UNIVERSIDAD DE BURGOS
DEPARTAMENTO DE QUÍMICA
AREA DE EDAFOLOGÍA Y QUÍMICA AGRÍCOLA

TESIS DOCTORAL

UTILIZACIÓN DE TÉCNICAS DE MINERÍA DE DATOS PARA LA PREDICCIÓN DEL COMPORTAMIENTO DE BIOSÓLIDOS APLICADOS EN AGRICULTURA

DANIEL PÉREZ ALONSO

BURGOS, 28 de Febrero de 2017

UNIVERSIDAD DE BURGOS

Utilización de técnicas de minería de datos para la predicción del comportamiento de biosólidos aplicados en agricultura

La tesis «Utilización de técnicas de minería de datos para la predicción del comportamiento de biosólidos aplicados en agricultura», que presenta D. Daniel Pérez Alonso para optar al título de doctor, ha sido realizada dentro del programa de «Tecnologías Industriales e Ingeniería Civil» de la Universidad de Burgos bajo la dirección de los doctores D. José Francisco Díez Pastor y D Juan Carlos Rad Moradillo. Los directores autorizan la presentación del presente documento como memoria para optar al grado de Doctor por la Universidad de Burgos.

Vo. Bo. del Director:	Vo. Bo. del Director:	El doctorando:
Dr. D. José Francisco Díez Pastor	Dr. D. Juan Carlos Rad Moradillo	D. Daniel Pérez Alonso

Burgos, 28 de Febrero de 2017

A mi hermano, a mi madre y a mis yayos

Agradecimientos

Me gustaría expresar mi agradecimiento a todas aquellas personas que han hecho posible que llegue hasta aquí. Quiero empezar por agradecer a mis directores de tesis, el Dr. José Francisco Díez Pastor y el Dr. Carlos Rad Moradillo, por su enorme ayuda, su paciencia, comprensión y ánimo para poder terminar este trabajo.

A los profesores del área de Edafología, a Javier, Mila, Belén, Juani y Yolanda. A Salvador González Carcedo que me dio la oportunidad de empezar a trabajar con ellos.

Al resto de compañeros del laboratorio: Julio, Bárbara, Casilda, Mario, Dani ("Dimitri"), Marta G., Marta A., Laura, Vanesa, Jorge, Ana, Rajaa, Evan y Xandra, a todos ellos que pasaron algún momento en el laboratorio conmigo y escucharon mis desahogos. A Susana, sin duda tu trabajo ha sido fundamental para que pudiera llegar este día, gracias de todo corazón.
A mi amiga del otro departamento, a Nadine, por ser una persona con un gran corazón, alguien con la que se aprende cada día y cuya constancia me ha servido de ejemplo en los momentos que pensaba dejar todo.

También quiero agradecer a las personas han estado a mi lado prestándome su apoyo incondicional, mis amigos, los de dentro y fuera de la Universidad, imposible nombrarlos a todos, que siempre han estado animándome en los momentos duros, celebrando conmigo las bondades de la vida y que nunca, nunca han dejado de confiar en que algún día conseguiría terminar este proyecto. Pero os aseguro que ya no será tan fácil hacer planes sin que yo me apunte.

Por supuesto mi agradecimiento máximo a mi familia, sin la cual, no hubiese podido superar todos los pesares de la vida. A mis tíos, Carlos y Mª Carmen, por su gran bondad, por su infinita paciencia y su manera de ser, a mis primas, mis "hermanas pequeñas", Ángela y Gema porque su juventud y su cariño me dan la fuerza para continuar, me animan y me acompañan siempre.

A todos
Muchas Gracias

ÍNDICE DE TABLAS

INDICE DE FIGURAS

1. Introducción

La comunidad de Castilla y León es la de mayor de extensión de España y una de las más extensas de la Unión Europea. Su superficie es de 9,4 millones de hectáreas de las cuales 3,5 millones son superficies de cultivo (MAGRAMA, 2014). Las provincias de Burgos, Valladolid y Palencia son las provincias con mayor superficie cultivada, principalmente en régimen de secano, por las escasas lluvias y el clima de la región. Un 10% de la superficie explotada se dedica al regadío, con parcelas de producción intensiva y con mayores rendimientos que las cultivadas de secano.

Los principales cultivos herbáceos son los cereales, siendo denominada como el "granero de España". Aunque el trigo ha sido el cultivo más extendido desde los años 60 del siglo pasado actualmente la cebada ha aumentado el terreno cultivado. El centeno y la avena son dos cereales que también se cultivan en las parcelas de secano de Castilla y León. También, además de las leguminosas, se ha extendido el cultivo del girasol, con más de 300.000 ha en la comunidad ocupa el 40% de la superficie nacional dedicada a este cultivo. Se trata de un cultivo eminentemente de secano pero el 11% se cultiva en regadío.

Por otro lado, la industria agroalimentaria de Castilla y León ocupa el 18% de la población activa y aporta el 25% del PIB. Castilla y León cuenta con más de 2.600 industrias agroalimentarias. La industria alimentaria derivada de la explotación agraria y ganadera, se ha consolidado como un sector clave para el desarrollo económico de Castilla y León. Esto conlleva un aumento del volumen de residuos orgánicos reciclados en la agricultura durante las últimas décadas.

Muchos de estos residuos orgánicos tienen en su composición altos contenidos de nutrientes como N, P, K u otros oligoelementos, por lo que pueden ser considerados como potenciales fertilizantes, además del efecto positivo que su contenido en materia orgánica proporciona al suelo (Hargreaves et al., 2008). La aplicación de residuos de origen urbano o industrial como enmienda orgánica puede constituirse en una eficaz medida para la recuperación de suelos degradados y con baja fertilidad al mejorar sus propiedades físicas y ser fuente de nutrientes para las plantas y la microbiota del suelo (García-Gil et al., 2000, Crecchio et al., 2004). No obstante, su aplicación debe realizarse con los debidos controles evitando los excesos de nutrientes o la toxicidad de algunos de sus constituyentes, pues pueden constituirse en un potencial peligro para la salud humana o de los ecosistemas receptores (Singh and Agrawal, 2008) ;(Smith, 2009).

Al igual que el desarrollo industrial ha permitido un incremento en las producciones agrícolas, de residuos de origen industrial y urbano, también se ha producido una mayor facilidad para acceder a medidas realizadas en los laboratorios. De esta manera se ha minimizado el tiempo para realizar los análisis y un incremento en la precisión de los mismos. También el incremento en la capacidad por parte de los procesadores de las últimas generaciones de ordenadores permite realizar cálculos imposibles hasta hace unas décadas en los centros de investigación. Así, por ejemplo, desde la creación del primer microprocesador en 1971 que podía realizar hasta 60000 operaciones por segundo (640 kiloinstrucciones por segundo) con un frecuencia de reloj de 700

kHz hemos pasado a aproximadamente 117160 millones de instrucciones por segundo de los actuales procesadores en el año 2011. En un intervalo de tiempo de tan solo 40 años ha existido un incremento exponencial de las operaciones que se pueden realizar con nuestros ordenadores y la cantidad de datos almacenados en bases disponibles para los investigadores ha generado un nuevo enfoque y una nueva manera de entender el tratamiento de los resultados mediante la Minería de Datos.

La Minería de Datos se podría definir como un campo de confluencia entre las ciencias de la computación y la estadística que estudia el proceso de descubrir patrones con volúmenes de datos elevados. La principal función de la Minería de Datos es el análisis de grandes cantidades de datos para extraer patrones y dependencias interesantes para posteriormente poder incorporar esos resultados en un flujo continuo de datos. La Minería de Datos abarca todo el proceso de descubrimiento de conocimiento a partir de la información localizada en bases de datos (*Knowledge Discovery in Databases*, KDD) y no sólo el diseño y la aplicación de algoritmos de aprendizaje. A partir de una definición del problema, se estudian y analizan los datos recogidos, para que una vez procesados en los formatos adecuados, varios algoritmos de aprendizaje puedan ser utilizados para obtener una información útil y realizar una interpretación de los resultados obtenidos.

El convenio firmado entre la Consejería de Medio Ambiente de Castilla y León y el Grupo de Investigación en Compostaje de la Universidad de Burgos (UBUCOMP) ha permitido la financiación de todas las actividades realizadas en campo, los consiguientes trabajos analíticos de laboratorio así como su exposición en diferentes simposios nacionales e internacionales. Gracias a los datos obtenidos de esas experiencias se ha podido realizar este estudio. Este trabajo se centra en la aplicación de técnicas de Minería de Datos para el estudio de los resultados de la adición agronómica de un compost de lodos de depuradora de origen urbano a diferentes cultivos de la provincia de Burgos. Los resultados de este trabajo de investigación se basan en tres de las experiencias que se realizaron en campo. En concreto la aplicación de un compost de lodos de depuradora en una rotación de secano localizada en el término municipal de Villafruela, comarca del Arlanza (Burgos). Los lodos de depuradora corresponden a la Estación Depuradora de Aguas Residuales (EDAR) de la ciudad de Burgos, la cual procesa de forma conjunta las aguas procedentes de la actividad urbana e industrial y que son posteriormente sometidos a un proceso de compostaje mediante pilas volteadas en el cual se incorporan cortezas de pino. Dicho proceso se realiza en la planta “Cuesta de Burgos” que gestiona el grupo Sufi SA (actualmente Valoriza). Los datos pertenecientes a otro tipo de residuos y de tipo de cultivo no se han incluido en este estudio.

2. Objetivos

El objetivo principal se centra en la búsqueda de nuevos modelos de representación e interpretación de los resultados obtenidos en una experiencia agronómica sobre utilización de biosólidos como fertilizante orgánico mediante el uso de nuevas técnicas de análisis de regresión.

Los objetivos parciales planteados en este estudio son:

I. Selección de un tipo de árbol de decisión que permite una fácil interpretación de los resultados.

II. Optimización de los parámetros de los árboles de decisión elegidos para una mejora en los coeficientes de correlación.

III. Interpretación de los árboles de decisión en las distintas experiencias contempladas en los conjuntos de datos disponibles.

IV. Extrapolación de los modelos obtenidos en el conjunto de datos a nuevos datos no incluidos en los modelos y su capacidad de predicción en las variables estudiadas.

V. Aplicación de métodos de regresores para lograr mejoras en las capacidades predictivas de los modelos.

3. Fundamento teórico

3.1. Uso del Descubrimiento de Conocimiento en Bases de Datos

El progreso tecnológico de las últimas décadas ha permitido una revolución digital donde se ha impulsado una era de la información acompañada de una accesibilidad a gran cantidad de datos para analizar. Las herramientas heurísticas para analizar esos datos de una manera automática y organizada son conocidas como Descubrimiento de Conocimiento en Bases de Datos (*Knowledge Discovery in Databases, KDD*).

El Descubrimiento de Conocimiento en Base de Datos (*KDD)* es un proceso que sigue distintas fases con el objetivo de llegar hasta el conocimiento oculto que esconden los grandes conjuntos de datos. Los pasos involucrados en todo el proceso *KDD* son:

1. Identificar el objetivo del proceso de *KDD*.
2. Seleccionar un conjunto de datos de destino o un subconjunto de muestras de datos en la que se pueda realizar el descubrimiento de conocimiento oculto.
3. Limpiar y procesar previamente los datos y decidir las estrategias para manejar los campos en blanco o alterar los datos según las necesidades.
4. Simplificar los conjuntos de datos mediante la eliminación de las variables no deseadas. Analizar las características útiles que pueden ser utilizadas para representar los datos, dependiendo del objetivo o tarea.
5. Elegir los algoritmos de Minería de Datos necesarios para el descubrimiento de patrones ocultos. Este proceso incluye decidir qué modelos y parámetros podrían ser apropiados para el proceso global de *KDD*.
6. Buscar patrones de interés estableciendo una forma de representación en particular, que pueden incluir reglas de clasificación o árboles de regresión.
7. Interpretar los conocimientos esenciales de los patrones extraídos.
8. Utilizar el conocimiento e incorporarlo en otro sistema para futuras experiencias.

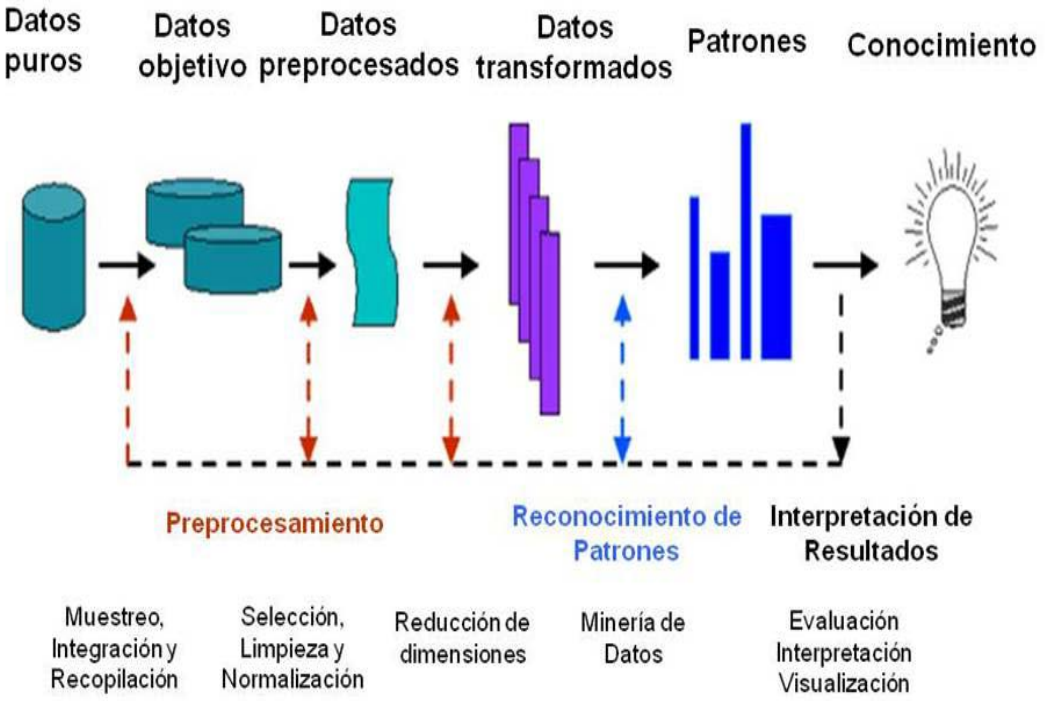


Figura 3-1 Esquema de un proceso KDD.

En el centro de los procesos de *KDD* se encuentra la Minería de Datos (*Data Mining*) que se puede definir como "*el proceso que consiste en la aplicación de análisis de datos y descubrimiento de algoritmos que, con unas limitaciones de eficiencia computacional aceptables, producen una particular enumeración de los patrones (o modelos) sobre los datos*". (Fayyad, 1996).

Esta nueva disciplina es aplicada a la exploración, análisis y extracción de patrones de comportamiento no accesibles con las herramientas de cálculo tradicionales. Esto no es algo realmente nuevo pues analistas, estadísticos, ingenieros y economistas han trabajado durante mucho tiempo con la idea de que los patrones subyacentes a los datos se pueden buscar de forma automática, ser identificados, validados y posteriormente utilizados en labores de predicción. Lo que es novedoso es el asombroso aumento de las oportunidades para encontrar patrones en los datos. Nuestra sociedad, cada vez más desarrollada, dispone de una enorme cantidad de información, potencialmente importante, localizada en numerosas bases de datos, accesible, pero que todavía no se ha acabado de tratar de la forma adecuada.

La Minería de Datos, que es un término relativamente nuevo, se basa en la articulación de la información útil de los datos mediante técnicas de análisis y extracción de patrones, previamente desconocidos, y si es posible su generalización para hacer predicciones exactas sobre los datos futuros. Por supuesto que estas nuevas metodologías también tienen sus problemas como

pueden ser la construcción de patrones demasiado simplistas o no resolutivos por sus coincidencias, exclusivas, con el conjunto de datos utilizado para un propósito determinado. Además, como con cualquier técnica para el estudio de datos, existirán inexactitudes debidas a los errores en el correcto funcionamiento de los aparatos, en el número de medidas realizadas o los datos perdidos. Por eso se deben utilizar algoritmos que sean lo suficientemente robustos como para poder extraer patrones a pesar las irregularidades de los patrones o la falta de exactitud de los datos.

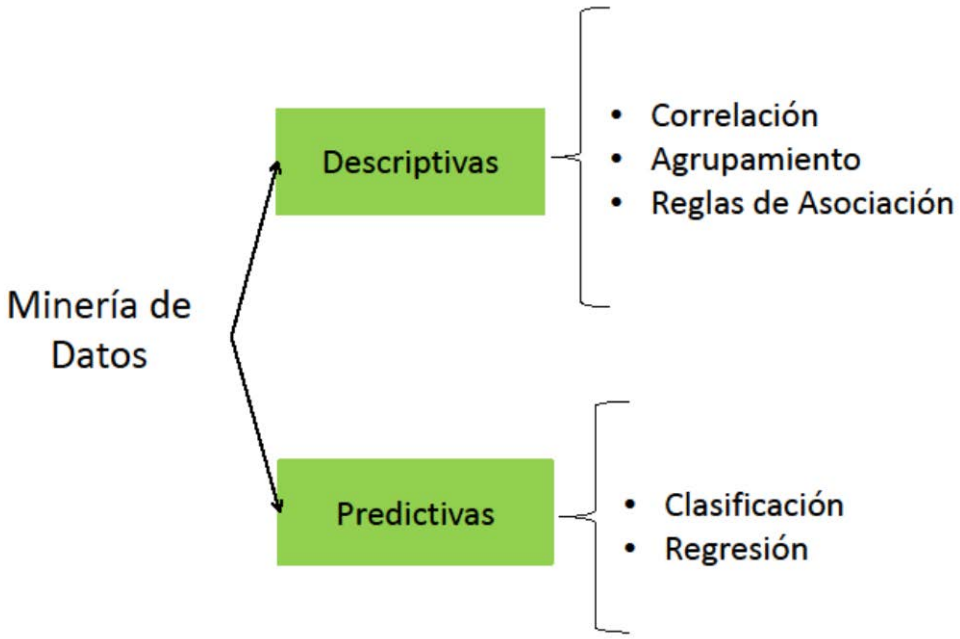
A pesar de su reciente creación, la Minería de Datos y la estadística no son dos disciplinas enfrentadas, sino más bien complementarias. A lo largo del tiempo, las dos han tenido lugares diferenciados de trabajo. Así por ejemplo, cuando se trata de hacer hincapié en encontrar una diferencia entre los datos con respecto a un factor experimental, la estadística se centraría en la prueba de afirmar o rechazar la hipótesis nula, mientras que la Minería de Datos se dedicaría a formular un proceso para la génesis de un patrón como una búsqueda a través de las posibles hipótesis de partida dadas. Otras aplicaciones de la Minería de Datos son la caracterización y clasificación de datos, pudiéndose construir modelos estadísticos que tengan un objetivo clasificatorio. Por otra parte, las tareas de Minería de Datos se pueden aplicar en la creación de modelos estadísticos, por ejemplo, usando la estadística para modelar los datos de ruido y valores perdidos. . En la Figura 3-2 se recogen las posibles aplicaciones de la Minería de Datos, desde las descriptivas, que incluyen la correlación, la agrupación y el descubrimiento de reglas de asociación de los datos, a las predictivas, centradas en la regresión y la clasificación.

La estadística es útil para establecer un conjunto de varios patrones desde los datos, así como para entender el mecanismo subyacente que genera y que afecta a los patrones. Los métodos estadísticos también se pueden utilizar para verificar los resultados de Minería de Datos. Por ejemplo, después de un modelo de clasificación o predicción el modelo extraído debe ser verificado mediante una evaluación estadística de hipótesis. Una prueba de hipótesis estadística (también llamado análisis de confirmación de los datos) toma decisiones estadísticas a partir de datos experimentales. Un resultado se llama estadísticamente significativo si es poco probable que haya ocurrido por casualidad. Si la clasificación o predicción del modelo es válida, entonces la estadística descriptiva del modelo aumenta la solidez del modelo.

La aplicación de métodos estadísticos en la Minería de Datos es un proceso complejo y resulta un reto enorme el conseguir un “escalado” del método estadístico desde un pequeño conjunto hasta una gran base de datos. Debido a la alta complejidad en el cálculo de muchos de los métodos estadísticos, los algoritmos deben ser diseñados y ajustados cuidadosamente, ya que, cuando tales métodos se aplican en grandes conjuntos de datos se puede incrementar el coste computacional. Los algoritmos diseñados trabajarán con datos distribuidos en distintos lugares, físicamente en distintas bases de datos, o con distintos tipos de variables (binarias, numéricas, fechas, ordinales, etc…) lo cual se complica, aún más, cuando se requiere la Minería de Datos para un análisis continuo y rápido donde el flujo de datos es en tiempo real.

A pesar de las diferencias señaladas, muchas de las técnicas que se han desarrollado para la Minería de Datos parten de esquemas similares a los utilizados por la estadística. Actualmente, ambas técnicas convergen desde un principio, tanto para perfeccionar el conjunto inicial de datos, como al construir la mayoría de los algoritmos de aprendizaje, los cuales utilizan pruebas estadísticas en la construcción de reglas o árboles, así como para la corrección de los modelos que están "sobre-ajustados" lo que depende también en gran medida de detalles en los ejemplos particulares utilizados para producirlos. Además, la Minería de Datos cuenta con numerosas herramientas para solucionar un determinado problema y así mismo, cada herramienta emplea diferentes algoritmos para su resolución.

Figura 3-2 Esquema aplicaciones de Minería de Datos.



- **Selección de Datos.**

La primera parte del proceso de Minería de Datos se basa en el planteamiento y análisis del problema que se desea estudiar, así como en la selección y análisis de los datos relevantes que mejor lo definen. Estos dos pasos permiten preparar de una manera adecuada los datos disponibles, para, finalmente, poder aplicar algoritmos de aprendizaje con altas probabilidades de éxito. Esta parte del proceso es la que determina, en gran medida, que las fases sucesivas sean capaces de extraer conocimiento útil y válido a partir de la información original.

Habitualmente cuando se trabaja con datos reales existen los denominados “datos impuros”, que conducen a la extracción de patrones y reglas poco útiles, por lo que es indispensable realizar tareas de pre-procesado. Cuando se

produce el pre-procesado de los datos se realizan modificaciones en la base de datos original: eliminación de anomalías, detección de valores erróneos o nulos, de valores que puedan contener un elevado componente de ruido, así como la inclusión de otras variables, no numéricas y definidas por el usuario, necesarias para poder hacer un seguimiento de las variables en función del tiempo. De esta manera, finalmente, se determina cuál es la información necesaria con la que se trabajará, eliminando aquella que es poco útil y completando la que falta.

Algunos algoritmos de Minería de Datos, especialmente algunos algoritmos de clasificación, requieren que los datos se encuentren en forma categórica por lo que, a menudo, es necesaria una transformación de los datos continuos en atributos categóricos (discretización) o la modificación de atributos discretos y continuos en un conjunto de nuevas variables con codificaciones binarias, mediante un proceso llamado binarización, el cual consiste en transformar una variable categórica con N posibles valores, en N variables categóricas con solo dos posibles valores: 0, 1 (variable binaria). Por otro lado cuando se tienen muchos valores distintos en atributos categóricos, o valores que se muestran en raras ocasiones, se pueden reducir los errores cometidos por los algoritmos mediante una reducción de las categorías a un número menor de valores posibles.

El Aprendizaje Automático a partir de la Minería de Datos puede realizarse de dos formas diferentes (Weiss and Indurkhya, 1998):

- El Aprendizaje Supervisado o Predictivo, es la tarea de aprendizaje automático que infiere una función a partir de datos de entrenamiento etiquetados (Mohri et al., 2012). En el aprendizaje supervisado, cada ejemplo es un par que consta de un objeto de entrada (típicamente un vector) y un valor de salida deseado (también llamada la señal de supervisión). Un algoritmo de aprendizaje supervisado analiza los datos de entrenamiento y produce una función deducida a partir de los mismos. Esta función servirá para colocar los nuevos ejemplos en el espacio que se ha creado a partir de los datos de entrenamiento. Así los nuevos datos se colocarán cercanos a aquellos datos del entrenamiento que son similares por sus características según la función establecida para la clasificación. Un escenario óptimo permitirá la elaboración de un algoritmo para determinar correctamente las etiquetas de clase para los casos que no se ven. Esto requiere que el algoritmo de aprendizaje generalice a partir de los datos de entrenamiento a situaciones que no se ven de una manera "razonable".

- El Aprendizaje No-supervisado o Descriptivo es la tarea de aprendizaje automático que infiere una función para describir la estructura oculta a partir de datos no etiquetados. Dado que los ejemplos que se dan para el aprendizaje son no marcados, no hay señal de error o recompensa para evaluar una posible solución. El aprendizaje no supervisado está estrechamente relacionado con el problema de la estimación de la densidad en

estadística (Bishop, 1996). Sin embargo el aprendizaje sin supervisión también abarca muchas otras técnicas que tratan de resumir y explicar las características clave de los datos.

3.2. Modelos en Minería de Datos

En muchos casos, el objetivo de la Minería de Datos es inducir un modelo predictivo. Al igual que los estadísticos, numerosos autores han estado interesados en la búsqueda de alternativas al modelo lineal restringido en el que el efecto de las variables explicativas es principalmente aditivo, por eso, se utilizan los árboles de decisión. Los modelos de predicción pueden ayudar para la toma de la mejor decisión. Los métodos supervisados intentan descubrir la relación entre los atributos de entrada (llamadas variables independientes) y un atributo de destino (variable dependiente). La relación que se establece es conocida como modelo. Por lo general, los modelos describen y explican fenómenos que están ocultos en el conjunto de datos y se pueden utilizar para predecir el valor del atributo de destino basado en los valores de los atributos de entrada.

Dentro de la Minería de Datos existen numerosos algoritmos que se pueden elegir en función del problema que se quiere resolver. Así mismo, la Minería de Datos cuenta con métodos para solucionar un determinado problema y cada método cuenta con diversos algoritmos o técnicas. Como se puede observar en la tabla 3-1 existen distintos algoritmos para las áreas de aplicación dependiendo de nuestros datos.

Tabla 3-1: Ejemplos de Algoritmos para clasificación y regresión.

Nombre	Predictivo		Descriptivo	
	Clasificación	Regresión	Agrupamiento	Asociación
Redes Neuronales	X	X	X	
Árboles de Decisión	X	X	X	X
Regresión Lineal y Logarítmica		X		
Regresión Logística	X			X

Existen dos tipos de modelos supervisados:

- Clasificación (clasificadores). Los clasificadores asignan la entrada de espacio en clases predefinidas.
- Regresión (regresores). En los modelos de regresión el mapa del espacio de entrada está en un dominio de valor real.

Ambos modelos comparten un enfoque de aprendizaje basado en las regularidades de las variables observadas por lo que las predicciones se hacen sobre las similitudes con situaciones previas. Los métodos en ambos casos difieren en la forma en que se expresa la similitud. Los árboles de regresión buscan hacer más importantes las propiedades compartidas explícitas, mientras que los enfoques basados en clasificadores, equiparan la disimilitud (falta de similitud) con una medida de la distancia.

Vamos a considerar sólo la tarea de aprendizaje supervisado, más específicamente de aprendizaje supervisado con modelos de estructura de árbol de decisión. Los métodos para la generación de árboles de decisión fueron iniciados por Hunt y sus compañeros de trabajo en la década de 1960 (Hunt, 1963), aunque su popularidad en las aplicaciones estadísticas se debe a la labor independiente de (Breiman et al., 1984) con dichas técnicas incorporadas en paquetes de software tales como *CART* y la evolución de *ID3* a *C4.5* (Quinlan, 1993).

Un árbol de clasificación o un árbol de regresión es un conjunto de reglas para predecir la clase o el valor numérico de un objeto a partir de los valores de sus variables de predicción. De esta manera, construir un árbol de decisión puede ser planteado como un problema recursivo llevado a cabo mediante la partición. El procedimiento de inducción del árbol es muy simple y eficiente. Además, los árboles de decisión y regresión no requieren realizar un costoso, en parámetros temporales, proceso inicial de cribado de los datos originales. Para ello, una vez planteado el problema, hay que analizar los datos, y por supuesto, la fiabilidad de la información que presentan. Entre los algoritmos existentes los árboles de decisión permiten una representación gráfica que no se puede realizar con las redes neuronales y tienen una estructura de separación que hace más sencilla la interpretación de los resultados frente a la regresión lineal.

Comparado con otros modelos existentes, como las máquinas de soporte vectorial, los modelos Gaussianos y las redes neuronales artificiales, los árboles de decisión tienen varias ventajas. En primer lugar, los algoritmos de aprendizaje permiten una distribución libre de los datos, es decir, no existen suposiciones previas especiales sobre la distribución que tienen. En segundo lugar, no requieren el ajuste de los parámetros o un periodo de entrenamiento computacionalmente costoso, como en el caso de máquinas de soporte vectorial. Debido a que se realiza una selección de los atributos en las pruebas de división, son bastante robustos a los datos considerados irrelevantes. Por estas razones, son fáciles de implementar y utilizar. Por último, el hecho de que los árboles de decisión sean fácilmente interpretables por los usuarios es lo que les proporciona su ventaja más importante. Cada árbol de decisión se puede representar con un conjunto de reglas que describen las dependencias entre las variables de entrada y la de destino.

3.3. Árboles de decisión y su funcionamiento

Dentro del análisis de cómo funcionan los árboles de decisión y cómo se ven influidos por los distintos algoritmos utilizados, se debe conocer qué árboles se adecuan a los parámetros establecidos por nosotros como satisfactorios para la finalidad propuesta, que no es otra que la de obtener una herramienta de fácil interpretación y sin un grado de complejidad excesiva en el resultado final. Por otro lado, un objetivo que el aprendizaje automático nos permite, es el lograr encontrar un modelo predictivo para futuros cambios realizados sobre las variables de entrada

Aquellos modelos que pueden servir como una herramienta explicativa entre objetos de diferentes clases, son los llamados modelos descriptivos. La función del modelo predictivo es, partiendo de un grupo de números reales o de valores de clase, crear un espacio de respuesta con un conjunto de variables descriptivas de diferentes tipos. Los modelos basados en árboles de decisión se emplean para predecir el valor de las variables respuesta, tanto numérica como nominal. Los tipos más simples de los modelos basados en árboles son árboles de clasificación y regresión (Rokach and Maimon, 2014) y por esto se encuentran entre los más empleados. Por un lado, los árboles de clasificación se utilizan para modelar conceptos representados con categorías simbólicas, mientras los árboles de regresión se utilizan para modelar las funciones definidas en el espacio de algunos o todos los atributos de entrada.

Un árbol de decisión permite mediante un conjunto de condiciones organizadas en una estructura jerárquica, llegar a la decisión final de tal manera que se puede tomar siguiendo una serie de condiciones que se cumplen desde la raíz del árbol hasta alguna de sus hojas. La elaboración de árboles de decisión se ha utilizado como método de Minería de Datos para representar los eventos que surgen a partir de una decisión. Los arboles de decisión resultan eficaces por proporcionar soluciones a problemas complejos de regresión no lineal, dado que se establece el enfoque de “*divide y vencerás*” aplicado sobre el espacio delimitado por las variables. Los árboles de regresión mejoran la

precisión para la variable de destino usando modelos cada vez más complejos en los nodos hoja por lo que su componente de sesgo del error es, normalmente, mucho menor. Las divisiones que se van llevando a cabo en el conjunto de los datos se hacen de manera recursiva mediante la elección de una partición que maximiza la reducción del error con respecto a las variables que se han separado en el nodo. Los procedimientos utilizados para construir árboles de decisión son por tanto algoritmos de arriba hacia abajo.

Tanto los árboles de clasificación como los de regresión buscan la predicción de una variable de respuesta *y*, dados los valores de un vector de variables predictoras *x*. Lo siguiente que debemos estudiar es cómo va a ser el crecimiento del árbol. Para ello, partimos de nuestros datos que constarán de T entradas y una respuesta para cada una de las N observaciones: es decir, (x_i, y_i) para i = 1, 2,. . . , N, con $x_i = (x_{i1}, x_{i2},..., x_{ip})$. El algoritmo tiene que decidir de forma automática cuáles son las variables y los puntos de división y también qué forma va a tener el árbol. Supongamos primero que tenemos una partición en M regiones $R_1, R_2,. . . , R_m$, y tenemos como resultado un c_m constante en cada región:

$$f(x) = \sum_{m=1}^{M} c_m \; I \; (x \; \in \; R_m)$$

El criterio para la selección de *f(x)* es por lo general el que establece la minimización del error de la suma de cuadrados

$$\sum (y_i - f(x_i))^2$$

Por eso, es fácil ver que el mejor c_m es el promedio de y_i en la región de R_m

$$\hat{c}_m = promedio \; (y_i | x_i \in \; R_m)$$

Dado que encontrar la mejor partición del espacio que establezca la suma mínima de cuadrados perfecta es computacionalmente imposible se procede partiendo de todos los datos, considerando una variable de división *j* y un punto de división *s* y esto nos define un par de semiplanos que dividen el espacio.

$$R_1(j, s) = \{X | Xj \leq s\}$$

y

$$\mathrm{R}_2(\mathrm{j}, \mathrm{s}) = \{\mathrm{X} | \mathrm{Xj} \leq \mathrm{s}\}$$

Luego se busca la variable *j* y el punto *s* que resuelva

$$min_{j,s}\left[min_{c1}\sum_{x_1\in R_{1(j,s)}}(y_i-c_1)^2+min_{c2}\sum_{x_2\in R_{2(j,s)}}(y_i-c_2)^2\right]$$

Y que para cada par de elecciones de *j* y *s* la minimización interior establece que

$$\hat{c}_1 = promedio\left(y_i|x_i \in R_{1(j,s)}\right)$$

y

$$\hat{c}_2 = promedio\left(y_i|x_i \in R_{2(j,s)}\right)$$

Para cada variable divisoria la determinación del punto de corte *s* puede hacerse fácilmente y de este modo se puede realizar una exploración con todos los puntos de entrada para establecer cuál es el mejor par: *j,s.*(Friedman et al., 2001).

Después de haber encontrado la mejor división, los datos resultantes quedan divididos entre las dos regiones y se procede a repetir el proceso de división en cada una de esas regiones. Este punto se repite con todas las regiones resultantes de cada división.

La partición de un espacio de variables en un serie de rectángulos para el ajuste a un modelo simple para cada una de ellas, es desarrollado por los métodos basados en los árboles de decisión. En cada separación que se establece en los nodos, se realiza una prueba para decidir si esa variable separa correctamente, comparando con un valor constante predeterminado.

Los árboles de decisión que utilizan divisiones tienen una forma de representación muy simple, por lo que es relativamente fácil para que el usuario entender el modelo establecido; pero por otro lado, representan una restricción a la expresividad del modelo. En general, cualquier restricción a una representación de árbol en particular puede restringir significativamente la forma funcional y por lo tanto, la aproximación más certera del modelo. Uno de los algoritmos de crecimiento para los árboles de decisión basado en divisiones univariadas es *ID3* de (Quinlan, 1993) y su versión mejorada *C4.5*.

El algoritmo *ID3* comienza con todas las muestras de entrenamiento en el nodo raíz del árbol. Un atributo se selecciona para dividir estas muestras. Para cada valor del atributo se crea una rama y el correspondiente subconjunto de muestras que tienen el valor del atributo especificado por la rama se mueve al nodo hijo recién creado. El algoritmo se aplica recursivamente a cada nodo hijo hasta que todas las muestras en un nodo pertenezcan a una misma clase. Cada ruta de la hoja en el árbol de decisión representa una regla de

clasificación. La decisión crítica para el algoritmo de generación del árbol de decisión actuando de arriba abajo, es la elección de un atributo en un nodo.

Los atributos de selección de algoritmos *ID3* y *C4.5* se basan en la minimización de la información de la entropía medida aplicada a los ejemplos en cada nodo. El enfoque basado en la entropía de información insiste en minimizar el número de pruebas que permiten a una muestra ser clasificada en el conjunto de datos. La selección de los atributos de ID3 se basa en la suposición de que la complejidad del árbol de decisión está fuertemente relacionada con la cantidad de información que tiene el atributo dado. Una información heurística selecciona el atributo que proporciona la ganancia de información más alta, es decir, el atributo que reduce al mínimo la información necesaria en el subárbol resultante de clasificar la muestra.

Una extensión de *ID3* es el algoritmo *C4.5*, el cual extiende el dominio de la clasificación de los atributos categóricos a los numéricos. La medida favorece a los atributos que resultan en la partición de los datos en subconjuntos que tienen una baja entropía de clase, es decir, cuando la mayoría de los ejemplos pertenecen a una misma clase. El algoritmo básicamente elige el atributo que proporciona el máximo grado de discriminación entre clases a nivel local.

3.4. Criterios y problemas de los árboles de decisión

Uno de los principales problemas que actualmente se tiene en los procesos de investigación es que, a pesar del aumento de datos medidos, también existen datos que pueden faltar por diferentes motivos, como pérdida, error, etc…Por eso, en el proceso de creación del árbol de decisión existen una serie de factores con los que el árbol debe tratar para mejorar. En primer lugar, por su naturaleza para el manejo de datos, es necesario ver el tratamiento que realiza con aquellos que son numéricos y por lo tanto cómo actúa frente a los valores perdidos. También esto conlleva preguntarse cómo el árbol procede con el ajuste frente a los datos, ¿se produce un aumento en el tamaño del árbol? ¿Es conveniente realizar una poda para evitar el sobreajuste? ¿Qué criterios se establecen para la poda?

Una de las formas mediante la que los árboles de decisión abordan el problema de los datos perdidos es tratarlos como un posible atributo más, siendo lo más apropiado cuando dichos datos perdidos no son significativos de alguna manera. El problema surge cuando estos datos perdidos poseen por si mismos una cantidad de información importante para el proceso de formación del árbol.

Algunos algoritmos que trabajan con datos perdidos proponen como solución rellenar los huecos con el valor más probable, o una distribución de probabilidad de todos los valores para el atributo dado. En los árboles C4.5, se acepta que las muestras con valores desconocidos se distribuyen probabilísticamente de acuerdo con la frecuencia relativa de valores conocidos.

a. Ganancia de información

El algoritmo *ID3* utiliza un criterio denominado ganancia para seleccionar el atributo que debe ser probado basándose en el concepto de teoría de la información: la entropía.

Supongamos que tenemos que seleccionar un posible test con *n* resultados (n valores para una determinada característica) que divide el conjunto *T* del conjunto de entrenamiento en subconjuntos $T_1, T_2, \ldots, T_n$. La única información disponible para poder seleccionar es la distribución de las clases en *T* y sus subconjuntos T_i. Si *T* es cualquier conjunto de muestras, y dada la frecuencia $freq(C_i,T)$ determinada por el número de muestras en *T* que pertenecen a la clase C_i (fuera de las k clases posibles), y dado $|T|$ que es el número de muestras en el conjunto *T* entonces el cálculo de la entropía del conjunto viene dado por:

$$Info\ (T) = -\sum_{i=1}^{k}\left(\left(\frac{freq(C_i,T)}{|T|}\right) * \ log_2\left(\frac{freq(C_i,T)}{|T|}\right)\right)$$

Considerando ahora una medida similar después de que *T* ha sido dividido con *n* resultados de un test atributo *X*. El resultado de la información esperada se puede dar como la suma ponderada de las entropías de los subconjuntos:

$$Info_x(T) = \sum_{i=1}^{k} ((|T_i|/|T| * Info(T_i))$$

Por lo tanto la ganancia mide la cantidad de información que se obtiene mediante la partición de *T* de acuerdo con el test *x*. El criterio de ganancia selecciona un test de *X* para maximizar la ganancia (*X*), es decir, este criterio selecciona un atributo con la mayor ganancia de información

$$Ganancia\ (x) = f\left(Info\ (T) - Info_x(T)\right)$$

Dada una cantidad de Información *Info(T)* e *Info$_X$(T)* calculada como se ha descrito anteriormente y en la que se tienen en cuenta sólo las muestras con valores conocidos de atributos, a continuación, el parámetro de ganancia se puede corregir razonablemente con un factor *F*, que representa la probabilidad de que un atributo determinado se conozca:

$$F = \frac{número\ de\ muestras\ en\ la\ base\ de\ datos\ con\ un\ valor\ conocido\ para\ un\ atributo}{número\ total\ de\ muestras\ en\ el\ conjunto\ de\ datos}$$

Del mismo modo, la información en los cortes del árbol, *Split-info (x)*, puede ser alterada por las muestras con valores desconocidos como un grupo adicional de división. Si la prueba de *x* tiene *n* resultados, su *Split-info(x)* se calcula como si la prueba dividiera el conjunto de datos en *n+1* subconjuntos. Esta modificación tiene una influencia directa sobre el valor final del criterio modificado *Ganancia-ratio(x)* y también se pueden aplicar a instancias parciales. Esto implica teóricamente dividir la instancia en partes más pequeñas, utilizando un valor numérico de ponderación, y el envío de parte de la instancia hacia abajo en cada rama en proporción al número de instancias de formación de esa rama de abajo. Con el tiempo, las diversas partes de la instancia llegarán a un nodo hoja, y las decisiones en estos nodos hoja deben ser recombinadas utilizando los pesos que se han filtrado a las hojas.

$$Split - info(X) = -\sum_{i=1}^{k} ((|T_i|/|T|) * log_2(|T_i|/|T|)$$

Esto representa la cantidad de información potencial que se genera dividiendo el conjunto *T* en *n* subconjuntos T_i. Por lo tanto podemos definir la nueva medida de ganancia como

$$Ganancia - ratio\ (X) = \frac{ganancia\ (X)}{Split - info\ (X)}$$

b. Formación del árbol

Otra cuestión es cómo se reparte el conjunto de entrenamiento, una vez que el atributo de división ha sido elegido, para permitir la aplicación recursiva del procedimiento de formación del árbol de decisión en cada uno de los nodos hijo. Se utiliza el mismo procedimiento de ponderación. Las instancias que faltan para aquellos valores relevantes del atributo se dividen teóricamente en partes, una parte para cada rama, en la misma proporción que los casos conocidos se separan por las diversas ramas. Cada parte de la instancia contribuye a las decisiones en los nudos inferiores en la forma habitual a través del cálculo del criterio de ganancia, aunque son ponderadas en consecuencia. Si se conocen todos los valores de atributos, el proceso es sencillo. A partir de un nodo raíz en un árbol de decisión, los test para el valor del atributo determinarán el recorrido a través del árbol y al final, el algoritmo terminará en uno de los nodos de hoja que definen únicamente la clase de un ejemplo de prueba (o con probabilidades, si el conjunto de entrenamiento tenía valores perdidos). Si el valor de un atributo relevante de prueba es desconocido, el resultado de la prueba no se puede determinar. A continuación, el sistema explora todos los posibles resultados de la prueba y combina la clasificación resultante aritméticamente. Puesto que puede haber varias rutas de acceso desde la raíz de un árbol o subárbol a las hojas, una clasificación es una distribución de clase en lugar de una sola clase. Cuando se ha establecido la distribución total de la clase para el caso de la prueba, la clase con la mayor probabilidad se asigna como la clase predicha.

c. Criterio de parada de los árboles

El segundo problema que deben resolver los árboles de decisión es el grado de ajuste del modelo. Los árboles de decisión totalmente expandidos a menudo contienen parte de estructura innecesaria, por lo que en general es aconsejable simplificarlos. La fase de crecimiento de los árboles continúa hasta que se desencadena un criterio de parada. Las siguientes condiciones son las reglas habituales de parada:

- Todas las instancias en el conjunto de entrenamiento pertenecen a un solo valor de *y*.
- Se ha alcanzado la profundidad máxima del árbol.

- El número de casos en el nodo terminal es menor que el mínimo número de casos para nodos padre.
- Si se divide el nodo, el número de casos en uno o más nodos secundarios será menor que el número de casos mínimo para nodos secundarios.
- El mejor criterio de división no es mayor que el valor de un cierto umbral.

El intento por parte del algoritmo de realizar el mejor ajuste conlleva el problema de sobreajuste. El sobreajuste sucede cuando el algoritmo de aprendizaje continúa desarrollando hipótesis que reducen el error del conjunto de entrenamiento a costa de un aumento del error del test. Esto se produce debido a dos razones principales:

- Ruido en los datos. La adaptación de cada rama del árbol a los datos estudiados con un error 0 conlleva un ajuste del modelo al ruido.
- Un número pequeño de ejemplos se asocia con un nodo hoja. En este caso, es muy posible que se seleccione un atributo particular para dividir muy bien los ejemplos con regularidades coincidentes, aunque no hay ninguna conexión real entre ese atributo y la función objetivo.

Dado que se ha establecido como criterio de crecimiento del árbol que cuando se supera cierto umbral con la disminución de la suma de cuadrados cesa la división, esto podría dejar debajo de ella otros puntos que establecen mejores divisiones posteriores. El establecer un criterio para la parada en el crecimiento del árbol es un elemento importante como parámetro, pues el elegir un valor de parada constante demasiado pequeño, puede crear un problema de sobreajuste de los datos. La complejidad de un árbol viene dada por alguno de los siguientes parámetros:

- Número total de nodos
- Número total de hojas
- Profundidad del árbol
- Número de atributos usados

Una vez realizado el árbol se procede a su poda estableciendo criterios basados en el incremento del coste y la disminución de la complejidad. El descarte de uno o más subárboles y su sustitución con hojas simplifica un árbol de decisión, siendo esta la tarea principal en la poda de árboles. Al sustituir el subárbol con una hoja, el algoritmo espera reducir la tasa de error predicho y aumentar la calidad del modelo.

El cálculo de la tasa de error no es simple. Una tasa de error basada solamente en un conjunto de datos de entrenamiento no proporciona una estimación adecuada. Una posibilidad para estimar la tasa de error predicho es el uso de un nuevo conjunto adicional de datos, no usados previamente, o utilizar la técnica de validación cruzada o *cross-validation*. Esta técnica divide las muestras inicialmente disponibles en bloques iguales de tamaño y, para cada bloque, el árbol se construye a partir de todas las muestras excepto dicho

bloque, probándose posteriormente con un bloque dado de muestras. Con las muestras de entrenamiento y de prueba disponibles, la idea básica de la poda de árbol de decisión es eliminar las partes del árbol (subárboles) que no contribuyen a la exactitud de clasificación de las muestras de prueba que no son usadas, produciendo un árbol menos complejo y por lo tanto más comprensible.

Pre-poda y poda

Existen dos enfoques para evitar que el método de partición recursiva, y por lo tanto el sobreajuste, pueda ser agrupado:

- Los enfoques que impiden crecer al árbol antes de que todos los datos se repartan perfectamente, estableciéndose un criterio de parada basado generalmente en algunas pruebas estadísticas, tales como la prueba de χ^2: Si no hay diferencias significativas en la precisión de la clasificación antes y después de la división, se representa el nodo actual como una hoja. Esta decisión se toma previamente a la división y por lo tanto, este enfoque se denomina pre-poda o *prepruning*.
- Enfoques que permiten que el árbol sobreajuste los datos, realizándose posteriormente una post-poda del árbol mediante la eliminación posterior de alguna estructura de árbol utilizando criterios de precisión seleccionados. La decisión en este proceso de post-poda o *postpruning* se hace después de que el árbol ha sido construido.

Estas dos vertientes se han debatido extensamente para establecer el mejor criterio, si bien no existe un gran consenso que evidencie el favor de cualquiera de ellos. *Pre-pruning* puede ser una alternativa viable cuando el tiempo de ejecución es una variable a tener en cuenta en el desarrollo de un estudio. (Hothorn et al., 2006) sugieren que el uso de las pruebas de significación como criterios de parada (es decir, pre-poda) podrían generar árboles que son similares a los árboles con podados óptimos. En su algoritmo *CTREES*, el criterio de parada se combina con la decisión de selección de división, y se formula como prueba de hipótesis. En particular, el atributo más significativo para la división es encontrado por el primero que rechaza la hipótesis H_0 que supone la no existencia de relación entre los atributos de entrada y el atributo de salida. Inicialmente se realiza un test de permutación estadística lineal para todas las variables predictoras. Se rechaza la hipótesis nula cuando el mínimo de los valores de *p*-ajustados (por un multiplicador de Bonferroni) es inferior a un umbral pre-especificado. Si no se rechaza la hipótesis nula, el árbol deja de crecer; si se rechaza el mejor atributo divisor es sugerido. Quinlan (1993) sugiere que no es apropiado utilizar la parada, es decir, los criterios pre-poda durante el crecimiento del árbol, porque esto influye en la calidad del predictor. En su lugar, recomienda en primer lugar el crecimiento del árbol, y después la poda del árbol. El *postpruning* además ofrece algunas ventajas como el efecto de combinación de presentan dos atributos que por separado no contribuyen al

predictor, pero que combinados de manera correcta son muy informativos. El resultado es un componente con mayor varianza del error, que está relacionado con la variabilidad resultante de la aleatoriedad de la muestra de aprendizaje.

C4.5 sigue el enfoque *postpruning*, pero utiliza una técnica específica para estimar la tasa de error predicho. Este método se llama poda pesimista, basado en la idea de que la cantidad del error que se estimó mediante el conjunto de entrenamiento no es lo suficientemente fiable. Este método busca evitar la necesidad de un conjunto de poda o la validación cruzada y utiliza el test de correlación estadística pesimista en su lugar. En este caso, la corrección de continuidad para la distribución binomial podría ser utilizada:

$$\varepsilon'(T,S) = \varepsilon(T,S) + \frac{|hojas(T)|}{2|S|}$$

El problema resultante de esta corrección es la aparición de un error de una tasa optimista, diferentes experiencias demostraron que el método de la poda de *C4.5* puede producir una innecesaria estructura adicional en el árbol final. El tamaño del árbol seguiría creciendo cuando más instancias sean añadidas a los datos de entrenamiento, incluso aunque esto no aumentara el rendimiento, por lo que Quinlan (1993) propuso realizar podas en los nodos internos *t* si esta tasa de error está dentro de un error estándar del árbol de referencia:

$$\varepsilon'(podado(T,t) \leq \varepsilon'(T,S) + \sqrt{\frac{\varepsilon'(T,S) * (1-\varepsilon'(T,S))}{|S|}}$$

Esta última condición se basa en el intervalo de confianza estadístico.
Siendo ε el error estándar del árbol y dado un *T* que es un subárbol cuya raíz es el nodo interno *t* y *S* es la parte de la conjunto de entrenamiento que se refiere al nodo *t*, se procederá a eliminar la rama convirtiéndolo en un nodo hoja siempre que el error estándar del árbol podado sea menor o igual que el valor obtenido para el error estándar del subárbol original sobre el conjunto del test.

El procedimiento de la poda pesimista realiza el recorrido de arriba hacia abajo sobre los nodos internos. Si se poda un nodo interno, entonces todos sus descendientes se eliminan del proceso de poda, lo que resulta una poda relativamente rápida.

Independientemente del método que se utilice, una pregunta clave es qué criterio se debe utilizar para determinar que el tamaño del árbol final sea el correcto. Desde la perspectiva del aprendizaje a partir de los flujos de datos, es difícil pensar en un tamaño final del árbol de decisión, ya que el proceso es

más bien dinámico y la fase de evaluación está intercalada con la fase de aprendizaje. El hecho de que haya una gran cantidad de datos de entrenamiento hace que sea muy poco probable que el sobreajuste pueda ocurrir debido a modelar datos con regularidades coincidentes. Por lo tanto, la única razón para que exista sobreajuste es la posible existencia de ruido en los datos.

3.5. Ventajas de los árboles de decisión

Las principales ventajas de los árboles de decisión son:

A. Los árboles de decisión, como se ha señalado anteriormente, requieren unas simples especificaciones para su funcionamiento:

   - División en 2 preguntas
   - Una regla de selección de la mejor división en cualquier nodo
   - Un criterio para elegir el árbol del tamaño adecuado

B. Los árboles son una poderosa y flexible herramienta para el tratamiento de datos.

   - Permite que sean aplicados a cualquier estructura de datos mediante un planteamiento adecuado de las preguntas a responder
   - Además, se pueden tratar datos categóricos y ordenados de manera simple. También hace un uso de información condicional en grupos no-homogéneos de relaciones de manera que, una vez que un nodo T se divide en estos dos nodos t_l y t_r (siendo t_l el nodo izquierdo y t_r es el nodo derecho) entonces se buscan de forma individual para la división más significativa.
   - Incluso hace una reducción automática paso a paso de la complejidad y de las variables. Esto es inherente a la estructura de crecimiento de los árboles dado que en cada nodo se busca la división más significativa. En este sentido, se asemeja a un procedimiento por pasos en lugar de un método del mejor de los subconjuntos. En cada etapa se intenta extraer la información más relevante de la parte del espacio que está trabajando.
   - Sin ningún costo computacional añadido establece una estimación de la probabilidad de errores de clasificación para el objeto, además de la clasificación del mismo.
   - Es invariante a las transformaciones operacionales realizadas a las variables ordenadas con una estructura de datos estándar.
   - Es muy robusto con respecto a los valores atípicos y mal clasificados.
   - Las representaciones de los árboles ofrecen información de fácil entendimiento e interpretación con respecto a la estructura predictiva de los datos.

El significado del árbol de "*tamaño adecuado*" y la exactitud de predicción óptima son los dos temas más importantes de un predictor de aprendizaje basado en árboles. Desde el punto de vista del aprendizaje como búsqueda, los algoritmos de aprendizaje del árbol realizan un movimiento desde lo simple hacia lo complejo. La búsqueda establece una escalada que comienza con un árbol vacío, considerando progresivamente hipótesis más elaboradas. El algoritmo mantiene una única hipótesis actual, ya que explora el espacio de todos los árboles posibles. Una vez que se selecciona una prueba de la división

y se refina la hipótesis actual, nunca retrocede para reconsiderar de nuevo esa opción. Por lo tanto, este método de búsqueda no tiene capacidad para determinar cuántos árboles de regresión alternativos son consistentes con los datos de entrenamiento disponibles. Por consiguiente, aunque eficiente, este procedimiento dará como resultado el primer árbol que se ajusta a los datos de entrenamiento y es susceptible de converger hacia soluciones óptimas locales.

4. Selección de modelos

4.1. Introducción

Actualmente se utilizan modelos estadísticos tanto para la explicación como para la predicción, y necesitan técnicas que sean lo suficientemente flexibles para expresar características típicas de sus datos, tales como no linealidades e interacciones.

La metodología empleada en el desarrollo de esta tesis doctoral, se ha basado la utilización de los árboles de regresión aplicados a los conjuntos de datos agronómicos obtenidos durante seis años en las que se aplicaron diferentes dosis de compost de biosólidos a diversas rotaciones de cultivos. Se buscan modelos con alto grado de correlación con los datos experimentales y dotados de un número de ramas adecuadas de forma que nos permita establecer comparaciones entre los distintos grupos estudiados. Para ello se han realizado modificaciones de los parámetros de desarrollo de los árboles de regresión, intentando obtener un consenso entre un grupo de valores suficientes para la explicación del mismo pero sin que esto implique una excesiva complicación. Además, se ha estudiado la posibilidad de combinar varios grupos de datos (como parámetros del suelo, parámetros del residuo, cultivos utilizados, etc…), lo cual conlleva una mayor precisión en las predicciones, incluyendo una reducción de la variabilidad y una mayor sensibilidad sobre los datos de entrenamiento seleccionados para validar los diferentes modelos.

