

CASO PRÁCTICO

Análisis de métodos de validación cruzada para la obtención robusta de parámetros biofísicos

Pérez-Planells, Ll.*¹, Delegido, J.², Rivera-Caicedo, J.P.², Verrelst, J.²

¹ Dpto. Física de la Tierra y Termodinámica, Facultad de Física, Universitat de València, 46100 València, Spain.

² Image Processing Laboratory (IPL), Parc Científic, Universitat de València, 46980 Paterna, Spain.

Resumen: Los métodos de regresión no paramétricos son una gran herramienta estadística para obtener parámetros biofísicos a partir de medidas realizadas mediante teledetección. Pero los resultados obtenidos se pueden ver afectados por los datos utilizados en la fase de entrenamiento del modelo. Para asegurarse de que los modelos son robustos, se hace uso de varias técnicas de validación cruzada. Estas técnicas permiten evaluar el modelo con subconjuntos de la base de datos de campo. Aquí, se evalúan dos tipos de validación cruzada en el desarrollo de modelos de regresión no paramétricos: *hold-out* y *k-fold*. Los métodos de regresión lineal seleccionados fueron: *Linear Regression* (LR) y *Partial Least Squares Regression* (PLSR). Y los métodos no lineales: *Kernel Ridge Regression* (KRR) y *Gaussian Process Regression* (GPR). Los resultados de la validación cruzada mostraron que LR ofrece los resultados más inestables, mientras KRR y GPR llevan a resultados más robustos. Este trabajo recomienda utilizar algoritmos de regresión no lineales (como KRR o GPR) combinando con la validación cruzada *k-fold* con un valor de *k* igual a 10 para hacer la estimación de una manera robusta.

Palabras clave: *Hold-Out*, *k-fold*, validación cruzada, aprendizaje automático, procesos gaussianos, regresión de Kernel Ridge.

Analysis of cross-validation methods for robust retrieval of biophysical parameters

Abstract: Non-parametric regression methods are powerful statistical methods to retrieve biophysical parameters from remote sensing measurements. However, their performance can be affected by what has been presented during the training phase. To ensure robust retrievals, various cross-validation sub-sampling methods are often used, which allow to evaluate the model with subsets of the field dataset. Here, two types of cross-validation techniques were analyzed in the development of non-parametric regression models: *hold-out* and *k-fold*. Selected non-parametric linear regression methods were least squares Linear Regression (LR) and Partial Least Squares Regression (PLSR), and nonlinear methods were Kernel Ridge Regression (KRR) and Gaussian Process Regression (GPR). Cross-validation results showed that LR performed most unstable, while KRR and GPR led to more robust results. This work recommends using a nonlinear regression algorithm (e.g., KRR, GPR) in combination with a *k-fold* cross-validation technique with *k*=10 to realize robust retrievals.

Key words: *Hold-Out*, *k-fold*, cross-validation, MLRA, Gaussian Process Regression, Kernel Ridge Regression.

* Autor para la correspondencia: Lluis.Perez@uv.es

1. Introducción

La obtención de parámetros biofísicos de la vegetación a partir de las observaciones de la Tierra desde satélite es un tema de gran interés debido a la importancia de estos parámetros, como son el índice de área foliar (LAI, por sus siglas en inglés) o la cantidad de clorofila en la hoja (Chl), ya que dan información directa sobre el estado fenológico y de salud tanto de bosques como de zonas de cultivo (Verrelst *et al.*, 2014a).

Para obtener los distintos parámetros biofísicos a partir de imágenes de satélite es necesario aplicar un modelo físico o estadístico que relacione las medidas de radiancia o reflectividad obtenidas con el parámetro de estudio (Lázaro-Gredilla *et al.*, 2014; Verrelst *et al.*, 2015). En el presente estudio vamos a centrarnos en métodos estadísticos no paramétricos, los cuales relacionan las medidas de satélite con el valor de la variable de interés. De esta manera se crean modelos que se ajustan para obtener la mejor estimación a partir de una serie de pares de datos de entrenamiento que contienen información para cada píxel de la medida de satélite y del valor del parámetro biofísico que se analiza (Verrelst *et al.*, 2012).

En función del modelo de regresión utilizado distinguimos dos tipos de métodos no paramétricos: lineales y no lineales. Dentro del grupo de métodos no lineales se encuentran los conocidos ‘*Machine Learning Regression Algorithms*’ (MLRAs). Estos constan de distintos algoritmos que permiten la obtención de modelos no paramétricos con bastante rapidez, hecho que, junto a la robustez de los modelos que se generan, hace que estos métodos sean de gran interés para su implementación en teledetección (Bazi *et al.*, 2014).

Con la finalidad de poder analizar los distintos métodos, se realiza una posterior validación del modelo obtenido (Guerschman *et al.*, 2015). Para ello, existen diversas técnicas, entre las cuales destacamos los métodos de validación cruzada. La validación cruzada hace uso de distintos subconjuntos de los datos disponibles para realizar el entrenamiento del modelo y su posterior validación (Arlot y Celisse, 2010). En este estudio, haremos uso de los métodos *hold-out* y *k-fold* para validar el modelo.

Los distintos métodos de regresión y de validación cruzada comentados pueden implementarse directamente desde ARTMO (*Automated Radiative Transfer Model Operator*), una toolbox con interfaz gráfica generada en MATLAB® (Rivera *et al.*, 2014a). ARTMO tiene incorporados distintos métodos de regresión no paramétricos y, además, permite aplicar a cada uno de los métodos las distintas maneras de validación cruzada comentadas, permitiendo comparar la validación de los distintos métodos utilizados.

El objetivo principal de este trabajo consiste en determinar el método de validación cruzada que ofrece un mejor resultado para distintos métodos de regresión no paramétricos. Para ello se aplicarán, para cuatro algoritmos no paramétricos distintos, el método de validación cruzada *k-fold* aplicando distintos valores de *k*, y el método *hold-out* aplicando distintos porcentajes para los datos de entrenamiento y de validación.

