*. Esta ofrece información sobre diferentes propiedades químicas y físicas del suelo en seis intervalos de profundidad desde los 5 cm hasta los 200 cm, y a una resolución espacial de 250 x 250 m (Batjes et al., 2020). En este estudio se utilizó la información del contenido de arena y de arcilla de la capa superficial, es decir, los primeros 5 cm, expresado en tanto por ciento. Los mapas de estas variables se remuestrearon a la malla de 0.25° del CCISM obteniendo un valor de cada variable para cada pixel.

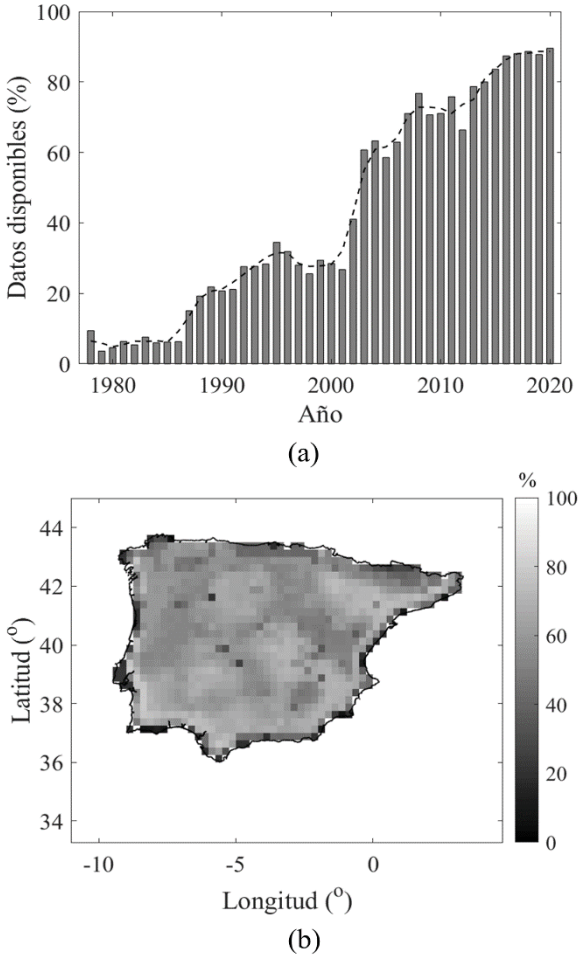

Fig. 1. Distribución espaciotemporal de los datos de CCISM. Porcentaje de la disponibilidad de datos anual en la Península Ibérica (a) y porcentaje de la disponibilidad de datos de cada pixel desde 1991 hasta 2020 (b).

3.- Metodología

En el estudio se evaluó la aplicabilidad de tres métodos de regresión basados en aprendizaje automático, aplicados al dominio espacial, para rellenar las lagunas de la base de datos de CCISM. Los tres métodos son: SVM, RF y GP, todos basados en principios estadísticos y matemáticos diferentes.

Para poder estimar los valores ausentes de SM, se escogieron como variables de entrada a los modelos, aquellas capaces de describir las propiedades del suelo y situar en el espacio los valores de SM. Esta información en conjunto ha demostrado mejorar la precisión de las estimaciones (Almendra-Martín et al., 2021b). Por tanto, las variables utilizadas como *inputs* de los modelos fueron

el contenido de arena y arcilla, las coordenadas geográficas y la humedad del día anterior.

Su aplicación al dominio espacial supone que, para cada día de la serie, se estima un modelo entrenado con los datos disponibles y se calculan los valores ausentes de SM. La evaluación de esos valores estimados con cada método se llevó a cabo mediante una validación cruzada *hold-out* con 9 repeticiones (Pérez-Planells et al., 2015). Es decir, los datos disponibles de un día se dividieron en dos conjuntos, el de entrenamiento con el 70% de los ellos y el de test con el 30%. Esta división se realizó de forma aleatoria, pero tratando de reproducir la distribución espacial de las lagunas de CCISM en el área de estudio. Además, en cada repetición la división de los datos fue la misma en todos los métodos para asegurar la consistencia en las intercomparaciones.

