Universidad de Valladolid

FACULTAD DE CIENCIAS

TRABAJO FIN DE GRADO

Grado en Física

Elaboración de modelos de calibrado para el cálculo de composiciones químicas con LIBS

Autor/a: Bogusz Ferens Michalek

Tutor/es/as: Guillermo López Reyes y José Antonio Manrique Martínez

Año 2023

Contenidos

1. Introducción

El instrumento SuperCam forma parte de la carga útil del rover Perseverance, lanzado el 30 de julio de 2020 como parte de la misión Mars 2020. El diseño del instrumento tiene como objetivo realizar un análisis de la superficie de Marte a un amplio rango de distancias alrededor del Rover, proporcionando información composicional de las diferentes muestras. Para ello, SuperCam implementa seis técnicas de observación: Laser Induced Breakdown Spectroscopy (LIBS), espectroscopía resuelta en el tiempo Raman o de luminiscencia (TRR/L), espectroscopía de reflectancia en el visible e infrarrojo cercano (VISIR), fotografía en color de alta resolución (RMI), y grabación acústica (MIC). Todas ellas con la capacidad de funcionar coalineadas y proporcionar datos de la misma muestra y en el mismo punto.

SuperCam es la evolución del instrumento anterior, ChemCam, del rover Curiosity. Toma un papel importante a la hora de cumplir los objetivos propuestos por la misión, como estudiar la geología del cráter Jerezo, comprobar la habitabilidad del planeta o buscar evidencias de la existencia de vida. El instrumento es capaz de realizar medidas de distintas características del terreno, en particular, la espectroscopía LIBS que permite la identificación de la composición química de las rocas presentes en el entorno. En este trabajo nos enfocaremos en el entrenamiento de modelos de regresión que

proporcionen estimaciones de la abundancia de diferentes elementos químicos a partir de los datos LIBS recolectados por SuperCam.

1.1. Espectroscopía LIBS

La espectroscopía LIBS utiliza un pulso laser de alta intensidad para inducir un plasma en la muestra a analizar. En esta técnica usamos el campo eléctrico de la luz proveniente de un láser pulsado para ionizar una pequeña fracción de la muestra a analizar, induciendo un plasma.

En un primer momento, cuando el pulso laser penetra en la muestra, esta se calienta e ioniza, generando un plasma que gradualmente se irá expandiendo y enfriándose. Como resultado, los diferentes iones inducidos irán volviendo a átomos neutros a medida que los electrones vuelven a ser capturados. Es en este proceso que se emiten fotones en longitudes de onda concretas, y dependientes de cada elemento químico. LIBS es, por tanto, una técnica capaz de detectar, virtualmente, cualquier elemento de la tabla periódica. Dada su rapidez, y baja necesidad de preparación de la muestra es un a técnica muy interesante que cuenta con diversos desarrollos y aplicaciones, siendo la exploración espacial la aplicación de interés en este trabajo.

El uso de datos LIBS para realizar modelos cuantitativos cuenta con bastantes ejemplos en la comunidad, y es base de muchos y diferentes trabajos publicados en muy diversas aplicaciones. En todo caso es importante remarcar las dificultades de la técnica a la hora de obtener resultados cuantitativos. En primer lugar, es una técnica que tiene un marcado carácter temporal. Como se ha indicado, el plasma presenta una evolución a medida que se expande y enfría. Los dos efectos más relevantes son:

- La evolución de la emisión de bremsstrahlung, que es un fondo de radiación continuo en el rango de medida que proviene de las diferentes aceleraciones de las partículas cargadas en el plasma. Este fondo va ligado a la temperatura del plasma y decae a medida que se enfría. Diferentes instrumentos con diferentes capacidades de resolución temporal captarán espectros con diferentes fondos.
- Las diferentes líneas espectrales de los diferentes elementos dependerán del instrumento que se use, siendo capaces de alcanzar diferentes niveles de ionización en función del diseño y potencia. Estas diferencias también existirán en las relaciones entre las intensidades de las líneas espectrales para un mismo elemento, y dependerán de la temperatura máxima alcanzada por el plasma, de hecho existe metodología para calcular esta temperatura a partir del análisis de las intensidades de las diferentes líneas de un determinado elemento (J. Chem. Educ., 2013).

Por todos estos elementos, unidos a las inestabilidades propias de los componentes de un instrumento, los modelos de regresión usando datos LIBS son procesos que se hacen para cada instrumento, y requieren de interpretación y análisis minucioso. En SuperCam o ChemCam, además, se incluye el factor de la distancia variable de la muestra, lo que produce una irradiancia diferente según distancia y los modelos deben adaptarse a estos factores.

1.2. SuperCam

En cuanto a los componentes de este instrumento, SuperCam tiene tres componentes principales (Figura 1):

- SuperCam Mast Unit (SCMU): El mástil donde se encuentran el láser para las técnicas espectroscópicas Raman, LIBS y Luminiscencia Resuelta en Tiempo, así como la cámara de imagen, Remote Micro Imager (RMI) y el espectrómetro de Infrarrojo..

- SuperCam Body Unit (SCBU): Esta unidad va montada dentro del cuerpo del vehículo, en ella se localizan los espectrómetros que se utilizan para las medidas, además de servir como fuente de alimentación para toda la MU.
- SuperCam Calibration Target (SCCT): Es la unidad que permite realizar pruebas de calibración de SuperCam en las mismas condiciones ambientales que las propias muestras que se analizan posteriormente.

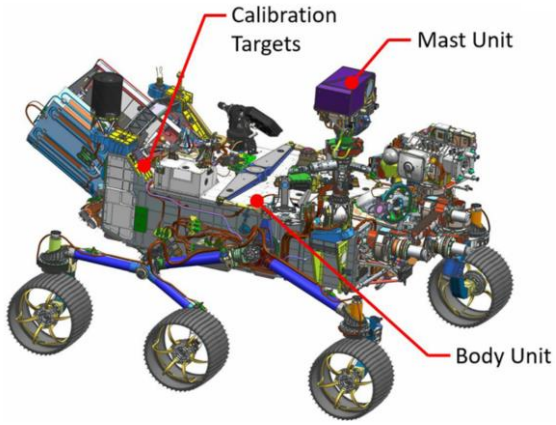

Figura 1: Componentes de SuperCam en el vehículo Perseverance. [2] Maurice et al. 2021.

1.2.1. Mast Unit (SCMU)

El telescopio de SuperCam se encuentra en la parte más elevada del mástil para poder realizar observaciones a distancia. El rango de acción nominal de SuperCam (2-7m) para las técnicas que dependan del láser, aunque la imagen y la espectroscopía pasiva pueden abordar distancias remotas. Junto con MastCamZ forma parte de los llamados instrumentos de ciencia remota. SuperCam colabora también con otros instrumentos situados en el brazo robótico, como PIXL o SHERLOC y que son los llamados instrumentos de ciencia de proximidad. De esta manera Perseverance tiene acceso a un área de trabajo de 7 metros de radio sin tener que desplazarse, aunque las medidas de RMI y VISIR se pueden realizar a distancias mucho más altas (de decenas de kilómetros), dependiendo de las condiciones atmosféricas en el planeta. Esta información es tenida en cuenta a la hora de planificar las actividades del Rover, y en la selección de muestras de interés para la ciencia de proximidad.

La unidad mástil se encuentra montada en el Remote Sensing Mast, (RSM), entre otros instrumentos como cámaras de navegación. Todo ello se encuentra en la Remote Warm Electronic Box (RWEB), con dos aperturas: Una para el micrófono y otra más grande para el telescopio. La apertura del telescopio está tapada por un cristal plano inclinado a 3.5º para evitar que el láser pueda reflejarse de vuelta al interior del telescopio. Toda la construcción de la unidad del mástil permite rotaciones de +-181º horizontalmente, y +-91º verticalmente, lo que permite el acceso al amplio área de medidas.

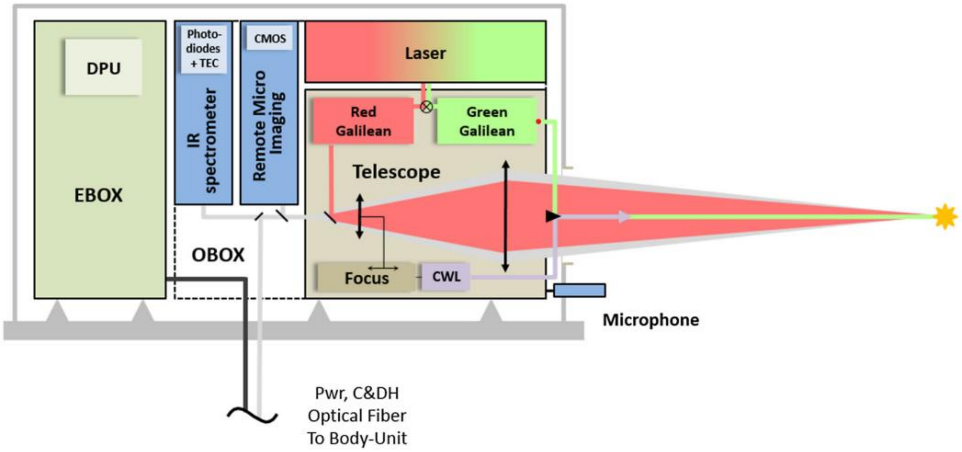

Figura 2: Esquema de la MU. Maurice et al. 2021.

El láser que se utiliza en SuperCam es una evolución del utilizado en ChemCam (En el rover Curiosity). Está diseñado para operar en temperaturas de entre -40ºC y 30ºC, y para sobrevivir en temperaturas de entre -50ºC y 55ºC. Para alcanzar las temperaturas necesarias dos calentadores llevan al laser a la temperatura necesaria.

El láser puede realizar disparos únicos, pero normalmente se utilizan ráfagas. Estas ráfagas son generalmente de entre 20 y 50 disparos realizados a 3Hz, sin embargo la frecuencia puede aumentar a 10Hz. El láser es un Nd-YAG cuya frecuencia base está en 1064nm, mediante un doblador de frecuencia se puede generar el segundo armónico para obtener 532nm (Maurice et al., 2021), longitud de onda requerida para su uso en la espectroscopía Raman y de luminiscencia, dado que la conversión nunca es del 100%, se utiliza un obturador que bloquea el camino del primer pulso para eliminar residuos de 1064 nm que puedan interferir con las medidas. Por simplicidad, el pulso de 1064nm se nombra como rojo y el de 532nm se nombra como verde. Adicionalmente, el láser tiene un tercer modo de operación de 852 nm para el láser de longitud de onda continua (CWL).

A diferencia de ChemCam, SuperCam incluye la capacidad de hacer espectroscopía Raman. Si comparamos la dispersión inelástica de la luz (efecto Raman) con la dispersión elástica de la luz, dispersión Rayleigh, ésta última es estadísticamente más probable que la dispersión Raman. Esto hace que la intensidad del espectro Raman esté varios órdenes de magnitud por debajo de la intensidad de luz de 532 nm recibida por el instrumento. Para poder registrar correctamente los espectros Raman el Mast Unit incorpora una serie de filtros interferométricos que dejan pasar la luz en el rango de medidas LIBS salvo en un rango alrededor de los 532 nm (Maurice et al., 2021),(Wiens et al., 2021). Este filtro, unido a otra serie de filtros en el demultiplexor del Body Unit, encargados de llevar cada rango espectral al espectrómetro adecuado, hacen que SuperCam no pueda registrar datos entre 465 nm y 535 nm, los rangos son diferentes en comparación con ChemCam.

El diseño óptico de la Optical Box (OBOX) permite realizar las funciones necesarias para las técnicas de espectroscopía: enfocar el láser rojo para LIBS, colimar el verde para TRR, y recolectar luz proveniente de la muestra. Para ello tenemos en el interior de la OBOX un telescopio Schmidt-Cassegrain, dos expansores de haz galileanos y un divisor de haz.

Puesto que SuperCam está diseñado para operar a distancias de entre 2 y 7 metros, el tamaño del foco en la muestra variará en función de esta distancia. El tamaño del “spot” puede ser desde 300 µm hasta 600 µm de diámetro (Maurice et al. 2021), este cambio supone un aumento de 4 veces el tamaño

del área irradiada por el láser. Es debido a esto que la distancia a la que se encuentra la muestra será un dato importante a la hora de analizar los espectros de emisión.

El objetivo focaliza la luz al divisor de haz, que distribuye la luz en función de su longitud de onda para los distintos análisis que se puedan realizar.

La luz proveniente de la muestra se transporta a la BU por medio de un cable de fibra óptica de 5.8m.

1.2.2. Body Unit (SCBU)

La Body Unit es la parte del aparato encargada de todo el análisis de la radiación proveniente de la muestra. Para realizar esta tarea está compuesta por varios componentes importantes:

- Un demultiplexor.
- Dos espectrómetros de reflexión Czerny-Turner que cubren entre 240 y 340 nm, en el caso del UV (Ultra Violeta) y entre 380 y 460 para el VIO (violeta), ambos usados para hacer LIBS.
- Un espectrómetro de transmisión para Raman, TRL, VIS y LIBS, que cubre entre 535 nm y 860 nm..
- Una caja electrónica (EBOX), al igual que en la MU.

El demultiplexor divide la luz en tres bandas espectrales, acoplando la luz a bandas de fibra óptica que la transmiten a las rendijas de cada espectrómetro. Consta de una lente colimadora, dos espejos dicroicos que separan la luz ultravioleta y violeta para sus respectivos espectrómetros, y un tercer espejo para la luz restante (Wiens et al., 2021).

En la parte final del demultiplexor, por la que pasa la tercera componente de la luz, se encuentra un filtro que elimina la luz del láser de 532nm. Este filtro es la segunda parte del sistema doble de filtros para eliminar esta luz. El primer filtro no eliminaba esta luz en su totalidad, por lo que aquí se utiliza un filtro Semrock que elimina la radiación restante.

Los espectrómetros de reflexión Czerny-Turner sólo se diferencian en sus espejos y sus redes de reflexión, ya que están diseñados para operar con rangos de longitudes de onda diferentes. Su funcionamiento de los espectrómetros viene representado en la figura 3: La luz proveniente del cable de fibra óptica entra al sistema por una rendija en la parte superior y queda colimada por el primer espejo. La red de reflexión de la parte superior derecha es la que logra la dispersión espectral de esta luz, que queda dirigida hacia segundo espejo, en la parte superior izquierda, que es el encargado de focalizarla sobre el detector de la parte derecha.

Para el análisis de la luz solo son de interés las reflexiones de primer orden, por lo que en las partes superior e inferior de la red de reflexión se sitúan deflectores ópticos que absorben las reflexiones de orden superior.

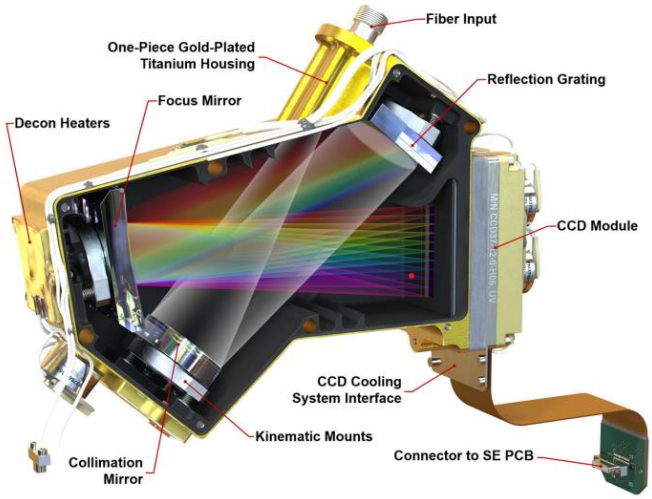

Figura 3: Espectrómetro Czerny-Turner. [3] Wiens et al. 2021.

La luz finalmente alcanza el dispositivo de carga acoplada (CCD), que transforma luz en impulsos eléctricos y aporta la información necesaria para construir el espectro de emisión de la banda de luz que hemos recibido. Este detector es capaz de operar con exposiciones como mínimo en el orden de los 8 milisegundos, este tiempo de exposición, muy superior a un evento de plasma en una medida LIBS, unido a no tener capacidad de introducir retardos entre el detector y el láser, hacen que SuperCam no pueda realizar medidas resueltas en tiempo con estos espectrómetros.

La luz restante que no ha sido interceptada por estos espectrómetros (con longitudes de onda entre 535nm y 850nm) pasa al tercer espectrómetro de transmisión. Este espectrómetro tiene una construcción diferente, ya que además de analizar espectros LIBS se encarga de realizar las técnicas Raman y TRL. La diferencia más notable entre este espectrómetro y los dos anteriores es que proyecta sobre la CCD tres bandas espectrales simultáneamente, posicionadas una sobre otra como se representa en la figura 3, y además incluye un intensificador. Esta disposición ofrece la posibilidad de optimizar la resolución en el rango de interés para el análisis Raman (535nm - 676nm), cubriendo un rango bastante amplio usando un solo detector.