2. Metodología

2.1. Datos utilizados

Los datos utilizados para realizar el trabajo pertenecen a la campaña SPARC-2003, que tuvo lugar durante el 12-14 de julio en Barrax, La Mancha, España (coordenadas 30°3’N, 28°6’W, 700 m de altitud). En esta campaña se hicieron medidas de diferentes parámetros biofísicos en distintos cultivos, entre ellas 110 muestras medidas con el Licor LAI (Delegido *et al.*, 2015) que usaremos en este estudio. Estas son comparadas con una imagen obtenida a partir de las medidas espectrales realizadas con el sensor hiperespectral aerotransportado HyMap (*Hyperspectral Mapper*), el cual está configurado con 125 bandas entre 450 nm y 2490 nm con un tamaño de píxel de 5 m (Delegido *et al.*, 2015). Esta imagen ha sido posteriormente corregida atmosférica y geométrica (Rivera *et al.*, 2014a).

2.2. Métodos de regresión no paramétricos

Entre los mejores métodos de regresión no paramétricos existentes destacan cuatro: dos lineales, el *Lineal Regression* (LR) y *Partial Least Squares Regression* (PLSR) y dos no lineales que son el

Kernel Ridge Regression (KRR) y el *Gaussian Process Regression* (GPR).

El método LR es el más simple de los métodos lineales existentes, basado en la obtención de relaciones lineales entre los pares de datos mediante ajustes de mínimos cuadrados (Shao, 1993). Consideramos X e y los datos de entrada y salida siendo matrices $n \times d$ y $n \times l$ respectivamente, donde d es la dimensión de los datos, en nuestro caso las longitudes de onda, n es el número de datos y l el número de parámetros biofísicos a estimar. Con estas matrices se pretende ajustar a un modelo lineal para predecir las variables de salida a partir de las características de los datos de entrada:

$$y = Xw \quad (1)$$

Donde w contiene los coeficientes de los modelos de regresión, con un tamaño $l \times d$.

El método PLSR generaliza y combina características del análisis de componentes principales y regresiones múltiples, de manera que realiza combinaciones lineales pero a partir de las componentes principales de los datos (Rivera *et al.*, 2014a).

En cuanto a los métodos no lineales utilizados, el método KRR minimiza los residuos del ajuste de mínimos cuadrados ampliando las dimensiones de las características del espacio. Se considera una aproximación a la versión *kernel* del método LR. En este caso los pesos se expresan como una combinación lineal de muestras localizadas:

$$w = \sum_{i=1}^n \alpha_i \varphi(x_i) \quad (2)$$

La predicción para una muestra de test x_* se obtiene como una función a partir de los pesos $\alpha = [\alpha_1, \dots, \alpha_n]^T$ (una por muestra), de la siguiente manera:

$$\mathbb{E}[f(x_*)] = k_{*,:}^T (K + \lambda I)^{-1} y = k_{*,:}^T \alpha \quad (3)$$

donde $k_{*,:}^T$ contiene las similitudes entre los datos test y los puntos de entrenamiento. Haciendo uso de la siguiente función matricial, con componentes $[K]_{ij}$:

$$K(x_i, x_j) = \exp\left(-\|x_i - x_j\|^2 / (2\sigma^2)\right) \quad (4)$$

siendo σ un parámetro a ajustar de la función.

En cuanto al método GPR, genera el modelo mediante un proceso similar al KRR, pero teniendo también en cuenta la varianza de los datos y la maximización del margen de probabilidad en los datos de entrenamiento. En este caso se hace uso de una función *kernel Gaussian anisótropa escalada*:

$$K(x_i, x_j) = \nu \exp\left(-\sum_{b=1}^B \frac{\|x_i^{(b)} - x_j^{(b)}\|^2}{2\sigma_b^2}\right) \quad (5)$$

donde ν es un factor de escalado, B es el número de bandas, y σ_b es un parámetro que controla la propagación de la relación entre cada una de las bandas espectrales (Verrelst *et al.*, 2014b,c; Wittenberghe *et al.*, 2014).

2.3. Validación cruzada

Existen diversas técnicas para validar los métodos de regresión, como son la comparación de los parámetros obtenidos con los obtenidos mediante modelos físicos teóricos o con simulaciones, utilizar nuevos conjuntos de datos conocidos para comparar con los obtenidos o el uso de técnicas de validación cruzada (Kozak and Kozak, 2003). Las técnicas que vamos a analizar en este trabajo pertenecen a este último grupo. Concretamente, vamos a ver los métodos *hold-out* y *k-fold*.

El método *hold-out* es el más sencillo de los distintos métodos de validación cruzada. Este separa el conjunto de datos disponibles en dos subconjuntos, uno utilizado para entrenar el modelo y otro para realizar el test de validación (Arlot y Celisse, 2010). De esta manera, se crea un modelo únicamente con los datos de entrenamiento. Con el modelo creado se generan datos de salida que se comparan con el conjunto de datos reservados para realizar la validación (que no han sido utilizados en el entrenamiento, por lo que no han sido utilizados para generar el modelo (Hawkins *et al.*, 2003)). Los estadísticos obtenidos con los datos del subconjunto de validación son los que nos dan la validez del método empleado en términos de error. Una aplicación alternativa de este método consiste en repetir el proceso *hold-out*, tomando distintos conjuntos de datos de entrenamiento (aleatorios) un determinado número de veces, de manera que se calculan los estadísticos de la regresión a partir de la media de los valores en cada una de las repeticiones.

El otro método utilizado, *k-fold*, está basado en el método anterior, pero con mayor utilidad cuando el conjunto de datos es pequeño (Yang y Huang, 2014). En este caso, el total de los datos se dividen en *k* subconjuntos, de manera que aplicamos el método *hold-out* *k* veces, utilizando cada vez un subconjunto distinto para validar el modelo entrenado con los otros *k-1* subconjuntos (Jung y Hu, 2015). El error medio obtenido de los *k* análisis realizados nos proporciona el error cometido por el método, permitiendo así evaluar su validez.