Los parámetros estadísticos calculados en la validación entre el conjunto de test y las estimaciones obtenidas fueron el coeficiente de correlación de Pearson (R), el sesgo o bias y el error cuadrático medio (RMSE), usados comúnmente en las validaciones de productos satelitales de SM (Entekhabi et al., 2010). También se obtuvo el tiempo de computación medio de cada uno de los métodos.

3.1. Support Vector Machines

Las SVM, desarrolladas inicialmente para problemas de clasificación, se basan en la teoría del aprendizaje estadístico (Vapnik, 1995). El algoritmo se define a partir de la idea de separar un conjunto de datos en dos clases mediante un hiperplano de separación óptimo que maximice el margen entre patrones (Dibike et al., 2001). Su aplicación a problemas de regresión consiste en encontrar, en lugar de un hiperplano, una función de regresión óptima, lo más plana posible y que minimice el riesgo empírico fijando un máximo de desviación ε del valor real (Smola y Schölkopf et al., 2004). Para tratar el problema de forma no lineal, se incorporan las funciones *kernel*, mediante las cuales, se mapean los datos a un espacio de mayor dimensión donde el problema puede tratarse de forma lineal. El algoritmo SVM aplicado a problemas de regresión se ha utilizado en el campo de la teledetección para diversas aplicaciones debido a su capacidad de generalizar, incluso con conjuntos de datos de entrenamiento pequeños (Mountrakis et al., 2011). La precisión de los modelos depende principalmente de la desviación máxima permitida ε, una constante de regularización C que compensa la complejidad de la función y su capacidad de generalizar y la función *kernel*. En este trabajo se utilizó una función *kernel* gaussiana con un factor de escala K del que también depende la precisión.

3.2. Random Forest

Los RF consisten en un conjunto de árboles de decisión independientes, entrenados con una muestra aleatoria distinta, pero con la misma distribución en todos los árboles (Breiman, 2001). Un árbol de decisión consiste en un modelo de decisiones binario en forma de árbol capaz de repartir las observaciones en función de sus atributos. La estimación final del RF se obtiene agregando las predicciones de todos los árboles individuales que forman el modelo (Amat-Rodrigo, 2017). Este algoritmo se ha aplicado en diferentes trabajos del campo de la teledetección, tanto para problemas de clasificación como de regresión (Mutanga et al., 2012). Su precisión depende de diferentes hiperparámetros que hay que considerar a la hora de definir los modelos como el número de árboles que lo componen y de las características de estos. En el diseño de los árboles de decisión se puede definir el número de nodos o número mínimo de observaciones por nodo. El *software* utilizado en este trabajo, generalmente, construye por defecto árboles profundos y el RF utiliza un total de 100 árboles.

3.3. Gaussian Processes

Los GP fueron implementados por primera vez para problemas de regresión por Williams y Rasmussen (1996). Se definen como una distribución de probabilidad gaussiana sobre funciones aleatorias y su ajuste se basa en el teorema de Bayes. Un GP es un proceso estocástico que puede definirse por su función media y su función de covarianza (Schulz et al., 2018). La función de covarianza expresa la similitud entre los valores predictores y los valores respuesta y se incorpora al modelo mediante una función *kernel*. La precisión en las estimaciones depende principalmente de la definición de esta función y de sus parámetros. Una característica atractiva de los GP es que permiten obtener no sólo las predicciones, sino también su intervalo de confianza. En este trabajo se utiliza una función *kernel* exponencial cuadrática, una de las funciones de covarianza más utilizadas. Estos parámetros se basan en la desviación estándar de la señal σ_f y la escala de longitud σ_l.

4.- Resultados y discusión

4.1. Hiperparámetros por defecto

Primero se evaluó la aplicabilidad de los tres métodos de regresión con los ajustes de parámetros que el software MATLAB utiliza por defecto. Los resultados obtenidos en la validación cruzada para cada modelo diario muestran, en general, una buena precisión (Fig. 2). Al analizar el valor de R, se observa un aumento de correlación en los últimos años de la serie, donde el porcentaje de datos disponibles es mayor y los valores de la serie de CCISM presentan menor incertidumbre (Almendra-Martín et al., 2021b). Los valores medios para todo el periodo son ligeramente mayores cuando se aplican los GP (R = 0,83) y más bajos con las SVM (R = 0,77). Sin embargo, las SVM presentan valores más estables con un menor número de *outliers*. Estos pueden estar relacionados con modelos sobreajustados de días donde los datos disponibles son escasos o no son representativos de la dinámica espacial de la SM. En las SVM no existe este problema por la buena capacidad del método para