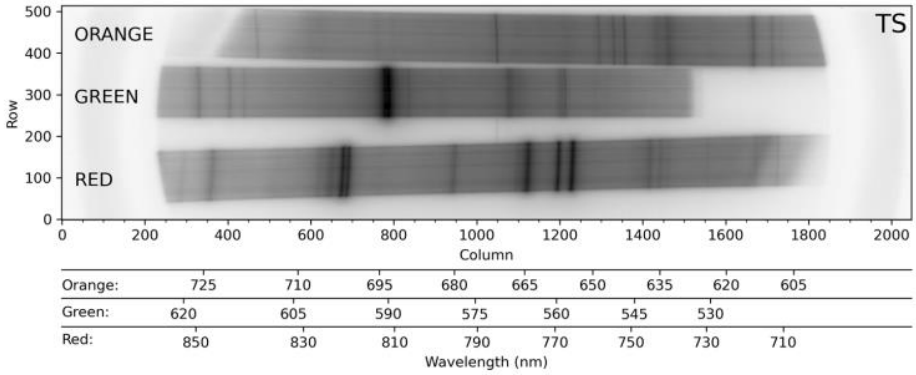

Figura 4: Detección de luz realizada por la CCD del espectrómetro de transmisión (TS). Wiens et al. 2021.

En figura 4 se representa el funcionamiento del espectrómetro de transmisión: La luz pasa por la rendija y se colima antes de pasar por un separador de haz dicroico, que separa la banda roja. Ambas bandas de luz viajan de forma paralela a dos redes de difracción, la de la banda roja produce su difracción directamente mientras que para el otro haz de luz se utilizan varias redes dispuestas una

detrás de la otra. Este segundo sistema logra la difracción de la luz naranja y la amarilla por separado, de modo que terminamos con tres bandas de luz que se focalizan sobre el intensificador antes de pasar finalmente a la CCD. Es este elemento, el intensificador, el que nos permite además adquirir espectros en tiempos muy cortos (tan cortos como 100 ns, para medidas Raman, o de 0.5 ms para medidas LIBS) y además nos da la capacidad de introducir retardos entre una señal de entrada, que en nuestro caso es una señal que avisa de que el laser se ha disparado, y el momento en que el intensificador está activo, permitiendo análisis resueltos en tiempo (Wiens et al., 2021).

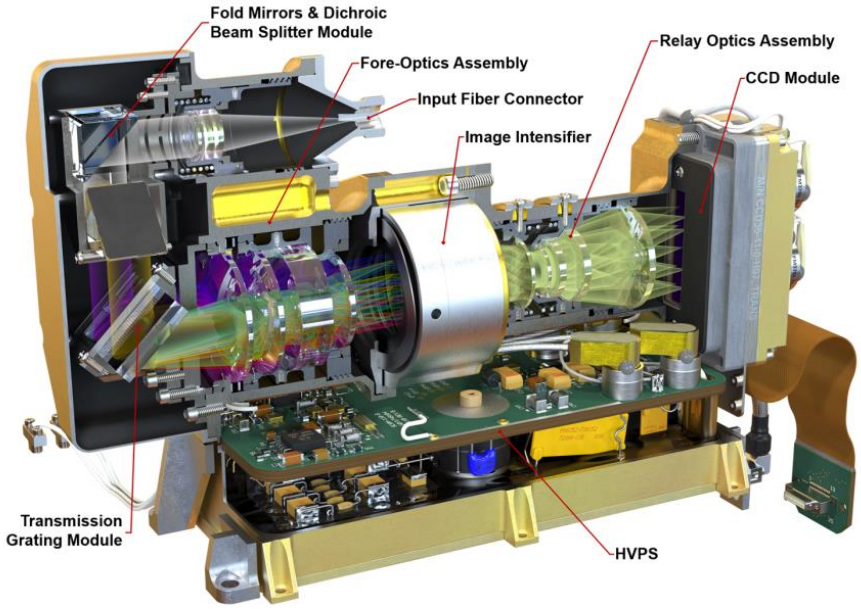

Figura 5: Espectrómetro de transmisión. Wiens et al. 2021.

La separación de la luz en estas tres bandas y sus diferentes características es importante, ya que más adelante será un motivo para considerar distintas normalizaciones de los espectros.

1.2.3. SuperCam Calibration Targets (SCCT)

Situados en la parte trasera del rover se encuentran las muestras de calibración, a una distancia fija de 1.56 metros del detector. A pesar de que los espectros de calibración se obtuvieron en la tierra simulando las condiciones de marte, es necesario poder realizar pruebas de calibración en el propio planeta para poder compararlas con las realizadas aquí, estos espectros son los que utilizaremos para entrenar los modelos de regresión.

La función de SCCT es poder controlar la evolución de SuperCam a lo largo de su misión. El instrumento se puede deteriorar al aterrizar, atravesar condiciones adversas o sencillamente deteriorarse con el tiempo por la acumulación de polvo. SCCT ofrece un objetivo conocido y constante sobre el que comprobar el funcionamiento del aparato en cualquier momento de la misión. También permite estudiar el funcionamiento de SuperCam en las condiciones cambiantes de Marte, e incluso relacionarlo con información de estas condiciones captada por otros instrumentos del Rover.

Se dispone de 36 muestras de calibración para el instrumento completo, 23 de ellas son dedicadas a LIBS. Esto es un aumento muy grande con respecto a las 10 que poseía ChemCam, lo que es una gran ventaja a la hora de estudiar un rango de composiciones más grande. En la Figura 2 se muestra una representación de las SCCT, las muestras desde 7 hasta 28 son las correspondientes a LIBS. La última

muestra de interés es la 33: se trata de una placa de titanio cuya función es la calibración de las longitudes de onda para varias técnicas de espectroscopía, entre ellas LIBS (Cousin et al., 2021).

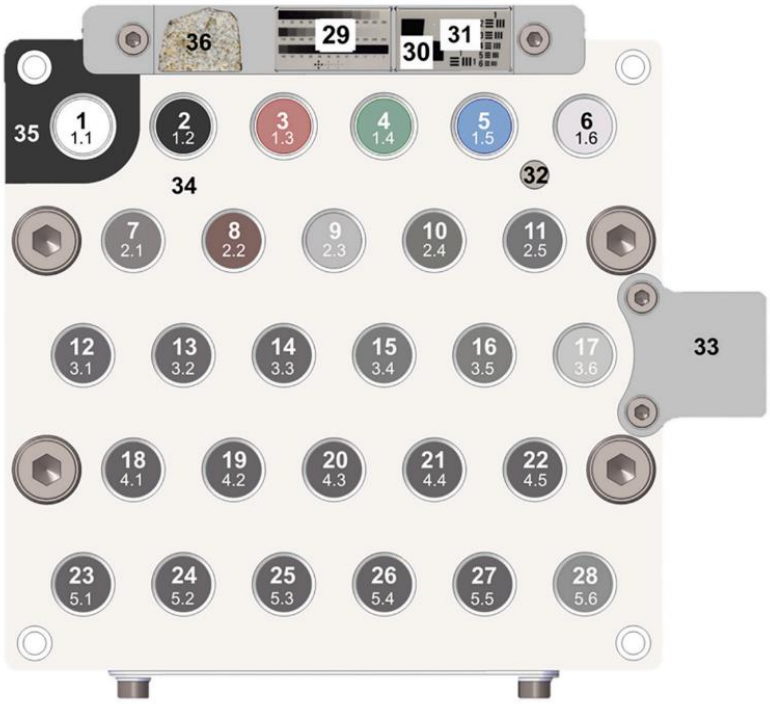

Figura 6: Muestras de calibración de SuperCam. [4] Cousin el al. 2021.

Estas muestras de calibración cumplen distintos propósitos a la hora de estudiar las composiciones del planeta. En general tenemos composiciones de minerales que se esperan encontrar en la zona de aterrizaje del rover al igual que silicatos dopados con elementos minoritarios (Cu, Cr, Zn, Li, Ni), todas estas muestras tienen como objetivo la detección de estos elementos o minerales (Manrique et al., 2020).

Otras muestras permiten detectar productos naturales del clima marciano o comparar la calibración con el modelo antiguo, Chemcam, ya que es una muestra original de ese modelo.

1.3. Normalización de datos y pretratamiento

En primer lugar, debemos normalizar nuestro set de datos para poder compararlos entre ellos. Esto se debe a que SuperCam obtiene espectros en un rango variable de distancias, y a que las muestras presentan diferentes características. Es necesario utilizar algún método de normalización para tratar de que, en la medida de lo posible, variaciones entre dos espectros sean debidas exclusivamente a diferentes composiciones. Para ello utilizaremos 3 tipos de normalización y podremos ver cuál de ellas nos dará el mejor resultado.

En primer lugar utilizamos la normalización más sencilla, la normalización total. Para cada espectro se toma las intensidades medidas y se dividen por la suma total de todas ellas, de esta manera el espectro queda normalizado de tal manera que la suma de todas sus intensidades es igual a 1. La suma se hace para todas las intensidades considerando el espectro como un único rango.

En segundo lugar, está la normalización por la línea del oxígeno. Esta normalización es útil ya que es un elemento con unas líneas de emisión fáciles de detectar, y suele estar presente en la atmósfera en la que estamos midiendo. Si simulamos las condiciones de marte habrá una concentración de CO_2

conocida en el aire alrededor de la muestra por lo que veremos las líneas de emisión del oxígeno en el espectro medido. Para hacer la normalización, tomamos de cada espectro el valor de la intensidad máxima medida en el rango de 1nm entorno al pico característico del oxígeno más intenso, que se encuentra en la longitud de onda de 777,35nm. A continuación dividimos las intensidades del espectro por dicho valor, de esta forma todos los espectros estarán normalizados de tal manera que el pico del oxígeno tomará siempre el valor 1. Esta normalización tiene en cuenta el hecho de que el CO2 marciano estará siempre presente en todas las medidas, pero puede verse afectada por otros factores como el oxígeno en la propia muestra, o las diferentes condiciones de irradiancia del láser que genera el plasma a las diferentes distancias.

Por último, podremos normalizar los espectros por partes, esta es la normalización utilizada por el equipo de SuperCam (A. Ryan et al., 2021). Como ya se vio anteriormente, el instrumento obtiene el espectro usando cinco rangos espectrales diferentes, que a su vez se obtienen por tres espectrómetros diferentes, cubriendo de manera discontinua un rango entre 250 a 850nm. Los rangos son aproximadamente 250-350, 350-460, 532-620, 620-712 y 712-850. Los dos primeros rangos se toman en los espectrómetros de tipo Czerny Turner, con 2048 píxeles en el detector, mientras que los tres últimos se toman en el espectrómetro de transmisión, que cuenta con una red de difracción holográfica, que proyecta tres tracks en el detector, siendo además este espectrómetro el único que cuenta con intensificador. En cada uno de estos las intensidades no se miden en intervalos iguales y además las intensidades relativas entre ellos no tienen por qué ser iguales. Esta normalización consiste en realizar la normalización total cinco veces, una para cada rango de longitudes de onda. Cada rango estará normalizado de tal manera que la suma de sus intensidades será igual a 1. En esta normalización conseguimos focalizarnos en cada rango por separado, haciendo las intensidades obtenidas por cada rango comparables de manera directa, aunque en cierto modo es muy similar a la normalización total.

1.4. Análisis de datos

Para manejar el conjunto de datos y generar los modelos utilizaremos el sistema Matrix Laboratory (MatLab) de manera que podremos realizar el tratamiento de datos, probando diferentes normalizaciones de los espectros, para su posterior análisis y reducción dimensional usando Principal Component Analysis (PCA) sobre ellos. Adicionalmente, usaremos la herramienta Regression Learner, que nos permitirá entrenar los modelos de regresión que nos proporcionen un cálculo de abundancia de cada elemento a partir de los datos LIBS.

1.4.1. PCA

PCA es una técnica estadística para describir un conjunto de datos en un nuevo sistema de variables en función de la varianza que describen. Con estas nuevas variables podremos seleccionar aquellas que nos aporten la mayor cantidad de información sobre nuestro sistema, de manera que será posible reducir enormemente la dimensionalidad del conjunto de datos. Esto es de especial interés en el caso que nos ocupa ya que un espectro LIBS de SuperCam, cubriendo un rango espectral entre 245–853 nm, está compuesto de 7933 puntos, es decir, 7933 variables por espectro.

Para hacer PCA, lo primero que debemos calcular es la covarianza entre todas nuestras variables. Cada covarianza para una pareja de variables nos indica de qué manera cambian una en función de la otra, es decir, es una medida de la relación que tienen estas dos variables en nuestro conjunto de datos.A partir de aquí, construiremos la matriz de covarianza, que incluye las covarianzas para cada pareja de variables que existen en nuestro sistema. Dado que la covarianza entre una variable x con y es la misma que la de y con x, la matriz de covarianza será cuadrada y simétrica. Esto significa que podremos encontrar los autovalores y autovectores de esta matriz, que nos darán la información necesaria para el cambio de base que estamos buscando.

Cada autovector es una nueva variable que usaremos para describir nuestro conjunto de datos en un nuevo sistema de coordenadas o ''componentes principales''. A cada autovector lo acompaña su autovalor, y en este caso los autovalores nos dan una medida de la cantidad de varianza del sistema completo que describe esta nueva coordenada. Sabiendo esto, ordenaremos los autovectores en orden decreciente de sus autovalores, de modo que las primeras componentes principales son las que más información nos ofrecen del sistema.

Como tenemos dos sistemas de coordenadas de igual dimensión, podremos transformar nuestro set de datos a las nuevas coordenadas, y como conocemos la varianza que describe cada coordenada, podremos seleccionar aquellas que nos aporten la mayor cantidad de información, reduciendo el número de variables. En nuestro caso partimos de casi 8000 canales de intensidad, pero tras hacer PCA podemos ver que con las 10 primeras componentes principales explicaremos más del 95% de varianza del sistema en todos los casos. ([5] Lindsey I Smith, 2002)

MatLab utiliza por defecto el algoritmo de descomposición en valores singulares (SVD).

Gracias a PCA podemos reducir el número de variables del sistema de casi 8000 a tan solo 10, a cambio de perder una pequeña cantidad de información.

Mas adelante veremos que en la práctica siempre seleccionaremos estas 10 primeras componentes principales para generar nuestros modelos. Sería posible hacer un estudio para ver el desempeño de los diferentes modelos entrenados en función del número de componentes que seleccionemos, ciertos elementos como el silicio pueden tener una alta varianza en todos los datos espectrales y una menor cantidad de PCs podrían servir para realizar los cálculos más eficientemente. Sin embargo, en este trabajo estandarizado el número de variables para centrarnos más en comparar los diferentes modelos y normalizaciones, y esta selección es una cantidad que ya explica más del 95% de la varianza del sistema.

1.4.2. Feature Selection

También comprobaremos la eficacia de ciertos algoritmos de selección de facetas como Maximum Relevance Minimum Redundancy (mRMR). Estos algoritmos toman las componentes ordenadas por su varianza explicada, y asignan valores en función de su relevancia real sobre el sistema, orientado a un objetivo específico, esto es, evaluando la relación de estas componentes principales con la abundancia del elemento en cuestión, creando una nueva clasificación. Esta clasificación puede ser útil, puesto que ciertas variables, a pesar de aportar mucha información sobre el sistema, pueden llegar a aportar la misma información que otras, es decir, son redundantes.

Esto es precisamente la metodología de estos algoritmos. En el caso de mRMR nuestra única decisión que tomamos es el número de facetas que queremos que seleccione, y esto será el número de iteraciones que realizara el algoritmo sobre todo el conjunto de variables. En cada iteración se selecciona una variable que será la más relevante de todas, y una vez realizada esta elección es permanente y no será considerada en las próximas iteraciones. La clave es que además de medir la relevancia de la variable para la medida que queramos realizar, el algoritmo compara su redundancia con las variables que ya han sido seleccionadas en iteraciones anteriores. Si a pesar de ser muy relevante la información proporcionada por una variable se solapa con la de una variable que hemos seleccionado anteriormente, esta no será seleccionada ya que además sabemos que la faceta ya seleccionada es más relevante por haber sido seleccionada anteriormente. Repitiendo este proceso conseguimos reordenar nuestras PCs para una selección más eficaz (Samuele Mazzanti, Towards Data Science, 2021).

De nuevo, por simplicidad, seleccionaremos las diez primeras variables seleccionadas por estos algoritmos en caso de que alguna de las PCs seleccionadas quedase fuera de las diez primeras.

1.4.3. Modelos de Regresión

Con nuestras variables seleccionadas en cada caso utilizaremos Regression Learner para entrenar distintos modelos de regresión, y finalmente poder obtener la información de la composición de una muestra a partir de su espectro.