De las diferentes metodologías aplicables a la elaboración de árboles de decisión, hemos decidido realizar la experiencia con aquellos que se encuentran disponibles en el software *Waikato Environment for Knowledge Analysis* (WEKA). WEKA es un software de libre distribución desarrollado en Java con un conjunto de algoritmos de aprendizaje automático para tareas de Minería de Datos. Está constituido por una serie de paquetes de código abierto con diferentes tareas como el procesamiento previo de los datos, clasificación, agrupamiento, asociación, visualización, así como facilidades para su aplicación y análisis de prestaciones cuando son aplicadas a los datos de entrada seleccionados. WEKA es un *software* de código abierto publicado bajo la Licencia Pública General de GNU. (Hall et al., 2009). WEKA es un entorno para experimentación de análisis de cualquier conjunto de datos. Para ello únicamente se requiere que los datos a analizar se almacenen con un cierto formato, conocido como *Attribute-Relation File Format* (*ARFF*). Para facilitar la utilización por los usuarios WEKA incluye una interfaz gráfica para acceder y configurar las herramientas disponibles sin necesidad de trabajar sobre la máquina virtual.

Weka dispone de una gran cantidad de métodos de clasificación y regresión, entre ellos nos hemos centrado en los árboles por su fácil interpretación y la posibilidad de utilizar grandes conjuntos de datos.

El árbol se desarrolla, según la conocida como "la división jerárquica" "partición-recursiva", "divisoria de grupo" o "segmentación". Inicialmente se toman todos los datos para analizar y se realiza una posterior división en dos grupos repetidas veces hasta acabar el análisis.

En este trabajo se va a analizar cuáles de los árboles de decisión de los que pueden trabajar con datos tanto numéricos como categóricos, son los que mejores resultados aportan en los distintos parámetros de salida. Estos parámetros son:

1. El coeficiente de correlación
2. La posibilidad de representación del árbol obtenido
3. El tamaño del árbol
4. El tipo de modelo que resuelve (modelos clasificatorios o de regresión).

Para poder realizar el proceso de evaluación hemos procedido al ajuste de los datos mediante la binarización de aquellos datos que son categóricos.

A continuación, se han evaluado los algoritmos de aprendizaje con el conjunto de datos. En este caso se han elegido 5 algoritmos y una variación. Esta última se realiza ya que el algoritmo de *M5P* no sólo establece un modelo lineal si no que permite la opción de realizar árboles de regresión, dado que es nuestro objeto de estudio, aunque se muestran los resultados de los dos tipos, *M5P* regresión lineal y árbol, para comprobar cuáles son los mejores según lo establecido en las premisas iníciales.

4.2. Árboles seleccionados para el estudio

- ***DecisionStump***. Es un algoritmo donde se produce una división binaria simple que genera un árbol de decisión, son árboles que sólo hacen una división a nivel de la raíz del árbol (Iba and Langley, 1992). El gasto computacional frente a los árboles de múltiples niveles es menor ya que no necesita preparar los datos para los siguientes pasos y además, el modelo de salida es más manejable, al tener solamente una división. Para manejar los datos nulos crea una tercera rama que nace del nodo raíz donde se clasifican todos los datos inclasificables. Este tipo de árbol puede realizar tanto clasificación como regresión, realizando esta última con el *RMSE*, que es la raíz cuadrada del error cuadrático medio producido, (Witten *et al.*, 1999).
- ***REPTree*** es un árbol que aprende tomando decisiones rápidas. Construye un árbol de regresión comprobando la varianza y la ganancia de información para realizar la división en los nodos. Y posteriormente lo poda usando sistemas de reducción de error en el podado. Está optimizado para resolver el problema rápidamente, separando valores para atributos numéricos sólo una vez. Además, en el caso de valores nulos funciona como el algoritmo *C4.5*, dividiendo las instancias correspondientes en grupos más pequeños. (Witten and Frank, 2005)
- ***RandomTree*** Es un algoritmo para la construcción de árboles, tomando una clase, que considera *k* atributos seleccionados al azar en cada nodo. Lleva a cabo la construcción del árbol sin poda. También tiene una opción que permite la estimación de las probabilidades de clase (o el

objetivo promedio en caso de regresión) basado en un conjunto de validación independiente (*backfitting*).

- ***M5P***. Es un algoritmo desarrollado por (Witten and Wang, 1997) como mejora del creado por Quinlan (Quinlan, 1992). En *M5P* se realiza una regresión lineal en todas las hojas del árbol de decisión, donde intervienen como variables todos los atributos que participan en las decisiones de división de los nodos que forman esa rama del árbol. M5P combina un árbol de decisión convencional con la posibilidad de funciones de regresión lineal en los nodos.
- ***Alternating Model Tree (AMTree)***. Es un algoritmo que crea un árbol modelo de alternancia mediante la minimización del error cuadrático. Los árboles modelo son árboles de decisión con modelos de regresión lineal múltiple en los nodos de la hoja. Al igual que en *Alternating Decision Tree* (Freund and Mason, 1999) para la clasificación, los *Alternating Model Tree* para la regresión contienen nodos divisores y de predicción, y utiliza funciones de regresión lineal simple en oposición a los predictores constantes en los nodos de predicción. Por otra parte, se aplica la regresión aditiva hacia adelante mediante el uso de modelos escalonados para hacer crecer el árbol en lugar de un algoritmo de refuerzo. El tamaño del árbol se determina mediante validación cruzada. Los resultados empíricos muestran que *Alternating Model Tree* logran un error cuadrático significativamente menor que los árboles modelo estándar en varios conjuntos de datos de regresión (Frank et al., 2015).

4.3. Conjunto de datos (*Dataset*) para el estudio

El conjunto de datos de suelo de la experiencia de Valdespinar se han tomado para el desarrollo del proceso de Minería de Datos buscando los valores de las variables estudiadas, mediante los modelos más simples y de más fácil interpretación. En la tabla 4.1 se muestran las variables que se han estudiado para realizar el estudio de selección de los modelos (propiedades del suelo y de los metales pesados en el suelo). También se muestran las variables categóricas que se han incluido así como las variables del biosólido añadido. En el esquema de la finca experimental (Figura 4-1) se puede observar el diseño con 5 parcelas (numeradas desde 1 a 5) con 5 tratamientos:

- Control
- Inorgánico
- Compost 1 (dosis baja 3,5 t/ha)
- Compost 2 (dosis media 7,5 t/ha)
- Compost 3 (dosis alta 17,5 t/ha)

El número de atributos tomados para el desarrollo de los árboles han sido 74 con un *data set* de N=450. Dentro de estos atributos se han incluido los datos de los años precedentes obtenidos como variables para ver la influencia que puede existir debido las adiciones anteriores en los cultivos actuales o en el suelo. Se han transformado alguna de estas variables desde nominal a binarias

para que los árboles puedan ser realizar separaciones en solamente dos nodos.

Estas son obtenidas de la siguiente manera:

- 3 medidas para cada muestra
- 5 tipos de tratamientos
- 5 réplicas por tratamiento
- 6 años experimentación
- Total: 3*x*5*x*5*x*6=450.

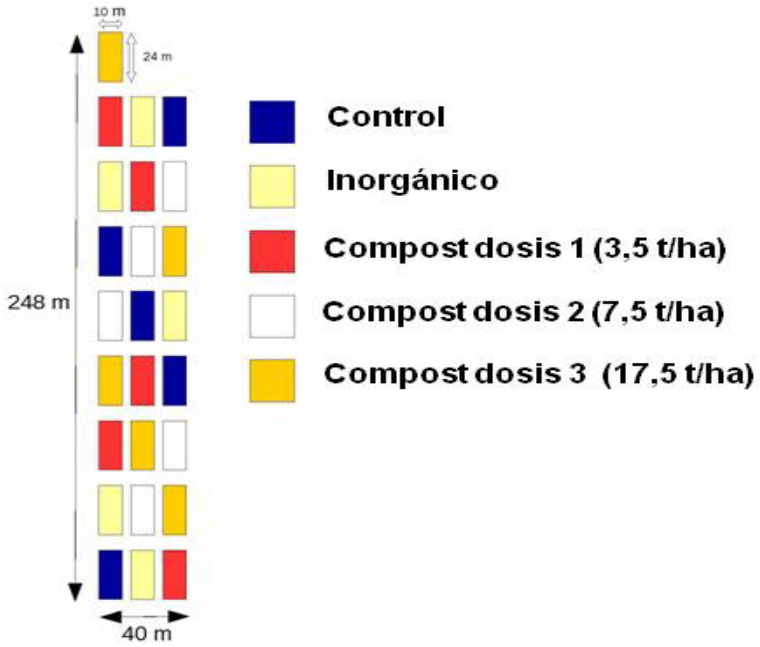


Figura 4-1 Esquema de la parcela de estudio.

Tabla 4-1: Variables utilizadas en el estudio.

Descripción	Nomenclatura	Nº de variables
Propiedades del suelo	pH-suelo, CE-suelo, MO-suelo, Nt-suelo, P-suelo, Na-suelo, K-suelo, Mg-suelo, Ca-suelo	9
Metales pesados del suelo	Cd-suelo, Cu-suelo, Cr-suelo, Fe-suelo, Mn-suelo, Ni-suelo, Pb-suelo, Zn-suelo	8
Propiedades del compost de biosólidos	pH-biosólido, CE-biosólido, MO-biosólido, Nt-biosólido, P-biosólido, Na-biosólido,K-biosólido, Mg-biosólido, Ca-biosólido, Cd-biosólido, Cu-biosólido, Cr-biosólido, Fe-biosólido, Mn-biosólido, Ni-biosólido, Pb-biosólido, Zn-biosólido	17
Otros parámetros	Tiempo desde última Adición, número de Adición, Adición, Parcela	4
Otros parámetros 2	Actual, Previo, Previo2,Fertilización, Temporada	29

4.4. Resultados y discussion

A continuación se muestran los coeficientes de correlación y los errores medios absolutos (MAE) de los distintos árboles probados para la selección de los más adecuados para el tratamiento de los conjuntos de datos de este trabajo.

Tabla 4-2: Tabla de características de los árboles estudiados (ej. variable cadmio).

Algoritmo	Representación	Tamaño	Modelo
AMTree	Sí		Lineal
DecisionStump	No		
M5P	No	1	Lineal
***M5P* (*RegressionTree*)**	Sí	22	Lineal
RandomTree	Sí	683	Numérico
REPTree	Sí	51	Numérico

Tabla 4-3: Coeficientes de correlación para las variables de estudio.

Coef.Corr.	AMTree	DecisionStump	M5P	M5P Regres.	RandomTree	REPTree
MO	**0,8833**	0,6696	0,8435	0,8049	0,8153	0,8093
Nt	0,9059	0,6416	**0,9347**	0,8559	0,8518	0,8965
P	**0,9356**	0,6910	0,9177	0,8340	0,8655	0,8348
Cd	0,9513	0,8845	**0,9622**	0,8929	0,9137	0,9354
Cr	0,9654	0,6448	**0,9744**	0,8816	0,9199	0,9082
Cu	0,9046	0,4421	**0,9476**	0,8536	0,8767	0,8780
Fe	0,7879	0,7102	**0,8312**	0,7622	0,7306	0,7722
Mn	0,8674	0,5362	**0,8903**	0,7567	0,7286	0,7347
Ni	0,9296	0,8003	**0,9434**	0,9107	0,8685	0,9090
Pb	0,8887	0,5979	**0,8950**	0,8309	0,7782	0,8079
Zn	0,8256	0,8340	**0,9448**	0,8620	0,8556	0,9096

Se muestran en negrita los coeficientes más elevados para cada variable.

Tabla 4-4: Error medio para las variables de estudio.

MAE	AMTree	DecisionStump	M5P	M5P Regres.	RandomTree	REPTree
MO	**0,1136**	0,195	0,137	0,1547	0,1448	0,1522
Nt	0,0154	0,0297	**0,013**	0,0203	0,0188	0,0166
P	**8,5453**	18,0529	9,4235	13,3476	9,9529	11,7246
Cd	0,1566	0,2448	**0,1473**	0,258	0,1926	0,1823
Cr	1,9279	5,9132	**1,6907**	3,7711	2,8377	3,0139
Cu	0,8791	2,2379	**0,7681**	1,3479	1,1027	1,148
Fe	1118,911	1321,8519	**1027,841**	1208,6435	1294,7453	1201,3544
Mn	17,7903	33,3581	**15,842**	25,2189	25,9207	25,6403
Ni	0,8182	1,4074	**0,7883**	1,0664	1,1398	1,0169
Pb	1,3803	2,5333	**1,3357**	1,7663	1,925	1,825
Zn	5,9198	9,0021	**4,3634**	7,6649	6,1773	5,5799

Se muestran en negrita los errores más bajos para cada variable.

En el caso del cadmio se ve que el algoritmo *M5P* es el mejor en el 81% de los casos y el coeficiente de mejora del algoritmo *M5P* frente al siguiente mejor está entre un 0,92-12,62%. En los casos en que resulta mejor otro algoritmo, ese es el *AMTree*, frente al segundo mejor que es M5P, con coeficientes entre 1,91 y 4,51% mejores. El árbol de regresión M5P creado se compone de numerosas hojas finales. Los valores de cada una de las hojas siguen una regresión lineal que da el valor final de la variable que cumple las condiciones de la hoja. Al realizarse una regresión lineal en cada hoja, es posible conocer el valor de las nuevas instancias a través de ecuaciones lineales.

Los resultados obtenidos por el algoritmo Decision Stump son los más bajos. Esto es lógico debido a que se trata de un algoritmo que genera un árbol bastante pequeño y no separa correctamente las variables de la base de datos que se ha utilizado en el ensayo.

Dado que nuestro interés se basa en encontrar modelos para el estudio cuyos algoritmos permiten la representación (*AMTree, M5P Reg, RandomTree* y *REPTree)* hemos realizado una serie de gráficos buscando un equilibrio entre una buena resolución, con mejores resultados de coeficientes de correlación, y un gráfico de fácil interpretación. Comparando los resultados de los coeficientes de correlación de las variables vemos que entre esos 4 algoritmos existe una mayoría de veces que *AMTree* es el que mejores coeficientes ofrece. Cuando comparamos los otros tres algoritmos existe una alternancia entre unos y otros para el coeficiente de correlación (6 de las 11 variables para *REPTree*).

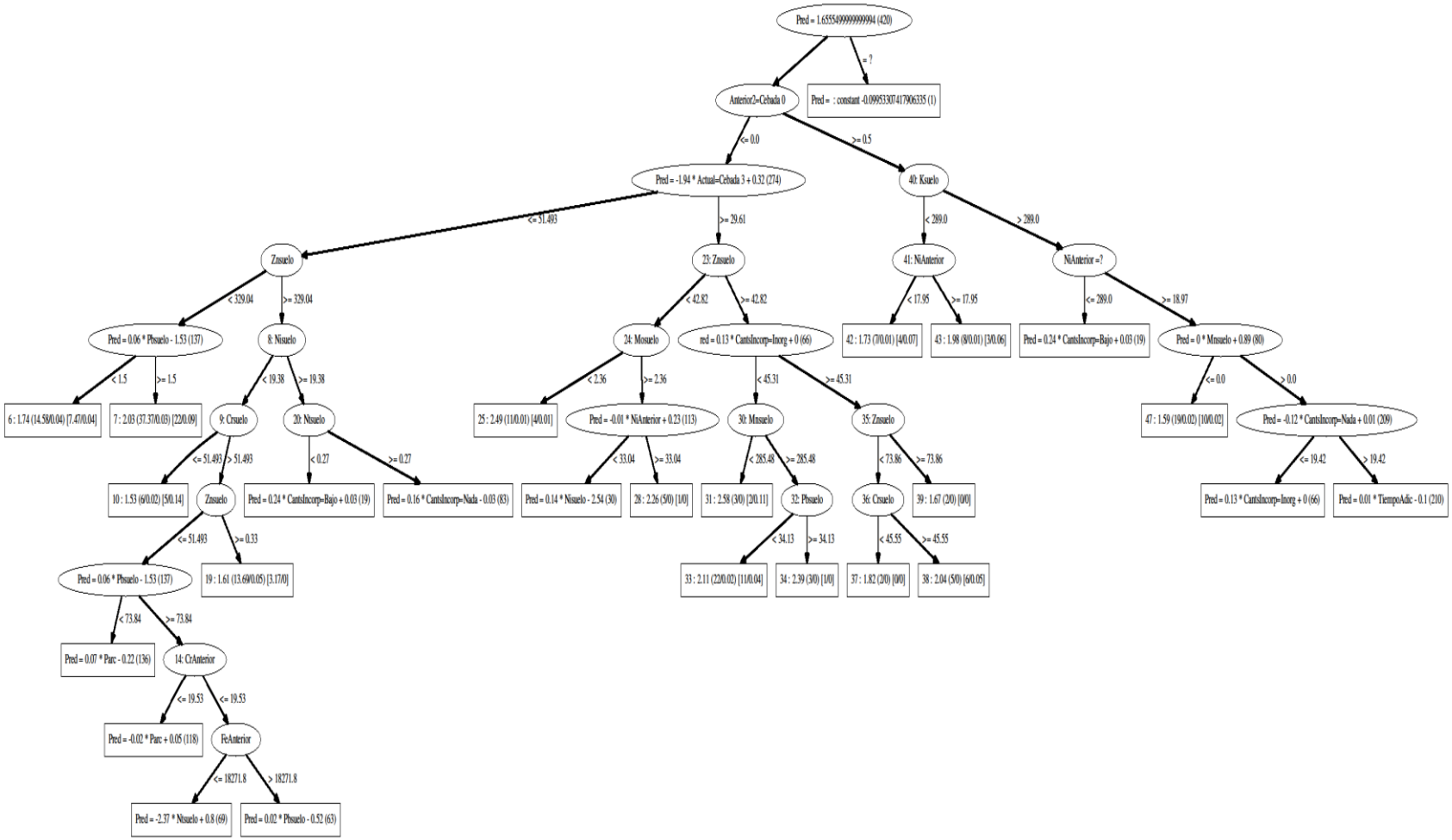


Figura 4-2 *AMTree.*

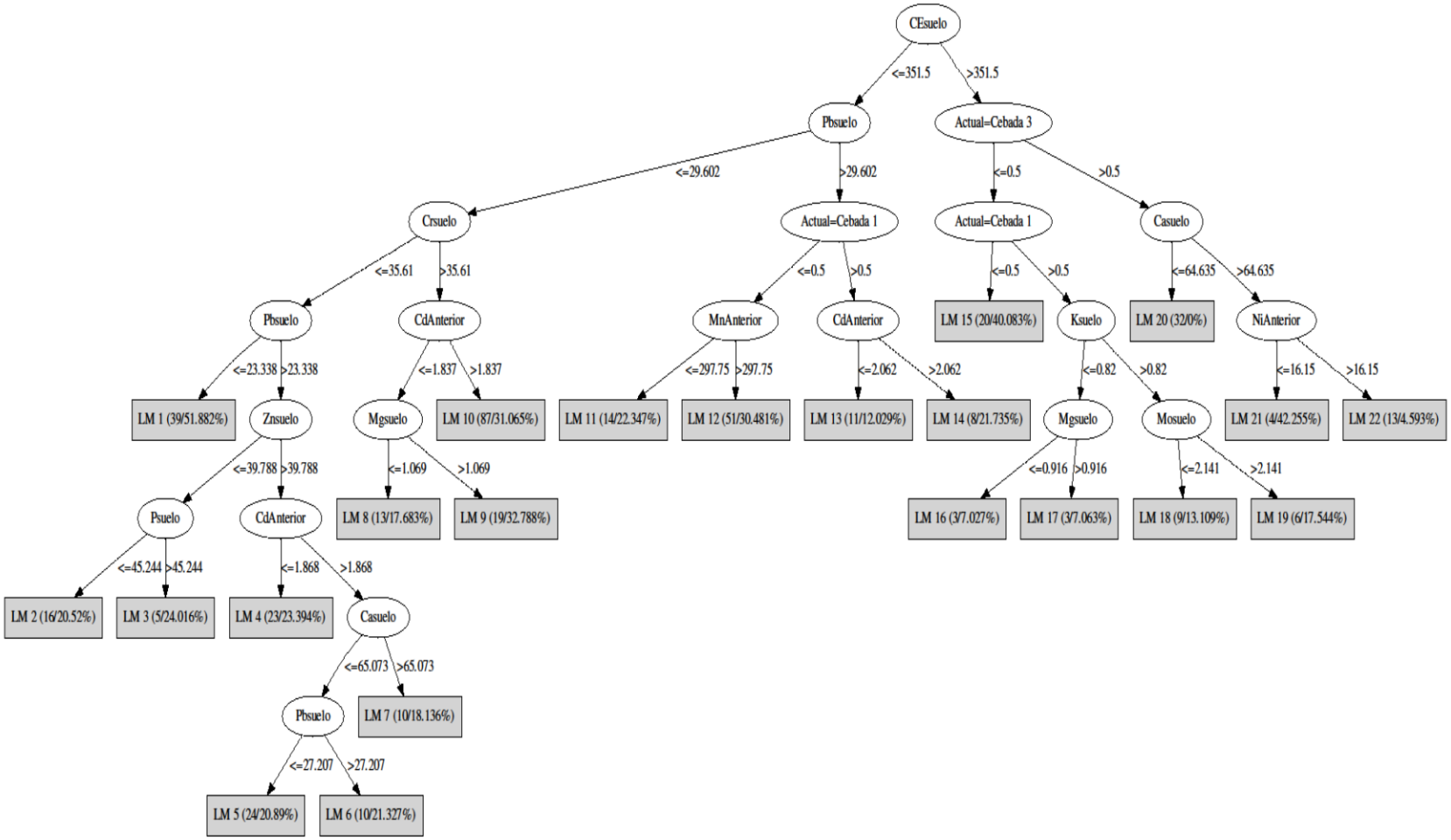


Figura 4-3 *M5P RegressionTree.*

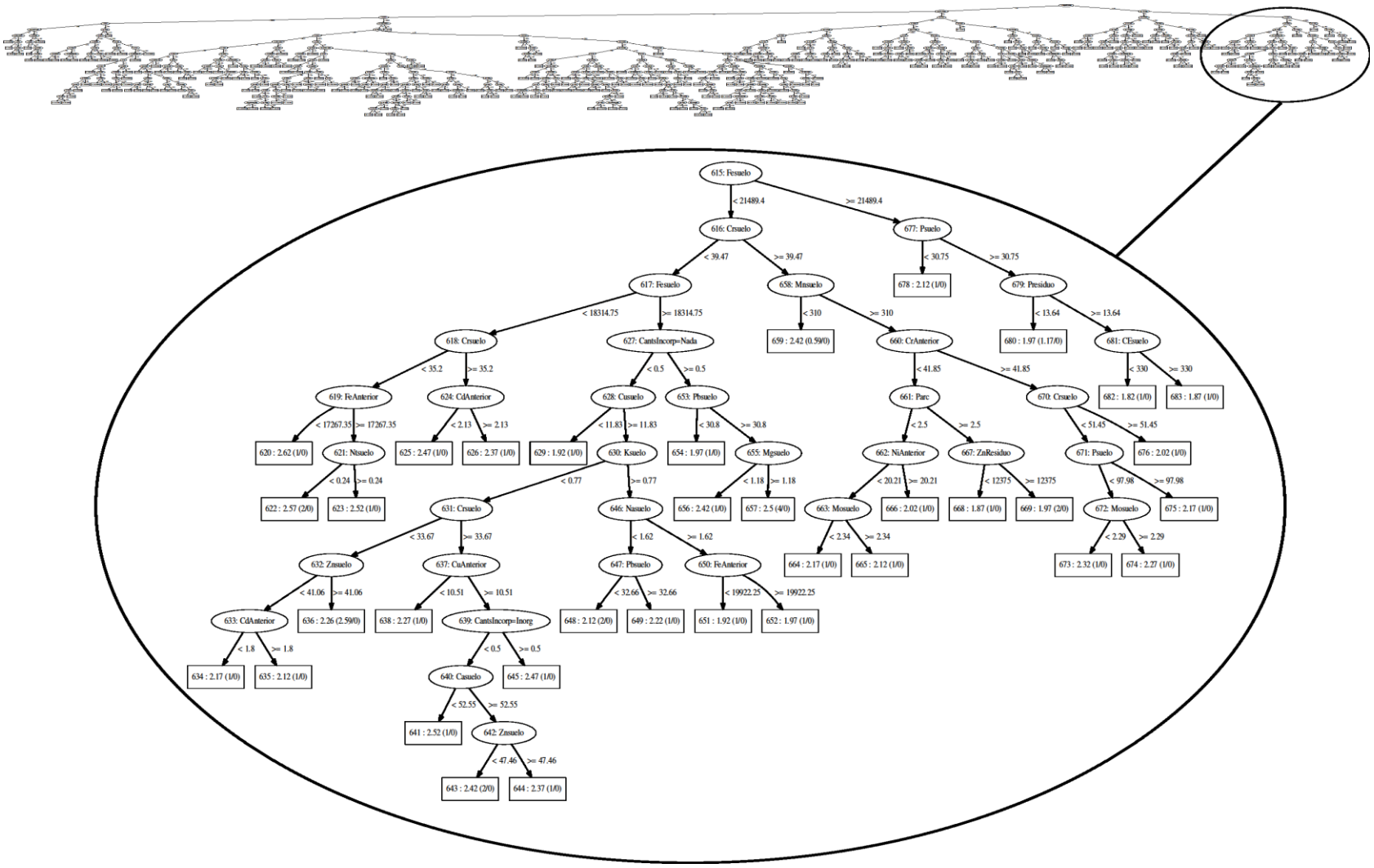


Figura 4-4 *RandomTree.*

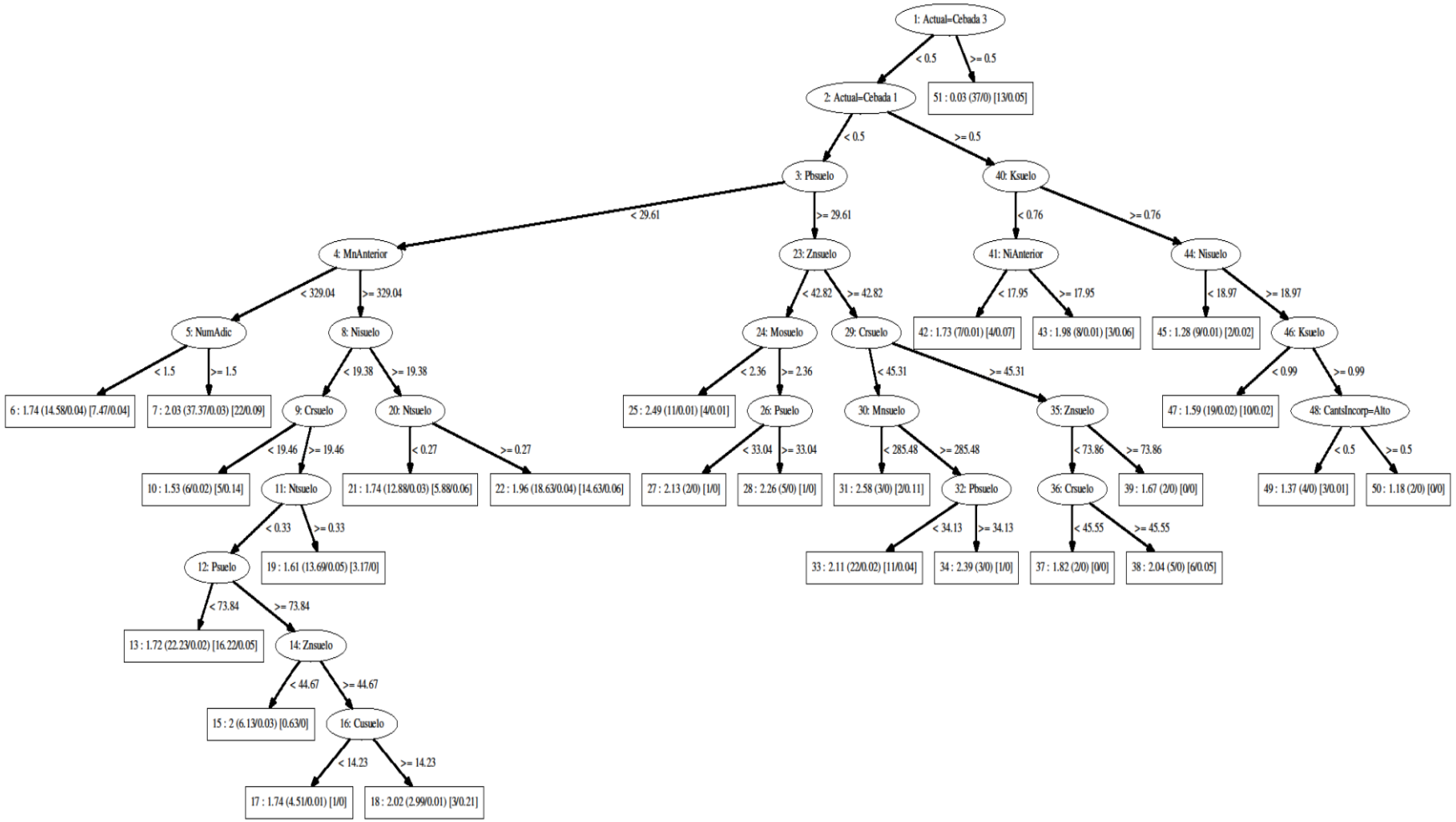


Figura 4-5 *REPTree*.

Se puede observar que es el algoritmo *AMTree* (Fig. 4-2) el que logra representaciones con los árboles que permiten una mejor interpretación, pero tiene el inconveniente de que la salida de los datos se realiza mediante un modelo lineal con múltiples variables y no permite una representación gráfica adecuada. Esta cuestión también afecta a los árboles que se obtienen mediante *M5P* en la opción de *Regression tree*, (Fig. 4-3) aunque cuando se usa sin esta opción, se obtienen los mejores resultados en el coeficiente de correlación. *RandomTree, (*Fig. 4-4*)* ofrece unos resultados gráficos con demasiadas ramas, lo cual dificulta la interpretación de los árboles representados. No obstante *RandomTree* no está pensado para usarse individualmente sino formando una combinación de regresores (*ensemble)* con multiples *RandomTrees*. Por otro lado observamos cómo, de los árboles representados, el algoritmo *REPTree* (Fig. 4-5) obtiene buenos resultados en cuanto a la cantidad de coeficiente de correlación reflejado por el modelo y su representación. Nos deja un árbol con una interpretación que no resulta excesivamente complicada.

En la representación de los coeficientes de correlación frente a los tamaños de las variables de cada algoritmo (Fig. 4-6), vemos una separación clara de los árboles por tamaño dado que aquellos que se conforman con el algoritmo *RandomTree* proporcionan árboles muy complejos (tamaños por encima de 500) para el fácil uso explicativo que se pretende hacer de ellos.

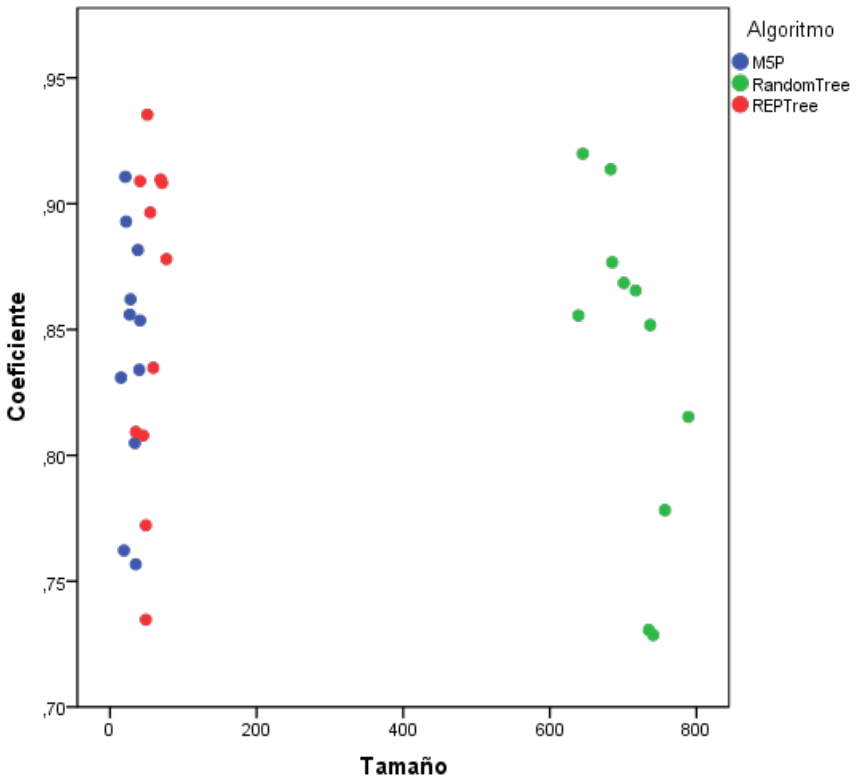


Figura 4-6 Gráfica de los coeficientes de correlación con los tamaños de las variables para cada algoritmo.

Sin embargo, al comparar en la Fig. 4-7 los algoritmos *REPTree* y *M5P*, vemos cómo, aunque el algoritmo *REPTree* aporta un poco más de complejidad a los árboles resueltos con los datos (tamaño de árbol mayor) los coeficientes se encuentran, en algunos casos, por encima de los obtenidos por el algoritmo *M5P*. Otro inconveniente que plantea el uso del algoritmo *M5P* (*Regression*) para los datos es la interpretación de de los árboles generados ya que son el resultado de regresiones lineales que podrían generar salidas con numerosas variables en cada nodo. Por lo tanto, partiendo de este algoritmo (*REPTree*), procederemos a una serie de modificaciones en los parámetros para, sin una pérdida excesiva de información, establecer un consenso entre la complejidad y la máxima explicación de las variables.

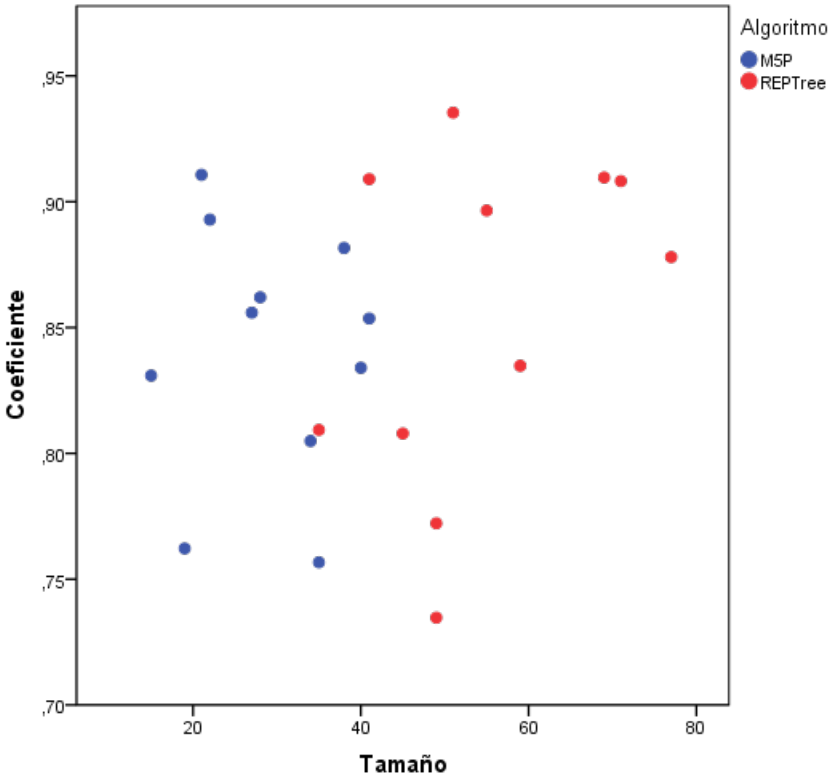


Figura 4-7 Gráfica de los coeficientes de correlación con los tamaños de las variables para algoritmos *REPTree* y *M5P*.

Las múltiples posibilidades que ofrece el algoritmo *REPTree* en la modificación de sus parámetros nos lleva a buscar un consenso entre complejidad y grado de satisfacción en la resolución. Para esto se realizan variaciones para que, con una pequeña pérdida de la determinación del coeficiente de correlación, se consiga la disminución del tamaño del árbol, lo cual redunda en una mejora en la visualización y la interpretación del mismo.

Para ello vamos a partir de nuestra base de datos y compararemos los valores obtenidos cuando se realiza un *REPTree* sin restricciones, un árbol sin límite

impuesto de ramas o de profundidad, con los obtenidos con variaciones en los parámetros que este algoritmo permite modificar.

La estructura y las predicciones de un árbol de decisión dependen del valor de los parámetros que regulan el proceso de construcción del árbol. En el caso de *REPTree,* estos parámetros son (Tabla 4-5):

- La profundidad máxima es la máxima profundidad del árbol.
- El peso mínimo definido como la suma mínima de peso de todos los casos en cada hoja; el algoritmo no permite crear una hoja con un grupo de instancias cuya suma de pesos es inferior a este número. Esto se debe a que en un caso extremo, se podría hacer un árbol con tantas hojas como de los casos, estas hojas corresponderían a la separación de todos los casos dando un árbol con valores puros y sin mezclas entre instancias, pero el árbol no tendría capacidad de generalización. En *REPTree* el peso de cada instancia es igual a 1 por lo que se puede hablar de número mínimo de instancias en cada hoja para simplificar.
- Datos para la poda es la cantidad de datos utilizados para la poda, esta cantidad se define en términos del número total de casos como el denominador de la fracción. Por ejemplo 3 significa que el número de instancias utilizadas para la poda es un tercio del total, 5 quinta y así sucesivamente.

Tabla 4-5: Parámetros modificables en el algoritmo *REPTree*.

Parámetro	**Valores**
Profundidad máxima	1,2,3 y 4
Mínimo peso de las instancias	8,16,24 y 32
Datos para la poda (en términos del total de instancias. Denominador de la fracción.)	3,5,7 y 9

En las siguientes gráficas (Figs. 4-8 y 4-9) se muestran los coeficientes de correlación y los errores absolutos medios de los árboles resultantes con las variables de estudio con las profundidades elegidas comparadas con el árbol sin restricciones. Se muestran en las gráficas las combinaciones de los parámetros descritos anteriormente y los valores que por defecto (*Sin rest*) ofrece el programa *Weka* para el algoritmo *REPTree*. Así, por ejemplo, vemos en el gráfico de coeficientes de correlación vs tamaño cómo la combinación de parámetros mínimo peso de instancias 32 y datos para la poda 3 con las cuatro profundidades probadas da unos resultados muy bajos.

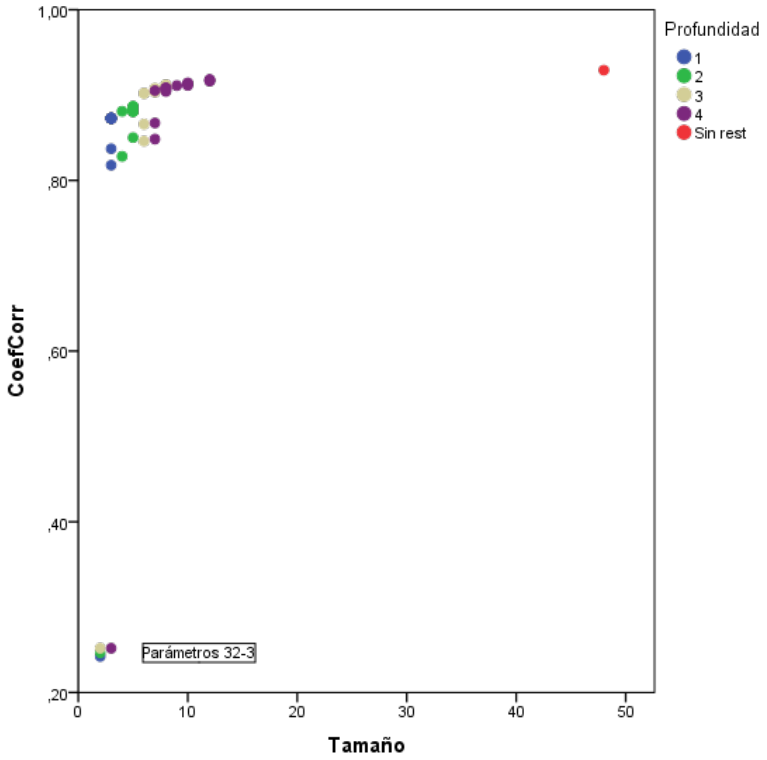


Figura 4-8- Ejemplo de gráfica con variaciones de parámetros del coeficiente de correlación del algoritmo *REPTree* para Cd.

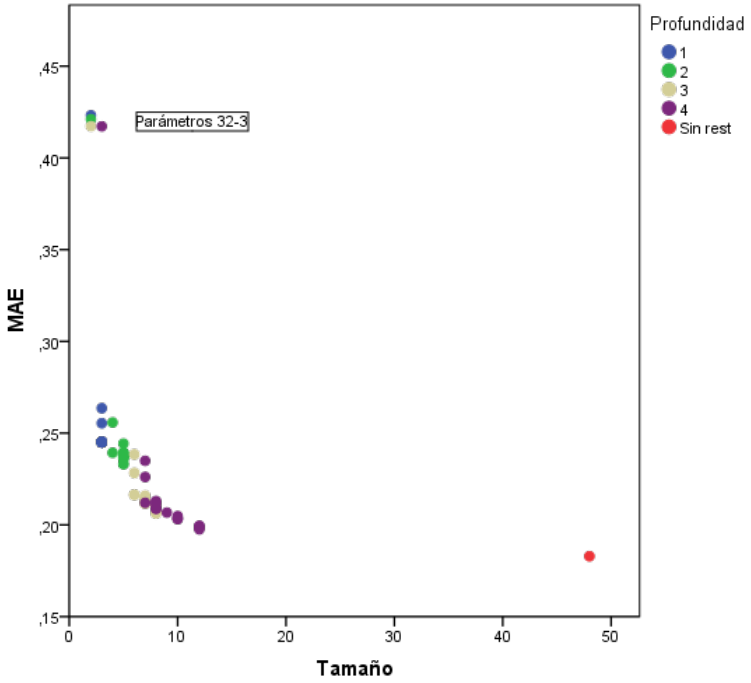


Figura 4-9 Ejemplo de gráfica con variaciones de parámetros del error medio del algoritmo *REPTree* para Cd.

Mediante estas gráficas se observa que existen una serie de parámetros que permiten la reducción del tamaño del árbol sin una gran pérdida de los

coeficientes de correlación y sin un aumento excesivo de los errores. Por otro lado, algunas de las combinaciones de los parámetros no tienen una capacidad suficiente para el desarrollo de un óptimo con las exigencias establecidas. En este caso, el principal problema de esta combinación es que toma un número de instancias excesivas para crear las hojas lo que conlleva una falta de exactitud pues se depositan demasiados datos en cada una y no proporciona un árbol con suficiente resolución. Por lo tanto, deberemos realizar un conjunto de experimentos con todos los parámetros que hemos definido anteriormente para encontrar el grupo de aquellos que permite una mejora en el coeficiente de correlación.

4.5. Conclusiones

La elección de árboles de decisión para una visión de todas las variables (tanto numéricas como categóricas) implicadas en un ensayo de campo durante sucesivos años permite una resolución de problemas complejos como el presente, además de ser una herramienta que no necesita muchas transformaciones ni tiene restricciones por el tipo de datos utilizados.

Parece razonable seleccionar el algoritmo *REPTree* para futuros ensayos e investigaciones por sus buenos resultados en cuanto a coeficientes de correlación, bajos errores y posibilidades de representación.

Un ajuste de los parámetros del algoritmo *REPTree* permite una obtención de árboles suficientemente robustos y con grados aceptables en los valores establecidos para nuestros propósitos (correlación, errores…).

5. Estudio de las fincas experimentales

5.1. Introducción

La realización de investigaciones en el ámbito de la agricultura conlleva una serie de problemas adicionales relacionados con los modelos estadísticos aplicables al análisis de los datos experimentales obtenidos, debido a que no siempre se cumplen los supuestos de normalidad y homogeneidad en las varianzas que exigen los análisis parámetricos. También es frecuente encontrar que los experimentos de campo están sometidos a determinadas variables ambientales que no pueden ser controladas por los investigadores: temperaturas excesivas en olas de calor, descenso en las temperaturas mínimas por debajo de los valores críticos para la planta, disminución en la cantidad de lluvia, etc…, elementos todos ellos que no se pueden obviar en la interpretación final de los resultados, por lo que han de ser incluidos o comprobar su influencia sobre las variables que han sido establecidas y medidas.

Aunque el uso agrícola es la utilización tradicional de los biosólidos o lodos generados en las estaciones depuradoras de aguas residuales (EDARs) (Singh et al., 2011), son pocos los modelos existentes que permitan la predicción de sus efectos en el suelo, en especial en lo referente a su contenido metálico, o su papel en la fertilidad y productividad de los diferentes cultivos (de Melo et al., 2007, Gabrielle et al., 2005). La Minería de Datos constituye una técnica estadística que mediante la aplicación de algoritmos matemáticos, muestra su utilidad en el descubrimiento de patrones internos de comportamiento en las bases de datos que puedan utilizarse como modelos. En la actualidad, la Minería de Datos está mostrando su utilidad en campos tan variados como la sociología, los estudios ambientales, la medicina, las ingenierías, etc., si bien su aplicación al estudio de sistemas agronómicos ha sido hasta el momento escasa. Este trabajo aborda la aplicabilidad de la Minería de Datos al estudio de la predicción de la evolución del sistema suelo-planta cuando se aplican al suelo fertilizantes orgánicos como son los composts de biosólidos (Westerman and Bicudo, 2005). El uso de Minería de Datos ha sido previamente propuesto como una herramienta para examinar las relaciones entre parámetros en este tipo de conjunto de datos complejos. (Cortet et al., 2011)

Actualmente el desarrollo de la tecnología de procesamiento de datos permite procesar una gran cantidad de información obtenida del desarrollo de experimentos complejos, como son las experiencias agronómicas. La Minería de Datos (Witten et al., 2011) trata de inferir ese conocimiento subyacente, partiendo de los datos experimentales y estableciendo modelos más complejos que permitan prever futuras actuaciones y tendencias. Los métodos de aprendizaje automático desarrollan y utilizan algoritmos que acceden al conocimiento usando herramientas estadísticas y computacionales y describen modelos satisfactorios para la predicción de valores futuros. Dentro de estas técnicas se encuentran las clasificatorias, en las que la instancia de salida es una clase, y las regresoras, donde el valor de salida es de tipo numérico.

5.2. Localización y procedimiento experimental

En este capítulo se han comparado varios modelos predictivos que determinan cuáles de los parámetros estudiados son las variables más importantes en la determinación de la fertilidad del suelo: materia orgánica, nitrógeno total y fósforo, y en la evaluación del riesgo ambiental, como es la variación en el contenido de metales pesados del suelo. Previamente, y tal como se ha explicado en el capítulo anterior, se ha realizado una adecuación de los datos para el entrenamiento de los modelos.

La experiencia agronómica se ha desarrollado en tres fincas experimentales localizadas en el término municipal de Villafruela (Burgos), Valdespinar 41º54'13'' N y 3º52'41'' O, Andadilla 41º56'16.77'' N y 3º56'5.35'' O y Serrana 41º56'15.94'' N y 3º50'58.78'' O perteneciente a la comarca del Arlanza, y caracterizadas por una pluviometría media anual de 457,7 mm y una temperatura media de 10,8 ºC. El suelo predominante es un *Cambisol Calcárico (CMc)* de escasa profundidad (25 cm) y textura franco-arcillosa dedicado al cultivo en secano. En ellas se ha realizado durante 6 campañas una rotación de cultivos cereal-oleaginosa-proteaginosa típica de la zona, con adición en años alternos de un compost de biosólidos procedente de la EDAR de Burgos.

El diseño experimental desarrollado en las fincas de Valdespinar, Andadilla y Serrana (Figura 5-1, Figura 5-2 y Figura 5-3) ha consistido en el establecimiento de 25 sub-parcelas de 10*x*24 m^2 y de 12*x*24 m^2, respectivamente, distribuidas en bloques incompletos no balanceados, sobre los que se introdujeron cinco tratamientos y cinco réplicas por tratamiento: control (C), fertilización inorgánica (I) y dosis crecientes de compost de lodo de 3,5; 7,5 y 17,5 t/ha (L_1, L_2 y L_3, respectivamente). Tras un cultivo precedente de cebada, los cultivos desarrollados en la finca de Valdespinar fueron: girasol, cebada, veza, cebada, guisante y cebada. En el caso de la finca de Andadilla, con cultivo precedente también de cebada, la secuencia de cultivos realizada fue: guisante, trigo, girasol, cebada, cebada y veza. Para la tercera de las fincas, denominada La Serrana, el diseño experimental consistió también en 25 sub-parcelas, en este caso de 12*x*48 m^2. Después del primer año de cultivo, cada una de las parcelas se subdividieron en dos; en una de estas submitades se continuará con la adición de las dosis correspondientes de compost de biosólidos de forma continua y en las otras sólo en años alternos. En esta tercera finca experimental, el cultivo precedente fue guisante, los cultivos desarrollados fueron: cebada, guisante, cebada, girasol, cebada y cebada.
Los muestreos de suelo se realizaron al finalizar la cosecha, mediante la extracción de 20 núcleos de suelo en cada subparcela a una profundidad de 0-20 cm, que fueron posteriormente secados al aire y tamizados a 2 mm. Los análisis físico-químicos que se realizaron en estas muestras compuestas de suelo fueron: contenido de humedad, pH (1:5, agua), conductividad eléctrica (1:5 agua 25 °C), nitrógeno total por combustión en un analizador elemental, fósforo asimilable (fósforo Olsen), materia orgánica (por vía húmeda), cationes de cambio y determinación de metales pesados. En el caso de los metales una

muestra de suelo se secó en estufa a 105 ºC, fue triturada en molino de bolas de ágata y tras digestión en ácido nítrico concentrado, su contenido metálico se determinó mediante espectrometría de absorción atómica de llama F-AAS. En planta se determinó el contenido en metales sobre materia seca a 60 ºC utilizando un procedimiento similar al descrito. En el caso de las muestras de grano de cada cosecha, se determinó además el contenido en nitrógeno y carbono totales de una muestra triturada pero sin secar. Todos los análisis se han realizado utilizando agua ultrapura tipo II, con reactivos de calidad PA y utilizando los métodos oficiales de análisis de suelos del (MAPA, 1994). Una descripción detallada de la metodología se encuentra en el anexo II siendo posible consultarlos con mayor detalle en la tesis de (Peña, 2013) de título "Capacidad fertilizante y riesgo metálico asociados a la utilización de residuos orgánicos en agricultura bajo diferentes condiciones de cultivo en la provincia de Burgos".

Las unidades de cada una de las variables aparecen en la tabla

Tabla 5-1: Variables utilizadas en el estudio y unidades de medida (sobre base seca).

Elemento	Unidades	Elemento	Unidades	Elemento	Unidades
pH		Conductividad Eléctrica (CE)	(dS m^{-1})	Materia Orgánica (MO)	(%)
Nitrógeno Total (Nt)	(%)	Fósforo (P)	mg Kg^{-1}	Sodio (Na)	mg Kg^{-1}
Potasio (K)	mg Kg^{-1}	Magnesio (Mg)	mg Kg^{-1}	Calcio (Ca)	mg Kg^{-1}
Cadmio (Cd)	mg Kg^{-1}	Cromo (Cr)	mg Kg^{-1}	Cobre (Cu)	mg Kg^{-1}
Hierro (Fe)	mg Kg^{-1}	Manganeso (Mn)	mg Kg^{-1}	Níquel (Ni)	mg Kg^{-1}
Plomo (Pb)	mg Kg^{-1}	Zinc (Zn)	mg Kg^{-1}		

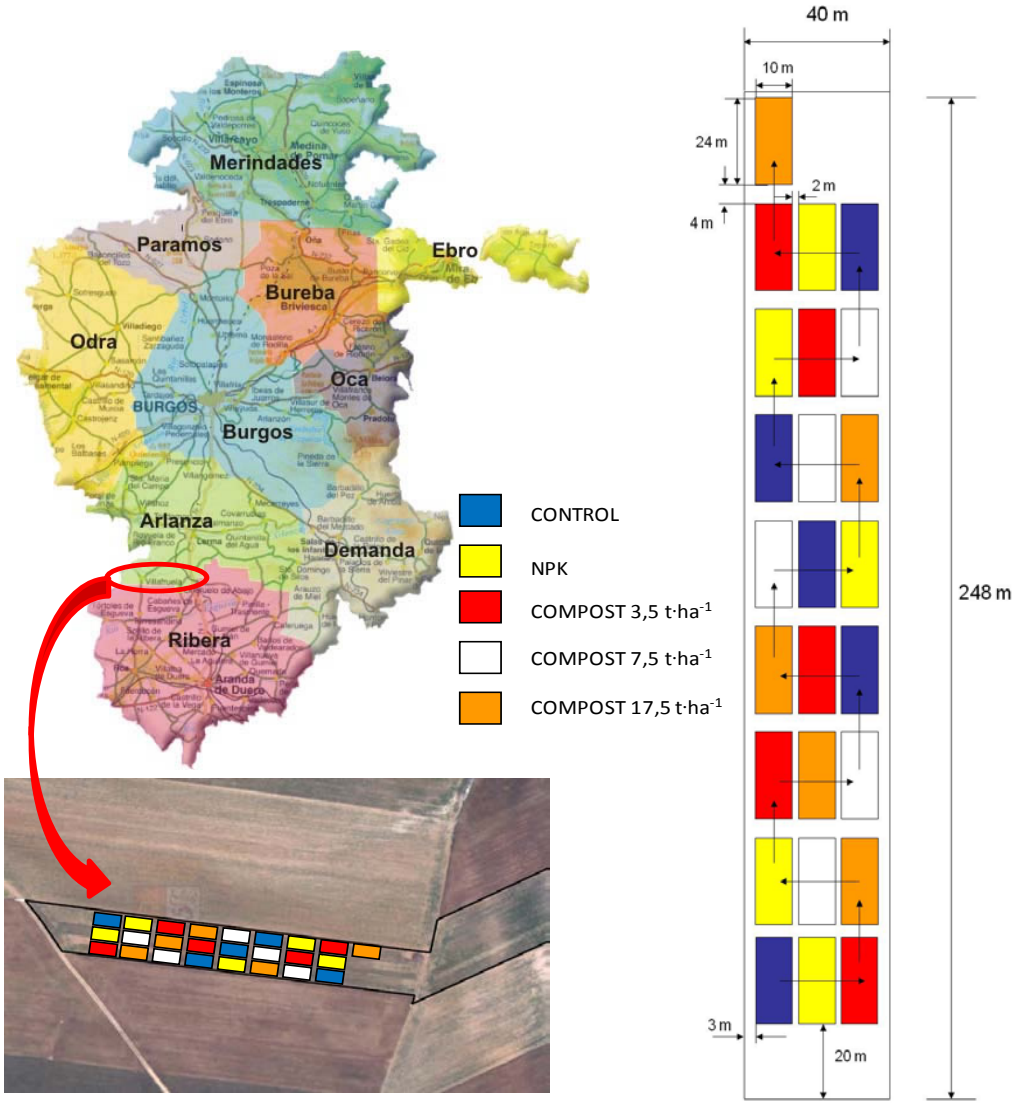


Figura 5-1 Esquema de la finca experimental de Valdespinar.