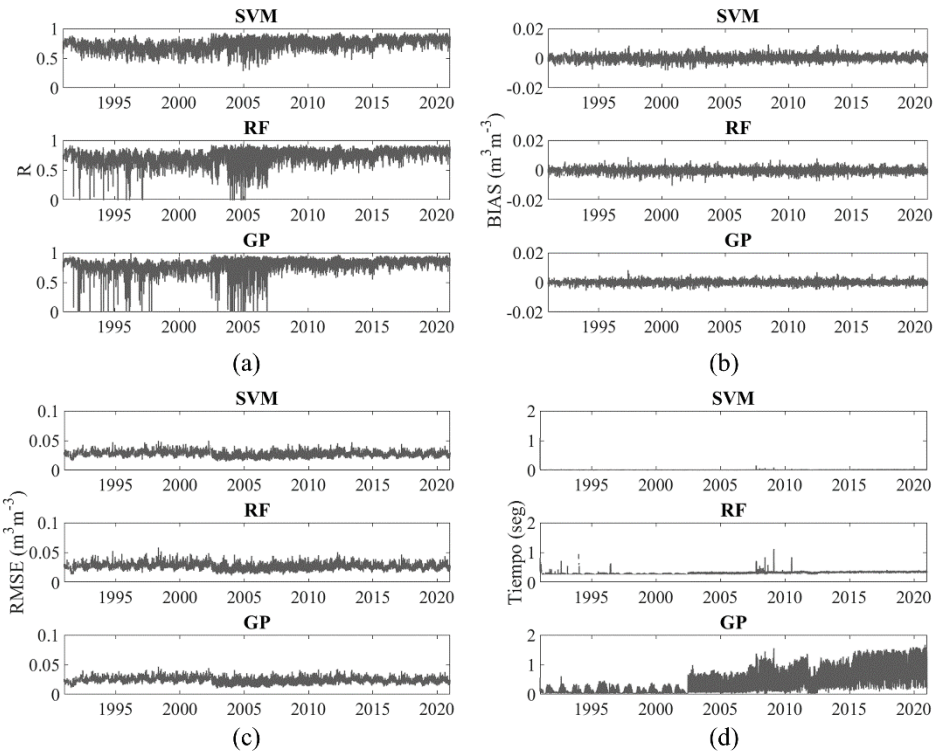

Fig. 2. Parámetros obtenidos en la validación cruzada *hold-out* realizada en las estimaciones de SM obtenidas con cada uno de los métodos de regresión para cada modelo diario. Parámetros de validación: coeficiente de correlación de Pearson (a), sesgo (b), RMSE (c) y tiempo medio de computación (d).

generalizar (Mountrakis et al., 2011). Los valores del sesgo y del RMSE obtenidos en todos los casos se aproximan al cero, lo que implica una buena precisión de los modelos, siendo esta ligeramente mayor con los GP (fig. 2b-c). Además, en el caso del RMSE, se observa una reducción notable en 2002, coincidente con la ruptura más marcada de la serie de CCISM provocada por la fusión de los diferentes sensores (Preimesberger et al. 2021) y con el aumento de datos disponibles en la serie (Fig. 1a). Sin embargo, aunque los GP muestran una precisión mayor, el tiempo de computación que necesitan para entrenar su modelo es muy superior al de otros métodos (Fig. 2d). También se observa un cambio a partir del 2002, en este caso ligado al aumento del número de datos.

4.2. Optimización de hiperparámetros

La precisión de los métodos de regresión aplicados en este trabajo depende de los valores asignados a diferentes hiperparámetros que definen el algoritmo. El software utilizado establece valores predeterminados para algunos de estos parámetros, sin embargo, en ocasiones su definición depende de los datos de entrada del modelo. Esto puede derivar en sobreajustes, si, por ejemplo, se tienen pocos datos y muy ruidosos al entrenar un modelo. Lo que se pretende es asignar valores únicos para todos los modelos diarios, pero que consigan la mayor precisión posible en las estimaciones de SM. Para optimizar los parámetros de cada modelo se escogió una muestra de forma aleatoria, pero representativa de los resultados obtenidos en la sección anterior, en lugar de calcular un modelo para cada día del periodo. Esta muestra consideró un 10% del período de estudio para reducir el tiempo de computación. Para estimar la precisión de esta muestra se realizó una validación cruzada 9-*fold* y se calculó el RMSE.