Para el entrenamiento de los diferentes modelos se han separado los espectros en dos conjuntos, uno de entrenamiento, que luego describiremos cómo se usa, y otro de prueba que se deja fuera del entrenamiento y sobre el que evaluaremos los modelos. Es importante remarcar que el conjunto de prueba que se deja fuera es, en general, el de las muestras que forman parte de la muestra de calibración de SuperCam que está en Marte. Esto se hace por dos motivos: por un lado, tener un set de muestras de comparación entre Marte y la Tierra, y por otro lado porque estas muestras fueron las únicas que se tomaron a diferentes distancias, dado el poco tiempo disponible para pruebas antes del lanzamiento. De esta manera se puede evaluar mejor el desempeño en Marte de modelos basados en datos de Tierra, y a la vez ver el impacto de la distancia en las diferentes normalizaciones y modelos entrenados.

Los modelos que utilizamos tienen todos el mismo objetivo: crear un modelo que predice el valor de una variable de destino en función de diversas variables de entrada. Pero el proceso utilizado para lograr esto difiere mucho entre ellos, a continuación explicaremos someramente el funcionamiento de los modelos más importantes que se utilizarán.

Tree

Consideremos primero una simple regresión lineal, esto es un modelo global en la que una única ecuación predice la información contenida por un cierto conjunto de variables. Cuando el número de variables aumenta y comienzan a interactuar entre ellas de maneras complejas y no lineales, aplicar un modelo tan sencillo resulta extremadamente complicado y pierde su efectividad a la hora de utilizarse. Por ello, una forma de solucionar este problema es dividir el conjunto de variables en clasificaciones de menor tamaño o particiones, donde las interacciones entre ellas es más predecible y manejable.

Este método utiliza como base para su funcionamiento un árbol de decisión, que representa esta sucesión de particiones y se crea utilizando un set de entrenamiento. Cada partición se nombra como una hoja o nodo, y todas las hojas se unen por medio de ramas. En cada nodo, se realiza una pregunta sobre el set de variables que tenemos, y cada posible respuesta es una de las ramas, de modo que de cada nodo obtenemos un número de nodos nuevos igual a todas las ramas que han surgido de él. De todas las ramas se selecciona aquella cuyo nodo minimiza su desviación típica, haciendo que el comportamiento de las variables sea cada vez más predecible. Este proceso continúa hasta que los nodos alcanzan un tamaño especificado o bien la desviación típica es cero.

En esta situación hemos llegado al llamado nodo terminal, donde el proceso termina y se aplica un modelo de regresión sencillo. Al haber utilizado un set de entrenamiento, el modelo estará sobreentrenado y se utiliza el conjunto de validación para ‘’podar’’ el árbol, lo que reduce su tamaño y complejidad.

Podar el árbol consiste en medir la relevancia de distintas ramas del árbol respecto al resultado final, con el objetivo de simplificar el modelo, aumentando la eficiencia y evitando el sobreentrenamiento. Este proceso se realiza por prueba y error eliminando ciertas ramas para ver cómo afecta esto al

resultado final del algoritmo, de esta forma se revela qué partes del árbol son redundantes respecto al resultado final (Wu et al., 2008).

Kernel

Utilizamos dos modelos, support vector machine (SVM) y Least squares regression (LSR) con la aproximación Kernel. Esta aproximación utiliza las llamadas funciones Kernel: se definen como una distribución de probabilidad que asignan el peso de un dato de entrenamiento en su entorno, cuanto más alejado del punto original, menor es el peso. Al atribuir una función Kernel a cada dato de entrenamiento podemos obtener en cualquier punto la suma de los pesos de todas las funciones Kernel, que nos ofrecerán una aproximación del conjunto de entrenamiento extendida a todo el continuo. (Niranjan Pramanik, Towards Data Science, 2019).

La aplicación de la aproximación Kernel a los modelos de regresión ofrece un camino más sencillo para general estos modelos, haciendo que sean más simples y rápidos de entrenar, con pérdidas mínimas en su precisión.

SVM

Este es un modelo que permite separar un conjunto de variables separando el espacio en el que se encuentran por medio de hiperplanos, dando lugar a subconjuntos de variables que las clasifican. El modelo busca el hiperplano que encuentre la mayor distancia entre sí mismo y cualquiera de las variables que clasifique, sin embargo en casos complicados este hiperplano puede ser muy complejo (Wu et al., 2008).

LSR

Least squares o mínimos cuadrados, es un tipo de regresión que busca encontrar una función contínua que minimize el error cuadrático medio entre dicha función y los datos que tenemos (Heinmueller et al., 2014).

Encontrar una función que se ajuste a los datos en casos complejos de alta dimensión puede ser una tarea difícil. De nuevo, el uso de funciones kernel nos hará más sencillo encontrar la función que buscamos.

Redes Neuronales Artificiales (Artificial Neural Networks, ANN)

Las redes neuronales son modelos de predicción que se inspiran en el funcionamiento de las neuronas en el cerebro. A simple vista son similares a los árboles de decisión mencionados previamente ya que funcionan enviando información entre distintos nodos o neuronas, pero su funcionamiento real es diferente.

Una red neuronal se estructura por medio de varias capas de neuronas, cada neurona se conecta con otra y tienen asignado un peso y un umbral. Al recibir datos, una neurona genera un valor como respuesta a la entrada, este valor de salida se ve a su vez multiplicado por el peso que se ha asignado a la neurona, atribuyéndose de esta forma lo importante que es respecto al resto. Si el valor de salida supera el valor umbral establecido, las neuronas envían esta información a la siguiente capa continuando el proceso hasta llegar a la capa de salida, que es la última de todas.

La forma de entrenar esta red de neuronas consiste en un algoritmo que utiliza como referencia continuamente la llamada función de coste, que es el error cuadrático medio de la predicción con el valor real de la medida. El algoritmo ajusta los valores de peso y umbral de las neuronas buscando

minimizar la función de coste, esto se consigue con el algoritmo de gradiente descendiente: un método que permite conocer en qué dirección ajustar los parámetros para alcanzar este mínimo.

La mayoría de redes neuronales funcionan en único sentido, avanzando por las capas de neuronas desde la entrada a la salida. Se pueden entrenar también modelos que pueden mover la información en la dirección opuesta, lo que permite atribuir a cada neurona su error asociado. Esto es muy útil para mejorar los parámetros del modelo de forma mucho más eficiente. Adicionalmente, en nuestro caso Regression Learner permite utilizar el modelo optimizable, que entrena numerosas veces la red, cambiando el número de capas y neuronas para encontrar el modelo más eficiente posible (What are neural networks?, 2021).

Gaussian Process Regression (GPR)

Este tipo de regresiones tienen un aspecto diferente a la idea general que tenemos. En lugar de buscar encontrar una función que nos dé el mejor modelo posible para predecir unos ciertos resultados, esta regresión aplica una distribución de probabilidad a todos los posibles valores de los parámetros de la función de regresión.

La forma de entrenar este modelo consiste en tomar una función de probabilidad inicial, y ajustarla utilizando los datos de entrenamiento. Cada nuevo espectro de entrenamiento aporta información nueva al entrenamiento que actualiza la función anterior utilizando la regla de Bayes. Las distribuciones de probabilidad calculadas nos permiten pesar todos los posibles valores de las predicciones que se calculan, es decir, cada vez que se hace una predicción GPR aplica todo el set de entrenamiento para hacerlo. Esto hace que GPR sea uno de los modelos que más coste computacional tiene (Hilarie Sit, 2019).

2. Objetivos

En este trabajo intento, usando el mismo set de datos que empleó el equipo de ciencia de SuperCam, entrenar diferentes modelos quimiométricos para algunos elementos mayoritarios de la mineralogía marciana, evaluando algunas aproximaciones diferentes a las que realizaron en el caso de SuperCam.

En primer lugar busco evaluar el funcionamiento de distintos métodos de machine learning para realizar predicciones de la composición de una muestra a partir del espectro medido por SuperCam. Algunos modelos fueron empleados por SuperCam, otros se emplean por primera vez con estos datos en este trabajo.

Además de buscar de entre todos los modelos aquel que mejor se ajuste a la realidad, éstos se entrenarán utilizando distintas normalizaciones con el objetivo de evaluar su eficacia y sus posibles aplicaciones. Esta normalización es necesaria a fin de eliminar la influencia de las diferentes distancias de medida en los espectros obtenidos, pero tiene impacto en los datos de entrada en los modelos a entrenar. Evalúo la mejor aproximación de normalización entre las tres más comunes empeñadas en ChemCam y SuperCam.

El uso de algoritmos de selección de facetas también se verá analizado. Siendo una aproximación que no se utilizó en el caso de elementos mayoritarios para SuperCam. En general en este trabajo usaremos como referencia los modelos que están siendo usados actualmente por Supercam y que se detallan en (A. Ryan et al., 2021).

3. Materiales y métodos

3.1. Espectros

Los espectros que utilizaremos para la calibración son los tomados por el instrumento SuperCam (J.A. Manrique et al. (2020). A. Cousin et al. (2022). J.M. Madrigara et al. (2022)). Los sets de datos fueron tomados en Tierra, durante el año 2020 previo al lanzamiento de la misión. Estos datos corresponden a una base de datos de minerales en la que gran parte de los minerales ya formaban parte de la base de datos mineralógica del instrumento ChemCam (Maurice et al. 2021. Wiens et al. 2021.) de la misión Mars Science Laboratory (Grotzinger et al. (2012)). Se trata de muestras caracterizadas cuidadosamente para tener una composición elemental lo más exacta posible, y sobre las que se han obtenido espectros LIBS a diferentes distancias, en el caso de las muestras correspondientes a la muestra de calibración de SuperCam, y a tres metros para la base de datos extendida. Estos datos son accesibles de forma pública en el Planetary Data System (PDS), un archivo de datos obtenidos en las misiones espaciales realizadas por la NASA.

Los espectros adquiridos por un instrumento como SuperCam, o ChemCam, en los que la distancia a la muestra varía, se ven afectados, como ya se discutió, por una variación del tamaño de spot de focalización y por tanto de la irradiancia máxima conseguida para la generación del plasma. Las diferentes normalizaciones se han seleccionado para tratar de estandarizar los datos obtenidos a diferentes distancias.

Dentro de las posibles variabilidades entre espectros hay que remarcar otra, que son las posibles diferencias existentes entre los espectros obtenidos en Marte y los usados en la generación de modelos tanto por el equipo de ciencia de SuperCam, como por este trabajo. En verdad, los espectros de esta librería de datos se obtuvieron por el instrumento SuperCam previamente a su lanzamiento, y con las muestras en condiciones marcianas de presión, temperatura y composición atmosférica. Es sobre este instrumento en estas condiciones, además, que se efectuó la corrección de la función instrumento de SuperCam (Carey Legett et al. 2022.). Por motivos evidentes, una vez el instrumento está en Marte, tanto la corrección de la respuesta del instrumento, como evaluaciones del efecto de la distancia no pueden realizarse. El único medio efectivo de comparación entre SuperCam en Tierra, cuando adquirió los datos usados para los modelos, y SuperCam en Marte, donde adquiere los datos de interés científico, es la muestra de calibración de SuperCam, SCCT.

La muestra de calibración de SuperCam cuenta, además de con el modelo de vuelo que está en Marte, con réplicas en la Tierra en el Laboratorio Nacional de Los Álamos (LANL), en EEUU, en el Instituto de Investigación en Astrofísica y Planetología (IRAP) en Toulouse y en la Universidad de Valladolid. Las muestras que forman parte de este componente han sido preparadas de manera que se garantice la homogeneidad en composición no solo dentro de la muestra, sino también entre las diferentes réplicas. Estas muestras de calibración son, por tanto, el elemento que se puede usar para comparar el comportamiento del instrumento en Marte y en la Tierra, o con las réplicas del instrumento.

Dada la limitación en tiempo previa al lanzamiento de la misión Mars2020, el tiempo disponible para la adquisición de la librería de datos y de actividades de caracterización del SuperCam fue limitado. En lo referente a la base de datos para la elaboración de modelos se procedió a la adquisición de espectros para la librería completa a una distancia de tres metros, más característica de las distancias habituales de operación de SuperCam, y dos sets adicionales en las muestras de calibración, uno a la distancia nominal de la muestra de calibración montada en la parte trasera del Perseverance (1.56 m) y otro a una distancia de cuatro metros.

En general, en la elección de los diferentes sets de entrenamiento y test, se ha procurado usar los espectros de estas muestras de calibración para el conjunto de test. Esto permite evaluar el desempeño de los modelos con un set de muestras común entre los datos de Marte, los datos de Tierra y los futuros datos obtenidos por réplicas de SuperCam, como ya se introdujo.

3.2. Elementos Mayoritarios

Los elementos mayoritarios que estudiaremos son Aluminio, Silicio, Calcio, Sodio, Hierro y Magnesio, que se presentan en nuestro set de datos formando parte de distintas sustancias: Al2O3, SiO2, CaO, Na2O, FeOT, MgO. El rango de concentraciones que podremos estudiar difiere entre estos elementos, y vienen representados en la siguiente gráfica.

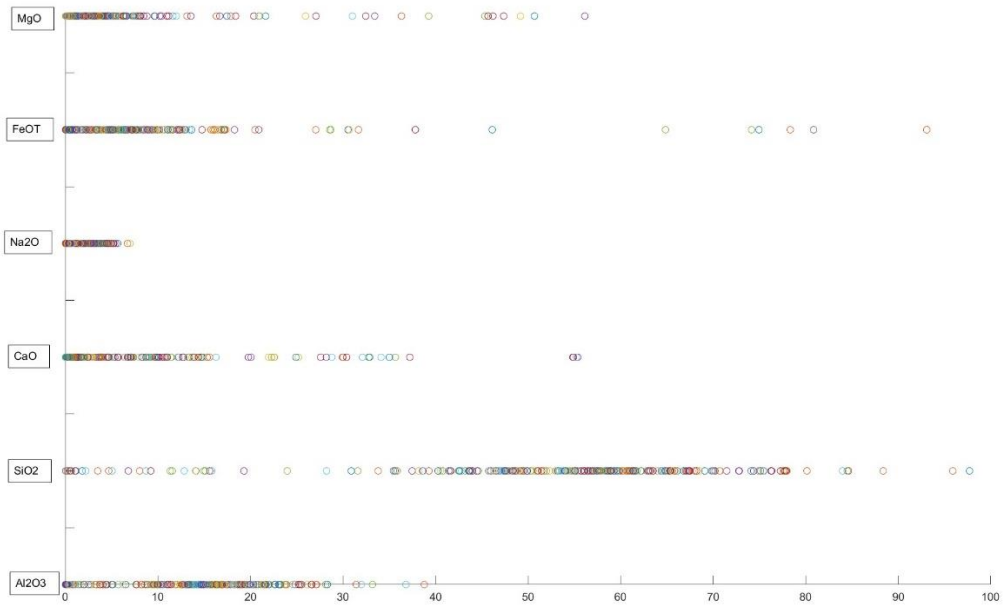

Figura 7

Por lo general en nuestros espectros tendremos concentraciones inferiores al 50% con la excepción del Silicio, para el que sus concentraciones suelen estar entre 50% y 70% para la mayoría de los espectros..

Estas composiciones deberían cubrir la mineralogía esperada en el cráter Jezero. Para verlo, podemos comparar las concentraciones de la figura 2 con la figura 3. En ella se representan las concentraciones de estas mismas sustancias presentes en varios tipos de minerales comunes en la geología del cráter. La información sobre estas composiciones se puede obtener de la página webmineral.

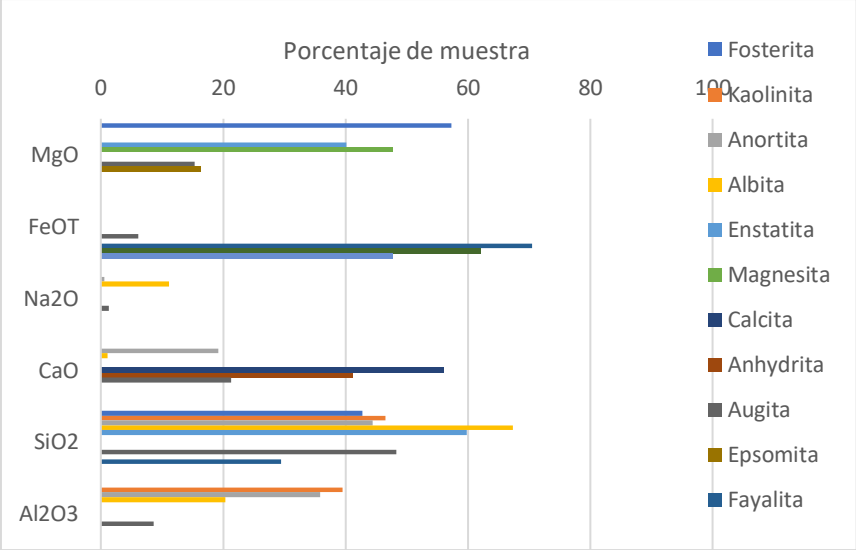

Figura 8: Concentraciones de óxidos en distintos tipos de minerales.