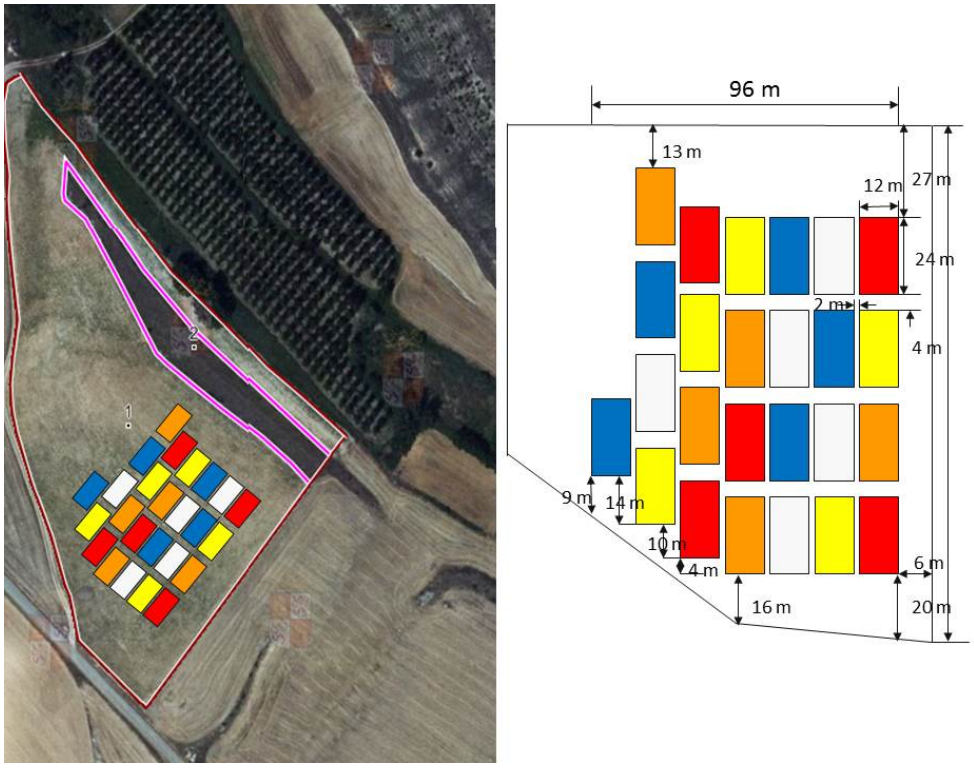


Figura 5-2 Esquema de la finca experimental de Andadilla.

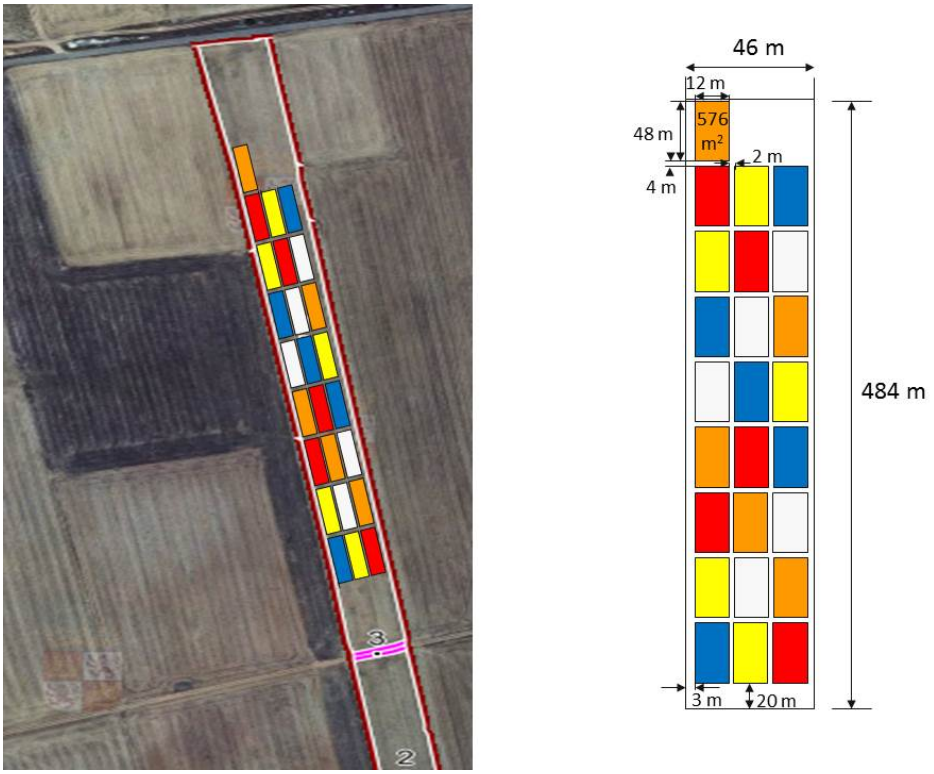


Figura 5-3 Esquema de la finca experimental de Serrana

5.3. Análisis estadístico mediante Minería de Datos

El objetivo del estudio es predecir los resultados de la adición de los biosólidos en distintos escenarios, por lo que se seleccionaron una serie de variables descriptivas de las propiedades de los suelos, tanto en cuanto a su fertilidad, como a la evolución de metales pesados y elementos traza, como del biosólido, del fertilizante inorgánico añadido y de los cultivos.

Se han utilizado el conjunto de datos correspondiente a los análisis de las muestras de suelo y planta recogidos durante las 6 campañas agrícolas realizadas en las tres fincas experimentales. El diseño experimental desarrollado consta de 5 tratamientos: control (C), fertilización inorgánica (I), y tres tipos de dosis de compost de biosólidos, baja (L1 3,5 t ha^{-1}), media (L2 7,5 t ha^{-1}) y alta (L3 17,5 t ha^{-1}), así como 5 réplicas por tratamiento.

Podemos establecer tres tipos de escenarios con objetivos diferenciados. En primer lugar, el escenario correspondiente al estudio de las propiedades relacionadas con la fertilidad del suelo. En este caso los parámetros a predecir serán el contenido de materia orgánica, el nitrógeno total y el fosforo disponible presente en el suelo, así como la producción obtenida tras la cosecha. El número de atributos tomados para el desarrollo de los árboles han sido 74 con un *data set* de N=450 para las fincas experimentales, analizadas por separado. Dentro de estos atributos se han incluido como variables, los datos obtenidos de los años precedentes para ver efecto acumulativo que pudieran tener las adiciones anteriores sobre los cultivos actuales o las propiedades del suelo.

Los dos escenarios siguientes tienen como objetivo predecir el efecto bioacumulativo de los metales pesados introducidos en el compost de biosólidos y su posible efecto sobre la calidad y la seguridad alimentaria de los cultivos; para ello se han medido las concentraciones de estos metales presentes en la raíz de las plantas y en el grano. En este caso los atributos han sido 82 y 92 respectivamente, con un número total de datos N =450. Cuando se realiza el estudio conjunto de las tres fincas experimentales se incluyó otra variable denominada “Lugar”, que incluyera la posible influencia de la localización de las fincas en las variables estudiadas. De este modo cuando se agrupan las tres parcelas el conjunto de datos es de N= 1350 con un número de atributos de 88 para las variables del suelo, 96 atributos para las variables de la raíz y de 106 para las relacionadas con el grano.

Con objeto de predecir la productividad se han utilizado las propiedades del suelo ya mencionadas, la concentración de metales pesados del suelo, las características del biosólido y otros atributos como la última adición, adiciones acumulativas, etc… La variable adiciones acumulativas tiene valores de sí o no, el tiempo desde la última adición es de 12 o 24 meses (dependiendo si coincide con una adición anual o no).

Cuando se realiza el estudio sobre la productividad, el cultivo actual, el cultivo anterior, o la temporada, no se usan porque queremos comparar tanto la productividad dentro de cada año en función de los diferentes tratamientos para

cada uno de los cultivos, así como las productividades globales, que mostraron variaciones importantes como consecuencia de las características climáticas de la campaña agronómica, lo que se muestra como un factor limitante que está siempre presente en una agricultura de secano.

El análisis estadístico se realizó mediante la utilización de árboles de regresión y su posterior análisis de significación mediante la realización de diez validaciones cruzadas. Este tipo de validación divide el conjunto de los datos iniciales en 10 subconjuntos, es decir, se toman todos los datos divididos en 10 grupos aleatorios. De estos subconjuntos se seleccionan 9 para entrenar el modelo dejando el subconjunto no utilizado para calcular el error de muestra. Este proceso se repite 10 veces usando cada uno de los subconjuntos de los que se dispone. Posteriormente, se calcula el error como la media aritmética de cada uno de los valores de los 10 errores que se han obtenido de las muestras parciales. De esta manera se tienen los valores de la media, el máximo, el mínimo y la desviación estándar de los 10 errores.

Para las técnicas de Minería de Datos se ha utilizado el software gratuito WEKA utilizando como herramienta de cálculo, el programa *REPTree*. WEKA es un *software* de código abierto publicado bajo la Licencia Pública General de GNU. (Hall et al., 2009) con un conjunto de algoritmos de aprendizaje automático para la realización de tareas relacionadas con la Minería de Datos. *REPTree* es un tipo de árbol de aprendizaje rápido, lo cual supone un bajo gasto computacional, caracterizado por aprender tomando decisiones rápidas.

Su proceso de actuación está optimizado para una resolución rápida del problema, elaborando en primer lugar un árbol de regresión, mediante el análisis de la varianza y la ganancia de información, lo cual se consigue mediante la división en los nodos. Posteriormente, realiza podas usando sistemas de reducción de error, seleccionando aquellas ramas en las que los errores se reducen más. Solamente separa valores para atributos numéricos una vez. Los valores perdidos son tratados dividiendo las instancias correspondientes en segmentos, igual que en el algoritmo *C4.5*.

El pseudocódigo de la operación básica de *REPTree* se muestra a continuación en los algoritmos 1 y 2 (Figura 5-3 y 5-4):

Para hacer crecer el árbol, el algoritmo encuentra el atributo que maximiza la ganancia de la varianza, dividiendo posteriormente los datos en dos nuevos nodos basados en el atributo seleccionado. Este paso se realiza de forma recursiva hasta que se alcanza alguna de estas condiciones de parada:

a. el nodo no contiene suficientes casos (al menos 2 × *minNum*), donde *minNum* es un parámetro del método;

b. el nodo es puro, es decir, los datos que llegan al nodo son homogéneos, o en otras palabras, la desviación estándar de los datos en el nodo es una pequeña proporción de la desviación estándar de todos los datos;

c. la profundidad máxima del árbol, *maxDepth*, se alcanza, donde *maxDepth* es también un parámetro del método.

Data: Dataset D with a set of attributes A

Result: A REPTree decision tree

if *use pruning* **then**

  Split D into training data D_t and pruning data D_p;

else

  Training data $D_t = D$;

end

Build a tree using D_t and A as shown in Algorithm 2;

if *use pruning* **then**

  Reduce error pruning using D_p;

end

Figura 5-4 Algoritmo 1: algoritmo REPTree.

a. Crecimiento

El algoritmo encuentra en cada nodo la mejor característica discriminante entre la clase y los demás atributos, dividiendo los datos en dos nuevos nodos basados en la característica elegida. Este paso es realizado recursivamente sobre los grupos de datos que van resultando hasta que puede realizarse una asignación de la clase en cada hoja. El proceso se repite hasta que la desviación estándar obtenida es una pequeña proporción de la desviación estándar de los datos iniciales (Fig. 5-5).

Data: Dataset D and a set of attributes A
Result: A decision tree $Tree$
if *no stop condition is reached* **then**
  Compute splitting criterion, SC(D,a_i), for each attribute $a_i \in A$;
  Find the best attribute a_b according to the splitting criterion;
  Using a_b, split D in n subsets;
  if *max* SC > 0 **and** $n > 1$ **then**
    foreach *of the n subset of* D_i **do**
      $Tree$ = BuildTree using D_i and A;
    end
  end
else
  $Tree$ = Create a leaf using D;
end

Figura 5-5 Algoritmo 2: algoritmo construcción del árbol (REPTree).

b. Poda

Una vez creado el árbol, se realiza un podado de éste, eliminando las ramas menos útiles, lo que supone realizar un balance entre la precisión y la complejidad del modelo.

Los parámetros que se pueden modificar y que contribuyen a una mejora en los resultados obtenidos para *REPTree* son:

- la profundidad máxima del árbol,
- la cantidad de datos usados para la poda (uno es usado para la poda y el resto para el crecimiento de las reglas)
- el mínimo peso total de las instancias en la hoja.

En nuestro estudio los parámetros ajustados son:

- Profundidad del árbol entre 1 y 4 con variaciones de 1 unidad.
- El mínimo número de instancias en la hoja varía entre 8 y 32 en intervalos de 8.
- Número de datos usados para la poda oscila entre 3 y 11 con variaciones de 2 unidades.
- De todos los árboles creados se eligieron aquellos con una correlación más elevada y se representaron para su posterior interpretación.

c. Interpretación

Los árboles de regresión que se obtienen deben ser interpretados según la pregunta "*qué ocurre si*" realizada a partir de la raíz de cada árbol; cuanto más cercana está la propiedad a la raíz del árbol, mayor es su influencia sobre la variable predicha por el modelo. En cada uno de los gráficos resume de forma eficaz la combinación de los mejores parámetros para encontrar el mejor coeficiente de correlación.

5.4. Resultados

Una vez construidos y entrenados los modelos, se obtuvieron los diferentes árboles de regresión. En la tabla se muestran los coeficientes de correlación seleccionados en cada variable del suelo estudiada para cada una de las fincas experimentales por separado y agrupadas.

Tabla 5-2: Coeficientes de correlación de los árboles de regresión obtenidos con las variables del suelo para las diferentes fincas experimentales y su agrupación.

Variables Suelo	**Coeficientes de correlación**			
	Valdespinar	Andadilla	Serrana	Tres Parcelas
Materia Orgánica	0,8004	0,4400	0,3514	0,6076
Nitrógeno	0,8760	0,5686	0,6900	0,6691
Fósforo	0,8240	0,8715	0,6869	0,7684
Cadmio	0,9243	0,8981	0,9037	0,8761
Cromo	0,8395	0,7578	0,8283	0,7559
Cobre	0,8463	0,8103	0,8431	0,8101
Hierro	0,7400	0,7525	0,8289	0,8496
Manganeso	0,7467	0,7772	0,7521	0,8899
Níquel	0,8917	0,8148	0,8233	0,8311
Plomo	0,8093	0,8566	0,8309	0,8296
Zinc	0,9054	0,8071	0,8162	0,8394

Tabla 5-3: Coeficientes de correlación de las de los árboles de regresión obtenidos con las muestras de raíz en las diferentes fincas experimentales y su agrupación.

Variables Raíz	Coeficientes de correlación			
	Valdespinar	Andadilla	Serrana	Tres Parcelas
Cadmio	0,7343	0,8154	0,4705	0,8064
Cromo	0,7901	0,9306	0,7608	0,8602
Cobre	0,8317	0,8019	0,6778	0,7754
Hierro	0,9041	0,9333	0,9187	0,9175
Manganeso	0,9353	0,9457	0,9532	0,9261
Níquel	0,8557	0,7597	0,8298	0,7526
Plomo	0,9015	0,9319	0,6906	0,9006
Zinc	0,8055	0,6758	0,3695	0,6760

Tabla 5-4: Coeficientes de correlación de los árboles de regresión obtenidos con las muestras de grano en las diferentes fincas experimentales y su agrupación.

Variables Grano	Coeficientes de correlación			
	Valdespinar	Andadilla	Serrana	Tres Parcelas
Carbono Total	0,9749	0,9496	0,9981	0,9655
Nitrógeno Total	0,9742	0,9206	0,9878	0,9473
Cadmio	0,7969	0,2956	0,3087	0,7633
Cromo	0,8232	0,7881	0,7771	0,8162
Cobre	0,8628	0,8101	0,9283	0,8596
Hierro	0,9195	0,1032	0,8421	0,8970
Manganeso	0,9051	0,9457	0,9325	0,9244
Níquel	0,8524	0,5623	0,9132	0,8411
Plomo	0,7710	0,6227	0,9457	0,8294
Zinc	0,8150	0,5985	0,9476	0,8448

Debido al gran número de árboles de regresión creados y estudiados, se ha decidido introducir buena parte de los mismos como material adicional. Esta subsección cubrirá sólo algunos de los ejemplos más relevantes obtenidos en las parcelas de estudio. El resto de los árboles y su interpretación se encuentran como material complementario en el anexo I.

Materia Orgánica en Valdespinar

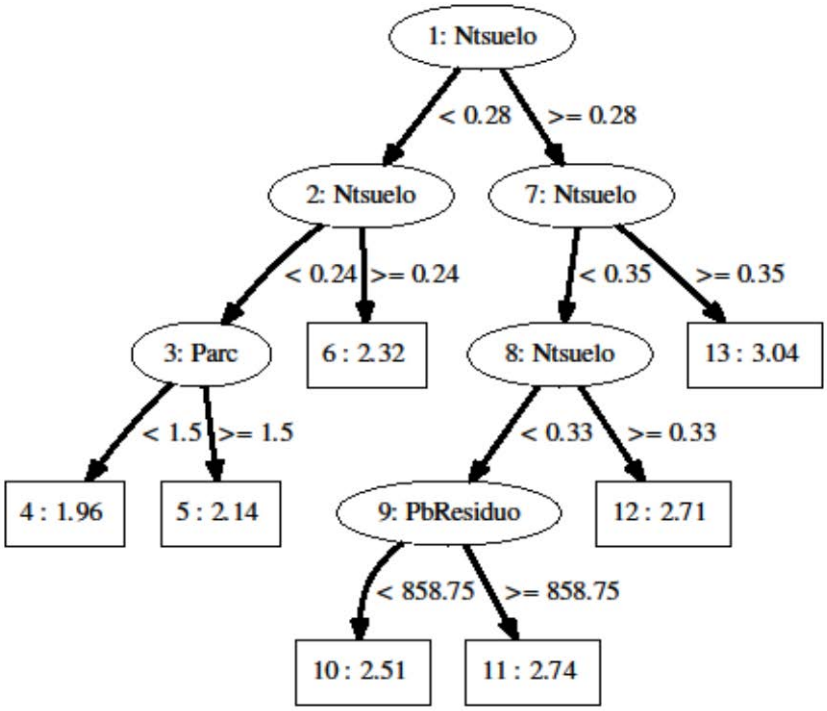


Figura 5-6 Árbol de regresión obtenido de materia orgánica del suelo en Valdespinar.

Parámetros del árbol. Profundidad máxima=4, peso mínimo=8, datos para la poda=3. Coeficiente de Correlación = 0,8004.

La interpretación del árbol de regresión obtenido y que aparece en la Figura 5-6, refleja la correlación que existe entre la materia orgánica del suelo y la cantidad de nitrógeno total que se encuentra presente en el suelo, llegando a alcanzar un valor máximo de MO de 3,04% sobre base seca. Una disminución del contenido de nitrógeno se refleja también en menores contenidos de materia orgánica en el suelo. También vemos cómo un incremento de la cantidad de plomo del residuo tiene influencia sobre ese contenido de materia orgánica en el suelo.

Materia Orgánica en Andadilla

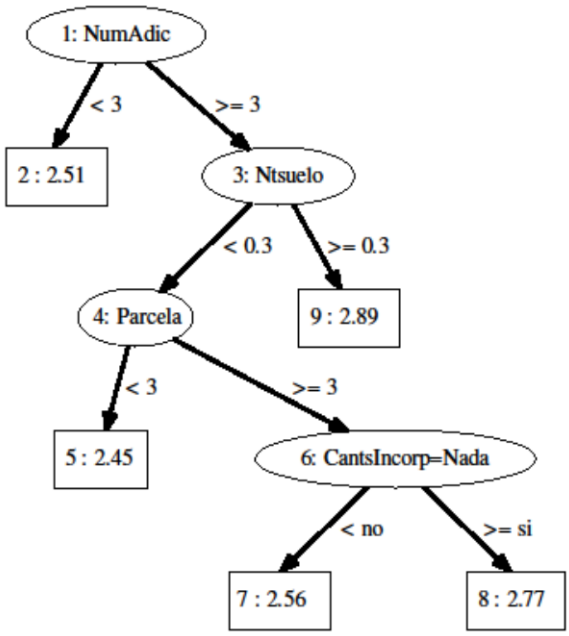


Figura 5-7 Árbol de regresión obtenido de materia orgánica del suelo en Andadilla.

Parámetros del árbol. Profundidad máxima=4, peso mínimo=8, datos para la poda=11. Coeficiente de Correlación = 0,4400.

Como en el caso anterior, los resultados obtenidos en con los datos de la Finca de Andadilla (Fig. 5-7) muestran que la materia orgánica se encuentra muy influida por la cantidad de nitrógeno total que se encuentra presente en el suelo, así como con el número de adiciones de biosólido realizadas, alcanzando el mayor valor de MO (2,89 %) cuando se completan tres adiciones. El menor valor también parece asociado a los menores contenidos de nitrógeno en el suelo (< 0,3 %).

Materia Orgánica en Serrana

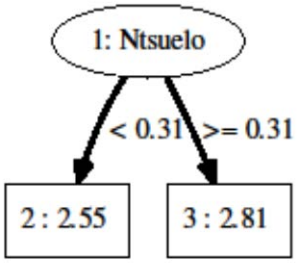


Figura 5-8 Árbol de regresión obtenido de materia orgánica del suelo en Serrana.

Parámetros del árbol. Profundidad máxima=4, peso mínimo=8, datos para la poda=3. Coeficiente de Correlación = 0,3514.

Sin embargo, en los datos de la Finca La Serrana (Fig. 5-8), la única variable que muestra influencia en la cantidad de materia orgánica es la cantidad de nitrógeno en el suelo. Con valores superiores a 0,31% de nitrógeno se encuentra un máximo de materia orgánica de 2,81%.

Nitrógeno total en Valdespinar

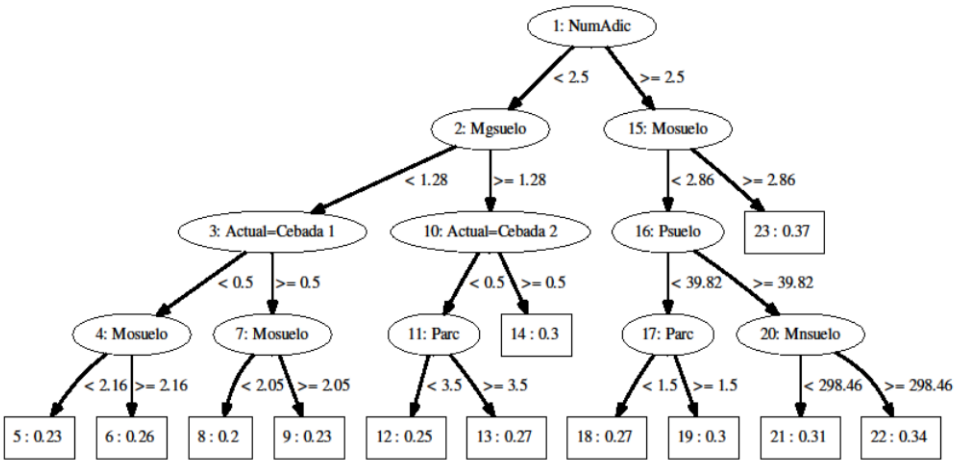


Figura 5-9 Árbol de regresión obtenido para nitrógeno del suelo.

Parámetros del árbol. Profundidad máxima=4, peso mínimo=8, datos para la poda=5. Coeficiente de Correlación = 0,876.

El árbol de regresión obtenido para el contenido en nitrógeno total del suelo (Fig. 5-9), muestra como principales variables que participan en este modelo a la cantidad de materia orgánica, tal y como se ha reflejado en los anteriores

árboles de regresión obtenidos para el contenido de MO. También se observan correlaciones positivas con otras variables del suelo como son la cantidad de manganeso, fósforo o magnesio, al número de adiciones de biosólido realizadas y si el cultivo fue cebada. Comprobamos que en las ramas donde aparece la materia orgánica, los valores crecientes de ésta muestran cantidades mayores de nitrógeno total. Como en el árbol de regresión obtenido para la MO, cuando el número de adiciones ha sido superior a 3 y la cantidad de materia orgánica presente en el suelo es superior a 2,86% en el suelo, se alcanzan los valores más elevados de nitrógeno en las medidas realizadas. Por otro lado, vemos que cuando el cultivo del año en estudio es la cebada por primera vez, los valores de nitrógeno en el suelo son los más bajos, estando ligados a la cantidad de materia orgánica del suelo.

Fósforo Olsen en Valdespinar

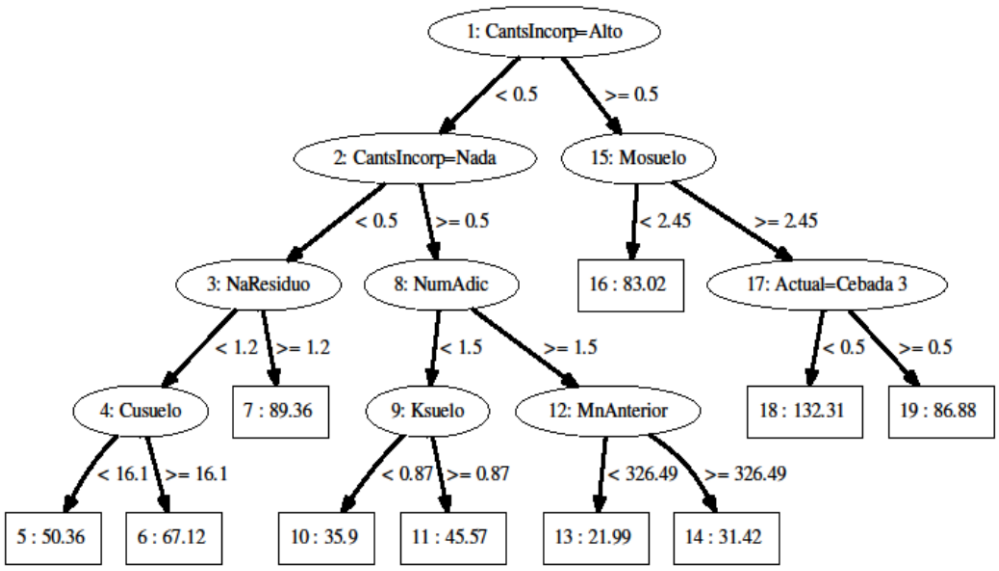


Figura 5-10 Árbol de regresión obtenido para el fósforo del suelo.

Parámetros del árbol. Profundidad máxima=4, peso mínimo=8, datos para la poda=5. Coeficiente de Correlación = 0,824.

El contenido de fósforo disponible del suelo está correlacionado con las mayores dosis de biosólido incorporado y con la materia orgánica presente en el suelo y tiene una correlación negativa cuando el cultivo ha sido cebada en el último año (Fig. 5-10). Otros factores que influyen en el aumento de la cantidad de fósforo son la cantidad de potasio disponible y cobre total en el suelo, así como una mayor cantidad de sodio en el residuo.

Los tratamientos de biosólidos para las dosis definidas como altas y la cantidad de fósforo en el residuo se corresponden con las mayores cantidades de fósforo en el suelo. Los menores valores se dan en el tratamiento control, (el cual no ha recibido ningún tipo de aporte).

Productividad en Valdespinar

Los árboles anteriores predicen las variables que tienen una fuerte influencia en la fertilidad del suelo pero no con la productividad en sí. Para el análisis de la productividad, dado que se desarrollaron cultivos diferentes y las condiciones de producción fueron muy variables, fundamentalmente por las características climáticas bajo las que se desarrolló cada campaña agrícola, se ha recurrido a una transformación previa de los valores en cinco rangos utilizando como referencia la producción de las parcelas control. Los rangos establecidos fueron;

- Igual, corresponde a la productividad en las parcelas control,
- Dos rangos superiores, con nombres “alta” y “muy alta”
- Dos rangos inferiores de nombres “baja” y “muy baja”.

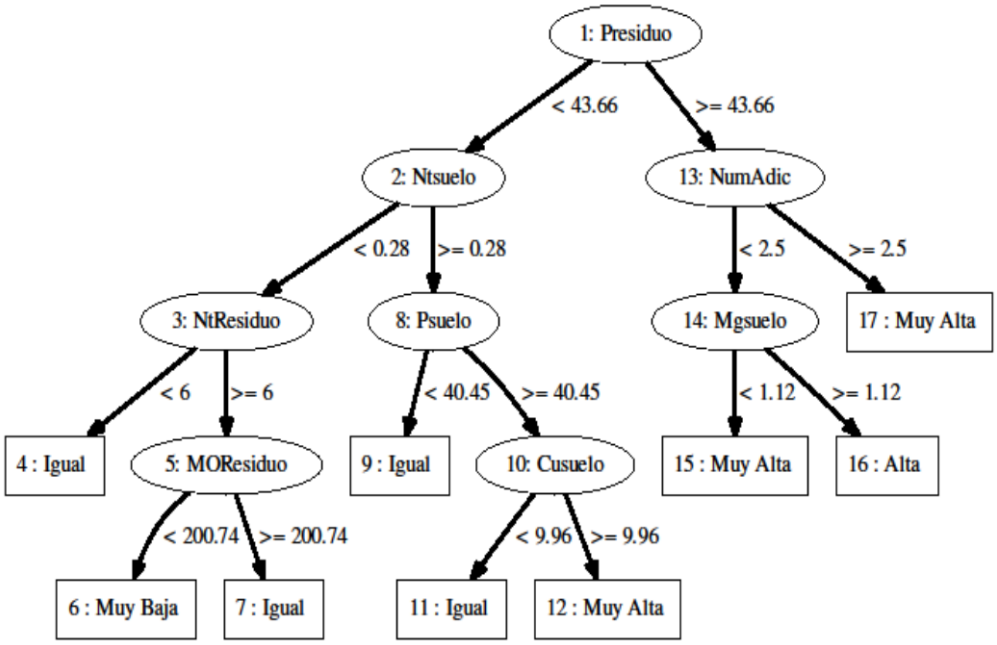


Figura 5-11 Árbol de regresión obtenido para la productividad de la finca de Valdespinar.

Parámetros del árbol. Profundidad máxima=4, peso mínimo=8, datos para la poda=5. Instancias clasificadas correctamente = 0,7689.

El árbol de regresión obtenido para la productividad (Figura 5-11), se observa su relación directa con la cantidad de nutrientes incorporados en el compost de biosólidos, principalmente con su contenido de P total y en una menor proporción con el contenido de nitrógeno en el suelo y en el compost. La aplicación de un exceso de este nutriente sobre la demanda que presenta el cultivo podría inducir una acumulación de fósforo total en el suelo, máxime en un suelo calcáreo y con valores de pH altos, lo que induce a su retrogradación a fosfato tricálcico de muy baja solubilidad.

Se observa que el árbol de decisión para la productividad no contiene ninguna hoja que prediga el valor "inferior", esto es debido a que el proceso de poda elimina este factor porque están en un nivel de profundidad del árbol demasiado alto.

Cadmio en Valdespinar

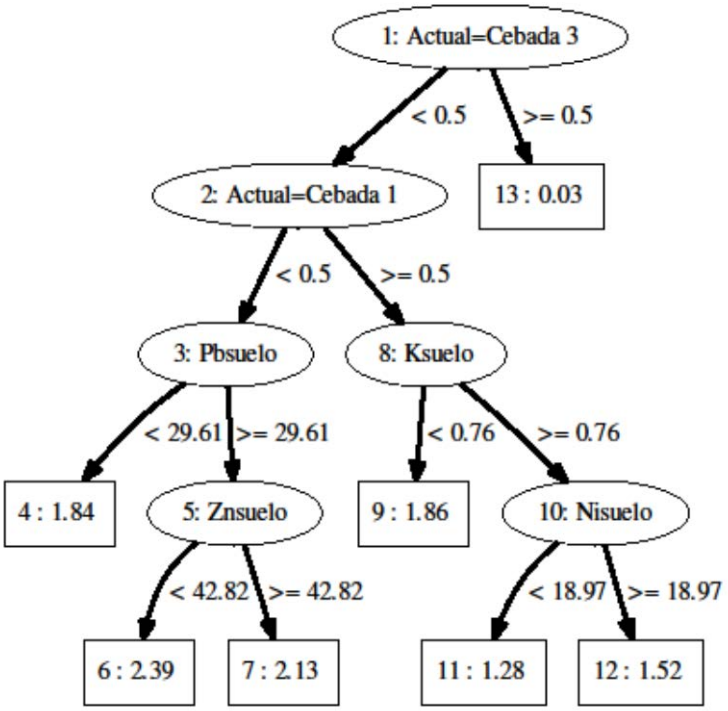


Figura 5-12 Árbol de regresión obtenido para el cadmio del suelo.

Parámetros del árbol. Profundidad máxima=4, peso mínimo=8, datos para la poda=3. Coeficiente de Correlación = 0,9243.

El árbol de regresión correspondiente al Cd (Fig 5-12), es el que mejor coeficiente de correlación (r= 0,9243) presenta de los estudiados. En cuanto a la presencia de este metal en el suelo, podemos ver cómo el cultivo de la cebada tiene una gran influencia sobre la cantidad de cadmio remanente en el suelo, lo cual se deriva del hecho de haberse cultivado en los años que no ha habido adición de enmiendas, lo que corresponde con menores cantidad de cadmio en el suelo. Los valores más bajos de Cd en el suelo corresponden al último año cultivado con cebada, en el cual no hubo adición. El cadmio aparece relacionado de manera positiva con cantidades de otros metales en el suelo como níquel, plomo y tiene correlaciones negativas con potasio y zinc. La cantidad de cadmio en el suelo tiene valores más bajos el último año.

Plomo en Valdespinar

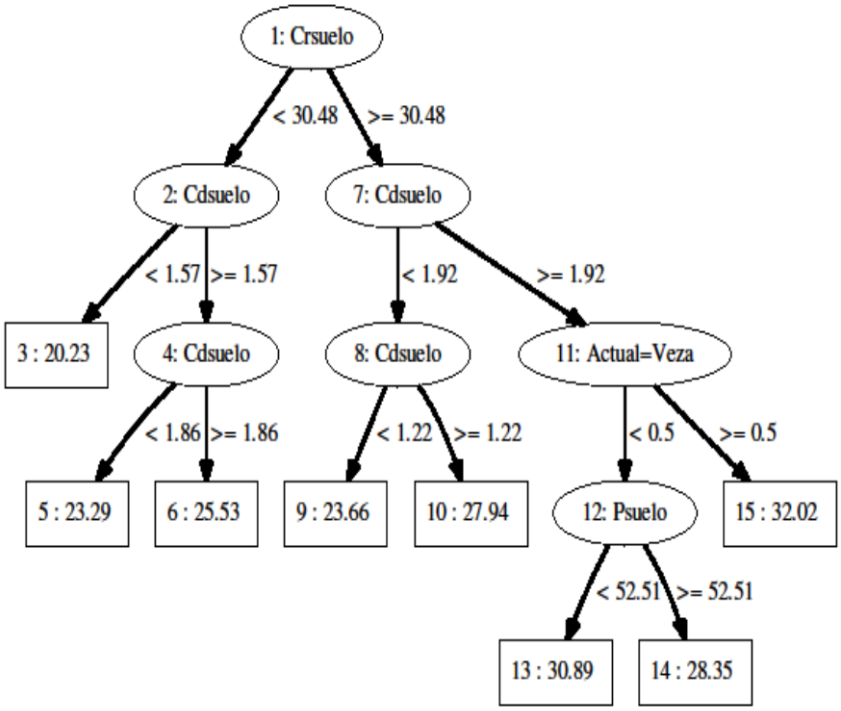


Figura 5-13 Árbol de regresión obtenido para el plomo del suelo.

Parámetros del árbol. Profundidad máxima=4, peso mínimo=8, datos para la poda=7. Coeficiente de Correlación = 0,8093.

En el caso del árbol de regresión para el Pb en el suelo (Fig. 5-13), se observa un árbol con una repetida influencia de la cantidad de cadmio en el suelo y donde los valores más altos de plomo aparecen asociadas con cantidades de cromo en el suelo superiores a 30,48 mg Kg^{-1}; además se asocia con valores altos de cadmio y cuando el cultivo actual es la veza. Por otra parte las menores cantidades de plomo presentes en el suelo también se asocian a valores bajos de cromo y cadmio en el suelo.

Zinc en Valdespinar

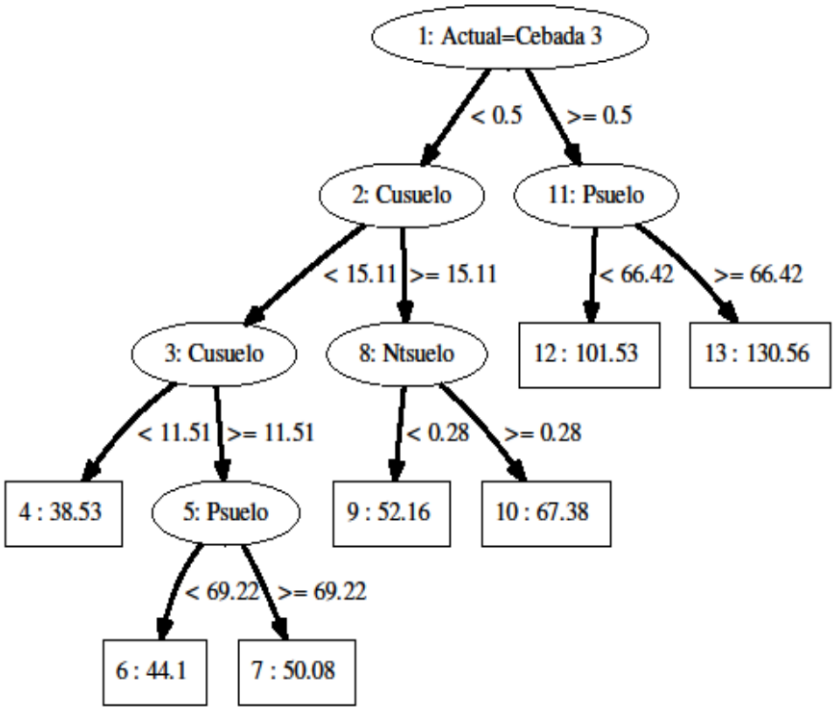


Figura 5-14 Árbol de regresión obtenido para el zinc del suelo.

Parámetros del árbol. Profundidad máxima=4, peso mínimo=8, datos para la poda=5. Coeficiente de Correlación = 0,9054.

Según los resultados mostrados en la Fig. 5-14, las variables que describen el árbol del Zn en el suelo son el cultivo actual de cebada y la cantidad de cobre, fósforo y nitrógeno en el suelo. Se observan los mayores valores de zinc en el suelo cuando el cultivo es la cebada por tercera vez y los menores cuando las cantidades de cobre en el suelo son inferiores a 11,51 mg Kg^{-1}.

Plomo en la raíz (Valdespinar)

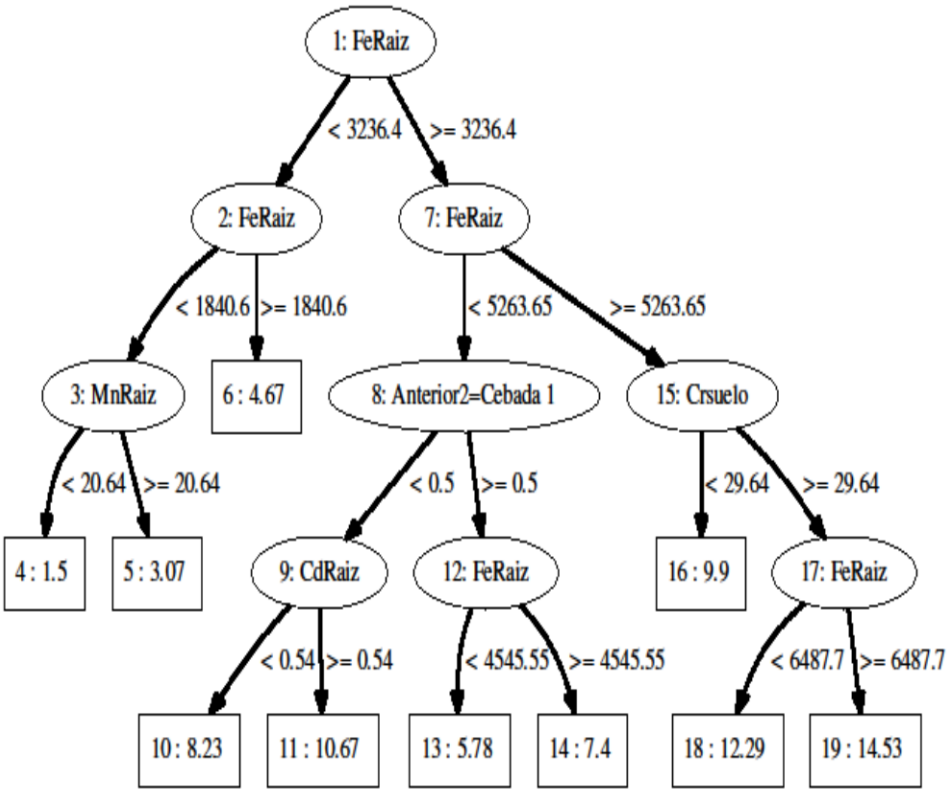


Figura 5-15 Árbol de regresión obtenido para el plomo en raíz.

Parámetros del árbol. Profundidad máxima=4, peso mínimo=8, datos para la poda=11. Coeficiente de Correlación = 0,9015.

El contenido de plomo presente en la raíz (Fig. 5-15), está relacionado con los contenidos de otros elementos en la raíz como son el hierro, manganeso y cadmio; también influye la cantidad de cromo presente en el suelo y cuando el cultivo que se realizó dos años atrás fue cebada. Si bien cabe destacar la influencia negativa de este último parámetro, ya que podemos observar cómo el cultivo de cebada hace disminuir la cantidad de plomo presente en el suelo. Los valores más bajos de plomo en la raíz tienen como variables determinantes al contenido de hierro y manganeso en la raíz. Los valores más altos tienen como factor principal los incrementos de hierro presente en la raíz y cantidades de cromo en el suelo de 29,64 mg Kg^{-1}.

Nitrógeno total en grano en Valdespinar

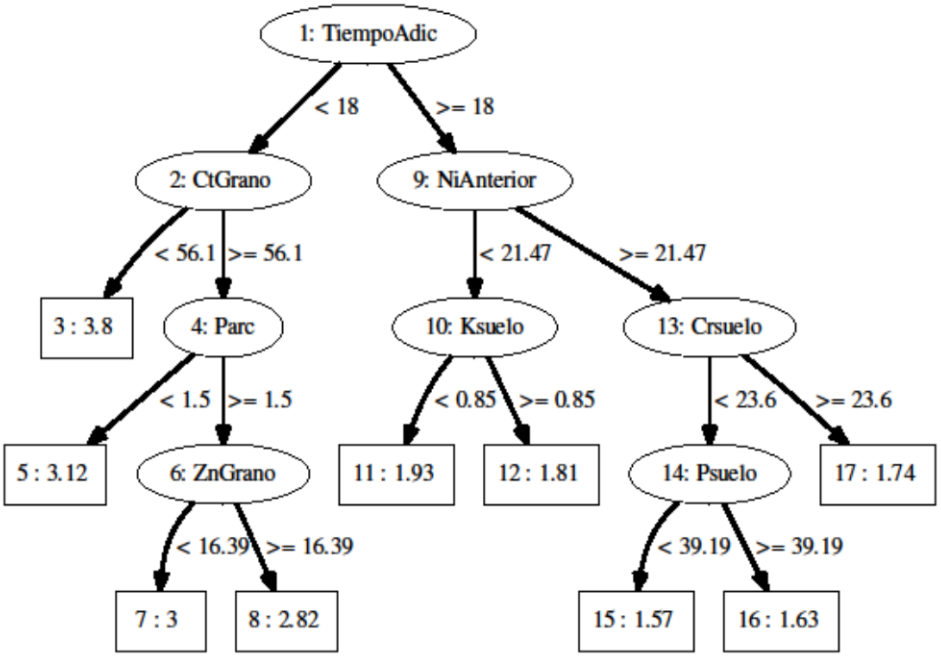


Figura 5-16 Árbol de regresión obtenido para el nitrógeno en grano.

Parámetros del árbol. Profundidad máxima=4, peso mínimo=8, datos para la poda=3. Coeficiente de Correlación = 0,9742.

El árbol de regresión obtenido para el contenido de Nitrógeno total del grano (Fig. 5-16), muestra una correlación con variables del suelo como la cantidad anterior de Ni, Cr y P disponible. Otros factores que influyen en el contenido de nitrógeno en grano es el contenido de C total del grano y el tiempo desde la adición del residuo. Sin embargo, vale la pena destacar la influencia negativa de Zn, puesse observa un aumento del contenido de nitrógeno en el grano al agregar las enmiendas del suelo. Cuando no se ha añadido biosólido en los últimos 18 meses, el aumento de la cantidad de N se produce de forma paralela a la reducción de las cantidades de P y Cr en el suelo.

5.5. Discusión

5.5.1 Fertilidad del suelo

Se han obtenido diferentes modelos mediante Minería de Datos para los parámetros relacionados con la fertilidad del suelo: N-total, Materia Orgánica del Suelo (MOS) y P-Olsen. Los árboles de regresión obtenidos muestran la

importancia de los parámetros iniciales del suelo, principalmente relacionado con las ratios de C/N en el caso de la materia orgánica y del nitrógeno.

La materia orgánica influye notablemente sobre la cantidad de nitrógeno presente en el suelo, como se ha observado en numerosos estudios. (Barrett and Burke, 2000) demostraron la estrecha relación entre los procesos de inmovilización de nitrógeno y los de mineralización del carbono, encontrando que la rápida inmovilización del nitrógeno se ve facilitada por una comunidad microbiana activa (Odlare et al., 2008) y la disponibilidad de un sustrato orgánico fácilmente mineralizable.

En el estudio que agrupa las tres parcelas conjuntas, la cantidad de materia orgánica en el suelo presenta un efecto antagonista con el zinc presente en el suelo tanto en el año de cultivo como en el anterior. En el trabajo de (Zahedifar et al., 2012) sobre la desorción de Zn en suelos, la liberación del complejo del zinc se ve disminuida con la aplicación de materia orgánica y aumenta con la cantidad de zinc incorporado. Como en casos anteriores, también se muestra una fuerte correlación con las cantidades de nitrógeno en el suelo (Barrett and Burke, 2000).

En el estudio del comportamiento del nitrógeno en el suelo la variable con más influencia es la cantidad de materia orgánica en el suelo. Así podemos ver que se observa un incremento en su contenido cuando existe un mayor número de adiciones de biosólidos, lo cual supone un aumento de la materia orgánica aportada.

Con el análisis de los datos de Valdespinar, se comprueba cómo el cultivo de cebada en el primer año da lugar a una menor cantidad de nitrógeno presente en el suelo aunque siempre se ve la influencia positiva de la materia orgánica, debido a los aportes de biosólidos, que permite incrementar las cantidades de nitrógeno. También vemos que el aumento de adiciones aporta una mayor cantidad de materia orgánica (rama derecha árbol) y de fósforo. Se ha demostrado que el compost suministra cantidades similares de P a un fertilizante inorgánico (NPK) (Mkhabela and Warman, 2005).

Por otra parte, el compost de biosólidos contiene importantes cantidades de P por lo que se puede utilizar como sustituto de los fertilizantes inorgánicos en el suministro de P disponible para la planta. El efecto de intercalar otros cultivos que aporten nitrógeno al suelo podría estar detrás de la explicación de las diferencias cuando se comparan los contenidos de nitrógeno en el suelo entre diferentes cultivos. El uso de la veza y otras leguminosas para incluir aportes de nitrógeno al suelo es conocido dentro de las prácticas de rotación. También se observa la influencia positiva de la materia orgánica, que permite valores crecientes de nitrógeno cuando no estamos utilizando el trigo como cultivo anual, en el caso de Andadilla. Paralelamente, también se observa que el aumento de fósforo en el suelo se corresponde con el incremento en las cantidades de nitrógeno en el suelo. Mayores cantidades de cromo y zinc anteriores en el suelo, por su acumulación, influyen en una mayor cantidad de nitrógeno en el mismo.

Los resultados obtenido en la finca de La Serrana, muestran un aumento de la cantidad de nitrógeno total en las parcelas que han recibido un aporte mayor de enmiendas (cantidad incorporada alta) y con relación a la cantidad de nitrógeno del residuo. También la cantidad de nitrógeno en el suelo aumenta cuando el tiempo de adición es de 12 meses, lo cual nos demuestra que las enmiendas disponen de nitrógeno fácilmente aprovechable por los cultivos, siendo un sustituto eficaz de los aportes de N necesarios por los cultivos. Sigua *et al.* (2005) muestran en su investigación con aplicación de lodos de depuradora durante años que son una fuente de nitrógeno y dieron una mejor producción frente al control. Finalmente podríamos decir que las relaciones anteriormente conocidas entre la materia orgánica y el nitrógeno en el suelo se ven reflejadas en los distintos árboles de regresión del estudio.

Con respecto al fósforo estudiado dentro de los parámetros de fertilidad del suelo, éste muestra correlaciones con la cantidad de residuo añadido y la cantidad de materia orgánica que se encuentra en el suelo para las distintas fincas experimentales, especialmente en Valdespinar. En otros trabajos relativos a la evolución del P en suelos tras la aplicación de enmiendas orgánicas, los resultados obtenidos en condiciones controladas, mostraron un aumento de P disponible, así como fuertes pero transitorios incrementos de actividades como la fosfatasa, que es otra vía para incrementar el P disponible, tras la aplicación de lodos de depuradora (Perucci 1990). Otros estudios de campo han demostrado la modificación de las propiedades del suelo debido al aumento de fósforo, tras la adición de lodos de aguas residuales y la aplicación de compost (Korboulewsky *et al.* 2002; Mantovi *et al.* 2005). Se ha confirmado que los metales pesados no son variables con mucha influencia en los árboles de regresión generados para el estudio de la fertilidad del suelo. Este hallazgo nos demuestra que la aplicación de ciertos biosólidos, en cantidades controladas con un seguimiento de su evolución en el suelo, ofrece una posibilidad para mejorar la fertilidad del suelo. El aumento de fósforo en el suelo también se corresponde con un aumento en la cantidad de Na en el lodo.

También con respecto al fósforo, con los resultados obtenido en Andadilla, dentro de los parámetros de fertilidad del suelo, éste muestra correlaciones con la cantidad de biosólido añadido y la cantidad de determinados metales en el suelo como son Zn, Cd o Cr. Una gran parte del fósforo aportado al suelo es debido a las incorporaciones de compost. Nuestra representación muestra una disminución del fósforo en el suelo que cuando se ha cultivado veza. Esta disminución es debida a la falta de aporte de biosólido durante dicho cultivo. Los metales presentes en el suelo, principalmente el cadmio, tienen correlaciones positivas con la cantidad de fósforo en el suelo. Mulla *et al.* (1980) en su estudio sobre una plantación de cítricos que fueron fertilizados con cantidades muy superiores a las tradicionalmente utilizadas encuentran una alta correlación (r=0,89) de las concentraciones de cadmio total y fósforo. Kidd *et al.* (2007) en su experiencia en suelos agrícolas enmendados con lodos de depuradora durante al menos 10 años en comparación con los suelos no enmendados (control) mostraban que la aplicación de lodos de aguas residuales provocó un aumento del pH del suelo, y las cantidades de N-total, P-Olsen, así como en concentraciones de Mg y K, Carbono Orgánico Disuelto

(DOC) y Ca intercambiable. Las concentraciones de Cu y Zn totales y extractables por EDTA del suelo fueron también significativamente mayores con la aplicación de lodos.

En Serrana, el incremento de plomo en el residuo y cromo en el suelo influye en una mayor cantidad de fósforo en el mismo. Esto puede ser debido a que la adición del biosólido conlleva un aumento de estos metales (Alloway y Jackson 1991) y como resultado, de dicha adición, una cantidad de fósforo aportado está disponible para el aprovechamiento por parte del cultivo. Esta observación se refuerza al considerar las ramas del árbol en la finca de la Serrana en las que aparece una menor cantidad de fósforo en el suelo cuando no se han añadido biosólidos (cantidades incorporadas=nada) y con el cultivo de cebada 4 que fue un año en el que no se adicionaron enmiendas en los suelos. Codling y Dao (2007) determinaron la eficacia del encalado, el P, y el hierro (Fe) en la reducción de la solubilidad de As y Pb en suelos, efecto que se incrementaba con el paso del tiempo. Los resultados de este experimento sugieren que la solubilidad de Pb está controlada no sólo por el pH, Fe y P, sino también posiblemente por materia orgánica y microorganismos del suelo.

En general los árboles de regresión obtenidos para el fósforo resultan complejos y con numerosas ramas, pero se puede observar como los componentes de los residuos (Na, Ca y Pb) en las distintas ramas producen incrementos en las cantidades de fósforo disponible en el suelo. La incorporación de cantidades de biosólido que permiten el aumento de fósforo en el suelo también aparece ligada a una mayor cantidad de sales de sodio y calcio en el residuo, según Mantovi *et al.* (2005). Los menores valores de fósforo en suelo aparecen directamente ligados a la rama con dosis control (ausencia de adiciones, representación de las tres parcelas) y con baja cantidad de calcio en el residuo (ningún aporte).

5.5.2 Carbono Total y Nitrógeno en Grano

La relación del carbono total con el nitrógeno en grano es muy alta y se ve una clara relación con el cultivo que se haya establecido en el año. Los árboles obtenidos para el carbono y nitrógeno en el grano se muestran definidos principalmente por los lugares de las experiencias o el tipo de cultivo sin apenas participación de otras variables.

En el estudio conjunto de las tres parcelas, la cantidad de carbono, que se encuentra en el grano tiene valores máximos con el girasol que principalmente se ha cultivado en Serrana o Andadilla. Los otros cultivos, cebada, veza, trigo, etc…tienen menores cantidades de carbono en su grano.