En el caso de las SVM, la precisión del modelo depende principalmente de la elección de tres hiperparámetros: la desviación máxima permitida ε, el factor de escala de la función kernel K, y la constante de regularización C. El factor de escala K es una constante independiente de los datos de entrada, el valor predeterminado es 1, pero se han estudiado los resultados para un rango de valores entre 0,1 y 4. La constante de regularización y la desviación máxima permitida se establece por defecto a partir de la distribución de la variable respuesta. En este caso, se ha escogido el rango de valores predeterminados más frecuentes en el periodo de estudio. Los resultados obtenidos con cada parámetro no han mostrado grandes cambios en los errores de las estimaciones de SM (Fig. 3). Se obtienen resultados menos precisos para valores de K menores a la unidad, pero para valores mayores, el error se estabiliza. La constante C también muestra un error ligeramente mayor para los valores más bajos del rango escogido y se estabiliza a medida que éste aumenta. Por el contrario, la desviación máxima en el rango escogido no parece influir en el error de la estimación.

Para los RF se estudió la influencia de tres hiperparámetros: el número de observaciones mínimo por nodo, estableciendo el rango entre 1 y 20, el mínimo de observaciones por nodo terminal, en un rango entre 1 y 10, y el número de árboles, desde 50 hasta 950. Los errores de las estimaciones obtenidos al variar el número

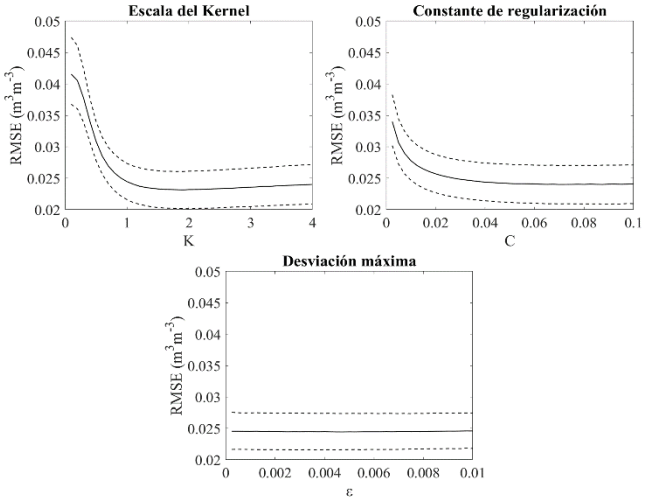

Fig. 3. Error obtenido en las estimaciones de SM con las SVM para cada valor de los hiperparámetros. Las líneas discontinuas indican los percentiles 25 y 75 y la línea continua la mediana de la muestra.

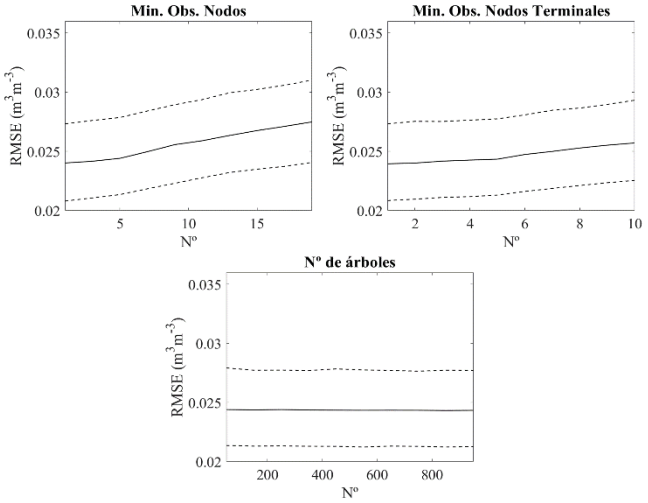

Fig. 4. Error obtenido en las estimaciones de SM con los RF para cada valor de los hiperparámetros. Las líneas discontinuas indican los percentiles 25 y 75 y la línea continua la mediana de la muestra.

