Universidad Autónoma de Baja California

Facultad de Ingeniería, Arquitectura y Diseño

Maestría y Doctorado en Ciencias e Ingeniería

ESPECTROSCOPÍA VISIBLE-INFRARROJO CERCANO PARA EVALUAR SUELOS CONTAMINADOS POR METALES PESADOS

TESIS

que para cubrir parcialmente los requisitos necesarios para obtener el

grado de

MAESTRO EN INGENIERÍA

Presenta

Daniel Miranda Salazar

Ensenada Baja California, México. Diciembre del 2012.

UNIVERSIDAD AUTÓNOMA DE BAJA CALIFORNIA

Facultad de Ingeniería, Arquitectura y Diseño
Unidad Ensenada

ESPECTROSCOPIA VISIBLE-INFRARROJO CERCANO PARA EVALUAR SUELOS CONTAMINADOS POR METALES PESADOS

TESIS

Que para obtener el grado de maestría en ingeniería presenta:

Daniel Miranda Salazar

Aprobada por:

Dr. Horacio Luis Martínez Reyes
Director de Tesis

Dr. Miguel Enrique Martínez Rosas
Miembro del Comité

Dr. Manuel Moisés Miranda Velasco
Miembro del Comité

Ensenada Baja California, México. Diciembre del 2012.

RESUMEN de la tesis de **Daniel Miranda Salazar**, presentada como requisito parcial para la obtención del grado de MAESTRO EN INGENIERÍA del programa de Maestría y Doctorado en Ciencias e Ingeniería (MYDCI) de la UABC. Ensenada Baja California, México, Noviembre de 2012.

ESPECTROSCOPÍA VISIBLE-INFRARROJO CERCANO PARA EVALUAR SUELOS CONTAMINADOS POR METALES PESADOS

Resumen aprobado por:

Dr. Horacio Luis Martínez Reyes
Director de Tesis

En este trabajo de tesis se presenta un método basado en espectroscopía de reflectancia Visible Infrarrojo-Cercano (VIS-NIR) que permite adquirir firmas espectrales del suelo para evaluar contaminación por metales pesados (Cobalto). Los niveles de concentración del contaminante en las muestras de suelo fueron determinados sometiendo los datos o firmas espectrales a métodos estadísticos (Regresión Parcial por Mínimos Cuadrados, PLSR). También, se hace una revisión de conceptos básicos y aspectos relacionados con el estudio de suelos, fibras ópticas, espectroscopía y quimiometría.

Palabras Clave: *Espectroscopía, Reflectancia espectral, firma espectral, VIS-NIR, Modelo de predicción, PLSR, Quimiometría.*

Dedicada

A mi mamá Juana Salazar Flores y a mi papá Felix Miranda González por ser un par de personas importantes en mi vida. Ustedes que siempre han estado a mi lado y me han mostrado una manera para ser una persona completamente independiente, y sobre todo, que me han mostrado el camino para ser una persona con principios y valores. Por todo esto, les doy las Gracias.

A mis hermanos Felix, Javier y Erika que como mis mejores amigos han alegrado mi vida tantos años. Ustedes quienes rompían mis momentos de concentración con *bulling* durante tareas e investigaciones en mi formación académica tornaron mi vida en una muy divertida. Por haber compartido tantos momentos y por seguir haciendo esto, con mucho cariño les dedico este trabajo.

A todos mis amigos que con su compañía y aprecio hicieron amena esta travesía.

Agradecimiento Especial

A mi director de tesis, el Dr. Horacio Luis Martínez Reyes por brindarme principalmente su amistad. Por poner a mi disposición todo su apoyo y sus conocimientos, así como también, por depositar toda su confianza en mis decisiones y tener paciencia en mí.

A mi comité Dr. Miguel Enrique Martínez Rosas y al Dr. Manuel Moisés Miranda Velasco por su completa confianza depositada en mí, por compartir sus conocimientos, sus alientos de ánimo y sus agradables consejos.

Al Maestro Edgar Arroyo Ortega por apoyarme completamente en una etapa importante de este proyecto, así como también, por llenarme de ánimos y confianza con esa actitud bien prendida.

A mi shuper amigo Javier Villagrana Mancilla por no decir que no en esos momentos importantes para la formación personal.

A todos mis amigos y colegas: Viris Silva Rodriguez, Diego A. Bustillos Iñiguez, Francisco Villalpando, Chema Alvares Murillo, Cathy Enriquez, Alejandro Chavéz Sanchéz, Luis León Luna, Jesus Ayala, Doc Raúl Martínez, Ismael Capuchín, Fernando Ramos y Ray Buen Rostro.

Gracias a todos por ser una parte de mi vida. Me hacen sentir bien padre. :-)

Agradecimientos

Al Dr. José de Jesus Zamarripa Topete por sus insistentes críticas que fueron pieza fundamental para la redacción de la tesis en este proyecto de investigación.

Al Consejo Nacional de Ciencia y Tecnologa (CONACyT) por el apoyo económico brindado.

A la Facultad de Ingeniera, Arquitectura y Diseño (FIAD) de la Universidad Autónoma de Baja California (UABC) por brindarme una oportunidad tan valiosa.

Índice general

Índice de figuras

Índice de tablas

Capítulo 1

INTRODUCCIÓN.

La espectroscopía visible cercano–infrarrojo (Vis-NIR, por sus siglas en inglés) se ha utilizado desde hace varios años en las ciencias del suelo para la medición de carbono, nitrógeno, óxidos metálicos, capacidad de intercambio cationico, puntos de marchitéz y tamaño de partículas. La técnica ha demostrado ser simple, rápida y no destructiva con el medio ambiente [Odlare et al., 2005]. Además ha demostrado ser más exacta que los métodos convencionales. Por ejemplo, la espectroscopía VIS puede ser más exacta que la digestión de dicromato para el análisis de carbón orgánico, entre otros tipos de análisis [Viscarrarossel et al., 2006].

Recientes investigaciones han demostrado que la espectrocopía de reflectancia difusa (DRS, por sus siglas en inglés) empleando la región (Vis-NIR, 400-2500 nm) del espectro electromagnético pueden proporcionar predicciones de bajo costo sobre las propiedades físicas, químicas y biológicas del suelo [Brown et al., 2006]. Por lo que, la espectroscopia se ha convertido en una técnica ampliamente utilizada en la industria para el estudio de suelos, así como también, para la identificación y/o caracterización de polímeros, farmacéuticos, petroquímica y otras áreas. El uso de la espectroscopía también ha ayudado en la detección y monitoreo de acidéz, salinidad, materia orgánica e inorgánica del agua. Esta técnica se ha estado empleando de una manera efectiva en

la detección de practicas que degradan el suelo, y que en consecuencia, reducen la productividad del mismo. Por lo que, se le ha considerado como una excelente herramienta para la protección ambiental. Por otro lado, la espectroscopía de reflectancia difusa (DRS) también permite obtener mayores beneficios en la producción de alimentos si existe una estrecha relación entre genetistas, agrónomos, ambientalistas y procesadores de alimentos.

Para este estudio, se prepararán 56 muestras de suelo contaminadas con nitrato de cobalto hexahidratado ($Co(NO_3)_2 * 6H_2O$) a diferentes concentraciones, las cuales se analizarán espectroscopicamente con una sonda de fibra óptica *R200-Angle Reflection Probe of Ocean Optics* para la obtención de firmas espectrales. Se elaborará un modelo de predicción de concentraciones utilizando procesos estadísticos de regresión parcial por mínimos cuadrados (PLSR por sus siglas en inglés) sobre los datos espectrales. Cabe mencionar, que antes de elaborar el modelo de predicción, será necesario aplicar análisis de varianza a los datos espectrales (firmas espectrales) para la detección del intervalo optimo de longitudes de onda para eliminar ruido espectral, y con esto determinar el intervalo de barrido. También, estos datos espectrales serán sometidos a nuevos análisis estadísticos de varianza para determinar el número óptimo de barridos que ayude a obtener toda la información cuantitativa de las muestras de suelos contaminadas.

Se implementará el algoritmo de regresión PLSR en la elaboración del modelo de calibración para determinar coeficientes de regresión que ayuden a realizar predicciones futuras de niveles de concentración de elementos contaminantes ($Co(NO_3)_2 * 6H_2O$). Una ves realizado lo anterior, se espera poder realizar predicciones de concentraciones a muestras nuevas (muestras de validación) contaminadas para conocer su estado químico.

1.1. Justificación.

El estudio de suelos para aplicaciones agrícolas siempre ha sido de gran importancia debido a la gran demanda de alimentos para satisfacer las crecientes necesidades de toda la población. Por lo cual, los métodos de análisis deben de ser cada vez más precisos en la predicción, por ejemplo, en contaminantes o deficiencia de nutrientes. Los métodos químicos convencionales para el análisis de suelos son muy exactos. Sin embargo, tienen el inconveniente de requerir de un análisis de laboratorio exhaustivo después de la obtención de muestras en el campo de estudio. Por otra parte, la reflectancia espectral (DRS) ha mostrado ser una poderosa herramienta para este tipo de estudios en aplicaciones agrícolas, ya que permite conocer el estado químico y físico del suelo, ofreciendo resultados en tiempo casi real y en el sitio de interés debido a su portabilidad. Además, de que este método puede ser ajustado para ofrecer resultados de más de un atributo del suelo con un solo análisis. Utilizando espectroscopía de reflectancia en la zona visible y una pequeña parte del infrarrojo cercano (VIS-NIR)(400-1000 nm) del espectro electromagnético se puede predecir la existencia de contaminantes comunes del suelo tales como los metales pesados (Ni, Cr, Co, Cu, Pb, etc.). Esto se logra utilizando firmas espectrales de las muestras de suelo para elaborar modelos de predicción con valores cuantitativos de los atributos del suelo, en este caso, contaminantes. Dichos modelos de predicción se desarrollan basándose en algoritmos estadísticos que ayudan a obtener información de interés del suelo bajo estudio.

1.2. Planteamiento del problema

En el análisis de suelos, en algunas ocasiones es necesario determinar condiciones de fertilidad o contaminación de suelos de manera rápida y eficiente, y que no involucre tener que tomar muestras y someter estas mismas a procesos químicos de laboratorio rigurosos. Lo cual implica mucho tiempo de análisis para obtener resultados.

Uno de los problemas se presenta cuando existen proporciones de suelos con condiciones de fertilidad desconocida en una área agrícola y se quiere conocer si el cultivo se encuentra en exposición con elementos contaminantes o existe deficiencia de minerales. También, cuando se pretende utilizar un espacio de tierra para aplicaciones agrícolas o urbanas y se requiere conocer su composición química, física y biológica.

Para los estudios de suelos se pretende utilizar una tecnología diferente a la convencional. En este caso, el estudio de firmas espectrales adquiridas con fibra óptica ha ofrecido datos confiables que son capaces de caracterizar el suelo, y con ello, determinar condiciones químicas, físicas y biológicas.

1.3. Objetivos

1.3.1. Objetivo general

Determinar la capacidad de la sonda de reflexión de fibra óptica *R200-Angle Reflection Probe* para ser utilizada como sensor en la evaluación de suelos contaminados por metales pesados mediante luz. En este caso, se requiere determinar valores cuantitativos de niveles de contaminación por Cobalto(Co) en el suelo.

1.3.2. Objetivo específico

Con el fin de lograr el objetivo general, se desprenden los siguientes objetivos específicos:

- Caracterizar la sonda de fibra óptica.
- Elaborar un sistema basado en espectroscopía de reflectancia difusa (DRS) para la adquisición de firmas espectrales que ayude a evaluar cuantitativamente la contaminación por metales pesados en suelos.

- Probar y validar dicho sistema.

1.4. Sinopsis de la tesis

Esta memoria de tesis de maestría se organiza en cinco capítulos como sigue: En el primer capítulo, se presenta una introducción del tema relacionado con este trabajo en el cual se plantean: la justificación para el desarrollo de esta investigación, el planteamiento del problema y los objetivos esperados en esta investigación. En el segundo capítulo, se muestra el desglose de los temas teóricos relacionados con el trabajo desarrollado en esta investigación para cumplir los objetivos finales. En el capítulo tercero, se muestran las descripciones detalladas de las etapas del proceso de experimentación desarrolladas en el laboratorio químico, así como también, en el laboratorio óptico. En el cuarto capítulo son mostrados los datos proporcionados por la etapa de experimentación, así como el proceso de estos mismos para la interpretación de resultados finales. Por último, se presenta el quinto capítulo, en el cual, se mencionan las conclusiones finales del trabajo desarrollado, y también, se mencionan las mejoras que pudiese tener el trabajo a futuro.

Capítulo 2

ANTECEDENTES.

En este capítulo se hace una descripción de los conceptos básicos necesarios para el desarrollo del presente trabajo: suelos, fibras ópticas, espectroscopía y quimiometría.

2.1. Suelos.

El termino *suelo*, que deriva del latín *solum*, y significa piso, puede definirse como la capa superior de la Tierra que se distingue de la roca solida y en donde las plantas crecen. Con este enfoque, los suelos deben considerarse como formaciones geológicas naturales desarrolladas bajo condiciones muy diversas de clima y materiales de origen, lo cual justifica su continua evolución y, en consecuencia, su gran variedad.

2.1.1. Importancia del suelo.

Los suelos son cuerpos naturales, dinámicos y vivos que desempeñan múltiples funciones en los ecosistemas terrestres, por lo que son un componente crítico de la biósfera [Casanellas, 2008].
Entre las principales funciones del suelo cabe destacar las siguientes:

- Producción de biomasa: alimentos, forrajes, fibras, biocombustibles, masas fores-

tales.

- Mantenimiento y mejora de la calidad del agua: filtrado, almacenamiento, intercambios iónicos.
- Regulación del ciclo hidrológico: almacenamiento y transferencia de agua.
- Transformación de sustancias.
- Fijación de gases de efecto invernadero: almacenamiento de carbono.
- Regulación del microclima al absorber la radiación solar e intervenir en la evaporación.
- Hábitat biológico y reserva genética al ser un medio poroso.
- Soporte físico de actividades humanas: vivienda, industrias, infraestructuras.
- Fuente de materias primas: arcilla, grava, arena, yeso, metales, minerales, etc.
- Fuente de información geológica y geomorfológica.

2.1.2. Contaminación de suelos.

La actividad industrial, agrícola, minera y las derivadas de la vida en grandes aglomeraciones urbanas son las principales fuentes de contaminación por metales pesados y otros elementos en el medio ambiente. Los suelos son uno de los componentes del medio impactados, que actúa a la vez como reservorio y fuente de estos metales. La contaminación de suelos generalmente es determinada por comparación con la concentración de estos elementos en sitios cercanos no afectados [Ruda de Schenquer et al., 2004].

La protección y recuperación de suelos contaminados por metales pesados o sustancias tóxicas esta demandando en los últimos años un gran esfuerzo en el desarrollo de técnicas de remediación y prevención, así como la búsqueda de materiales susceptibles de ser usados como absorbentes de contaminantes específicos [Alcalá, 2007].

2.1.3. Métodos para determinar niveles de concentración de metales pesados en el suelo.

La determinación de micronutrientes (hierro, manganeso, zinc y cobre) disponibles y metales contaminantes (plomo, cadmio y níquel) en el suelo, se realiza a través de un método llamado AS-14, el cual es descrito en la Norma Oficial Mexicana NOM-021-SEMARNAT-2000 que establece las especificaciones de fertilidad, salinidad y clasificaciones de suelos, estudio y análisis. [Norma Oficial Mexicana., 2002].

Principios y aplicación.

Los procedimientos analíticos tendientes a evaluar la disponibilidad de algún metal, tal como: zinc, cobre, hierro, manganeso, plomo, cadmio o níquel; fundamentalmente se asocian a su capacidad para disolver o extraer alguna forma química del metal presente en el suelo. La eficiencia de extracción dependerá de la capacidad de cada solución para poder recuperar parte de aquellas formas de metales presentes en el suelo, las cuales generalmente se asocian a la cantidad de metal que es absorbido por los cultivos.

Entre las sustancias utilizadas para recuperar a los metales del suelo, destacan aquellas que emplean a compuestos orgánicos con la capacidad para formar complejos estables, tal es el caso del DTPA (ácido del dietilen-triamino-pentaacético) y del EDTA (ácido del etilen-diamino-tetraacético). Las soluciones complejantes, como el DTPA y el EDTA, tienen como finalidad el recuperar elementos metálicos que se encuentran en forma intercambiable, ligados a la materia orgánica y disolver formas precipitadas. [Norma Oficial Mexicana., 2002].

2.2. Fibras ópticas.

En los últimos años, los sensores de fibra óptica se han desarrollado y revolucionado a partir de las investigaciones y experimentaciones en los laboratorios, lo que ha permi-

tido desarrollar tecnologías para diversas aplicaciones prácticas. Las tecnologías basadas en fibras ópticas se pueden dividir en dos grandes categorías de sensores: intrínsecos y extrínsecos. Los sensores intrínsecos son usados en medicina, defensa, y aplicaciones aeroespaciales, y ellos pueden ser utilizados para medir temperatura, presión, humedad aceleración y deformación. Los sensores extrínsecos son utilizados en las telecomunicaciones para el monitoreo del estado y el rendimiento de las fibras ópticas dentro de una red [Yin et al., 2008].

Las fibras ópticas están constituidas por hilos conductores de luz fabricados a base de vidrio o plástico. Por cuyo interior pueden circular millones de impulsos lumínicos, que son transportadores de la información, generados por una fuente luminosa. Dicha señal es recogida en el otro extremo de la fibra, por un receptor demodulador o convertidor. La fibra óptica esta formada fundamentalmente por tres elementos:

- El núcleo, que es el lugar por donde circula físicamente la luz. El diámetro del núcleo permite que la luz sea transmitida guiando un solo modo óptico (monomodo) o permitiendo el guiado de varios modos (multimodo). El perfil del núcleo se fabrica de la forma de salto de índice o de índice gradual en las fibras multimodo.