El nitrógeno en grano tiene relaciones muy importantes con los tipos de cultivo utilizados mostrando diferencias entre las distintas fincas experimentales. También la presencia de ciertos metales en grano influye de manera negativa en la cantidad de N-total, mostrando por un lado que las acumulaciones que pueden resultar tóxicas y que cuando los metales tienen mayores concentraciones en el grano existe una menor cantidad de nitrógeno.

Parte del contenido de N total en los abonos orgánicos como el compost de lodos de aguas residuales, no está inmediatamente disponible para las plantas y necesita ser liberado después de la mineralización de los compuestos orgánicos. Eso podría producir un tiempo que transcurre entre la aplicación o la enmienda y su asimilación por parte de la cosecha que a su vez, se acumula en el suelo para cultivos futuros o es lixiviado dentro del suelo. En diversos trabajos se ha comprobado que la aplicación de compost maduro y estabilizado, como el obtenido con lodos de EDAR, incrementó en un 6-8% del contenido de N orgánico en el suelo después de aplicaciones sucesivas; en contraposición, estiércoles o compost de residuos urbanos muestran mayor degradabilidad de sus componentes orgánicos (Chalhoub *et al.* 2013). Sin embargo, la presencia en ellos de metales pesados, tales como Cr o Ni, puede disminuir el contenido de N en el grano como consecuencia de su efectos tóxicos. Sharma *et al.* (1995) encontraron que un incremento en el suministro de Cr (VI) de 0,05 a 1,0 mM afectó severamente a la producción de grano en el trigo, con una concentración reducida de nitrógeno proteico y un aumento en la reducción de azúcares. Llegaron a la conclusión que el Cr (VI) es un potente inhibidor metabólico y un factor fitotóxico para el trigo. Sin embargo, en experiencias como las de Xue *et al.* (2014), para otros metales, encontraron que un buen suministro de nitrógeno a la planta aumentó el transporte de Zn y su acumulación en los granos de trigo. En nuestro caso, en el árbol representado para los datos de Valdespinar, el Zn contenido en el grano mostró un efecto negativo sobre su contenido de nitrógeno total.

Gran parte del nitrógeno fijado por las leguminosas se elimina con la cosecha de las semillas, dado su alto valor proteico, por lo que las contribuciones residuales de N fijado a los suelos agrícolas en los restos de la cosecha pueden ser relativamente pequeñas. Diversos autores como Beck *et al.* (1991); Peoples y Craswell (1992); Ravuri y Hume (1993) comprobaron que entre un 45 y un 75 % del N en la biomasa aérea de las leguminosas de grano se elimina en el grano durante la cosecha. Aun así, la inclusión de leguminosas en una secuencia de cultivo generalmente mejora la productividad de los cultivos siguientes. Así podemos ver que en Andadilla, el nitrógeno presente en el grano se nos muestra en los árboles de regresión asociado a los cultivos de guisante o también de girasol, que presentan los mayores contenidos de nitrógeno, apareciendo un alto grado de asociación entre ambos parámetros en grano.

Cuando se estudia la finca de la Serrana, el árbol relativo al contenido de nitrógeno en grano corrobora que los mayores valores de nitrógeno en grano se dan con el cultivo de guisante, algo lógico dado que es una leguminosa. Jensen (1996) en una experiencia de 4 años, comparó los rendimientos del cultivo mixto de guisante forrajero (*Pisum sativum* L.) y cebada de ciclo corto (*Hordeum vulgare* L.), con respecto al monocultivo, con y sin aporte nitrogenado. El intercalado de cultivo sin aporte de N suministrado por fertilizantes, también tuvo rendimientos similares a los del monocultivo, pero la ventaja de los intercalados en el cultivo (guisante-cebada) se debe principalmente al uso gratuito de N de las fuentes inorgánicas del suelo por los

componentes intercalados, lo que resulta en una menor competencia por N inorgánico.

Peoples *et al.* (2009), muestran en su meta-análisis con datos recopilados de todo el mundo, que por cada tonelada de materia seca producida de legumbres, el promedio equivalente sobre toda la planta en brotes y raíces noduladas, llega a 30-40 kg de nitrógeno (N). Aunque, gran parte del N_2 fijado por la leguminosa se elimina normalmente en la cosecha de semillas de alto valor proteico, las contribuciones residuales netas de N fijado a los suelos agrícolas después de la cosecha pueden ser relativamente pequeñas, si bien, la inclusión de leguminosas en una secuencia de cultivo generalmente mejora la productividad de los cultivos siguientes. Del mismo modo, las rotaciones y cultivos intercalados de plantas que son más capaces de acceder y movilizar elementos minerales con baja solubilidad pueden ser utilizados para aumentar sus concentraciones en los tejidos y el rendimiento del cultivo (Rengel *et al.*, 1999; Jolley *et al.* 2004; Graham *et al.* 2007;Inal *et al.* 2007).

Según Cechin y de Fátima Fumis (2004) la aplicación de concentraciones de nitrógeno altas dieron como resultado una mayor producción de materia seca en los brotes por planta y el efecto fue evidente a partir de 29 días después de la siembra. En nuestro caso, en la finca de La Serrana, la variable del tiempo transcurrido desde la adición nos muestra una mayor cantidad de carbono total en grano en el primer año de aplicación, mientras que en el árbol de regresión correspondiente al contenido de nitrógeno aparece cuando es superior a ese tiempo. En el estudio de Warman y Termeer (2005), comparando la eficacia fertilizante de los lodos con respecto a fertilizantes inorgánicos, encontraron que inicialmente la concentración de N fue mayor en el forraje de hierba de las parcelas con fertilizantes inorgánicos y fue en el segundo año cuando el contenido de N aumentó en el forraje enmendado con lodos de depuradora y el compost de lodos. Esta propuesta está relacionada con lo que aparece representado en el árbol de regresión, si bien puede ser debido también a la contribución del cultivo del guisante en la misma, lo que hace que supone que su efecto se manifieste cuando el tiempo de adición sea mayor a 12 meses.

Cuando representamos el árbol del conjunto de las tres fincas, se puede destacar que la cantidad de nitrógeno en grano aumenta con el número de adiciones y los cultivos separan granos con mayor contenido como el guisante y otros más bajos en nitrógeno como la cebada.

5.5.3 Productividad

Desde el punto de vista de los resultados obtenidos al analizar los niveles de productividad, se observa que las mayores producciones aparecen asociadas a la concentración de P-total en el compost, a un mayor número de adiciones, así como al aumento de N-total y P-Olsen en el propio suelo y, en menor medida, a la composición en C y N del compost. Efecto que es lógico, pues la aplicación

de biosólidos supone un aporte importante de nutrientes y deja un efecto de fertilización residual en el suelo.

En el árbol de regresión obtenido no se reflejan efectos adversos significativos sobre la producción por parte de los metales pesados presentes en el compost, ni tampoco ningún efecto relevante de la rotación de cultivos empleada, lo que valida el empleo como fertilizante orgánico de este compost de lodos EDAR. Otros estudios realizados, como el de Mishra *et al.* (2009), o las investigaciones de Ayari *et al.* (2010), en los cuales se aplicaron dosis de compost consecutivas durante varios años, también mejoraron las propiedades del suelo y la productividad del cultivo, con un aumento en la concentración de Cd, Cu, Cr, Ni, Pb y Zn en todas las partes de la planta y en menor medida en el grano.

5.5.4 Cadmio

El cadmio es uno de los metales que más preocupa por su movilidad en cultivos como los cereales y su posible transferencia a través de la cadena trófica.

Podemos observar en los árboles de regresión obtenidos para el contenido de cadmio en el suelo la influencia del contenido de plomo, relación que también se encuentra en los árboles de regresión del contenido de plomo en el suelo. La transferencia de estos metales a las diferentes partes de la planta y en especial al grano no es muy alta debido a su toxicidad para las plantas. Wang *et al.* (2006) establecieron los factores de transferencia de Cd y Pb, en verduras, observando una disminución con el aumento de las concentraciones de Cd y Pb en el suelo, lo que indica que la proporción de Cd y Pb absorbido disminuyó a medida que el Cd y Pb total del suelo aumentaba. En las relaciones establecidas entre el cadmio y otras propiedades del suelo, se observó una correlación negativa con el contenido de potasio en el suelo. Hao *et al.* (2012) encontraron que la proporción de la relación Cd/Zn en la raíz del girasol se vio afectada por el potasio presente en la raíz, lo que produjo una traslocación a la parte aérea del cadmio.

Stuczynski *et al.* (2003) refieren que la movilidad de metales en suelos enmendados no pudo describirse en función de simples parámetros como el pH o la cantidad de materia orgánica. La reacción de Pb con el suelo provocó una fuerte inmovilización, con menos del 1% del Pb recuperado mediante extracción con $CaCl_2$ 0,01 M. La inmovilización de Cd también fue significativa, mientras que la inmovilización del Zn aplicado fue mucho más débil que la de Cd o Pb. Por otro lado, se observa un efecto antagónico entre la materia orgánica sobre el cadmio, He y Singh (1993) encuentran en un estudio con rye-grass, que los rendimientos en materia seca que no se vieron afectados por la adición de materia orgánica, pero las concentraciones de Cd en distintas partes de la planta disminuyeron con cantidades crecientes de materia orgánica agregada.

También, en la representación del cadmio en la raíz, aparece una correlación con el contenido de plomo. Según Hassett *et al.* (1976) en una experiencia con maíz, la interacción entre el Pb y el Cd en la raíz produce una disminución en la elongación radicular. Dado que en la superficie de la raíz, Cd^{2+} y Pb^{2+} se unen a los grupos carboxi- de los ácidos urónicos del mucílago y la capacidad de mucílago para unir metales pesados disminuye en la serie de cationes: $Pb^{2+} > Cu^{2+} > Cd^{2+} > Zn^{2+}$, la unión con el mucílago restringe la captación de metal en la raíz y establece una barrera importante para su protección. Si bien algunos de los metales son liberados cuando el mucílago se biodegrada. (Morel *et al.* 1986).

En el caso del cadmio en la raíz el máximo se conforma con variables como los contenidos de Fe, Cu, Mn y Zn en la raíz. Los mínimos valores aparecen relacionados con la cantidad de Fe en la raíz y de Pb en el suelo, en el estudio de las tres parcelas. Según Hernández *et al.* (1998) en sus experimentos, el Cd se acumula principalmente en las raíces (90% de la concentración total de la planta) y se inhibe casi completamente la absorción de Mn. Existen algunos trabajos que estudian las interacciones entre el Mn y el Cd en plantas como son los de Wu *et al.* (2003a) y En el primero el Cd redujo las concentraciones de Mn y Cu en granos, raíces y brotes, descubriendo una correlación negativa significativa entre las concentraciones de Zn, Cu, Mn y la concentración de Cd en diferentes órganos de la planta, lo que sugiere la posibilidad de frenar la acumulación de Cd en plantas de cebada mediante la aplicación de estos microelementos en los suelos contaminados con Cd. Si bien, en el segundo trabajo, cuando el suministro de Mn era adecuado, las plantas mostraron pocos síntomas de toxicidad por Cd, aunque la absorción y la acumulación de Cd aumentaron. Nuestra experiencia obtiene unos resultados más cercanos a este último estudio, con cantidades crecientes de Cd cuando existen aumentos de Fe o Mn en raíz.

El Cd es motivo de mayor preocupación que otros metales pesados debido a su movilidad en cultivos como los cereales (Doyle 1977), de ahí el interés en probar su transferencia hacia el grano. Aunque la baja correlación del Cd en el grano y la falta de representatividad puede deberse a la falta de concentraciones significativas de acumulación en estas partes de la planta como se ve en Adams *et al.* (2004). Estos autores muestran que las concentraciones de Cd eran mucho más bajas en el grano de cebada que en el grano de trigo en condiciones de suelo comparables y los modelos realizados no daban buenas correlaciones. El aumento de carbono en el grano tiene un efecto antagónico sobre el Cd en el grano. Este efecto nos lleva a pensar en la relación inversa, explorada en trabajos anteriores Zia-ur-Rehman *et al.* (2015) ; Gao *et al.* (2011), donde sus autores señalaron que aunque no se encontró relación en el estudio entre la cantidad de Cd y Zn en el grano, ambos metales se correlacionaron negativamente con el rendimiento del grano, pues la absorción de Cd, incluso a bajas concentraciones, fue perjudicial para la mayoría de las plantas, al provocar una reducción en su crecimiento. Aunque el estudio de François *et al.* (2009), donde se compararon diferentes enfoques para estimar las concentraciones de Cd y Zn en granos de trigo basado en la especiación de estos elementos en la solución del suelo, las cantidades

adicionadas y diversos parámetros del suelo, muestra que las predicciones basadas en la química de la solución del suelo logran correlaciones positivas entre los valores predichos y reales de Cd, aunque fue más difícil predecir las concentraciones de Cd de grano. La predicción de Cd en grano en los cultivos de trigo es mejor a través del modelado empírico de los parámetros del suelo y los insumos del suelo, que a través de estimaciones de las fracciones biodisponibles en la solución del suelo.

5.5.5 Cromo

El cromo afecta gravemente al crecimiento de las plantas y a la productividad por sus efectos tóxicos para los cultivos, lo cual supone un riesgo para una agricultura sostenible.

Analizando el árbol de regresión para la concentración de Cr en suelo podemos deducir su estrecha relación con la cantidad de Cr presente en la enmienda. El árbol de Cr en suelo muestra la relación entre la cantidad presente de este metal anteriormente en dicho lugar así como con las cantidades de otros metales poco móviles como son el Pb y otros como Zn o Cu.

Para corroborar este hecho podemos ver las ramas del árbol de regresión para los datos de Valdespinar, en las que la incorporación de altas cantidades de biosólido o tras una primera adición con cultivo del girasol, se produce una mayor acumulación de Cr. UdDin *et al.* (2015) hablan del incremento de las concentraciones de Cr mientras se reducen las concentraciones de Ca, Mg y K en la raíz, tallos y exudados de la raíz, aunque en nuestra representación sólo se ve un aumento del Cr en la raíz relacionado con el Ca del suelo.

Según Stewart *et al.* (2003) los efectos del envejecimiento muestran que la bioaccesibilidad del Cr disminuye después de los primeros 50 días, hecho que se relaciona con una mayor estabilidad de Cr en la superficie del suelo, lográndose una bioaccesibilidad constante a partir de los 200 días. Así Fendorf *et al.* (2004) observan una disminución exponencial de la bioaccesibilidad durante los primeros 100 días de incubación que se acompaña de una disminución en la fracción intercambiable de estos elementos traza, lo que sugiere que hay, de hecho, un cambio en el mecanismo de retención que opera a través de este período de tiempo. Para la finca de Serrana con un tiempo de 12 meses desde la última adición, podemos observar un incremento en la acumulación de Cr en el suelo, esto también ocurre cuando el número de adiciones se incrementan. En la rama derecha del árbol se ve cómo con 2 o más adiciones, las cantidades de Cr presentes en el suelo son mayores. Según Banks *et al.* (2006) la retención de Cr en los suelos se correlacionó con el contenido en Cr(III). La absorción de cromo por parte de la planta representaba menos del 1 % del cromo retirado del suelo. En general, la adición de biosolidos tiene una influencia fuerte en los aumentos de las cantidades de Cr y su posible movilidad dado que las plantas carecen de un sistema de transporte específico para Cr, que es tomado por los portadores de iones esenciales tales como el sulfato o el hierro.

Las representaciones de Cr en la raíz muestran en alguna de las fincas poca variabilidad y no existe representación (Serrana). En aquellas que existen representaciones se ve la presencia del níquel como metal determinante. Liu y Kottke (2003) muestran como, de hecho, el Cr y el Ni se localizaron en los precipitados de las células de las raíces expuestas a sólo 100 μmol L^{-1} de una mezcla Cr-Ni, aunque las mayores cantidades de ambos elementos se acumularon principalmente en las paredes celulares y vacuolas de la cuarta o quinta capa cortical de la raíz.

De igual modo si se miran los árboles que representan los metales en grano, el Cr aparece relacionado con metales con poca movilidad como es el Ni o el Zn. En el estudio de Lübben y Sauerbeck (1991) se realizaron experimentos en macetas para examinar la distribución de metales pesados en el trigo después de 18 años de adiciones de enmiendas que mostraban contaminación por metales pesados. Los contenidos de Cd, Zn, Pb y Cr del grano fueron sensiblemente inferiores a los de la paja. Sin embargo, los niveles de Ni y Cu en el grano excedieron a los de la paja. Las raíces se enriquecieron en Cd, Zn, Ni y Cu, en comparación con el suelo. Sin embargo, Pb y Cr apenas fueron absorbidos por las raíces.

El efecto acumulador y la poca movilidad del Cr hacen que sólo en muy bajas proporciones pasen hacia el grano. En el estudio de Chatterjee y Chatterjee (2000), se concluye que la traslocación del Cr desde las raíces a la parte aérea de la planta es mínimo por eso los valores que aparecen de Cr en el grano no son elevados. Esto nos da idea de la baja movilidad que tiene este metal. Golovatyj y Bogatyreva (1999) mostraron que el Cr tiene un carácter estable que no depende de las propiedades del suelo así como que la máxima cantidad de este elemento contaminante siempre estaba contenida en las raíces mientras que una pequeña cantidad pasa a otros órganos de la planta, solamente un 0,1% del Cr acumulado se encontraba en el grano frente al 98% acumulado en la raíz.

Por otro lado en la finca de la Serrana, en la representación de Cr en el grano vemos un efecto antagónico mostrado por la cantidad de materia orgánica de los residuos, lo que puede relacionarse con lo expuesto por Bartlett y James (1988), para los que la mayor parte del Cr (VI) fue reducido a Cr (III) y retenido por las raíces fuertemente unido mediante un complejo de forma orgánica, insoluble o soluble pero que no se transloca a la parte aérea. En la representación del árbol en esta parcela, así como en la de Andadilla, vemos la correlación con el cultivo de girasol; según Shahandeh y Hossner (2000), la mostaza india (*Brassica juncea*) y el girasol fueron de las especies que más Cr acumularon entre las especies de plantas agrícolas.

5.5.6 Cobre

La representación de este árbol muestra las dinámicas del Cu relacionado principalmente con el Zn y la influencia de las cantidades incorporadas de biosólido, que conducen a la acumulación de Ni, Zn y Cr en el suelo. Según Ashworth y Alloway (2004), en un estudio con un suelo franco arenoso mediante columna de lixiviación para determinar el grado de movilidad de Cu, Ni, Zn y la materia orgánica disuelta (DOM) liberada de lodos de depuradora, mostraron la importancia de la DOM en el incremento de la movilidad de ambos metales, Cu y Ni, lo que se evidenció por su falta de movilidad cuando se añade a la columna de suelo como formas inorgánicas. Cantidades altas de Zn y Ni en el suelo proporcionaban valores elevados de Cu en las medidas realizadas. Por otro lado vemos como, los valores más bajos en la cantidad de los metales anteriormente comentados permiten tener unas menores cantidades de Cu en el suelo. Podríamos concluir que el contenido en Cu aumenta al aumentar la dosis añadida. (Morera *et al*. 2002; Luo y Rimmer (1995).

Se observa un antagonismo en el árbol de Andadilla del Fe presente en el suelo con el Cu. La presencia de concentraciones de Fe precedentes en la rama del árbol de Cu en el suelo con un efecto antagonista, nos da una idea de cómo un incremento de Fe en el suelo provoca una disminución en la movilidad del Cu. Willis (1936) mostró el efecto antagonista del Fe con el Cu, Zn y Mn, mientras en el estudio realizado sobre la disponibilidad del Fe en suelos calcáreos del sur de Irán de Ghasemi-Fasaei y Ronaghi (2008), la aplicación Fe aumentó la absorción de Fe, pero disminuyó la de Mn, Zn y Cu en cultivos de trigo.

El Cu presente en la raíz tiene un árbol de regresión que muestra la influencia de variables como el Fe, el Mn y el Zn presente en la raíz y metales en el suelo como son el Cr y Pb, cuya cantidad en el suelo se ve incrementada por la adición de biosólidos. Zhou y Wong (2001) mostraron cómo con residuos orgánicos añadidos a suelos calcáreos, la mayor cantidad de materia orgánica del residuo producía un incremento del Cu en la raíz. Los residuos orgánicos en gran medida podrían facilitar la pérdida por lixiviación de Cu, debido a la formación de complejos de DOM-metálicos solubles. Así por ejemplo nuestro estudio de las tres parcelas conjuntas, se muestra un incremento de la cantidad de Cu en la raíz potenciado por las cantidades más altas de biosólido añadido.

Silveira *et al.* (2002) estudiaron la adsorción de Cu en tres muestras de suelos naturales brasileños, relacionándola con el contenido de materia orgánica y con la eliminación de óxidos de Fe. Para valores de pH más altos, los óxidos contribuyeron de manera efectiva para la adsorción de Cu, ya que su eliminación fue seguida por una disminución en la retención de Cu. Los óxidos de Fe y Mn participan en la retención en las raíces del Cu presente en los biosólidos como se muestra en las representaciones de las parcelas de Valdespinar y Serrana. Según el trabajo realizado por Degryse *et al.* (2008), el

Cu y el Zn tomados por la planta se explica mediante la formación de complejos con los exudados radiculares. Se concluyó en esta investigación que los exudados radiculares de las plantas dicotiledóneas son el principal mecanismo para la movilización de Cu y Zn. Martínez y Motto (2000) explican cómo la solubilidad de agregados de Pb, Zn, y Cu aumentó con una disminución en el pH. Esto sugiere que la solubilidad metálica puede ser controlada por propiedades del suelo, tales como el pH, el contenido de materia orgánica y la mineralogía.

Entre las variables más destacables en la representación del grano podemos centrarnos en las que presentan una influencia negativa sobre la acumulación de Cu en grano como son: el tiempo desde la última adición, lo que puede indicar una disminución de formas disponibles de Cu en el suelo, el K de cambio, la cantidad de P-total adicionado en el residuo o el pH del suelo. Este último factor es importante pues el suelo en estudio tiene un pH fuertemente alcalino, lo que supone una movilidad metálica reducida; la adición de un residuo orgánico tiene un ligero efecto neutralizador del pH del suelo, lo que puede manifestarse en un incremento de la movilidad metálica de elementos como el Cu (Smith 1994). Doelman y Haanstra (1984) encontraron que el contenido en Fe era el principal factor abiótico que afecta negativamente a la toxicidad del Cu, mientras que el Mn fue encontrado como segundo factor. Si bien el Cu no aparece con efectos antagónicos en nuestras representaciones, los árboles muestran las relaciones de ambos metales en la raíz y el grano que hemos obtenido. Según Kabata-Pendias (2010) la interacción Zn-Cu es considerada como un antagonismo, lo que significa que la presencia de uno disminuiría la absorción del otro. Esta afirmación no se puede asumir con nuestras representaciones que se encuentran más cercanas al estudio realizado en cebada de Luo y Rimmer (1995), donde un análisis de los mecanismos implicados en la asimilación del metal por la planta, concluyó que el crecimiento era controlado principalmente por la cantidad de Zn disponible para las plantas, que dependía de las cantidades de las cantidades de Zn y Cu añadidas.

5.5.7 Hierro y Manganeso

Ambos metales aparecen relacionados en los árboles de las tres partes en estudio (suelo, raíz y grano).

Según Fadigas *et al.* (2010) en su estudio de retención metálica en suelos de Brasil, la mejor correlación se obtuvo entre los contenidos de Fe, Mn, arcilla y limo, y Cd, Co, Cu, Zn, Cr y Ni en el suelo. El contenido de Mn se incluyó en el modelo de regresión lineal múltiple para Cd y Co, debido a su asociación con estos últimos metales. En nuestras representaciones, las relaciones de los metales, además de las propias de Fe y Mn entre ellos, se asocian con las adiciones y otros metales como son el Ni o el Pb. El Mn en el suelo tiene un incremento mayor cuando se cultiva la cebada el primer año, lo que corresponde al inicio de la experiencia con adición de biosólidos.

En varias de las representaciones de los árboles del hierro en raíz y grano vemos las influencias del manganeso y viceversa. Se observa una gran correlación entre ambos metales e incluso se pueden ver semejanzas en la estructura de las ramas de los árboles representados. Estas relaciones concuerdan con experiencias como los datos obtenidos por Gallardo-Lara *et al.* (2006).

En el caso del Fe en la raíz, de Andadilla, se puede ver un efecto antagónico debido al incremento de Zn en el suelo y en la Serrana del Cr en el suelo para con el Mn. Shaheen y Rinklebe (2014) encontraron en perfiles de suelos, cómo las concentraciones de Cr, Cu y Zn se correlacionaban significativamente con los óxidos y amorfos de Mn y Fe.

En el caso del manganeso en el grano, en Valdespinar, se observa el efecto antagonista del cobre en el grano y de la cantidad de nitrógeno en el suelo. Según Lastra *et al.* (1988) en su trabajo con cultivos hidropónicos de pinos , la absorción de Mn mostró un antagonismo con el Cu, siendo las acículas el sitio de mayor concentración de Mn.

5.5.8 Níquel

El níquel es uno de los metales con menor movilidad en los suelos. En Kashem *et al*. (2007) se muestra un estudio de la movilidad de Cd, Ni y Zn en suelos de Bangladesh. Este estudio demostró que el Ni era el metal con la menor movilidad entre estos tres metales estudiados. En nuestro caso, los árboles de regresión para Ni en el suelo aparecen relacionados con las adiciones de biosólidos. Las representaciones muestran la influencia de la incorporación de residuos en los cultivos. Así tanto en Valdespinar con la veza como en Andadilla con el girasol y la temporada tercera en el conjunto de las tres fincas se han añadido biosólidos. De igual modo, en la representación del árbol de La Serrana el número de adiciones corrobora la relación entre la cantidad de Ni presente en el suelo y los aportes de este metal debidos a las adiciones de enmiendas. Shewry y Peterson (1976) analizaron las raíces de tres especies de plantas acumuladoras y las concentraciones de Ni, mostrando cómo llegaban a acercarse o incluso superar las concentraciones del suelo, por lo que la traslocación de ese elemento en el conjunto de la planta no es muy frecuente.

En Yin *et al.* (2002) se muestra la distribución de Ni y Zn en la solución de los suelos, así como la actividad de Cu^{2+} libre . Estos estaban estrechamente relacionados con la materia orgánica del suelo (SOM), un hecho que podemos observar en nuestros resultados que muestran la presencia de estos metales en las ramas del árbol de regresión para la finca de la Serrana. También vemos que los valores de Ni en suelo aparecen ligados a cantidades crecientes de Pb en el suelo. Según Asensio *et al.* (2016), no se observó una competencia significativa entre los metales Ni, Pb y Zn, por lo que concluyeron que los residuos utilizados tienen una alta capacidad de adsorción que evitaría que los

metales estuvieran en forma móvil a pesar de las posibles adiciones a partir de residuos.

Existe una correlación positiva del Mn y del Fe con el Ni en el suelo y en la raíz, tal como se refleja en los árboles de Valdespinar, Andadilla y las tres fincas agrupadas. En el estudio de Antić-Mladenović *et al.* (2011) explican cómo afecta a la solubilidad del Ni en suelos, la presencia de Mn, la disolución y precipitación de óxidos de Fe/Mn y la adsorción en distintos sólidos. En nuestros resultados, la formación de óxidos en la raíz y la baja movilidad del Cr y del Ni, explica su presencia en la representación.

En el trabajo de Bhattacharyya *et al.* (2008) se estudió el efecto de la adición de compost de residuos sólidos urbanos (MSWC) en el contenido de Co y Ni en cultivos de arroz. Ambos metales se unen de forma significativa a la materia orgánica y a los óxidos de Fe y Mn en los residuos. En nuestro caso, en la representación del níquel en la raíz, vemos una relación clara con el Cr en raíz. Fargašová (2012) indican que el efecto adverso del Ni sobre la planta era más fuerte que el de Cr (III) y Cr (VI) y que las raíces de las plantas fueron más sensibles que los brotes, aunque el transporte de Ni a los brotes fue mayor que con el Cr. Por otra parte, en el estudio con adiciones de nutrientes enriquecidos con Cr, Ni y Cd de Zurayk *et al.* (2000), con doce especies hidrófilas mediterráneas del Líbano, la acumulación de metales y bioconcentración se produjo preferentemente en las raíces.

Observando los árboles de este metal podemos ver cómo las principales cantidades del metal se distribuyen entre el suelo y la raíz, sin existir unos aportes peligrosos para la salud que puedan transferirse al grano. Dentro de los árboles representados para el grano se observa la correlación entre Ni y Cu. En Ashworth y Alloway (2004), se encontraron similitudes entre el comportamiento del Cu y el Ni, lo que sugiere que la migración se produjo como complejos orgánicos. En cuanto al grano en el trabajo sobre arroz de Tang y Miller (1991) se determinó que existía una interacción sinérgica entre Cu y Ni en términos de toxicidad para el arroz y la concentración de dichos metales en los tejidos vegetales son descritos con modelos cuadráticos.

Estudios como el de Mamindy-Pajany *et al.* (2013) establecen que el carbono orgánico disuelto generado por el lodo aplicado en suelos calcáreos podría facilitar la lixiviación de Ni debido a la formación de complejos de Ni-orgánicos solubles. Los aportes de biosólidos incrementan la materia orgánica en el suelo así como las cantidades de níquel y de otros metales como el cromo o el plomo que a pesar de su escasa movilidad pueden llegar hasta el grano aunque sea en cantidades pequeñas. Según el estudio de Melo *et al.* (2007), el níquel apareció ligado a las fracciones más estables de la materia orgánica y aunque el níquel añadido aumentó la concentración en los brotes no se observó aumento del metal en el grano. A pesar de esto en la representación del níquel presente en el grano en la Finca la Serrana se observa una mayor cantidad de níquel con el cultivo de cebada en el inicio de la experiencia. Ese año se procedió a la incorporación, por primera vez, del biosólido al suelo. Esto indica

un incremento de níquel inicial que en los siguientes cultivos no ha sido alcanzado.

5.5.9 Plomo

La toxicidad del plomo en el suelo es un gran problema aunque su reducida movilidad y disponibilidad permiten un mayor control. En su estudio Gigliotti *et al.* (1996) observaron una movilidad muy limitada de Pb en plantas de maíz cultivadas en suelos calcáreos (pH 8,3), hecho que se corrobora en otros estudios como los de Hernando *et al.* (1989) ; Smith (1992); Warman *et al.* (1995); Shaheen (2009). Por un lado, en nuestro caso, aparecen incrementos de los valores Pb en el suelo cuando ha habido adiciones de biosólido. Así, por ejemplo, cuando miramos las ramas de los árboles de los datos conjuntos de las tres fincas aparece la temporada cuarta en la representación. En esa temporada no ha habido adición de biosólidos y los valores de plomo en el suelo son menores. Además los aportes de biosólidos incrementan las cantidades de los metales en el suelo. Algunos metales poco móviles como Ni y Cr, claramente ligados a las adiciones de enmiendas, aparecen en el nodo raíz en el árbol de regresión, como en Valdespinar y en Serrana. Según Bahmanyar (2008) el uso de aguas residuales en el riego de arroz incrementó las cantidades disponibles de Cd, Ni, Cr y Pb en el suelo. La acumulación de metales pesados fue mayor en las raíces de arroz (control y tratamiento) que en los brotes y grano de arroz.

Los óxidos de hierro y manganeso que se pueden formar en la raíz producen retenciones de los metales pesados. Destaca la fuerte relación entre el Pb y el Fe mostrada dentro de la representación del Pb en raíz de Valdespinar y entre el Pb y el Mn en Andadilla. Los incrementos de materia orgánica da lugar concentraciones más altas de Pb y mayores cantidades de Pb disuelto o labil como se expone en Sauvé *et al.* (1998). Por otro lado, el trigo acumula mayores cantidades de Pb en la raíz, en la representación de las tres fincas, a pesar de que ese año no coincida con la adición de enmiendas, esta misma tendencia se observa cuando el tiempo de adición es superior a 12 meses, dado que el Pb se acumulará en el suelo y acabará posteriormente pasando a las raíces. Podemos concluir que la adición de este metal al suelo provoca un incremento en las cantidades de Pb en suelos pero que no resulta fácil que se transfiera a otras partes de la planta.

En las representaciones obtenidas para el Pb en el grano podemos ver la relación del Pb con el Cr y el Fe y con el tipo de cultivo. Así por ejemplo, cuando hay cultivo de cebada en Valdespinar la cantidad de Pb en el grano es baja aunque haya habido tres adiciones anteriores y sin embargo en Serrana, el cultivo de la cebada tiene valores más altos dado que coincide con la primera adición de biosólido y se incrementa la cantidad de Pb aportado. Aunque se trata de un metal poco móvil, existe una leve transferencia al grano. Si bien, esto no se corresponde con el estudio de Alegría *et al.* (1991) para el Pb donde

se encontró una correlación negativa entre el contenido de metal extraíble del suelo y el encontrado en hojas-tallos.

5.5.10 Zinc

En las representaciones de este metal en el suelo se puede ver la influencia de los cultivos en distintos años. Los aportes de biosólidos en las fincas de nuestro estudio van incrementando la cantidad de Zn en el suelo. Se puede inferir que existe una tendencia a la acumulación dada la correlación positiva del Zn en el suelo con el cultivo final, después de las adiciones, en Valdespinar. En Andadilla y Serrana se ven menores cantidades de Zn en el suelo con los cultivos iniciales, debido a que es el primer aporte de biosólidos (Kabata-Pendias 2010). Algunos modelos realizados en condiciones controladas destacan la disminución de la movilidad del Zn en suelos a los cuales se les han añadido enmiendas orgánicas (Battaglia et al. (2007); Zhang et al. (2016). Zahedifar *et al*. (2012) en suelos calcáreos tratados con materia orgánica y Zn muestran que hubo una rápida tasa de desorción durante las primeras 4 h seguido por un ritmo más lento durante las 12 h siguientes. La desorción de Zn aumentó a medida que se aplicó Zn, mientras que disminuyó con la materia orgánica aplicada.

Nan *et al.* (2002) estudian en suelos calcáreos tratados con Cd y Zn la transferencia a los tejidos de trigo y maíz. Según los autores existe un mecanismo de interacción Cd/Zn donde los efectos de los dos metales son sinérgicos entre sí en condiciones de campo, y que el aumento de los contenidos de Cd y Zn en los suelos podrían aumentar la acumulación de Zn o Cd en los cultivos de trigo y maíz estudiados. Wu *et al.* (2003b) muestran en su estudio, que el Cd añadido al medio disminuyó no sólo significativamente las concentraciones de Zn en todos los tejidos de las plantas, sino que también inhibe su translocación desde las raíces a los brotes, lo que lleva a una mayor proporción de Zn en raíz/tallo en las plantas tratadas con Cd. Este mecanismo de interacción parece mostrar su presencia con las representaciones en los árboles de Andadilla en el suelo y de Valdespinar y Andadilla en la raíz con la presencia Cd (anterior) en el suelo. Gao *et al.* (2011) en un estudio con distintas variedades de trigo, sugieren que el cultivo de especies conocidas puede mejorar la concentración de Zn sin afectar necesariamente la concentración de Cd en el grano.

El fósforo en el suelo tiene una correlación negativa con el Zn? presente en la raíz, en Valdespinar, produciéndose una disminución de Zn cuando las cantidades de fósforo aumentan, lo cual ya venía detallado en Verma y Minhas (1987); Singh y Steenberg (1975); Adriano *et al.* (1971); François *et al.* (2009).

De los resultados obtenidos se concluye que el Zn presente en el grano viene representado por múltiples variables y se muestran relaciones con metales presentes en el suelo y la raíz como el Ni o el número de adiciones de biosólido efectuadas. Alloway (2009) muestra la transferencia de distintos metales entre

las diferentes partes de la planta, así mientras Cu y Zn tienen un movimiento selectivo hacia partes como el grano, otros metales como Cd y Pb encuentran restringido su movimiento a estos órganos quedando en partes como la cáscara. La aparición de las cantidades de Zn en los nodos principales del árbol del Cu en grano nos lleva a pensar en una relación directa en los transportes de ambos metales (Cu y Zn) entre las distintas partes de la planta. Las relaciones entre el Cu y el Zn en el suelo, así como su posterior reflejo en el árbol del grano vienen relacionadas con los efectos que tenía el Cu añadido dado que aumenta la toxicidad del Zn, según Luo y Rimmer (1995). La posible sinergia entre los metales puede aumentar el efecto tóxico del Zn en el grano, que se reflejaría en una diminución de la cantidad de nitrógeno en el mismo. Así mismo la presencia dentro de los gráficos para el Zn en grano, de la concentración de este metal en la raíz, nos hace pensar que existe una fácil captación por parte de las raíces del Zn que, posteriormente, es traslocado al grano cuando se forma.

5.6. Conclusiones

Los árboles de regresión reproducen los resultados obtenidos por estudios con otras herramientas como las regresiones lineales. Este estudio describe el uso de árboles de decisión para determinar las relaciones entre variables que representan diferentes propiedades del suelo, raíz o grano. Además se puede decir que, según los resultados obtenidos, se producen mejoras para el suelo por la adición de biosólidos con incrementos de la materia orgánica, el nitrógeno y el fósforo en el suelo. Sin embargo se debe tener un control del tipo de biosólido añadido para evitar efectos adversos como la acumulación de metales pesados o un exceso de salinidad en el suelo por los aportes hechos. Las correlaciones entre metales poco móviles como cromo, níquel y plomo son mostradas con esta nueva técnica. Así mismo las relaciones entre los óxidos de hierro y manganeso presentes en el suelo y los formados en las raíces de las plantas son reflejadas mediante los árboles de regresión. Las representaciones obtenidas mediante los árboles de regresión aportan una visión de fácil interpretación que permite una mirada más atenta a los procesos que ocurrieron cuando se trabaja con matrices temporales y con relaciones complejas entre los numerosos datos estudiados. La combinación de los parámetros de poda permite generar gráficos de un tamaño adecuado, algo deseable para la interpretación y manejo de estos gráficos. Otra ventaja de estos algoritmos es su capacidad de trabajar con variables numéricas y nominales al mismo tiempo.

6. Validación del modelo

6.1 Introducción

Los estudios realizados hasta el momento nos han ayudado para obtener una serie de predictores en las distintas experiencias. La construcción de un predictor puede tener dos propósitos:

1. Para entender las relaciones estructurales entre la respuesta y las variables medidas.

2. Para predecir la variable de respuesta correspondiente a los futuros vectores de medición con la mayor precisión posible.

Aunque estos modelos resultan útiles por su fácil interpretación y la visión, en forma gráfica, de las relaciones entre las variables implicadas, se necesita dar un paso más. Este paso sería poder realizar una predicción. La predicción es una operación bien conocida en el análisis de datos. Para ello se establecen dos fases; en una fase de formación (la fase inductiva) se utiliza para desarrollar un modelo en el que de alguna manera se resumen las relaciones entre las variables, y que luego se puede aplicar a los datos nuevos para deducir una predicción a partir de ellos (la fase deductiva).

La fase inductiva se lleva a cabo en varios pasos que son:

1. Un paso de formación (o fase de entrenamiento), tiene lugar con una muestra de individuos cuya clasificación, o regresión, es conocida, y que se extraerá al azar de la población para ser modelada.
2. Una fase de validación interna, para comprobar el modelo resultante de la formación con otra muestra de individuos cuya clasificación se conoce y que se extraerá al azar de la misma población que la muestra de entrenamiento. Este paso nos permite seleccionar el mejor de los modelos creados en la etapa de formación, evitando al mismo tiempo el sesgo optimista que sería causado por una prueba sobre la misma muestra que la utilizada para la formación.
3. Una etapa opcional de validación con una tercera muestra, (fase de test) cuyos valores se conocen, para medir el rendimiento del modelo seleccionado como mejor en los dos pasos anteriores. Los propósitos de este paso es predecir la calidad de los resultados obtenidos cuando el modelo es aplicado.
4. Una etapa de aplicación, en el que se aplica el modelo resultante de la totalidad de la población para ser modelado.

En la práctica, los tres pasos - entrenamiento, prueba y validación - pueden estar todos en el mismo conjunto de datos, o pueden estar en conjuntos de datos separados. Se puede proceder con una división de los datos de prueba y entrenamiento en una proporción de 50:50 pero no resulta una buena división de los datos. Es mejor utilizar más de la mitad de los datos para el entrenamiento (Picard and Cook 1984) aunque eso suponga disminuir los datos

de prueba. Sin embargo, existe una técnica estadística con variaciones simples llamada validación cruzada. En la validación cruzada, se decide sobre un número fijo de particiones de los datos. Supongamos que usamos tres. Entonces los datos se dividen en tres particiones aproximadamente iguales; Cada uno a su vez se utiliza para las pruebas y el resto se utiliza para entrenamiento. Es decir, utilizar dos tercios de los datos para el entrenamiento y un tercio para probar, y se repite el procedimiento tres veces para que al final, cada instancia se ha utilizado exactamente una vez para la prueba. Esto se llama triple validación cruzada.

La forma estándar de predecir la tasa de error de una técnica de aprendizaje dada una única, fija muestra de datos es utilizar diez veces la validación cruzada. Los datos son divididos aleatoriamente en 10 partes en las que la clase está representada en, aproximadamente, las mismas proporciones que en el conjunto de datos completo. Una de las partes se retiene y se realiza el aprendizaje con los nueve décimos restantes. Después se calcula la tasa de error en el conjunto y así con el resto de conjuntos. El procedimiento de aprendizaje se ejecuta un total de 10 veces en diferentes conjuntos de entrenamiento. Finalmente, las diez estimaciones de error se promedian para obtener una estimación del error global.

Es necesario poder extrapolar estos resultados y realizar seguimientos en conjuntos de datos que no sean conocidos para comprobar su grado de ajuste con la realidad de las investigaciones dado que el objetivo de los modelos ha sido estimar el valor de una variable desconocida como función de un cierto número de otras variables independientes.

6.2 Métodos para la predicción

El objetivo es determinar cuál de los modelos con coeficientes de correlación más altos ajustan mejor con los datos que se han medido en laboratorio. Una vez construidos y entrenados los modelos, se realiza un testeo con datos previamente separados de la base de datos de entrenamiento. Este testeo de datos se realiza con los modelos que presentan un mejor comportamiento en cada uno de los grupos estudiados, de acuerdo a los criterios de selección analizados.

En la siguiente tabla podemos ver distintos métodos utilizados en Minería de Datos para la predicción

Tabla 6-1: Métodos usados en Minería de Datos.

	Precisión	Robustez	Concisión	Interpretación	Pocos datos	Valores perdidos	*Outliers*	Variables correlacionadas	Velocidad
Regresión lineal	+	+	+	+	+	-	-	-	+
Análisis lineal discriminante	+	+	+	+	+	-	-	-	+
Regresión logística	+	+	+	+	+	=	=	-	=
Árboles de decisión	=	-	=	+	-	+	+	+	+
Redes neuronales	+	-	-	-	-	-	=	=	-
Support vector machines	+	=	-	-	-	-	=	=	-

Si queremos un modelo preciso, vamos a preferir regresión lineal, análisis discriminante y regresión logística, y tal vez las *SVM* (*support vector machines*) y redes neuronales, teniendo cuidado para evitar el sobreajuste (asegurándose de que no hay más de una capa oculta, y no tener demasiadas unidades en la capa oculta). Para una mayor robustez, debemos evitar los árboles de decisión y ser cuidadoso de las redes neuronales. Para la concisión del modelo, debemos preferir regresión lineal, o discriminante. También el análisis de regresión logística, y en cierta medida árboles de decisión, siempre que los árboles no tienen demasiadas hojas. Para facilitar la lectura de las normas, debemos preferir los árboles de decisión y evitar las redes neuronales y *SVM*. La regresión logística, análisis discriminante y regresión lineal también proporcionan modelos fáciles de interpretar. Cuando se dispone de datos con valores perdidos, podemos tratar de usar un árbol, o la regresión logística, con una codificación de los valores perdidos como una clase especial. Los valores extremos (*outliers*) de las variables continuas no afectan a los árboles de decisión y son no demasiado de un problema para la regresión logística .Si las variables independientes son muy numerosas o muy correlacionadas, se eligen árboles de decisión. Si tenemos grandes volúmenes de datos, lo mejor es evitar las redes neuronales y las *SVM*. Las redes neuronales son más útiles cuando la estructura de los datos no es clara. Si estamos buscando un método para la aplicación directa a los datos, sin tener que preocuparse por la preparación (por la normalización, discretización, la transformación o la selección de variables), o la homogeneización de los mismos (cuando son de diferentes tipos), o con ajustes costosos de los parámetros, se deben tener en cuenta árboles de decisión.

Cada uno de ellos tiene sus ventajas y sus inconvenientes pero resulta que uno de los pocos métodos que pueden trabajar con valores perdidos y *outliers*, sin verse penalizado, y con la posibilidad de ser interpretado son los árboles de decisión.

En la regresión, los predictores por lo general se han construido utilizando un enfoque paramétrico. Según esto se busca un valor dentro de un conjunto finito de parámetros que nos minimiza la función de partida. En la regresión lineal, la práctica común es utilizar una selección por etapas o un algoritmo mejorado de subconjuntos. Esta selección de variables por etapas o de mejores métodos de regresión subconjuntos invalida el modelo inferencial, en el que los supuestos necesarios para el buen desempeño son mucho más estrictos, su comportamiento ha sido ampliamente estudiado, y las herramientas de diagnóstico para comprobar la bondad de ajuste están disponibles, y los programas de proporcionan mayor robustez a los modelos están progresando. Al no considerar la inclusión de todas las variables a la vez, no son necesarios los cumplimientos de todos los requisitos estadísticos clásicos. Por eso, cuando se realizan métodos de predicción existen una serie de parámetros que nos indican la calidad de estas técnicas en su función.

En predicción se usa como indicador de la calidad de la regresión R^2, coeficiente de determinación, que debe estar lo más próximo a 1 y se calcula

elevando al cuadrado el coeficiente de correlación. Este indica el grado de precisión del modelo que hemos desarrollado.

- Coeficiente de Correlación (CORR)

Este índice mide el grado de covariación entre distintas variables relacionadas linealmente indicando el grado de relación entre las variables estudiadas. En este caso los valores del coeficiente de correlación van de -1 a 1, siendo -1 una correlación perfecta negativa y +1 una correlación positiva perfecta entre las variables.

$$CORR = \frac{\sum_{k=1}^{n} \frac{(p_k - \bar{p})(m_k - \bar{m})}{n-1}}{\sqrt{\sum_{k=1}^{n} \frac{(p_k - \bar{p})^2}{n-1} \sum_{k=1}^{n} \frac{(m_k - \bar{m})^2}{n-1}}}$$

Donde m y p son, respectivamente, el valor medido y el valor predicho, es el número puntos de la base de datos usada para validar los modelos,

$$\bar{m} = \frac{1}{n}\sum_{k=1}^{n} m_k$$

y

$$\bar{p} = \frac{1}{n}\sum_{k=1}^{n} p_k$$

Otras veces se puede usar el coeficiente de R^2 ajustado en el cual se tienen en cuenta la cantidad de las variables involucradas en el modelo para ver el grado de ajuste óptimo. Este coeficiente siempre es menor que el valor del R^2

$$R^2 ajustado = 1 - \frac{(1 - R^2)(n-1)}{n-p-1}$$

Donde *n* es el tamaño de muestra y *p* el número de variables

- Error absoluto medio (*MAE*).

Una de las medidas de los errores en Minería de Datos es el *MAE* (Abramowitz and Stegun 1964). El *MAE* mide la magnitud media de los errores en un conjunto de pronósticos, sin tener en cuenta su dirección. Mide la precisión para las variables continuas. El *MAE* es la media de la muestra de verificación de los valores absolutos de las diferencias entre las previsiones y la observación correspondiente. El *MAE* es una puntuación lineal que significa que todas las diferencias individuales tienen el mismo peso en el promedio. Indica las desviaciones producidas por los valores obtenidos por el modelo con respecto al valor real. Cuanto menor sea el error más exactitud tiene el modelo en predecir nuevos valores. Se calcula como la media de los errores individuales sin tener en cuenta el signo. Este índice evita el posible problema que hace que al elevar al cuadrado el error se dé demasiado peso a los errores más extremos, afectando al resultado final.

$$MAE = \frac{1}{n}\sum_{k=1}^{n}|m_k - p_k|$$

- Error cuadrático medio (*RMSE*).

El *RMSE* es una regla de puntuación cuadrática que mide la magnitud media del error. Es la diferencia entre la previsión y los valores observados correspondientes al cuadrado y luego dividido sobre la muestra. Por último, se toma la raíz cuadrada de la media. Dado que los errores se elevan al cuadrado antes de que se promedien, el *RMSE* da un peso relativamente alto a grandes errores. Esto significa que el *RMSE* es más útil cuando grandes errores son particularmente indeseables. Es la medida más común y utilizada. Se realiza la raíz cuadrada ya que la suma de los cuadrados puede evitar conocer la magnitud real de los errores, puesto que elevar al cuadrado la divergencia da mayor peso a los errores elevados.

$$RMSE = \sqrt{\frac{1}{n}\sum_{k=1}^{n}(m_k - p_k)^2}$$

o su raíz cuadrada, el *RMSE*, que tendría la misma unidad que el parámetro de pronóstico. Al igual que con el *MAE*, su alcance es de cero a infinito, con una puntuación perfecta de 0. *MSE* es la diferencia cuadrada entre los pronósticos y las observaciones. Debido a la segunda potencia, el *MSE* y *RMSE* son mucho más sensibles a grande los errores de pronóstico que el *MAE*. Esto puede ser especialmente perjudicial en presencia de valores anómalos potenciales en el conjunto de datos y, en consecuencia, al menos con los conjuntos de datos pequeños o limitados, se prefiere el uso de la *MAE*.

MAE también es más práctico desde la intuición de los pronosticadores de destino, ya que muestra los errores en la misma unidad y escalar como el parámetro de sí mismo. Los valores de *MAE* van desde cero a infinito y, como en el error medio, una puntuación perfecta es igual a 0. Las medidas del *MAE* y del *RMSE* pueden ser usados juntos para comprobar la variación en los errores de un conjunto de predicciones. Son puntuaciones de orientación negativa, es decir, son mejores aquellos valores que son más bajos. El *RMSE* siempre será mayor o igual al *MAE*; la gran diferencia entre ambos es que una mayor variación entre ellos, provocará una mayor variación en los errores individuales de la muestra.

Si el RMSE = MAE, entonces, todos los errores son de la misma magnitud.

El uso del error absoluto medio (*MAE*), puede ayudar para la determinación de la exactitud de las predicciones de un modelo. Servirá para ver cómo se acerca la precisión de las previsiones cuando no tenemos acceso al modelo original utilizado para producir el pronóstico. Pero puede ocurrir con este método que se basa en la media de error que se puede subestimar el impacto de los errores grandes, pero poco frecuentes. Si nos centramos demasiado en la media, se producirá una gran variación debida a un gran error poco frecuente. Para el ajuste de estos grandes errores poco frecuentes, se calcula el error cuadrático medio (*RMSE*). Elevando al cuadrado los errores antes calculamos su media y luego tomando la raíz cuadrada de la media, se llega a una medida de la magnitud del error que da más peso a los grandes pero poco frecuentes errores que la media. Es por eso que se deben comparar *RMSE* y *MAE* para determinar si el pronóstico contiene errores grandes pero poco frecuentes.

- Robustez

Este segundo parámetro nos muestra que las variables del conjunto de entrenamiento, cuando se establece un modelo, producen una dependencia que dicho modelo debe ser capaz de generalizar para otras muestras. Por lo tanto, debe tener la menor sensibilidad posible a las fluctuaciones aleatorias producidas por ciertas variables así como a los datos perdidos. El modelo generado debe seguir siendo aplicable durante un tiempo razonable por lo que la estabilidad en el tiempo deber ser probada. Esto debería hacerse con muestras que cubran periodos de tiempo no establecidos en el conjunto de entrenamiento. Por lo que el modelo no debe aplicarse a variables que son dudosas, difíciles de obtener, o inestables de un período a otro.

- Concisión

Es relativamente sencillo de entender que un modelo con muchas variables producirá un mayor acercamiento a la realidad que se está estudiando. El problema surge cuando la cantidad de variables que participan provocan que se sufra una imposibilidad de entender el modelo de una manera concreta. Por eso, las reglas del modelo deben ser lo más simples posible y el número de reglas deben ser lo más pequeño posible. Esto nos asegurará que son más fáciles de entender y de controlar, y con mayor capacidad de generalización a otras poblaciones que la de la muestra de entrenamiento.

6.3 Construcción del *dataset*

El *dataset* construido corresponde a los 6 años de cultivo realizados en la parcela experimental de Valdespinar, Serrana y Andadilla y la validación de los modelos se realizó mediante validación cruzada de 10 veces. Este tipo de validación divide el conjunto de los datos iníciales en 10 subconjuntos. De estos subconjuntos se seleccionan 9 para entrenar el modelo dejando el subconjunto no utilizado para calcular el error de muestra. Este proceso se repite 10 veces usando cada uno de los subconjuntos de los que se dispone. Posteriormente, se calcula el error como la media aritmética de cada uno de los valores de los 10 errores que se han obtenido de las muestras parciales. De esta manera se tienen los valores de la media, el máximo, el mínimo y la desviación estándar de los 10 errores.

El escenario correspondiente a la investigación se basa en el estudio de las propiedades que afectan a la fertilidad del suelo. En este caso los parámetros a predecir serán la materia orgánica, el nitrógeno total y el fosforo presente en el suelo y los metales pesados que se encuentran en el mismo. El número de atributos tomados para el desarrollo de los árboles han sido 52 con un data set de N=1350.

El modelo utilizado para cada una de las variables ha sido *REPTree*. Según los parámetros establecidos en los estudios realizados previamente se han utilizado aquellos parámetros óptimos que den mejores resultados con un mayor coeficiente de correlación. Una vez obtenidos los parámetros según el criterio seleccionado para cada una de las variables de estudio se generará un modelo que posteriormente será utilizado sobre el conjunto de datos seleccionado para la predicción.

El conjunto de datos utilizado para la validación de los modelos generados se trata de una subdivisión de la parcela de La Serrana en la cual se estableció un régimen continuo de adiciones de biosolidos durante 5 años seguidos a partir del primer año de adición. Al igual que la finca experimental, se han mantenido las dosis establecidas en la experiencia y el mismo tipo de cultivo anual. Se tienen tres tipos de tratamientos (bajo, medio y alto) El conjunto de datos tiene el mismo número de atributos que el data set utilizado para la consecución del modelo y el número de instancias es de 225 (N=225).

6.4 Resultados

En la siguiente tabla se muestran los datos obtenidos por el conjunto de datos utilizado para la obtención del modelo y los resultados de las variables de estudio en el modelo de validación. Aunque no se incluyen los resultados de hierro y manganeso como metales pesados, debido a la alta dependencia que tienen estos metales con el entorno del suelo, ambas variables participan en los modelos estudiados.