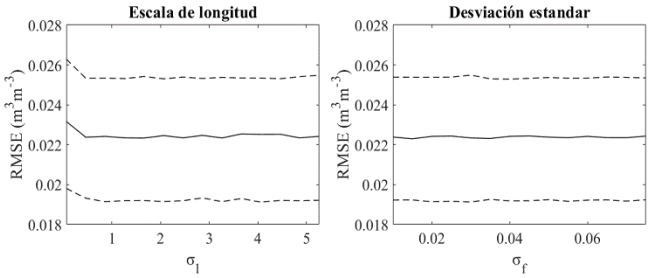

Fig. 5. Error obtenido en las estimaciones de SM con los GP para cada valor de los hiperparámetros. Las líneas discontinuas indican los percentiles 25 y 75 y la línea continua la mediana de la muestra.

de árboles no varían de forma significativa. Por el contrario, se observa una precisión menor en el cálculo de SM al incrementar el número de observaciones mínimas en los nodos (Fig. 4).

Por último, se optimizaron los hiperparámetros de los GP. Estos fueron la desviación estándar de la señal σ_f y la escala de longitud σ_l de la función *kernel*. Los valores predeterminados de ambos parámetros dependen de las variables de entrada, por lo que el rango analizado se escogió a partir de los valores por defecto más frecuentes en el periodo de estudio, al igual que en las SVM. Así, el rango seleccionado de la desviación estándar de la señal fue de 0,01 a 0,075 y el de la escala de longitud de 0,06 a 5,6. Los errores obtenidos con estos valores no mostraron mucha variación en el rango seleccionado. Únicamente se observa un ligero incremento del RMSE con el valor más bajo de σ_l (Fig. 5).

Del análisis realizado con cada método de regresión se han obtenido los valores de los hiperparámetros que optimizan el error de las estimaciones de la serie de CCISM. En los casos en los que se observó una disminución del error, se seleccionó el valor del hiperparámetro con el que se obtenía el mínimo de RMSE. En aquellos hiperparámetros en los que no se observó una mejora significativa, pero el valor predeterminado depende de los datos de entrada, se seleccionó el valor por defecto más común en todos los modelos diarios. Los valores finales seleccionados se pueden ver en la Tabla 1.

Tabla 1. Valores de los hiperparámetros escogidos para optimizar los modelos con cada método de regresión.

Método	Hiperparámetro	Valor
SVM	K	1,5
	C	0,0375
	ε	0,0038
RF	Min. observaciones por nodo	1
	Min. observaciones por nodo terminal	2
	Nº arboles	100
GP	σ_f	0,0275
	σ_l	4,3

4.3. Hiperparámetros optimizados

Los tres algoritmos de regresión evaluados para el relleno de lagunas de la base de datos de CCISM fueron entrenados de nuevo, pero esta vez utilizando los valores de los hiperparámetros de la Tabla 1. Los resultados obtenidos en la validación cruzada *hold-out* se muestran en la figura 6. En general los resultados obtenidos son más precisos, el valor medio de R es mayor para las SVM (R = 0,80) y los GP (R = 0,92), mientras que en los RF se mantiene (R = 0,78). Sin embargo, el coeficiente de correlación en los GP y los RF ya no presenta *outliers* (Fig. 6a). En el caso del sesgo y el RMSE, únicamente se observa una mejora notable en los GP (Fig. 6b-c). Si se observa el tiempo de computación, aunque la magnitud expresada en este trabajo está condicionada por el *hardware* usado, se ve que los métodos más rápidos fueron las SVM y los RF (Fig. 6d). Los GP, aunque ofrecen resultados notablemente mejores, requieren un coste computacional mayor, que además se incrementa ligeramente con los valores optimizados de los hiperparámetros.