- El revestimiento, que es una capa, también de vidrio o plástico, protege al núcleo y no permite que haya fugas de las señales lumínicas al chocar contra la pared de la interfaz núcleo-revestimiento.

- El recubrimiento, que es una funda a base de distintos materiales y protege a la fibra de cualquier agente externo que pudiera afectar la integridad del núcleo.

2.2.1. Fibra monomodo

Este tipo de fibras utiliza un núcleo con diámetro mas pequeño para reducir la dispersión y con lo cual se logra una propagación del rayo en un solo modo. Los diámetros

del núcleo de este tipo de fibras oscilan entre 5 y 10 μm, y el diámetro del revestimiento es de 125 μm. Estas fibras logran tener un ancho de banda de 50 a 100 GHz-Km, en donde esta capacidad regularmente está limitada por los equipos electrónicos [Luna, 2010].

2.2.2. Fibra multimodo

Indice escalonado

En este tipo de fibras, la luz es reflejada en múltiples trayectorias o modos. Lo que origina que la longitud de estas trayectorias sean diferentes provocando que los tiempos de desplazamiento de los rayos sean mayores o menores, dando como resultado que las señales que al mismo tiempo entran a la fibra salgan en tiempos distintos. Esto causa que el pulso óptico sufra un ensanchamiento llamado dispersión modal como resultado de los diferentes modos en la fibra [Luna, 2010].

Indice gradual

En este tipo de fibras, se pretende reducir la dispersión modal que tienen las fibras de índice escalonado. Para esto, el núcleo se fabrica con capas concéntricas de vidrio, en donde cada capa a partir del eje central del núcleo presenta un índice de refracción menor al anterior, lo que permite que el haz se refracte continuamente dando como resultado un patrón casi sinusoidal. Por lo tanto, los rayos que viajan en las capas lejanas al eje central del núcleo se desplazan a velocidades mayores que los que viajan a las capas cercanas, lo que da como resultado que todos los rayos tiendan a llegar al mismo tiempo al final de la fibra. El índice gradual reduce la disperión modal hasta 1 ns/Km o menos. Estas fibras comúnmente emplean diámetros de núcleo de 50, 62.5 o 85 μm y 125 μm para el revestimiento. La mas usada regularmente es la que tiene la relación 65.5/125 μm [Luna, 2010] [DeCusatis and DeCusatis, 2006].

2.3. Espectroscopía.

Las interacciones de la radiación con la materia son el tema de la ciencia denominada espectroscopía. Los métodos analíticos espectroscópicos se fundamentan en medir la cantidad de radiación que producen o absorben las moléculas o átomos. Los métodos espectroscópicos se pueden clasificar según la región del espectro electromagnético utilizada en un análisis. Entre las regiones del espectro utilizadas se encuentran: los rayos gamma (γ), los rayos X, la radiación ultravioleta (UV), la radiación infrarroja (IR), las microondas y radiofrecuencias (RF). También, existen técnicas espectroscópicas que ni siquiera abarcan la radiación electromagnética, como la espectroscopía acústica, de masas y de electrones [Luna, 2010].

De hecho, la espectroscopia tiene una función importante en el desarrollo de la teoría atómica. Además, de que los métodos espectroquímicos quizás se han convertido en una herramienta muy utilizada para explicar la estructura molecular, y también para la determinación cuantitativa y cualitativa de compuestos orgánicos e inorgánicos [Luna, 2010].

2.3.1. Espectro electromagnético.

El espectro electromagnético es un intervalo continuo de ondas que va desde las ondas de radio hasta los rayos gamma. En el vacío, las ondas electromagnéticas se mueven a la misma rapidez, y difieren entre si por la frecuencia. La clasificación de las ondas electromagnéticas por su frecuencia es el espectro electromagnético (Figura 2.1).

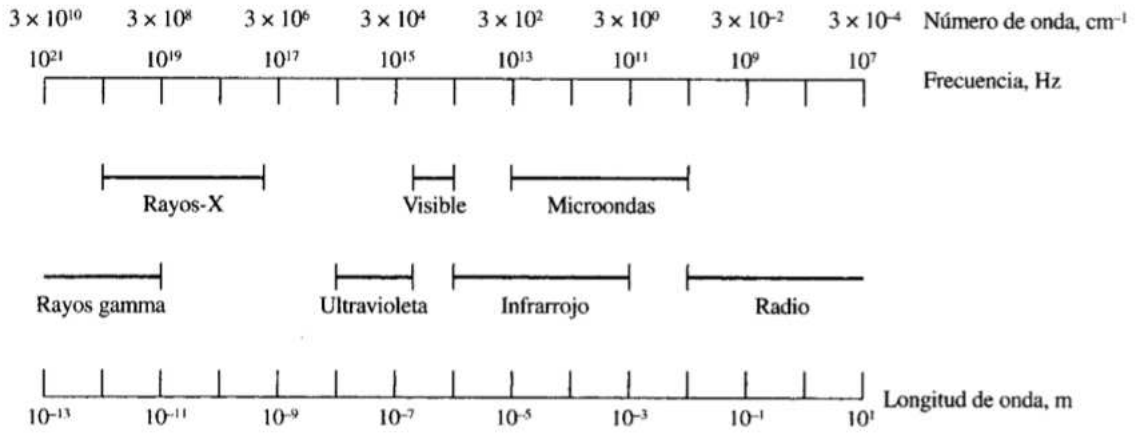


Figura 2.1: Espectro electromagnético.

Luz blanca

Cuando todas las ondas simples que componen una luz tienen la misma frecuencia, la luz se llama *monocromática*. La luz blanca es difícil de definir en el caso mas general, por lo que entenderemos por luz blanca la constituida por un espectro continuo que recoge los colores básicos de la naturaleza y en el que ninguna frecuencia predomina por su intensidad, lo que también es conocido como luz o radiación de espectro equienergético [García et al., 1999].

Infrarrojo

La región infrarroja se extiende aproximadamente de 3 x 10^{11} Hz hasta alrededor de 4 x 10^{14} Hz. El infrarrojo, o IR, a menudo se subdivide en cuatro regiones; el IR cercano, es decir, cerca del visible (780–3000 nm), el IR intermedio (3000–6000 nm), el IR lejano (6000–15,000) y el IR extremo (15,000 nm–1 mm).

Cualquier material es capaz de absorber e irradiar IR por agitación térmica de sus moléculas constitutivas. Además del espectro continuo emitido por gases, líquidos y sólidos; las moléculas aisladas termicamente excitadas pueden emitir IR en rangos angostos

específicos. Debido a las vibraciones y rotaciones de estas moléculas, las emisiones son características de los enlaces químicos involucrados [García et al., 1999].

2.3.2. Absorbancia.

La absorbancia espectral es una manera de medir la cantidad de luz que es capaz de absorber una muestra. Para la mayoría de las muestras, la absorbancia relaciona linealmente la concentración de un elemento o analito de una sustancia. La absorbancia (A_λ) se calcula utilizando la ecuación 2.1.

$$A_\lambda = -log_{10}(\frac{S_\lambda - D_\lambda}{R_\lambda - D_\lambda}) \tag{2.1}$$

Donde:

- S_λ = Intensidad del espectro de la muestra en la longitud de onda λ.
- D_λ = Intensidad del espectro de la referencia oscura en la longitud de onda λ.
- R_λ = Intensidad del espectro de la referencia en la longitud de onda λ.

La absorbancia también es proporcional a la concentración de una sustancia al interactuar con la luz (conocido como *Ley de Beer*). Las aplicaciones comunes de absorción incluyen la cuantificación de concentraciones químicas en muestras acuosas o gaseosas [Ocean Optics Inc., 2007].

2.3.3. Transmitancia.

La transmitancia espectral es utilizada para determinar el porcentaje de energía que atraviesa una muestra relativa a la cantidad de la luz que atraviesa una muestra de referencia. La transmitancia puede también proporcionar la cantidad de luz reflejada de una muestra. Por lo que, la transmitancia y reflectancia utilizan los mismos cálculos matemáticos. Comúnmente se suele expresar la transmisión como un porcentaje

($\%T_\lambda$) relativo a una sustancia estándar (tal como el aire). Una manera de calcular la transmitancia y la reflectancia se logra con la ecuación 2.2 [Ocean Optics Inc., 2007].

$$\%T_\lambda = \frac{S_\lambda - D_\lambda}{R_\lambda - D_\lambda} * 100\,\% \tag{2.2}$$

Donde:

- S_λ = Intensidad del espectro de la muestra en la longitud de onda λ.
- D_λ = Intensidad del espectro de la referencia oscura en la longitud de onda λ.
- R_λ = Intensidad del espectro de la referencia en la longitud de onda λ.

2.3.4. Reflectancia.

La reflexión es el retorno de la radiación por una superficie, sin que exista un cambio en la longitud de onda. La reflexión puede ser:

- *Especular.*

En donde el ángulo de incidencia es igual al ángulo de reflexión, como se ilustra en la figura 2.2.

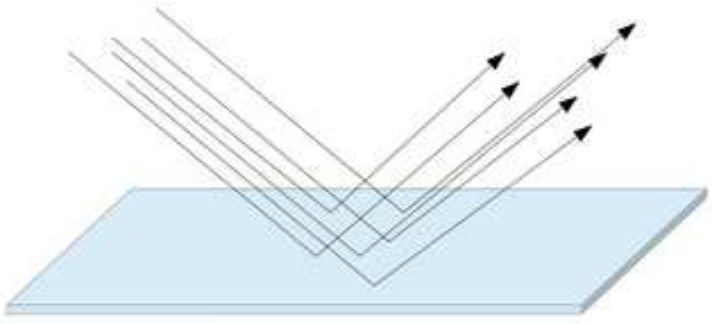


Figura 2.2: Reflexión especular.

- *Difusa.*

El ángulo de incidencia no es igual al ángulo de reflexión, como se ilustra en la figura 2.3.

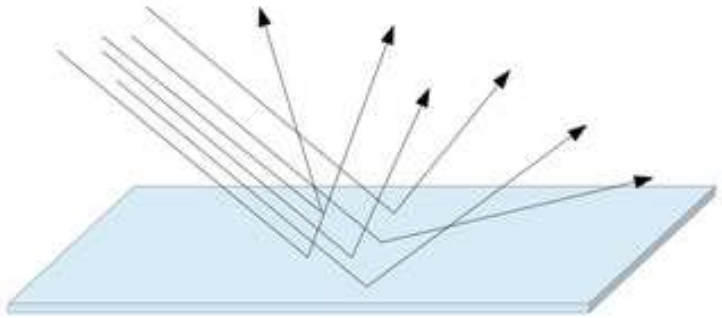


Figura 2.3: Reflexión difusa.

Cada superficie presenta ambas reflexiones difusa y especular. Algunas superficies pueden retornar más reflexión especular, mientras que otras reflejan más difusamente. La reflexión especular incrementa proporcionalmente con la cantidad de brillo en una superficie.

La reflexión se suele representar como un porcentaje ($\%R_\lambda$) relativa a la reflexión de una superficie estándar de referencia, como se indica en la ecuación 2.3.

$$\%R_\lambda = \frac{S_\lambda - D_\lambda}{R_\lambda - D_\lambda} * 100\,\% \tag{2.3}$$

Las aplicaciones mas comunes de reflexión incluyen medidas de propiedades de superficies sólidas para determinar condiciones físicas, químicas y hasta incluso biológicas. Como por ejemplo, determinar propiedades visuales en pinturas, plásticos, y productos alimenticios [Ocean Optics Inc., 2007].

2.3.5. Espectroscopía de reflectancia difusa (DRS, por sus siglas en inglés).

Este es una técnica de detección rápida y no destructiva que ha demostrado tener la habilidad para identificar y cuantificar propiedades del suelo en sitio. Esta es considerada como una próxima tecnología de detección en el estudio de suelos, ya que se adecua de manera perfecta para el estudio rápido de los mismos debido a que se ha estado usando con equipo portatil, lo que permite realizar análisis en sitio para cuantificar suelo orgánico, carbón inorgánico, y contenido de arcilla. A la fecha, pocos estudios han reportado sobre el uso de VisNIR DRS para caracterizar suelos contaminados [Chakraborty et al., 2010].

2.3.6. Firma espectral.

Las firmas espectrales de los materiales se definen por su reflectancia, transmitancia o absorbancia como un conjunto de datos o gráficas que proporcionan una relativa intensidad de radiación como una función de la longitud de onda. Bajo condiciones controladas, las firmas espectrales se derivan de las transiciones electrónicas de los átomos y vibraciones de estiramiento y flexión del grupo estructural de los átomos que forman las moléculas y cristales. Las vibraciones fundamentales de la mayoría de los suelos pueden ser encontradas en la región del medio-infrarrojo, con sobretonos y combinaciones encontradas en la región del infrarrojo-cercano [Brown et al., 2006].
Estas vibraciones son fundamentalmente de dos tipos:

a) *Vibraciones de tensión.*

Se producen por movimientos oscilatorios de los átomos a lo largo del eje del enlace que los une, de forma que la distancia interatómica aumenta o disminuye como se muestra en la figura 2.4, es decir, que en este tipo de enlaces se dilatan y se contraen, sin que varíen los ángulos de los mismos [Ramos and Madero, 1979].

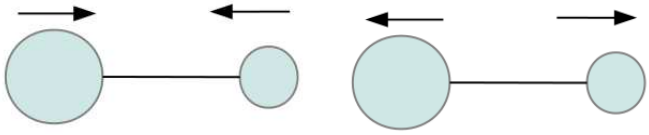


Figura 2.4: Vibración de tensión.

b) *Vibraciones de flexión.*

Se producen por movimiento de los átomos en dirección perpendicular al enlace que los une, como se muestra en la figura 2.5. Estas vibraciones se deben a deformaciones de los ángulos de enlace, sin que varíen las longitudes de los mismos [Ramos and Madero, 1979].

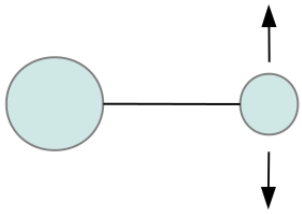


Figura 2.5: Vibración de flexión.

Una vibración de flexión puede consistir en un cambio de los ángulos entre enlaces que tienen un átomo común o también el movimiento de un grupo de átomos con respecto al resto de la molécula, sin movimientos relativos de unos átomos respecto a otros dentro del grupo.

2.4. Quimiometría.

La quimiometría es una disciplina metrológica que relaciona la química y las matemáticas, especialmente la estadística para extraer de los datos experimentales la mayor cantidad posible de información y extender el conocimiento del sistema químico [Fernández, 2005].

La quimiometría nace a principios de los años 70 del siglo pasado, con el desarrollo de la instrumentación y el gran auge de la microinformática que permite almacenar y tratar un elevado número de datos analíticos [Fernández, 2005]. Por lo que, se ha convertido en una parte muy importante de la química analítica y su uso ha ido en aumento aplicándose al tratamiento de todo tipo de datos espectroscópicos.

2.4.1. Modelos de predicción.

Actualmente los modelos multivariables de regresión son parte importante de la investigación en el estudio de suelos; ya sea para encontrar coeficientes, para realizar predicciones cuantitativas de atributos, o para generar nuevas hipótesis. El objetivo principal de estos modelos es cuantificar elementos o compuestos químicos contenidos en las muestras sometidas a estudio, en este caso, muestras de suelo. Los modelos de regresión tienen en general una estructura común que debe resultar familiar a la mayoría, generalmente con el patrón dado por la ecuación 2.4.

$$respuesta = ponderacion_1 * predictor_1 + ponderacion_2 * predictor_2 + ...ponderacion_k * predictor_k \tag{2.4}$$

La recolección de datos tiene el propósito de explicar las interrelaciones que existen entre ciertas variables o de determinar los factores que afectan a la presencia o ausencia de algún atributo en el objeto de estudio. Los modelos son un instrumento útil al suministrar una explicación matemática simplificada de dicha relación.

2.4.2. Etapas del proceso de modelado.

Para la obtención de un modelo robusto capaz de predecir, en este caso, concentraciones de nuevas muestras, se deben de seguir las siguientes etapas: selección de un

conjunto de calibración, determinación de la propiedad a cuantificar por métodos de referencia, obtención de la señal (datos espectrales), cálculo del modelo, y validación del mismo. A continuación se explica a detalle cada una de las etapas antes mencionadas para el desarrollo del modelo [Vázquez, 2004].

Selección del conjunto de calibración.

En el ámbito de la química analítica la calibración es definida como el proceso que permite establecer la relación entre la repuesta de los instrumentos y una propiedad determinada de la muestra, que suele ser la concentración [Vázquez, 2004].

Se debe seleccionar un conjunto de muestras limitada que represente la variabilidad química y física que pueda darse durante un análisis de rutina. Estas muestras son llamadas conjunto o muestras de calibración (*training set*) y deben de incorporar variabilidad de distinta naturaleza según la finalidad del modelo.

También, se debe seleccionar un método de calibración para el procesamiento de los datos y determinar los coeficientes de relación entre las señales analíticas y las propiedades a determinar en las muestras. Existen diferentes algoritmos basados en la reducción de variables para este proceso de determinación, como lo son:

- *Regresión parcial por mínimos cuadrados (PLRS, por sus siglas en inglés)*
- *Regresión por componentes principales (PCR, por sus siglas en inglés)*

Determinación por métodos de referencia.

En este etapa se utilizan métodos de referencia para determinar los valores de concentración o propiedades de las muestras. Estos métodos deben ser capaces de determinar de manera exacta y precisa los valores de concentración de las muestras de calibración, ya que de ello dependerá la exactitud del modelo de predicción calculado.

Obtención de la señales analíticas.

Es necesario obtener la señales analíticas utilizando el equipo adecuado para la extracción de los datos. Esta investigación esta relacionada con análisis espectroscópicos, por lo que se deben capturar los datos espectrales de la muestras para posteriormente realizar el tratamiento de los mismos.

Cálculo del modelo.