Tabla 6-2: Tabla de las variables de estudio de modelo y validación.

Modelo (REPTree)	M.Orgánica	Nitrógeno	Fósforo	Cadmio	Cromo	Cobre	Níquel	Plomo	Zinc
Coef.Corr.	0,6076	0,6691	0,7684	0,8761	0,7559	0,8101	0,8311	0,8296	0,8394
Coef.Det.R^2	0,3692	0,4477	0,5904	0,7676	0,5714	0,6563	0,6907	0,6882	0,7046
MAE	0,1881	0,0203	11,325	0,1919	3,8391	1,2429	1,3895	1,8098	5,1549
RMSE	0,2531	0,0281	17,418	0,2501	5,6102	1,6435	2,0612	2,3482	8,2365
Validación	**M.Orgánica**	**Nitrógeno**	**Fósforo**	**Cadmio**	**Cromo**	**Cobre**	**Níquel**	**Plomo**	**Zinc**
Coef.Corr.	0,2376	-0,0091	0,5569	0,8757	0,8521	0,7258	0,7721	0,8823	0,5312
Coef.Det.R^2	0,0565	0,0001	0,3101	0,7669	0,7261	0,5268	0,5961	0,7785	0,2822
MAE	0,2493	0,0389	18,474	0,1954	5,1813	1,6562	2,0802	1,9861	8,3368
RMSE	0,3725	0,0534	22,86	0,2708	7,132	2,276	2,5075	2,6202	11,436

Los coeficientes de correlación del conjunto de datos de modelado son altos con algunas excepciones como la materia orgánica o el nitrógeno en el suelo. Los coeficientes de correlación del conjunto de validación ofrecen unos resultados aceptables para la mayoría de los metales pesados pero muy poco asumibles para las variables más ligadas a la fertilidad de los suelos como son materia orgánica o nitrógeno. La comparación de los errores del conjunto de datos para el modelo y el de validación nos muestra diferencias entre los valores a predecir y los reales. Así, cuando los valores de *MAE* y *RMSE* se encuentran más próximos la diferencia entre los errores medios y aquellos que se pueden dar por valores extremos son menores. Podemos ver que en los datos del modelo los errores *MAE* y *RMSE* son menores que los datos de validación lo cual es lógico dado que el conjunto de datos de validación tiene coeficientes de correlación más bajos. Se observa un incremento en los errores *MAE* y *RMSE* importante cuando los coeficientes de correlación de los datos de validación tienen un descenso importante como es el caso de las variables materia orgánica, nitrógeno, fósforo y zinc mientras que en las variables con coeficientes de correlación elevados (con mejores ajustes) las diferencias entre los errores no resultan tan acusadas. Aunque es cierto que los incrementos en los errores de tipo poco frecuentes (*RMSE*) se muestran también en las variables con buenos ajustes.

A continuación se representan los valores de las distintas variables mediante gráficos de dispersión o *scatterplots*. Para poder realizar una comparativa del ajuste esta representación es de los resultados predichos frente a los observados (*scatterplots*) dado que es una de las herramientas más simples para la verificación. La ordenada y la abscisa deben tener la misma escala, en cuyo caso la perfección es representada por los puntos de la línea de 45 grados para el que observado es igual al pronosticado. La correspondencia entre la línea de regresión y la línea de 45 grados es simplemente la medida de la fiabilidad.

Se ha realizado una comparativa de diversos metales pesados presentes en el suelo así como nutrientes como son la materia orgánica, el nitrógeno y el fósforo. Especialmente, los resultados para el cromo y el plomo muestran una capacidad de predicción mejor que otros modelos desarrollados. Se puede observar que los modelos que mejores resultados presentaban eran los que usan árboles de regresión para metales considerados principalmente como poco móviles. Tal como se ha dicho anteriormente es el caso del cromo, plomo pero también del níquel y del cadmio aunque en estos dos últimos el coeficiente de correlación del modelo de predicción es un poco más bajo que el del conjunto de origen. Por otro lado se puede ver que metales como son el zinc y el cobre, tienen los ajustes del modelo más bajos considerando los metales presentes en el suelo. Especialmente se observa una caída del coeficiente de correlación en el modelo del zinc muy importante.

Por otro lado si nos centramos en los valores de los modelos de factores del suelo los coeficientes de correlación no muestran unos buenos resultados con los datos de validación. El fósforo tiene el ajuste más alto dentro de estas variables estudiadas seguido por la materia orgánica y el nitrógeno. El

coeficiente de correlación de la materia orgánica se encuentra por debajo de 0,5 (0,2376) y el ajuste del nitrógeno es prácticamente 0.

Los gráficos de dispersión de los valores predichos frente a los observados permite conocer cómo de alejados están los puntos de la recta que representa la exactitud ideal *y=x* y nos dan una idea de las tendencias que los valores pueden establecer (sesgo).

La representación del error frente a los valores predichos nos mostrará el tipo de sesgo que se puede ver en los valores. En (Reynolds 1984) se consideran los procedimientos estadísticos para el problema de determinar de qué manera un modelo realiza bien la predicción de los valores de las variables que se observan en un sistema real de interés.

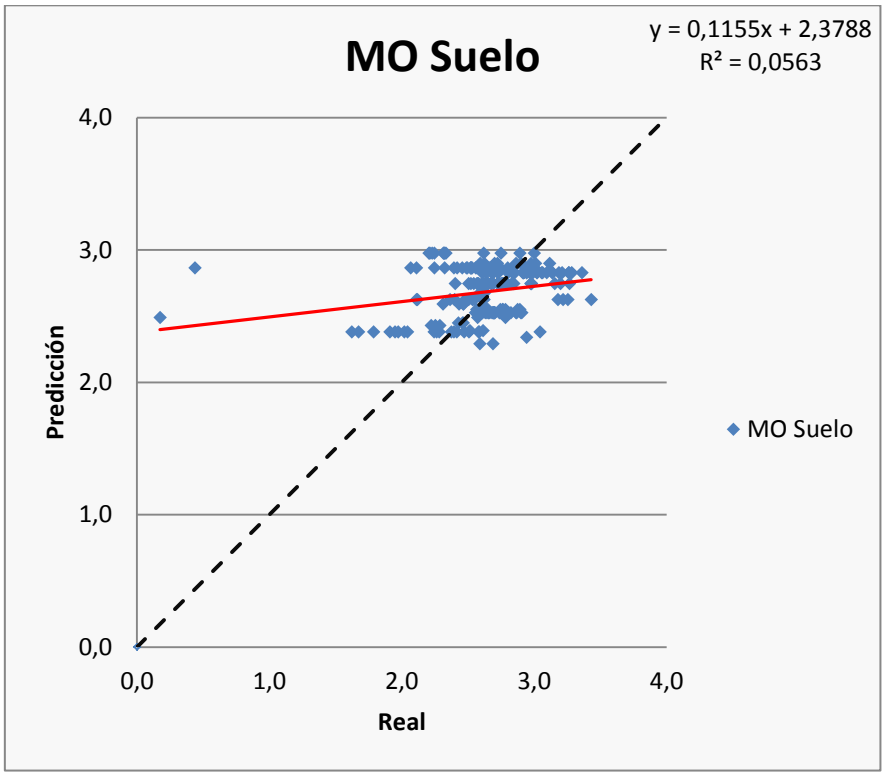


Figura 6-1 *Scatterplot* Materia orgánica del suelo.

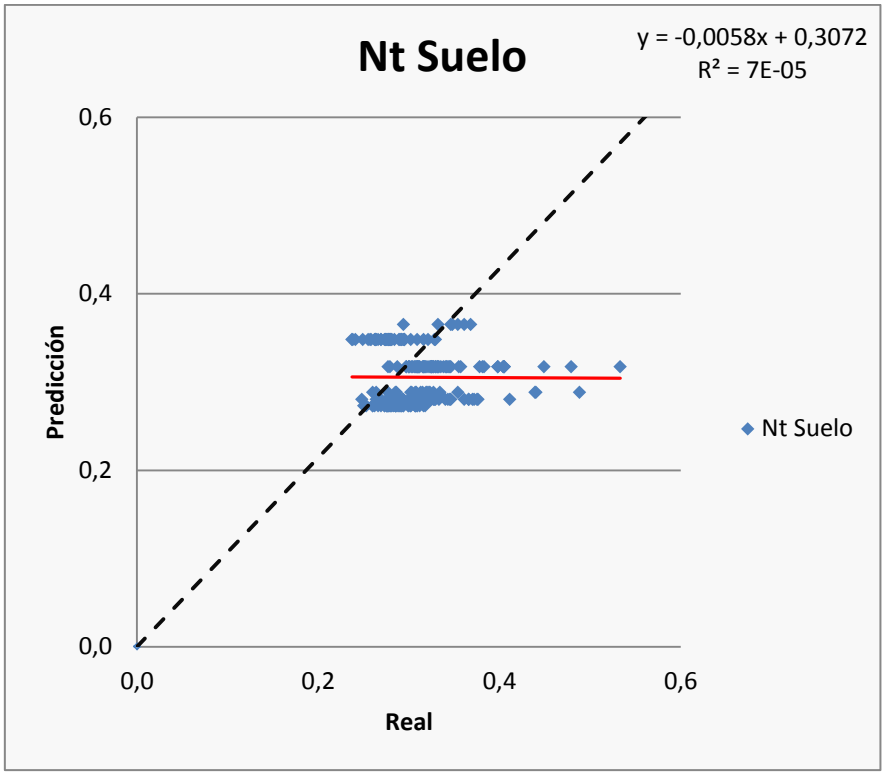


Figura 6-2 *Scatterplot* Nitrógeno del suelo.

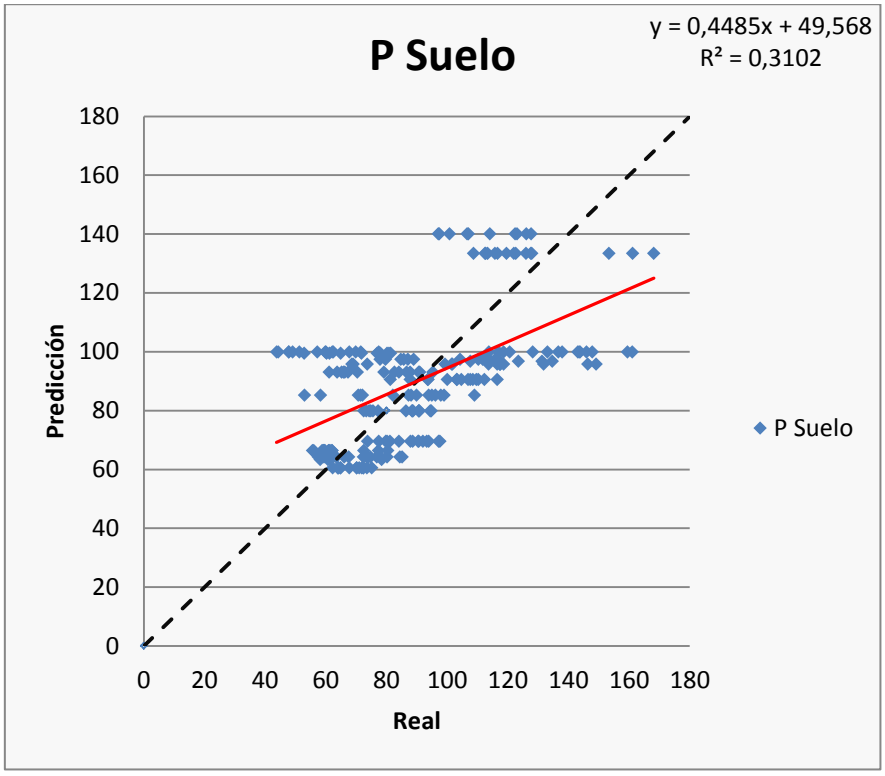


Figura 6-3 *Scatterplot* Fósforo del suelo

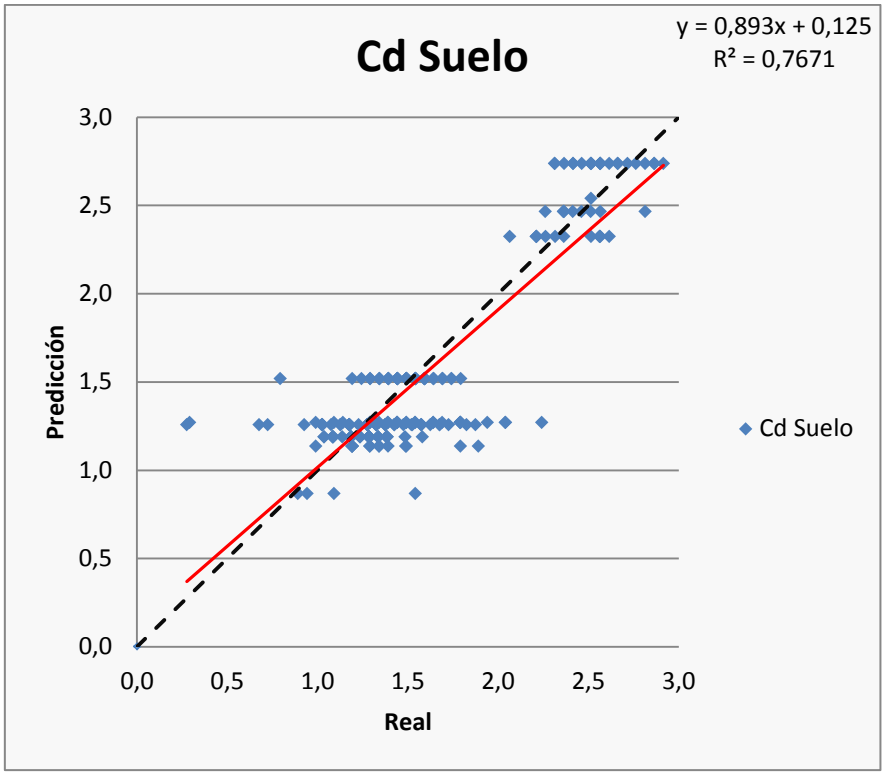


Figura 6-4 *Scatterplot* Cadmio del suelo.

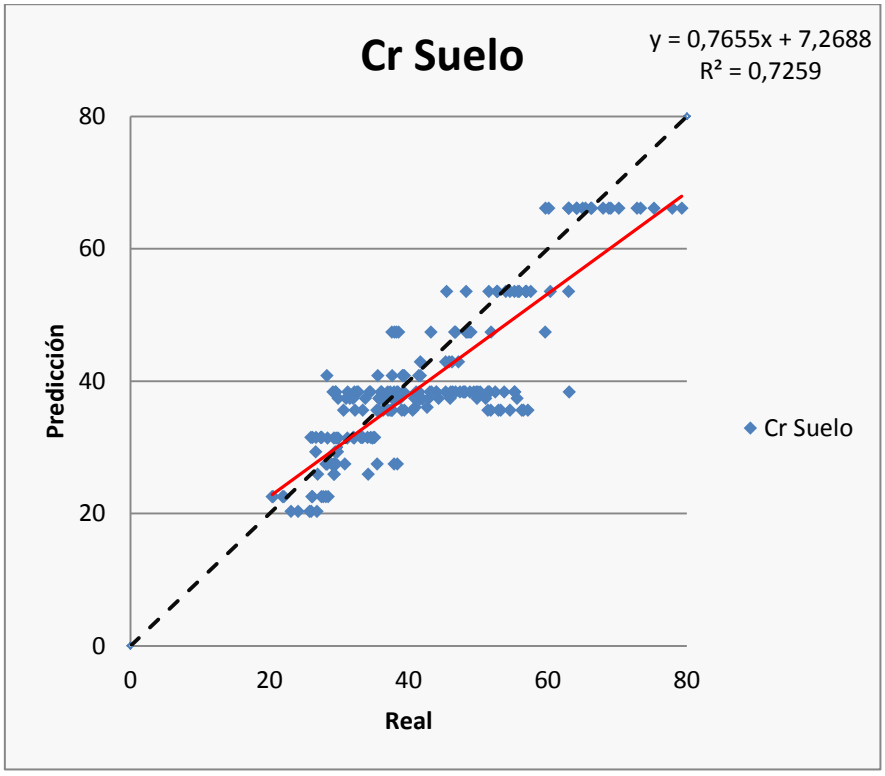


Figura 6-5 *Scatterplot* Cromo del suelo.

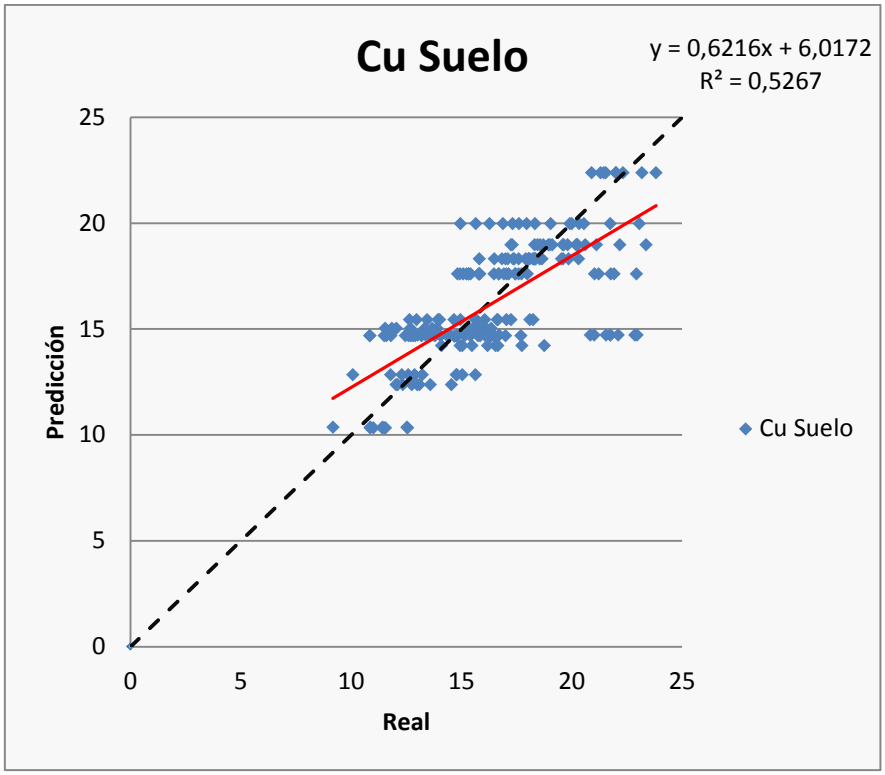


Figura 6-6 *Scatterplot* Cobre del suelo.

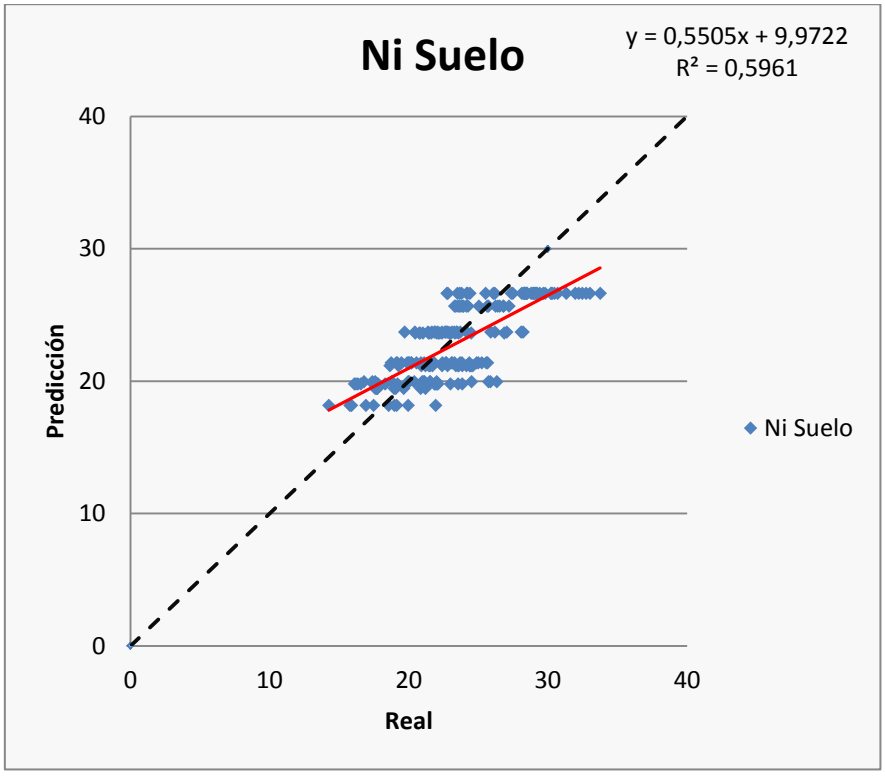


Figura 6-7 *Scatterplot* Níquel del suelo.

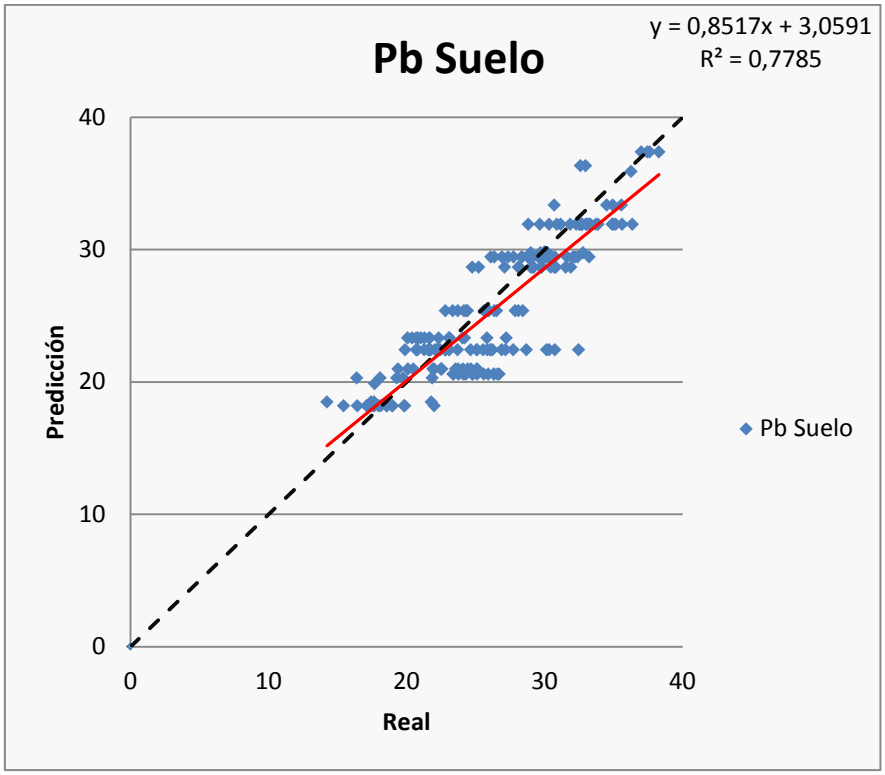


Figura 6-8 *Scatterplot* Plomo del suelo.

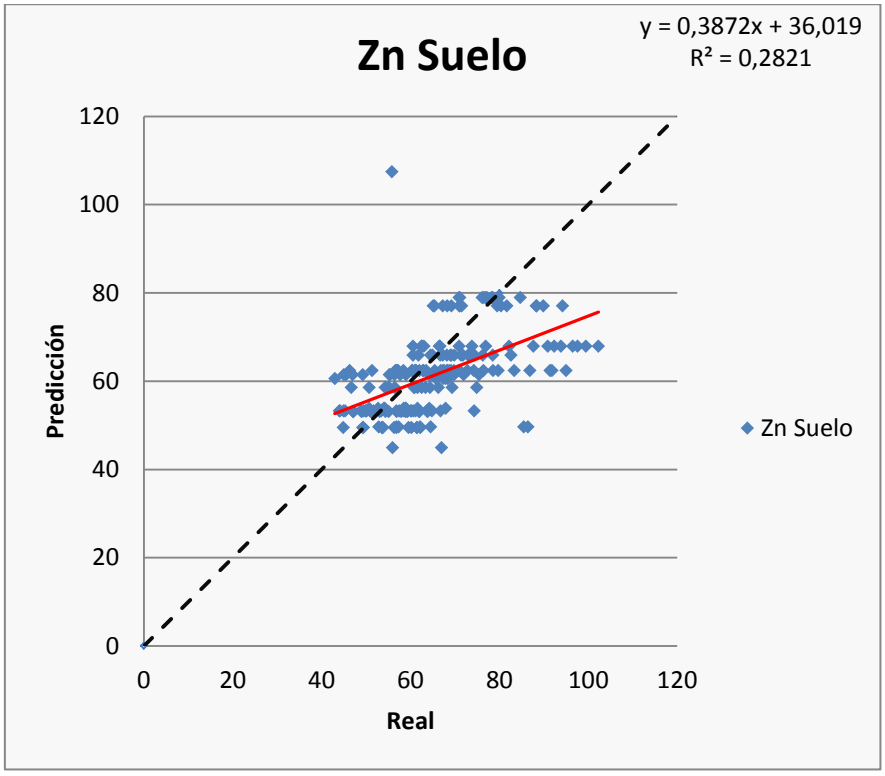


Figura 6-9 *Scatterplot* Zinc del suelo.

6.5 Discusión

Si las previsiones fueran perfectas, la recta representada en cada uno de los gráficos entre los valores de predicción y los observados coincidiría con la línea de 45 grados, lo cual supone una pendiente de 1. La correspondencia entre la línea de regresión y la línea de 45 grados es simplemente la medida de la fiabilidad. Una comparación de la orientación de la línea de regresión y la línea de 45 grados da una representación visual de la calidad relativa de los pronósticos. A medida que las disminuciones de calidad (comparando el análisis y los puntos extremos de pronóstico en las parcelas), son mayores la línea de regresión tiende más hacia la horizontal. Una línea horizontal significa que las observaciones son independientes de la previsión, o, dicho de otro modo, el valor medio de las observaciones es tan buen predictor como el método de pronóstico.

Los aportes de enmiendas orgánicas durante 6 años seguidos junto con la escasez de movilidad de estos metales permiten que los modelos no se vean afectados por cambios entre los lugares de experimentación y favorecen una predicción de las cantidades que se encuentran presentes en el suelo. La capacidad del cromo VI para la adsorción (Bradl 2004) se encuentra ligada a la presencia de hierro y manganeso forma de óxidos (dos variables incluidas en nuestro conjunto de datos del estudio). De manera similar para otros metales cuyo potencial tóxico es importante, como son el cadmio, plomo y el níquel, las discrepancias entre los modelos producidos por los datos originales y los de predicción son bajas, lo cual hace ver que son buenos predictores en los modelos. En este estudio de Bradl (2004) muestra con sus modelos como la presencia de materia orgánica del suelo también juega un papel importante en la adsorción de plomo. La materia orgánica del suelo puede inmovilizar Pb a través de reacciones de adsorción específicas, mientras que la movilización de Pb puede ser facilitada también por la materia orgánica disuelta o ácidos fúlvicos. Por otro lado, los metales involucrados en los procesos de los sistemas enzimáticos de las plantas, y por lo tanto más proclives a ser movilizados por necesidades de la planta en su desarrollo y funcionamiento, como son el zinc y cobre nos muestran disminuciones en los coeficientes de correlación especialmente en el caso del zinc (Chen *et al.* 2006). La posibilidad de que la planta permita la absorción y traslocación de estos metales desde el suelo a otras partes puede que establezca una caída en los ajustes de los coeficientes de correlación de estos modelos. Así, por ejemplo, en Impellitteri *et al.* (2003), las ecuaciones que establecieron de cobre eran dependientes de la cantidad de materia orgánica y el pH en un 70% de la variabilidad y en el zinc el pH en un 75% y en Wu *et al.* (2010) establecen que la relación de disponibilidad de cobre presentaba una tendencia de distribución espacial en altas proporciones dependiendo de las regiones de la zona de estudio. Las propiedades del suelo que influyen en la distribución espacial de la disponibilidad de Cu fueron SOM (materia orgánica del suelo) y pH, además de la concentración de Cu disponible. Este último punto de la concentración disponible es importante porque no es algo incluido en el conjunto de datos que nuestro modelo haya podido incorporar. Aunque en Bradl (2004) establecen

que los factores importantes que determinan la concentración de Zn en los suelos, son el pH, el contenido mineral de arcilla, CEC, la materia orgánica del suelo y el tipo de suelo lo cual debería ayudar a que nuestro modelo se ajustara bien a los datos, esto no ocurre.

En el caso de los componentes del suelo, principalmente en el caso del nitrógeno se produce un desajuste del modelo con caídas del coeficiente de correlación hasta 0, lo cual nos indica que las variables que participan en este modelo deben ser revisadas y sustituidas por aquellas que no muestren cambios tan bruscos con las parcelas experimentales. En el estudio de Walker *et al.* (2001) las relaciones resultantes entre la tasa de crecimiento relativo de la planta (RGR) y la concentración de N variaron entre formas lineales y curvilíneas (o posiblemente bi-lineales) dependiendo de las condiciones de tratamiento. El N recién adquirido se canalizó directamente a los sitios de mayor demanda, donde se asimiló rápidamente. Debido a estas asimilaciones rápidas las medidas de nitrógeno que se tiene en nuestro conjunto de datos no permiten tener unos diagramas de dispersión que aseguren un buen modelo de predicción (Ros *et al.* 2009). Una explicación sobre las variaciones, tan grandes, entre los valores de los modelos de los datos de origen con los de validación puede verse en los árboles de regresión que se han mostrado en capítulos anteriores. En dichos árboles se ve como las variables que más se relacionan para la determinación de la materia orgánica es el nitrógeno y viceversa, es por esto que, las variaciones existentes entre las parcelas experimentales al ser aplicadas sobre un solo conjunto de datos provoca caídas en los ajustes. En el estudio sobre modelos matemáticos para el carbono y el nitrógeno en el suelo realizado por Manzoni y Porporato (2009), donde se lleva a cabo un análisis histórico, revela que la complejidad, el grado y el número de no linealidades en general aumentaron con los años, mientras que disminuyeron con el aumento de la escala espacial y temporal de trabajo. También se encontró que las formulaciones matemáticas desarrolladas específicamente para ciertas escalas a menudo tienden a ser utilizados también en otras escalas espaciales y temporales diferentes a las originales, posiblemente resultando en inconsistencias entre formulaciones teóricas y de aplicación del modelo. Y concluye que, es crítico que los futuros esfuerzos de modelado representen la escala de la dependencia de sus formulaciones matemáticas con cuidado. Así, mientras que la precisión del pronóstico puede decirnos mucho sobre el pasado, como ocurría con los modelos desarrollados con los datos de origen, se debe recordar estas limitaciones cuando se utiliza pronósticos para predecir el futuro.

Observando los coeficientes del fósforo también, existe una disminución en el ajuste, pero resulta menos acusada que la ocurrida con la materia orgánica y el nitrógeno. Podemos asumir que está dentro de unos rangos adecuados para la interpretación y desarrollo de modelos propuestos.

6.6 Conclusiones

Los modelos de predicción para conocer las cantidades presentes en el suelo con adiciones de biosolidos en cultivos ajustan muy bien para metales poco móviles.

Aquellos metales que resultan más móviles entran dentro de un modelo de ajuste con valores razonablemente adecuados para la predicción.

Deben establecerse nuevos desarrollos de ajustes para una adecuada predicción de las variables del suelo ligadas a la calidad del mismo, materia orgánica, nitrógeno, e incluso el fósforo, debido a sus discrepancias entre los valores de predicción y los valores obtenidos en laboratorio (reales).

En resumen, se puede decir que los buenos resultados obtenidos por los modelos desarrollados para este estudio pueden ser usados como una estrategia eficaz para la gestión y del control de la adición de residuos en los cultivos.

7. Métodos basados en combinación de regresores para la mejora en las predicciones

7.1. Análisis de los resultados mediante la comparación de métodos.

Una de las maneras de enriquecer el conocimiento sobre un tema de estudio y de mejorar los datos obtenidos es realizar un contraste de las fuentes, como por ejemplo, realizando medidas con distintos aparatos, tener unos patrones para las calibraciones, y también mediante comparación de los resultados que se dan. En numerosas disciplinas del conocimiento se establece una búsqueda de segundas o terceras opiniones, de forma previa a la toma de decisiones definitivas sobre ciertos asuntos. Este tipo de consultas se hacen sopesando los consejos dados por un conjunto de expertos o comité (Dietterich 2000)(Rokach, 2010). La idea de construir un modelo mejorado mediante la combinación de modelos más simples y el conseguir automatizar la valoración de los expertos con objeto de mejorar los resultados es el objetivo de los métodos basados en *ensembles*.(Polikar 2006).

Los métodos *ensemble* utilizan una combinación de las predicciones obtenidas por los diferentes algoritmos de aprendizaje para dar como resultado una sola predicción con el fin de mejorar la generalización (posibilidad de extrapolar a otros conjuntos de datos) o la robustez sobre el estimador (Opitz y Maclin 1999). Esto puede suponer una mejora inmediata en los resultados obtenidos, basada principalmente en un aumento del coeficiente de correlación del modelo.

Dentro de los numerosos métodos de *ensemble* de aplicación podemos diferenciar dos tipos:

1. Métodos de promediado (*Averaging methods*). Este tipo de métodos se basan en la construcción de varios estimadores de manera independiente y posteriormente se hace un promedio de las predicciones obtenidas por ellos. El algoritmo resultante de la combinación habitualmente establece, una mejora de los algoritmos de partida porque se produce una reducción de la varianza. (Ejs: *Bagging, Random Forest*)

2. Métodos de refuerzo (*Boosting methods*). Estos métodos se construyen con unos algoritmos de base elaborados secuencialmente buscando la disminución del sesgo del algoritmo de combinación. Se busca el establecer un conjunto resultante más potente por la combinación de modelos menos robustos. (Ej: *Adaboost*) (Freund y Schapire 1997).

De igual modo, según el enfoque que siga el *ensemble* se puede clasificar en:

1. *Ensembles* con el fin de disminuir la varianza (Ej:*Bagging*)

2. *Ensembles* para disminuir el sesgo (Ej:*Boosting*) basado en la idea de que es más fácil encontrar y promediar muchas reglas básicas, que

encontrar una sola y altamente exacta regla de predicción (Schapire 2003).

3. *Ensembles* para la mejora de la fuerza predictiva (Ej: *Stacking*)

Por otro lado dependiendo del tipo de predictores que participan en el *ensemble* se pueden clasificar en:

1. *Ensembles* homogéneos. Aquellos en que los predictores base son del mismo tipo.

2. *Ensembles* heterogéneos. En los predictores base que forman el *ensemble* son de distinto tipo (Ej: *Vote, Stacking*)

La mejora de los resultados obtenidos se produce cuando existe una diversidad significativa entre los modelos utilizados. De nada sirve tener un conjunto de expertos si todos ellos opinan lo mismo y por lo tanto podrían ser reemplazados por un solo.

7.2. Métodos basados en combinación de regresores

AdaBoost. Es un algoritmo de refuerzo (o *boosting*) que tiende a reducir el sesgo de cada observación del conjunto de entrenamiento *m*. Se busca que en cada iteración el peso de las instancias bien clasificadas disminuya y el peso de las instancias incorrectamente clasificadas aumente. Debido al cambio de los pesos el clasificador se centrará en aquellas instancias que son más difíciles de clasificar. En este caso se utilizará *AdaBoostR* que es capaz de funcionar con conjuntos de datos de regresión.(Freund y Schapire 1996)

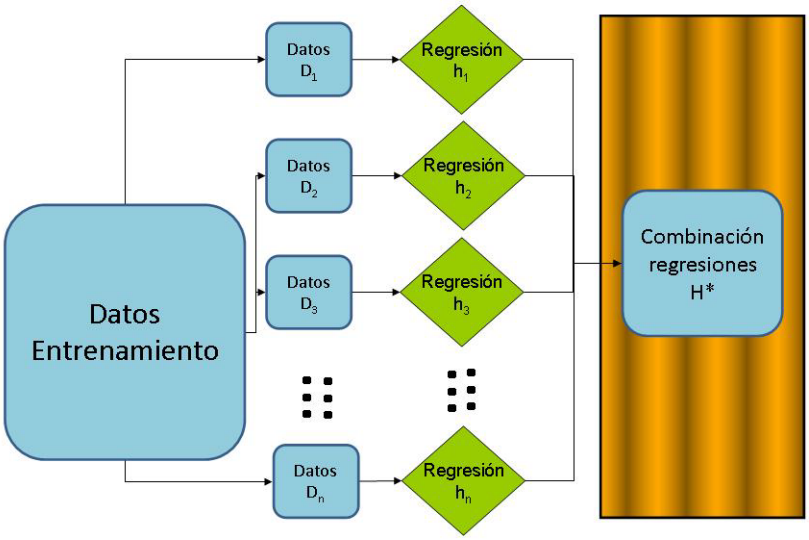


Figura 7-1 Esquema de funcionamiento de *AdaBoost.*

AdditiveRegression (*AdReg*). Es un metaclasificador que mejora el rendimiento mediante de un conjunto de modelos base aplicados a un problema de regresión. En cada iteración se ajusta un modelo según los residuos que se obtienen de aplicar mínimos cuadrados en cada iteración anterior (Friedman, 1999). Los residuos obtenidos primeramente son el gradiente de la función de pérdida que es minimizada con respecto a los valores del modelo en cada punto. La predicción se lleva a cabo mediante la adición de las predicciones de cada clasificador. Este metaclasificador impulsa modelos de regresión mediante el ajuste de una función simple parametrizada de forma secuencial (*Base learner*) para actualizar los "pseudo" residuos por mínimos cuadrados en cada iteración. Estos pseudo-residuos son el gradiente de la función de pérdida que se reduce al mínimo, con respecto a los valores del modelo en cada punto de los datos de entrenamiento evaluados en el paso actual. Se muestra que tanto la precisión de aproximación y la velocidad de ejecución de gradiente *Boosting* se pueden mejorar sustancialmente mediante la incorporación de la asignación al azar en el procedimiento. En concreto, en cada iteración una submuestra de los datos de entrenamiento se extraerá al azar (sin reemplazo) a partir del conjunto de datos de entrenamiento completo. Esta submuestra seleccionada al azar se utiliza a continuación en lugar de la muestra completa para adaptarse al *Base learner* y calcular la actualización de modelo para la iteración actual. Este enfoque aleatorio también aumenta la robustez contra el exceso de capacidad del Base learner. La reducción de la contracción de parámetros (tasa de aprendizaje) ayuda a prevenir el sobreajuste (*Overfitting*) y tiene un efecto de suavizado pero aumenta el tiempo de aprendizaje.

El funcionamiento del clasificador es empezar con un *ensemble* vacío, para ir añadiendo elementos secuencialmente. En el primer paso se utiliza un algoritmo de regresión, por lo cual se obtienen unos residuos, que serán la diferencia entre el valor predicho y el real. Después se utiliza otro algoritmo de predicción, que en esta ocasión trata de predecir los valores de los residuos. Al finalizar el cálculo, se añaden las predicciones hechas por el segundo modelo a las del primero. Si continúan existiendo errores, se vuelve a realizar la operación con un tercer algoritmo y así sucesivamente.

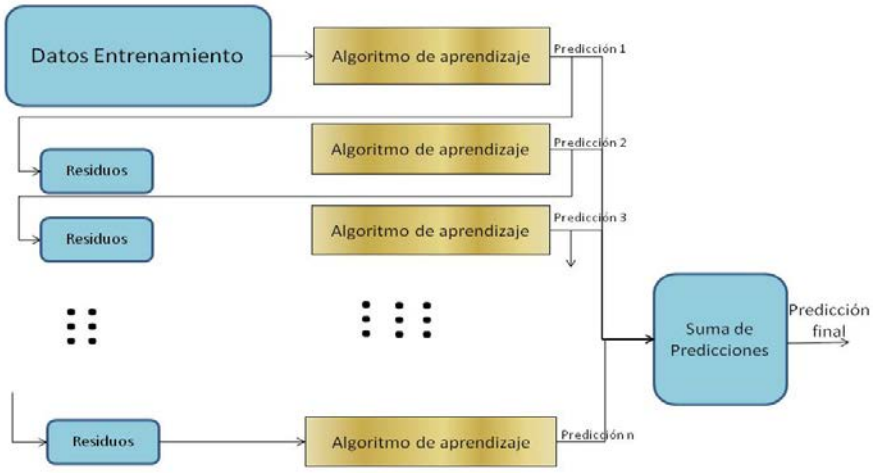


Figura 7-2 Esquema de funcionamiento de *AdditiveRegression*.

Bootstrap aggregating (*Bagging*). Es un método para generar múltiples versiones de un predictor que posteriormente se combinan para producir una decisión final (Breiman, 1996). Las múltiples versiones se forman entrenando cada uno de los modelos base con versiones modificadas del conjunto de entrenamiento original obtenidas mediante muestreo con reemplazamiento (*bootstrap)*. Este método mejora su precisión respecto al uso de un único modelo base dado que aporta estabilidad en el proceso de generación de nuevos *datasets* ya que las perturbaciones que pueden causar cambios significativos en el predictor construido mediante *bagging* disminuyen. Después de la construcción de varios modelos de regresión, se realiza una media de los valores obtenidos para calcular una predicción final.

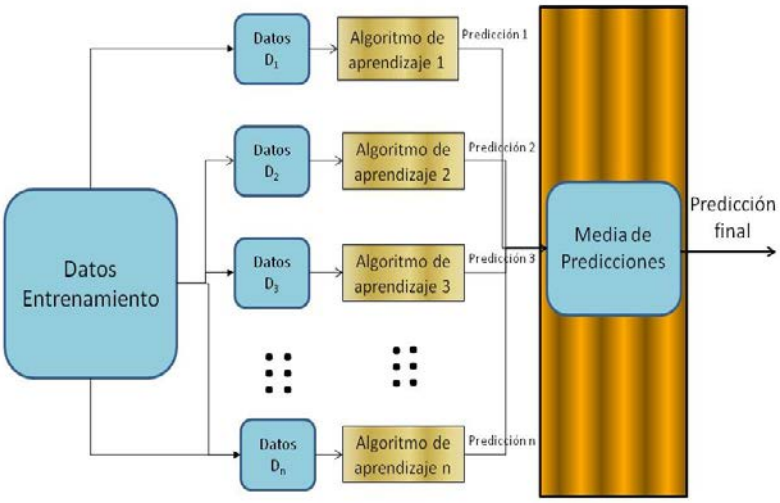


Figura 7-3 Esquema de funcionamiento de *Boostrap aggregating*.

CVParameterSelection. Es un método para la realización de la selección de parámetros de validación cruzada para cualquier clasificador. Este meta-

clasificador puede optimizar a través de un número arbitrario de parámetros, con sólo un inconveniente (la posibilidad de una enorme cantidad de posibles combinaciones de parámetros): sólo permite opciones únicas y directas del clasificador base (Kohavi y John 1997). A diferencia de los otros métodos descritos en esta lista, en este método se crean múltiples predictores o modelos base, pero la predicción se obtiene exclusivamente de uno de ellos, aquel que tenga el mejor rendimiento de acuerdo a una validación cruzada interna. Así, por ejemplo, si el funcionamiento de un predictor base se rige por 3 parámetros. El primero puede tomar 10 valores diferentes, el segundo 10 y el tercero 3. La validación cruzada para evaluar el modelo tiene 5 folds, en total se van a construir 10x10x3x5=1500 modelos diferentes.

Random Forest. Es un método que combina un conjunto de árboles de decisión tal que cada árbol depende de los valores de un vector aleatorio probado independientemente y con la misma distribución para cada uno de estos. La idea principal de este método es obtener numerosos árboles aunque sean producto de modelos ruidosos y posiblemente imparciales que al verse combinados produce una reducción de los errores. Se puede elegir el número de árboles que participa en el modelo en cada combinación para, con incremento del gasto computacional, aumentar la mejoría en los resultados. *Random Forest* se puede ver como *bagging* de *RandomTrees* y los *RandomTrees* son árboles que en cada nodo eligen el mejor atributo entre un subconjunto aleatorio del total de atributos.

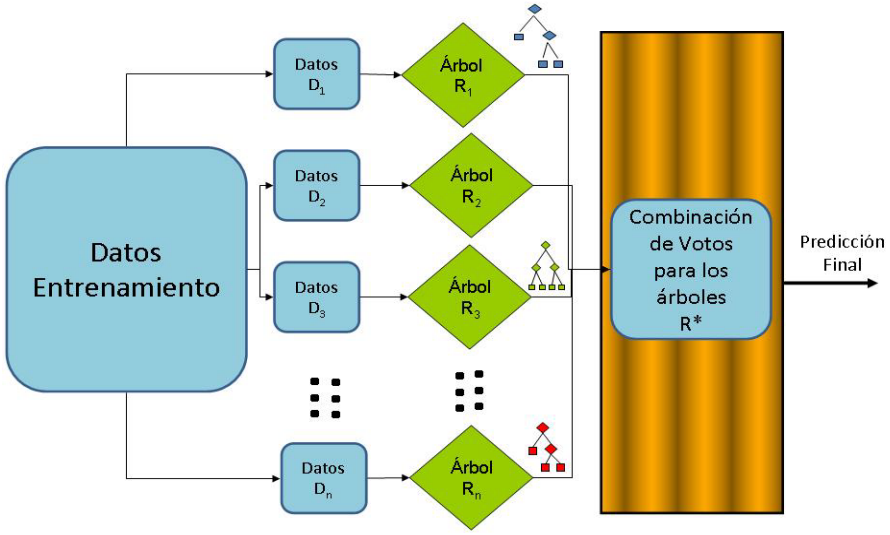


Figura 7-4 Esquema de funcionamiento de *Random Forest*.

Stacking. Este método combina múltiples modelos generados por diferentes algoritmos de aprendizaje a partir de un mismo conjunto de datos, usándose sobre todo con una gran variedad de modelos diferentes. Siendo su objetivo, alcanzar la mayor precisión generalizando el modelo. Este método utiliza un algoritmo de aprendizaje en un nivel diferente (*Metalearner*, nivel 2) para

determinar cuál es la mejor combinación de salida de los algoritmos utilizados como regresores (base, nivel 1) (Wolpert 1992);(Witten y Ting 1999). El algoritmo de aprendizaje del nivel 2 no necesita ser un algoritmo muy complejo ya que la mayor parte del trabajo es realizado en el nivel 1.De esta manera en el nivel 1, a partir de un mismo conjunto de datos, se obtienen predicciones diferentes según la utilización de los diferentes algoritmos de aprendizaje. Posteriormente, el nivel 2 tiene tantos atributos de entrada como algoritmos de aprendizaje tenía el nivel 1. A partir de este nivel 2 se producen las combinaciones para las salidas predictivas de los algoritmos de aprendizaje del nivel 1 mediante un algoritmo que obtiene la predicción final. Se trata de seleccionar cada uno de los algoritmos del nivel 1 cuando es conveniente usarlo, evitando el uso de sistemas de votos que puede ampliar el error de predicción, ya que con el procedimiento de voto no se determina correctamente cuál es el algoritmo con más precisión. Los problemas aparecen para adaptar las salidas del nivel 1 como entradas en el nivel 2.

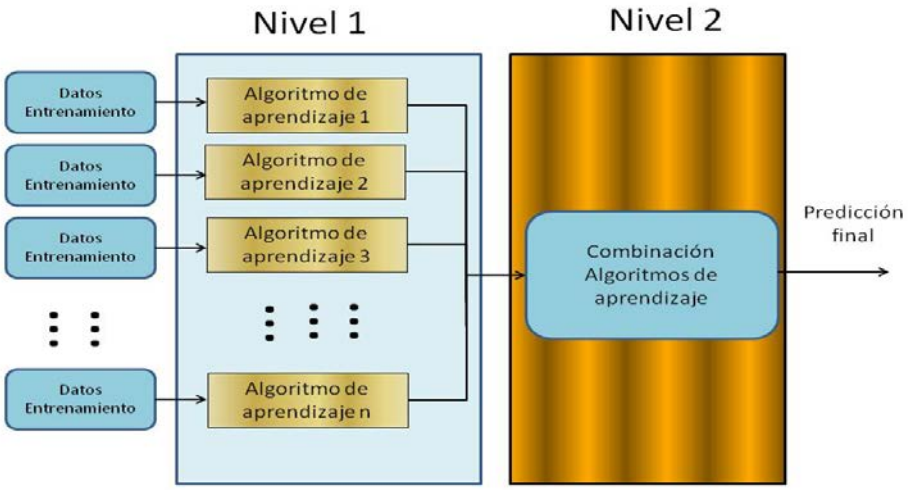


Figura 7-5 Esquema de funcionamiento de *Stacking.*

Vote. En este caso la manera de combinarse es a través de una combinación simple de predictores. Se construye un esquema de combinación basado en un voto, utilizando diferentes patrones, a partir de varios algoritmos de aprendizaje base (Kuncheva 2004);(Kittler *et al.* 1998). Utilizando diferentes algoritmos base, se obtienen diferentes valores de regresión para el problema planteado, y como es normal, entre los modelos calculados se obtienen mejores y peores resultados. Para obtener la mejor predicción de una nueva instancia se aplica otro algoritmo, cuyas entradas son los resultados obtenidos con los modelos regresores base, y cuya salida, basada en un sistema de voto, devuelve un resultado mucho más adaptado al problema. En casos de regresión, el esquema de combinación de predicciones más usado es la realización de medias. Esta combinación es simple, se seleccionan todas las predicciones de los modelos calculados y se calcula su media.

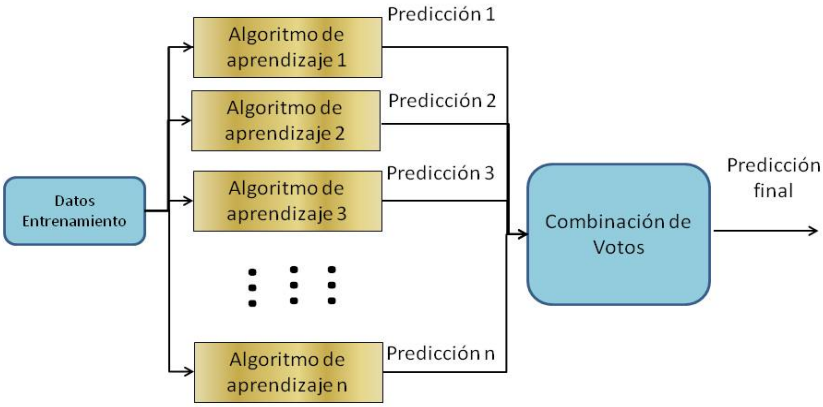


Figura 7-6 Esquema de funcionamiento de *Vote*.

7.3. Aplicación de los métodos ensemble para los datos de estudio.

Este estudio proporciona una comparativa entre los distintos métodos de *ensemble* para los árboles de regresión. Los árboles de regresión mediante estos métodos combinan los puntos fuertes de los algoritmos. El modelo final puede ser entendido como un modelo de regresión mejorado.

Con los métodos basados en combinación de regresores descritos anteriormente se procede a su utilización con cada una de las variables estudiadas en los datos. Se realiza una comprobación de las posibles mejoras en los resultados obtenidos mediante estos métodos frente a los inicialmente estudiados. Dado que nuestro principal propósito es seleccionar los mejores modelos de predicción, no nos condiciona que se obtengan un gran número de árboles sin interpretación visual posible.

En este capítulo se comparan los métodos *ensembles* con el conjunto de datos buscando cuál puede ser más beneficioso para la obtención de un mejor modelo para cada variable frente a los árboles de decisión usados, estos algoritmos generan componentes individuales en los sistemas proporcionando disminución en los errores de las predicciones. Para esta experiencia hemos partido del conjunto de datos de las tres parcelas experimentales analizadas en los capítulos anteriores, comparando la combinación de parámetros que mejores resultados daba, con respecto al coeficiente de correlación, y una combinación de parámetros en la cual se limita la profundidad a 1 y cuyo resultado da un valor más bajo de dicho coeficiente. La selección de

profundidad 1 es debida a que todos los coeficientes resultantes siempre son menores que los obtenidos por la mejor combinación de parámetros. Todos los modelos han sido probados y se ha realizado una validación cruzada de 10.

Los métodos *ensemble* que se han probado han sido:

1. *AdaBoost*
2. *Additive Regression*
3. *Bootstrap aggregating(Bagging)*
4. *CVParameterSelection*
5. *Random Forest*
6. *Stacking*
7. *Vote*

En los casos estudiados se ha visto que un cambio en el valor en la profundidad del árbol era suficiente para que el uso de ensembles satisfaga la premisa de estudio. Para el método de *Random Forest*, por las características de sus parámetros, se han variado el número de árboles (10, 20, 30 y 40) usados para el análisis.

En los métodos *Stacking* y *Vote* cuyo funcionamiento se basa en combinaciones de varios regresores se han utilizado:

1. *RandomTree*
2. *M5P*
3. *REPTree* con dos variantes, aquella con el mejor resultado de coeficiente de correlación y otra con idénticos parámetros cambiando la profundidad a 1.
4. *DecisionStump*

En Vote se ha utilizado como regla de combinación *Average of probabilities* (la cual devuelve la media de las probabilidades para cada uno de los clasificadores base seleccionados)

Debido a que *Stacking* es un metaclasificador con dos niveles, en el segundo se ha seleccionado una regresión lineal por el suavizado en las fronteras de decisión que este método permite y algunos autores recomiendan (Breiman 1996). Esto hará que el resultado de salida sea una combinación lineal de los algoritmos que se usan en el nivel 1.

7.4. Resultados

En primer lugar se muestran las tablas con las variables del suelo. Junto a ellas se incluye un ranking de los algoritmos que fueron mejores. La tabla de clasificación muestra el número de victorias estadísticamente significativas (de acuerdo al test t de student con una confianza de 0,05) que cada algoritmo ha tenido en comparación con todos los demás algoritmos del conjunto de datos, en este caso se ha usado el coeficiente de correlación del *REPTree* como base de comparación. Una victoria (>), significa una precisión que es mejor que la exactitud de otro algoritmo y que la diferencia fue estadísticamente significativa. Por otro lado también se realiza una aproximación a las predicciones de cada modelo y su representación gráfica para poder observar las diferencias con respecto a los datos del algoritmo inicial.

Tabla 7-1: Aplicación a la materia orgánica del suelo.