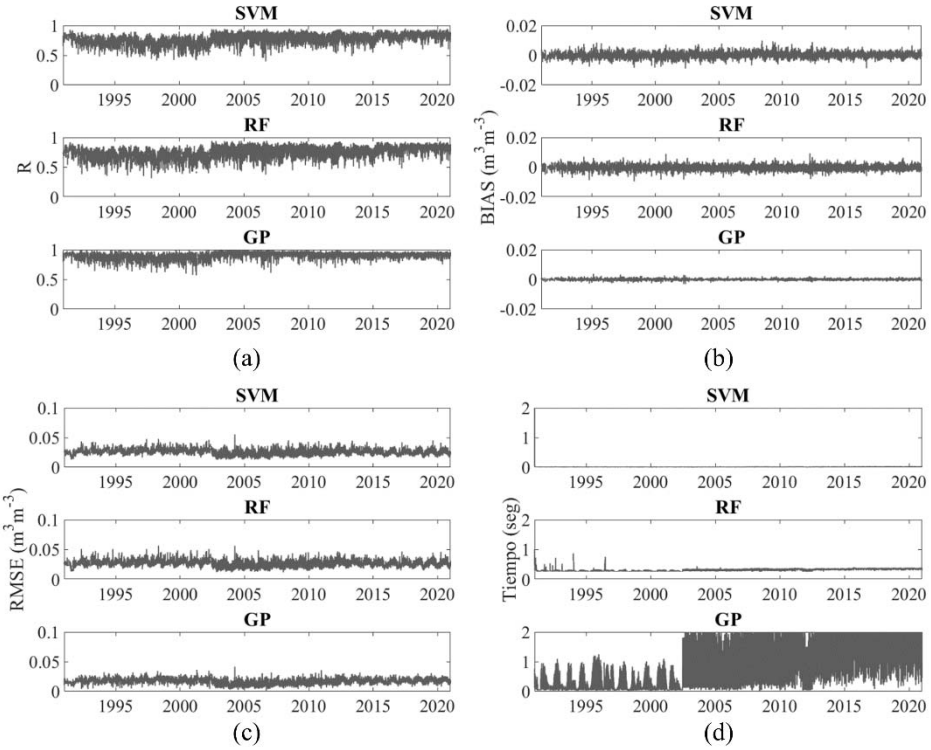

Fig. 6. Parámetros obtenidos en la validación cruzada *hold-out* realizada en las estimaciones de SM obtenidas con cada uno de los métodos de regresión optimizado para cada modelo diario. Parámetros de validación: coeficiente de correlación de Pearson (a), sesgo (b), RMSE (c) y tiempo medio de computación (d).

Estos resultados revelan la adecuación de los métodos de regresión basados en el aprendizaje automático para el relleno de lagunas de CCISM. Además, su implementación en el dominio espacial, ha mostrado ser igual o más precisa que la implementación en el dominio temporal, aunque su uso sea menos frecuente (Almendra-Martín et al. 2021b). Los mejores resultados se han obtenido con los GP antes y después de la optimización. Las SVM, se han optimizado hasta presentar estimaciones más precisas que los RF.

Los mejores resultados obtenidos, los de los GP, son comparables con los obtenidos por Llamas et al., 2020. En su trabajo obtuvieron una correlación de 0,88 aplicando *kriging* de regresión, también un proceso gaussiano (Shadrin et al., 2021). Sin embargo, la forma de aplicar los GP propuesta en este trabajo permite proporcionar al modelo más información, como la textura de suelo y la SM del día antes, variable que ha demostrado ser clave para obtener estimaciones de SM adecuadas en métodos de regresión (Almendra-Martín et al., 2021b).

5.- Conclusiones

Monitorizar la humedad del suelo es una tarea clave para la gestión de los recursos hídricos y para el estudio del ciclo hidrológico. La base de datos de CCI es una herramienta muy útil para esta tarea ya que ofrece la serie satelital de SM más larga disponible. Sin embargo, las lagunas de datos plantean serias limitaciones para determinadas aplicaciones. En este trabajo se han propuesto tres métodos de regresión basados en el aprendizaje automático para abordar este problema en la Península Ibérica. Las SVM, los RF y los GP se han aplicado al dominio espacial utilizando como variables de entrada a los modelos la textura del suelo, las coordenadas geográficas y la humedad del día anterior. Las estimaciones una vez evaluadas mediante una validación cruzada *hold-out* con 9 repeticiones mostraron que los GP eran capaces de estimar los valores de CCISM de una forma más precisa. Sin embargo, las SVM a pesar de presentar en la validación los valores medios de correlación más bajos, fue el método que menos *outliers* presentó. Probablemente se deba a la capacidad de generalizar, con pocos datos de entrenamiento, que tiene el modelo.