Para el cálculo del modelo es necesario realizar un tratamiento previo de los datos espectroscópicos y posteriormente encontrar la relación más simple entre la señal analítica y la propiedad a determinar, ya sea estableciendo la relación entre la concentración del analito (compuesto o elemento a determinar) o con parámetros físicos de la muestra. Una vez corregidos los datos, se puede llevar a cabo el desarrollo del modelo teniendo en cuenta las bases teóricas que explican la relación entre la magnitud física de la señal analítica con la propiedad a medir (reflectancia, absorbancia, etc.). Para el cálculo del modelo se utiliza una gran variedad de algoritmos basados en técnicas estadísticas para evaluar la calidad del mismo. Entre las técnicas que destacan, se encuentran las técnicas de *Regresión lineal múltiple (MLR)*, *Regresión en componentes principales (PCR)* y *Regresión parcial por mínimos cuadrados(PLSR)*, las cuales se explicaran a detalle más adelante.

Validación del modelo.

En esta etapa, se debe de utilizar un conjunto de muestras (*test set*) ajenas a las muestras de calibración las cuales deben ser sometidas al modelo desarrollado. Los valores predichos por el modelo se deben de comparar con los valores de concentración conocidos de estas muestras de validación. Los valores predichos deben ser muy similares a los determinados por el método de referencia para demostrar la robustez del modelo. Una vez validado el modelo, éste deberá ser utilizado para realizar análisis de nuevas

muestras con concentraciones desconocidas, lo cual será una segunda comprobación de la capacidad predictiva del modelo y deberá proporcionar valores de predicciones aceptables.

2.4.3. Reducción de variables

Debido a la gran cantidad de información que se maneja en los análisis espectroscópicos, es necesario utilizar técnicas que ayuden a concentrar la mayor parte de esa información en grupos mas pequeños sin que exista pérdida relevante de la información. El objetivo de esta reducción del volumen de información es poder agilizar el proceso de los datos. Uno de los métodos más utilizados para la reducción de variables es la descomposición de los datos en componentes principales (PCA, por sus siglas en inglés). Las técnicas quimiométricas más utilizadas para la elaboración de modelos de predicción se basan en este tipo de análisis PCA, las cuales son: *Regresión Parcial por mínimos cuadrados (PLSR, por sus siglas en inglés)* y *Regresión por componentes principales (PCR, por sus siglas en inglés)*.

Pre-tratamiento de los datos

Los procedimientos de reducción de variables no suelen ser aplicados a los datos originales (Matriz de datos X), por lo que son previamente tratados para eliminar posibles efectos que puedan afectar la descomposición de los datos. El tratamiento mas utilizado en el análisis de datos espectroscópicos es el centrado de los datos.

Por ejemplo, si tenemos una matriz X de datos espectroscópicos, en donde cada fila representa a cada espectro (firma espectral) y cada columna a cada longitud de onda (variable), entonces, el centrado convierte a la matriz X de la siguiente manera:

Para el centrado se calcula la media de cada columna o variable ($\bar{x}_k$) de la matriz de calibración (matriz X), y se resta este valor a cada elemento de la columna (x_{ik}) como se muestra en la ecuación 2.5.

$$x_{ik}^{centrado} = x_{ik} - \bar{x}_k \tag{2.5}$$

$x_{ik}^{centrado}$: Matriz de firmas espectrales centradas.

$\bar{x}_k$: Vector medio de la columna o variable k (longitud de onda).

x_{ik}: Elemento i (espectro) de cada columna k.

De la ecuación 2.5, el valor medio corresponde al centro del modelo, y todos los valores de la variables están referidos a dicho centro. Este método permite seguir manteniendo las unidades originales [Vázquez, 2004].

En la gráfica de la figura 2.6 se aprecia que las funciones de un grupo de datos se encuentra desplazadas con respecto del origen del sistema de ejes. A este grupo de datos se le conoce como grupo de datos originales (no centrados).

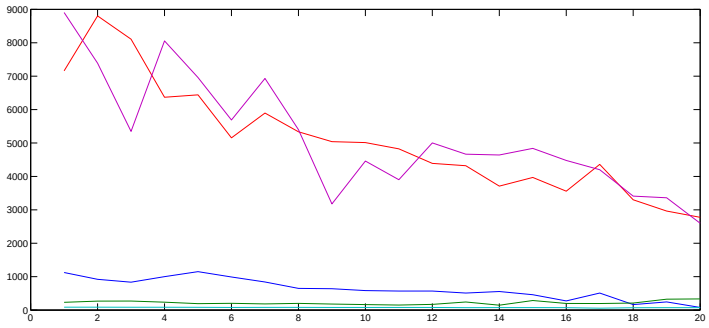


Figura 2.6: Datos originales.

En la gráfica de la figura 2.7 se aprecia el graficado del resultado de aplicar el método de centrado a los datos originales, en este sistema de ejes los datos están agrupados y referidos en un mismo punto.

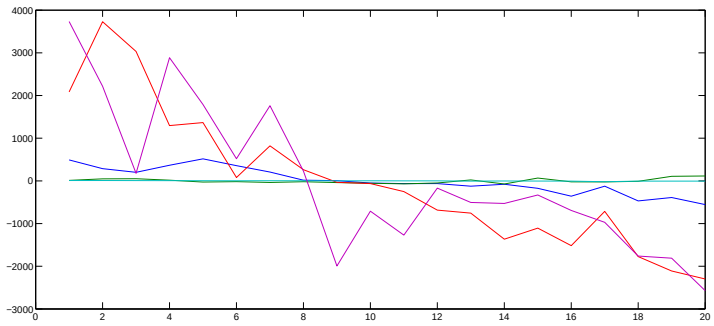


Figura 2.7: Datos originales centrados.

Análisis en componentes principales (PCA).

Análisis en componentes principales es una técnica estadística que transforma linealmente un conjunto original de variables en un conjunto sustancialmente más pequeño de variables correlacionadas. Su objetivo es reducir las dimensiones del conjunto de datos originales. Un conjunto menor de variables correlacionadas es mucho más fácil de entender y de utilizar en un posterior análisis que tener un conjunto de datos demasiado grande [Dunteman, 1992].

Las firmas espectrales de las muestras se registran con **k** longitudes de onda las cuales pueden ser representadas por un vector de **k** coeficientes. Este vector puede ayudar a construir un espacio con **k** dimensiones de tal manera que cada longitud de onda esté representado por estas dimensiones, en donde se pueden representar las muestras como un punto en este espacio. Por lo tanto, si se tienen **m** muestras las cuales se pueden representar como un punto en el espacio de **k** dimensiones. Si estas **m** muestras tienen algo en común, aparecerán agrupadas en el espacio, de lo contrario, si estas no tienen ninguna relación, estas aparecerían dispersas en este mismo espacio.

El objetivo de este análisis es encontrar las direcciones que expliquen la mayor variabilidad de las muestras y utilizarlas como nuevos ejes de coordenadas, llamados componentes principales (PC's). Con los cuales se reduce el espacio de **k** dimensiones

(**a**<**k**), manteniendo intacta la información importante de los datos. Donde **a** representa un conjunto de nuevos vectores llamados Componentes principales (PC, por sus siglas en inglés).

Estos nuevos ejes se definen matemáticamente utilizando los *Loadings*, que son los cosenos de los ángulos que forman los nuevos ejes con los originales. Las coordenadas de las muestras en estos nuevos ejes son llamados *Scores*.

Una matriz de datos **X**, la cual contiene firmas espectrales, numéricamente se descompone como se puede apreciar en la ecuación 2.6 en el producto de dos matrices: una es la matriz de *scores* **T** y la otra es una matriz de *loadings* **P**, y un residual contenido en la matriz **E** [Rosipal, 2006].

$$X = TP^t + E \tag{2.6}$$

Existen diferentes algoritmos capaces de obtener sólo los primeros PC's sin necesidad de calcular todos los vectores propios de una matriz. De otra forma, la representación completa de la matriz **X** implica **k** vectores de *loadings* y *scores*, lo cual no es necesario debido a que los primeros componentes describen el mayor porcentaje de variabilidad de los datos, en donde cada componente contiene información de diferente relevancia.

Como se mencionó anteriormente, el objetivos de PCA es la descomposición de los datos originales para la reducción de dimensiones del sistema, y como resultado de este algoritmo, la matriz de datos **X** se representa con un número menor de vectores **a** de la forma dada en la ecuación 2.7.

$$X = t_1 p_1^t + t_2 p_2^t + \ldots + t_a p_a^t + E \tag{2.7}$$

Por lo que, ahora la matriz de datos **X** queda descrita por un conjunto de vectores (**PC'**s) no correlacionados entre si en un nuevo sistema de ejes ortogonales [Vázquez, 2004].

2.4.4. Regresión parcial por mínimos cuadrados (PLSR).

Este es un método que posibilita cuantificar un compuesto o elemento en una mezcla sin necesidad de conocer los compuestos de la misma. Este método esta basado en la reducción de variables para concentrar la mayor cantidad de información en nuevas variables (componentes principales) sin tener pérdida relevante de información. Para el proceso de calibración, la regresión no se hace sobre las variables originales, si no que se realiza a estas nuevas variables debido a su dimensión, simplificando la elaboración del modelo y la interpretación de resultados.

PLSR aprovecha las propiedades de la descomposición en componentes principales (PCA), realizando una regresión múltiple inversa de la propiedad a determinar sobre los *scores* obtenidos en el PCA en lugar de realizarla en los datos originales.

Por lo tanto, si tenemos un conjunto de **P** elementos absorbentes, tendremos **P** variables $y_1, y_2, y_3, \ldots, y_p$, las cuales representan la concentración de cada elemento absorbente. El espectro o firma espectral de este conjunto registra **K** longitudes de onda que forma un conjunto de **K** variables $x_1, x_2, x_3, \ldots, x_k$ que pueden ser representadas en forma de un vector **x**. Si se cuenta con un grupo de **M** objetos (muestras), se pueden agrupar los vectores que contienen la información de cada una de ellas en dos matrices: la matriz **Y**, que contiene las concentraciones de cada elemento en cada una de las muestras, con dimensiones de **M** x **P**; y la matriz **X**, que contiene las firmas espectrales de cada muestra, de dimensiones de **M** x **K**. En donde la matriz **X** concentra la propiedad a determinar de cada una de las muestras en cada una de sus filas, mientras que las columnas contienen la información de cada una de las variables (longitudes de onda) para todas las muestras contenidas en dicha matriz **X** [Rosipal, 2006].

2.4.5. Calibración del modelo.

Durante la etapa de calibración, el algoritmo PLSR utiliza la información contenida en las dos matrices de datos: la matriz contenedora de los datos espectroscópicos (matriz **X**) dada por la ecuación 2.8 y la matriz contenedora de la información de la propiedad a determinar (matriz **Y**) dada por la ecuación 2.9, obteniéndose unas variables auxiliares llamadas variables latentes, factores o componentes PLS [Vázquez, 2004].

$$X = TP^t + E = \sum_{a=1}^{A} t_a p_a^t + E \tag{2.8}$$

$$Y = UQ^t + E = \sum_{a=1}^{A} u_a q_a^t + F \tag{2.9}$$

La ecuación 2.8 es el resultado de la descomposición de los datos espectroscópicos (matriz X), en donde **T** es la matriz de *scores* (Xscores), **P** la matriz de *loadings* (Xloadings) y la matriz **E** es la de los residuales. La ecuacion 2.9 es el resultado de la descomposición de los datos de respuesta (vector Y) a los que se ajustara el modelo, en donde **U** es la matriz de *scores* (Yscores), **Q** la matriz de *loadings* (Yloadings) y **F** la matriz de los residuales.

Las dimensiones de estas nuevas matrices estarán determinadas por **M** número de muestras, **A** factores, **K** variables (longitudes de onda) y **P** propiedades a determinar (analitos). De tal manera que T y U tendrán dimensiones (M x A), P^t (A x K) y Q^t (A x P).

La descomposición de las matrices X y Y no es independiente, sino que se realiza manteniendo una relación dada por la ecuación 2.10 entre las matrices T (Xscores) y U (Yscores). Las cuales son utilizadas para realizar la regresión y con ello determinar coeficientes de correlación (b_a) para cada uno de los factores.

$$u_a = b_a * t_a \tag{2.10}$$

En la tabla 2.1 se describe el proceso de calibración implementando PLSR con el paquete de programación Matlab. Este algoritmo es conocido como PLS1 debido a que ayuda a determinar la concentración de un solo analito o compuesto.

Tabla 2.1: Proceso de calibración.

Calibración empleando Matlab	
Load Datos	Se cargan los datos. Matrices X, Y y validación (MV)
[Xl,Yl,Xs,Ys,beta,PctVar, MSE,stats]= plsregress(X,Y,ncomp)	Se aplica la función *plsregress* a los datos **X** y **Y** para determinar las variables necesarias para elaborar el modelo con el número de componentes PLS deseado (ncomp).
Procedimiento para el calculo de los coeficientes de predicción *beta* (β)	
$b = W * (Xl^T * stat.W) * Yl$	Se calcula el vector b, el cual contiene coeficientes de regresión para cada variable λ
$b_0 = Y - X * b$ $media = mean(b_0)$	Se calcula el termino constante de los coeficientes de regresión en el vector beta (β)
$\beta = cat(1, mean(Y - X * b), b$ $\beta = cat(1, media, b)$	Se concatena b_0 con b para obtener el vector de coeficientes $\beta_0, \beta_1, \beta_2, \ldots, \beta_K$.

2.4.6. Validación del modelo.

Una vez que se obtiene el vector con los coeficientes de predicción (vector β), es posible realizar predicciones a nuevas muestras de suelo con concentraciones desconocidas para la validación del modelo. El modelo es el resultado del cálculo de los coeficientes de predicción, los cuales se emplean para realizar nuevas predicciones de concentración del analito o propiedad a determinar. A continuación se muestran algunas maneras

representativas de ecuaciones para realizar predicciones de concentración en muestras nuevas ajenas a las muestras de calibración empleando los coeficientes del vector β [Mark and Workman, 2007].

Concentración=término constante+
(Coef. de regresión 1)•(Reflectancia en long. de onda 1)+
(Coef. de regresión 2)•(Reflectancia en long. de onda 2)+...+
(Coef. de regresión K)•(Reflectancia en long. de onda K)

También, descrita en la ecuación 2.11.

$$Concentracion = \beta_0 + \beta_1 A_1 + \beta_2 A_2 + \ldots + \beta_K A_K + e \tag{2.11}$$

Y también en matriz expandida como:

$$C = \begin{bmatrix} c_1 \\ c_2 \\ c_3 \\ \vdots \\ c_N \end{bmatrix}$$

$$A = \begin{bmatrix} A_{11} & A_{12} & A_{13} & \ldots & A_{1K} \\ A_{21} & A_{22} & A_{23} & \ldots & A_{2K} \\ A_{31} & A_{32} & A_{33} & \ldots & A_{3K} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ A_{N1} & A_{N2} & A_{N3} & \ldots & A_{NK} \end{bmatrix}$$

$$\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_K \end{bmatrix}$$

$$e = \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ \vdots \\ e_N \end{bmatrix}$$

y su notación matricial reducida, se muestra en la ecuación 2.12.

$$c = a\beta + e \tag{2.12}$$

Las ecuaciones 2.11 y 2.12 muestran la manera de realizar la operaciones matemáticas para encontrar o determinar concentraciones en otras muestras utilizando los datos espectroscópicos de dichas muestras y los coeficientes de correlación obtenidos con las muestras de calibración [Mark and Workman, 2007].

En la tabla 2.2 se muestra el procedimiento para realizar la predicción de concentración a nuevas muestras (muestras de validación) empleando el programa computacional *Matlab*. Aquí se describen las herramientas utilizadas, así como también, la manipulación de las variables contenedoras de los datos espectroscópicos de las muestras de validación.

Tabla 2.2: Proceso de validación.

Validación empleando Matlab	
Load Datos	Se cargan los datos de validación. Matriz X* o MV
$MV = cat(2, ones(n,k), MV);$	Se agrega una fila de uno's al inicio de la matriz de predicción MV debido a la longitud del vector beta (K+1 longitudes de onda (variables)).
Load beta	Se carga el vector beta (β) calculado en la calibración. $\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_K \end{bmatrix}$ (coeficientes de regresión para cada variable λ)
$\hat{Y} = MV^T * \beta;$	Se aplica la siguiente multiplicación matricial para obtener el vector con las predicciones de concentración ($\hat{Y}$) de las muestras de validación (MV).

2.4.7. Evaluación de la capacidad predictiva del modelo.

El principal objetivo en la calibración de un modelo es obtener parámetros o valores que nos permitan determinar propiedades de un analito que difieran lo menos posible de los valores reales.

Este tipo de modelos de cuantización, proporcionan valores cuantitativos de los

resultados esperados. Por lo que, se debe emplear algún parámetro que permita evaluar la capacidad predictiva de dicho modelo, para esto, se debe utilizar un parámetro que permita evaluar el error medio de toda la población y no solo de una muestra. Para resolver este problema, se puede optar por utilizar la sumatoria de los cuadrados de los residuales $\sum(\overline{y}_{ij} - y_{ij})^2$, conocido regularmente como PRESS (Predicted Residual Error Sum of Squares) o su valor medio, el cual se obtiene dividiendo PRESS por el número de muestras de predicción/validación (np), conocido como MSEP (Mean Square Error of Prediction) $\sum(\overline{y}_{ij} - y_{ij})^2/n_p$. También se puede emplear la raíz cuadrada del valor medio (MSEP), conocido como RMSEP (Root Mean Square Error of Prediction) $\sqrt{\sum(\overline{y}_{ij} - y_{ij})^2/n_p}$ [Brown et al., 2006].

2.4.8. Método de selección del número de componentes principales PLS

Validación cruzada

Para llevar a cabo la construcción de un modelo de predicción comúnmente se utilizan dos grupos de muestras (calibración y validación). Si este número de muestras es relativamente pequeño es conveniente utilizar la validación cruzada (*cross validation*) para determinar el ajuste de los datos de calibración en el modelo. Esto se realiza dividiendo el conjunto de muestras de calibración en varios bloques, y el modelo se construirá tantas veces como número de bloques se haya elegido, utilizando un bloque para comprobar resultados (validar) y los restante para elaborar el modelo. De esta manera, se pueden calcular lo errores de predicción que ayudan a comprobar el ajuste de los datos.