	CORR	MAE	RMSE	RAE	RRSE	Ranking Métodos		
						≥	>	<
RandomForest 40	**0,8437*** **	0,1403	0,1917	52,4959	56,2012	14	14	0
RandomForest 30	**0,8396*** **	0,1413	0,1932	52,8675	56,625	12	13	1
RandomForest 20	**0,8330*** **	0,143	0,1954	53,5146	57,2665	10	12	2
RandomForest 10	**0,8132*** **	0,1482	0,202	55,4973	59,2421	8	11	3
Stacking	**0,7799*** **	0,1546	0,2129	57,8764	62,4665	6	10	4
Vote	**0,7314*** **	0,1817	0,2414	67,9823	70,733	4	9	5
AdReg 4.9.8	**0,6852*** **	0,1874	0,2481	70,2105	72,8614	0	6	6
Bagging 4.9.8	**0,6673*** **	0,1931	0,2556	72,2089	74,8897	0	6	6
Adaboost 4.9.8	**0,6510** **	0,2008	0,2625	75,0947	76,9616	0	6	6
CVParameter 4.9.8	**0,6125**	0,2024	0,2691	75,7108	78,9012	-5	4	9
REPTree 4.9.8	0,6076	0,2035	0,2709	76,1453	79,433	-5	4	9
AdReg 1.9.8	0,5268**	0,2189	0,2903	81,9018	85,1047	-10	1	11
Bagging 1.9.8	0,5019***	0,2264	0,2973	84,6336	87,1041	-10	1	11
Adaboost 1.9.8	0,4904***	0,2426	0,3093	90,7344	90,6962	-10	1	11
CVParameter 1.9.8	0,4234***	0,2359	0,3085	88,214	90,4261	-14	0	14

Resultados obtenidos de testear los modelos seleccionados frente a *REPTree*. En negrita los modelos mejores y las diferencias significativas en CORR son indicadas ***, P< 0,001; **, P< 0,01; *, P < 0.05. CORR es Coeficiente de correlación, MAE Error absoluto medio; RMSE Raíz del error cuadrático medio; RAE Error absoluto relativo; RRSE Raíz del error cuadrático relativo.

Tabla 7-2: Aplicación al nitrógeno en el suelo.

	CORR	MAE	RMSE	RAE	RRSE	Ranking Métodos		
						≥	>	<
RandomForest 40	**0,8857*** **	0,016	0,0235	45,3356	48,8109	10	10	0
RandomForest 30	**0,8830*** **	0,0162	0,0237	45,7181	49,1834	10	10	0
RandomForest 20	**0,8784*** **	0,0164	0,0239	46,2693	49,7241	10	10	0
RandomForest 10	**0,8675*** **	0,0169	0,0247	47,6051	51,2106	10	10	0
Stacking	**0,8471*** **	0,0172	0,0255	48,6528	53,0093	8	8	0
Vote	**0,8122*** **	0,022	0,0309	61,9415	64,086	4	8	4
Adaboost 4.5.8	**0,7919*** **	0,0221	0,0302	62,4095	62,8935	2	6	4
Bagging 4.5.8	**0,7466** **	0,0236	0,0326	66,5402	67,7959	-3	3	6
AdReg 4.5.8	**0,7237** **	0,0235	0,0335	66,254	69,5316	-3	3	6
REPTree 4.5.8	0,6691	0,0256	0,0358	72,3704	74,4066	-5	2	7
CVParameter 4.5.8	0,6491	0,0261	0,0366	73,6248	76,0266	-5	2	7
AdReg 1.5.8	0,5698	0,0289	0,0391	81,6472	81,2758	-7	0	7
Adaboost 1.5.8	0,5219**	0,0322	0,0423	90,937	87,922	-9	0	9
Bagging 1.5.8	0,4664***	0,0307	0,0432	86,4079	89,5574	-10	1	11
CVParameter 1.5.8	0,3860***	0,0314	0,0445	88,3942	92,252	-12	0	12

Resultados obtenidos de testear los modelos seleccionados frente a *REPTree*. En negrita los modelos mejores y las diferencias significativas en CORR son indicadas ***, P< 0,001; **, P< 0,01; *, P < 0.05. CORR es Coeficiente de correlación, MAE Error absoluto medio; RMSE Raíz del error cuadrático medio; RAE Error absoluto relativo; RRSE Raíz del error cuadrático relativo.

Tabla 7-3: Aplicación al fósforo en el suelo.

	CORR	MAE	RMSE	RAE	RRSE	Ranking Métodos		
						≥	>	<
RandomForest 40	**0,9293*** **	8,837	12,4005	35,0054	38,3689	12	12	0
RandomForest 30	**0,9282*** **	8,877	12,4711	35,1631	38,5864	12	12	0
RandomForest 20	**0,9257*** **	8,9703	12,6434	35,5255	39,1085	11	11	0
RandomForest 10	**0,9194*** **	9,2272	12,9999	36,5435	40,2183	7	10	3
Stacking	**0,9137*** **	9,2526	13,0837	36,7013	40,5951	8	10	2
Vote	**0,8707*** **	13,404	17,7215	53,0583	54,8137	4	9	5
Adaboost 4.5.8	**0,8459*** **	13,7816	17,7358	54,614	54,9495	1	7	6
AdReg 4.5.8	**0,8263*** **	13,3415	18,3766	52,8977	56,9578	0	6	6
Bagging 4.5.8	**0,8195*** **	14,2757	18,8901	56,5669	58,4647	-1	6	7
REPTree 4.5.8	0,7684	15,5327	20,7027	61,5608	64,1402	-6	3	9
CVParameter 4.5.8	0,7635	15,6145	20,8846	61,8532	64,6736	-6	3	9
AdReg 1.5.8	0,7437	16,1555	21,7568	64,0115	67,3174	-6	3	9
Bagging 1.5.8	0,5416***	21,6075	27,5374	85,4859	85,1698	-11	1	12
Adaboost 1.5.8	0,5294***	22,1812	27,8382	87,7939	86,1802	-11	1	12
CVParameter 1.5.8	0,4699***	22,3508	28,5766	88,46	88,4365	-14	0	14

Resultados obtenidos de testear los modelos seleccionados frente a *REPTree*. En negrita los modelos mejores y las diferencias significativas en CORR son indicadas ***, P< 0,001; **, P< 0,01; *, P < 0.05. CORR es Coeficiente de correlación, MAE Error absoluto medio; RMSE Raíz del error cuadrático medio; RAE Error absoluto relativo; RRSE Raíz del error cuadrático relativo.

Tabla 7-4: Aplicación al cadmio en el suelo.

	CORR	MAE	RMSE	RAE	RRSE	Ranking Métodos		
						≥	>	<
RandomForest 40	**0,9606*** **	0,1547	0,1972	27,8433	28,4691	14	14	0
RandomForest 30	**0,9598*** **	0,1557	0,1989	28,0189	28,7112	11	12	1
RandomForest 20	**0,9586*** **	0,1575	0,2014	28,3395	29,0722	11	12	1
RandomForest 10	**0,9540*** **	0,1633	0,2103	29,3945	30,354	7	10	3
Stacking	**0,9491*** **	0,1637	0,2164	29,4789	31,2877	7	10	3
Adaboost 4.9.8	**0,9254*** **	0,2142	0,2699	38,5482	38,9719	2	7	5
AdReg 4.9.8	**0,9247*** **	0,2029	0,2631	36,5616	38,04	2	7	5
Vote	**0,9187*** **	0,2601	0,3212	46,7262	46,2933	2	7	5
Bagging 4.9.8	**0,8970** **	0,2417	0,3058	43,5275	44,1832	-2	6	8
CVParameter 4.9.8	**0,8769**	0,2567	0,3313	46,2156	47,8584	-6	3	9
REPTree 4.9.8	0,876	0,2575	0,3322	46,3721	47,999	-6	3	9
AdReg 1.9.8	0,8444	0,2977	0,3655	53,5037	52,7102	-6	3	9
Adaboost 1.9.8	0,7756***	0,4123	0,5	74,0098	72,0062	-10	2	12
Bagging 1.9.8	0,6382***	0,4396	0,5544	78,8999	79,8186	-12	1	13
CVParameter 1.9.8	0,5093***	0,4754	0,598	85,3838	86,1141	-14	0	14

Resultados obtenidos de testear los modelos seleccionados frente a *REPTree*. En negrita los modelos mejores y las diferencias significativas en CORR son indicadas ***, P< 0,001; **, P< 0,01; *, P < 0.05. CORR es Coeficiente de correlación, MAE Error absoluto medio; RMSE Raíz del error cuadrático medio; RAE Error absoluto relativo; RRSE Raíz del error cuadrático relativo.

Tabla 7-5: Aplicación al cromo en el suelo.

	CORR	MAE	RMSE	RAE	RRSE	Ranking Métodos		
						≥	>	<
RandomForest 40	**0,9478*** **	2,3981	3,6187	27,3553	32,8804	13	13	0
RandomForest 30	**0,9464*** **	2,4196	3,6523	27,6002	33,1852	12	12	0
RandomForest 20	**0,9442*** **	2,4574	3,7115	28,0311	33,7403	11	12	1
RandomForest 10	**0,9368*** **	2,5875	3,8914	29,5234	35,3977	7	10	3
Stacking	**0,9315*** **	2,572	3,9412	29,3784	35,876	7	10	3
Vote	**0,8721*** **	4,3736	5,8718	49,8923	53,5799	3	8	5
Adaboost 4.9.8	**0,8621*** **	4,3041	5,7771	49,1145	52,7467	2	7	5
Bagging 4.9.8	**0,8478*** **	4,352	6,0025	49,6248	54,7254	1	7	6
AdReg 4.9.8	**0,7945** **	4,8064	6,6549	54,977	60,9111	-3	5	8
CVParameter 4.9.8	**0,7587**	5,2101	7,1125	59,4953	65,0187	-6	3	9
REPTree 4.9.8	0,7559	5,2356	7,1507	59,8304	65,4074	-6	3	9
AdReg 1.9.8	0,751	5,401	7,1254	61,8135	65,2944	-5	3	8
Adaboost 1.9.8	0,5943***	7,3176	9,1394	83,4655	83,6111	-11	1	12
Bagging 1.9.8	0,5722***	7,0862	9,0586	80,8238	82,8279	-11	1	12
CVParameter 1.9.8	0,5055***	7,3996	9,4402	84,4171	86,3568	-14	0	14

Resultados obtenidos de testear los modelos seleccionados frente a *REPTree*. En negrita los modelos mejores y las diferencias significativas en CORR son indicadas ***, P< 0,001; **, P< 0,01; *, P < 0.05. CORR es Coeficiente de correlación, MAE Error absoluto medio; RMSE Raíz del error cuadrático medio; RAE Error absoluto relativo; RRSE Raíz del error cuadrático relativo.

Tabla 7-6: Aplicación al cobre en el suelo.

	CORR	MAE	RMSE	RAE	RRSE	Ranking Métodos		
						≥	>	<
RandomForest 40	**0,9417*** **	0,8477	1,1167	33,8754	35,3714	14	14	0
RandomForest 30	**0,9401*** **	0,855	1,1264	34,1724	35,6855	12	13	1
RandomForest 20	**0,9374*** **	0,8686	1,1436	34,7233	36,2413	9	11	2
RandomForest 10	**0,9302*** **	0,907	1,191	36,2577	37,7501	7	10	3
Stacking	**0,9301*** **	0,8638	1,1548	34,5482	36,6392	8	10	2
Adaboost 4.3.8	**0,8847*** **	1,1912	1,5372	47,5482	48,6598	2	7	5
Vote	**0,8847*** **	1,3008	1,668	51,878	52,7793	2	7	5
Bagging 4.3.8	**0,8778*** **	1,2128	1,5636	48,4079	49,4934	2	7	5
AdReg 4.3.8	**0,8524*** **	1,2963	1,6728	51,8229	53,0722	-2	6	8
AdReg 1.3.8	**0,8196**	1,4257	1,8253	56,9756	57,8692	-6	3	9
REPTree 4.3.8	0,8101	1,4437	1,8491	57,671	58,6283	-6	3	9
CVParameter 4.3.8	0,8007	1,4737	1,8915	58,8506	59,9441	-6	3	9
Adaboost 1.3.8	0,6672***	1,9641	2,4803	78,3287	78,5851	-11	1	12
Bagging 1.3.8	0,6469***	2,0191	2,5698	80,481	81,3109	-11	1	12
CVParameter 1.3.8	0,4678***	2,1946	2,7955	87,5702	88,5718	-14	0	14

Resultados obtenidos de testear los modelos seleccionados frente a *REPTree*. En negrita los modelos mejores y las diferencias significativas en CORR son indicadas ***, P< 0,001; **, P< 0,01; *, P < 0.05. CORR es Coeficiente de correlación, MAE Error absoluto medio; RMSE Raíz del error cuadrático medio; RAE Error absoluto relativo; RRSE Raíz del error cuadrático relativo.

Tabla 7-7: Aplicación al hierro en el suelo.

	CORR	MAE	RMSE	RAE	RRSE	Ranking Métodos		
						≥	>	<
RandomForest 40	**0,9342*** **	1033,49	1427,61	33,0877	36,3096	13	13	0
RandomForest 30	**0,9332*** **	1038,17	1434,82	33,2436	36,5006	13	13	0
RandomForest 20	**0,9309*** **	1056,83	1454,33	33,8393	36,9875	9	11	2
Stacking	**0,9253*** **	1080,56	1484,36	34,6188	37,7537	8	10	2
RandomForest 10	**0,9243*** **	1100,94	1508,55	35,2477	38,3732	7	10	3
Vote	**0,9016*** **	1387,04	1880,33	44,3525	47,7016	3	8	5
Adaboost 4.3.8	**0,8963*** **	1351,63	1788,93	43,2707	45,4837	3	8	5
Bagging 4.3.8	**0,8829*** **	1381,8	1852,24	44,2372	47,0802	-1	6	7
AdReg 4.3.8	**0,8771*** **	1414,04	1889,59	45,3673	48,1693	-1	6	7
REPTree 4.3.8	0,8496	1547,58	2066,72	49,5792	52,5726	-6	3	9
CVParameter 4.3.8	0,8489	1548,72	2075,76	49,6246	52,7935	-6	3	9
AdReg 1.3.8	0,8471	1575,2	2099,55	50,4638	53,4463	-6	3	9
Bagging 1.3.8	0,6795***	2294,44	3005,64	73,3622	76,2175	-11	1	12
Adaboost 1.3.8	0,6477***	2545,45	3082,33	81,3907	78,3448	-11	1	12
CVParameter 1.3.8	0,5574***	2478,59	3273,59	79,3109	83,1	-14	0	14

Resultados obtenidos de testear los modelos seleccionados frente a *REPTree*. En negrita los modelos mejores y las diferencias significativas en CORR son indicadas ***, P< 0,001; **, P< 0,01; *, P < 0.05. CORR es Coeficiente de correlación, MAE Error absoluto medio; RMSE Raíz del error cuadrático medio; RAE Error absoluto relativo; RRSE Raíz del error cuadrático relativo.

Tabla 7-8: Aplicación al manganeso en el suelo.

	CORR	MAE	RMSE	RAE	RRSE	Ranking Métodos		
						≥	>	<
Stacking	**0,9568*** **	20,4176	27,1761	25,8762	29,0175	11	11	0
RandomForest 40	**0,9548*** **	21,2841	28,6267	26,9331	30,5133	12	12	0
RandomForest 30	**0,9539*** **	21,4208	28,8587	27,1074	30,763	12	12	0
RandomForest 20	**0,9522*** **	21,7279	29,2591	27,4949	31,1912	9	11	2
RandomForest 10	**0,9470*** **	22,7224	30,5352	28,7517	32,5534	6	10	4
Vote	**0,9201*** **	28,5215	37,4185	36,0994	39,893	2	7	5
Adaboost 4.3.8	**0,9168*** **	28,7619	37,7575	36,4062	40,2515	1	6	5
AdReg 4.3.8	**0,9141*** **	29,0951	38,0474	36,8433	40,5927	1	6	5
Bagging 4.3.8	**0,9104*** **	29,0386	38,8212	36,7684	41,4046	0	6	6
REPTree 4.3.8	0,8899	32,0776	42,6369	40,6192	45,4821	-6	3	9
CVParameter 4.3.8	0,888	32,3633	42,9731	40,9892	45,8367	-6	3	9
AdReg 1.3.8	0,8876	33,2333	43,4897	42,0645	46,3615	-6	3	9
Adaboost 1.3.8	0,8033***	44,1563	56,3517	55,8732	60,0568	-10	2	12
Bagging 1.3.8	0,8001***	43,7755	56,3258	55,3981	60,0285	-13	0	13
CVParameter 1.3.8	0,8001***	43,7709	56,3278	55,3922	60,0309	-13	0	13

Resultados obtenidos de testear los modelos seleccionados frente a *REPTree*. En negrita los modelos mejores y las diferencias significativas en CORR son indicadas ***, P< 0,001; **, P< 0,01; *, P < 0.05. CORR es Coeficiente de correlación, MAE Error absoluto medio; RMSE Raíz del error cuadrático medio; RAE Error absoluto relativo; RRSE Raíz del error cuadrático relativo.

Tabla 7-9: Aplicación al níquel en el suelo.

	CORR	MAE	RMSE	RAE	RRSE	Ranking Métodos		
						≥	>	<
RandomForest 40	**0,9406*** **	0,8519	1,3248	28,1043	33,9974	12	12	0
Stacking	**0,9401*** **	0,869	1,2971	28,7344	33,4615	10	10	0
RandomForest 30	**0,9394*** **	0,8616	1,3344	28,4249	34,252	11	11	0
RandomForest 20	**0,9376*** **	0,875	1,3505	28,8714	34,6807	10	11	1
RandomForest 10	**0,9311*** **	0,923	1,4084	30,4558	36,2074	7	10	3
Vote	**0,8884*** **	1,2862	1,8317	42,4137	47,1861	4	9	5
Adaboost 4.5.8	**0,8699*** **	1,4144	1,9275	46,7212	49,8055	1	7	6
Bagging 4.5.8	**0,8604*** **	1,3937	1,9767	46,0245	50,9244	1	7	6
AdReg 4.5.8	**0,837**	1,5166	2,1223	50,0983	54,7855	-4	4	8
CVParameter 4.5.8	**0,8319**	1,5354	2,1476	50,7411	55,412	-4	4	8
REPTree 4.5.8	0,8311	1,535	2,1538	50,6872	55,5521	-4	4	8
AdReg 1.5.8	0,8054*	1,6793	2,3008	55,4717	59,4395	-8	3	11
Bagging 1.5.8	0,7041***	2,0094	2,748	66,256	70,9806	-12	0	12
CVParameter 1.5.8	0,7041***	2,0091	2,7476	66,2451	70,9688	-12	0	12
Adaboost 1.5.8	0,6957***	2,1142	2,8559	69,6524	73,7161	-12	0	12

Resultados obtenidos de testear los modelos seleccionados frente a *REPTree*. En negrita los modelos mejores y las diferencias significativas en CORR son indicadas ***, P< 0,001; **, P< 0,01; *, P < 0.05. CORR es Coeficiente de correlación, MAE Error absoluto medio; RMSE Raíz del error cuadrático medio; RAE Error absoluto relativo; RRSE Raíz del error cuadrático relativo.

Tabla 7-10: Aplicación al plomo en el suelo.

	CORR	MAE	RMSE	RAE	RRSE	Ranking Métodos		
						≥	>	<
RandomForest 40	**0,9468*** **	1,3163	1,7351	32,6871	30,394	13	13	0
RandomForest 30	**0,9460*** **	1,3257	1,7463	32,8971	30,6091	12	12	0
RandomForest 20	**0,9449*** **	1,3378	1,7606	33,1682	30,8869	11	12	1
RandomForest 10	**0,9400*** **	1,3903	1,8238	34,3696	32,1122	7	10	3
Stacking	**0,9379*** **	1,3746	1,8314	34,5449	31,7636	7	10	3
Adaboost 4.3.8	**0,9035*** **	1,8322	2,3382	44,0828	42,3218	4	9	5
Vote	**0,8873*** **	2,0401	2,6463	49,8252	47,0417	0	6	6
Bagging 4.3.8	**0,8859*** **	1,9215	2,4979	47,0731	44,3708	0	6	6
AdReg 4.3.8	**0,8789*** **	1,9418	2,5517	48,1344	44,8625	0	6	6
AdReg 1.3.8	**0,8384**	2,2733	2,903	54,7533	52,5189	-6	3	9
CVParameter 4.3.8	**0,8311**	2,2602	2,9476	55,604	52,221	-6	3	9
REPTree 4.3.8	0,8296	2,2706	2,9646	52,4316	55,9104	-6	3	9
Adaboost 1.3.8	0,7444***	3,0989	3,8385	72,2708	71,4269	-10	2	12
Bagging 1.3.8	0,5997***	3,4086	4,2727	80,4718	78,6069	-12	1	13
CVParameter 1.3.8	0,5608***	3,5065	4,3996	82,8778	80,8775	-14	0	14

Resultados obtenidos de testear los modelos seleccionados frente a *REPTree*. En negrita los modelos mejores y las diferencias significativas en CORR son indicadas ***, P< 0,001; **, P< 0,01; *, P < 0.05. CORR es Coeficiente de correlación, MAE Error absoluto medio; RMSE Raíz del error cuadrático medio; RAE Error absoluto relativo; RRSE Raíz del error cuadrático relativo.

Tabla 7-11: Aplicación al zinc en el suelo.

	CORR	MAE	RMSE	RAE	RRSE	Ranking Métodos		
						≥	>	<
Stacking	**0,9379*** **	3,703	6,2961	28,5402	34,1963	10	10	0
RandomForest 40	**0,9364*** **	3,8227	6,5404	29,4456	35,4523	10	10	0
RandomForest 30	**0,9355*** **	3,8479	6,5852	29,6352	35,6795	10	10	0
RandomForest 20	**0,9332*** **	3,9283	6,6927	30,2576	36,2797	10	10	0
RandomForest 10	**0,9270*** **	4,1287	6,9647	31,8015	37,7186	10	10	0
Vote	**0,8904*** **	6,411	9,0036	49,3589	48,9445	0	5	5
AdReg 4.3.8	**0,8858*** **	5,7663	8,5347	44,531	46,5057	0	5	5
Adaboost 4.3.8	**0,8747*** **	6,6018	9,2182	50,9048	50,1042	0	5	5
AdReg 1.3.8	**0,8700*** **	6,3566	9,0554	49,0788	49,3558	-1	4	5
Bagging 4.3.8	**0,8694*** **	6,429	9,1294	49,6094	49,6985	0	5	5
CVParameter 4.3.8	**0,8419**	7,0665	9,8921	54,5354	53,9128	-6	3	9
REPTree 4.3.8	0,8394	7,1094	10,0004	54,8539	54,4456	-7	3	10
Adaboost 1.3.8	0,6716***	11,0688	14,1781	85,2165	77,1056	-10	2	12
Bagging 1.3.8	0,6103***	11,0348	14,4787	84,9483	78,7948	-13	0	13
CVParameter 1.3.8	0,6071***	11,0617	14,5097	85,1522	78,9664	-13	0	13

Resultados obtenidos de testear los modelos seleccionados frente a *REPTree*. En negrita los modelos mejores y las diferencias significativas en CORR son indicadas ***, P< 0,001; **, P< 0,01; *, P < 0.05. CORR es Coeficiente de correlación, MAE Error absoluto medio; RMSE Raíz del error cuadrático medio; RAE Error absoluto relativo; RRSE Raíz del error cuadrático relativo.

7.5. Discusión

Cuando se disminuye la profundidad de los árboles en los *ensembles* los resultados obtenidos ofrecen una reducción de los coeficientes de correlación, así como un aumento en los errores, Todos los métodos con profundidades de 1 tienen menores valores que su mismo método con profundidad de 4. Esto tiene sentido, ya que con una profundidad de 1 el árbol sólo tiene un nivel, por lo que las instancias se han de dividir en dos hojas en el caso de los atributos numéricos. A todas las instancias de la misma hoja les asigna la misma salida con lo cual los árboles de profundidad 1 son muy poco precisos.

Los métodos *Random Forest (RF)* y *Stacking* ofrecen un claro incremento del coeficiente de correlación con todas las variables. Cuando se utilizan los métodos *Random Forest* el aumento en la cantidad de árboles (desde 10 a 40) permite una mejoría en los resultados. *Random Forest*, a menudo (9 de 11

veces) da los mejores resultados en los coeficientes de correlación con tamaño de 40 salvo en las variables de manganeso y zinc. En esas dos variables el mejor resultado se da cuando se usa el método *Stacking*. En el caso de los Random Forest el incremento del número de árboles hace que aumente la precisión del método hasta el punto en que se produce una estabilización y no se producen mejoras. Cada vez más en las ciencias de la vida se están usando los métodos *Random Forest* para el reconocimiento de patrones en datos a escala ómica (Touw *et al.* 2013), dado que *RF* proporciona dos aspectos que son muy importantes para la Minería de Datos: alta precisión de predicción e información sobre importancia variable para la clasificación.

Cuando se utilizan los métodos con profundidades 1 los resultados de los coeficientes de correlación obtenidos con el *AdReg* son los mejores mientras que los resultados más bajos son los obtenidos con *CVParamete*r, excepto para el manganeso y el níquel que es *AdaBoost*.

Comparando dos métodos como son *AdaBoost* y *Bagging* se observa que *AdaBoos*t suele mejorar los resultados obtenidos, estos métodos incorporan importantes ventajas de los métodos basados en árboles, ya que manejan diferentes tipos de variables predictoras y acomodan los datos faltantes. No tienen necesidad de transformación previa de datos o eliminación de valores atípicos, pueden encajar relaciones no lineales complejas y manejar automáticamente los efectos de interacción entre predictores. Elith *et al.* (2008) muestran que el ajuste de árboles en *Boosting* supera el mayor inconveniente de los modelos de árbol único: su desempeño predictivo relativamente pobre y en nuestro estudio los resultados son similares.

En el estudio de Prasad *et al.* (2006) evaluando cuatro técnicas, *RF* y *Bagging* fueron superiores al reproducir las distribuciones de valor de importancia actual (una medida de área basal además de la abundancia) para las cuatro especies de árboles. Las estimaciones futuras de hábitat adecuado después del cambio climático fueron visualmente más razonables con *BT* y *RF*, con un desempeño ligeramente mejor por RF evaluado por las estadísticas de *Kappa*, las estimaciones de correlación y la distribución espacial de los valores de importancia. Aunque en nuestro estudio *Random Forest* obtiene un mejor resultado que los métodos *Bagging* o *AdaBoost*.

Al comparar los métodos vemos como *AdReg* funciona mejor cuando el árbol de decisión que se va a formar es muy pequeño, y que un incremento en el tamaño del árbol apenas mejora el resultado final, mientras que *CVParameter* tiene un aumento elevado en los coeficientes de correlación cuando se incrementa la profundidad del árbol. Dado que *AdReg* funciona haciendo que el siguiente regresor entrene con los residuos del regresor anterior, esto hace que si los resultados de salida de un árbol base son malos el margen de mejora en cada paso es mayor que si de inicio se tiene un árbol con buenos resultados de inicio. Cuando un árbol base es muy bueno, los residuos son muy pequeños y el nuevo árbol tiene que aprender con datos que tienen salidas muy uniformes y próximas a cero, esto dificulta las mejoras posteriores.

En los métodos que permiten profundidades de 4, los menores valores de coeficientes de correlación se dan con el método de *CVParameter* y en algunos casos como con el manganeso y el cobre, este método no mejora los valores originales del *REPTree*. Los mejores resultados se obtienen con el método de *Adaboost* excepto para las variables de materia orgánica y de zinc, para los que el método de *AdReg* mejora los valores de los coeficientes de correlación. *Adaboost* trabaja bien con conjuntos de datos con poca cantidad de ruido y es posible que en las dos variables anteriores se produzca un aumento de los ejemplos mal etiquetados. Debido a lo cual, para las mismas variables, el método de *AdReg*, dado que trabaja como metaclasificador, reduce la cantidad de parámetros de aprendizaje y tiene un efecto de suavizado frente a posibles datos ruidosos.

En todos los casos, realizando una comparativa entre los parámetros originales de estudio (profundidad, número de hojas,…) los métodos *ensembles* permiten un aumento en los coeficientes de correlación de entre el 0,04% (para el zinc) y 29,26% (para el nitrógeno) y una disminución de los errores (Raíz del error cuadrático medio, RMSE) entre 1,57% (zinc) y 103,29% (cromo).

Por otro lado, los mejores resultados se han obtenido, en conjunto, con los *Random Forest* permitiendo incrementos medios de los coeficientes entre 6,1% y 28,70%, respecto a los *REPTree*. (Breiman 2001; Breiman 2002) diseña los *Random Forest* para producir unas predicciones exactas en las que no tenga lugar un sobreajuste de los datos. También cabe destacar los buenos resultados que se obtienen con los métodos *Stacking*, si bien, estos resultan un poco más costosos en tiempo computacional por la complejidad de los modelos que incorporan en los dos niveles y su ajuste mediante regresiones lineales. En los métodos estudiados se puede comprobar la relación que existe entre la mejora de los coeficientes de correlación y la disminución de los errores.

Dadas las características propias de los métodos *ensemble* se pueden incluir algoritmos distintos de los árboles de decisión, si bien, en este caso nos hemos ceñido a los mismos por ser el objeto inicial de estudio. Añadir otros algoritmos diferentes puede permitir la mejora de los resultados aunque se ha visto como, para el conjunto de datos, algunas combinaciones ofrecen coeficientes de correlación aceptables para el estudio. Se puede comprobar que métodos de combinación de varios algoritmos, como *Stacking*, a pesar de realizar sólo una combinación de algoritmos de árboles de decisión consiguen una notable mejora de los resultados comparado con los árboles de decisión con parámetros modificados iniciales.

A continuación se muestran, mediante diagramas de dispersión, cómo se comportan los modelos con las distintas variables de estudio y la dependencia de la precisión de los valores de variables que hemos obtenido.

Cuando se observa las representaciones de los diagramas de dispersión (o *scatterplots*) de los datos reales frente a los calculados, se distinguen algunas tendencias que son útiles para la interpretación de los resultados. Así por ejemplo, es posible ver que sí existe una ligera tendencia a que los valores representados se sitúen por encima de la línea de 45º, lo que indica un sesgo

positivo, mientras que si los datos se sitúan por debajo, el sesgo será negativo. También se pueden establecer unos umbrales en los extremos. Cuando los datos observados están dentro de los rangos extremos y el valor de predicción es 0, se producen los llamados falsa alarma, y si no hay ningún valor de predicción y se encuentra en un rango extremo son los llamados eventos perdidos. En algunos casos cuando los valores objetivos son cero o cercanos a ese valor, existe el problema que con este tipo de valores, los errores relativos pueden ser altos, incluso si la predicción del método ha sido buena. Cuando esto ocurre, se puede resolver el problema transformando los datos (añadiendo una constante). El caso de los eventos perdidos se establecería que nuestro modelo predice valores por encima de los valores reales medidos.

Para los datos clasificados como “falsa alarma” puede ser que nuestro modelo haga una predicción cero o cercana a ese valor cuando la realidad es un valor más elevado. En el caso de una predicción sobre la cantidad de metales pesados en el suelo debidos a las dosis de biosólidos añadidas supondría un riesgo por la posibilidad de acumulación o de sobrepasar las cantidades de metales pesados permitidas por la legislación.

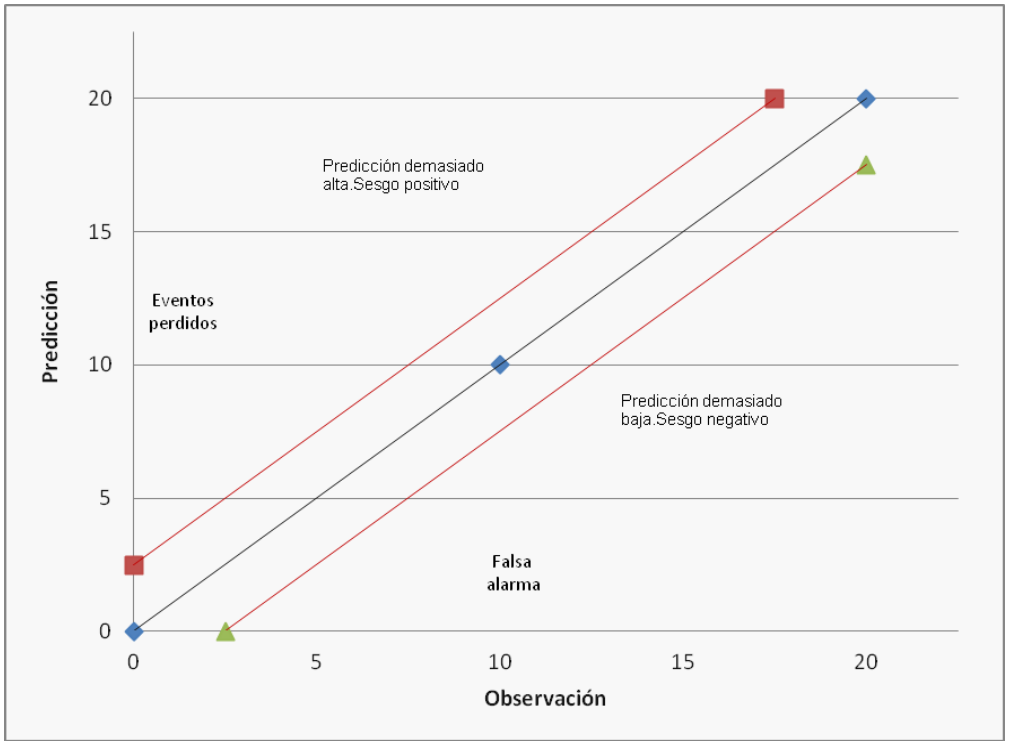


Figura 7-7 Ejemplo de funcionamiento de diagramas de dispersión.

Los métodos utilizados de predicción para las variables como materia orgánica muestran unos ajustes bajos, a pesar de la mejoría de los resultados obtenidos con *REPTree* originalmente. Y también se puede observar como los datos obtenidos con el método *REPTree* tiene una tendencia a pronosticar valores demasiado bajos frente a los reales. La dispersión que se produce de los pronósticos en el nitrógeno en suelo produce unos ajustes de los coeficientes de correlación apartados de los valores óptimos (Línea de 45º). En el caso del

fósforo los métodos utilizados muestran una predicción bastante buena con los datos empleados.

El método *Vote* un gran número de datos perdidos donde la predicción está por encima de los valores reales que se han obtenido en las muestras de campo principalmente en los metales que son más móviles y pueden ser tomados por las plantas, es el caso cobre y zinc, aunque también se observa esta falta de ajuste entre predicción y valores reales con el hierro, manganeso y plomo.

Podemos ver unos buenos ajustes de un gran número de métodos con variables como materia orgánica, fósforo, cadmio y cromo. Por otro lado, se ven discrepancias entre valores reales y de predicción (con muchos puntos dispersos) en el nitrógeno y en el zinc. Aun siendo razonable este resultado por la capacidad de traslocación de estas variables en el suelo, esto nos hace pensar que se deberían buscar otros modelos o modificar los parámetros para un mejor ajuste como, por ejemplo, ampliar el número de árboles de predicción (de 40 a 100) en el método *Random Forest*, o incluir otros algoritmos distintos de árboles (*ej*, redes neuronales, regresión lineal, etc…) para el método *Stacking* dado que son de los que mejores puntuaciones han conseguido en el estudio de las variables.

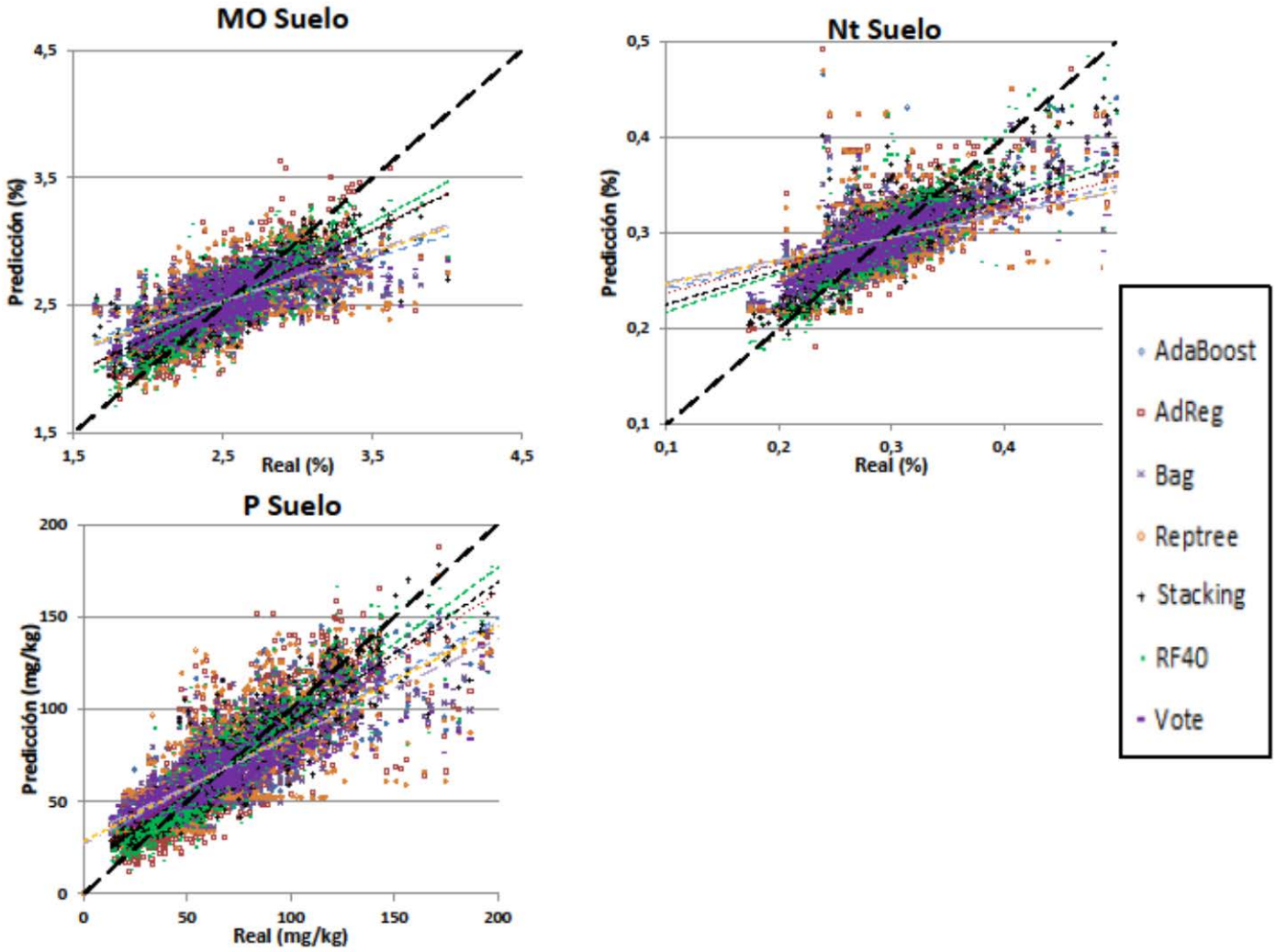


Figura 7-8 Diagramas de dispersión materia orgánica, nitrógeno y fósforo.

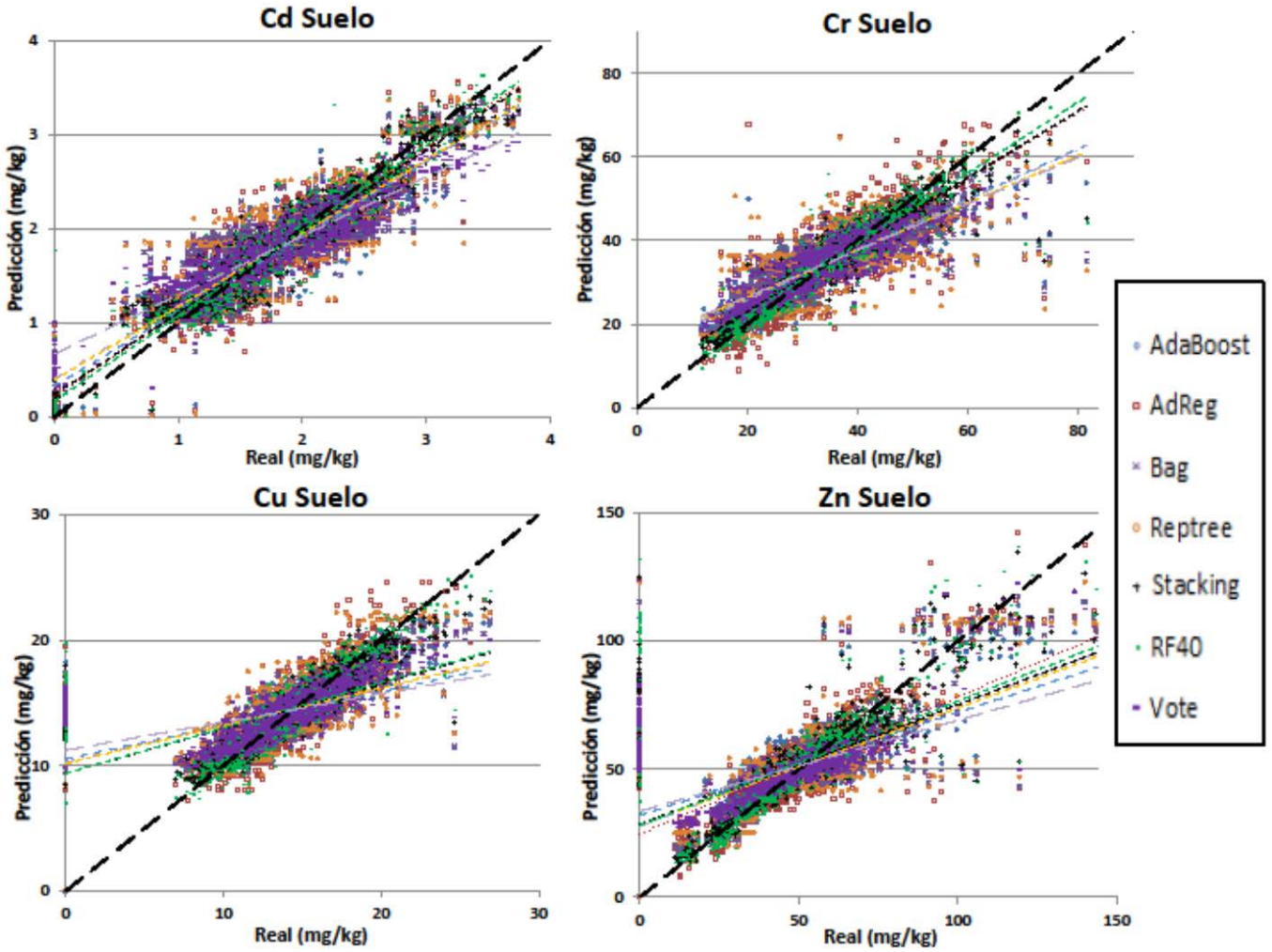


Figura 7-9 Diagramas de dispersión Cd, Cr, Cu y Zn.

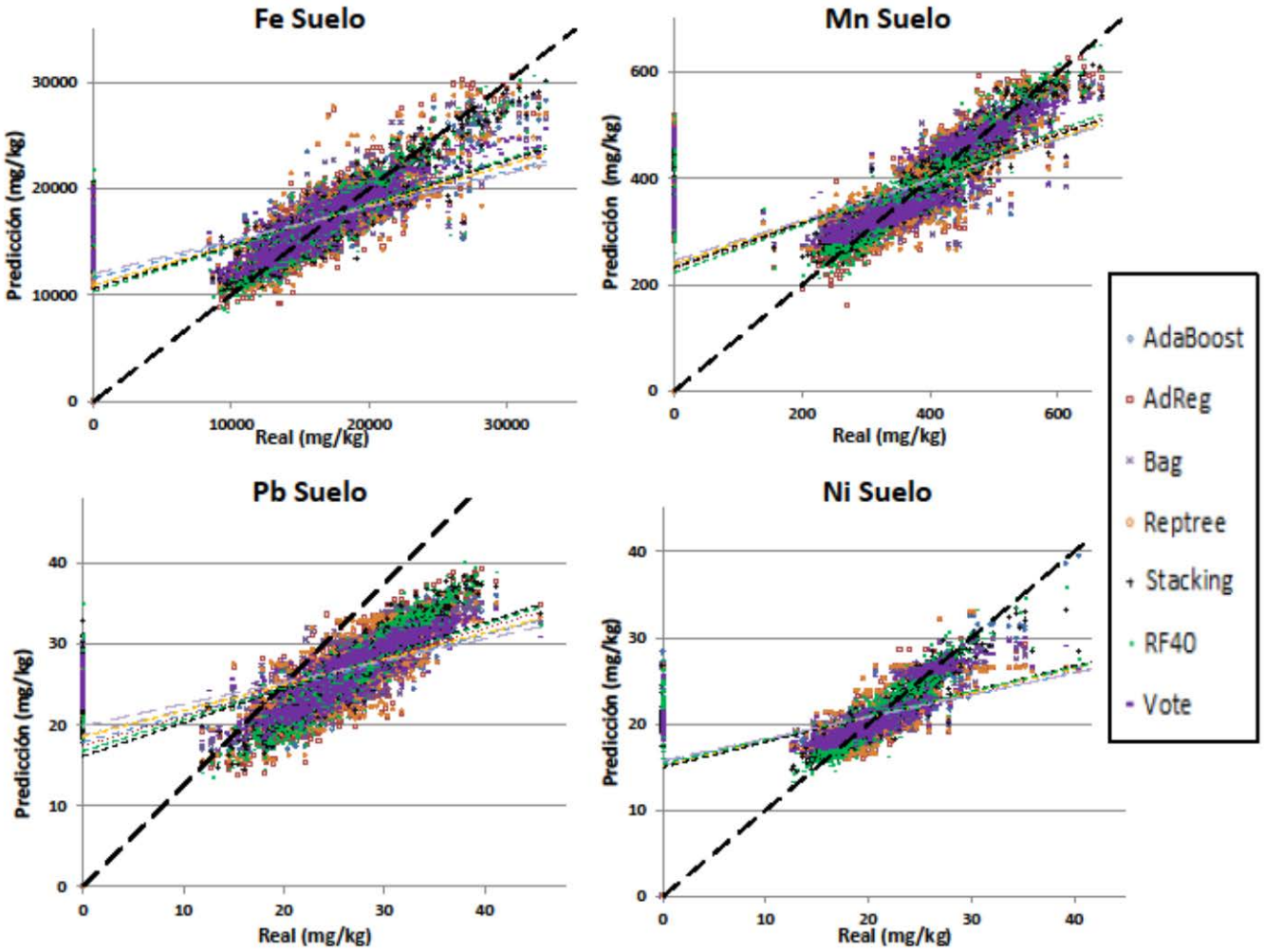


Figura 7-10 Diagramas de dispersión Fe, Mn, Pb y Ni.

7.6. Conclusiones

En conclusión, los métodos basados en combinación de regresores pueden utilizarse para realizar mejoras en los resultados de los árboles de regresión. La capacidad predictiva superior de *Random Forest* puede usarse para aumentar los coeficientes de correlación, en la mayor parte de los casos, así como reducir los errores obtenidos en los datos. Principalmente, estas técnicas pueden usarse para mejorar cualquier variable de respuesta recogida y para entender qué predictores se pueden impulsar con un mayor nivel de confianza que con otros métodos. Estas mejoras, con mejores coeficientes de correlación de las variables incorporadas a los modelos y unos menores errores, abren la posibilidad a unos datos de predicción más fiables.

La restricción de uso de algoritmos de árboles de decisión para los métodos con varios niveles permite una mejora significativa en los coeficientes de correlación aunque conlleva un incremento en el tiempo de computación. (*Stacking*)

Se puede decir que unos ajustes en los parámetros de partida de los árboles de regresión y la combinación adecuada de métodos *ensembles* permiten unos resultados obtenidos satisfactorios. El inconveniente que surge es la dificultad de interpretación de los árboles de regresión obtenidos, incluso en el caso de *Stacking* que establece una combinación lineal de los algoritmos de nivel 1.

8. Bibliografía

Abramowitz M, Stegun IA (1964) Handbook of Mathematical Functions: with formulas, graphs, and mathematical tables. vol 55. Courier Corporation.

Adams ML, Zhao FJ, McGrath SP, Nicholson FA, Chambers BJ (2004) Predicting Cadmium Concentrations in Wheat and Barley Grain Using Soil Properties. Journal of Environmental Quality 33:532-541.

Adriano DC, Paulsen GM, Murphy LS (1971) Phosphorus-Iron and Phosphorus-Zinc Relationships in Corn (Zea mays L.) Seedlings as Affected by Mineral Nutrition. Agronomy Journal 63:36-39.

Alegría A, Barberá R, Boluda R, Errecalde F, Farré R, Lagarda MJ (1991) Environmental cadmium, lead and nickel contamination: possible relationship between soil and vegetable content. Fresenius' Journal of Analytical Chemistry 339:654-657.

Alloway BJ (2009) Soil factors associated with zinc deficiency in crops and humans. Environmental Geochemistry and Health 31:537-548.

Alloway BJ, Jackson AP (1991) The behaviour of heavy metals in sewage sludge-amended soils. Science of the Total Environment 100:151-176.

Antić-Mladenović S, Rinklebe J, Frohne T, Stärk H-J, Wennrich R, Tomić Z, Ličina V (2011) Impact of controlled redox conditions on nickel in a serpentine soil. Journal of Soils and Sediments 11:406-415.

Asensio V, Forján R, Vega FA, Covelo EF (2016) Nickel, Lead and Zinc sorption in a reclaimed settling pond soil. Pedosphere 26:39-48.

Ashworth DJ, Alloway BJ (2004) Soil mobility of sewage sludge-derived dissolved organic matter, copper, nickel and zinc. Environmental Pollution 127:137-144.

Ayari F, Hamdi H, Jedidi N, Gharbi N, Kossai R (2010) Heavy metal distribution in soil and plant in municipal solid waste compost amended plots. International Journal of Environmental Science & Technology 7:465-472.

Bahmanyar MA (2008) Cadmium, Nickel, Chromium, and Lead Levels in Soils and Vegetables under Long-Term Irrigation with Industrial Wastewater. Communications in Soil Science and Plant Analysis 39:2068-2079.

Banks MK, Schwab AP, Henderson C (2006) Leaching and reduction of chromium in soil as affected by soil organic content and plants. Chemosphere 62:255-264.

Barrett JE, Burke IC (2000) Potential nitrogen immobilization in grassland soils across a soil organic matter gradient. Soil Biology and Biochemistry 32:1707-1716.

Bartlett R, James B (1988) Mobility and bioavailability of chromium in soils. En: Chromium in the Natural and Human Environments . JO Nriagu, E Nieboer (Eds.), Wiley, New York.

Battaglia A, Calace N, Nardi E, Petronio BM, Pietroletti M (2007) Reduction of Pb and Zn bioavailable forms in metal polluted soils due to paper mill sludge addition: Effects on Pb and Zn transferability to barley. Bioresource Technology 98:2993-2999.

Beck DP, Wery J, Saxena MC, Ayadi A (1991) Dinitrogen fixation and Nitrogen balance in cool-season food legumes. Agronomy Journal 83: 334-341.

Bhattacharyya P, Chakrabarti K, Chakraborty A, Tripathy S, Kim K, Powell MA (2008) Cobalt and nickel uptake by rice and accumulation in soil amended with municipal solid waste compost. Ecotoxicology and Environmental Safety 69:506-512.

Bishop CM (1996) Neural networks for pattern recognition. Oxford University Press.

Bradl HB (2004) Adsorption of heavy metal ions on soils and soils constituents. Journal of Colloid and Interface Science 277:1-18.

Breiman L (1996) Stacked regressions. Machine learning 24:49-64.

Breiman L (2001) Random forests. Machine learning 45:5-32.

Breiman L (2002) Manual on setting up, using, and understanding random forests v3. 1. Statistics Department University of California Berkeley, CA, USA.

Breiman L, Friedman J, Stone CJ, Olshen RA (1984) Classification and Regression Trees. Taylor & Francis,Londres.

Cechin I, de Fátima Fumis T (2004) Effect of nitrogen supply on growth and photosynthesis of sunflower plants grown in the greenhouse. Plant Science 166:1379-1385.

Codling EE, Dao TH (2007) Short-term effect of Lime, Phosphorus, and Iron amendments on water-extractable Lead and Arsenic in Orchard soils. Communications in Soil Science and Plant Analysis 38:903-919.

Cortet J, Kocev D, Ducobu C, Dzeroski S, Debeljak M, Schwartz C (2011) Using data mining to predict soil quality after application of biosolids in agriculture. Journal of Environmental Quality 40:1972-1982.

Crecchio C, Curci M, Pizzigallo MDR, Ricciuti P, Ruggiero P (2004) Effects of municipal solid waste compost amendments on soil enzyme activities and bacterial genetic diversity. Soil Biology and Biochemistry 36:1595-1605.

Chalhoub M, Garnier P, Coquet Y, Mary B, Lafolie F, Houot S (2013) Increased nitrogen availability in soil after repeated compost applications: Use of the PASTIS model to separate short and long-term effects. Soil Biology and Biochemistry 65:144-157.

Chatterjee J, Chatterjee C (2000) Phytotoxicity of cobalt, chromium and copper in cauliflower. Environmental Pollution 109:69-74.

Chen GC, He ZL, Stoffella PJ, Yang XE, Yu S, Yang JY, Calvert DV (2006) Leaching potential of heavy metals (Cd, Ni, Pb, Cu and Zn) from acidic sandy soil amended with dolomite phosphate rock (DPR) fertilizers. Journal of Trace Elements in Medicine and Biology 20:127-133.

de Melo WJ, de Stéfani Aguiar P, Maurício Peruca de Melo G, Peruca de Melo V (2007) Nickel in a tropical soil treated with sewage sludge and cropped with maize in a long-term field study. Soil Biology and Biochemistry 39:1341-1347.

Degryse F, Verma VK, Smolders E (2008) Mobilization of Cu and Zn by root exudates of dicotyledonous plants in resin-buffered solutions and in soil. Plant and Soil 306:69-84.

Dietterich TG Ensemble methods in machine learning. In: International workshop on multiple classifier systems., 2000. Springer, pp 1-15.

Doelman P, Haanstra L (1984) Short-term and long-term effects of cadmium, chromium, copper, nickel, lead and zinc on soil microbial respiration in relation to abiotic soil factors. Plant and Soil 79:317-327.

Doyle J (1977) Effects of low levels of dietary cadmium in animals—A review. Journal of Environmental Quality 6:111-116.

Elith J, Leathwick JR, Hastie T (2008) A working guide to boosted regression trees. Journal of Animal Ecology 77:802-813.

Fadigas FS, Amaral Sobrinho NMBd, Anjos LHCd, Mazur N (2010) Background levels of some trace elements in weathered soils from the Brazilian Northern region Scientia Agricola 67:53-59.

Fargašová A (2012) Plants as models for chromium and nickel risk assessment. Ecotoxicology 21:1476-1483.

Fayyad UM (1996) Data Mining and Knowledge Discovery: Making Sense Out of Data IEEE Expert: Intelligent Systems and Their Applications 11:20-25.

Fendorf S, La Force MJ, Li G (2004) Temporal Changes in Soil Partitioning and Bioaccessibility of Arsenic, Chromium, and Lead. Journal of Environmental Quality 33: 2049-2055.