Las bajas concentraciones de sodio en nuestras muestras se justifican con las bajas concentraciones encontradas en minerales como anortita, del 0,5% o bien albita que llega al 10%. Nuestros espectros de Magnesio cubren la epsomita y augita, ambas con composiciones de MgO del 15%. Otros casos específicos pueden ser la anhidrita que se caracteriza por una composición del 40% de óxido de calcio o bien la siderita, con un 60% de óxido de hierro. Los contenidos de Si son alrededor del 40 para minerales ígneos como el olivino (mezcla de forsterita y fayalita) o más cercanos al 50 % en filosilicatos como la kaolinita. Los contenidos de Aluminio son de 20 % en el caso de la albita y 35% para kaolinita y anortita, que forma parte del rango cubierto por nuestro set de muestras. El rango abarcado por las muestras para el óxido de aluminio cubren el rango esperado de las rocas del cráter, 20 % en el caso de la albita y 35% para kaolinita y anorita.

En ciertos casos podemos ver que dos minerales distintos poseen prácticamente la misma concentración de un compuesto que estudiamos. Esto es un elemento importante a la hora de entrenar nuestros modelos, ya que podremos evitar imprecisiones debido a efectos de matriz.

Las líneas de emisión de ciertos minerales, a pesar de tener la misma composición de un elemento, pueden ser diferentes, ya que el elemento se encuentra en matrices diferentes. Esto se puede deber a dos motivos, uno de ellos es por motivos físicos. Debido a la textura, dureza u otras propiedades físicas de la matriz, la acción del láser para generar el plasma puede cambiar, de modo que las líneas de emisión serán diferentes. El otro motivo es el químico, ya que debido a la presencia de ciertos elementos en la matriz pueden alterar la emisión espectral del elemento que nos interesa estudiar. Esto se puede producir por ejemplo por absorciones de líneas espectrales por otros compuestos de la matriz, entre muchos otros (Bret C. Windom and David W. Hahn, 2009).

Entrenando nuestros modelos utilizando matrices diferentes logramos que aprendan a discernir entre datos que son útiles, y aquellos que pueden cambiar debido a estos efectos.

Aparte de ciertos casos interesantes, los rangos de concentraciones en las muestras actúan en conjunto para poder identificar todos los distintos tipos de minerales que encontraríamos en el cráter, que se componen de mezclas de los distintos óxidos que estudiamos.

3.3. Definición de los diferentes sets de datos por elemento

Una vez tenemos nuestro set de datos normalizado de tres formas distintas, seleccionamos los espectros que tengan una concentración no nula del elemento que queramos entrenar, ya que los

minerales de la base de datos han sido caracterizados por diversos equipos. Además se ha utilizado una selección uniforme con la seguida por el equipo de ciencia de SuperCam, eliminando algunas muestra que son problemáticas. Como ya se ha explicado, dejamos fuera un set de muestras para probar los modelos. Sobre este nuevo set de datos de entrenamiento, haremos PCA en cada normalización y en cada elemento, obteniendo una matriz de transformación de intensidades a componentes principales en cada caso, así como las componentes principales del set de entrenamiento. Adicionalmente, entrenaremos ciertos modelos en los que sobre nuestro conjunto de PCs utilizaremos un algoritmo de selección de facetas, MRMR. A fin de prevenir sobreentrenamientos en los modelos, esto es que el algoritmo se especialice en identificar los espectros del set de entrenamiento pero no sea útil en espectros nuevos, utilizamos la validación k-folding durante el entrenamiento.

K-Folding

Esta validación consiste en dividir el set de entrenamiento en un número k de folds, subconjuntos de los datos escogidos de forma aleatoria. El entrenamiento se realiza en varias iteraciones, en cada una de ellas el entrenamiento se realiza con todos los folds excepto uno, que se usa como validación del entrenamiento. El algoritmo realiza estas iteraciones hasta que todos los folds han sido utilizados de validación, al validarse un mayor número de veces y con conjuntos de datos diferentes conseguimos mejorar la valoración que le asignamos a nuestro modelo.

Figuras de mérito

Para conocer qué combinaciones de Modelo, Normalización y selección de facetas nos da mejores resultados nos guiaremos con el error cuadrático medio medido al aplicar los modelos sobre el conjunto de validación (ECMV), pero sobre todo sobre el conjunto de Test (ECMT), ya que éste nos dará la información más similar a una medida real con nuestro modelo.

Este error cuadrático medio nos dará un indicador único, y global, sobre el desempeño de cada modelo. Pero para poder realizar un mejor análisis usaremos también la representación gráfica de los valores reales contra las predicciones. Esta representación nos permitirá ver posibles tendencias, o incluso apreciar en qué rangos de concentraciones funcionan mejor nuestros modelos.

Evaluaremos también el efecto de la distancia a la que se tomaron las medidas a la hora de aplicar nuestros modelos, para comprobar si esta variación produce efectos apreciables sobre su precisión.

4. Resultados

En primer lugar, se representan en la gráfica una comparación entre las concentraciones utilizadas para entrenar los modelos y las concentraciones utilizadas para la fase de Test.

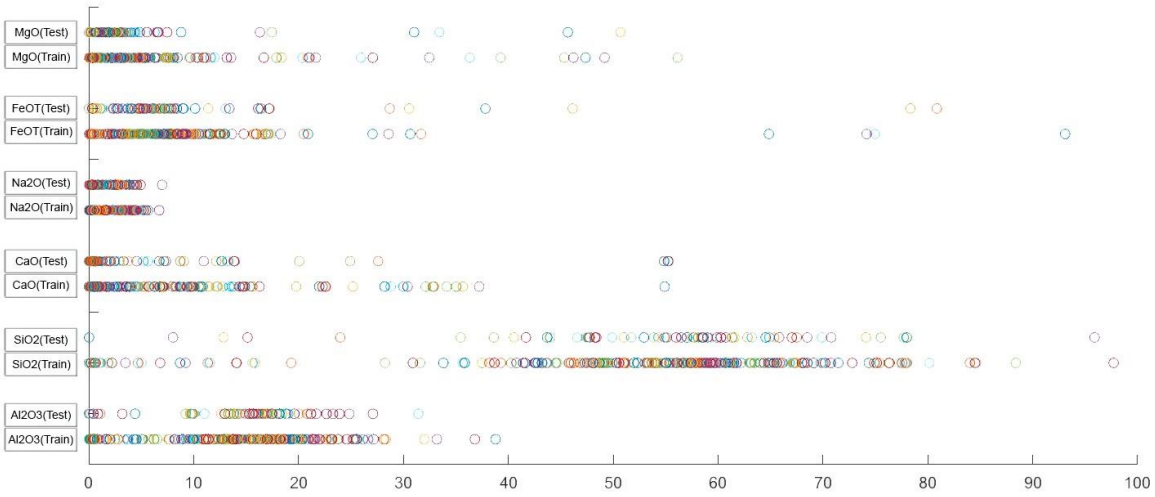

Figura 9

Como nos podemos fijar las concentraciones se distribuyen de manera similar entre ambos conjuntos, los rangos de concentraciones que tienen mayor densidad de espectros en la fase de test tienen atribuidos también una gran concentración de puntos para el set de entrenamiento. Esto nos permite afirmar que los modelos han sido entrenados correctamente para la prueba que realizaremos para saber lo fiables que son. Se busca también evitar, en la medida de lo posible, que nuestro conjunto de test tenga concentraciones fuera del rango de concentraciones empleadas para el entrenamiento. Sería esperable, lógicamente, que los modelos no se comportasen correctamente fuera del rango de entrenamiento.

A pesar de esto, es importante añadir que en todos los casos las concentraciones más altas para cada compuesto son escasas, tanto en entrenamiento como test. Al tener pocos casos con los que entrenar los modelos en las concentraciones más elevadas de las que disponemos es más probable que los resultados sean poco precisos en estos límites. De hecho, en todos los modelos las predicciones en los límites superiores subestiman la concentración. Sin embargo, estas imprecisiones no se verán reflejadas con tanta intensidad en el ECMT, ya que tendremos pocos casos en los que realizar el test de estas concentraciones en comparación con las demás.

Adicionalmente, podemos fijarnos en ciertas discrepancias que tendrán interés más adelante. Para todos los compuestos estudiados, el rango abarcado entre el valor máximo y mínimo de concentraciones en el set de entrenamiento, contiene al rango abarcado por el set de test. Es decir, nunca tendremos en la fase de pruebas una concentración más pequeña que el mínimo entrenado ni una más alta que el máximo entrenado, con la excepción de dos casos: El calcio y el sodio, con valores de test ligeramente superiores al máximo del conjunto de entrenamiento

Tanto en el caso del Na2O como del CaO, el conjunto de prueba contiene una o varias concentraciones superiores al máximo entrenado, para visualizarlo mejor se representarán estos dos casos por separado del resto en sus respectivos apartados.

Otro punto que es de interés analizar es comprobar la eficacia de nuestros modelos con respecto a la distancia a la que se tomaron los espectros. Los modelos se entrenan con el set de calibrado tomado siempre a una distancia fija de 3 metros, pero el set de test tiene adicionalmente espectros tomados a distancias de 1.5 y 4 metros. Representaremos en cada caso las predicciones teniendo en cuenta estas distancias.

Resultados para Al2O3:

Modelo	Normalización	Feature selection	RMSE (Validation)	RMSE (Test)
Op. SVM	Norm5	Top 10 MRMR	4.7769	256.36
Op. GPR	Norm5	Top 10 MRMR	1.6639	5.4812
Kernel (SVM)	Norm5	Top 10 MRMR	5.0714	7.0473
Kernel (LSR)	Norm5	Top 10 MRMR	3.4569	7.3913
Op. Ensemble	Norm5	Top 10 MRMR	3.2808	5.3905
Op. ANN	Norm5	Top 10 MRMR	3.7358	6.4973
Op. SVM	Norm5	Top 10 PCA	3.066	3.0638
Op. GPR	Norm5	Top 10 PCA	1.0566	2.817
Kernel (SVM)	Norm5	Top 10 PCA	4.7775	5.0256
Kernel (LSR)	Norm5	Top 10 PCA	2.7983	5.4255
Op. Ensemble	Norm5	Top 10 PCA	2.1918	3.0675
Op. ANN	Norm5	Top 10 PCA	1.8915	3.5606
Op. SVM	TotalNorm	Top 10 MRMR	2.3955	3.6669
Op. GPR	TotalNorm	Top 10 MRMR	1.8121	3.1739
Kernel (SVM)	TotalNorm	Top 10 MRMR	5.6075	6.1612
Kernel (LSR)	TotalNorm	Top 10 MRMR	3.9766	6.3657
Op. Ensemble	TotalNorm	Top 10 MRMR	2.716	4.2609
Op. ANN	TotalNorm	Top 10 MRMR	2.9428	6.3777
Op. SVM	TotalNorm	Top 10 PCA	2.4919	5.3504
Op. GPR	TotalNorm	Top 10 PCA	1.3358	2.8573
Kernel (SVM)	TotalNorm	Top 10 PCA	5.1608	5.482
Kernel (LSR)	TotalNorm	Top 10 PCA	3.2796	5.5533
Op. Ensemble	TotalNorm	Top 10 PCA	2.4874	3.321
Op. ANN	TotalNorm	Top 10 PCA	2.6537	3.0342
Op. SVM	OxNorm	Top 10 MRMR	7.7757	8.8689
Op. GPR	OxNorm	Top 10 MRMR	1.4707	5.0178
Kernel (SVM)	OxNorm	Top 10 MRMR	4.9406	6.4088
Kernel (LSR)	OxNorm	Top 10 MRMR	3.4594	6.2002
Op. Ensemble	OxNorm	Top 10 MRMR	2.5334	4.004
Op. ANN	OxNorm	Top 10 MRMR	1.8892	5.8914
Op. SVM	OxNorm	Top 10 PCA	2.6679	14.344
Op. GPR	OxNorm	Top 10 PCA	1.24	5.622
Kernel (SVM)	OxNorm	Top 10 PCA	4.9062	5.6229
Kernel (LSR)	OxNorm	Top 10 PCA	3.2127	5.936
Op. Ensemble	OxNorm	Top 10 PCA	2.4226	3.6507
Op. ANN	OxNorm	Top 10 PCA	2.2791	4.1353

Tabla 1: Al2O3

El modelo que mejores resultados da tanto en la fase de validación como de test es Gaussian Process Regression, otros modelos que funcionan muy bien son también las redes neuronales, Ensemble o SVM.

Podemos ver que una norma general que se sigue en esta tabla es que si comparamos los resultados de un modelo en una cierta normalización, pero escogiendo las componentes seleccionadas por MRMR, el resultado empeora. Esto no se cumple en todos los casos, pero sí en aquellos que mejores resultados tienen en la fase de test. En cuanto a las normalizaciones, la normalización por la línea del oxígeno obtiene resultados algo peores que las otras dos, que se comportan de forma muy similar en este caso siendo la mejor Norm5.

Si nos centramos en el modelo con mejores resultados, podemos utilizarlo sobre nuestro conjunto de espectros de test y representar las concentraciones predichas frente a las concentraciones reales en esos espectros. En la figura 11 se representa primero todo el set completo, y a continuación los resultados clasificados en función de la distancia a la que se tomó la medida.

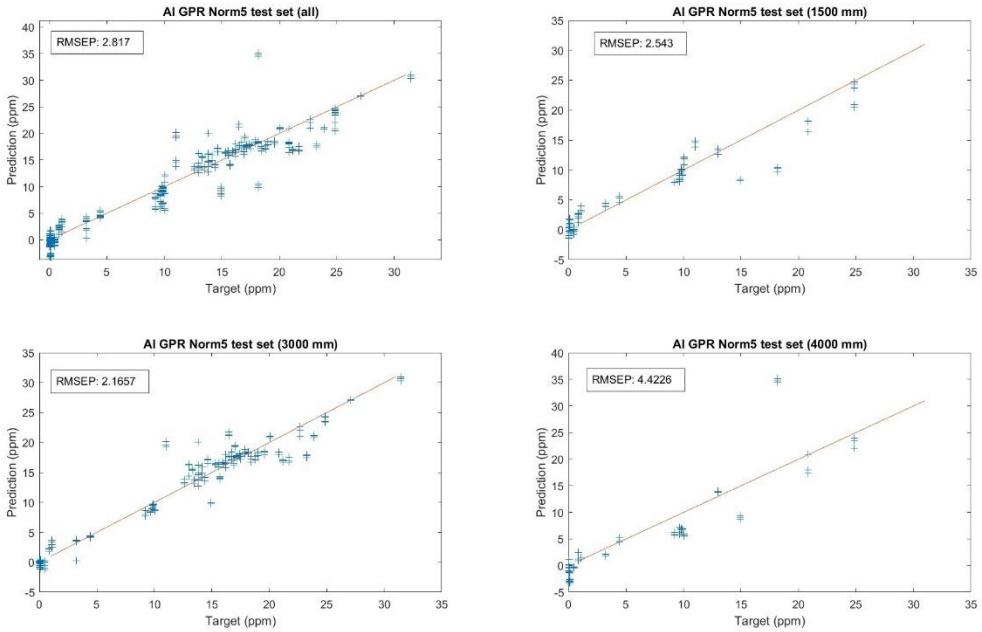

Figura 10

Las predicciones se ajustan bien a los valores reales a lo largo de todo el rango de concentraciones, con algunas desviaciones en ciertos puntos que se encuentran sobre todo en el rango intermedio de 10%-20%. Este caso funciona muy bien en los espectros de test de concentraciones máximas de algo más del 32%, algo que no sucederá con los modelos siguientes. Este hecho lo podemos justificar observando las concentraciones utilizadas para entrenar este modelo, vistas en la figura (). A diferencia de los demás compuestos, para el Al2O3 tenemos una mayor cantidad de espectros con concentraciones más altas que 32% llegando incluso a alcanzar casi un 40%.

Como es de esperar, a una distancia de 3 metros, que es a la que ha sido entrenado el modelo, el error es el menor de todos. A distancias inferiores a tres metros, el modelo funciona de una forma bastante similar, con un error algo mayor, pero sigue siendo inferior al error total del modelo. Como se puede comprobar, el problema surge al tomar medidas a distancias más altas, donde el error aumenta considerablemente.

Veremos en el resto de casos que estos efectos en la distancia son consistentes, a distancias superiores a la entrenada, el modelo reduce considerablemente su precisión.