Selección optima del número ncomPLS

El punto clave para una optima calibración en el desarrollo de un modelo predictivo basado en la reducción de variables se encuentra en la adecuada selección del número de componentes PLS (ncomPLS). Existen diferentes maneras para la reducción de este número de valores o componentes PLS que están basadas en el análisis de error de predicción al utilizar un diverso número de componentes PLS. El método mas popular para realizar el modelo es el de la validación cruzada por la manera de seleccionar el número de componentes. Este método consiste en calcular la sumatoria de los cuadrados de los residuales (PRESS), representarlos en función del número de componentes y buscar el componente que ofrece el valor PRESS más bajo. De esta forma, se consigue que al momento de incrementar el número de ncomPLS, el error disminuye hasta que se llega a un punto donde el valor del error (PRESS) empieza a aumentar debido al sobre ajuste del modelo con el cálculo de nuevos ncomPLS, los cuales solo aportan ruido.

Este popular método auxilia en la determinación del número de componentes PLS realizando una validación cruzada en la que se va calculando la sumatoria del error cuadrado de los residuales (PRESS) para un número creciente de ncomPLS.

La gráfica de la figura 2.8 muestra los resultados de los valores MSEP en función del número de componentes PLS, de donde se puede deducir que el valor óptimo para la elaboración de un modelo en cuestión es utilizando 5 ncomPLS.

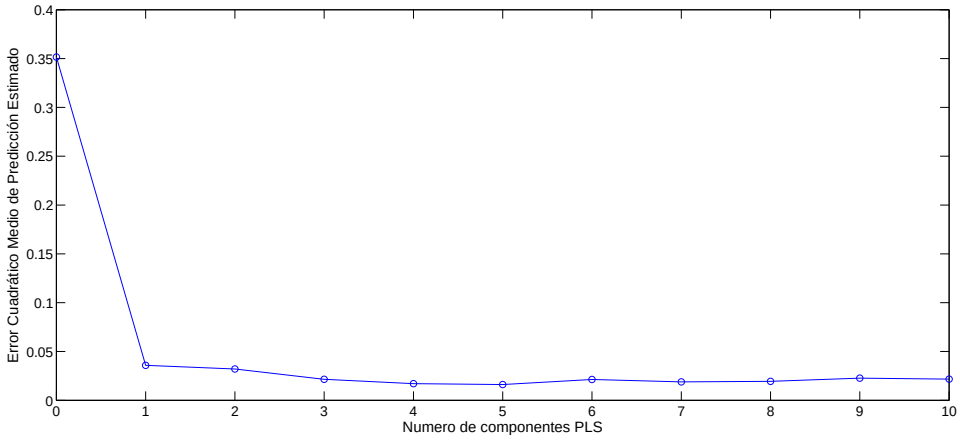


Figura 2.8: Selección del número óptimo de ncomPLS.

En la tabla 2.3 se presenta el proceso para la selección del numero de componentes PLS utilizando las herramientas de estadísticas (*Statistics Toolbox*) del programa *Matlab*. Se aprecia que después de cargar los datos de calibración se implementa la función *plsregress* para determinar la matriz MSEP con dimensiones de 2 x (ncomPLS+1) la cual contiene los valores del error medio cuadrado de predicción para los modelos con 0 a 10 ncomPLS. La primera columna de la matriz MSEP contiene los errores medios cuadrados para las variables predictivas en X, y la segunda columna contiene los errores medios cuadrados para las variables de respuesta en Y.

Tabla 2.3: Proceso de selección del número de componentes PLS.

Selección de ncomPLS empleando Matlab	
Load Datos	Se cargan los datos de calibración. Matrices X y Y.
[Xl,Yl,Xs,Ys,beta,PctVar,**MSEP**] = plsregress (X,Y,10,'cv',10);	Se aplica la función *plsregress* a los datos **X** y **Y** para determinar la variable MSEP empleando validación cruzada (cv) con 10 iteraciones para cada uno de los 10 ncomPLS.
plot(0:10,MSEP(2,:));	Se realiza el graficado de los errores MSEP como función de los componentes PLS (Figura 2.8).

El método utilizado para calcular la matriz MSEP fue la validación cruzada utilizando 10 iteraciones para dividir el conjunto de datos de calibración y así realizar 10 veces la validación (10-*bold cross valdation*) para cada uno de los modelos de 0 a 10 ncomPLS.

Por lo tanto, para seleccionar el numero optimo de ncomPLS en la elaboración del modelo, es necesario realizar el graficado de la segunda columna de la matriz MSEP en función de los modelos de 0 a 10 ncomPLS para visualizar los errores de predicción para cada uno de los componentes PLS, y así poder seleccionar el modelo con menor error MSEP como se aprecia en la gafica de la figura 2.8 [The MathWorks, 2010].

Capítulo 3

EXPERIMENTACIÓN.

En este capitulo de describe la preparación de las muestras contaminadas de suelo en el laboratorio químico. Así como, el montaje experimental y la adquisición de firmas espectrales en el laboratorio de comunicaciones ópticas.

3.1. Preparación de las muestras de suelo.

El material fue extraído de una zona vinícola a una profundidad aproximada de 35 cm, el cual se clasifica como suelo franco por estar constituido por arena, limo y arcilla. Las muestras de suelo utilizadas en la experimentación fueron tamizadas para luego ser preparadas en el laboratorio químico de la FIAD. El tamizado se realizó con un tamiz no. 30 ($<$ 2 mm de apertura) en el laboratorio de ingeniería civil de la FIAD como se muestra en la figura 3.1(a). Posteriormente las muestras se almacenaron en contenedores de 90 ml como se puede observar en la figura 3.1(b).

3.1.1. Análisis del compuesto contaminante.

Algunas investigaciones han demostrado que la técnica de espectroscopía utilizando la región VIS-NIR es capáz de detectar metales pesados a concentraciones mayores

(a) (b)

Figura 3.1: Tamizado y almacenamiento.

de 1 gr/kg con gran exactitud [Wu et al., 2007], [Celerino, 2008]. Por lo que, para la manipulación del compuesto $Co(NO_3)_2{*}6H_2O$, fue necesario conocer su masa molecular para determinar las cantidades a utilizar, y así poder obtener un gramo del metal "Cobalto" (Co) del compuesto..

- Masa molecular.

$$\begin{aligned} Co = 58.93u * 1 = &\quad 58.93 \\ N = 14.00 * 2 = &\quad 28.00 \\ O = 16.00 * 12 = &\quad 192.00 \\ H = 1.00 * 12 = &\quad 12.00 \\ \hline &\approx 292uma \end{aligned}$$

Con este valor de masa, se calculó el porcentaje de cobalto contenido en x cantidad del compuesto o en un mol del mismo.

- Porcentaje de cobalto.

$$\%Co = 58.93/292 = 0.201815 \approx 20\,\%$$

Con estos datos del compuesto, se propusieron 40 concentraciones algebraicamente calculadas con incrementos de 5 centésimas de gramos (Tabla 3.1) en soluciones de 13 ml de agua destilada. También, se propusieron cinco muestras más para realizar la validación del modelo de predicción. Estas nuevas concentraciones (Tabla 3.2) cubrieron todo el rango de las 40 concentraciones propuestas.

Tabla 3.1: Concentraciones propuestas para el desarrollo del modelo de predicción.

Muestras de calibración (concentraciones)	Compuesto $Co(NO_3)_2 * 6H_2O$ g/13 ml	Cobalto (Co) g/13 ml
1	0.5	0.1
2	0.75	0.15
3	1	0.2
4	1.25	0.25
5	1.5	0.3
6	1.75	0.35
7	2	0.4
8	2.25	0.35
9	2.5	0.5
10	2.75	0.55
.	.	.
.	.	.
.	.	.
40	10.25	2.05

Tabla 3.2: Concentraciones de validación. Primer grupo.

Muestras de Validación (concentraciones)	Compuesto $Co(NO_3)_2 * 6H_2O$ g/13 ml	Cobalto (Co) g/13 ml
1	0.6	0.12
2	3.9	0.78
3	5.8	1.16
4	7.5	1.5
5	9.9	1.98

Tabla 3.3: Concentraciones de validación. Segundo grupo.

Muestras de Validación (concentraciones)	Compuesto $Co(NO_3)_2 * 6H_2O$ g/13 ml	Cobalto (Co) g/13 ml
1	0.5	0.1
2	0.8	0.16
3	1.4	0.28
4	1.6	0.32
5	1.8	0.36
6	2.8	0.56
7	4.7	0.94
8	5.6	1.12
9	6.6	1.32
10	8.4	1.68
11	11.3	2.26

3.1.2. Proceso de laboratorio químico.

Con los niveles de concentración propuestos (para calibración y validación) se procedió a desarrollar el trabajo en el laboratorio químico de la FIAD utilizando el siguiente material (Figura 3.2(a) y 3.2(b)):

- 1 Báscula de precisión digital.

- 1 Báscula de precisión analógica.
- 1 Espátula.
- 1 Mortero.
- 1 Embudo de cristal.
- 1 Matraz Erlenmeyer.
- 10 Cápsulas petri.
- 100 Recipientes contenedores de 90 ml.
- 10 Vasos de precipitado de 50 ml.
- 2 Vasos de precipitado de 250 ml.
- 1 Vaso de precipitado de 100 ml.
- 2 Probetas de 10 ml.
- 1 Frasco lavador.
- Lentes protectores.
- Guantes de latex.
- Cubre bocas.

Se prepararón 56 concentraciones propuestas (calibración y validación) en agua destilada, después se realizó la mezcla entre cada una de las soluciones preparadas y porciones de 100 gr de suelo para obtener las muestras contaminadas (Figura 3.3).

(a) (b)

Figura 3.2: Material de laboratorio químico.

Figura 3.3: Mezcla entre el contaminante y suelo.

Posteriormente se realizó un homogenizado de la mezcla del contaminante y el suelo (Figura 3.4), obteniéndose una masa arcillosa. Todas la muestras fueron colocadas en recipientes de vidrio petri-dish sobre una charola e introducidas en un horno de resistencia a 125 $°C$ durante 24 horas aproximadamente como se muestra en la figura 3.5.

Figura 3.4: Proceso de homogenizado.

Figura 3.5: Proceso de secado.

Después de finalizar el proceso de secado, las muestras contaminadas fueron molidas/granuladas utilizando un mortero, como se muestra en la figura 3.6.

Figura 3.6: Proceso de granulado.

Finalmente, todas las muestras fueron colocadas en recipientes contenedores de 90 ml con 100 gr de suelo. Dichos contenedores fueron etiquetados con la información de

la concentración del contaminante, y el número de muestra.

3.2. Espectrometría del suelo contaminado.

En este sección se presenta la experimentación llevada a cabo para obtener las firmas espectrales de las muestras de suelo contaminadas. Se muestra el montaje experimental, la calibración, y la adquisición de la información.

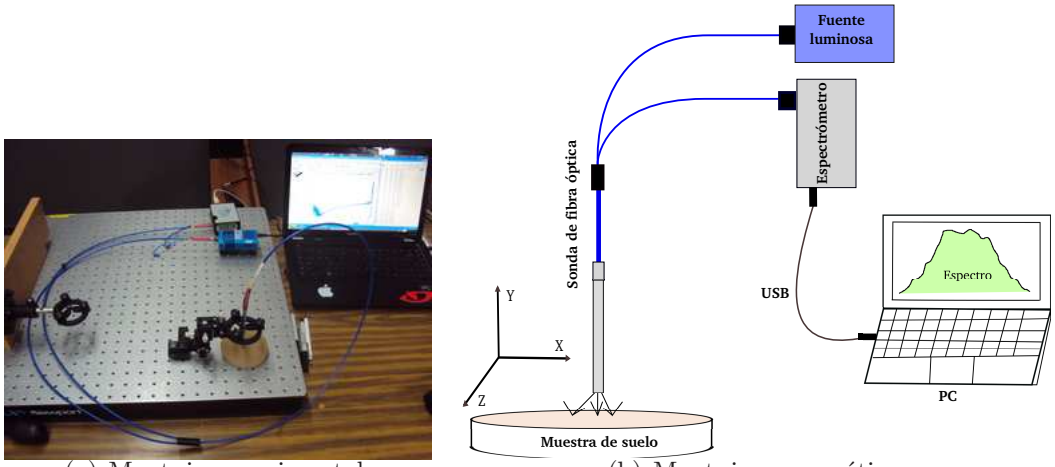
3.2.1. Material y equipo.

Para el análisis espectral se utilizó un sistema integrado con un espectrómetro (*USB4000 de Ocean Optics*) responsable de convertir una señal luminosa (análogica) en señal digital, la cual es adquirida por una computadora portátil con la ayuda del programa *Spectra Suite OceanOptics* que contiene los controladores del espectrómetro para su correcto funcionamiento. El espectrómetro se utiliza en conjunto con una sonda de fibra óptica (*R200-Angle Reflection Probe de Ocean Optics*), la cual esta compuesta por un conjunto de siete fibras, cada una de ellas con un diámetro de 200 μm aproximadamente, de las cuales, seis fibras son para canalizar la luz proveniente de la fuente luminosa, y la fibra restante es para la captura del reflejo luminoso. Asimismo, fue necesario utilizar el programa *MatLab* para importar los datos y realizar un procesamiento de la información.

El montaje experimental para el análisis espectral con el sistema mencionado anteriormente se muestra en las figuras 3.7(a) y 3.7(b).

3.2.2. Firma espectral de referencia.

Para el análisis espectroscópico es necesario utilizar una firma espectral de referencia, la cual debe ser adquirida de un material con una superficie blanca con capacidad



(a) Montaje experimental. (b) Montaje esquemático.

Figura 3.7: Sistema de medición para el análisis espectral.

de reflejar la luz incidente de manera difusa. Para este proyecto de análisis de firmas espectrales, se analizaron diferentes materiales con superficies claras y uniformes. Se observó que el papel fotográfico proporciona firmas espectrales adecuadas para ser utilizadas como señales de referencia. De esta forma, para el análisis espectral de las muestras de suelo se utilizó papel fotográfico mate para adquirir la referencia necesaria (Figura 3.8). La firma espectral de referencia fue capturada y almacenada por el programa *Spectra Suite de Ocean Optics* para validar las firmas espectrales de las muestras de suelo. Los datos de la firma espectral de referencia almacenados por el programa fuerón exportados a *MatLab* y se muestran en la figura 3.9 en función de la longitud de onda.

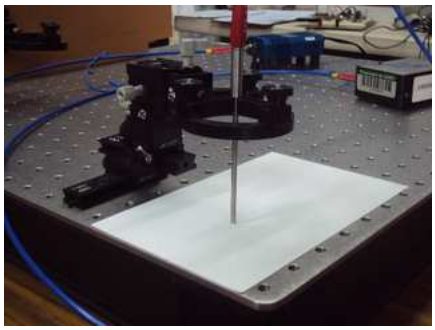


Figura 3.8: Adquisición de la firma espectral de referencia.

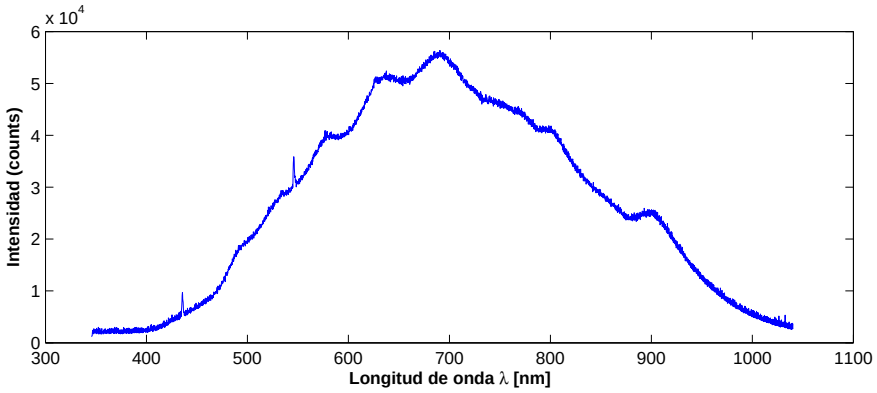


Figura 3.9: firma espectral de referencia.

3.2.3. Obtención del intervalo óptimo de longitudes de onda.

El intervalo del espectro electromagnético utilizado en este estudio estuvo limitado por el espectrómetro debido a que su intervalo de operación es de 345 nm a 1040 nm, con una resolución aproximada de 0.3 nm.

En los primeros barridos realizados a las muestras de suelo se observó que las firmas espectrales contenían valores de reflectancia para 3646 longitudes de onda. Por consiguiente, se procedió a capturar 100 espectros con frecuencia de 1 espectro por segundo. Estos espectros fueron almacenados en una matríz de datos con dimensiones de 100 x 3646 (Figura 3.10) en donde las filas representan el número de espectros obtenidos a una muestra dada y las columnas representan cada una de las 3646 longitudes de onda.

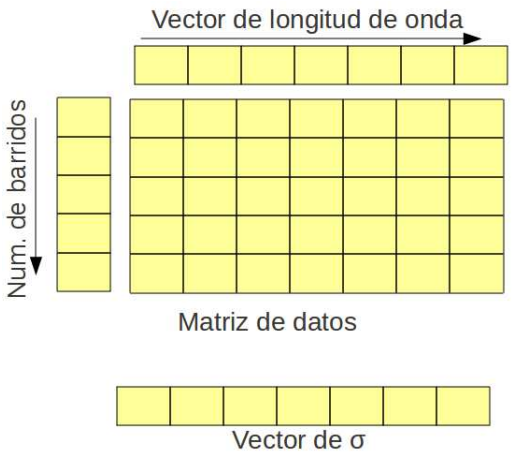


Figura 3.10: Matríz de datos.