El *software* utilizado, MATLAB, utiliza unos valores predeterminados para los hiperparámetros de los modelos que no siempre son los óptimos y cuando estos valores dependen de las características de los datos de entrada, esto puede provocar un sobreajuste en el modelo. La optimización de los hiperparámetros se realizó en una muestra aleatoria pero representativa en los tres métodos estudiados. Los errores obtenidos con la variación de parámetros no mostraron cambios muy significativos. Sin embargo, sí se pudo determinar un conjunto de valores con los que se mejoraron los algoritmos.

En la evaluación de los tres métodos una vez optimizados, no se obtuvieron valores anómalos en el

coeficiente de correlación y se mejoraron las estimaciones en todos los casos, pero de manera más pronunciada en los GP. Además, se volvió a observar como este método estimaba los valores de la serie de CCISM con mayor precisión, aunque el coste computacional es superior.

Los GP aplicados al dominio espacial utilizando únicamente la información espacial, como el *kriging*, han demostrado ser técnicas fiables para la estimación de series espacialmente continuas en otros trabajos. Sin embargo, la forma de aplicar los GP propuesta en este trabajo permite proporcionar al modelo más información espacial y de esta forma captar y recrear mejor la dinámica de la SM.

Agradecimientos. Este estudio fue financiado por el Ministerio de Ciencia, Innovación y Universidades de España (Proyectos ESP2017-89463-C3-3-R y RTI2018-096765-A-100), la Junta de Castilla y León (Proyecto SA112P20), el Fondo Europeo de Desarrollo Regional (FEDER) y el proyecto Unidad de Excelencia CLU-2018-04, cofinanciado por FEDER y la Junta de Castilla y León.

6.- Bibliografía

Almendra-Martín, L., J. Martínez-fernández, A. González-Zamora, P. Benito-Verdugo y C.M. Herrero-Jiménez, 2021a. Agricultural drought trends on the Iberian Peninsula: an analysis using modeled and reanalysis soil moisture products. Atmosphere. *12*. 236.

Almendra-Martín, L., J. Martínez-Fernández, M. Piles y Á. González-Zamora, 2021b. Comparison of gap-filling techniques applied to the CCI soil moisture database in Southern Europe. Remote Sens. Environ. *258*, 112377.

Amat-Rodrigo J., 2017. Árboles de decisión, random forest, gradient boosting y C5.0. *https://www.cienciadedatos.net/* [último acceso junio, 2021].

Batjes, N. H., E. Ribeiro y A. V. Oostrum, 2020. Standardised soil profile data to support global mapping and modelling (WoSIS snapshot 2019). *Earth Syst. Sci. Data*, *12*, 299-320.

Beck, H. E., M. Pan, D. G. Miralles, R. H. Reichle, W. A. Dorigo, S. Hahn, J. Sheffield, L. Karthikeyan, G. Balsamo, R. M. Parinussa, A.I.J.M. van Dijk, J. Du, J. S. Kimball, N. Vergopolan y E. F. Wood, 2021. Evaluation of 18 satellite-and model-based soil moisture products using in situ measurements from 826 sensors. *Hydrol Earth Syst Sci*, *25*, 17-40.

Breiman, L., 2001. Random forests. *Mach. Learn*. *45*, 5–32.

Cui, Y., C. Zeng, J. Zhou, H. Xie, W. Wan, L. Hu, W. Xiong, X. Chen, W. Fan y Y. Hong, 2019. A spatio-temporal continuous soil moisture dataset over the Tibet plateau from 2002 to 2015. *Sci. data 6*, 247.

Dibike, Y.B., S. Velickov, D. Solomatine, M.B. Abbott, 2001. Model Induction with Support Vector Machines: Introduction and Applications. *J. Comput. Civ. Eng*. *15*, 3, 208–216.

Dorigo, W., W. Wagner, C. Albergel, F. Albrecht, G. Balsamo, L. Brocca, D. Chung, M. Ertl, M. Forkel, A. Gruber, E. Haas, P.D. Hamer, M. Hirschi, J. Ikonen, R. de Jeu, R. Kidd, W. Lahoz, Y.Y. Liu, D. Miralles, T. Mistelbauer, N. Nicolai-Shaw, R. Parinussa, C. Pratola, C. Reimer, R. van der Schalie, S.I. Seneviratne, T. Smolander y P. Lecomte, 2017. ESA CCI Soil Moisture for improved Earth system understanding: State-of-the art and future directions. *Remote Sens. Environ*. *203*, 185–215.