A pesar de lo que nos indica el error, viendo las gráficas podemos darnos cuenta de que el ajuste para las distancias más altas no sería mucho peor que los demás, si no fuera por la predicción de varios puntos de alrededor del 18%, los cuales el modelo predice que tienen una concentración de un 35%. Haciendo el mismo estudio para el segundo mejor modelo, SVM, volvemos a tener el mismo problema, estos puntos conllevan un error muy grande. Sin embargo, esto no sucede en el caso de usar Ensemble:

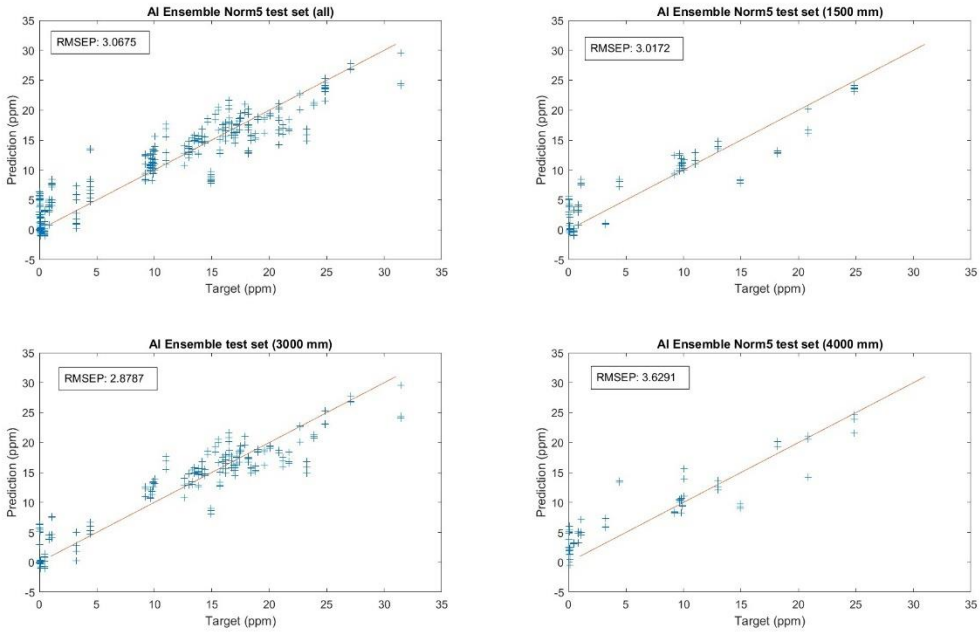

Figura 11

El error cometido es superior a los modelos anteriores, pero podemos ver que a una distancia de 4 metros este modelos se ajusta mejor a la realidad, particularmente con los espectros de 18% que estaban dando problemas. Podría ser que Ensemble es un modelo que funciona mejor para predecir medidas a distancias superiores a la de entrenamiento, aun así, el error sigue aumentando ligeramente que a distancias menores al igual que en todos los casos ya que nos hemos centrado tan solo en unos pocos casos aislados.

Algo que será de interés en casos posteriores, pero merece la pena introducir, es el fallo de la normalización por oxígeno. Ya se ha mencionado que esta normalización funciona peor que las demás, y es algo que veremos como norma general en todos los compuestos con la excepción del FeOT. Comparamos a continuación los resultados a distintas distancias, para GPR pero con Oxnorm:

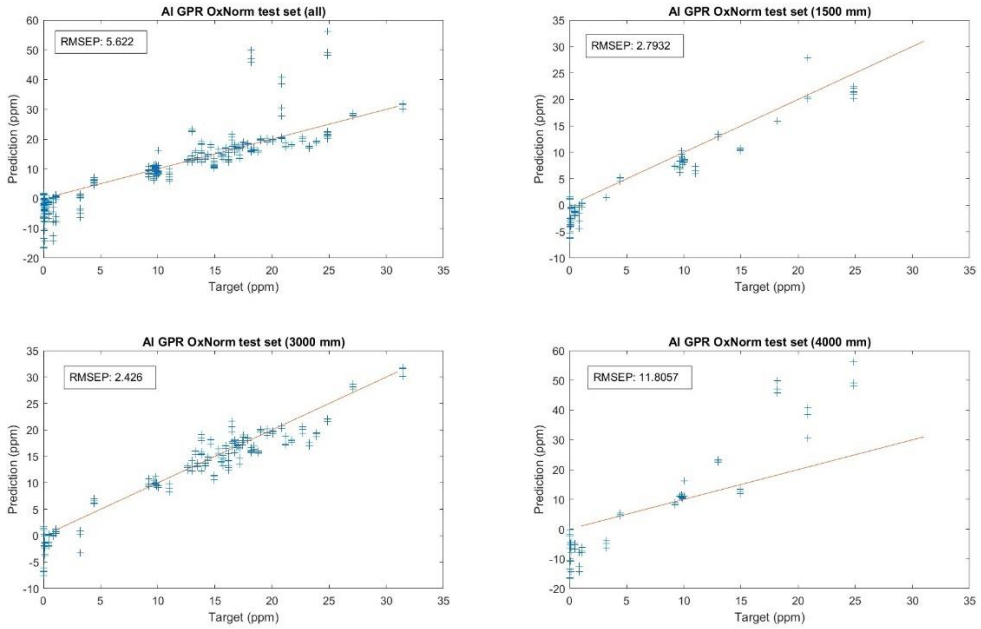

Figura 12

Podemos ver un resultado muy interesante: a pesar de ser algo peor, el modelo tiene resultados bastante similares al mejor modelo que se ha encontrado para estos óxidos, con la excepción de el set a distancias de 4 metros, donde las predicciones no se ajustan en absoluto a la realidad.

Este mal resultado tan definido a distancias más altas nos puede servir para discutir más aun la eficacia de ensemble para estos casos particulares:

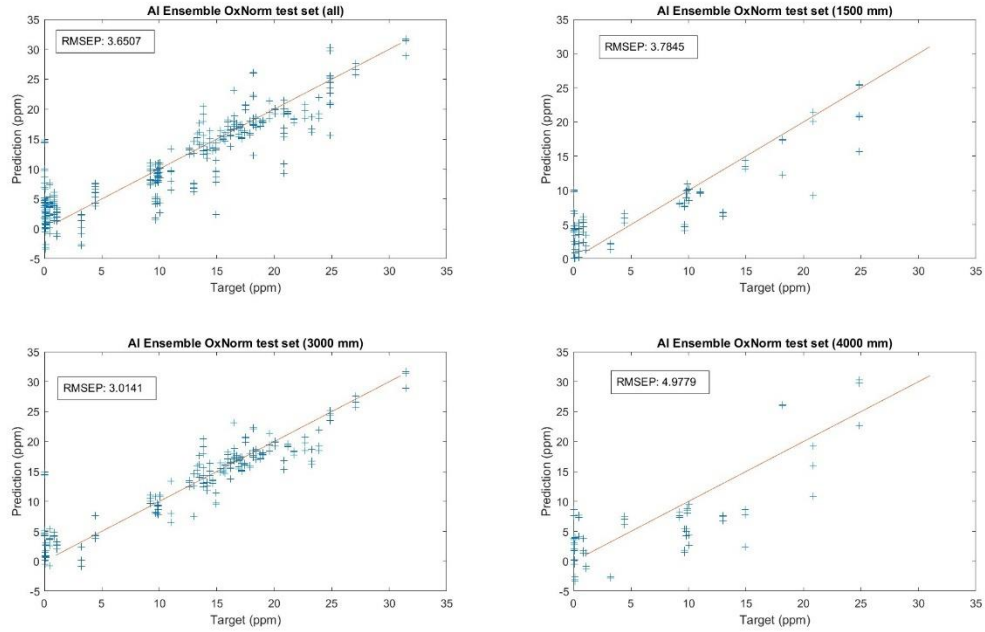

Figura 13

Utilizando Ensemble para la normalización por oxígeno parece solucionar en gran parte este problema. Obviamente el error total aumenta y también lo hacen los cometidos a distancias de 1.5 y

3 metros, pero hemos conseguido disminuir el error cometido en 4 metros hasta una cantidad más asumible, por lo menos en el contexto de las demás cifras.

Resultados para el SiO2:

Modelo	Normalización	Feature selection	RMSE (Validation)	RMSE (Test)
Op. SVM	Norm5	Top 10 PCA	6.1887	10.855
Op. GPR	Norm5	Top 10 PCA	3.654	8.3862
Kernel (SVM)	Norm5	Top 10 PCA	15.188	16.414
Kernel (LSR)	Norm5	Top 10 PCA	8.0645	12.051
Op. ANN	Norm5	Top 10 PCA	3.5074	10.043
Op. Ensemble	Norm5	Top 10 PCA	4.7302	10.506
Op. SVM	Norm5	Top 10 MRMR	13.122	13.361
Op. GPR	Norm5	Top 10 MRMR	5.3234	10.764
Kernel (SVM)	Norm5	Top 10 MRMR	15.804	17.148
Kernel (LSR)	Norm5	Top 10 MRMR	9.3068	14.036
Op. Ensemble	Norm5	Top 10 MRMR	8.7766	13.802
Op. ANN	Norm5	Top 10 MRMR	8.2996	18.632
Op. SVM	TotalNorm	Top 10 PCA	5.8859	17.045
Op. GPR	TotalNorm	Top 10 PCA	2.4665	9.2046
Kernel (SVM)	TotalNorm	Top 10 PCA	15.367	16.815
Kernel (LSR)	TotalNorm	Top 10 PCA	7.4528	14.058
Op. Ensemble	TotalNorm	Top 10 PCA	4.3539	9.8914
Op. ANN	TotalNorm	Top 10 PCA	4.6188	12.174
Op. SVM	TotalNorm	Top 10 MRMR	14.474	17.05
Op. GPR	TotalNorm	Top 10 MRMR	6.0637	15.358
Kernel (SVM)	TotalNorm	Top 10 MRMR	16.462	18.694
Kernel (LSR)	TotalNorm	Top 10 MRMR	10.481	17.662
Op. Ensemble	TotalNorm	Top 10 MRMR	9.0893	16.37
Op. ANN	TotalNorm	Top 10 MRMR	12.065	35.428
Op. SVM	OxNorm	Top 10 PCA	5.3489	26.258
Op. GPR	OxNorm	Top 10 PCA	3.2983	10.535
Kernel (SVM)	OxNorm	Top 10 PCA	14.643	17.389
Kernel (LSR)	OxNorm	Top 10 PCA	7.4161	15.741
Op. Ensemble	OxNorm	Top 10 PCA	5.5111	9.8688
Op. ANN	OxNorm	Top 10 PCA	8.463	15.152
Op. SVM	OxNorm	Top 10 MRMR	13.319	12.573
Op. GPR	OxNorm	Top 10 MRMR	5.3001	14.79
Kernel (SVM)	OxNorm	Top 10 MRMR	15.473	17.75
Kernel (LSR)	OxNorm	Top 10 MRMR	9.1219	17.673
Op. Ensemble	OxNorm	Top 10 MRMR	7.9025	18.307
Op. ANN	OxNorm	Top 10 MRMR	7.3299	27.882

Tabla 2: SiO2

Las primeras conclusiones que podemos obtener de este conjunto de resultados es que de nuevo GPR es el modelo que mejores resultados obtiene y que la mejor selección de facetas son las diez primeras PCs.
Al igual que antes, fijándonos en los mejores modelos, utilizar algoritmos de selección de facetas tiene como resultado un aumento en el ECMT. Para modelos menos precisos la precisión sí aumenta, pero no llega a reducir el error lo suficiente como para ser comparable con los mejores modelos ya que en ningún caso llega a ser menor de 10. La normalización con mejores resultados vuelve a ser Norm5, el cambio en la normalización no provoca un efecto demasiado abrupto en el error, pero lo empeora.

Si comparamos el ECMV y ECMT veremos que se produce un gran aumento, pasa de 3.654 a 8.3862. Esto significa que el modelo está sobreentrenado, al pasar a utilizarse en casos fuera de su conjunto de entrenamiento el error aumenta más del doble ya que está utilizando información que no proviene de la presencia de SiO2, sino de patrones que ha aprendido a partir de los espectros de entrenamiento.

Representamos las concentraciones obtenidas por el modelo frente a las reales:

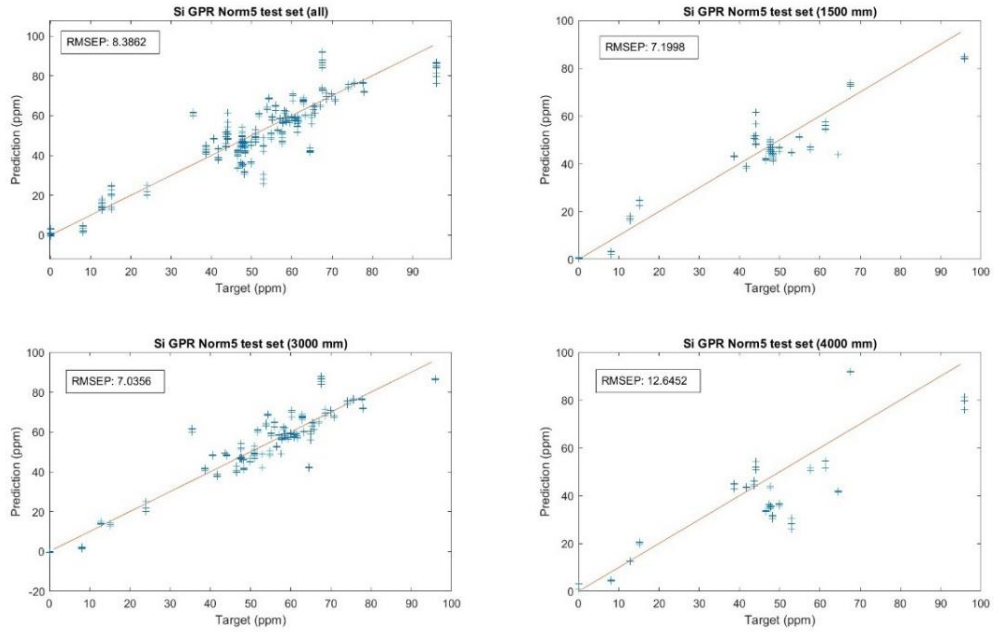

Figura 14

Los resultados tienen una dispersión esperable ya que el error es de 8.3862, este error es más alto que en el modelo anterior. Las concentraciones que abarca son mucho más amplias, desde 0 hasta el 95%, lo que puede explicar este aumento en el error.

Lo importante es que no hay una zona destacable en la que no se ajustan bien a la realidad, con la excepción de las concentraciones más altas, en las que el modelo predice siempre una concentración menor de la que debería. Como ya se mencionó anteriormente, esta discrepancia se puede explicar viendo las concentraciones utilizadas para entrenar el modelo, que son reducidas en las más altas. A pesar de esto el error que se comete en esta zona no es mucho mayor que el cometido en el resto, simplemente tiene una dirección mucho más clara.

De nuevo, al intentar aplicar el modelo a espectros a distancias de 4 metros, las predicciones fallan considerablemente. En este caso, utilizar Ensemble para intentar mejorar la predicción a 4 metros no funciona, de hecho empeora considerablemente el resultado. Otros modelos tampoco consiguen mejorar esta predicción a 4 metros, y a distancias menores tampoco consiguen superar la eficacia de GPR. La opción restante es buscar una solución en la normalización, para OxNorm, los modelos que se asemejan al anterior son GPR y Ensemble:

GPR con OxNorm es un caso muy interesante, los resultados a distancias de 1.5 y 3 metros son muy similares a los obtenidos con Norm5, llegando incluso a mejorar en la menor de las dos. Todo esto a costa de, de nuevo, unas predicciones muy malas a 4 metros.

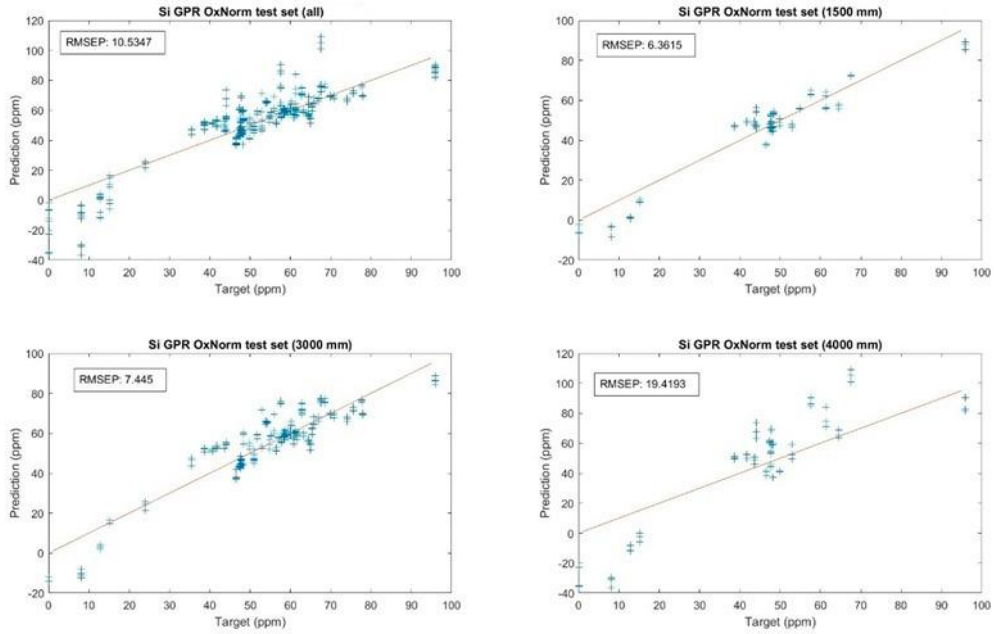

Figura 15

Si utilizamos Ensemble, solucionamos este problema, es más, las predicciones a 4 metros son algo mejores que las realizadas con GPR Norm5. No obstante, el modelo en general funciona peor a las demás distancias.