Si comparamos los dos métodos, el método *k-fold* tiene la ventaja de que todos los datos son utilizados para entrenar y validar, por lo que se obtienen resultados más representativos *a priori*. Mientras que, para el método *hold-out*, se realiza el proceso *n* veces de manera aleatoria, lo que no garantiza que los casos de entrenamiento y validación no se repitan.

2.4. Procedimiento experimental

Para el desarrollo de la parte experimental se han realizado cinco experimentos distintos:

- a) Obtención de LAI sin validación cruzada. Antes de analizar los valores estadísticos que se obtienen al aplicar los distintos métodos de validación cruzada, es interesante conocer los obtenidos sin aplicar dichos métodos, ya que son la referencia para saber si la aplicación de los métodos de validación cruzada nos da mejores estadísticos sobre el modelo generado. En este caso, se ha obtenido el valor de RMSE obtenido al generar los modelos mediante los métodos de regresión no paramétricos haciendo uso de distintos porcentajes de datos para entrenar y validar. Se ha calculado este término haciendo variaciones de 10% en la cantidad de datos de entrenamiento, variando estos desde el 10% hasta el 90%.
- b) Obtención de LAI con validación cruzada *hold-out*. Tomando la alternativa del método *hold-out* explicada en la sección 2.3, se ha aplicado el método *hold-out* para distintos números de repeticiones del proceso, concretamente para repeticiones desde 5 hasta 30, de 5 en 5. Además, dado que el porcentaje de datos de entrenamiento se mantiene constante durante las repeticiones, se ha repetido todo el proceso para distintos porcentajes de datos de entrenamiento (50%, 60% y 70%).

- c) Obtención de LAI con validación cruzada con método *k-fold*. En el caso de *k-fold*, se ha aplicado para los distintos métodos no paramétricos generando distintos subconjuntos de datos de entrenamiento-test, variando *k* desde 5 hasta 30, de 5 en 5.
- d) Comparación entre *hold-out* y *k-fold* utilizando el mismo porcentaje de datos de entrenamiento. Para realizar una comparación más realista entre los estadísticos obtenidos con ambos métodos de validación cruzada, dado que en los experimentos anteriores se han utilizado distintos porcentajes de datos, se ha aplicado el método de validación cruzada *hold-out* con 10 repeticiones y los porcentajes de datos correspondientes a los distintos valores de *k* (implementados en el método *k-fold*). En este caso, la comparación se realiza utilizando los métodos de regresión no paramétricos no lineales (KRR y GPR).
- e) Aplicación de los mejores métodos a una imagen HyMap. Para las mejores configuraciones de cada uno de los métodos de validación cruzada obtenidos, se ha aplicado el modelo generado a una imagen obtenida con el sensor HyMap, obteniendo de esta las imágenes de LAI y la desviación estándar.

Todas las comparaciones se han hecho haciendo uso del RMSE como medida estadística de referencia.

3. Resultados

3.1. Datos sin validación cruzada

La Figura 1 muestra los valores de RMSE obtenidos tras aplicar los distintos métodos no paramétricos para realizar el entrenamiento con distintos porcentajes de datos y realizando la validación con el subconjunto de datos restante escogido al azar.

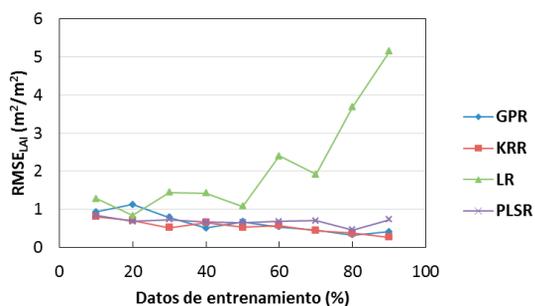


Figura 1. RMSE obtenido del entrenamiento-test sin hacer uso de la validación cruzada.

Observamos que los valores de RMSE de los métodos PLSR, KRR y GPR son considerablemente menores que los obtenidos con el método LR.

Se puede ver que el valor de RMSE disminuye ligeramente a medida que aumenta el porcentaje de datos de entrenamiento utilizado para los métodos PLSR, KRR y GPR. Para PLSR y GPR el mejor valor se obtiene haciendo uso del 80% de datos de entrenamiento, mientras que para KRR se obtiene con el 90% de datos de entrenamiento.

Cabe destacar que en este caso los resultados dependen del subconjunto escogido, y no del total de los datos, por lo que estadísticamente los resultados son menos fiables que en los casos posteriores, ya que además, no disponen de un rango de incertidumbre.

3.2. Validación cruzada *hold-out*

Se ha aplicado la validación cruzada *hold-out* a cada uno de los métodos no paramétricos empleados, con distinto número de repeticiones *hold-out* y haciendo uso del 50%, 60% y 70% de datos de entrenamiento. Estos resultados se muestran en la Figura 2, donde podemos analizar el impacto en términos de RMSE en el caso del modelo de regresión GPR.

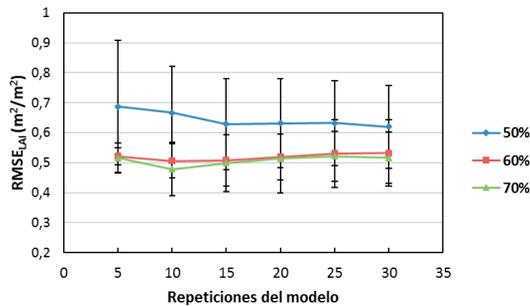


Figura 2. RMSE obtenido de la validación cruzada *hold-out* y el método de regresión GPR utilizando el 50 %, 60 % y 70 % de datos para entrenar el modelo.

De la Figura 2 podemos ver que para los modelos obtenidos haciendo uso del 70% de datos para entrenar el modelo se obtienen valores de RMSE menores.

En la Figura 3 se muestran los valores de RMSE obtenidos haciendo uso del 70% de los datos

disponibles para entrenar el modelo, y del 30% restante para validar.