François M, Grant C, Lambert R, Sauvé S (2009) Prediction of cadmium and zinc concentration in wheat grain from soils affected by the application of phosphate fertilizers varying in Cd concentration. Nutrient Cycling in Agroecosystems 83:125-133.

Frank E, Mayo M, Kramer S (2015) Alternating model trees. Paper presented at the Proceedings of the 30th Annual ACM Symposium on Applied Computing., Salamanca, Spain.

Freund Y, Mason L (1999) The Alternating Decision Tree Learning Algorithm. Paper presented at the Proceedings of the Sixteenth International Conference on Machine Learning.

Freund Y, Schapire RE Experiments with a new boosting algorithm. In: Icml., 1996. pp 148-156.

Freund Y, Schapire RE (1997) A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. Journal of Computer and System Sciences 55:119-139.

Friedman J, Hastie T, Tibshirani R (2001) The Elements of Statistical Learning Vol 1. Springer Series in Statistics, Springer, Berlin.

Gabrielle B, Da-Silveira J, Houot S, Michelin J (2005) Field-scale modelling of carbon and nitrogen dynamics in soils amended with urban waste composts. Agriculture, Ecosystems & Environment 110:289-299.

Gallardo-Lara F, Azcón M, Polo A (2006) Phytoavailability and Fractions of Iron and Manganese in Calcareous Soil Amended with Composted Urban Wastes. Journal of Environmental Science and Health, Part B 41:1187-1201.

Gao X, Mohr RM, McLaren DL, Grant CA (2011) Grain cadmium and zinc concentrations in wheat as affected by genotypic variation and potassium chloride fertilization. Field Crops Research 122:95-103.

García-Gil JC, Plaza C, Soler-Rovira P, Polo A (2000) Long-term effects of municipal solid waste compost application on soil enzyme activities and microbial biomass. Soil Biology and Biochemistry 32:1907-1913.

Ghasemi-Fasaei R, Ronaghi A (2008) Interaction of Iron with Copper, Zinc, and Manganese in Wheat as Affected by Iron and Manganese in a Calcareous Soil. Journal of Plant Nutrition 31:839-848.

Gigliotti G, Businelli D, Giusquiani PL (1996) Utilisation of Waste Organic MatterTrace metals uptake and distribution in corn plants grown on a 6-year urban waste compost amended soil. Agriculture, Ecosystems & Environment 58:199-206.

Golovatyj SE, Bogatyreva A, Minsk (Belarus)) (1999) [Effect of levels of chromium content in a soil on its distribution in organs of corn plants]

Graham RD et al. (2007) Nutritious Subsistence Food Systems. In: Donald LS (ed) Advances in Agronomy, vol Volume 92. Academic Press, pp 1-74.

Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH (2009) The WEKA data mining software: an update SIGKDD Explor Newsl 11:10-18.

Hao X-Z, Zhou D-M, Li D-D, Jiang P (2012) Growth, Cadmium and Zinc Accumulation of Ornamental Sunflower (Helianthus annuus L.) in Contaminated Soil with Different Amendments. Pedosphere 22:631-639.

Hargreaves JC, Adl MS, Warman PR (2008) A review of the use of composted municipal solid waste in agriculture. Agriculture, Ecosystems & Environment 123:1-14.

Hassett JJ, Miller JE, Koeppe DE (1976) Interaction of lead and cadmium on maize root growth and uptake of lead and cadmium by roots. Environmental Pollution (1970) 11:297-302.

He QB, Singh BR (1993) Effect of organic matter on the distribution, extractability and uptake of cadmium in soils. European Journal of Soil Science 44:641-650.

Hernández LE, Lozano-Rodríguez E, Gárate An, Ramón C-R (1998) Influence of cadmium on the uptake, tissue accumulation and subcellular distribution of manganese in pea seedlings. Plant Science 132:139-151.

Hernando S, Lobo MC, Polo A (1989) Advances in Humic Substances Research Effect of the application of a municipal refuse compost on the physical and chemical properties of a soil. Science of The Total Environment 81:589-596.

Hothorn T, Hornik K, Zeileis A (2006) Unbiased Recursive Partitioning: A Conditional Inference Framework. Journal of Computational and Graphical Statistics 15:651-674.

Hunt EB (1963) Concept learning: An information processing problem. vol 8. Behavioral Science, vol 4. John Wiley & Sons, Ltd., New York.

Iba W, Langley P (1992) Induction of one-level decision trees. Paper presented at the Proceedings of the ninth international workshop on Machine learning, Aberdeen, Scotland, United Kingdom.

Impellitteri CA, Saxe JK, Cochran M, Janssen GM, Allen HE (2003) Predicting the bioavailability of copper and zinc in soils: modeling the partitioning of potentially bioavailable copper and zinc from soil solid to soil solution. Environmental Toxicoology and Chemistry 22:1380-1386.

Inal A, Gunes A, Zhang F, Cakmak I (2007) Peanut/maize intercropping induced changes in rhizosphere and nutrient concentrations in shoots. Plant Physiology and Biochemistry 45:350-356.

Jensen ES (1996) Grain yield, symbiotic N_2 fixation and interspecific competition for inorganic N in pea-barley intercrops. Plant and Soil 182:25-38.

Jolley VD, Hansen NC, Shiffler AK (2004) Nutritional and management related interactions with iron-deficiency stress response mechanisms. Soil Science and Plant Nutrition 50:973-981 Kabata-Pendias A (2010) Trace elements in soils and plants. CRC press, Boca Raton FL.

Kashem MA, Singh BR, Kawai S (2007) Mobility and distribution of cadmium, nickel and zinc in contaminated soil profiles from Bangladesh. Nutrient Cycling in Agroecosystems 77:187-198.

Kidd PS, Domínguez-Rodríguez MJ, Díez J, Monterroso C (2007) Bioavailability and plant accumulation of heavy metals and phosphorus in agricultural soils amended by long-term application of sewage sludge. Chemosphere 66:1458-1467.

Kittler J, Hatef M, Duin RP, Matas J (1998) On combining classifiers. IEEE transactions on pattern analysis and machine intelligence 20:226-239.

Kohavi R, John GH (1997) Wrappers for feature subset selection. Artificial intelligence 97:273-324.

Korboulewsky N, Dupouyet S, Bonin G (2002) Environmental risks of applying sewage sludge compost to vineyards: carbon, heavy metals, nitrogen, and phosphorus accumulation. Journal of Environmental Quality 31:1522-1527.

Kuncheva LI (2004) Combining Pattern Classifiers: Methods and Algorithms. Wiley-Interscience,New York.

Lastra O, Chueca A, Lachica M, López Gorgé J (1988) Root Uptake and Partition of Copper, Iron, Manganese, and Zinc in Pinus radiata Seedlings Grown under Different Copper Supplies. Journal of Plant Physiology 132:16-22.

Liu D, Kottke I (2003) Subcellular localization of chromium and nickel in root cells of *Allium cepa* by EELS and ESI. Cell Biology and Toxicology 19:299-311.

Lübben S, Sauerbeck D (1991) The uptake and distribution of heavy metals by spring wheat. Water, Air, and Soil Pollution 57:239-247.

Luo Y, Rimmer DL (1995) Zinc-copper interaction affecting plant growth on a metal-contaminated soil. Environmental Pollution 88:79-83.

Mamindy-Pajany Y, Sayen S, Guillon E (2013) Impact of sewage sludge spreading on nickel mobility in a calcareous soil: adsorption–desorption through column experiments. Environmental Science and Pollution Research 20:4414-4423.

Mantovi P, Baldoni G, Toderi G (2005) Reuse of liquid, dewatered, and composted sewage sludge on agricultural land: effects of long-term application on soil and crop Water Research 39:289-296.

Manzoni S, Porporato A (2009) Soil carbon and nitrogen mineralization: Theory and models across scales. Soil Biology and Biochemistry 41:1355-1379.

MAPA (1994) Métodos oficiales de análisis, vol. 3: Plantas, suelos, aguas para riego, fertilizantes orgánicos e inorgánicos.

Martínez CE, Motto HL (2000) Solubility of lead, zinc and copper added to mineral soils. Environmental Pollution 107:153-158.

Mishra M, Sahu RK, Sahu SK, Padhy RN (2009) Growth, yield and elements content of wheat (*Triticum aestivum*) grown in composted municipal solid wastes amended soil. Environment, Development and Sustainability 11:115-126.

Mkhabela MS, Warman PR (2005) The influence of municipal solid waste compost on yield, soil phosphorus availability and uptake by two vegetable crops grown in a Pugwash sandy loam soil in Nova Scotia. Agriculture, Ecosystems & Environment 106:57-67.

Mohri M, Rostamizadeh A, Talwalkar A (2012) Foundations of machine learning. MIT press.

Morel JL, Mench M, Guckert A (1986) Measurement of Pb^{2+}, Cu^{2+} and Cd^{2+} binding with mucilage exudates from maize (*Zea mays* L.) roots. Biology and Fertility of Soils 2:29-34.

Morera MT, Echeverría J, Garrido J (2002) Bioavailability of heavy metals in soils amended with sewage sludge Canadian Journal of Soil Science 82:433-438.

Mulla DJ, Page AL, Ganje TJ (1980) Cadmium Accumulations and Bioavailability in Soils From Long-Term Phosphorus Fertilization. Journal of Environmental Quality 9: 408-412.

Nan Z, Li J, Zhang J, Cheng G (2002) Cadmium and zinc interactions and their transfer in soil-crop system under actual field conditions. Science of The Total Environment 285:187-195.

Odlare M, Pell M, Svensson K (2008) Changes in soil chemical and microbiological properties during 4 years of application of various organic residues Waste Management 28:1246-1253.

Opitz D, Maclin R (1999) Popular ensemble methods: An empirical study. Journal of Artificial Intelligence Research 11:169-198.

Peña S (2013) Capacidad fertilizante y riesgo metálico asociados a la utilización de residuos orgánicos en agricultura bajo diferentes condiciones de cultivo en la provincia de Burgos. Tesis Doctoral Universidad de Burgos, España.

Peoples MB et al. (2009) The contributions of nitrogen-fixing crop legumes to the productivity of agricultural systems. Symbiosis 48:1-17.

Peoples MB, Craswell ET (1992) Biological nitrogen fixation: Investments, expectations and actual contributions to agriculture Plant and Soil 141:13-39.

Perucci P (1990) Effect of the addition of municipal solid-waste compost on microbial biomass and enzyme activities in soil. Biology and Fertility of Soils 10:221-226.

Picard RR, Cook RD (1984) Cross-Validation of Regression Models. Journal of the American Statistical Association 79:575-583.

Polikar R (2006) Ensemble based systems in decision making. IEEE Circuits and systems magazine 6:21-45.

Prasad AM, Iverson LR, Liaw A (2006) Newer Classification and Regression Tree techniques: Bagging and Random Forests for ecological prediction. Ecosystems 9:181-199.

Quinlan JR Learning with continuous classes. In: 5th Australian joint conference on artificial intelligence, 1992. Singapore, pp 343-348.

Quinlan JR (1993) C4. 5: Programs for Machine Learning.

Ravuri V, Hume DJ (1993) Soybean Stover Nitrogen Affected by Dinitrogen Fixation and Cultivar Agronomy Journal 85: 328-333.

Reynolds MR (1984) Estimating the error in model predictions. Forest Science 30:454-469.

Rokach L, Maimon O (2010) Ensemble-based classifiers. Artificial Intelligence Review 33:1-39.

Rokach L, Maimon O (2014) Data mining with decision trees: theory and applications. World scientific.

Ros GH, Hoffland E, van Kessel C, Temminghoff EJM (2009) Extractable and dissolved soil organic nitrogen – A quantitative assessment. Soil Biology and Biochemistry 41:1029-1039.

Sauvé S, McBride M, Hendershot W (1998) Soil solution speciation of Lead(II): Effects of Organic Matter and pH. Soil Science Society of America Journal 62:618-621.

Schapire RE (2003) The boosting approach to machine learning: An overview. In: Nonlinear estimation and classification. Springer, pp 149-171.

Shahandeh H, Hossner LR (2000) Plant Screening for Chromium Phytoremediation International Journal of Phytoremediation 2:31-51.

Shaheen SM (2009) Sorption and lability of cadmium and lead in different soils from Egypt and Greece. Geoderma 153:61-68.

Shaheen SM, Rinklebe J (2014) Geochemical fractions of chromium, copper, and zinc and their vertical distribution in floodplain soil profiles along the Central Elbe River, Germany Geoderma 228–229:142-159.

Sharma DC, Chatterjee C, Sharma CP (1995) Chromium accumulation and its effects on wheat (*Triticum aestivum* L. cv. HD 2204) metabolism Plant Science 111:145-151.

Shewry PR, Peterson PJ (1976) Distribution of Chromium and Nickel in Plants and Soil from Serpentine and Other Sites. Journal of Ecology 64:195-212.

Sigua GC, Adjei MB, Rechcigl JE (2005) Cumulative and residual effects of repeated sewage sludge applications: forage productivity and soil quality implications in South Florida, USA. Environmental Science and Pollution Research International 12:80-88.

Silveira MLA, Alleoni LRF, Camargo OA, Casagrande JC (2002) Copper adsorption in oxidic soils after removal of organic matter and iron oxides. Communications in Soil Science and Plant Analysis 33:3581-3592.

Singh BR, Steenberg K (1975) Interactions of Micronutrients in Barley Grown on Zinc-polluted Soils1. Soil Science Society of America Journal 39:674-679.

Singh RP, Agrawal M (2008) Potential benefits and risks of land application of sewage sludge Waste management (New York, NY) 28:347-358.

Singh RP, Singh P, Ibrahim MH, Hashim R (2011) Land application of sewage sludge: Physicochemical and microbial response. In: Whitacre DM (ed) Reviews of Environmental Contamination and Toxicology. Springer New York, New York, NY, pp 41-61.

Smith SR (1992) Sewage sludge and refuse composts as peat alternatives for conditioning impoverished soils: Effects on the growth response and mineral status of *Petunia grandiflora*. Journal of Horticultural Science 67:703-716.

Smith SR (1994) Effect of soil pH on availability to crops of metals in sewage sludge-treated soils. I. Nickel, copper and zinc uptake and toxicity to ryegrass. Environmental Pollution 85:321-327.

Smith SR (2009) A critical review of the bioavailability and impacts of heavy metals in municipal solid waste composts compared to sewage sludge. Environmental International 35:142-156.

Stewart MA et al. (2003) Effects of Contaminant Concentration, Aging, and Soil Properties on the Bioaccessibility of Cr(III) and Cr(VI) in Soil. Soil and Sediment Contamination: An International Journal 12:1-21.

Stuczynski T, McCarty G, Siebielec G (2003) Response of soil microbiological activities to cadmium, lead, and zinc salt amendments. Journal of Environmental Quality 32:1346-1355.

Tang T, Miller DM (1991) Growth and tissue composition of rice grown in soil treated with inorganic copper, nickel, and arsenic. Communications in Soil Science and Plant Analysis 22:2037-2045.

Touw WG, Bayjanov JR, Overmars L, Backus L, Boekhorst J, Wels M, van Hijum SAFT (2013) Data mining in the Life Sciences with Random Forest: a walk in the park or lost in the jungle? Briefings in Bioinformatics 14:315-326.

UdDin I, Bano A, Masood S (2015) Chromium toxicity tolerance of Solanum nigrum L. and Parthenium hysterophorus L. plants with reference to ion pattern, antioxidation activity and root exudation Ecotoxicology and Environmental Safety 113:271-278.

Verma TS, Minhas RS (1987) Zinc and phosphorus interaction in a wheat-maize cropping system. Fertilizer Research 13:77-86.

Walker RL, Burns IG, Moorby J (2001) Responses of plant growth rate to nitrogen supply: a comparison of relative addition and N interruption treatments. Journal of Experimental Botany 52:309-317.

Wang G, Su M-Y, Chen Y-H, Lin F-F, Luo D, Gao S-F (2006) Transfer characteristics of cadmium and lead from soil to the edible parts of six vegetable species in southeastern China. Environmental Pollution 144:127-135.

Warman PR, Muizelaar T, Termeer WC (1995) Bioavailability of As, Cd, Co, Cr, Cu, Hg, Mo, Ni, Pb, Se, and Zn from Biosolids Amended Compost. Compost Science & Utilization 3:40-50.

Warman PR, Termeer WC (2005) Evaluation of sewage sludge, septic waste and sludge compost applications to corn and forage: yields and N, P and K content of crops and soils. Bioresource Technology 96:955-961.

Weiss SM, Indurkhya N (1998) Predictive Data Mining: A Practical Guide. Morgan Kaufmann Publishers Inc.

Westerman PW, Bicudo JR (2005) Management considerations for organic waste use in agriculture. Bioresource Technology 96:215-221.

Willis LG (1936) The function of copper in soils and its relation to the availability of iron and manganese Journal of Agricultural Research 52:467-476.

Witten I, Wang Y Induction of model trees for predicting continuous classes. In: Proc. Poster Papers Europ. Conf. Machine Learning, 1997.

Witten IH, Frank E (2005) Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann, Massachusetts.

Witten IH, Frank E, Hall MA (2011) Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann Publishers Inc., Massachusetts.

Witten IH, Ting KM (1999) Issues in Stacked Generalization. Journal of Artificial Intelligence Research 10 271-289.

Wolpert DH (1992) Stacked generalization. Neural Networks 5:241-259.

Wu C, Luo Y, Zhang L (2010) Variability of copper availability in paddy fields in relation to selected soil properties in southeast China. Geoderma 156:200-206.

Wu F, Zhang G, Yu J (2003b) Interaction of Cadmium and four microelements for uptake and translocation in different barley genotypes. Communications in Soil Science and Plant Analysis 34:2003-2020.

Xue Y-F, Zhang W, Liu D-Y, Yue S-C, Cui Z-L, Chen X-P, Zou C-Q (2014) Effects of nitrogen management on root morphology and zinc translocation from root to shoot of winter wheat in the field. Field Crops Research 161:38-45.

Yin Y, Impellitteri CA, You S-J, Allen HE (2002) The importance of organic matter distribution and extract soil:solution ratio on the desorption of heavy metals from soils. Science of The Total Environment 287:107-119.

Zahedifar M, Karimian N, Yasrebi J (2012) Influence of applied zinc and organic matter on zinc desorption kinetics in calcareous soils. Archives of Agronomy and Soil Science 58:169-178.

Zhang J, Hua P, Krebs P (2016) The influences of dissolved organic matter and surfactant on the desorption of Cu and Zn from road-deposited sediment. Chemosphere 150:63-70.

Zhou LX, Wong JWC (2001) Effect of Dissolved Organic Matter from Sludge and Sludge Compost on Soil Copper Sorption. Journal of Environmental Quality 30: 878-883.

Zia-ur-Rehman M, Sabir M, Rizwan M, Saifullah, Ahmed HR, Nadeem M (2015) Chapter 13 - Remediating Cadmium-Contaminated Soils by Growing Grain Crops Using Inorganic Amendments. In: Soil Remediation and Plants. Academic Press, San Diego, pp 367-396.

Zornoza P, Sánchez-Pardo B, Carpena RO (2010) Interaction and accumulation of manganese and cadmium in the manganese accumulator *Lupinus albus*. Journal of Plant Physiology 167:1027-1032.

Zurayk R, Sukkariyah B, Baalbaki R (2000) Common hydrophytes as bioindicators of nickel, chromium and cadmium pollution. Water, Air, and Soil Pollution 127:373-388.

9. Anexo I

MATERIAL COMPLEMENTARIO

Valdespinar

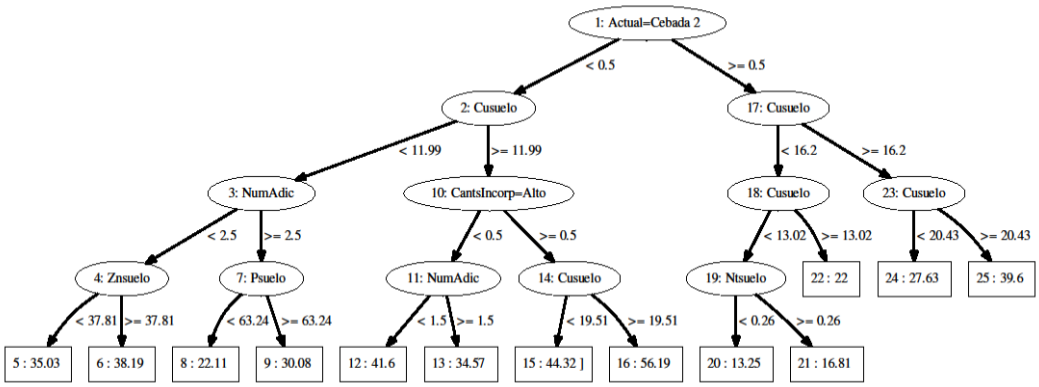


Figura 9-1 Árbol de regresión obtenido para el cromo del suelo.

Parámetros del árbol. Profundidad máxima=4, peso mínimo=8, datos para la poda=9.

En este árbol podemos ver las relaciones entre las cantidades de cromo con el cobre en el suelo, el nitrógeno, los cultivos, el número de adiciones y las cantidades incorporadas. Todas las variables tienen una influencia positiva con la cantidad de cromo en el suelo excepto el número de adiciones, ya que cuando estas son 3 la cantidad del metal es menor en el suelo.
El cultivo de la cebada permite una primera separación en la cual si existe este cultivo las cantidades de cromo en el suelo son menores. El valor más bajo se obtiene cuando el cultivo es la cebada por segunda vez y las cantidades de cobre en el suelo son inferiores a 13,02 mg Kg^{-1} y el nitrógeno no es superior a 0,26%.
En el otro lado del árbol vemos como cuando el cultivo actual no es la cebada entonces las adiciones altas de enmiendas permiten los valores más altos de cromo en el suelo.

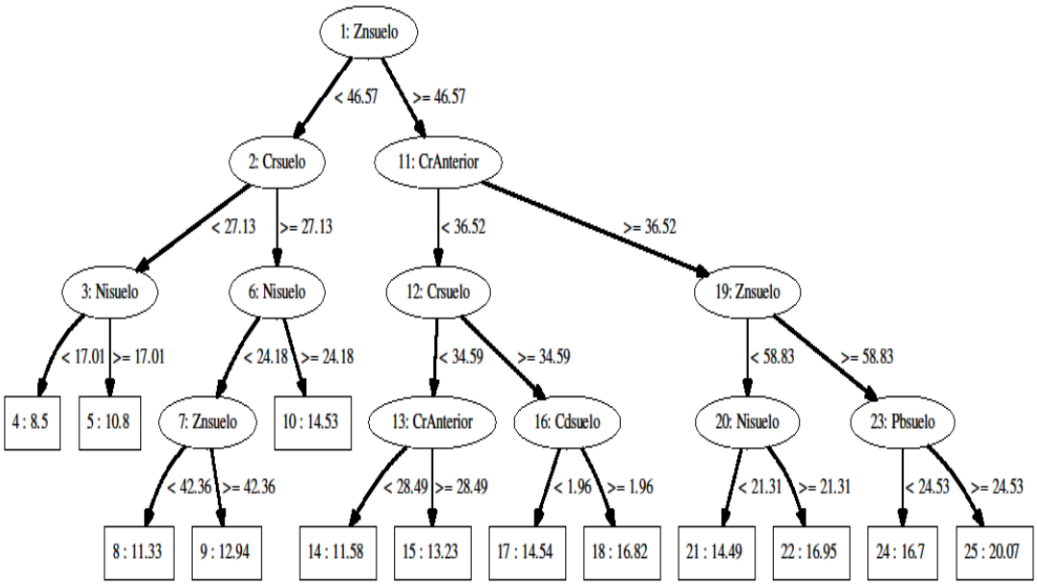


Figura 9-2 Árbol de regresión obtenido para el cobre del suelo.

Parámetros del árbol. Profundidad máxima=4, peso mínimo=8, datos para la poda=9.

El árbol de regresión obtenido para el contenido de cobre en el suelo (*r* = 0,846), muestra como principales variables que participan en este modelo a las concentraciones de Zn, Cr y Ni del suelo. También la presencia de cadmio, de plomo y de cantidades de cromo anterior en el suelo discriminan en las contribuciones de cobre en el árbol.

Los valores más altos se dan con cantidades de zinc en el suelo de 46,57 mg Kg^{-1}, de cromo anterior por encima de 36,52, de zinc por encima de 58,83 y de plomo en el suelo mayores a 24,53 mg Kg^{-1}. Mientras que los menores valores vienen determinados por valores inferiores de 27,13 de cromo y de níquel por debajo de 17,01 mg Kg^{-1}.

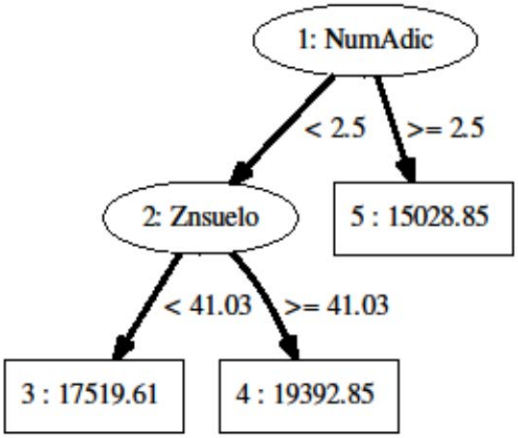


Figura 9-3 Árbol de regresión obtenido para el hierro del suelo.

Parámetros del árbol. Profundidad máxima=4, peso mínimo=8, datos para la poda=5.

Los valores en las correlaciones del hierro, así como el manganeso, en los árboles de regresión son menores que los demás metales que han sido estudiados. Podemos ver como dos variables nos definen la interpretación de este árbol. Se muestran relaciones con el número de adiciones y el zinc presente en el suelo. Las cantidades más bajas de hierro en el suelo se obtienen cuando el número de adiciones ha sido 3. Cuando tenemos una o dos adiciones de enmiendas podemos observar que existe una relación con la cantidad de zinc. El zinc tiene una correlación positiva con el hierro presente en el suelo.

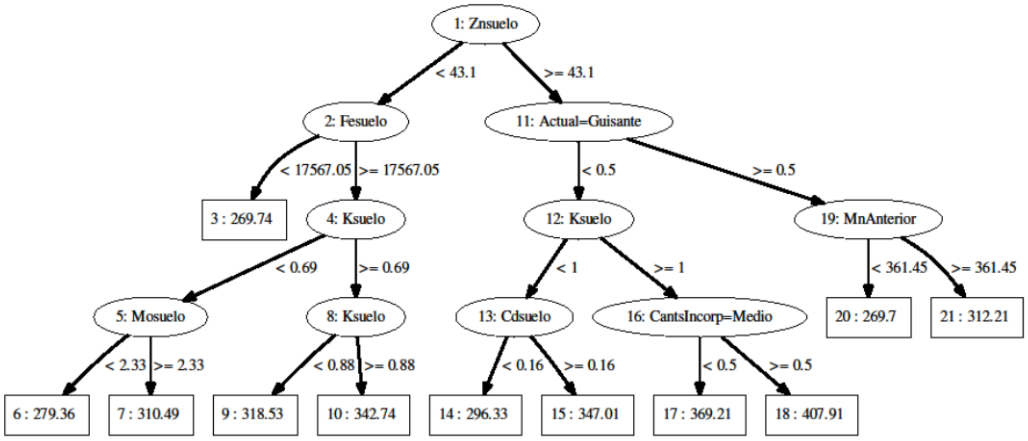


Figura 9-4 Árbol de regresión obtenido para el manganeso del suelo.

Parámetros del árbol. Profundidad máxima=4, peso mínimo=8, datos para la poda=7.

El árbol de regresión obtenido para el contenido de manganeso en el suelo muestra como principales variables que participan en este modelo a las concentraciones de Zn, Fe , la cantidad de K y Cd, de materia orgánica del suelo ,la aplicación de dosis medias de enmienda y tener como cultivo actual guisante. Precisamente este último parámetro es el único que tiene un efecto antagonista sobre las cantidades de manganeso en el suelo. Vemos como la incorporación de cantidades de biosólido en dosis medias permiten los mayores valores de manganeso medidos en el suelo.

Destaca la relación entre el hierro y el manganeso que aparece reflejada en la rama izquierda del árbol donde los menores valores de hierro conllevan las cantidades más bajas de manganeso presente en el suelo. La presencia del zinc en las principales ramas en los árboles representados de hierro y manganeso puede ser una confirmación de la destacada relación de los óxidos de hierro y manganeso en la estructura del suelo.

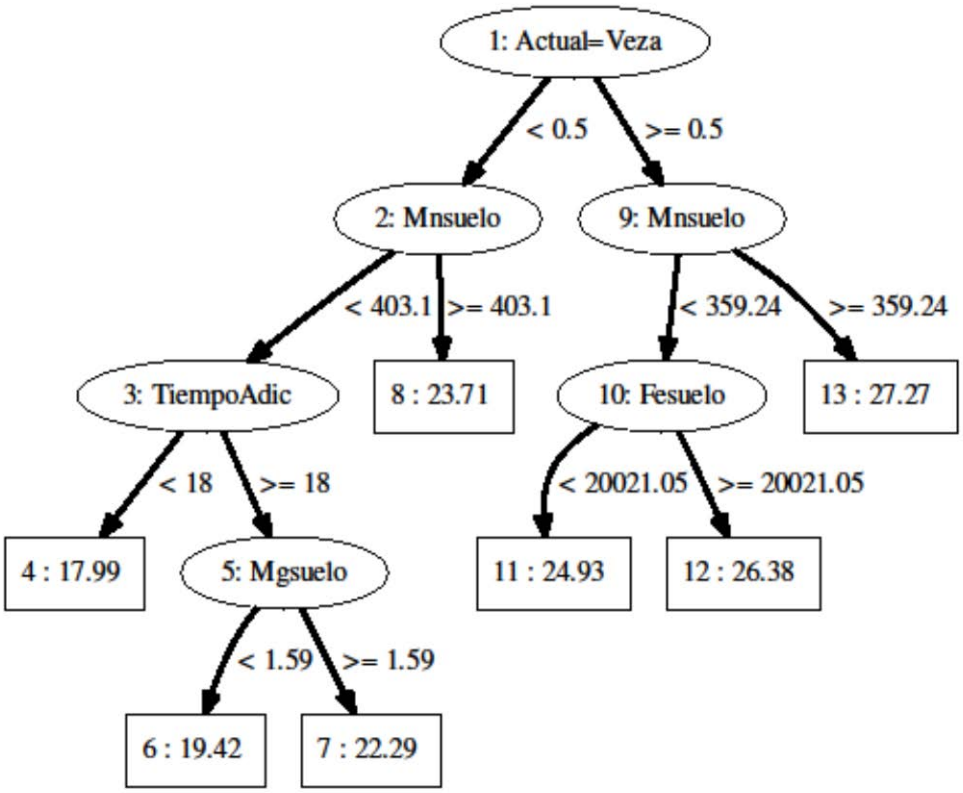


Figura 9-5 Árbol de regresión obtenido para el níquel del suelo.

Parámetros del árbol. Profundidad máxima=4, peso mínimo=8, datos para la poda=5.

En las medidas de níquel en el suelo las variables más influyentes vemos que son el cultivo actual de veza, el manganeso, hierro y magnesio del suelo y el tiempo transcurrido desde la adición. Se observa una fuerte relación entre el níquel en el suelo y la cantidad de manganeso.
Los valores más bajos se tienen cuando no se ha cultivado veza en ese año, la cantidad de manganeso es menor de 403,1 mg Kg^{-1} y han pasado 12 meses de

la adición al suelo de residuos. Cuando se ha cultivado veza y la cantidad de manganeso en el suelo es de 359,24 mg Kg^{-1} el níquel tiene un valor de 27,27 mg Kg^{-1}.

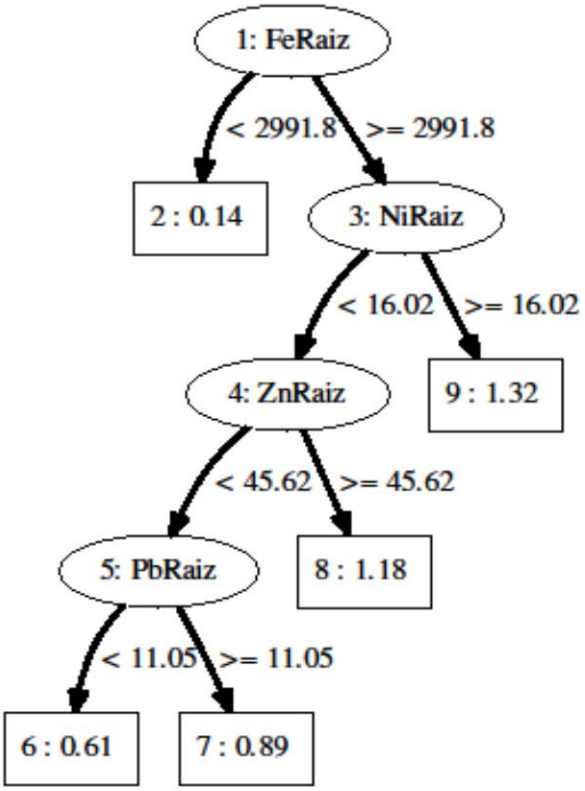


Figura 9-6 Árbol de regresión obtenido para el cadmio de la raíz.

Parámetros del árbol. Profundidad máxima=4, peso mínimo=8, datos para la poda=5.

La cantidad de cadmio en la raíz se ve influida por algunos metales presentes en la raíz. Estos metales son el hierro, níquel, zinc y plomo. El níquel con valores superiores a 16,02 mg Kg^{-1} dan lugar a las mayores cantidades de cadmio en raíz.

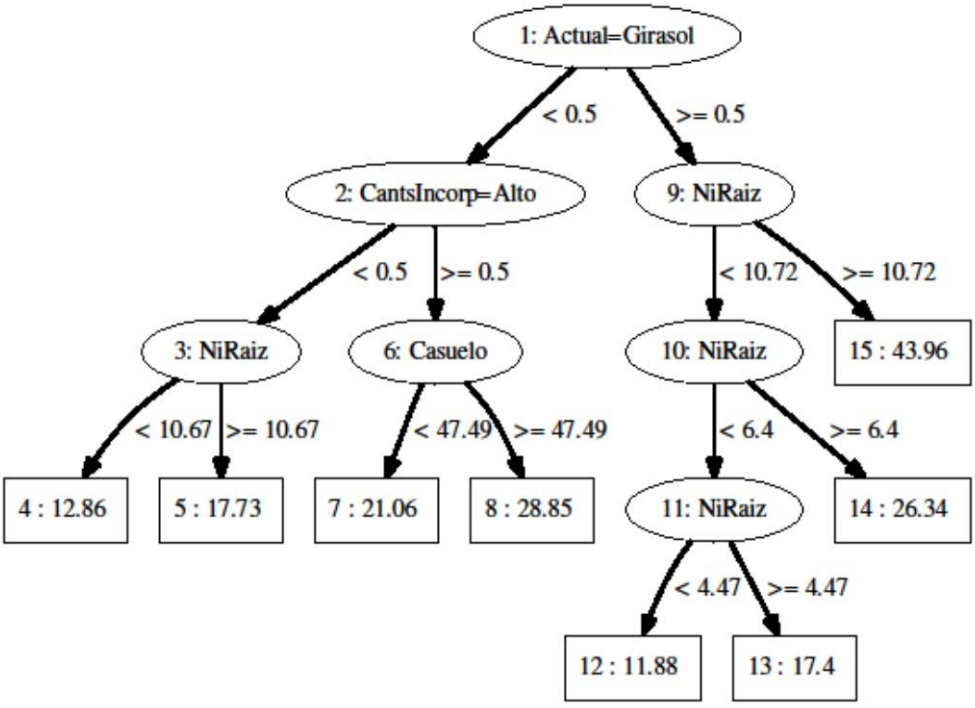


Figura 9-7 Árbol de regresión obtenido para el cromo de la raíz.

Parámetros del árbol. Profundidad máxima=4, peso mínimo=8, datos para la poda=9.

Las cantidades de cromo medidas en la raíz dependen de si el cultivo en el año ha sido el girasol, si la cantidad de residuo incorporada ha sido alta, la cantidad de níquel encontrado en la raíz y en menor medida del calcio presente en el suelo.

Los mayores valores de cromo en la raíz se obtienen cuando se tiene como cultivo actual el girasol y la cantidad de níquel en la raíz tiene valores superiores a 10,72 mg Kg^{-1}. Podemos ver cómo, si no se ha cultivado girasol en ese año la incorporación de cantidades altas de enmiendas al suelo determina los mayores valores de cromo en la raíz con una influencia positiva del calcio presente en el suelo.

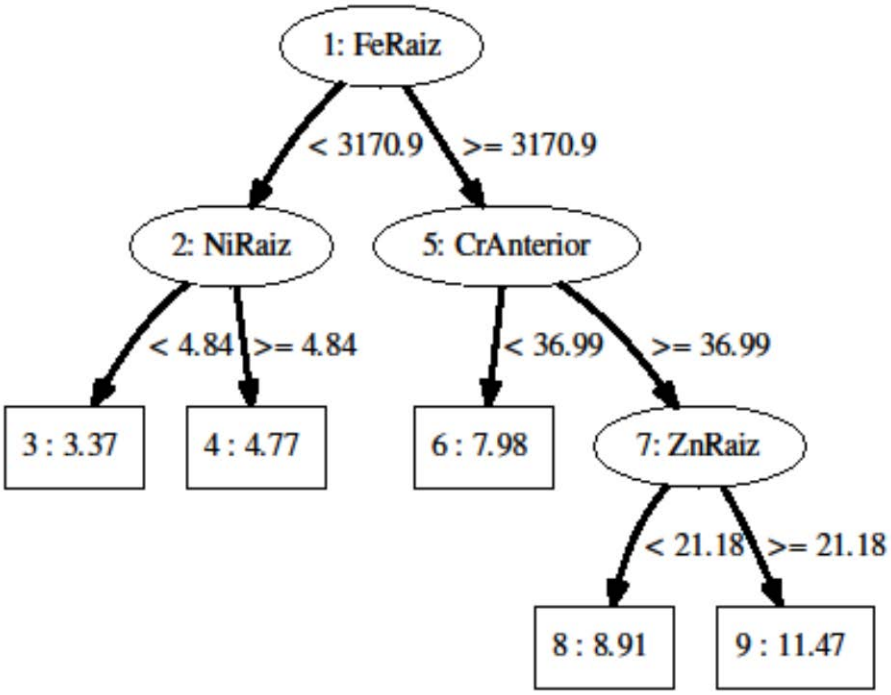


Figura 9-8 Árbol de regresión obtenido para el cobre de la raíz.

Parámetros del árbol. Profundidad máxima=4, peso mínimo=8, datos para la poda=7.

El árbol de regresión obtenido para el contenido de Cu en la raíz presenta como principales variables explicativas a las cantidades de otros metales como Fe, Ni y Zn en la raíz, así como el contenido de Cr acumulado en el suelo debido a la aplicación de biosólidos en años anteriores.

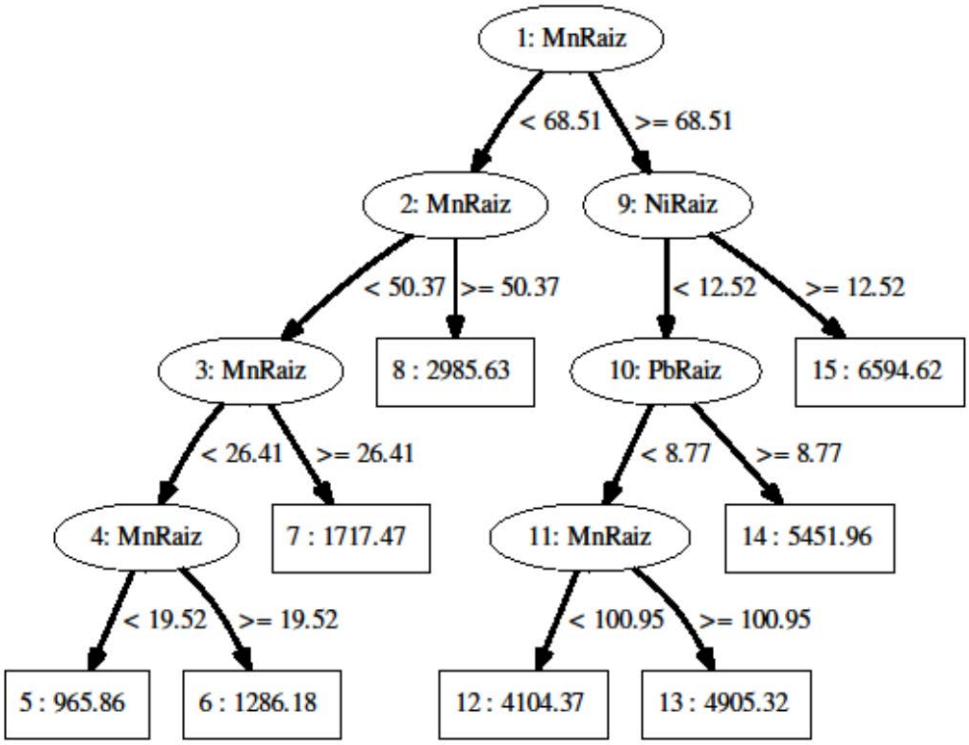


Figura 9-9 Árbol de regresión obtenido para el hierro de la raíz.

Parámetros del árbol. Profundidad máxima=4, peso mínimo=8, datos para la poda=3.

El hierro en la raíz tiene como variables más relevantes, en la formación del gráfico del árbol, al manganeso, el níquel y el plomo que se encuentran en la raíz.

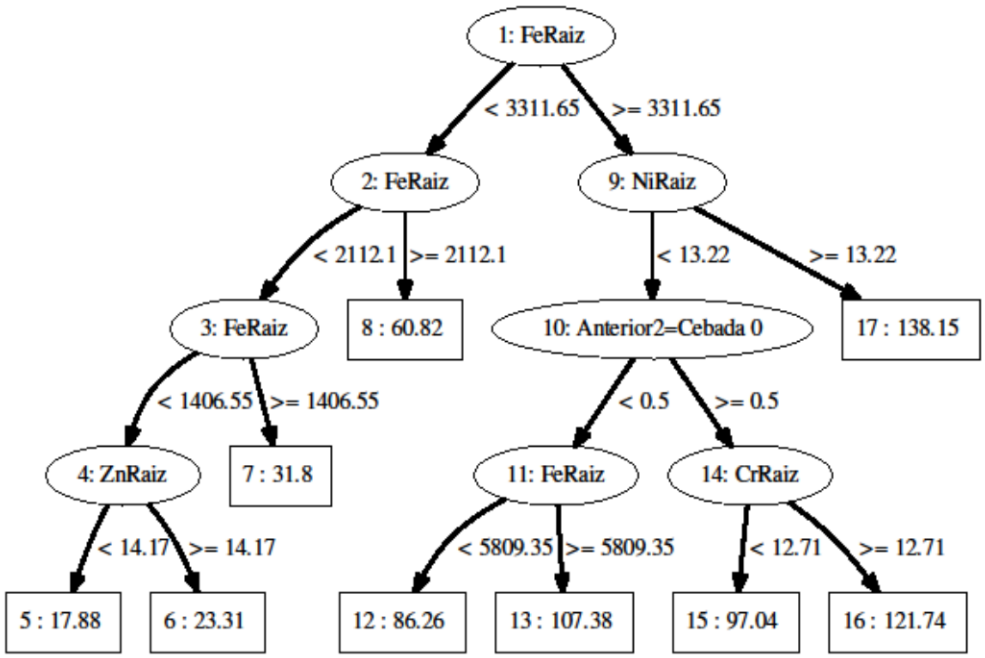


Figura 9-10 Árbol de regresión obtenido para el manganeso de la raíz.

Parámetros del árbol. Profundidad máxima=4, peso mínimo=8, datos para la poda=5.

Este árbol es el que mejor coeficiente de correlación muestra de los estudiados en cuanto a la presencia de metales en la raíz. Al igual que el árbol que representaba los valores en el suelo vemos la gran relación entre el hierro con el manganeso. Las variables que aparecen en el árbol son el hierro, zinc, níquel y cromo en la raíz y el cultivo de cebada de inicio después de dos años. Los valores más bajos de manganeso en la raíz se dan cuando los valores de hierro en la raíz son menores 1406,55 y el zinc está por debajo de 14,17. La mayor cantidad de manganeso aparece con valores de hierro en raíz de 3311,65 y níquel de 13,22 mg Kg^{-1}.

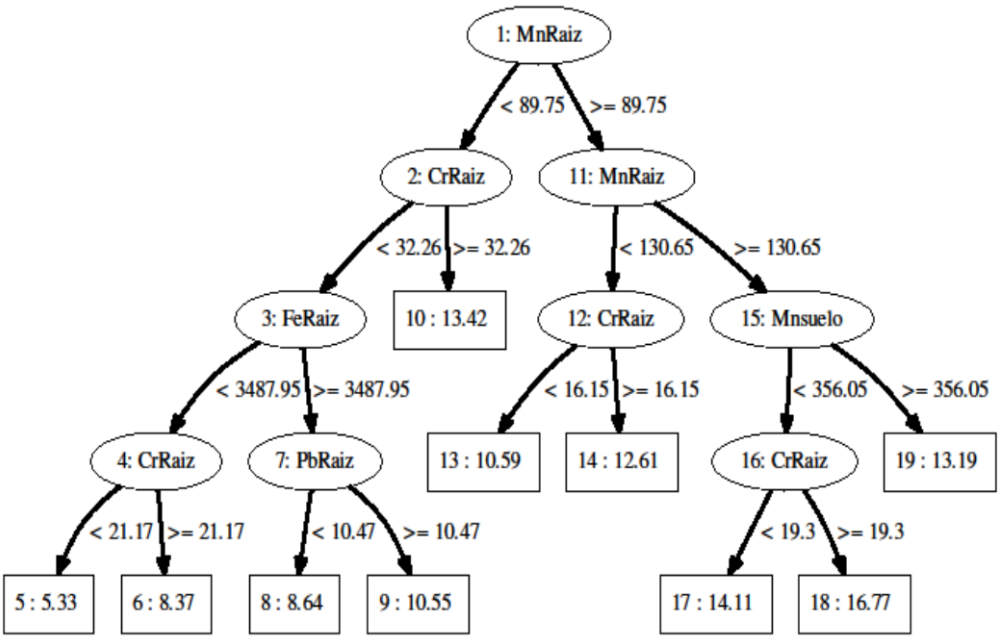


Figura 9-11 Árbol de regresión obtenido para el níquel de la raíz.

Parámetros del árbol. Profundidad máxima=4, peso mínimo=8, datos para la poda=3.

El níquel en la raíz aparece correlacionado con variables de los metales en la raíz como son el manganeso, el cromo, el hierro y el plomo así como el manganeso en el suelo. La rama de mayores valores de níquel en la raíz muestra un efecto antagónico debido a la cantidad de manganeso de suelo, esto es, se disminuyen los valores níquel en la raíz cuando los valores de manganeso en el suelo aumenta y se produce un incremento de los valores de níquel cuando hay mayores de cromo en la raíz ya que el cromo tiene una correlación positiva con la cantidad de níquel en la raíz.

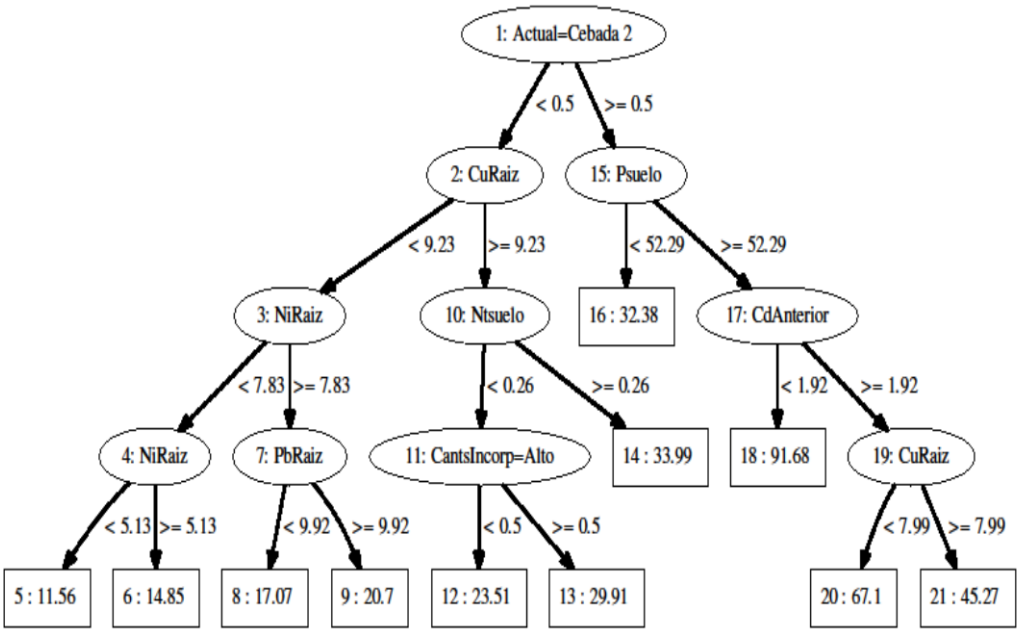


Figura 9-12 Árbol de regresión obtenido para el zinc de la raíz.

Parámetros del árbol. Profundidad máxima=4, peso mínimo=8, datos para la poda=7.

El zinc presente en la raíz tiene una estructura compleja con numerosas variables incluidas en su interpretación. Estas variables son: el cultivo de cebada por segunda vez, metales en la raíz como cobre , níquel, plomo, variables del suelo como son, la cantidad de nitrógeno, de fósforo y cadmio presente anteriormente y una adición alta de biosólido al suelo. La cantidad más alta de zinc en la raíz aparece cuando el cultivo ha sido la cebada y la cantidad de cadmio anteriormente en el suelo es inferior a 1,92 mg Kg^{-1}. Esta última variable tiene un efecto antagónico sobre la cantidad de zinc al igual que el cobre presente en la raíz.

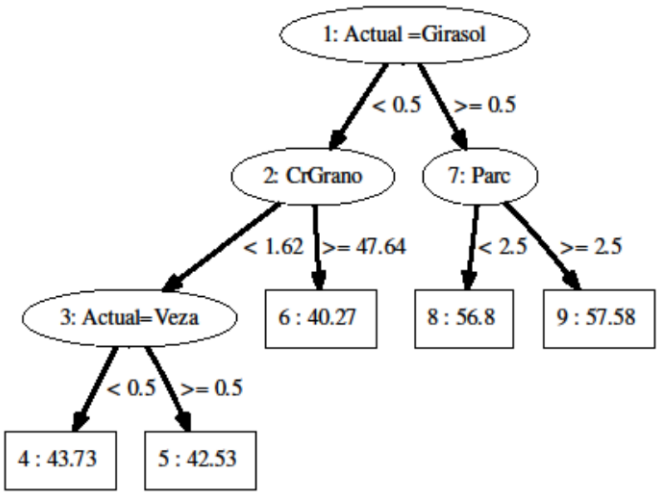


Figura 9-13 Árbol de regresión obtenido para el carbono del grano.

Parámetros del árbol. Profundidad máxima=4, peso mínimo=8, datos para la poda=11.

El árbol de regresión obtenido para el contenido de Carbono total del grano presenta como principales variables explicativas con efecto positivo al cultivo de girasol en el año actual y determinadas parcelas. Se observa una disminución de la cantidad de carbono en el grano cuando el cultivo actual ha sido la veza y la cantidad de cromo en el grano aumenta.

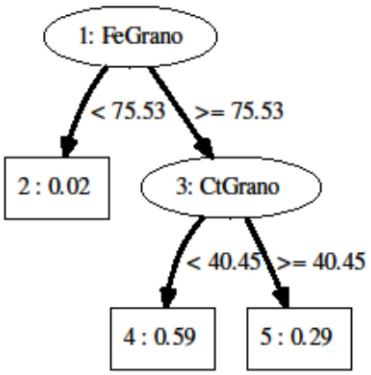


Figura 9-14 Árbol de regresión obtenido para el cadmio en el grano.

Parámetros del árbol. Profundidad máxima=2, peso mínimo=8, datos para la poda=7.

En este árbol vemos la relación entre la cantidad de hierro presente en el grano y el aumento de cadmio en el mismo. Por otro lado vemos como una mayor presencia de carbono en grano produce un efecto de disminución de la cantidad de cadmio en el grano.

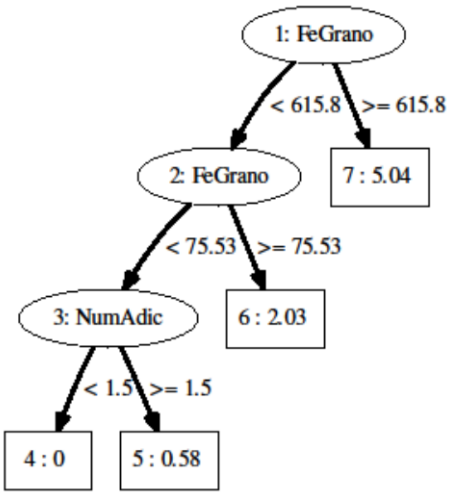


Figura 9-15 Árbol de regresión obtenido para el cromo en el grano.

Parámetros del árbol. Profundidad máxima=4, peso mínimo=8, datos para la poda=9.

El árbol de regresión obtenido para el contenido de cromo en grano presenta como principales variables explicativas la cantidad de hierro en el grano y el número de adiciones. Estas variables tienen una correlación positiva ofreciendo los menores valores cuando las cantidades hierro en grano son inferiores a 75,53 mg Kg^{-1} y el número de adiciones es menor a dos.

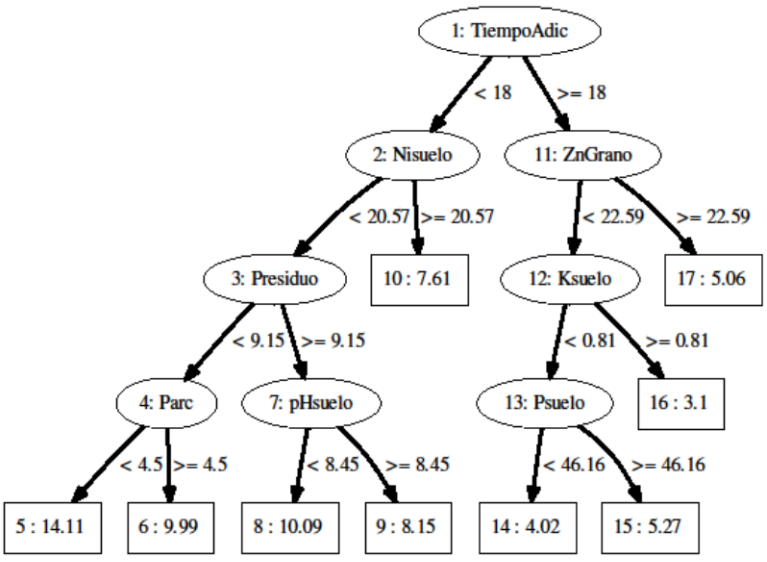


Figura 9-16 Árbol de regresión obtenido para el cobre en el grano.