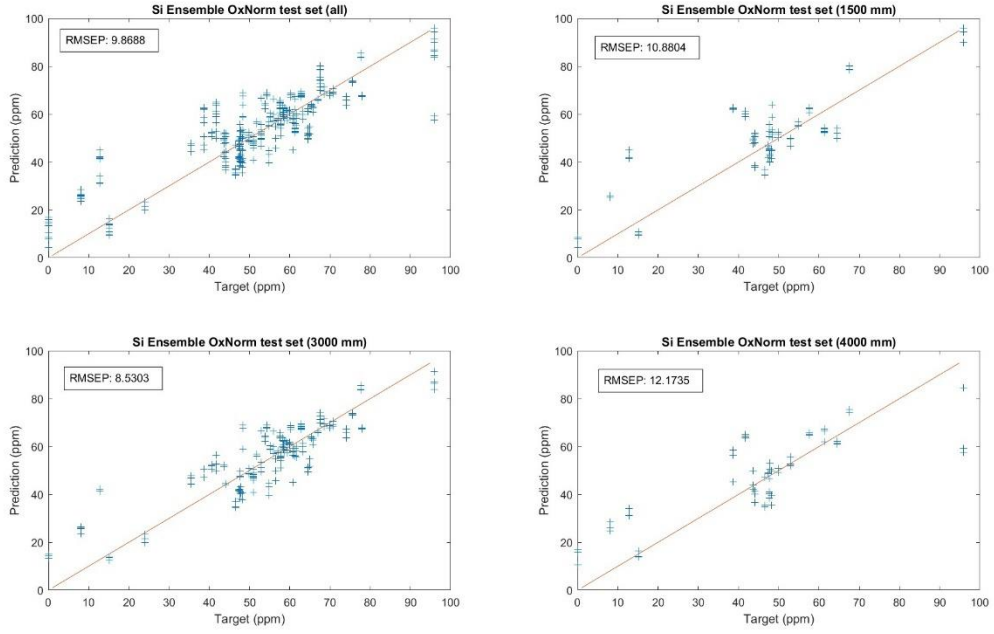

Figura 16

Resultados para el CaO:

Modelo	Normalización	Feature selection	RMSE (Validation)	RMSE (Test)
Op. SVM	Norm5	Top 10 PCA	1.6059	4.0418
Op. GPR	Norm5	Top 10 PCA	0.79535	4.4206
Kernel (SVM)	Norm5	Top 10 PCA	5.0894	12.649
Kernel (LSR)	Norm5	Top 10 PCA	2.3919	7.4133
Op. Ensemble	Norm5	Top 10 PCA	1.1342	4.3153
Op. ANN	Norm5	Top 10 PCA	1.5742	5.0082
Op. SVM	Norm5	Top 10 MRMR	3.2514	7.1861
Op. GPR	Norm5	Top 10 MRMR	0.94513	4.9836
Kernel (SVM)	Norm5	Top 10 MRMR	5.989	12.962
Kernel (LSR)	Norm5	Top 10 MRMR	2.6265	10.717
Op. Ensemble	Norm5	Top 10 MRMR	1.6678	4.7414
Op. ANN	Norm5	Top 10 MRMR	1.8679	6.7096
Op. SVM	OxNorm	Top 10 PCA	2.4307	9.237
Op. GPR	OxNorm	Top 10 PCA	0.91664	4.8218
Kernel (SVM)	OxNorm	Top 10 PCA	6.1347	13.767
Kernel (LSR)	OxNorm	Top 10 PCA	2.8014	11.081
Op. Ensemble	OxNorm	Top 10 PCA	1.4358	6.0107
Op. ANN	OxNorm	Top 10 PCA	1.5855	5.7189
Op. SVM	OxNorm	Top 10 MRMR	8.6761	14.139
Op. GPR	OxNorm	Top 10 MRMR	1.4663	12.305
Kernel (SVM)	OxNorm	Top 10 MRMR	5.9601	13.637
Kernel (LSR)	OxNorm	Top 10 MRMR	3.271	12.27
Op. Ensemble	OxNorm	Top 10 MRMR	1.9665	10.543
Op. ANN	OxNorm	Top 10 MRMR	2.5027	10.602
Op. SVM	TotalNorm	Top 10 PCA	1.7308	3.4353
Op. GPR	TotalNorm	Top 10 PCA	0.85774	5.1057
Kernel (SVM)	TotalNorm	Top 10 PCA	6.4889	12.802
Kernel (LSR)	TotalNorm	Top 10 PCA	2.8832	9.9176
Op. Ensemble	TotalNorm	Top 10 PCA	1.4545	5.4933
Op. ANN	TotalNorm	Top 10 PCA	1.622	5.9943
Op. SVM	TotalNorm	Top 10 MRMR	3.3935	15.6
Op. GPR	TotalNorm	Top 10 MRMR	1.5432	5.1127
Kernel (SVM)	TotalNorm	Top 10 MRMR	6.5555	13.287
Kernel (LSR)	TotalNorm	Top 10 MRMR	3.7733	12.151
Op. Ensemble	TotalNorm	Top 10 MRMR	2.1125	6.3512
Op. ANN	TotalNorm	Top 10 MRMR	2.8019	5.866

Tabla 3: CaO

Nos encontramos ante un caso que se diferencia del resto ligeramente, el mejor resultado en validación es GPR con Norm5, sin embargo su resultado en test no es muy bueno. El mejor resultado obtenido en test se logra con SVM en la normalización Total. La gráfica de la predicción (Figura 17) nos indica que el modelo tiene mayores problemas en predecir las concentraciones altas, esto era de esperar puesto que el máximo entrenado es del 54.9%, pero el máximo de test es del 55.38% (Figura 18).

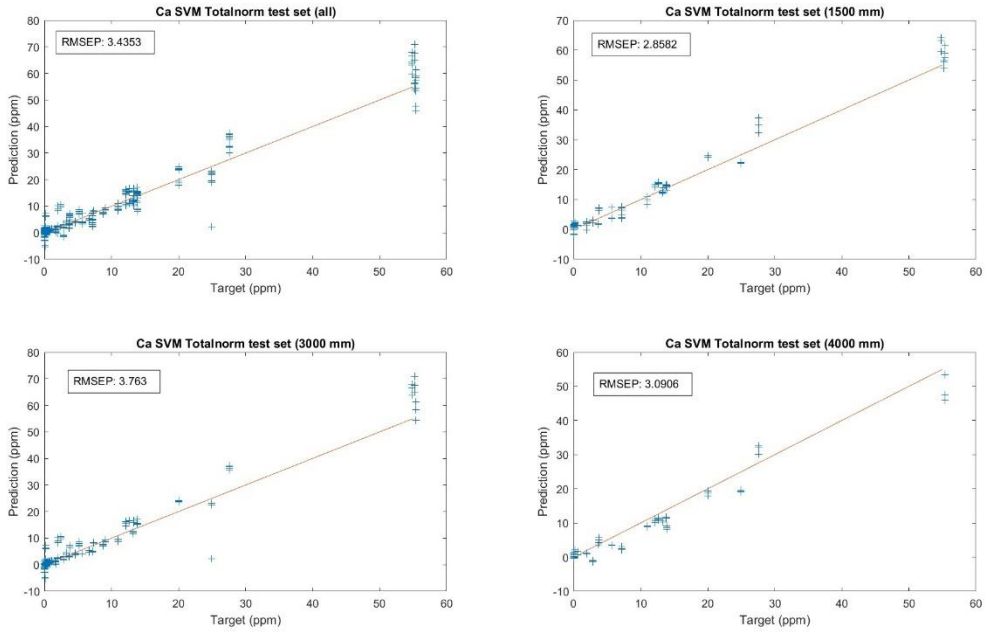

Figura 17

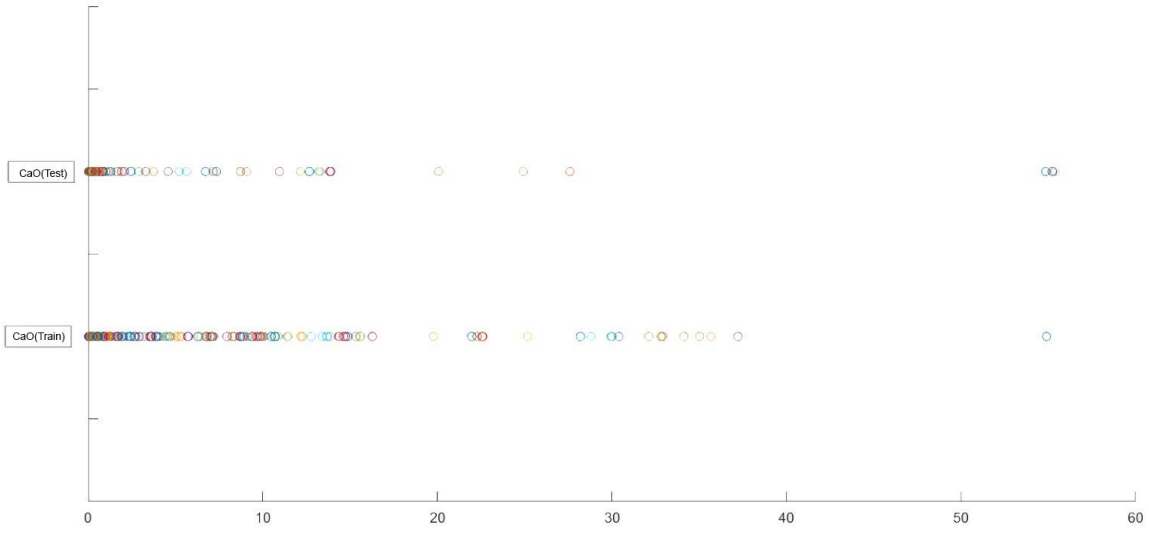

Figura 18

El efecto de cambiar la distancia provoca una reducción del error, a diferencia de los demás modelos que hemos visto. Podemos ver que la mayoría de puntos de concentraciones altas se encuentran en el set de 3 metros, las demás concentraciones ofrecen resultados decentes, aunque a valores intermedios son algo peores también debido a la escasez de puntos. También es interesante ver que este modelo no subestima las concentraciones más altas, a diferencia de lo que hemos visto en general hasta este momento. Parece ser que SVM es el único modelo con buenos resultados que no comete este error. De hecho la dispersión total de las predicciones a las concentraciones máximas parece estar centrado correctamente sobre el valor real.

Resultados para el Na2O:

Modelo	Normalización	Feature selection	RMSE (Validation)	RMSE (Test)
Op. GPR	Norm5	Top 10 PCA	0.27706	0.91049
Kernel (SVM)	Norm5	Top 10 PCA	0.54614	1.1891
Kernel (LSR)	Norm5	Top 10 PCA	0.535	1.2623
Op. Ensemble	Norm5	Top 10 PCA	0.43363	0.85711
Op. ANN	Norm5	Top 10 PCA	0.38541	0.93259
Op. SVM	Norm5	Top 10 PCA	0.38638	1.2637
Op. GPR	Norm5	Top 10 MRMR	0.44489	0.89276
Kernel (SVM)	Norm5	Top 10 MRMR	0.69391	1.4824
Kernel (LSR)	Norm5	Top 10 MRMR	0.68459	1.5043
Op. Ensemble	Norm5	Top 10 MRMR	0.5562	1.0997
Op. ANN	Norm5	Top 10 MRMR	0.68995	1.0329
Op. SVM	Norm5	Top 10 MRMR	0.80005	0.95194
Op. SVM	OxNorm	Top 10 MRMR	1.6999	1.8071
Op. GPR	OxNorm	Top 10 MRMR	0.51842	1.3084
Kernel (SVM)	OxNorm	Top 10 MRMR	0.95039	1.6361
Kernel (LSR)	OxNorm	Top 10 MRMR	0.91425	1.6698
Op. Ensemble	OxNorm	Top 10 MRMR	0.77096	1.2506
Op. ANN	OxNorm	Top 10 MRMR	0.81356	2.1153
Op. SVM	OxNorm	Top 10 PCA	0.64695	1.5704
Op. GPR	OxNorm	Top 10 PCA	0.5478	1.3248
Kernel (SVM)	OxNorm	Top 10 PCA	1.0115	1.4
Kernel (LSR)	OxNorm	Top 10 PCA	0.98336	1.4397
Op. Ensemble	OxNorm	Top 10 PCA	0.83725	1.1692
Op. ANN	OxNorm	Top 10 PCA	0.68662	2.5702
Op. SVM	TotalNorm	Top 10 MRMR	1.4095	1.7321
Op. GPR	TotalNorm	Top 10 MRMR	0.44515	1.469
Kernel (SVM)	TotalNorm	Top 10 MRMR	0.98759	1.7271
Kernel (LSR)	TotalNorm	Top 10 MRMR	0.92438	1.7622
Op. Ensemble	TotalNorm	Top 10 MRMR	0.64615	1.4747
Op. ANN	TotalNorm	Top 10 MRMR	0.93233	2.3435
Op. SVM	TotalNorm	Top 10 PCA	1.0599	2.7325
Op. GPR	TotalNorm	Top 10 PCA	0.45836	1.3864
Kernel (SVM)	TotalNorm	Top 10 PCA	0.91969	1.3149
Kernel (LSR)	TotalNorm	Top 10 PCA	0.78637	1.3334
Op. Ensemble	TotalNorm	Top 10 PCA	0.77557	1.2682
Op. ANN	TotalNorm	Top 10 PCA	0.64051	1.5978

Tabla 4:Na2O

Esta vez el mejor modelo es Ensemble, y no GPR. El mejor modelo lo podemos encontrar con Norm5 y a pesar de que las demás normalizaciones parecen funcionar igual de bien, esto es un efecto de lo pequeño que es el error. Si comparamos los promedios de los errores cometidos en cada normalización veremos que Norm5 es la mejor con diferencia: Norm5 = 1.11, OxNorm = 1.60, TotalNorm = 1.69.

Fijándonos en otros modelos, pasar a utilizar selección de facetas no tiene un efecto claro, en algunos casos el modelo mejora y en otros no. Lo consistente es que estas variaciones en general son muy reducidas debido a que el error es siempre muy bajo. Los únicos casos excepcionales los podemos encontrar en Norm5, donde las variaciones son más grandes, y además utilizar MRMR parece aumentar el error casi siempre.

Representamos las concentraciones obtenidas por el modelo frente a las reales:

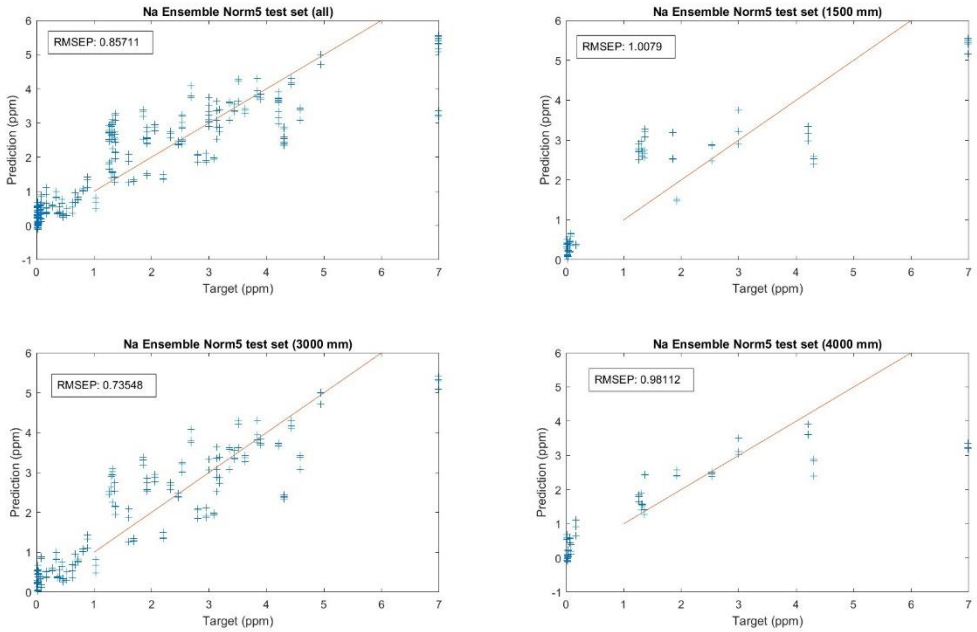

Figura 19

En este caso ocurre lo contrario que para el SiO2, al estar trabajando en un rango de concentraciones mucho más reducido (0% - 7%), el error disminuye enormemente. En la gráfica parece ser mayor pero de nuevo es porque estamos representando los puntos solamente en este rango pequeño, realmente el error cometido es el más pequeño que veremos en cualquier modelo.