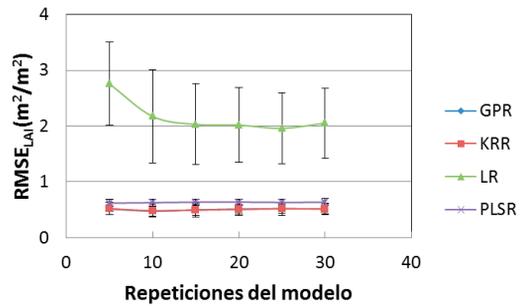


Figura 3. RMSE obtenido de la validación cruzada *hold-out* con el 70 % de datos para entrenar los modelos GPR, KRR, LR y PLSR.

Para poder visualizar mejor las diferencias entre los métodos que tienen valores más cercanos, se muestra en la Figura 4 los mismos resultados que en la Figura 3 pero, en este caso, tan solo para los modelos GPR, KRR y PLSR.

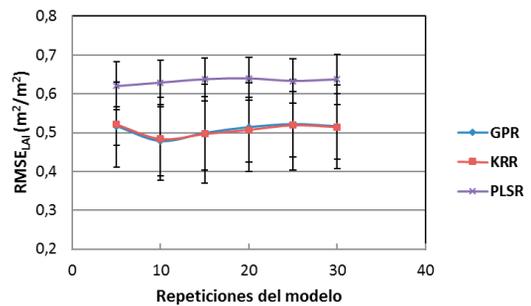


Figura 4. RMSE obtenido de la validación cruzada *hold-out* con el 70 % de datos para entrenar los modelos GPR, KRR y PLSR.

En las figuras 3 y 4 podemos ver que las diferencias entre los valores de RMSE obtenidas entre los dos métodos lineales son considerables, siendo bastante menor el obtenido con el método PLSR, que presenta valores cercanos a los obtenidos con los métodos no lineales. En cuanto a los métodos no lineales, se puede observar que no hay diferencias apreciables entre ambos métodos en términos de RMSE. Sin embargo, sí hay que destacar la diferencia temporal de ejecución del modelo, donde el MLRA KRR es más rápido que

el GPR. Esta diferencia en el tiempo de ejecución podemos apreciarla en la Tabla 1, donde se muestra el tiempo que ha tardado cada uno de los modelos en generarse y validarse según el número de repeticiones realizadas.

Tabla 1. Tiempo de ejecución del método *hold-out* para los métodos de regresión GPR y KRR en las distintas repeticiones realizadas.

	5	10	15	20	25	30
GPR (s)	30,13	30,62	27,65	26,31	28,99	24,82
KRR (s)	0,49	0,62	0,51	0,43	0,49	0,43

Se observa que a partir de las 10 repeticiones del modelo, el valor de RMSE en el método LR se estabiliza a un valor RMSE de 2 m²/m², mientras que para el resto de métodos no tiene una gran influencia el número de repeticiones empleadas, ya que las variaciones que se obtienen en términos de RMSE son muy pequeñas. De la misma manera, las barras de error muestran la desviación estándar de las medidas, obteniendo valores cercanos a 0,7 m²/m² en el método LR. En cambio, para los otros tres métodos las desviaciones obtenidas son muy pequeñas comparando con las de LR, obteniendo valores menores en el caso del método lineal PLSR, alrededor de 0,05 m²/m², y próximos a 0,1 m²/m² para los métodos GPR y KRR, aunque son menores en el primero de ellos.

Adicionalmente, se muestra en la Figura 5 la representación gráfica de cada uno de los puntos

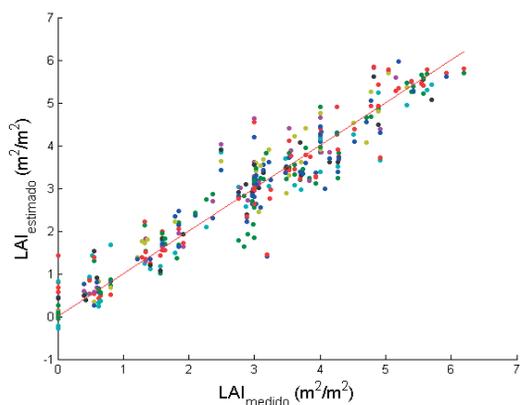


Figura 5. Representación gráfica de los valores estimados de LAI comparados con los medidos al aplicar la validación cruzada *hold-out*, haciendo uso de 10 repeticiones con el método de regresión GPR y el 70 % de datos de entrenamiento. Los puntos de un mismo color muestran los valores de una misma repetición.

utilizados en la validación cruzada con 10 repeticiones del modelo para GPR.

Cada uno de los colores mostrados en la Figura 5 representa las medidas de un subconjunto al que se le aplica el método *hold-out*. Debido a esto se diferencian 10 colores, ya que la gráfica mostrada corresponde al caso de 10 repeticiones. Observamos que la mayoría de medidas se aproximan bastante a la recta 1:1. Visualmente se observa una desviación similar a lo largo de todos los valores y conjuntos de datos.

3.3. Validación cruzada *k-fold*

Se ha realizado la validación cruzada con el método *k-fold* para valores de *k* entre 5 y 30 en cada uno de los métodos utilizados. En la Figura 6 se muestran los valores de RMSE obtenidos al aplicar cada valor de *k*. Igual que ocurría con el método *hold-out*, el método LR obtiene valores de RMSE considerablemente superiores a los otros MLRAs. El método lineal PLSR tiene un error ligeramente superior a los métodos no lineales, aunque con valores aceptables. En cuanto a los métodos no lineales, nuevamente obtienen valores similares, siendo la principal diferencia el tiempo de ejecución del método.

Para poder visualizar mejor las diferencias entre los métodos que tienen valores más cercanos, se muestra en la Figura 7 los mismos resultados que en la Figura 6 pero, en este caso, tan solo para los modelos GPR, KRR y PLSR.