Estos datos fueron sometidos a análisis estadísticos que ayudaron a identificar las longitudes de onda que concentran la información de cada muestra contaminada con el compuesto $Co(NO_3)_2 * 6H_2O$. Con este análisis realizado a los datos, se tuvo la capacidad de poder eliminar a las longitudes de onda que solo aportaron ruido, con lo cual, se logró reducir las dimensiones de las matrices utilizadas para contener los datos espectroscópicos, reduciendo la cantidad de información, y lo mas importante, agilizando el tiempo de procesamiento por el ordenador para realizar el proceso de análisis en la elaboración del modelo de predicción.

Este análisis estadístico consistió en calcular y almacenar la desviación estandar (σ) para cada longitud de onda utilizando la matríz de datos, dichos valores de σ fueron almacenados en un nuevo vector llamado vector de desviación estándar (σ) como se muestra en la figura 3.10. Debido a que el espectrómetro proporciona valores de intensidad para cada una de las 3646 longitudes de onda, este vector σ contiene 3646 valores. A cada longitud de onda le corresponden 100 valores correspondientes a los 100 barridos realizados a la muestra de suelo. Considerando que cada barrido significa adquirir una firma espectral.

Para poder determinar el intervalo óptimo de longitudes de onda (vector λ) a utilizar

en el análisis, el vector λ fue graficado contra el vector de desviación estándar como se muestra en la figura 3.11. A partir de esta gráfica, se pudieron observar los picos correspondientes a las señales de ruido aportadas por el equipo óptico (espectrómetro). Con esta información fuimos capaces de determinar estas longitudes de onda ruidosas, y como consecuencia, acotar el intervalo del vector λ.

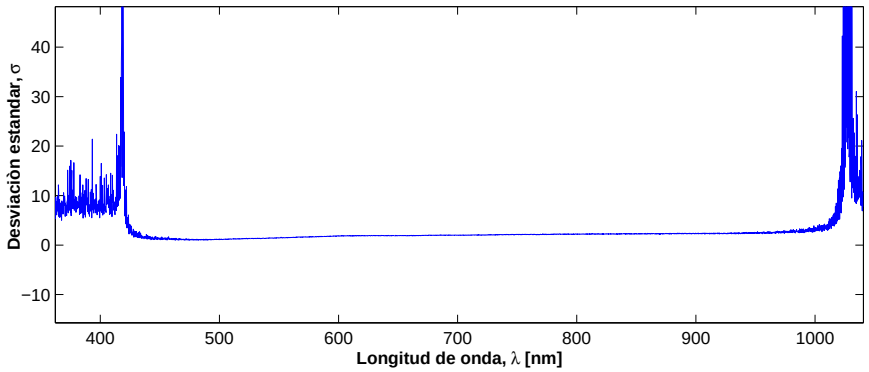


Figura 3.11: Vector σ en función del vector λ.

De la figura 3.11, se puede apreciar una región constante de la desviación estándar entre los 450 nm y los 980 nm. De esta forma, se considera a esta región como un intervalo óptimo que concentra la información cuantitativa del contaminante en las muestras, o que por lo menos, no aporta ruido. Por lo tanto, como resultado se determinó el intervalo de análisis del vector de longitud de onda λ de 450 nm a 980 nm.

Como consecuencia de este análisis, obtuvimos un nuevo rango espectral que se redujo de 3646 longitudes de onda (350 nm – 1040 nm) a solo 2811 (450 nm – 980 nm) (Figura 3.12).

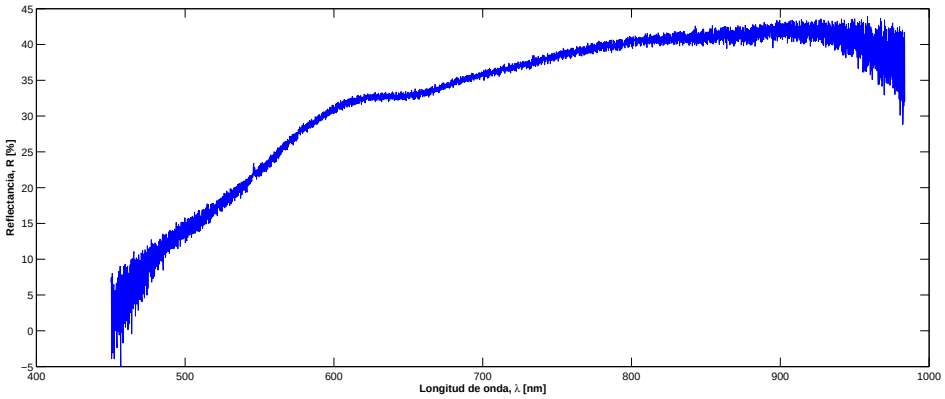


Figura 3.12: Firma espectral con 2811 longitudes de onda λ. Intervalo acotado.

Un espectro (firma espectral) sin este proceso estadístico es mostrado en la figura 3.13, en la cual se aprecia que las longitudes de onda que aportan ruido, no fueron eliminadas en comparación al espectro con intervalo acotado mostrado en la figura 3.12.

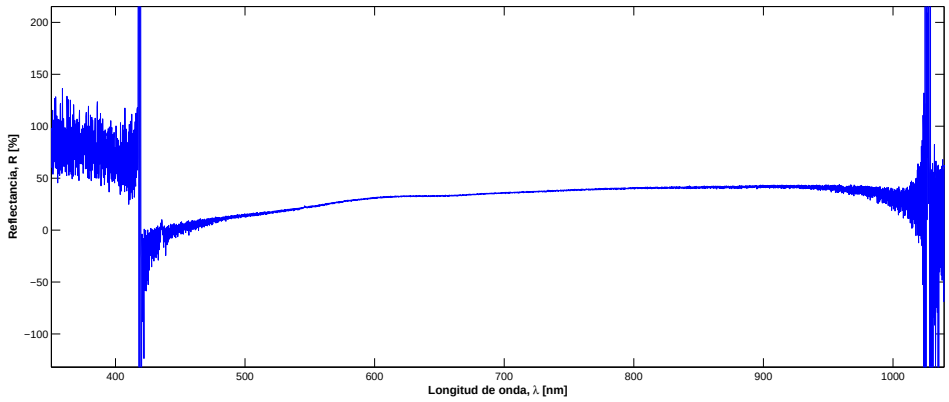


Figura 3.13: Firma espectral con 3646 longitudes de onda λ. Intervalo no acotado.

3.2.4. Determinación del número de barridos.

Para la elaboración del modelo de predicción se analizaron tres muestras de suelo con la sonda de fibra óptica con el fin de determinar el número óptimo de firmas espectrales

necesarias para obtener toda la información cuantitativa del compuesto contaminante en una muestra. El procedimiento consistió en capturar 100 firmas espectrales de una muestra utilizando la sonda, dichos espectros fueron adquiridos a una frecuencia de un espectro por segundo. Por lo que, para obtener 100 firmas espectrales, fue necesario someter la muestra a un análisis espectroscópico durante aproximadamente 2 minutos realizando barridos aleatorios por toda la superficie de la muestra, la cual se colocó en un recipiente "petri-dish" como se muestra en la figura 3.14.

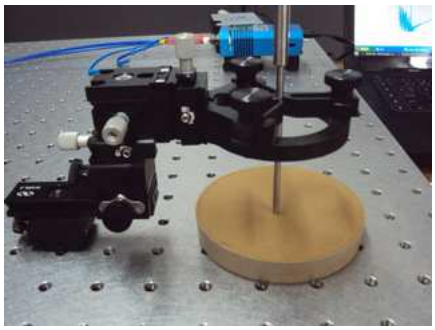


Figura 3.14: Análisis espectroscópico para determinar el número de barridos.

Una vez capturadas las 100 firmas espectrales, estas fueron almacenadas en una nueva matríz de datos como la mostrada anteriormente en la figura 3.10.

El análisis consistió en calcular la desviación estandar (σ) para 100 grupos de espectros, en donde el primer grupo solo contuvo un espectro, el segundo contuvo a dos espectros, y así sucesivamente hasta el grupo número 100 el cual contuvo 100 espectros, correspondiente al número de barridos, tomando en cuenta que cada barrido es una firma espectral o espectro obtenido de la muestra en análisis. Como resultado, se obtuvieron 100 valores de desviación estándar los cuales fueron graficados contra cada uno de los 100 grupos como se ilustra en la figura 3.15.

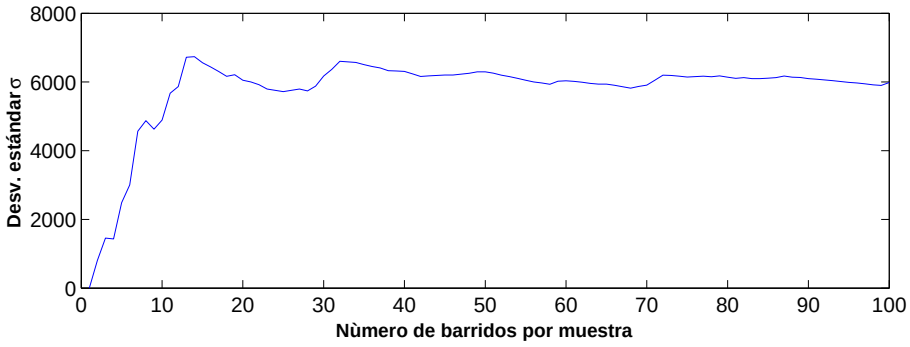


Figura 3.15: Determinación del número de barridos.

En la gráfica de la figura 3.15 se logra apreciar que a partir de los 40 barridos se alcanza a apreciar un intervalo o región constante en la función, lo cual implica que después de los 40 barridos por muestra se puede obtener la información de la misma adecuadamente. Con el fin de asegurar la captura de la información de las muestras, se determinó la cantidad de 60 barridos aleatorios sobre la superficie de una muestra como suficientes para obtener la información cuantitativa del nivel de contaminación por el compuesto $Co(NO_3)_2 * 6H_2O$.

3.3. Adquisición de firmas espectrales.

Una vez realizados los procesos anteriores de varianza (σ^2), se procedió a analizar espectroscópicamente las 45 muestras de suelo contaminadas con el compuesto $Co(NO_3)_2 * 6H_2O$. Recordando que 40 de estas muestras se propusieron para elaborar una base de datos, con el fin de obtener un modelo de predicción con la capacidad de determinar niveles de contaminación por metales pesados en muestras nuevas (muestras de validación). Las cinco muestras restantes, fueron para validar el modelo de predicción elaborado a partir del análisis a las 40 muestras primeramente propuestas. Finalmente, se puede mencionar que se obtuvieron 60 firmas espectrales de cada una de las 45 mues-

tras de suelo contaminadas. Cabe mencionar, que las 60 firmas espectrales obtenidas de cada una de las muestras de suelo fueron almacenadas en una matriz de datos, y estas mismas firmas fueron promediadas utilizando el programa computacional *Matlab* para obtener una sola firma espectral por muestra (Figura 3.16). Por lo que, si tenemos 45 muestras, entonces solo tendremos 45 firmas espectrales, las cuales son las que proporcionan la información de cada una de las muestras mencionadas. Para el análisis espectroscópico, la sonda fue colocada normalmente a la superficie de la muestra, la cual se colocó en un recipiente "petri-dish" con una separación de ≈2 mm entre la sonda y la superficie de la muestra.

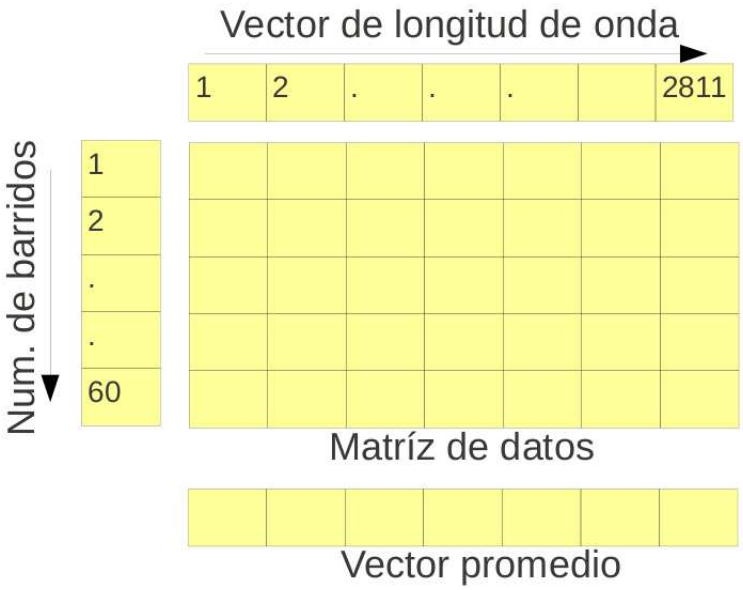


Figura 3.16: Arreglo matricial para cada una de las muestras de calibración y validación.

Capítulo 4

RESULTADOS.

En este capítulo se muestran los resultados obtenidos en el laboratorio de las firmas espectrales procesadas. Se muestran dos validaciones del modelo de predicción y se comparan por medio del error cuadrático medio de predicción.

4.1. Procesamiento de las firmas espectrales.

Las firmas espectrales extraídas de las muestras de suelo contaminadas fueron importadas y procesadas utilizando el programa computacional *Matlab*, como se muestra en el diagrama a bloques de la figura 4.1. Esto se realizó para poder elaborar un modelo de predicción que fuese capaz de proporcionar valores exactos de los niveles de contaminación por el compuesto $Co(NO_3)_2 * 6H_2O$ en muestras nuevas de suelo (muestras de validación).

Este proceso se integró por tres etapas como se aprecia en la figura 4.1, en donde la primera de estas etapas fue el montaje y la configuraciónón del equipo óptico. La siguiente etapa fue la encargada de la adquisicion de las firmas espectrales de las muestras de suelo utilizando el programa computacional *SpectraSuite*. En la última de estas etapas; se importaron los datos obtenidos en la etapa de adquisición de datos, se reali-

zó el proceso de suavizado, y se elaboró el modelo de predicción utilizando algoritmos estadísticos mediante el programa computacional *Matlab*.

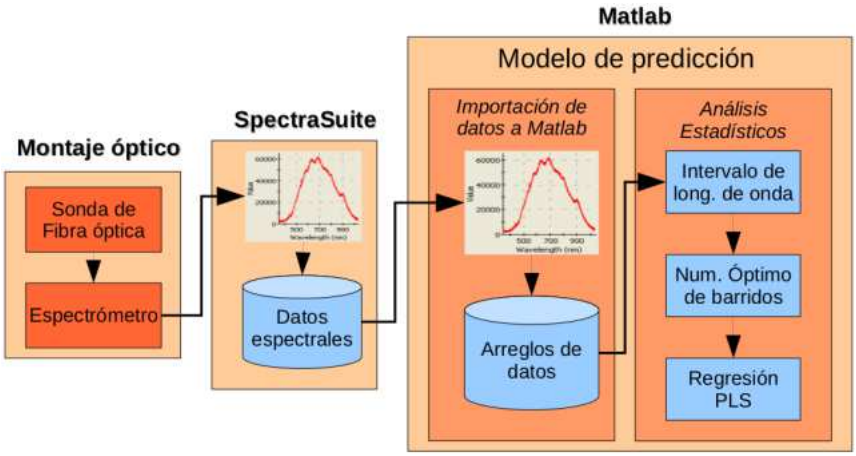


Figura 4.1: Diagrama a bloques del procesamiento de datos.

4.1.1. Suavizado de los datos.

Los espectros de las muestras fueron almacenados en matrices con dimensiones de 60 x 2811 como las que se han mostrado en la figura 3.16, en donde, cada una de las filas corresponden a cada una de las firmas espectrales, y las columnas corresponden a las longitudes de onda. En total, se obtuvieron 45 arreglos matriciales, los cuales almacenaron a los espectros de cada una de las muestras. Cada una de estas 45 matrices fueron el resultado de un promediado que dio como resultado vectores del tamaño del vector de longitud de onda (λ) y que corresponden a las firmas espectrales de cada una de las 45 muestras contaminadas bajo análisis. Después de haber obtenido las 45 firmas espectrales, estas fueron suavizadas utilizando la herramienta "smooth" de Matlab. Por medio de la función "smooth" se analizaron las firmas espectrales en modo de puntos para crear series de promedios. Para esta investigación se utilizó un "span" igual a cinco, con lo cual, cada punto promedio de la serie, fue el resultado de cinco valores de un dato original.

Esta herramienta de suavizado se utilizó para obtener espectros con contornos más limpios, en este caso, para eliminar el ruido en las firmas espectrales, lo cual se puede apreciar comparativamente en las figuras 4.2 y 4.3. Sin embargo, aplicar demasiado suavizado a los datos puede significar perdida de información, por lo que no es conveniente aplicar un suavizado muy pronunciado a los datos.

En este trabajo se desarrolló un código de programación utilizando *Matlab* para realizar la importación de los datos, así como también, para realizar el procesado de los datos de manera automática.

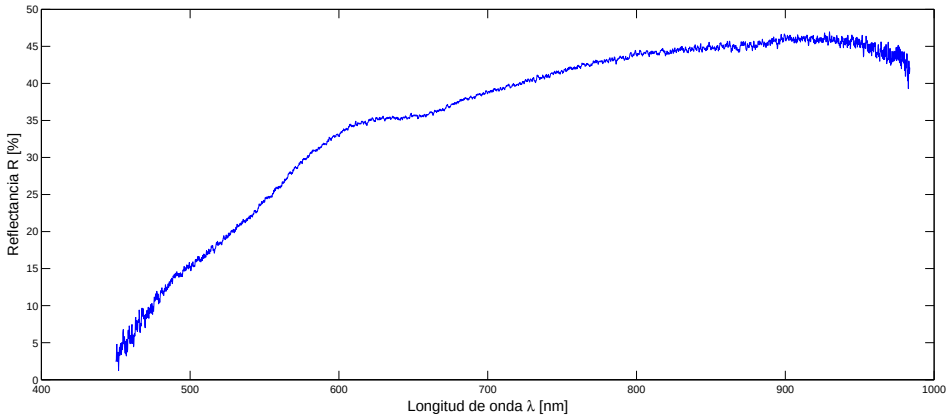


Figura 4.2: Espectro suavizado.