Como en todos los casos, las predicciones en 7% son siempre menores de lo que deberían. Los espectros que utilizamos para entrenar los modelos de Na2O no alcanzan concentraciones tan altas. Los modelos del Na2O pasan la fase de prueba utilizando para comprobarlos espectros que incluyen una concentración del 7%, mientras que han sido entrenados utilizando como máximo espectros con una concentración del 6.72% lo que podría explicar este efecto:

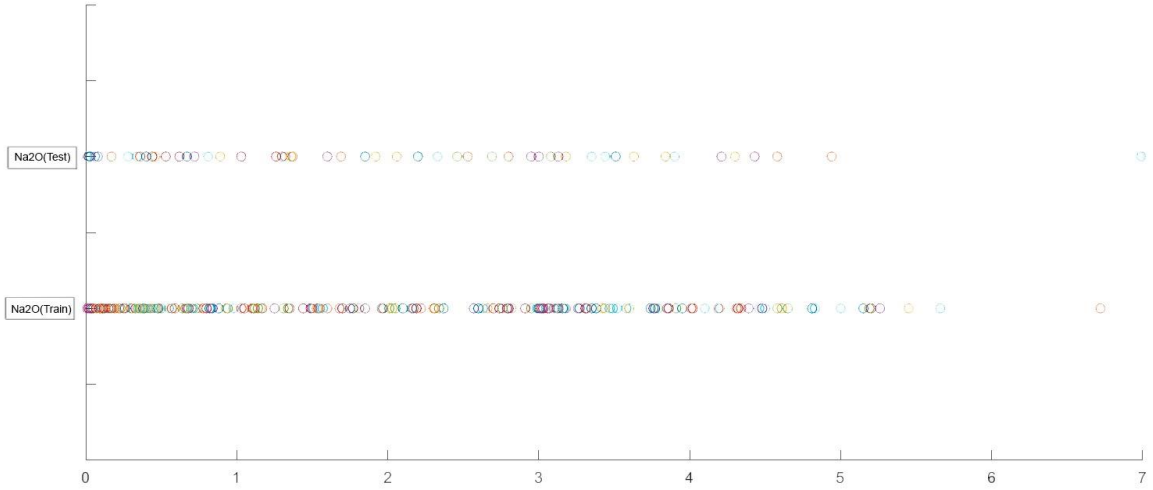

Figura 20

Este es un caso especial, cambiar la distancia a la que toma la medida no afecta prácticamente al resultado: la distancia de 3 metros es la mejor, pero nos encontramos ante un caso en el que a 4 metros los resultados son mejores que a 1.5m. Sin embargo, parece que el modelo predice peor las

concentraciones más bajas a 4 metros, mientras que obtiene bastante buenos resultados a concentraciones algo más altas. Lo contrario ocurre a distancias menores. A concentraciones altas se subestima la predicción independientemente de la distancia, y esto ocurre en general para todos los modelos.

En este caso con un rango de concentraciones reducido, cambiar el modelo o la normalización no tiene ningún efecto marcado para mejorar alguna parte de la predicción. Es importante destacar aun así que GPR no deja de ser una opción muy eficaz ya que su uso solo vio un ligero aumento en el error, su comportamiento con las distintas distancias es también prácticamente idéntico al de Ensemble.

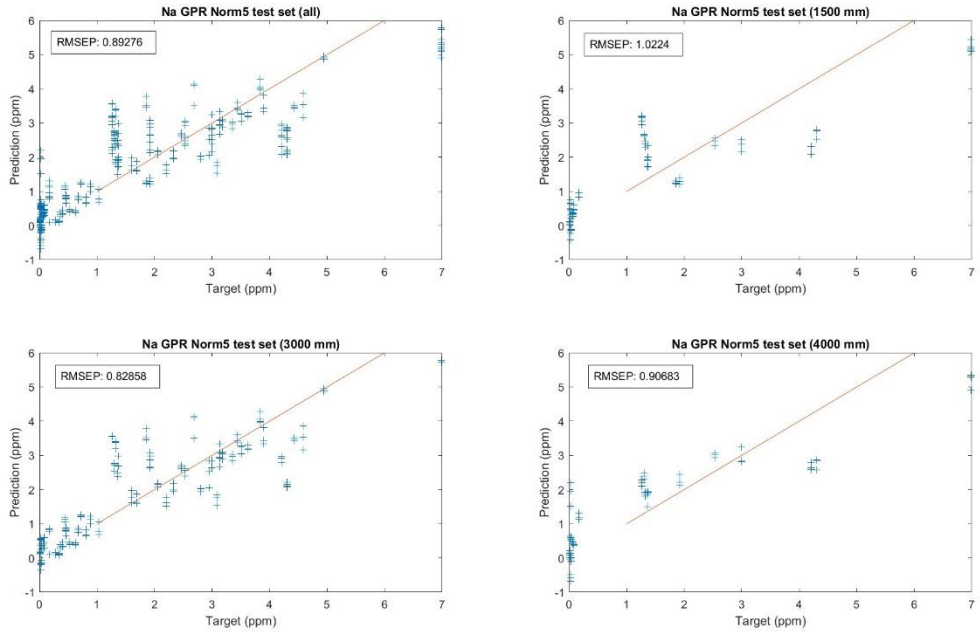

Figura 21

Resultados para el FeOT:

Modelo	Normalización	Feature selection	RMSE (Validation)	RMSE (Test)
Op. SVM	Norm5	Top 10 PCA	1.2411	6.9542
Op. GPR	Norm5	Top 10 PCA	1.0999	8.6587
Kernel (SVM)	Norm5	Top 10 PCA	9.3648	14.562
Kernel (LSR)	Norm5	Top 10 PCA	3.524	10.429
Op. Ensemble	Norm5	Top 10 PCA	2.0079	10.512
Op. ANN	Norm5	Top 10 PCA	1.9625	8.9526
Op. SVM	Norm5	Top 10 MRMR	9.0164	12.786
Op. GPR	Norm5	Top 10 MRMR	1.9816	7.7906
Kernel (SVM)	Norm5	Top 10 MRMR	9.4009	14.7
Kernel (LSR)	Norm5	Top 10 MRMR	4.8452	11.439
Op. Ensemble	Norm5	Top 10 MRMR	3.882	9.3514
Op. ANN	Norm5	Top 10 MRMR	3.7966	12.32
Op. SVM	OxNorm	Top 10 PCA	1.7376	6.7867
Op. GPR	OxNorm	Top 10 PCA	1.1286	7.2389
Kernel (SVM)	OxNorm	Top 10 PCA	9.6133	14.949
Kernel (LSR)	OxNorm	Top 10 PCA	5.8451	12.002
Op. Ensemble	OxNorm	Top 10 PCA	1.7712	10.85
Op. ANN	OxNorm	Top 10 PCA	1.311	4.6547
Op. SVM	OxNorm	Top 10 MRMR	6.8894	10.195
Op. GPR	OxNorm	Top 10 MRMR	1.6787	9.2667
Kernel (SVM)	OxNorm	Top 10 MRMR	9.5825	14.862
Kernel (LSR)	OxNorm	Top 10 MRMR	6.1325	14.293
Op. Ensemble	OxNorm	Top 10 MRMR	2.1027	11.159
Op. ANN	OxNorm	Top 10 MRMR	1.7697	8.7591
Op. SVM	TotalNorm	Top 10 PCA	4.1813	6.8046

Op. GPR	TotalNorm	Top 10 PCA	2.1263	5.9394
Kernel (SVM)	TotalNorm	Top 10 PCA	8.9138	14.059
Kernel (LSR)	TotalNorm	Top 10 PCA	4.2524	9.4846
Op. Ensemble	TotalNorm	Top 10 PCA	2.921	7.9151
Op. ANN	TotalNorm	Top 10 PCA	2.3534	7.0756
Op. SVM	TotalNorm	Top 10 MRMR	6.4204	23.476
Op. GPR	TotalNorm	Top 10 MRMR	2.7118	8.9896
Kernel (SVM)	TotalNorm	Top 10 MRMR	9.2829	14.14
Kernel (LSR)	TotalNorm	Top 10 MRMR	4.4038	10.218
Op. Ensemble	TotalNorm	Top 10 MRMR	3.0805	9.3245
Op. ANN	TotalNorm	Top 10 MRMR	3.0263	9.8301

Tabla 5: FeOT

Este es el primer y único caso en el que Norm5 no es la mejor normalización, de hecho, es por norma general, la normalización por oxígeno es la que peor funciona de las tres. En casi todos los casos, MRMR empeora el resultado. Hay algunos casos en los que mejora, pero su error ya de por si es bastante alto, en los casos de interés, el efecto es que el error aumenta drásticamente. El modelo más preciso en la fase de test es la red neuronal, y no GPR. Este caso especial nos puede aportar mucha información respecto a cómo se comportan los modelos con respecto a las normalizaciones. Norm5 obtiene resultados muy malos en todos los casos, pero podemos fijarnos en los resultados de TotalNorm. Si nos restringimos a esta normalización, el mejor modelo vuelve a ser GPR y en Norm5 es también mejor que las redes neuronales, pero no llega a ser el mejor. Considerando todo esto, puede ser que en OxNorm el modelo GPR no es tan eficaz como las redes neuronales, cosa que no sucede en las demás normalizaciones. Esto podría explicar por qué es mejor utilizar las redes neuronales, ya que para el FeOT las normalizaciones en las que GPR es por lo general el mejor modelo no funcionan bien, lo que nos restringe a utilizar OxNorm. De hecho, el segundo mejor modelo es justamente GPR pero utilizando TotalNorm.

Representamos las concentraciones obtenidas por el modelo frente a las reales:

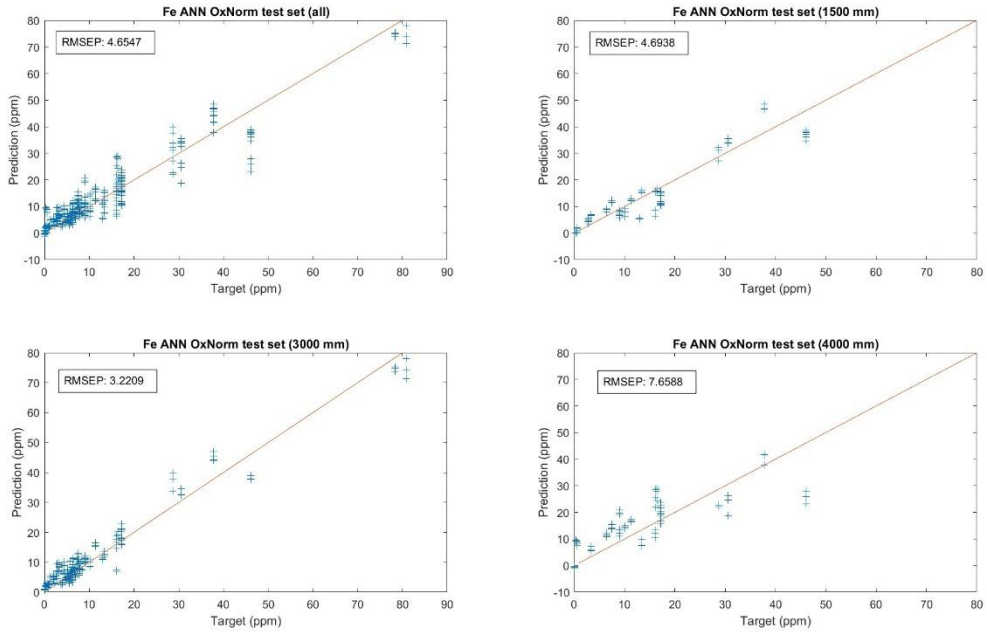

Figura 22

Al igual que en todos los casos el modelo subestima los valores más altos con un error no mayor que en el resto de concentraciones. Para los casos con menores concentraciones el modelo funciona muy bien, y a medida que nos alejamos, a partir del 15% empeora considerablemente. Esto es algo de

esperar, si nos fijamos en la figura 9, los espectros de entrenamiento utilizados para este caso están casi todos localizados justamente en el rango de 0 a 15%. Esto hace que el modelo funcione mucho mejor en esta franja de concentraciones, que no es una desventaja ya que los espectros de test también se encuentran en gran parte por esta zona.

Cabe destacar por último un efecto adverso de la distribución de los espectros de entrenamiento: Aunque los espectros de test se encuentren en la misma zona, restringirnos a ella puede provocar menor variedad y por ello sobreentrenamiento, el cual podemos ver muy claramente comparando cualquier ECMV con su ECMT correspondiente. En particular, para el mejor modelo, el error entre validación y test aumenta casi cuatro veces su valor. Esto también es parcialmente el efecto de los espectros medidos a 4 metros de distancia, que aumentan el error del conjunto de test, pero veremos que los medidos a las demás distancias igualmente tienen un error similar al ECMT que estamos discutiendo.
A pesar de esto, puede el sobreentrenamiento sea un precio aceptable, ya que el error aun así es bastante bajo, considerando el rango de valores que estamos midiendo (0% - 80%). Si lo comparamos con un modelo que abarcaba un rango similar como es el caso del SiO2, vemos que ahí el ECMT es 8.38, casi el doble que en este caso.

Como podemos comprobar, el cambio en la distancia sí provoca un aumento del error. En este caso también es interesante destacar que las predicciones de concentraciones altas no se desvían mucho de la realidad en comparación con otros modelos que hemos visto, podemos comprobar que es principalmente debido a que los espectros de test de estas concentraciones fueron todos medidos a la distancia entrenada, 3 metros. Por ello la predicción es buena.

Manteniendo la misma normalización, utilizar otros modelos tiene resultados muy pobres. En ningún caso logramos mejorar los resultados a 4 metros, y los demás resultados también empeoran.

Resultados para el MgO:

Modelo	Normalización	Feature selection	RMSE (Validation)	RMSE (Test)
Op. SVM	Norm5	Top 10 PCA	2.2489	3.4512
Op. GPR	Norm5	Top 10 PCA	0.71328	2.3151
Kernel (SVM)	Norm5	Top 10 PCA	6.9639	11.222
Kernel (LSR)	Norm5	Top 10 PCA	2.6935	7.216
Op. Ensemble	Norm5	Top 10 PCA	1.9223	4.068
Op. ANN	Norm5	Top 10 PCA	1.2886	2.8947
Op. SVM	Norm5	Top 10 MRMR	2.5439	11.49
Op. GPR	Norm5	Top 10 MRMR	1.3953	11.555
Kernel (SVM)	Norm5	Top 10 MRMR	8.0814	13.07
Kernel (LSR)	Norm5	Top 10 MRMR	4.2567	11.216
Op. Ensemble	Norm5	Top 10 MRMR	2.1301	13.144
Op. ANN	Norm5	Top 10 MRMR	1.9773	13.658
Op. SVM	OxNorm	Top 10 PCA	2.2529	28.848
Op. GPR	OxNorm	Top 10 PCA	0.79765	9.2596
Kernel (SVM)	OxNorm	Top 10 PCA	7.3443	12.461
Kernel (LSR)	OxNorm	Top 10 PCA	3.1645	11.008
Op. Ensemble	OxNorm	Top 10 PCA	1.863	5.2175
Op. ANN	OxNorm	Top 10 PCA	1.4698	5.8535
Op. SVM	OxNorm	Top 10 MRMR	8.7578	12.394
Op. GPR	OxNorm	Top 10 MRMR	1.4907	7.1186
Kernel (SVM)	OxNorm	Top 10 MRMR	7.6053	12.473
Kernel (LSR)	OxNorm	Top 10 MRMR	3.9475	10.84
Op. Ensemble	OxNorm	Top 10 MRMR	1.9725	7.0548
Op. ANN	OxNorm	Top 10 MRMR	2.7352	12.202
Op. SVM	TotalNorm	Top 10 PCA	1.9099	3.8494
Op. GPR	TotalNorm	Top 10 PCA	0.87761	2.8573
Kernel (SVM)	TotalNorm	Top 10 PCA	7.3887	11.725
Kernel (LSR)	TotalNorm	Top 10 PCA	3.4535	7.9048
Op. Ensemble	TotalNorm	Top 10 PCA	1.39	4.1336
Op. ANN	TotalNorm	Top 10 PCA	1.4701	4.8534

Op. SVM	TotalNorm	Top 10 MRMR	3.3872	12.872
Op. GPR	TotalNorm	Top 10 MRMR	1.2047	3.2636
Kernel (SVM)	TotalNorm	Top 10 MRMR	7.6367	11.02
Kernel (LSR)	TotalNorm	Top 10 MRMR	3.8915	7.6235
Op. Ensemble	TotalNorm	Top 10 MRMR	1.6288	4.4103
Op. ANN	TotalNorm	Top 10 MRMR	1.7636	3.3161

Tabla 6: MgO

Al igual que en otros casos, el mejor modelo vuelve a ser GPR usando Norm5 con las 10 primeras PCs. También se vuelve a repetir la tendencia general de que MRMR empeora el resultado, sobre todo en los casos con menor error.