Observamos en las figuras 6 y 7 observamos que para LR se obtienen valores de RMSE en un rango entre 2 y 3 m²/m², entre 0,56 y 0,62 m²/

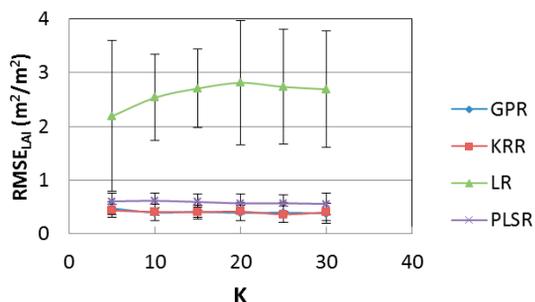


Figura 6. RMSE obtenido de la validación cruzada *k-fold* con el 70 % de datos para entrenar los modelos GPR, KRR, LR y PLSR.

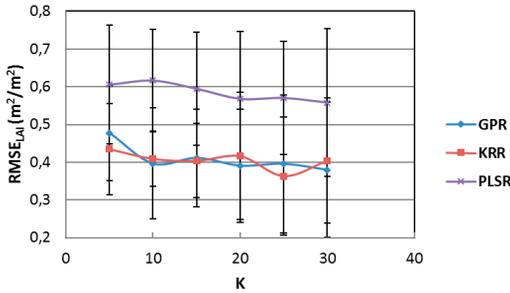


Figura 7. RMSE obtenido de la validación cruzada *k-fold* con el 70 % de datos para entrenar los modelos GPR, KRR y PLSR.

m^2 para PLSR, entre 0,36 y 0,43 m^2/m^2 para KRR y entre 0,38 y 0,48 m^2/m^2 para GPR. En este caso, para $k=10$ el valor de RMSE para LR empieza a estabilizarse. Por otra parte, aunque para $k=5$ el valor es menor, hay que tener en cuenta que la dispersión es mucho mayor, por lo que no se considera que sea un buen valor de k para utilizar el método LR. En cuanto a los otros tres métodos, podemos ver en las figuras 6 y 7 que la diferencia que se observa para cada k es mínima. Además, los k con menor RMSE obtienen un mayor valor de desviación estándar, siendo 10 y 15 los valores de k con menor desviación estándar.

En la Figura 8 se muestra la representación gráfica de los valores de LAI estimados comparados con los medidos aplicando los datos de validación cruzada *k-fold* para k igual a 10.

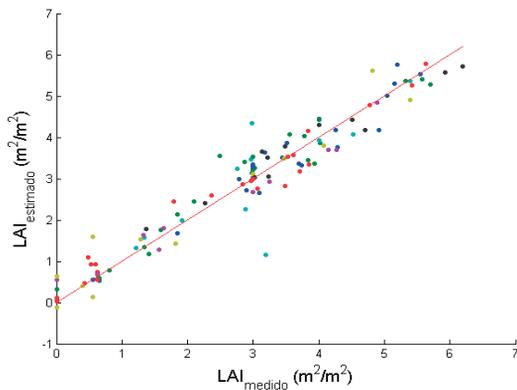


Figura 8. Representación gráfica de los valores estimados de LAI comparados con los medidos al aplicar la validación cruzada *k-fold*, haciendo uso de $k = 10$ para el método de regresión GPR y el 70 % de datos de entrenamiento. Los puntos de un mismo color muestran los valores de un mismo subconjunto.

Cada uno de los colores de la Figura 8 pertenece a uno de los k subconjuntos distintos utilizados para validar. Concretamente, debido a que se muestra la representación para k igual a 10, se observan 10 colores distintos. Se puede observar que en este caso la desviación de los puntos es menor que el que se apreciaba en la Figura 5. Tan solo un punto azul, localizado aproximadamente en el punto (3,1), se separa en mayor medida de la línea 1:1.

3.4. Comparación entre métodos *hold-out* y *k-fold* utilizando el mismo porcentaje de datos de entrenamiento

Dado que los porcentajes de datos de entrenamiento utilizados en cada uno de los casos anteriores no son iguales, se ha realizado un nuevo cálculo en el que se ha aplicado el método *hold-out* para realizar la validación con el porcentaje de datos de entrenamiento equivalente al utilizado en cada una de las distintas medidas realizadas con el método *k-fold*, es decir, el porcentaje equivalente a utilizar un número distinto de k . Se ha aplicado el método de validación cruzada *hold-out* con 10 repeticiones, obteniendo así la gráfica mostrada en la Figura 9.

Vemos en la Figura 9 que los valores de RMSE para LR se encuentran entre 2 y 3 m^2/m^2 , entre 0,51 y 0,56 m^2/m^2 para PLSR, entre 0,35 y 0,45 m^2/m^2 para KRR y entre 0,39 y 0,42 m^2/m^2 para GPR. Si comparamos los resultados de la Figura 9 con los de la Figura 6, observamos que para el método LR, en este caso es para $k = 5$ donde obtenemos el RMSE más bajo y con menor dispersión, es decir,

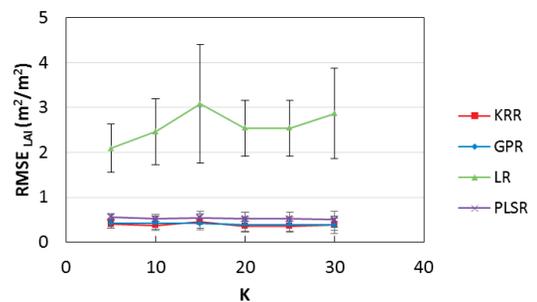


Figura 9. RMSE obtenido tras aplicar el método *hold-out* con 10 repeticiones para los porcentajes de datos de entrenamiento equivalentes a los utilizados cuando se hace uso del método *k-fold*.

haciendo uso del 80% de datos para entrenar el modelo, mientras que para el resto de métodos nuevamente observamos que el RMSE obtenido difiere muy poco al variar el porcentaje de datos de entrenamiento. Además, se obtiene que los mejores métodos son KRR y GPR, mientras que los errores obtenidos con PLSR son siempre un poco superiores a estos.