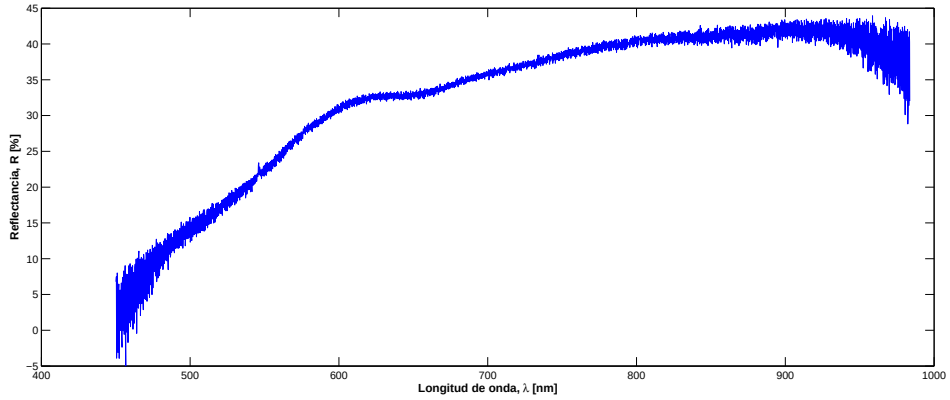


Figura 4.3: Espectro sin suavizado.

4.2. Modelo de predicción.

Después de haber obtenido las firmas espectrales de las 45 muestras de suelo contaminadas, las cuales fueron representadas por 45 vectores. Cuarenta de estos vectores se propusieron para elaborar el modelo de predicción (muestras de calibración). Estos 40 vectores fueron concatenados para formar una matríz con dimensiones de 40 filas por 2811 columnas como se muestra en la figura 4.4, a la cual se le denominó con el nombre de *matríz X*. Esta matríz X contiene toda la información de las 40 muestras de suelo contaminadas, y es utilizada para elaborar el modelo de predicción, el cual es útil para determinar niveles de contaminación en nuevas muestras (muestras de validación).

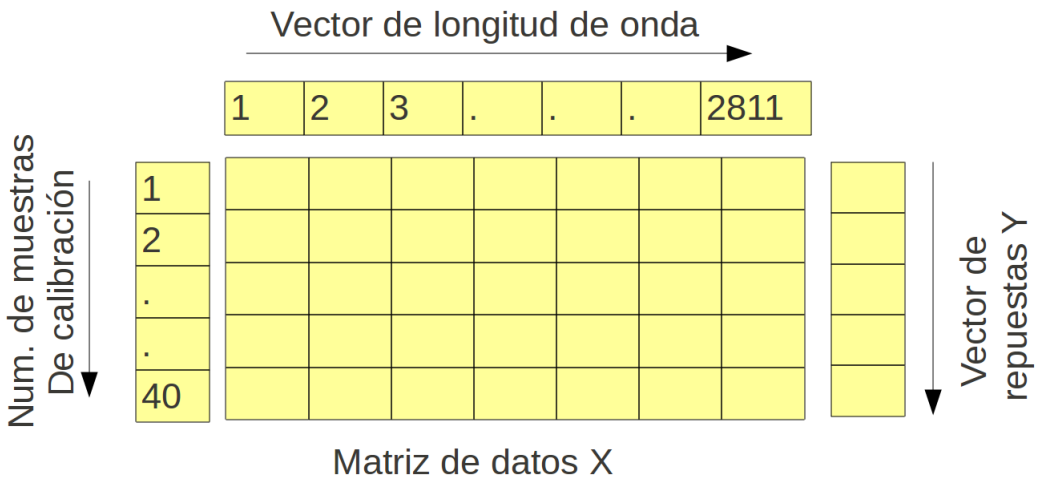


Figura 4.4: Matríz de datos y vector de repuestas.

En la figura 4.4 se logra apreciar un vector llamado *vector de respuesta Y* con dimensiones de 40 x 1, el cual almacena las cantidades correspondientes a los niveles de concentración del compuesto $Co(NO_3)_2 * 6H_2O$ utilizado en la preparación de las muestras de suelo. Por lo tanto, los valores de concentración mostrados en la tabla 3.1, son los correspondientes a los valores contenidos por el vector de respuesta "Y".

Las firmas espectrales de las muestras de validación fueron almacenadas en un arreglo matricial llamado *Matríz X** (Figura 4.5). Se debe de tener en cuenta que los valores

o niveles de concentración del contaminante ($Co(NO_3)_2 * 6H_2O$) en cada muestra de validación están representados por la tabla 3.2.

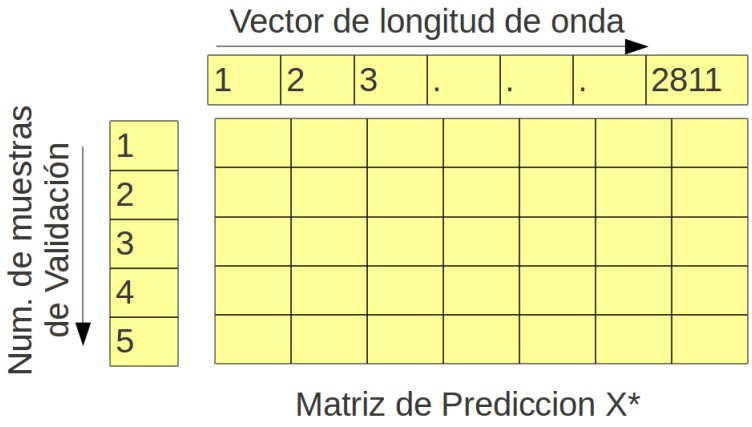


Figura 4.5: Matríz de datos de validación.

Esta matríz de validación (matríz X^*) tiene las dimensiones de 5 filas por 2811 columnas, en donde cada una de las filas representa a la información de cada muestra, y cada columna representa a cada longitud de onda del vector λ. Esta matríz fue utilizada después de la elaboración del modelo de predicción para la validación del mismo, en donde el objetivo fue predecir los valores de contaminación mostrados en la tabla 3.2.

Para la elaboración del modelo, se empleó una técnica basada en conceptos estadísticos llamada *Partial Least Square Regression (PLSR)*, la cual utiliza los datos de las muestras de calibración, y los valores de concentración utilizados en las muestras del grupo de calibración, para obtener coeficientes de correlación aplicando regresión lineal (PLSR) a este grupo de arreglos matriciales (matríz X y vector Y).

4.3. Regresión parcial por mínimos cuadrados (PLSR) (*Calibración*).

Para poder encontrar los coeficientes de correlación entre la matríz de datos X y el vector de respuesta Y que se muestran en la figura 4.4 se implementó el algoritmo de regresión (PLSR) utilizando las herramientas estadísticas del programa *Matlab*. Se utilizó la función *plsregress* para realizar la calibración del modelo como se muestra en el diagrama de flujo de la figura 4.6.

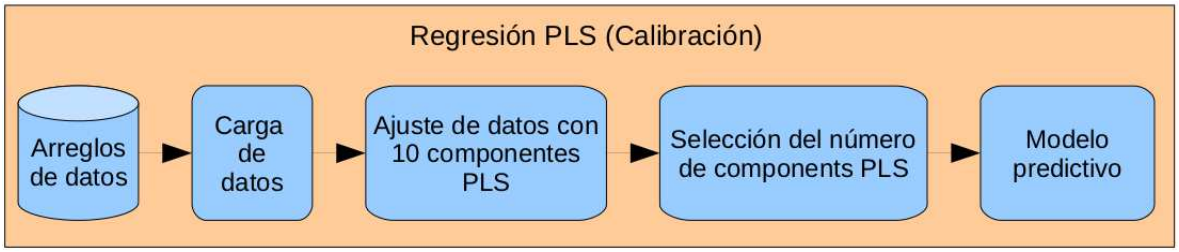


Figura 4.6: Diagrama de flujo para la calibración del modelo predictivo.

En este diagrama de flujo (Figura 4.6) se muestran los procesos necesarios para realizar el modelo y la calibración del mismo. Por medio de estos procesos se realiza la manipulación de los datos espectrales procesados anteriormente (espectros sin ruido y suavizados) con los cuales se ha elaborado y calibrado el modelo de predicción. A continuación se describen los procesos mencionados.

4.3.1. Arreglo de datos.

Los datos espectrales se almacenaron en las variables *matrizX_vectorY* y *validacion*. Estas variables contienen a la matríz de datos X y al vector de respuesta Y, así como también, al vector de longitud de onda (λ). La variable *validacion* contiene los datos espectrales correspondientes a las muestras que se emplearon para validar el modelo calibrado (modelo de predicción). Estos arreglos son los utilizados en la regresión lineal,

la cual posteriormente fue de utilidad para encontrar a las variables de correlación entre ambas matrices.

4.3.2. Carga de datos.

Este proceso implementa algoritmos de programación, el cual tiene la función de cargar los datos almacenados en la base de datos contenedora de las variables necesarias para la elaboración y calibración del modelo. Una vez cargados los datos, estos se graficaron para conocer las firmas espectrales de las muestras de calibración, así como se visualiza en la gráfica de la figura 4.7. También, se graficó la matriz de validación para visualizar las firmas espectrales (Figura 4.8).

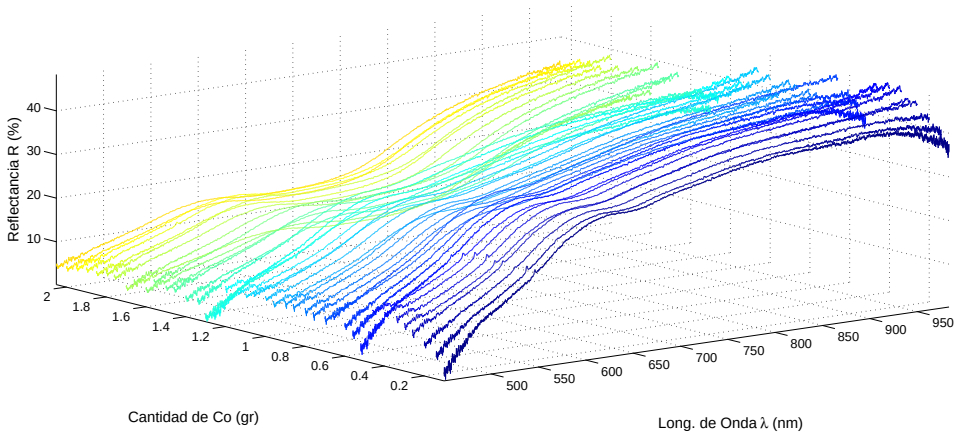


Figura 4.7: Firmas espectrales de las muestras de calibración.

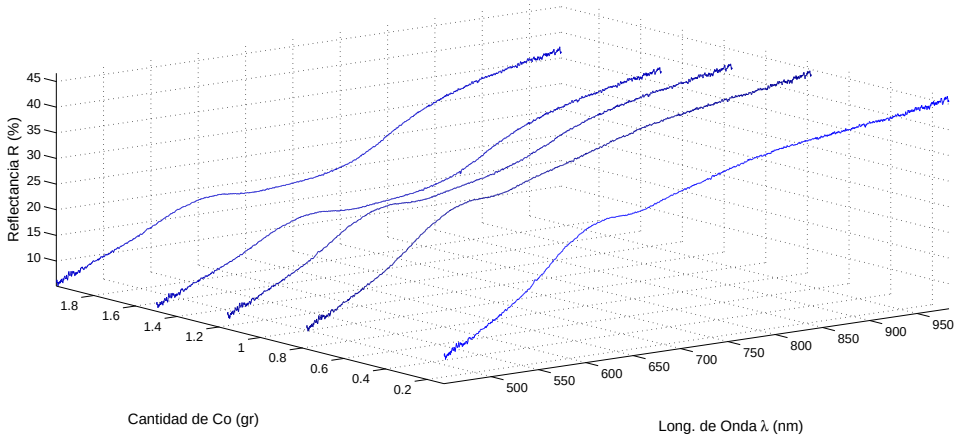


Figura 4.8: Firmas espectrales de muestras de validación del primer grupo.

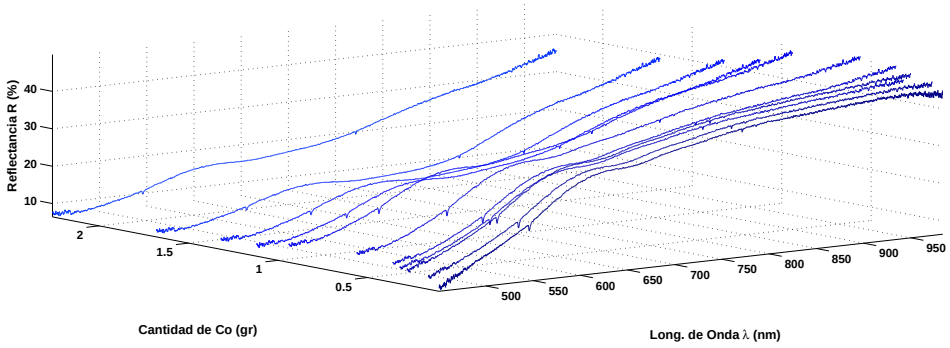


Figura 4.9: Firmas espectrales de muestras de validación del segundo grupo.

Las gráficas de las figuras 4.7, 4.8 y 4.9 muestran las firmas espectrales de las 56 muestras propuestas para elaborar y calibrar el modelo de predicción (Figura 4.7), así como las firmas espectrales para la validación del mismo (Figuras 4.8 y 4.9).

4.3.3. Ajuste de datos con 10 componentes PLS.

Este método está basado en la reducción de variables, llamadas "componentes PLS", y con las que podemos reducir la cantidad de datos espectrales sin pérdida de información utilizando una cantidad de componentes PLS adecuada [Vázquez, 2004]. Una vez

que se han cargado los datos, se aplica la función *plsregress* a la matríz de datos X y al vector de respuesta Y, como se puede observar en la función con la siguiente sintaxis:

$$[XL, YL, XS, YS, BETA, PCTVAR, MSEP, stats] = \textbf{PLSREGRESS}(\textbf{X,Y,ncomp,...})$$

Esta función se ejecutó sustituyendo "ncomp" por el numero 10, con lo cual se indicó el número de componentes principales PLS a utilizar. Se logra apreciar que la función *plsregress* utiliza la matríz de datos X y el vector de respuesta Y. Cuando se ejecuta esta función, el programa arroja como resultado un grupo variables, que permiten el cálculo de los coeficientes de predicción.

$$[XL, YL, XS, YS, BETA, \textbf{PCTVAR}, MSEP, stats]$$

Una vez obtenidas estas variables, se utilizó el vector $PCTVAR$ el cual contiene valores que indican el porcentaje de varianza explicada en el vector de respuesta Y. Este vector tiene dimensiones de 2 filas x (ncomp+1) columnas, lo que es igual a 2 filas por 11 columnas. Este vector es de suma importancia debido a que ayudó a determinar el número óptimo de componentes PLS con los cuales se obtuvo un mejor ajuste de los datos de la matriz X sobre el vector Y, lo cual se describe más adelante.

La función plsregress, desarrolla una regresión PLS del vector Y sobre la matríz X, utilizando n-componentes PLS. Estos componentes PLS son los que determinan el tamaño de las variables reducidas a partir de X y Y, estas variables reducidas son llamadas variables de carga (XL y YL) y estas son las que concentran la mayor parte de la información útil para elaborar el modelo de predicción. Por lo tanto, entre menor sea el número de componentes PLS, las dimensiones de estas variables de carga serán menores. Con lo cual se agiliza el tiempo de proceso de los datos y de respuesta.

Una manera de determinar el número óptimo del componentes PLS fue graficando el vector de varianza $PCTVAR$ como función del número de componentes *ncomp* utilizado en la función *plsregress* como se muestra en la gráfica de la figura 4.10.

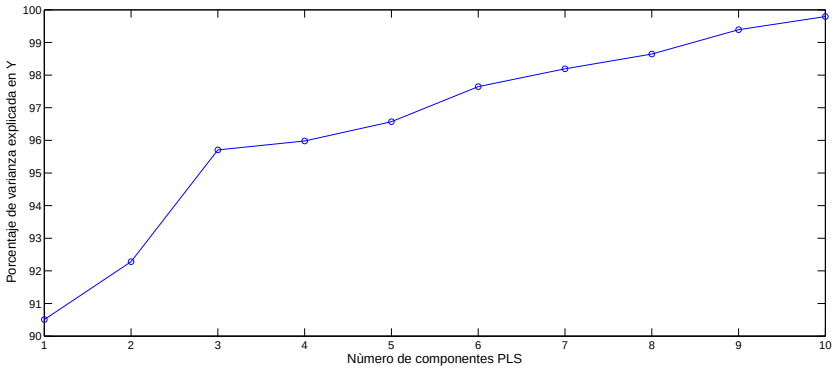


Figura 4.10: Varianza explicada en Y.

En la gráfica de la figura 4.10 se muestra el porcentaje de varianza en Y con respecto al número de componentes PLS. Se puede apreciar en la gráfica que al utilizar tan solo un componente PLS se puede explicar el 90 % de la varianza en Y, lo cual es un porcentaje muy bueno. Este porcentaje es capáz de indicar que tan ajustados estarán los datos espectrales a los valores del vector Y, el cual contiene los valores de concentración utilizados en la preparación de las muestras de calibración. Por lo tanto, el poder explicar la varianza arriba del 90 % indica que los datos espectrales podrán tener un muy buen ajuste con los datos del vector Y.

Otra manera que también se utilizó para seleccionar el *ncompPLS* adecuado, fue utilizando la variable *Error cuadrático medio de predicción (MSEP por sus siglas en inglés)* proporcionado por la misma función *plsregress* con la siguiente sintaxis:

$$[XL, YL, XS, YS, BETA, PCTVAR, \mathbf{MSEP}, stats] = PLSREGRESS(X, Y, ncomp, \mathbf{'cv',10})$$

Esta variable MSEP es el producto de realizar una *validación cruzada* con 10 iteraciones utilizando los datos de la matriz X y los del vector de respuesta Y como se ilustra en la figura 4.11. Así, se puede determinar la precisión del modelo en la predicción de las muestras de validación.