Representamos las concentraciones obtenidas por el modelo frente a las reales:

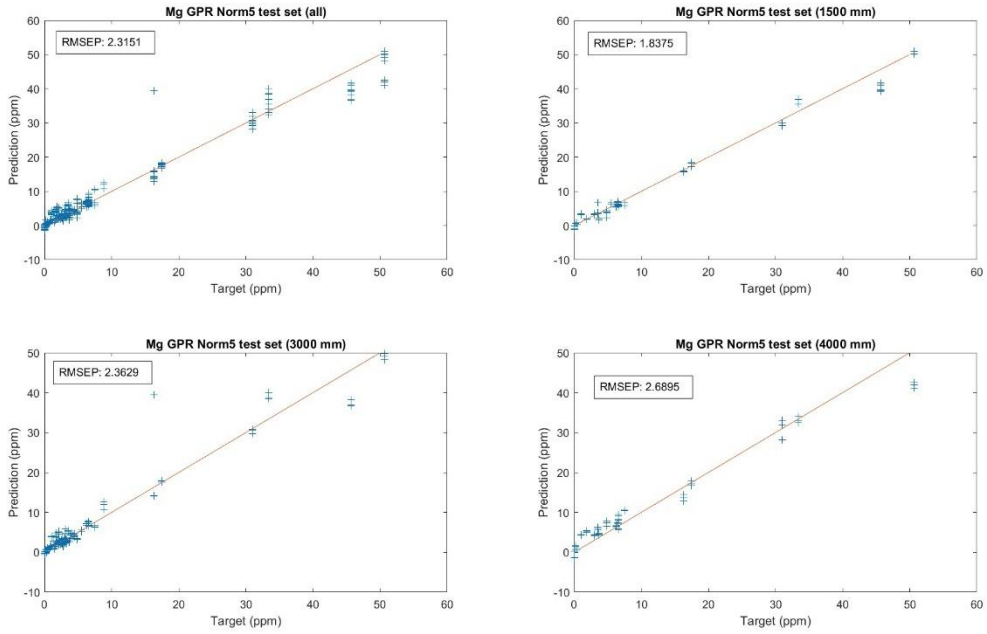

Figura 23

El modelo ofrece unos muy buenos resultados en la zona de menor concentraciones donde además encontramos la mayoría de los espectros (ver figura ()). El comportamiento del modelo es muy favorable también en el resto de concentraciones, siendo uno de los casos que mejores predicciones ofrece en el límite superior de las concentraciones que se estudian. En este caso los espectros de entrenamiento a pesar de estar concentrados en los valores bajos, también tienen una gran cantidad en las regiones más altas en comparación con los de test. Es por este motivo que este modelo funciona muy bien en todo el rango de valores.

Las variaciones en la distancia no afectan demasiado a los resultados, tenemos una buena predicción en todos los casos. Aunque el ajuste que hemos encontrado no parece tener un punto débil en este caso, los resultados del magnesio nos pueden aportar información sobre cómo se comportan los distintos modelos con las normalizaciones:

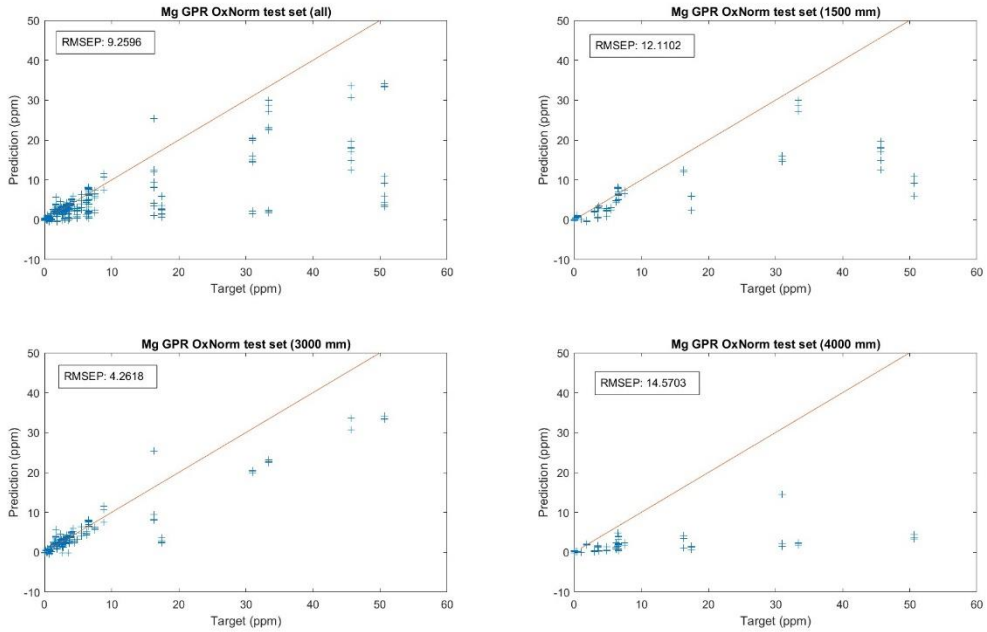

Figura 24

De nuevo, el uso de GPR con OxNorm tiene resultados desastrosos en cuanto cambiamos la distancia, esto nos sirve para ver la utilidad de otros modelos para estos casos específicos. Es importante ver que el aumento entre el ECMV y ECMT que tenemos en la tabla de datos no se debe principalmente al sobreentrenamiento, los sets de espectros a distancias diferentes de 3 metros son los culpables. Como vemos a continuación, tanto Ensemble (Figura 25) como las redes neuronales (Figura 26) obtienen unos resultados mucho mejores. A pesar de tener un resultado global algo peor, las redes neuronales obtienen un resultado bastante mejor a la distancia de 4 metros que Ensemble.

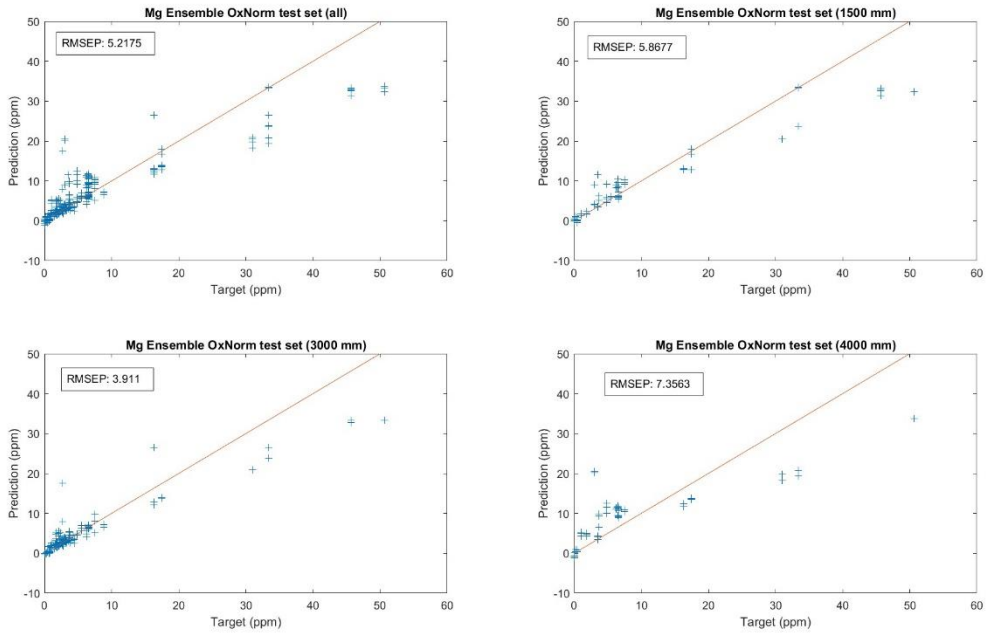

Figura 25

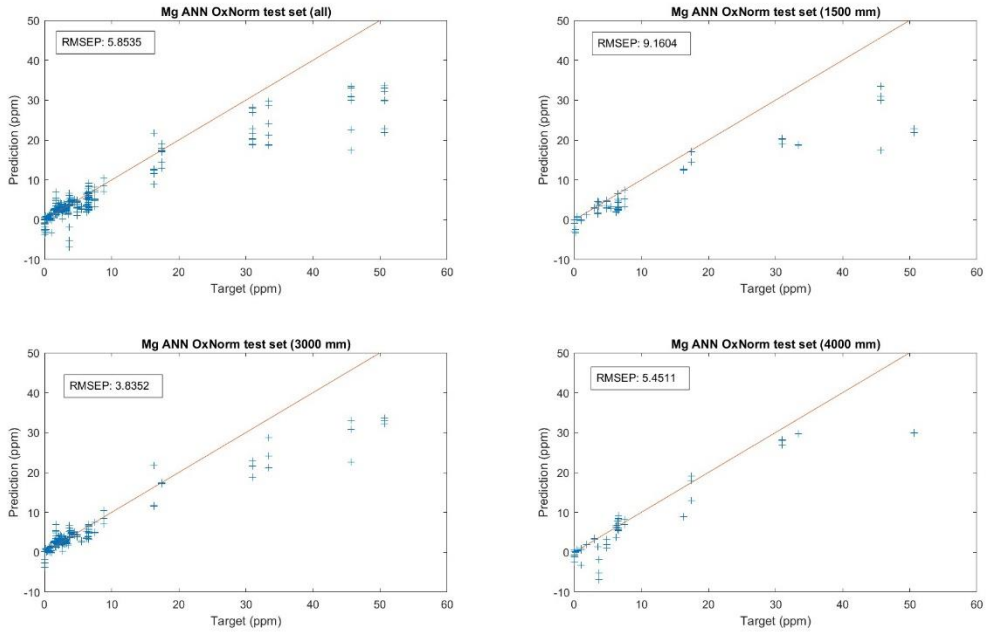

Figura 26

5. Conclusiones

Hemos visto como resultado claro que Norm5 es la normalización que mejor homogeniza los datos en la mayoría de los casos, es un resultado consistente incluso diferenciando las distancias a las que se ha tomado la medida. La diferencia entre esta normalización y la total es pequeña, aunque la primera obtiene en general mejores resultados, ya que ambas funcionan de forma similar pero al normalizar por partes estamos considerando el funcionamiento del detector, cosa que no hace Totalnorm.

Nuestros modelos tienen baja precisión para concentraciones altas, al ser entrenados aprenden de la información que les ofrecen los espectros de una concentración en concreto y también de las que la rodean. Esto significa que al acercarnos al extremo, ya no hay más datos de concentraciones superiores para entrenar el modelo, y por eso la precisión se ve reducida.

Oxnorm obtiene peores resultados por lo general ya que usa como referencia el CO2 del aire, pero debido a la presencia de óxidos en la muestra puede perder valor como elemento de comparación y normalización entre diferentes muestras. Aun con este problema, Oxnorm consiste en utilizar una referencia relativamente constante de entre todas las líneas emitidas, el oxígeno de la atmósfera. Es por esto que se ha podido ver que en algunos casos esta normalización tiene mejores resultados. Para los óxidos de hierro Oxnorm parece ser la mejor normalización con diferencia, esto puede deberse a que existe una mayor correlación entre los picos de Fe y O que para otros casos. Otro motivo es la presencia de otros elementos cuyas líneas de emisión son mucho más intensas que las del Hierro, si se da esta situación Norm5 actuará de tal forma que dará mucha mayor prioridad a estos elementos intensos, por lo que las líneas que nos interesan para entrenar nuestro modelo tendrán una intensidad menor en proporción. Oxnorm ignora este problema puesto que se fija solamente en la línea del oxígeno, y puede ser el motivo por el que funcione mucho mejor en este caso.

El modelo que mejor funciona de forma global es GPR, seguido por las redes neuronales, Ensemble o SVM. Si decidimos utilizar la normalización por oxígeno, este modelo deja de ser el óptimo para ser sustituido por los siguientes. En particular, ensemble parece funcionar bien con esta normalización

para ajustar los espectros a distancias superiores a la entrenada pero ANN también suele ser una buena opción. El modelo cambia debido a que la información con la que aprende es diferente, con Norm5 o totalnorm estamos normalizando en función del pico más alto del espectro, o de una parte de él. Esto significa que podríamos tener espectros en los que para una misma concentración de un elemento el pico característico es mas pequeño que en otros casos, ya que hay otro más intenso que antes no había. Esta diferencia puede ser lo que provoque que estos otros modelos puedan funcionar peor, ya que quedan despistados por estos cambios en la información, mientras que GPR parece ser capaz de lidiar con ello.

Sin embargo, lo que buscamos como objetivo final son modelos que funcionen bien en promedio para todas las distancias y concentraciones posibles. Es por ello que seleccionamos como mejor modelo en todos los casos aquel que obtenga un menor ECMT global. Una posible aplicación de estos casos particulares sería crear modelos que utilicen medias ponderadas de distintos modelos, o se apliquen por partes dependiendo de la distancia o concentración. Esto sería un tema mucho más amplio y complejo al que no nos podemos adentrar en este trabajo, estos modelos ya se utilizan actualmente por SuperCam. Un ejemplo es el utilizado para Al2O3, que se trata en realidad de 4 modelos diferentes trabajando en conjunto (Ryan B. Anderson et al. 2021.).

En cuanto a Feature Selection, está claro que es un algoritmo que no mejora los resultados comparado con la selección de las primeras 10 componentes principales. Al estar tratando con elementos mayoritarios con muchas líneas espectrales, tendrán bastante varianza asociada que ya se explica con las primeras PCs. Cambiar las componentes nos hace perder esta información. Además, los algoritmos que utilizamos son capaces de aprender la relación entre las cantidades de elementos, ya que si tenemos una alta concentración de un componente el resto tendrá que ser menor. Utilizar selección de facetas puede provocar que se pierda esta información, ya que estos algoritmos solo buscan relacionar las variables con la concentración de uno de los elementos. Es importante comentar también, que en una aplicación como la de este trabajo, no tenemos limites de cálculo y hemos podido usar un número elevado de variables (10 PCs). Si hubiese que limitar el número de variables a usar por los modelos, de manera que la varianza explicada por las n primeras PCs (donde n<10) se reduzca drásticamente, los algoritmos de selección de facetas podrían mejorar los resultados.

Los resultados que hemos obtenido en cuanto al ECMT son algo mayores que los obtenidos en un estudio hecho previamente, descrito en “Post-Landing Major Element Quantification Using SuperCam Laser Induced Breakdown Spectroscopy”, (Ryan B. Anderson et al. 2021.) Además las diferencias entre los errores obtenidos para distintos elementos siguen la misma tendencia.

6. Referencias

- A. Cousin et al. “SuperCam calibration targets on board the perseverance rover: Fabrication and quantitative characterization”, Space Science Reviews (2022).
- Bret C. Windom and David W. Hahn. “Laser ablation—laser induced breakdown spectroscopy (LA-LIBS): A means for overcoming matrix effects leading to improved analyte response” Journal of Analytical Atomic Spectrometry, 2009. DOI 10.1039/b913495f.
- Carey Legett et al. "Optical calibration of the SuperCam instrument body unit spectrometers," Appl. Opt. 61, 2967-2974 (2022).
- Grotzinger, J.P., Crisp, J., Vasavada, A.R. et al. Mars Science Laboratory Mission and Science Investigation. Space Sci Rev 170, 5–56 (2012).
- Hainmueller, J., & Hazlett, C. (2014). Kernel regularized least squares: Reducing misspecification bias with a flexible and interpretable machine learning approach. Political Analysis, 22(2), 143-168.

- Hilarie Sit, “Quick Start to Gaussian Process Regression” Towards Data Science, 2019.
- J.A. Manrique et al. “SuperCam Calibration Targets: Design and Development”, Space Science Reviews (2020).
- J.M. Madariaga et al. “Homogeneity assessment of the SuperCam calibration targets onboard rover perseverance” Space Science Reviews (2022).
- Lindsay I Smith “A tutorial on Principal Component Analysis,” (2002).
- Maya L. Najarian and Rosemarie C. Chinni. “Temperature and Electron Density Determination on Laser-Induced Breakdown Spectroscopy (LIBS) Plasmas: A Physical Chemistry Experiment”. Journal of Chemical Education (2013).
- Niranjan Pramanik, “Kernel Regression — with example and code” Towards Data Science, 2019.
- Roger C. Wiens et al. “The SuperCam Instrument Suite on the NASA Mars 2020 Rover: Body Unit and Combined System Tests”, Space Science Reviews (2021).
- Ryan B. Anderson et al. “Post-Landing Major Element Quantification Using SuperCam Laser Induced Breakdown Spectroscopy”, (2021).
- Samuele Mazzanti “MRMR Explained Exactly How You Wished Someone Explained to You” Towards Data Science, 2021.
- S. Maurice et al. “The SuperCam Instrument Suite on the Mars 2020 Rover: Science Objectives and Mast-Unit Description”, Space Science Reviews (2021).
- What are neural networks? (2021) En IBM. https://www.ibm.com/topics/neural-networks.
- Wu, X., Kumar, V., Ross Quinlan, J. et al. Top 10 algorithms in data mining. Knowl Inf Syst 14, 1–37 (2008).