Se observa por tanto que las diferencias son mínimas para un mismo método no paramétrico de regresión comparando los dos tipos de validación cruzada empleados. En cambio, sí que se observan notables diferencias entre los distintos métodos de regresión utilizados, especialmente entre los métodos lineales y los no lineales. Entre los métodos lineales utilizados, el LR ofrece términos de RMSE muy elevados, tal como se ha visto, por lo que su uso no es recomendable en estas aplicaciones. Mientras que en el caso del PLSR, aunque con resultados de RMSE mayores que los métodos no lineales, la diferencia con éstos últimos es relativamente pequeña, por lo que puede ser un algoritmo a tener en cuenta. Por otra parte, la comparación entre los métodos no lineales KRR y GPR no muestra diferencias a destacar entre ambos métodos, ya que los dos métodos ofrecen resultados en el mismo rango, tal como puede observarse en la Tabla 2.

La Tabla 2 muestra los datos numéricos de RMSE mostrados gráficamente en las secciones 3.1 y 3.2 para los dos casos de análisis actuales.

Tabla 2. RMSE obtenido para los métodos de validación cruzada: *hold-out* con 10 repeticiones y 10-Fold.

MLRA	<i>hold-out</i> (m ² /m ²) (10 repeticiones)	10-fold (m ² /m ²)
GPR	0,48	0,40
KRR	0,48	0,41
LR	2,17	2,54
PLSR	0,63	0,62

En la Tabla 2 se observa que el método que obtiene un menor valor de RMSE es el GPR con la validación cruzada 10-fold, aunque con una diferencia muy pequeña respecto al método KRR. En el caso de la validación cruzada *hold-out*, los métodos GPR y KRR tienen el mismo valor. Hay que destacar que en el caso del LR el valor de RMSE obtenido con *hold-out* es menor que con 10-fold, pero es el único método con el que ocurre este hecho.

3.5. Aplicación a una imagen HyMap

Una vez analizados por separado los distintos métodos, se va a realizar la comparación entre la mejor configuración de cada uno de ellos. Por ello, se va a realizar la comparación entre el método de validación cruzada *hold-out* con 10 repeticiones y el método *k-fold* con *k* igual a 10 (10-fold). Los modelos obtenidos con estos métodos han sido aplicados a una imagen HyMap para ver el efecto de su implementación en una imagen.

En las figuras 10 y 11 se muestra la comparación entre estos dos métodos de validación cruzada, aplicando el MLRA GPR con la validación cruzada 10-fold y *hold-out*, respectivamente, sobre una imagen de 500×500 píxeles tomada con el sensor HyMap.

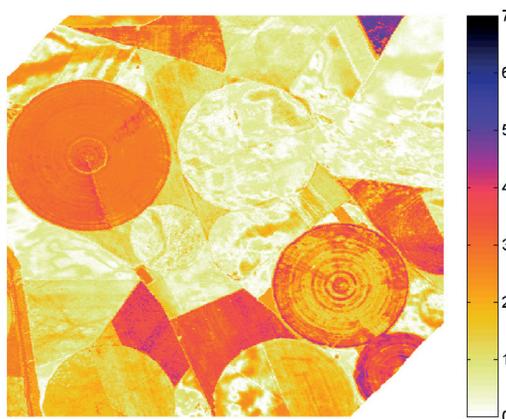


Figura 10. Mapa de LAI (m²/m²) obtenida aplicando la validación cruzada 10-fold con el MLRA GPR.

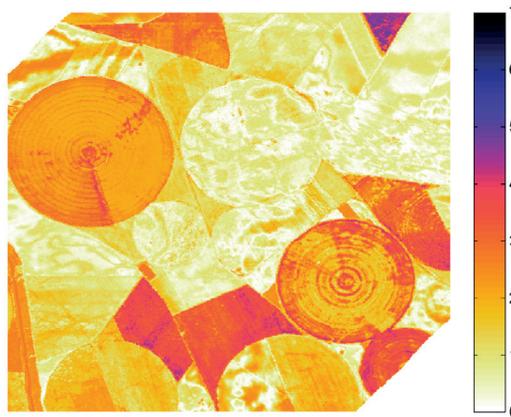


Figura 11. Mapa de LAI (m²/m²) obtenida aplicando la validación cruzada *hold-out* (10 repeticiones) con el MLRA GPR.

En las figuras 10 y 11 podemos observar que, en ciertas zonas con vegetación (valores de LAI alrededor de $3 \text{ m}^2/\text{m}^2$), el valor representado en la Figura 8 es mayor que en la Figura 9, por lo que puede decirse que en estos casos el modelo generado con el método 10-fold da como resultado valores ligeramente mayores que el generado con el método *hold-out*.

Las figuras 12 y 13 muestran la desviación estándar obtenida al elaborar el mapa de LAI con los diferentes métodos (10-fold y *hold-out*, respectivamente).

Tal como se puede observar en las figuras 12 y 13, se obtienen imágenes muy similares a simple vista, lo que implica que las diferencias son pequeñas.

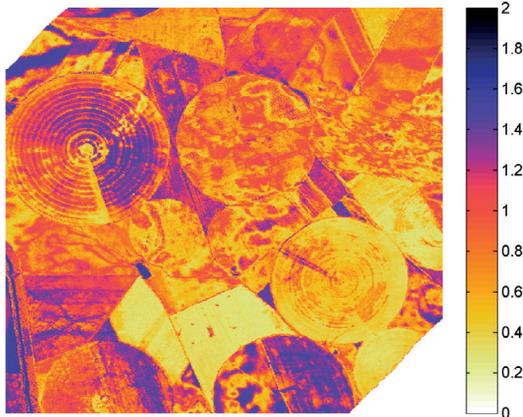


Figura 12. Desviación estándar del mapa de LAI (m^2/m^2) obtenida aplicando la validación cruzada 10-fold con el MLRA GPR.