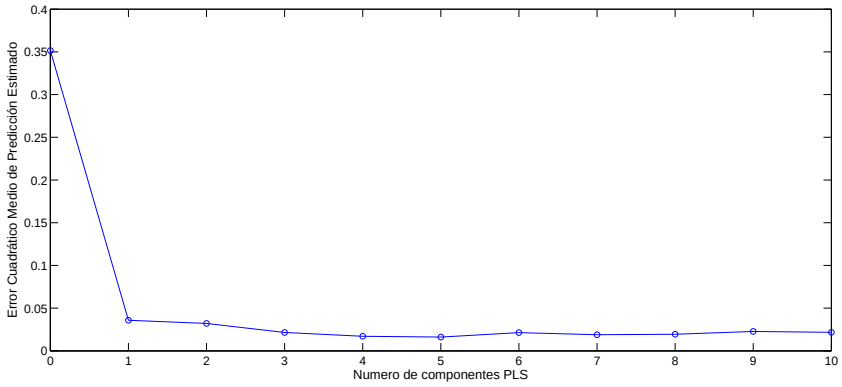


Figura 4.11: Error cuadrático medio en función del número de componentes PLS.

En la gráfica de la figura 4.11 se aprecia que existe un error de predicción relativamente pequeño (cercanos a cero) en los 5 componentes PLS en comparación con los demás componentes PLS de la gráfica. Hay que recordar que el valor del MSEP es la media de los errores cuadráticos de predicción de cada una de las 10 iteraciones utilizadas en la validación cruzada. (*10-fold cross validation*).

4.3.4. Selección del número de componentes principales PLS.

En este trabajo se consideró utilizar 5 componentes principales para realizar el modelo y con los cuales se explica cerca del 97 % de la varianza en Y.

En la gráfica de la figura 4.12 se aprecia que los datos o la información de las firmas espectrales están ajustadas a la recta, la cual representa a los valores del vector de respuesta Y, y en el cual se almacenan los valores de concentración propuestos para la contaminación de las 40 muestras de calibración. Cada uno de los puntos en la gráfica (figura 4.12) representa a la informacion de cada una de las 40 muestras de calibración barridas con la fibra óptica. Se puede apreciar en esta misma gráfica que se tiene un valor $R^2 = 0.9657$, el cual representa a la correlación existente entre la respuesta ajustada y la observada. A este valor se le conoce como *coeficiente de correlación*, y se puede apreciar

que este valor es muy cercano a la unidad, lo cual significa que las firmas espectrales de las muestras de calibración son capaces de proporcionar información para la elaboración del modelo predictivo con el cual se analizaron nuevas muestras contaminadas (muestras de validación).

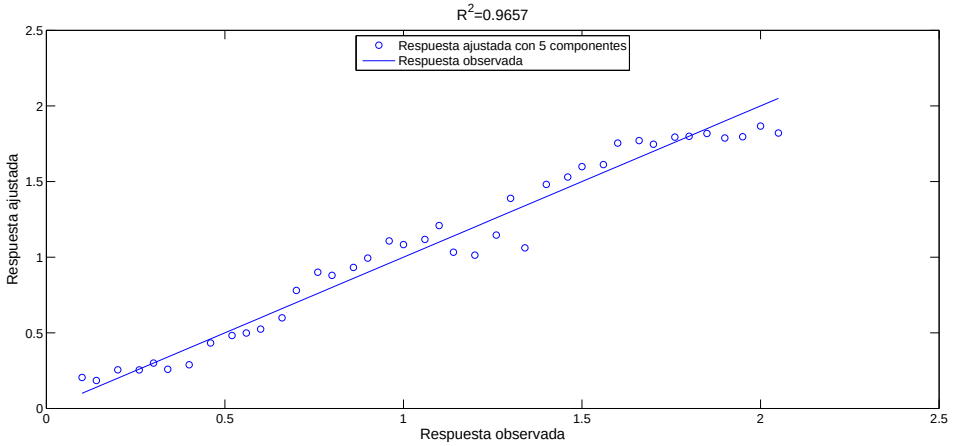


Figura 4.12: Datos espectroscópicos ajustados con 5 componentes PLS.

Para obtener una respuesta ajustada de las firmas espectrales en función de la respuesta observada (valores de concentración del compuesto, tabla 3.1), se utilizó la variable *BETA* obtenida en la función *plsregress* como se muestra en la ecuación 4.1.

$$Y = [X] * [\beta] + [Residuals] \tag{4.1}$$

Esta variable *BETA* (β) es un vector con la misma longitud que el *vector de longitud de onda + 1*, y contiene coeficientes de correlación para cada una de las 2811 longitudes de onda del vector de longitud de onda (λ). Esta variable *BETA* es la que determina predicciones del compuesto $Co(NO_3)_2 * 6H_2O$ en firmas espectrales de suelo contaminado con este mismo. Por lo tanto, la manera de visualizar el ajuste de las firmas espectrales de calibración y conocer el coeficiente de correlación fue utilizando el resultado de la ecuación 4.1, en donde se aprecia la multiplicación de la matriz X

(datos de calibración) y la variable $BETA$. De la ecuación 4.1 se puede determinar el vector de respuesta Y con la ayuda de los *residuales*.

4.3.5. Modelo predictivo.

Se determinó utilizar el modelo de predicción con 5 componentes PLS debido a su robustez, ya que mostró tener un valor de error de predicción bajo ($MSEP < 0.05$) empleando una validacion cruzada con 10 iteraciones(*10-fold cross validation*) con las muestras de calibración. También, debido a que se obtuvo un coeficiente de correlación cercano a la unidad ($R^2 = 0.9657$) lo cual demostró la existencia de una elevada correlación entre las firmas espectrales de la matríz X y los valores de concentración del vector de respuesta Y como se muestra en la gráfica de la figura 4.12.

A partir de estos datos, se determinó utilizar la variable "BETA" (β) obtenida en la función *plsregress* con 5 componentes PLS para realizar predicciones a muestras ajenas a las muestras de calibración utilizadas para generar el modelo, en este caso, hablamos de las muestras de validación propuestas en la tabla 3.2.

4.4. Validación del modelo de predicción.

En esta sección se muestran las validaciones realizadas al modelo de predicción utilizando dos conjuntos de muestras de validación con distintas concentraciones de contaminación propuestas. En la primera validación fue utilizado el conjunto de muestras representadas en la tabla 3.2, mientras que la segunda validación fue realizada utilizando los niveles de contaminación representados en la tabla 3.3.

4.4.1. Primera validación.

Para validar nuestro modelo de predicción, se utilizó la matriz X^* la cual contiene las firmas espectrales de las muestras de validación con las concentraciones propuestas en

la tabla 3.2. Recordemos que la dimensión de esta matríz es de *5 filas x 2811 columnas*, en donde las filas representan a las cinco firmas espectrales de cada una de las muestras de validación, y las columnas representan a cada una de las longitudes de onda del vector de longitud de onda (λ).

El objetivo de este proceso de validación fue utilizar la matríz de validación X^* (Figura 4.5) para predecir los valores de concentración del compuesto ($Co(NO_3)_2 * 6H_2O$) contenido en cada una de las muestras de validación, los cuales corresponden a los valores mostrados en la tabla 3.2. Para esto se utilizó la ecuación 4.1 con el cambio de variable que se muestra en la ecuación 4.2.

$$\hat{Y} = [X^*] * [\beta] \tag{4.2}$$

En donde $\hat{Y}$ es la variable que almacena las predicciones realizadas por el modelo a las cinco muestras de validación, y la variable X^* corresponde a la matríz contenedora de las firmas espectrales de validación. La variable BETA (β) de la ecuación 4.2 es la variable obtenida de la función *plsregress* utilizando 5 componentes principales PLS ejecutada como se presenta a continuación:

$$[XL, YL, XS, YS, \mathbf{BETA}] = PLSREGRESS(X, Y, 5)$$

Esta variable BETA (β) es un vector que contiene 2812 valores que corresponden a los coeficientes de correlación para cada longitud de onda del vector λ, el cual almacena 2811 valores correspondientes a las longitudes de onda. Para poder realizar la operación de la ecuación 4.2 fue necesario concatenar una fila de unos a la izquierda de la matríz de predicción X^* para tener la misma cantidad de filas que el vector β con 2812 valores. Una vez desarrollada la operación de multiplicación en la ecuación de predicción (Ecuación 4.2), se obtuvieron los resultados dados en la tabla 4.1:

Tabla 4.1: Tabla comparativa de resultados.

	Valores calculados	Valores predichos
Muestras de Validación	Cobalto (Co) (gr)	Modelo $\hat{Y}$ (gr)
1	0.12	0.2263
2	0.78	0.7302
3	1.16	1.3278
4	1.5	1.6150
5	1.98	1.7284

Error de predicción del modelo.

En la tabla de resultados (Tabla 4.1) se muestran las predicciones ($\hat{Y}$) del modelo y las concentraciones del compuesto utilizadas en las muestras de suelo que fueron utilizadas para la validación. Fue notable la exactitud del modelo, ya que los valores predichos por este mismo en las muestras 2, 3, 4 y 5 fueron muy cercanos a las concentraciones del compuesto preparadas en en el laboratorio químico, lo cual también se puede apreciar en la siguiente gráfica (Figura 4.13).

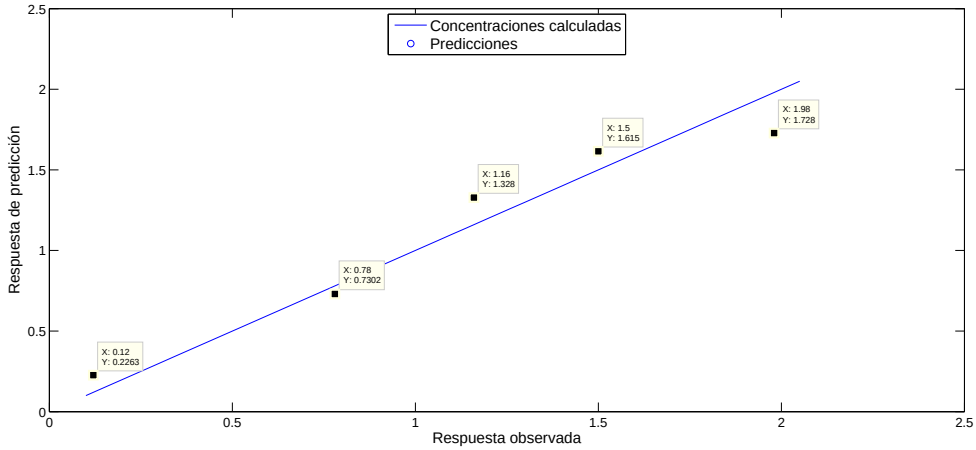


Figura 4.13: Predicciones de las muestras de validación.

La mejor manera de conocer el error de predicción del modelo en muestras ajenas a las utilizadas en la calibración, es calculando el MSEP dado por la ecuación 4.3.

$$MSEP = [\sum_{n}(\hat{y} - y)^2]/n = 0.02369 \tag{4.3}$$

En la ecuación del MSEP se emplean tanto los valores propuestos o calculados en laboratorio (Tabla 3.2), así como también, los predichos por el modelo ($\hat{Y}$) en la tabla 4.1. Como se puede apreciar, el error de predicción MSEP es relativamente pequeño, por lo que concluimos que se elaboró un modelo robusto capáz de predecir el compuesto $Co(NO_3)_26H_2O$ con exactitud en muestras de suelo.

4.4.2. Segunda validación.

Para esta validación se utilizaron distintas muestras contaminadas. El objetivo de esta segunda validación fue comprobar el correcto funcionamiento del modelo predictivo realizado con las 40 muestras de calibración. Por lo que se propusieron 11 nuevas

concentraciones del compuesto $Co(NO_3)_26H_2O$ para contaminar muestras de suelo, las cuales se muestran en la tabla 3.3.

Para esta validación se realizó el mismo procedimiento utilizado en la validación anterior. Se utilizó la matriz de predicción (o matriz de validación) X^* mostrada en la figura 4.5, pero ahora con dimensiones de *11 filas x 2811 columnas* para almacenar las firmas espectrales de las nuevas 11 muestras para validar el modelo predictivo, y las cuales se muestran en la figura 4.9.

Se empleó la ecuación 4.1 utilizada en la validación anterior para determinar las predicciones ($\hat{Y}$) del modelo. También se empleó la misma variable BETA (β), la cual se había calculado utilizando la función *plsregress* de Matlab empleando 5 componentes principales. Como resultado de la operación, se obtuvieron los nuevos valores predichos ($\hat{Y}$) mostrados en la tabla 4.2.

Tabla 4.2: Tabla comparativa de resultados.

	Valores calculados	Valores predichos
Muestras de Validación	Cobalto (Co) (gr)	Modelo $\hat{Y}$ (gr)
1	0.1	0.2110
2	0.16	0.2391
3	0.28	0.2988
4	0.32	0.3446
5	0.36	0.3948
6	0.56	0.5893
7	0.94	1.1644
8	1.12	1.5024
9	1.32	1.6286
10	1.68	1.7475
11	2.26	1.6488

Error de predicción del modelo.

La tabla 4.2 muestra las predicciones ($\hat{Y}$) del modelo junto con los valores de concentración propuestos del compuesto contaminante. Se logra apreciar que las mejores predicciones se realizaron en las muestras de la número 3 hasta la número 10 debido a que el error de predicción es muy pequeño en comparación con las muestras 1, 2 y 11, las cuales introducen un error de predicción mucho mayor al sistema.

Para calcular el error cuadrático medio de predicción (MSEP) en esta validación, se utilizó la ecuación 4.3, y solo se tomaron en cuenta los valores calculados y predichos para las muestras de la número 3 hasta la número 10 debido a que las primeras dos muestras (muestras de validación 1 y 2) contienen concentraciones muy bajas del contaminante y no es posible que el sistema las detecte con exactitud. Por otra parte, la muestra número 11 contiene una concentración de 2.26 gr de cobalto que no puede ser predecible por el modelo debido a que está preparado para analizar concentraciones menores a los 2 gr de cobalto sobre 100 gr de suelo. Por lo tanto, depués de haber desarrollado la ecuación 4.3, se obtuvo el valor de MSEP que se muestra en la ecuación 4.4.

$$MSEP = [\sum_{n}(\hat{y} - y)^2]/n = 0.2994035/8 = 0.03742543 \tag{4.4}$$

Del resultado anterior se observa que el error MSEP en este segunda validación sigue siendo pequeño, pero un poco mayor que el error de la primera validación. Por lo tanto, se concluyó que los datos obtenidos después de la predicción del modelo aplicado a las muestras de validación confirmaron los resultados de otras investigaciones [Wu et al., 2007] y [Celerino, 2008], aunque el modelo mostró problemas para predecir las concentraciones de cobalto cercanas a 0.1 gr/100 gr de suelo, su comportamiento mejoró para las predicciones de las muestras con mayor concentración de Cobalto. Es-

to demostró que los metales pesados son detectables con espectroscopía en el rango VIS-NIR en concentraciones mayores 1 g/Kg en suelos.

Como experimentación final, durante la segunda validación se propuso integrar muestras de calibración en la validación del modelo, para esto se emplearon 11 muestras que se muestran en la tabla 4.3 y las cuales también fueron utilizadas para la calibración del modelo. En total, incluyendo las muestras de las validaciones anteriores, se utilizaron 27 muestras para la validación. La gráfica de la figura 4.14 muestra los resultados del modelo sobre estas muestras en donde se pueden observar las predicciones para las muestras de la primera y segunda validación incluyendo las muestras de calibración.

Tabla 4.3: Tabla comparativa de resultados.

	Valores calculados	Valores predichos
Muestras de Validación	Cobalto (Co) (gr)	Modelo $\hat{Y}$ (gr)
1	0.14	0.2054
2	0.3	0.3596
3	0.52	0.5208
4	0.66	0.6816
5	0.7	0.8289
6	0.9	1.0536
7	1.14	1.5589
8	1.3	1.5286
9	1.46	1.5186
10	1.56	1.6048
11	1.7	1.5776

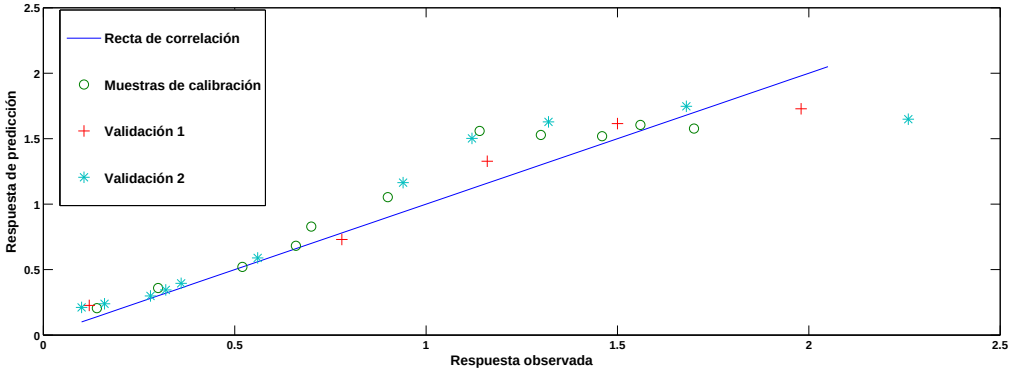


Figura 4.14: Predicciones de las muestras de validación.