Entekhabi, D., R.H. Reichle, R.D. Koster, W.T. Crow, 2010. Performance metrics for soil moisture retrievals and application requirements. *J. Hydrometeorol*. *11*, 832–840.

González-Zamora, A., L. Almendra-Martín, M. De Luis, y J. Martínez-fernández, 2021. Influence of soil moisture versus climatic factors in Pinus halepensis growth variability in Spain: a study with remote sensing and modeled data. Remote Sens. *13*,757.

Gruber, A., W.A. Dorigo, W. Crow, W. Wagner, 2017. Triple Collocation-Based Merging of Satellite Soil Moisture Retrievals. *IEEE Trans. Geosci. Remote Sens. 55*, 6780–6792.

Gruber, A., T. Scanlon, R. van der Schalie, W. Wagner, W. Dorigo, 2019. Evolution of the ESA CCI Soil Moisture climate data records and their underlying merging methodology. *Earth Syst. Sci. Data 11*, 717–739.

Llamas, R.M., M. Guevara, D. Rorabaugh, M. Taufer y R. Vargas, 2020. Spatial gapfilling of ESA CCI satellite-derived soil moisture based on geostatistical techniques and multiple regression. *Remote Sens*. *12*, 665.

Martínez-Fernández, J., A. González-Zamora y L. Almendra-Martín, 2021. Soil moisture memory and soil properties: an analysis with the stored precipitation fraction. J. Hydrol. *593*, 125622.

Mountrakis, G., J. Im y C. Ogole, 2011. Support vector machines in remote sensing: A review. *ISPRS J. Photogramm. Remote Sens*. *66*, 247–259.

Mutanga, O., E. Adam y M. A. Cho, 2012. High density biomass estimation for wetland vegetation using WorldView-2 imagery and random forest regression algorithm. *Int. J. Appl. Earth Obs. Geoinf*. *18*, 399-406.

Pérez-Planells, L., J. Delegido, J. P. Rivera-Caicedo y J. Verrelst, 2015. Analysis of cross-validation methods for robust retrieval of biophysical parameters. *Rev. Teledetección*, *44*, 55-65.

Preimesberger, W., T. Scanlon, C.H. Su, A. Gruber y W. Dorigo, 2020. Homogenization of Structural Breaks in the Global ESA CCI Soil Moisture Multisatellite Climate Data Record. *IEEE Trans Geosci Remote Sens. 59*, 2845-2862.

Schulz, E., M. Speekenbrink, y A. Krause, 2018. A tutorial on Gaussian process regression: Modelling, exploring, and exploiting functions. *J. Math. Psychol*. *85*, 1-16.

Shadrin, D., A. Nikitin, P. Tregubova, V. Terekhova, R. Jana, S. Matveev y M. Pukalchik, 2021. An Automated Approach to Groundwater Quality Monitoring—Geospatial Mapping Based on Combined Application of Gaussian Process Regression and Bayesian Information Criterion. *Water*, *13*, 400.

Smola, A.J. y B. Schölkopf, 2004. A tutorial on support vector regression. *Stat. Comput*. *14*, 199–222.

Vapnik, V.N., 1995. The Nature of Statistical Learning Theory. Springer New York, New York, NY.

Wang, G., D. Garcia, Y. Liu, R. de Jeu y A.J. Dolman, 2012. A three-dimensional gap filling method for large geophysical datasets: application to global satellite soil moisture observations. *Environ. Model. Softw*. *30*, 139–142.

Williams, C.K., y C.E. Rasmussen, 1996. Gaussian processes for regression. *MIT Press*.

Xiao, Z., L. Jiang, Z. Zhu, J. Wang y J. Du, 2016. Spatially and temporally complete satellite soil moisture data based on a data assimilation method. *Remote Sens*. *8*.

Xing, C., N. Chen, X. Zhang y J. Gong, 2017. A machine learning based reconstruction method for satellite remote sensing of soil moisture images with in situ observations. *Remote Sens*. *9*.