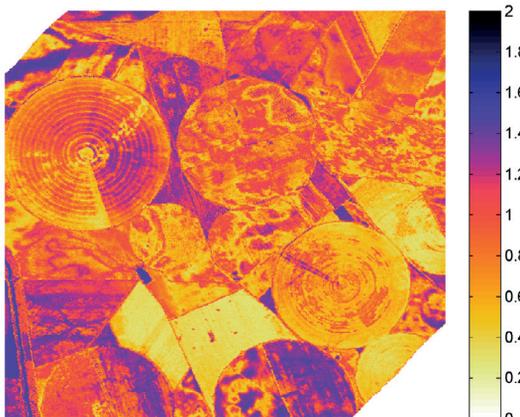


Figura 13. Desviación estándar del mapa de LAI (m^2/m^2) obtenida aplicando la validación cruzada *hold-out* (10 repeticiones) con el MLRA GPR.

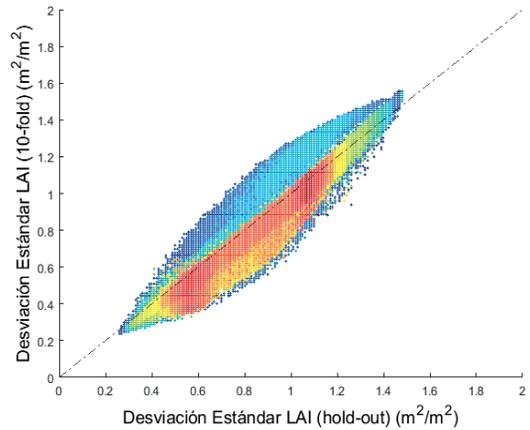


Figura 14. Desviación estándar de la imagen obtenida con 10-fold frente a *hold-out*. Los puntos con tonos más cálidos indican una mayor densidad de puntos mientras que los tonos más fríos indican una menor densidad de puntos.

Para analizar con una mayor claridad que método ofrece valores de desviación estándar menores, representamos en la Figura 14 la desviación estándar obtenida mediante el método 10-fold frente a la obtenida con el método *hold-out*.

Se observa en la Figura 14 que existe una mayor densidad de píxeles con una diferencia sistemática en la que la desviación estándar obtenida mediante el método 10-fold da generalmente valores menores que los obtenidos con el método *hold-out*.

4. Discusión

En el presente trabajo se han evaluado distintos métodos de regresión MLRAs haciendo uso de las técnicas de validación cruzada existentes en la toolbox ARTMO. En concreto se han analizado los métodos lineales LR y PLSR y los no lineales KRR y GPR, haciendo uso de un conjunto de datos de la campaña SPARC-2003 realizada en Barrax (España). Estos datos contienen medidas de LAI de diferentes cultivos e información de reflectividad de las 125 bandas espectrales del sensor hiperespectral HyMap.

La principal diferencia entre los métodos de regresión GPR y KRR reside en las propias características de los modelos, ya que el método GPR permite conocer las bandas más relevantes así como los intervalos de confianza de la predicción realizada (Rasmussen and Williams, 2006).

Los resultados obtenidos están en conformidad con los obtenidos en Rivera *et al.* (2014b), donde se consideran los MLRAs KRR (con RMSE de 0,42 m²/m² para la imagen HyMap) y GPR (con RMSE de 0,30 m²/m² para la imagen HyMap) como los algoritmos de regresión más robustos, mientras que los algoritmos no paramétricos lineales como el PLSR se consideran sistemáticamente inferiores a estos.

Este trabajo ha sido realizado con datos del sensor hiperespectral HyMap, por lo que es de interés realizar el análisis en sensores con otras características espectrales, así como analizar con mayor profundidad la aplicación de los métodos no paramétricos no lineales, los cuales ofrecen mejores resultados (Rivera *et al.*, 2014b).

5. Conclusiones

Conocida la importancia de la obtención de parámetros biofísicos para el seguimiento de la fenología y salud de la vegetación, en este trabajo se ha analizado mediante distintos métodos de validación cruzada el error obtenido al aplicar diversos métodos de regresión no paramétricos. Los métodos estudiados han sido el LR y PLSR como ejemplos lineales y KRR y GPR como no lineales, todos ellos desde la toolbox ARTMO.

Del análisis realizado, se ha obtenido que, por una parte, los métodos de regresión MLRAs no lineales ofrecen resultados con errores menores que los métodos lineales. Además, entre los dos métodos no lineales analizados, no existe diferencia en cuanto a valores de RMSE, pero la desviación estándar del RMSE es mayor en el caso del MLRA KRR respecto al GPR, sin embargo, este último necesita más tiempo para ejecutarse.

Por otro lado, en referencia a los métodos de validación cruzada aplicados, el método *k-fold* ofrece valores de RMSE para cada uno de los métodos ligeramente mejores, por lo que realiza la validación con una mayor precisión. Más concretamente, en el caso del LR obtenemos el mejor resultado para un valor de *k* igual a 10, mientras que para los otros métodos el valor de *k* tiene una menor importancia ya que todos obtienen un error similar. Si analizamos el método *hold-out*, aunque los valores de RMSE que obtiene son un poco mayores, estos siguen siendo aceptables para los métodos no lineales y para el PLSR, especialmente cuando

utilizamos el 70% de los datos para entrenar y el 30% restante para validar. En este caso, el método LR ofrece sus mejores valores cuando utilizamos el método *hold-out* repitiéndolo 10 veces.

Para la mejor configuración de cada uno de los métodos, es decir, haciendo uso de *k=10* para *k-fold*, y 10 repeticiones y un 70% de los datos para entrenar en el caso de *hold-out*, se han aplicado los modelos obtenidos a una imagen HyMap, llegando a determinar que en zonas con mayor valor de LAI, los valores obtenidos con 10-fold son mayores que con *hold-out*, así como la desviación estándar es menor, lo que nos lleva a determinar que el método 10-fold ofrece resultados con una incertidumbre menor que los obtenidos con *hold-out*.

Por lo tanto, podemos concluir que a pesar del coste computacional que requiere aplicar el método *k-fold*, es conveniente su uso ya que ofrece mejores resultados, con menor incertidumbre, especialmente cuando el conjunto de datos disponible no es excesivamente grande.

Agradecimientos

Este estudio ha contado con financiación de la Generalitat Valenciana (a través del contrato Gerónimo Forteza de LI. Pérez, FPA/2015/081) en el marco del proyecto CGL2011-30433-C02-02 financiado por el Ministerio de Economía y Competitividad.