El objetivo de introducir muestras del conjunto de calibración, las cuales habían sido utilizadas para la elaboración del modelo fue para visualizar el error cuadrático medio de predicción (MSEP) y con esto poder realizar comparaciones con los demás errores de predicción de las otras validaciones. Se encontró que los errores MSEP de las dos validaciones son casi iguales que el error determinado en las predicciones de las muestras de calibración (Ecuación 4.5). Recordemos que existe un conjunto de 40 muestras de calibración con las cuales se elaboró el modelo predictivo, por lo que, si optamos por utilizar estas muestras para validar el modelo, como resultado obtendríamos idealmente una predicción con error de cero.

$$MSEP = [\sum_{n}(\hat{y} - y)^2]/n = 0.0270 \tag{4.5}$$

Capítulo 5

CONCLUSIONES.

El primer objetivo de esta investigación consistió en determinar atributos que contribuyeran a evaluar niveles de fertilidad en suelos, calculando cuantitativamente minerales esenciales para la vida vegetal en el suelo. Debido al alcance del equipo óptico del laboratorio del cuerpo académico de la FIAD , esto no fue posible, ya que se requería de equipo que fuese capáz de operar en toda la región del infrarrojo, y nuestro equipo solo se limita a operar en la región visible y una pequeña parte del infrarrojo cercano. Por lo tanto, se optó por la evaluación de suelos contaminados por metales pesados.

Debido a los extraordinarios resultados de esta investigación, se redactó un artículo, el cual fue sometido y presentado con éxito en el Congreso Internacional de Investigación en Ingeniería Eléctrica "ENIINVIE" con el nombre *Visible-Near Infrared Spectroscopy to Assess Soil Contaminated with Cobalt* [Salazar et al., 2012].

5.1. Aportaciones

Fue posible confirmar la habilidad de la espectrocopía de reflectancia espectral (empleando el rango visible e infrarrojo cercano) para la evaluación de metales pesados en suelos. Por lo que, se puede considerar que esta técnica es una buena alternativa para

evaluar suelos contaminados. Por lo tanto, puede ser útil para la solución al problema de la contaminación que va en aumento proporcionalmente con el crecimiento de la población en el mundo.

La espectroscopía VIS-NIR, por lo tanto, se convierte en una muy buena opción de evaluación debido a que no es un método invasivo ni destructivo con el suelo sometido a esta técnica, sin dañar al medio ambiente. Hay que resaltar que en esta investigación se utilizaron las firmas espectrales para determinar un solo analito en el suelo. Sin embargo, un solo espectro es capáz de proporcionar valores cuantitativos de varios analitos en una muestra con una sola firma espectral.

También se demostró que los métodos estadísticos son una herramienta muy poderosa en los análisis espectroscópicos debido a que los datos pueden someterse a estas técnicas, ya sea para conocer propiedades químicas, físicas o biológicas en alimentos, frutos, suelo, agua, combustibles, etc.

Como resultado de esta investigación se logró implementar una técnica que basada en algoritmos estadísticos fue capáz de procesar datos espectroscópicos con resultados muy aceptables del modelo.

5.2. Trabajo a futuro

Con los resultados y conclusiones de esta investigación, se dedujo que esta técnica o método puede escalarse o ajustarse a cualquier equipo óptico con distintas cualidades o alcance para conocer infinidad de atributos de diferentes materiales organicos e inorgánicos dependiendo del problema que se quiera atacar.

Por lo que, queda a futuro realizar un modelo mucho mas robusto capáz de poder discernir varias propiedades del suelo con una sola firma espectral, probar la escalabilidad del algoritmo con otros equipos ópticos de distintas cualidades para analizar el desempeño del algoritmo del sistema, así como también, unificar los códigos de pro-

gramación de *Matlab* utilizados en los diferentes procesos del análisis espectral para optimizar el sistema.

También, debido a que el sistema esta integrado por equipo práctico y de fácil movilidad, queda como trabajo a futuro, optimizar este sistema para realizar análisis del suelo en campo y/o en sitio, y obtener resultados de evaluación en tiempo real.

Por otra parte, cabe mencionar que durante la etapa experimentación, las muestras de suelo fueron barridas con la fibra óptica manualmente, por lo que, podría realizarse una mejora en este proceso, ya que se podría automatizar este mismo para agilizar el análisis y agregar homogeneidad al mismo.

Bibliografía

[Alcalá, 2007] Alcalá, M. C. G. (2007). *La Contaminación de Suelos y Aguas: Su Prevención con Nuevas Sustancias Naturales*. Serie Ciencias. Universidad de Sevilla.

[Brown et al., 2006] Brown, D., Shepherd, K., Walsh, M., Dewaynemays, M., and Reinsch, T. (2006). Global soil characterization with VNIR diffuse reflectance spectroscopy. *Geoderma*, 132(3-4):273–290.

[Casanellas, 2008] Casanellas, J. P. (2008). *Introducción a la edafología: Uso y protección del suelo*. Ediciones Mundi-Prensa.

[Celerino, 2008] Celerino, Q. L. (2008). Aplicación de la espectroscopía de reflectancia infrarrojo cercano (NIRS) en el análisis de suelos.

[Chakraborty et al., 2010] Chakraborty, S., Weindorf, D. C., Morgan, C. L., Ge, Y., Galbraith, J. M., Li, B., and Kahlon, C. S. (2010). Rapid Identification of Oil-Contaminated Soils Using Visible Near-Infrared Diffuse Reflectance Spectroscopy. *Journal of Environment Quality*, 39(4):1378.

[DeCusatis and DeCusatis, 2006] DeCusatis, C. and DeCusatis, C. J. S. (2006). *Fiber Optic Essentials*. Electronics & Electrical. Academic Press, illustrate edition.

[Dunteman, 1992] Dunteman, G. H. (1992). *Principal Components Analysis*. Quantitative Applications in the Social Sciences. Sage, 2 edition.

[Fernández, 2005] Fernández, C. M. (2005). *Quimiometría*. Universitat de València.

[García et al., 1999] García, J. R., Virgós, J. M., and Rovira, J. M. V. (1999). *Fundamentos de óptica ondulatoria*. Servicio de Publicaciones de la Universidad de Oviedo.

[Luna, 2010] Luna, J. L. L. (2010). *Caracterización de suelo por medio de luz*. Maestría en ingeniería, Universidad Autónoma de Baja California.

[Mark and Workman, 2007] Mark, H. and Workman, J. (2007). *Chemometrics in Spectroscopy*. Academic Press.

[Norma Oficial Mexicana., 2002] Norma Oficial Mexicana. (2002). SECRETARÍA DE MEDIO AMBIENTE Y RECURSOS NATURALES. *Diario Oficial*, (ESTABLECE LAS ESPECIFICACIONES DE FERTILIDAD, SALINIDAD Y CLASIFICACIÓN DE SUELOS, ESTUDIO, MUESTREO Y ANÁLISIS.):85.

[Ocean Optics Inc., 2007] Ocean Optics Inc. (2007). SpectraSuite Spectrometer Operating Software.

[Odlare et al., 2005] Odlare, M., Svensson, K., and Pell, M. (2005). Near infrared reflectance spectroscopy for assessment of spatial soil variation in an agricultural field. *Geoderma*, 126(3-4):193–202.

[Ramos and Madero, 1979] Ramos, E. P. and Madero, M. C. M. (1979). *Problemas de Determinación Estructural Orgánica Por Espectroscopía IR*. Anales de la Universidad Hispalense Series. Universidad de Sevilla.

[Rosipal, 2006] Rosipal, R. (2006). Overview and recent advances in partial least squares. *Subspace, Latent Structure and Feature Selection*, pages 34–51.

[Ruda de Schenquer et al., 2004] Ruda de Schenquer, E. E., Mongiello, A., and Acosta, A. (2004). *Contaminacion y salud del suelo*. Universidad Nac. del Litoral.

[Salazar et al., 2012] Salazar, D. M., Reyes, H. M., Martínez-Rosas, M., Velasco, M. M., and Ortega, E. A. (2012). Visible-near infrared spectroscopy to assess soil contaminated with cobalt. *Procedia Engineering*, 35:245–253.

[The MathWorks, 2010] The MathWorks (2010). MATLAB R2010a Documentation software.

[Vázquez, 2004] Vázquez, F. M. G. (2004). Desarrollo de un método de selección de variables para datos espectroscópicos en el infrarrojo cercano. Technical report, Escola Tècnica Superior d'Enginyeria, Cataluña, España.

[Viscarrarossel et al., 2006] Viscarrarossel, R., Walvoort, D., Mcbratney, A., Janik, L., and Skjemstad, J. (2006). Visible, near infrared, mid infrared or combined diffuse reflectance spectroscopy for simultaneous assessment of various soil properties. *Geoderma*, 131(1-2):59–75.

[Wu et al., 2007] Wu, Y., Chen, J., Ji, J., Gong, P., Liao, Q., Tian, Q., and Ma, H. (2007). A Mechanism Study of Reflectance Spectroscopy for Investigating Heavy Metals in Soils. *Soil Science Society of America Journal*, 71(3):918.

[Yin et al., 2008] Yin, S., Ruffin, P. B., and Yu, F. T. (2008). *Fiber optic sensors*. Optical Science and Engineering. CRC Press, 6000 Broken Sound Parkway NW, Suite 300 Boca Raton, FL 33487-2742, second edition.

Apéndice A

Códigos del programa

A.1. Código para importación y análisis de datos.

```
columna=1;
for i =  0:149  %Este número corresponde al numero de escaneo realizados por
% muestra (150 firmas espectrales).
% Con esta función examinamos todos los archivos TXT adquiridos por SpectraSuite
% cuyo nombre de archivos son s1.txt, s2.txt, s3.txt,...,s150.txt.
    for j=columna
        % concatenando
        co1 = [ 's' num2str( i ) '.txt' ];
        load (co1);
        c=load (co1);
        lon=c(:,1);% la variable "lon" contiene el vector de longitudes de onda
%       m=length(lon');% esta es la longitud del vector "lon".
%       plot (lon,c(:,2));%% Plotea todos los espectros
        col=c(:,2);% Extraer columnas de cada espectro
        col_smooth=smooth(c(:,2));
%       plot (lon,col_smooth);%% Plotea todos los espectros suavizados
%       vector=zeros(m,j);
%       v=vector(:,j);
```

```
        for fila = 1:length(lon)
            vector(fila,columna)=col(fila,1);% aquí se almacenan todos los espectros.
            vector_s(fila,columna)=col_smooth(fila,1);% espectros suavizados
%           v=vector(:,columna);
        end
%se reduce la matriz de datos "vector" quedando de 450.18nm hasta 983.68nm
    mr=vector(490:3300,:);  % Matriz reducida: Se eliminan las componentes de ruido de
     %los espectros
    lonr=lon(490:3300,:); % vector "lon" reducido.
    mr_s=vector_s(490:3300,:); % Matriz reducida "SUAVIZADA": se eliminan las
     %componentes de ruido de los espectros
%   ploteo
    plot(lonr,mr);%matriz de datos sin suavizado (normal)
    figure;
    plot(lonr,mr_s);%matriz de datos con suavizado

    %%%%%%%%%% parte estadística para la matriz reducida

%     %%%%%              "sin suavizado"
%     m=mean(mr,2);% calculo de la media para cada espectro en la matriz de datos.
%     ds=std(mr,0,2); % desviación estándar para cada observación en la matriz de
% datos "vector".
%     mr_t=mr'; % calcula la traspuesta de la matriz c para que cada fila corresponda
% a cada espectro.
%     v=var(mr_t);%se calcula la varianza de cada fila de la matriz c.
%     v_t=v'; % traspuesta de la var
%     cv=ds./m; % coeficiente de variación
%     % m_ds=m+ds;%esta es la suma de la media mas la desviación estándar.
%     suma_cv=sum(cv);  %realiza la suma de todos los elementos del vector para
% realizar el graficado
%     suma_ds=sum(ds);%realiza la suma de todos los elementos del vector para realizar
```

```
% el graficado
%     suma_var=sum(v_t);
%     % plot(lon,v_t,'r');

     %%%%%              "con suavizado"

    m=mean(mr_s,2);% calculo de la media para cada espectro en la matriz de datos.
    ds=std(mr_s,0,2);% desviación estándar para cada observación en la matriz de datos
% "vector".
    mrs_t=mr_s'; % calcula la traspuesta de la matriz c para que cada fila corresponda
% a cada espectro.
    v=var(mrs_t);%se calcula la varianza de cada fila de la matriz c.
    v_t=v'; % traspuesta de la varianza
    cv=ds./m; % coeficiente de variación
    % m_ds=m+ds;%esta es la suma de la media mas la desviación estándar.
    suma_cv=sum(cv);  %realiza la suma de todos los elementos del vector para
% realizar el graficado
    suma_ds=sum(ds);%realiza la suma de todos los elementos del vector para realizar
% el graficado
    suma_var=sum(v_t);
    % plot(lon,v_t,'r');

%datos estadisticos para la determinación del numero de barridos (escaneos) por muestra
    vrz(1,columna)=suma_var;%matriz de varianzas
    d_s(1,columna)=suma_ds;%matriz de desviación estándar
    c_v(1,columna)=suma_cv;% matriz de coef. de variación.
    %v=vector(:,columna);


    end
    columna=columna+1;

end
```

```
% figure
plot(lonr,m,'r');
%%%%Gráficas para la determinación del num. de barridos por muestra
figure
plot((1:length(d_s)),c_v);
figure
plot((1:length(d_s)),d_s);
figure
plot((1:length(d_s)),vrz);
```

A.2. Determinar el número de componentes principales PLS

```
%Este codigo nos permite visualizar gráficamente el numero optimo de
%componentes principales a utilizar para explicar la máxima varianza en Y
%para el primer modelo de predicción utilizando el intervalo de
%                          450nm---983nm


clear all
clc
load matrizX_vectorY  %Esta variable almacenas las firmas espectrales del conjunto de
%calibración y los valores calculados de concentración del elemento
%contaminante. Esta variable se crea a partir del programa
%''importación y análisis de datos''. Mostrado en este anexo.
load validacion %Esta variable almacena las firmas espectrales de las muestras
%de validación,
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%%Esta sección es para calcular el num. de componentes.%%%%%

[Xloadings,Yloadings,Xscores,Yscores,beta,PctVar,MSEP] = plsregress(X,Y,10,'cv',10);
% Esta el mejor manera de seleccionar el ncompPLS. Aquí se utiliza el Error
```

```
% Medio Cuadrático de Predicción (MSEP)%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
plot(0:10,MSEP(2,:), 'b-o');
xlabel('Numero de componentes PLS');
ylabel('Error Cuadrático Medio de Predicción Estimado');
%legend({'PLSR'},'location','N');

%%%%%%%%%%%%%%%% Esta es otra manera para seleccionar el ncompPLS adecuada
figure;
plot(1:10,cumsum(100*PctVar(2,:)),'-bo');
xlabel('Numero de componentes PLS');
ylabel('Porcentaje de varianza explicada en Y');
% xlabel('Number of PLS components');
% ylabel('Percent Variance Explained in y');
```

A.3. Calibración del modelo

```
%Este codigo es para visualizar la gráfica de los valores ajustados y
%observados en Y para el primer modelo de predicción utilizando el
%intervalo de 450nm---983nm
clear all
clc
load matrizX_vectorY
load validacion

[dummy,h] = sort(Y);
[n,p] = size(X);
[Xloadings,Yloadings,Xscores,Yscores,beta,PLSPctVar,MSE,stats]=plsregress(X,Y,5,'cv',10);

yfit = [ones(n,1) X]*beta;
TSS = sum((Y-mean(Y)).^2);
RSS_PLS = sum((Y-yfit).^2);
rsquaredPLS = 1 - RSS_PLS/TSS;
```

```
plot(Y,yfit,'bo',Y,Y);
title('R^2=0.9657');
xlabel('Respuesta observada');
ylabel('Respuesta ajustada');
% xlabel('Observed Response');
% ylabel('Fitted Response');
legend({'Respuesta ajustada con 5 componentes' 'Respuesta observada'},'location','N');
% figure;
```

A.4. Modelo predictivo

```
clear all
clc
load matrizX_vectorY
load validacion

[dummy,h] = sort(Y);
set(gcf,'DefaultAxesColorOrder',jet(60));
plot3(repmat(lonr',40,1)',repmat(Y(h),1,length(lonr))',X(h,:)');
set(gcf,'DefaultAxesColorOrder','default');
xlabel('Long. de Onda \lambda (nm)'); ylabel('Cantidad de Co (gr)');
zlabel('Reflectancia R (%)'); axis('tight');
%2D
% plot(lonr,X);
% xlabel('Long. de Onda \lambda (nm)'); ylabel('Reflectancia R (%)');
grid on
[n,p] = size(X);
[Xloadings,Yloadings,Xscores,Yscores,beta] = plsregress(X,Y,5);

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%%%%%% validacion con muestras externas al modelo de calibracion %%%%%%
```

```
 mv=[mv1, mv2, mv3, mv4, mv5]';
 p=(mv)*(beta);% muestras de validacion. valores observados. 'predicciones'
prediccion=p

%%%%%Graficado de la respuesta de predicción contra la respuesta
%%%%%calculada.
figure
y=[0.12; .78; 1.16; 1.5; 1.98];%muestras de validación. valores calculados
plot(Y,Y, y, prediccion,'bo');
xlabel('Respuesta de observada');
ylabel('Respuesta predicción');
% xlabel('Observed Response');
% ylabel('Fitted Response');
legend({'Concentraciones calculadas' 'Predicciones'},'location','N');

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%%%%%%Ploteo de las firmas de muestras de validación 3D y 2D%%%%%%%%%%%%%%%%
% figure
% y=[0.12;0.78;1.16;1.5;1.98];
% [dummy,h] = sort(y);
% set(gcf,'defaultaxescolororder',jet(60));
% m=(mv(:,2:2812));
% plot3(repmat(lonr',5,1)',repmat(y(h),1,length(lonr))',m(h,:)');
% set(gcf,'defaultaxescolororder','default');
% xlabel('long. de onda \lambda (nm)'); ylabel('cantidad de co (gr)');
% zlabel('reflectancia r (%)'); axis('tight');
% grid on
%2D
%plot(lonr,(mv(:,2:2812)));
% xlabel('Long. de Onda \lambda (nm)'); ylabel('Reflectancia R (%)');
```