Referencias

- Arlot, S., Celisse, A. 2010. A Survey of Cross-validation Procedures for Model Selection. *Statistics Surveys*, 4, 40-79. <http://dx.doi.org/10.1214/09-SS054>
- Bazi, Y., Alajlan, N., Melgani, F., AlHichri, H., Yager, R.R. 2014. Robust estimation of water chlorophyll concentrations with gaussian process regression and IOWA aggregation operators. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 7(7), 3019-3028. <http://dx.doi.org/10.1109/JSTARS.2014.2327003>
- Delegido, J., Verrels, J., Rivera, J.P., Ruiz-Verdú, A., Moreno, J. 2015. Brown and green LAI mapping through spectral indices. *International Journal of Applied Earth Observation and Geoinformation*, 35, 350-358. <http://dx.doi.org/10.1016/j.jag.2014.10.001>

- Guerschman, J.P., Scarth, P.F., McVicar, T.R., Renzullo, L.J., Malthus, T.J., Stewart, J.B., Rickards, J.E., Trevithick, R. 2015. Assessing the effects of site heterogeneity and soil properties when unmixing photosynthetic vegetation, non-photosynthetic vegetation and bare soil fractions from Landsat and MODIS data. *Remote Sensing of Environment*, 161, 12-26. <http://dx.doi.org/10.1016/j.rse.2015.01.021>
- Hawkins, D.M., Basak, S.C., Mills, D. 2003. Assessing Model Fit by Cross-Validation. *Journal of Chemical Information and Computer Sciences*, 43(2), 579-586. <http://dx.doi.org/10.1021/ci025626i>
- Jung, Y., Hu, J. 2015. A k-fold averaging crossvalidation procedure. *Journal of Nonparametric Statistics*, 27(2), 167-179. <http://dx.doi.org/10.1080/10485252.2015.1010532>
- Kozak, A., Kozak, R. 2003. Does cross validation provide additional information in the evaluation of regression models? *Canadian Journal of Forest Research*, 33(6), 976-987. <http://dx.doi.org/10.1139/x03-022>
- Lázaro-Gredilla, M., Verrelst, J., Camps-Valls, G. 2014. Retrieval of Biophysical Parameters With Heteroscedastic Gaussian Processes. *IEEE Geoscience and Remote Sensing Letters*, 11(4), 838-842. <http://dx.doi.org/10.1109/LGRS.2013.2279695>
- Rasmussen, C.E., Williams, C.K.I. 2006. *Gaussian Processes for Machine Learning*. The MIT Press, New York, US.
- Rivera, J.P., Verrelst, J., Delegido, J., Veroustraete, F., Moreno, J. 2014a. On the Semi-Automatic Retrieval of Biophysical Parameters Based on Spectral Index Optimization. *Remote Sensing*, 6, 4927-4951. <http://dx.doi.org/10.3390/rs6064927>
- Rivera, J.P., Verrelst, J., Muñoz-Marí, J., Moreno, J., Camps-Valls, G. 2014b. Toward a Semiautomatic Machine Learning Retrieval of Biophysical Parameters. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 7(4), 1249-1259. <http://dx.doi.org/10.1109/JSTARS.2014.2298752>
- Shao, J. 1993. Linear Model Selection by Cross-Validation. *Journal of the American Statistical Association*, 88(422), 486-494. <http://dx.doi.org/10.1080/01621459.1993.10476299>
- Verrelst, J., Alonso, L., Camps-Valls, G. 2012. Retrieval of Vegetation Biophysical Parameters Using Gaussian Process Techniques. *IEEE Transactions on Geoscience and Remote Sensing*, 50(5), 1832-1843. <http://dx.doi.org/10.1109/TGRS.2011.2168962>
- Verrelst, J., Rivera, J.P., Leonenko, G., Alonso, L., Moreno, J. 2014a. Optimizing LUT-Based RTM Inversion for Semiautomatic Mapping of Crop Biophysical Parameters from Sentinel-2 and -3 Data: Role of Cost Functions. *IEEE Transactions on Geoscience and Remote Sensing*, 52(1), 257-269. <http://dx.doi.org/10.1109/TGRS.2013.2238242>
- Verrelst, J., Alonso, L., Rivera, J.P., Moreno, J., Camps-Valls, G. 2014b. Gaussian Process Retrieval of Chlorophyll Content from Imaging Spectroscopy Data. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 6(2), 867-874. <http://dx.doi.org/10.1109/JSTARS.2012.2222356>
- Verrelst, J., Rivera, J.P., Moreno, J., Camps-Valls, G., 2014c. Gaussian Processes uncertainty estimates in experimental Sentinel-2 LAI and leaf Chlorophyll content retrieval. *ISPRS Journal of Photogrammetry and Remote Sensing*, 86, 157-167. <http://dx.doi.org/10.1016/j.isprsjprs.2013.09.012>
- Verrelst, J., Camps-Valls, G., Muñoz-Marí, J., Rivera, J.P., Veroustraete, F., Clevers, J., Moreno, J. 2015. Optical remote sensing and the retrieval of terrestrial vegetation bio-geophysical attributes – A review. *ISPRS Journal of Photogrammetry and Remote Sensing*, 108, 273-290. <http://dx.doi.org/10.1016/j.isprsjprs.2015.05.005>
- Wittenberghe, S.V., Verrelst, V., Rivera, J.P., Alonso, L., Moreno, J., Samson, R. 2014. Gaussian processes retrieval of leaf parameters from a multi-species reflectance, absorbance and fluorescence dataset. *Journal of Photochemistry and Photobiology B: Biology*, 134, 37-48. <http://dx.doi.org/10.1016/j.jphotobiol.2014.03.010>
- Yang, Y. y Huang, S., 2014. Suitability of five cross validation methods for performance evaluation of nonlinear mixed-effects forest models – a case study. *Forestry*, 87, 654-662. <http://dx.doi.org/10.1093/forestry/cpu025>