Parámetros del árbol. Profundidad máxima=4, peso mínimo=8, datos para la poda=11.

El análisis de los contenidos de Cu en grano permiten la obtención de un árbol de regresión que muestra una dependencia de variables como el tiempo

transcurrido desde la última adición, el valor de pH del suelo, la concentración de Ni, K cambiable y P-Olsen en el suelo, así como el P-total presente en el biosólido y la presencia de otros metales en el grano como el Zn. Algunas de estas variables presentan una influencia negativa sobre la acumulación de Cu en grano como son: el tiempo desde la última adición, lo que puede indicar una disminución de formas disponibles de Cu en el suelo, el K de cambio, la cantidad de P-total adicionado en el residuo o el pH del suelo.

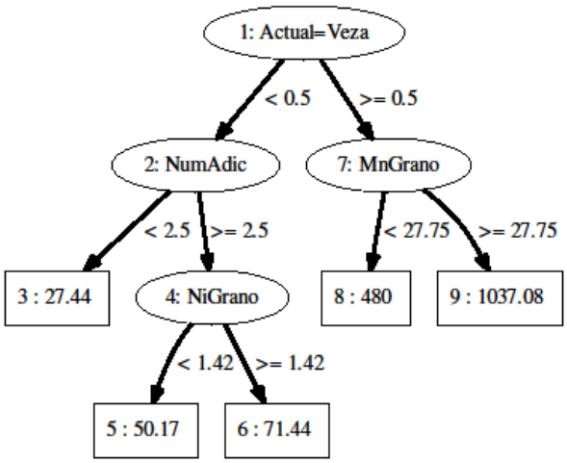


Figura 9-17 Árbol de regresión obtenido para el hierro en el grano.

Parámetros del árbol. Profundidad máxima=4, peso mínimo=8, datos para la poda=7.

El árbol de regresión obtenido para el contenido de hierro en grano presenta como principales variables explicativas el cultivo de veza, la cantidad de manganeso y níquel en el grano y el número de adiciones. Estas variables tienen una correlación positiva ofreciendo los menores valores cuando no se cultiva veza y el número de adiciones es menor a 3 mientras que los valores más altos se dan con el cultivo de veza y cantidades de manganeso en grano por encima de 27,75 mg Kg^{-1}.

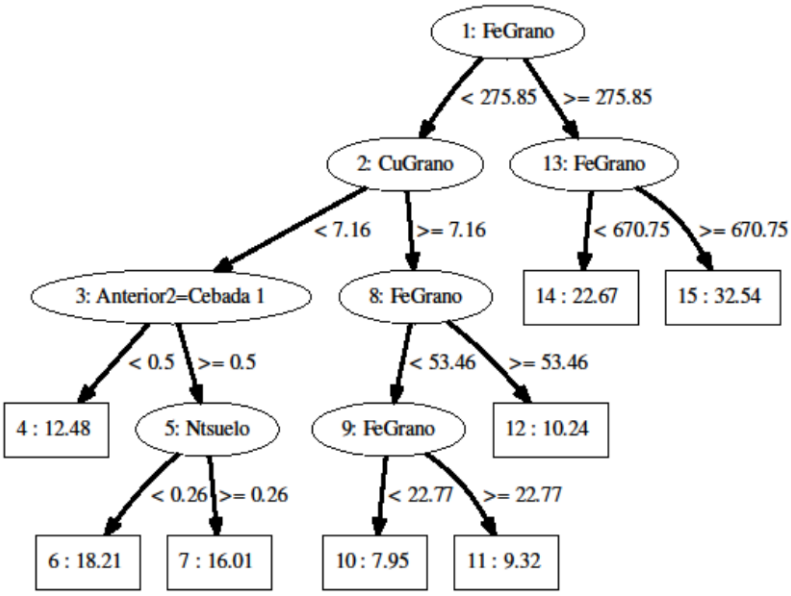


Figura 9-18 Árbol de regresión obtenido para el manganeso en el grano.

Parámetros del árbol. Profundidad máxima=4, peso mínimo=8, datos para la poda=9.

El árbol de regresión obtenido para el manganeso en grano muestra una dependencia de variables como el Fe y el Cu en grano, el cultivo de cebada dos años antes y la cantidad de nitrógeno presente en el suelo. Se observa un efecto antagonista por parte del cobre en grano y del nitrógeno en el suelo que disminuye la cantidad de Mn. El valor más alto de manganeso (32,54 mg Kg^{-1}) está determinado por la presencia de Fe en grano en cantidades superiores a 670,75 mg Kg^{-1}. El más bajo de manganeso (7,95 mg Kg^{-1}) también tiene como variable principal de nuevo al hierro, esta vez con cantidades menores a 22,77 mg Kg^{-1} y con valores de cobre en grano por encima de 7,16 mg Kg^{-1}.

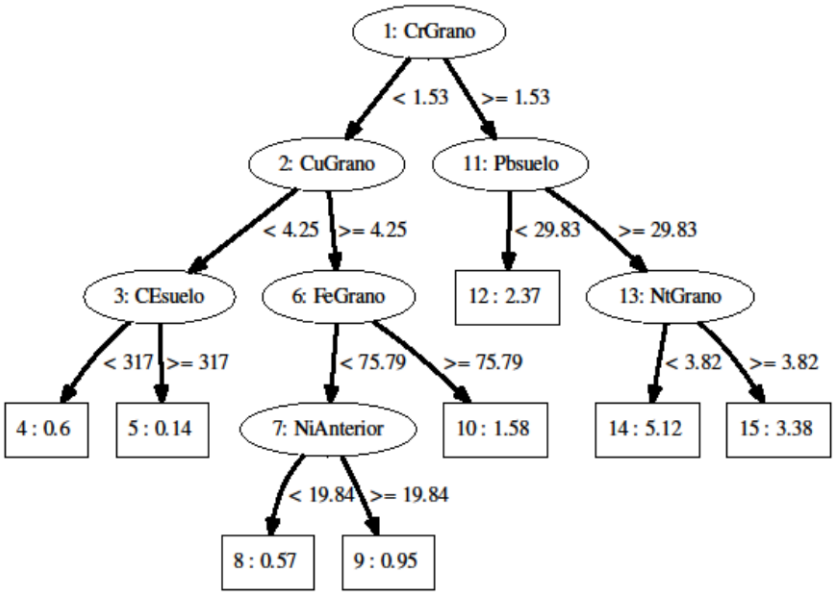


Figura 9-19 Árbol de regresión obtenido para el níquel en el grano.

Parámetros del árbol. Profundidad máxima=4, peso mínimo=8, datos para la poda=3.

En la representación del níquel en el grano podemos ver una participación de metales como son el cromo y cobre junto con el nitrógeno acumulado en el grano. Por otro lado aparecen dos variables pertenecientes al suelo como son la cantidad de plomo y la conductividad eléctrica. Las mayores cantidades de níquel en el grano se dan cuando el cromo está por encima de 1,53 y el plomo en el suelo es igual o superior a 29,83 mg Kg^{-1}. Se observan correlaciones negativas entre la cantidad de níquel y el nitrógeno presente en el grano. También vemos un aumento del níquel en el grano cuando se tienen conductividades inferiores a 317 $\mu S \cdot cm^{-1}$ en el suelo. En la otra rama de la variable conductividad vemos los menores valores (0,14 mg Kg^{-1}) de níquel en el grano.

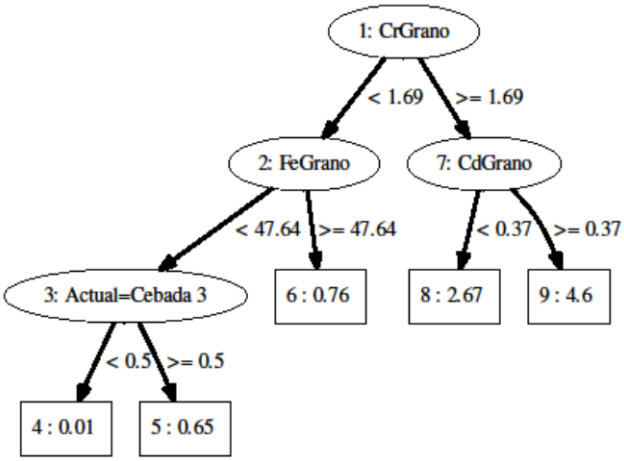


Figura 9-20 Árbol de regresión obtenido para el plomo en el grano.

Parámetros del árbol. Profundidad máxima=4, peso mínimo=8, datos para la poda=3.

Las variables que intervienen en la representación de este árbol son cromo, hierro y cadmio en grano y la cebada como cultivo por tercer año. Todas estas variables tienen efectos de incremento en la cantidad de plomo en el grano. Se establece una alta correlación entre el plomo y el cromo en el grano siendo ambos metales conocidos por su baja movilidad, además cuando las cantidades de cromo son menores 1,69 mg Kg^{-1}, se ve la participación en el árbol del hierro presente en el grano incrementando la cantidad de acumulación de plomo en el grano. Los menores valores de plomo se obtienen cuando el cromo y el hierro en el grano están por debajo de 1,69 y de 47,64 respectivamente y se ha cultivado cebada por tercera vez.

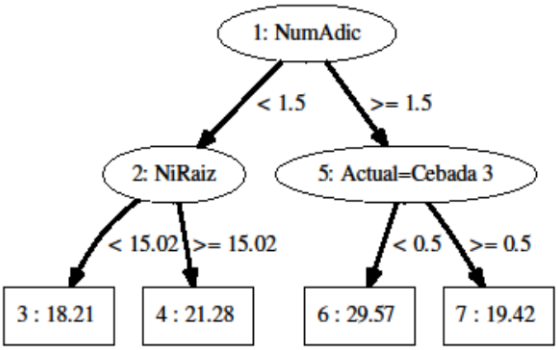


Figura 9-21- Árbol de regresión obtenido para el zinc en el grano.

Parámetros del árbol. Profundidad máxima=4, peso mínimo=8, datos para la poda=9.

Las cantidades de zinc en el grano están influidas por factores no estrictamente dependientes del grano como en otros árboles estudiados anteriormente. Podemos comprobar que en este caso la mayor cantidad de zinc en el grano se encuentra en presencia de cultivos que no sean la cebada y si el número de

adiciones es superior a 1. El cultivo de la cebada por tercer año tiene un efecto antagonista produciendo una disminución de la cantidad de zinc en el grano.

Andadilla

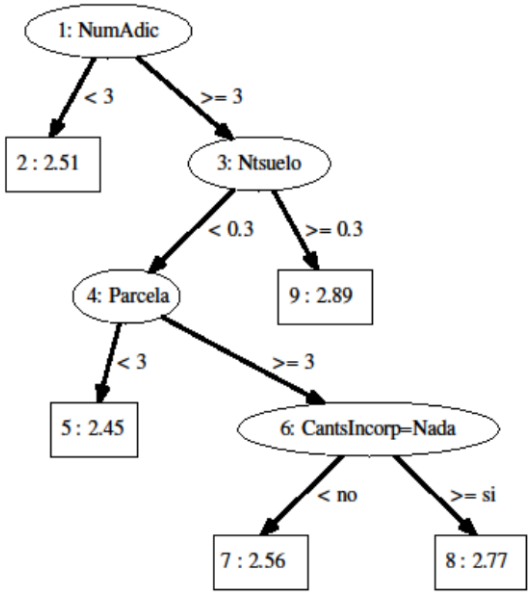


Figura 9-22 Árbol de regresión obtenido de materia orgánica del suelo.

Parámetros del árbol. Profundidad máxima=4, peso mínimo=8, datos para la poda=11.

La interpretación que se puede dar al siguiente árbol sería que la materia orgánica se encuentra muy influida por la cantidad de nitrógeno total que se encuentra presente en el suelo y el número de adiciones, que cuando su valor es de 3 da como resultado el mayor valor 2,89 %. La disminución del nitrógeno conlleva menores cantidades de materia orgánica en el suelo.

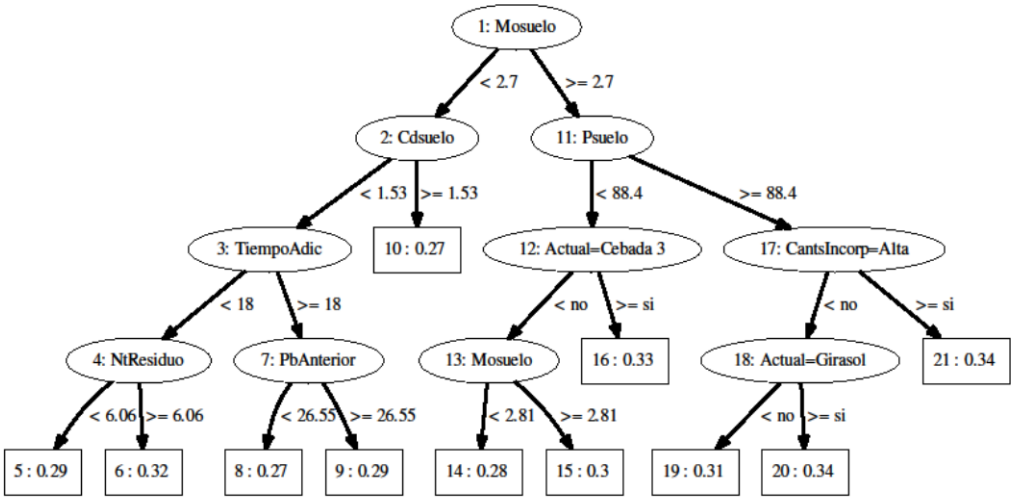


Figura 9-23 Árbol de regresión obtenido de nitrógeno en el suelo.

Parámetros del árbol. Profundidad máxima=4, peso mínimo=8, datos para la poda=7.

El árbol de regresión obtenido para el contenido de nitrógeno en el suelo muestra como principales variables que participan en este modelo si el cultivo ha sido girasol y la cantidad de materia orgánica que está presente en el suelo. También observamos correlaciones positivas con otras variables como son la cantidad de cadmio o plomo presente en el suelo el año anterior, el fósforo presente en el suelo y cantidades altas de adiciones. La representación de este árbol también muestra una estrecha relación entre la materia orgánica y el nitrógeno total, dependiendo del número de adiciones de compost así como del fósforo presente en el suelo.

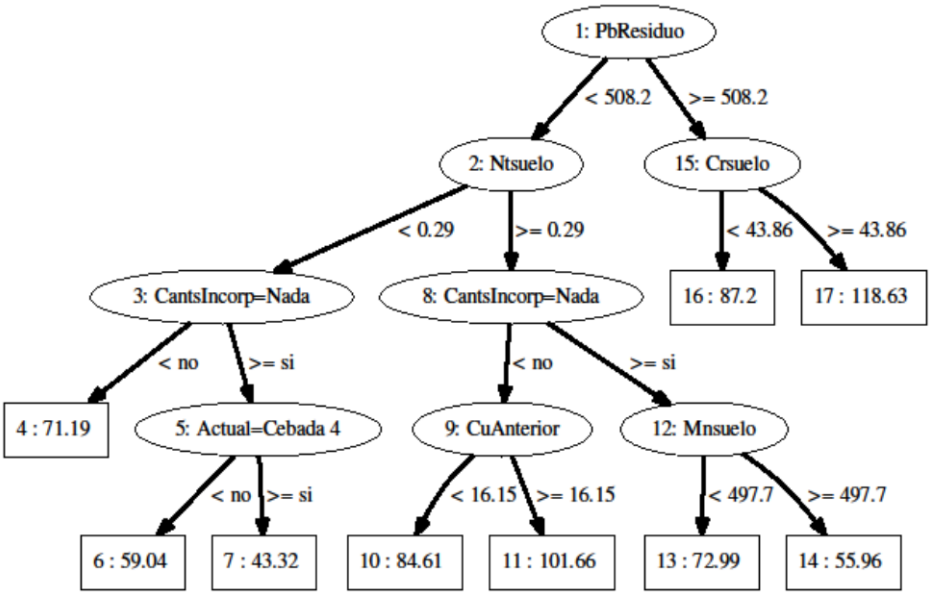


Figura 9-24 Árbol de regresión obtenido del fósforo en el suelo.

Parámetros del árbol. Profundidad máxima=4, peso mínimo=8, datos para la poda=9.

El contenido de P del suelo está correlacionado con cantidades de Pb del residuo incorporado, el cultivo en el año de cebada y con el cobre y cromo presente en el suelo.

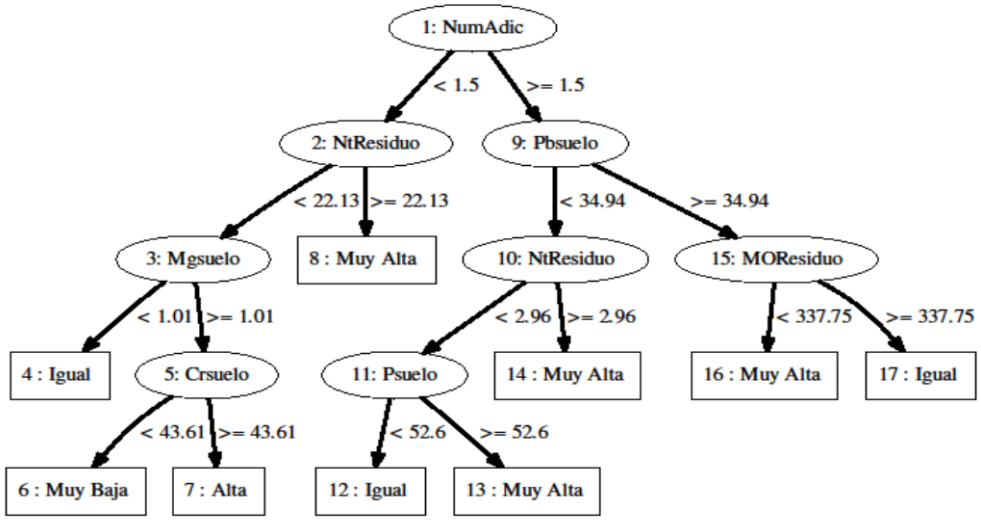


Figura 9-25- Árbol de regresión obtenido de productividad de los cultivos.

Parámetros del árbol. Profundidad máxima=4, peso mínimo=8, datos para la poda=5. Instancias clasificadas correctamente = 0,7933.

La productividad está directamente relacionada con la cantidad de nutrientes incorporados en el compost principalmente por el número de adiciones, con su contenido de nitrógeno y con el contenido de P en el suelo y la materia orgánica del residuo.

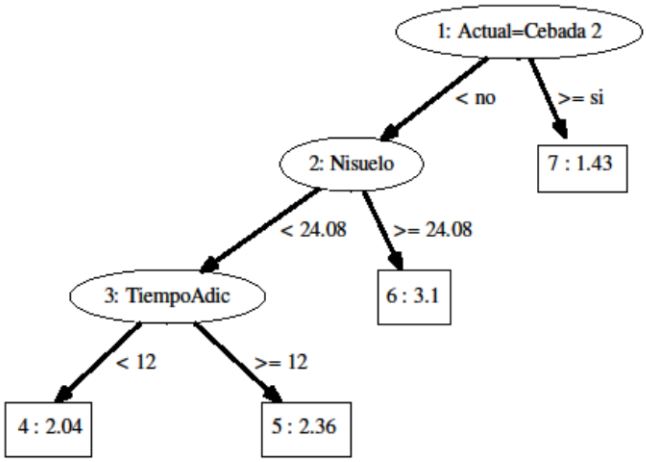


Figura 9-26 Árbol de regresión obtenido de cadmio en el suelo.

Parámetros del árbol. Profundidad máxima=4, peso mínimo=8, datos para la poda=3.

Se observa cómo el cultivo de la cebada, el níquel en el suelo y el tiempo de adición tienen una influencia en la cantidad de cadmio en el suelo. Los valores más bajos se dan cuando se ha cultivado cebada por segunda vez.

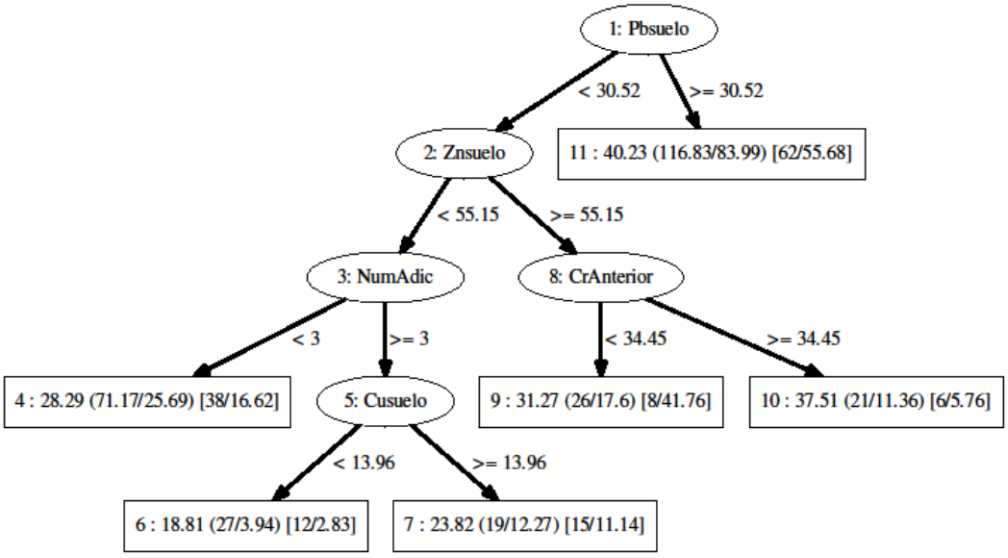


Figura 9-27 Árbol de regresión obtenido de cromo en el suelo.

Parámetros del árbol. Profundidad máxima=4, peso mínimo=8, datos para la poda=3.

En este árbol podemos ver las relaciones entre las cantidades de cromo con el plomo en el suelo, el zinc, el cobre, el número de adiciones y las cantidades presentes de cromo en el suelo anteriormente. Todas las variables dependientes de los metales tienen una influencia positiva con la cantidad de cromo en el suelo mientras el número de adiciones tiene correlación negativa con la cantidad de cromo en el suelo.

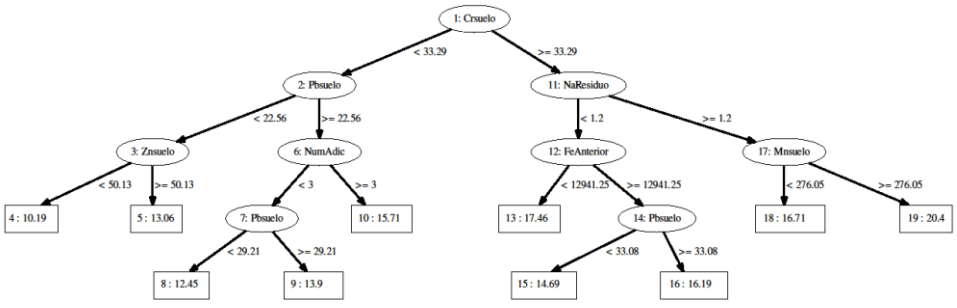


Figura 9-28 Árbol de regresión obtenido de cobre en el suelo.

Parámetros del árbol. Profundidad máxima=4, peso mínimo=8, datos para la poda=7.

El árbol de regresión obtenido para el contenido de cobre en el suelo muestra como principales variables que participan en este modelo a las concentraciones de Cr, Pb, Zn y Mn del suelo, también la cantidad de Fe presente los años

anteriores en el suelo, el número de aplicaciones del biosólido y la cantidad de Na presente en el residuo. Los valores más altos se dan con cantidades de cromo en el suelo de 33,29 mg Kg^{-1} y de manganeso por encima de 276,05 y con una cantidad de Na en el residuo mayor de 1,2 mg Kg^{-1} mientras que los menores valores vienen determinados por valores inferiores de 33,29 de cromo, de plomo por debajo de 22,56 mg Kg^{-1} y de zinc en el suelo menor de 50,13 mg Kg^{-1}.

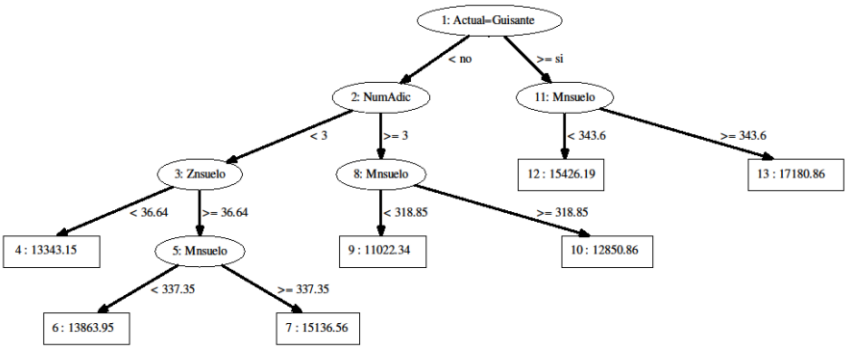


Figura 9-29 Árbol de regresión obtenido de hierro en el suelo.

Parámetros del árbol. Profundidad máxima=4, peso mínimo=8, datos para la poda=3.

El árbol de regresión obtenido para el contenido de hierro en el suelo muestra como principales variables que participan en este modelo a las concentraciones de Mn y Zn del suelo, el número de aplicaciones de enmienda y tener como cultivo actual guisante.

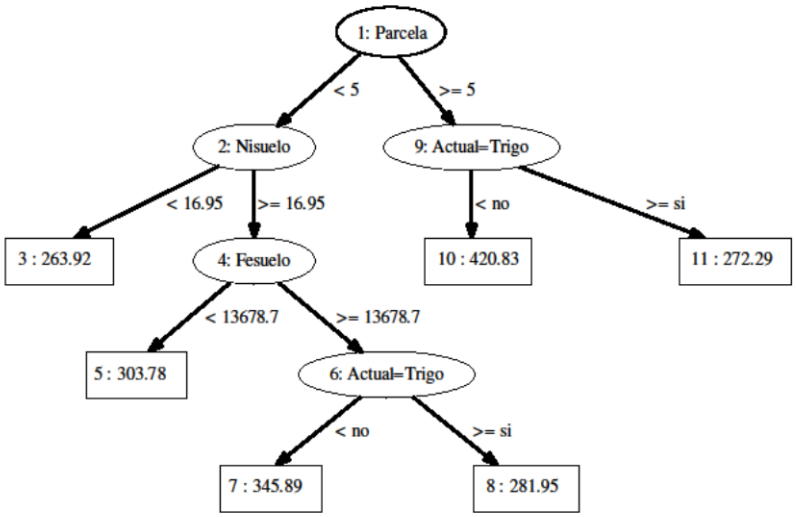


Figura 9-30 Árbol de regresión obtenido de manganeso en el suelo.

Parámetros del árbol. Profundidad máxima=4, peso mínimo=8, datos para la poda=11.

El árbol de regresión obtenido para el contenido de manganeso en el suelo muestra como principales variables que participan en este modelo a las concentraciones de Fe, Ni y tener como cultivo actual trigo. Precisamente este

último parámetro es el único que tiene un efecto antagonista sobre las cantidades de manganeso en el suelo. Se observa diferencias entre las cantidades de manganeso entre la parcela 5 del resto.

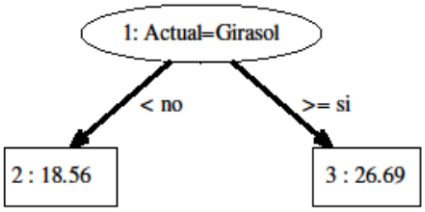


Figura 9-31 Árbol de regresión obtenido de níquel en el suelo.

Parámetros del árbol. Profundidad máxima=4, peso mínimo=8, datos para la poda=3.

En las medidas de níquel en el suelo la única variable influyente ha sido el cultivo de girasol en el año.

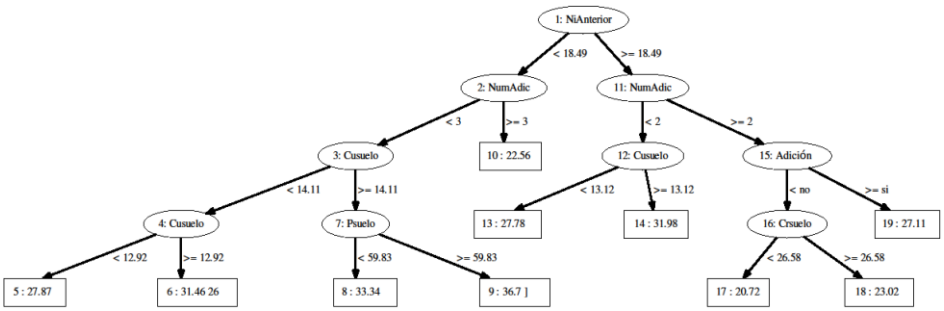


Figura 9-32 Árbol de regresión obtenido de plomo en el suelo.

Parámetros del árbol. Profundidad máxima=4, peso mínimo=8, datos para la poda=11.

Se observa un árbol con una influencia de la cantidad de materia orgánica en el suelo, número de adiciones y dosis y de cantidades de metales presentes en el suelo, como el cobre y el cromo.

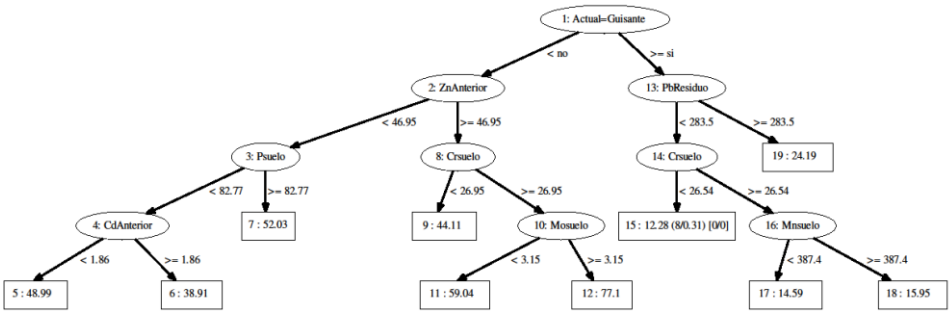


Figura 9-33 Árbol de regresión obtenido de zinc en el suelo.

Parámetros del árbol. Profundidad máxima=4, peso mínimo=8, datos para la poda=7.

Las variables que describen este árbol son el cultivo actual, la cantidad de plomo, cromo, manganeso, fósforo y materia orgánica en el suelo, también tienen su participación en el árbol dos variables referidas a las cantidades en años anteriores de zinc y cadmio. Se observa el efecto antagónico de las cantidades de cadmio anterior sobre el zinc presente en el suelo. Los valores más bajos se dan cuando se ha cultivado guisante, la cantidad de plomo en el residuo es inferior a 283,5 mg Kg^{-1} y el cromo en el suelo es menor a 26,54. Por otro lado la mayor cantidad de zinc en el suelo resulta de la combinación de cantidades de materia orgánica y de cromo en el suelo con valores de 3,15 y 26,95 mg Kg^{-1}, respectivamente, junto con valores de zinc en los años anteriores de 46,95 y no haber cultivado guisante ese año.

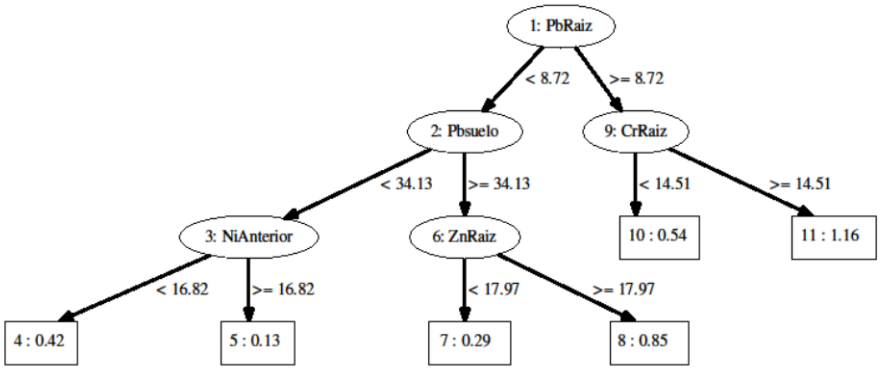


Figura 9-34 Árbol de regresión obtenido de cadmio en la raíz.

Parámetros del árbol. Profundidad máxima=4, peso mínimo=8, datos para la poda=11.

La cantidad de cadmio en la raíz se ve influida por algunos metales presentes en la raíz. Estos metales son el plomo, el cromo y el zinc en la raíz, plomo en el suelo y la cantidad de níquel anteriormente presente en el suelo. El plomo en raíz con valores superiores a 8,72 mg Kg^{-1} da lugar a las mayores cantidades de cadmio en raíz. Los menores valores de cadmio se observan cuando el plomo en raíz es menor a 8,72, el plomo en el suelo es menor a 34,13 y el níquel anteriormente es de 16,82 mg Kg^{-1}.

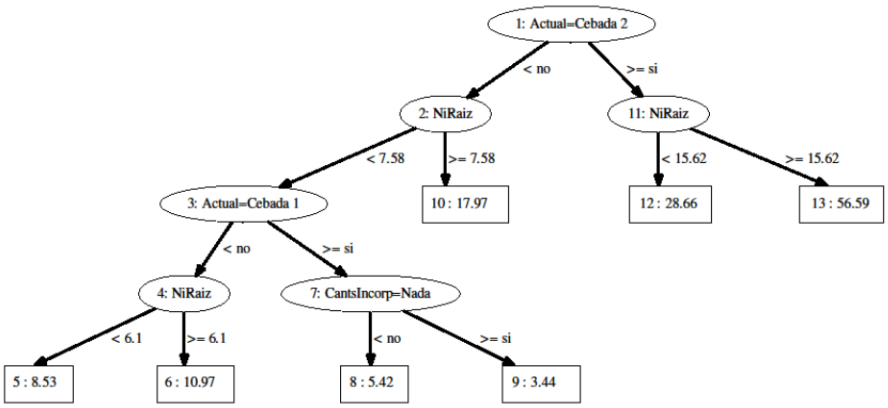


Figura 9-35 Árbol de regresión obtenido de cromo en la raíz.

Parámetros del árbol. Profundidad máxima=4, peso mínimo=8, datos para la poda=5.

Las cantidades de cromo medidas en la raíz dependen de la cantidad de níquel presente en la raíz, el cultivo de cebada y las cantidades incorporadas. Los mayores valores de cromo en la raíz se obtienen cuando se tienen cantidades de níquel en la raíz con valores superiores a 15,62 mg Kg^{-1}. Los menores valores aparecen con las parcelas control.

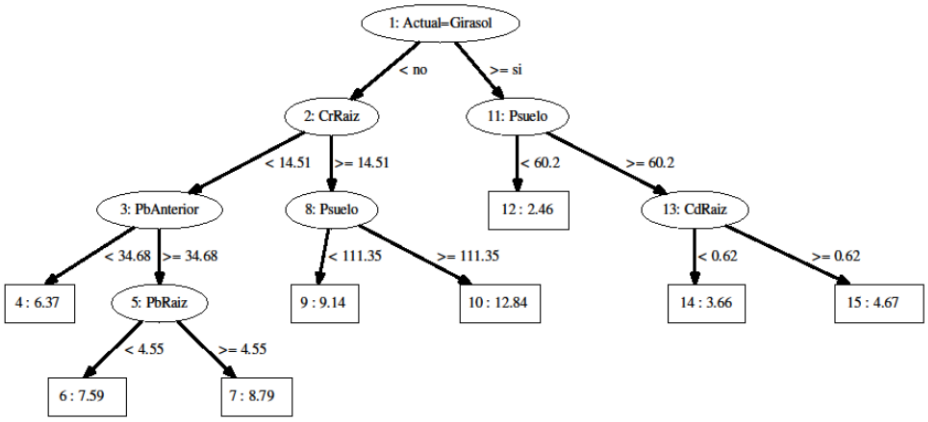


Figura 9-36 Árbol de regresión obtenido de cobre en la raíz.

Parámetros del árbol. Profundidad máxima=4, peso mínimo=8, datos para la poda=11.

El árbol de regresión obtenido para el contenido de Cu en la raíz presenta como principales variables explicativas al cultivo actual girasol, el fósforo en el suelo y el cromo en la raíz. Tenemos otros metales en la raíz como son el cadmio y plomo, este último también en cantidades presentes en el suelo. Las menores cantidades se dan con el cultivo del girasol y con cantidades de fósforo en el suelo inferiores a 60,2 mg Kg^{-1}. Los valores máximos vienen definidos por el no cultivo de girasol y por cantidades de cromo en raíz de 14,51 mg Kg^{-1} y de fósforo en suelo de 111,35 mg Kg^{-1}.

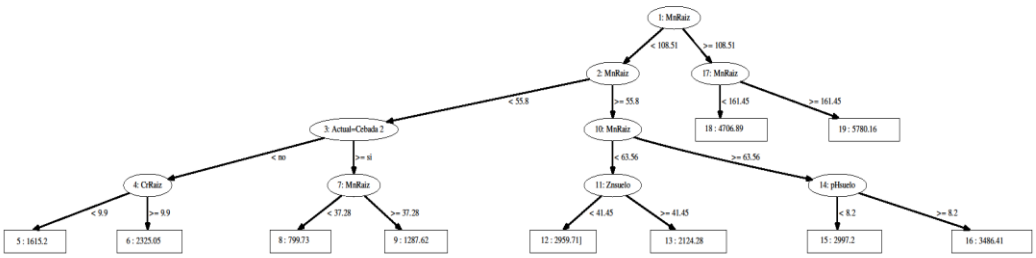


Figura 9-37 Árbol de regresión obtenido de hierro en la raíz.

Parámetros del árbol. Profundidad máxima=4, peso mínimo=8, datos para la poda=9.

Este árbol es el que mejor coeficiente de correlación, junto con el manganeso, muestra de los estudiados en cuanto a la presencia de metales en la raíz. El hierro en la raíz tiene como variables más relevantes, en la formación del gráfico del árbol, al manganeso, y otras variables como el cromo en la raíz, el cultivo de cebada por segunda vez, y dos variables del suelo como son el pH y la cantidad de zinc. Es destacable que en el caso del zinc en el suelo tiene un efecto antagonista con la cantidad de hierro de la raíz, los aumentos en la cantidad de zinc en el suelo producen una disminución en la cantidad de hierro encontrado en la raíz.

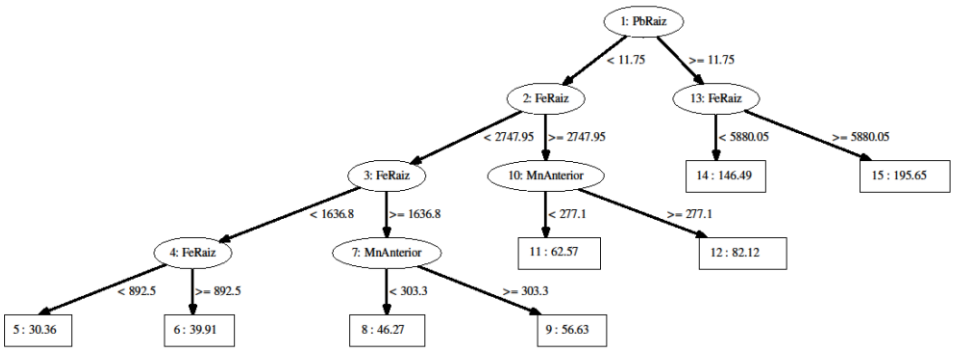


Figura 9-38 Árbol de regresión obtenido de manganeso en la raíz.

Parámetros del árbol. Profundidad máxima=4, peso mínimo=8, datos para la poda=5.

Al igual que el árbol que representaba los valores en el suelo vemos la gran relación entre el hierro con el manganeso. Las variables que aparecen en el árbol son el hierro y plomo en la raíz. También las cantidades de manganeso presentes en el suelo anteriormente tienen presencia entre las variables. Los valores más bajos de manganeso en la raíz se dan cuando los valores de hierro en la raíz son menores 892,5 mg Kg^{-1} y el plomo está por debajo de 11,75 mg Kg^{-1}. La mayor cantidad de manganeso aparece con valores de hierro en raíz de 5880,05 mg Kg^{-1}.

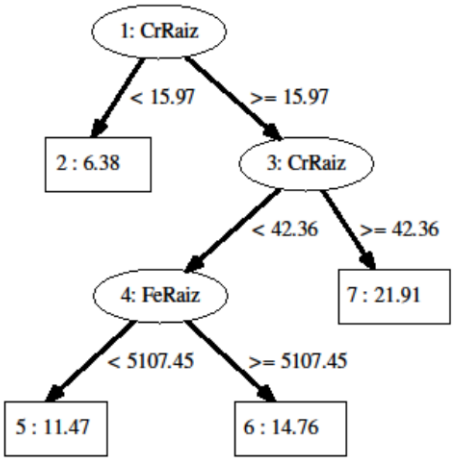


Figura 9-39 Árbol de regresión obtenido de níquel en la raíz.

Parámetros del árbol. Profundidad máxima=4, peso mínimo=8, datos para la poda=5.

El níquel en la raíz aparece correlacionado con variables de los metales en la raíz como son el cromo y el hierro. La rama de mayores valores de níquel en la raíz se presenta con valores superiores a 42,36 mg Kg^{-1} de cromo en la raíz. El menor valor de níquel en raíz, 6,38 mg Kg^{-1}, se produce con valores inferiores a 15,97 mg Kg^{-1} de cromo en la raíz.

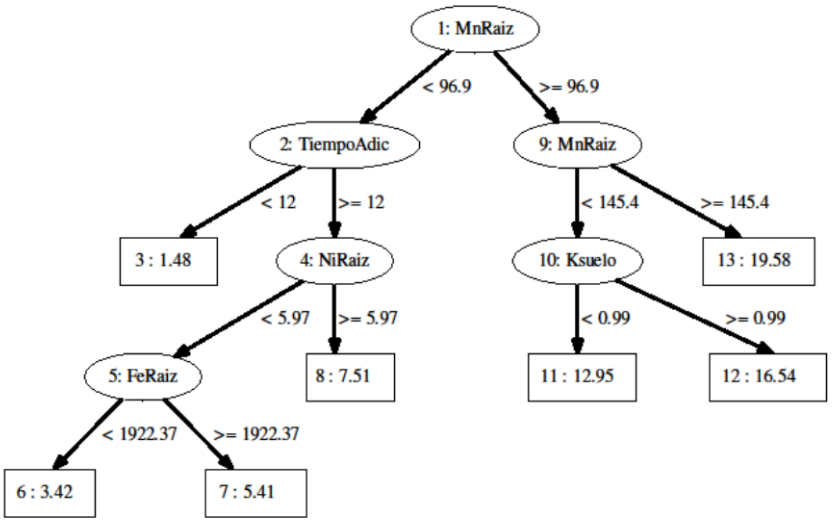


Figura 9-40 Árbol de regresión obtenido de plomo en la raíz.

Parámetros del árbol. Profundidad máxima=4, peso mínimo=8, datos para la poda=11.

El plomo presente en la raíz está relacionado con variables de la raíz como son la cantidad de manganeso, hierro, níquel, de potasio en el suelo y el tiempo de adición. Los valores más bajos de plomo en la raíz tienen como variables determinantes al manganeso en raíz (96,9 mg Kg^{-1}) y tiempo de adición de 12

meses. Los valores más altos, 19,58 mg Kg^{-1}, tienen como factor principal los incrementos de manganeso en la raíz.

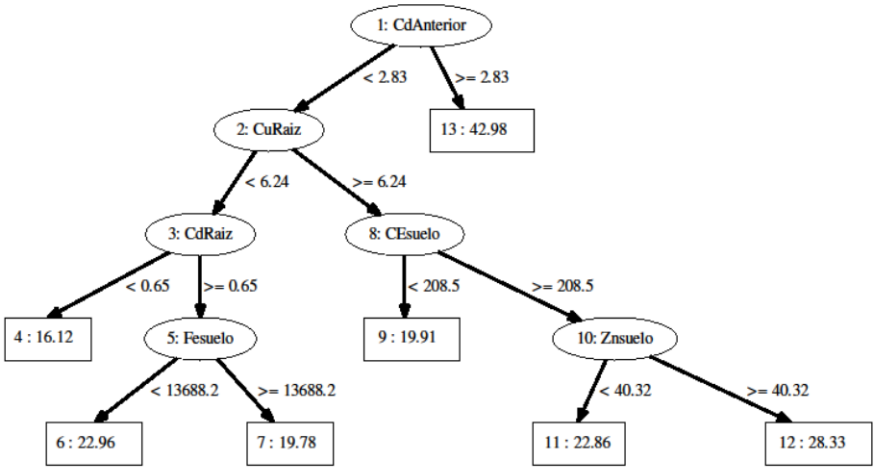


Figura 9-41 Árbol de regresión obtenido de zinc en la raíz.

Parámetros del árbol. Profundidad máxima=4, peso mínimo=8, datos para la poda=7.

El zinc presente en la raíz tiene como factores principales el cobre en la raíz y el cadmio anterior en suelo. Otras variables que participan en la representación son la conductividad eléctrica, el hierro y el zinc del suelo y el cadmio de la raíz. El hierro en el suelo tiene un efecto antagonista. La cantidad más alta de zinc en la raíz aparece cuando el cadmio anterior en el suelo es mayor de 2,83 mg Kg^{-1}. Las menores cantidades de zinc se presentan con cadmio por debajo de 0,65 mg Kg^{-1} y con cobre por debajo de 6,24 mg Kg^{-1}, ambos en raíz.

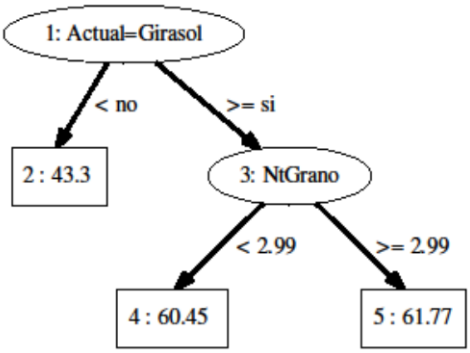


Figura 9-42 Árbol de regresión obtenido de carbono en el grano.

Parámetros del árbol. Profundidad máxima=4, peso mínimo=16, datos para la poda=7.

Vemos un incremento de la cantidad de carbono en grano cuando el cultivo es el girasol y una correlación leve con la cantidad de nitrógeno en grano. El menor valor (43,3 mg Kg^{-1}) se da cuando no se ha cultivado girasol y el mayor

(61,77 mg Kg^{-1}) cuando sí se cultiva y el nitrógeno en grano está por encima de 2,99 %.

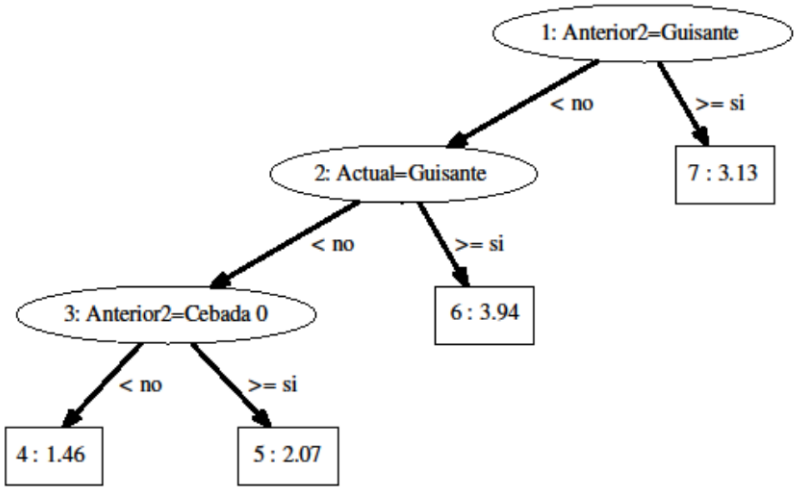


Figura 9-43 Árbol de regresión obtenido de nitrógeno en el grano.

Parámetros del árbol. Profundidad máxima=3, peso mínimo=8, datos para la poda=11.

El árbol de regresión obtenido para el contenido de nitrógeno total del grano presenta como principales variables explicativas el cultivo de guisante en el año actual, que además fue el primero en el que se produjo la adición de enmiendas, y hace dos años y de la cebada previa al ensayo dos años antes. Se observa una disminución de la cantidad de nitrógeno en el grano (1,46%) cuando no se ha cultivado guisante. El cultivo de guisante proporciona valores más altos de nitrógeno en grano, estos valores también corresponden a los años en los que se ha incorporado enmiendas orgánicas.

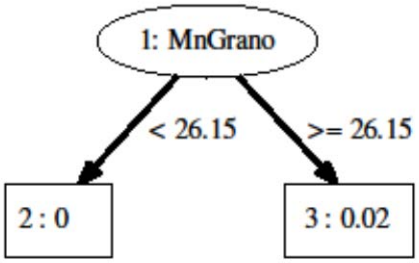


Figura 9-44 Árbol de regresión obtenido de cadmio en el grano.

Parámetros del árbol. Profundidad máxima=3, peso mínimo=8, datos para la poda=7.

El árbol creado sólo tiene una variable que influya siendo esta la cantidad de manganeso en el grano.

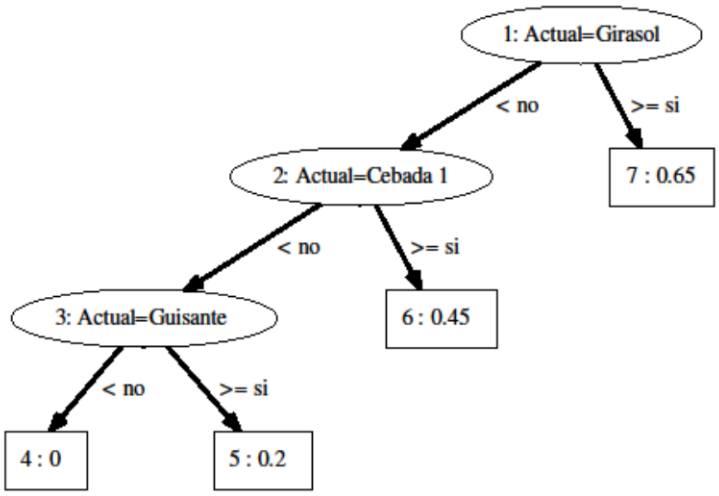


Figura 9-45 Árbol de regresión obtenido de cromo en el grano.

Parámetros del árbol. Profundidad máxima=3, peso mínimo=8, datos para la poda=9.

La cantidad de cromo presente en el grano muestra correlaciones con los cultivos actuales, girasol, cebada 1 y guisante. Los mayores valores se observan con el cultivo del girasol y los más bajos cuando no se cultiva guisante. No hay grandes diferencias entre los valores de las ramas.

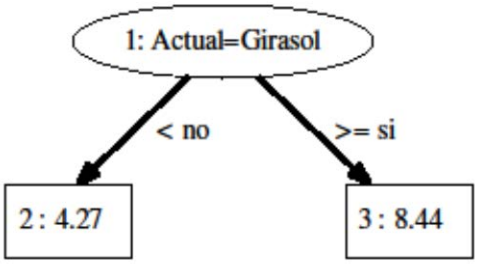


Figura 9-46 Árbol de regresión obtenido de cobre en el grano.

Parámetros del árbol. Profundidad máxima=4, peso mínimo=8, datos para la poda=7.

La cantidad de cobre presente en el grano establece correlaciones con el cultivo del girasol.

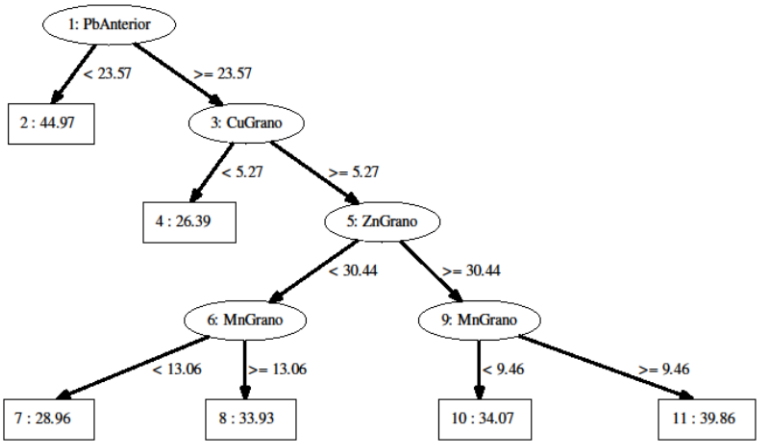


Figura 9-47 Árbol de regresión obtenido de hierro en el grano.

Parámetros del árbol. Profundidad máxima=4, peso mínimo=8, datos para la poda=3.

Las cantidades de plomo anterior en el suelo es la variable principal. Otras variables que se ven en el árbol son el cobre, manganeso y zinc en el grano. El coeficiente de correlación de este metal en grano es muy pequeño dándonos un árbol poco representativo.

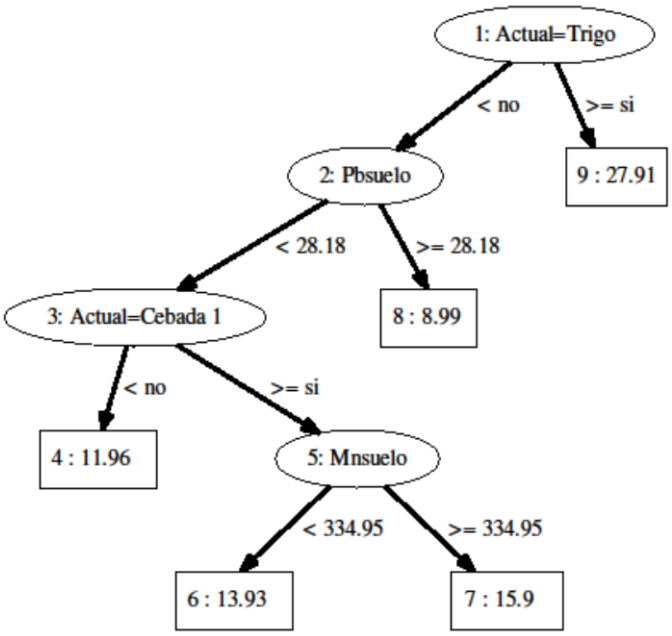


Figura 9-48 Árbol de regresión obtenido de manganeso en el grano.

Parámetros del árbol. Profundidad máxima=4, peso mínimo=8, datos para la poda=11.

El árbol de regresión obtenido para el manganeso en grano tiene un r = 0,9457 y muestra una dependencia de variables como son el cultivo actual de cebada 1 y de trigo y el plomo y el manganeso en el suelo. El valor más alto de

manganeso (27,91 mg Kg^{-1}) está determinado por el cultivo de trigo en el año actual. La presencia de plomo en el suelo en cantidades de 28,18 mg Kg^{-1} ofrece un resultado de 8,99 mg Kg^{-1} de manganeso en el grano siendo este el valor más bajo.

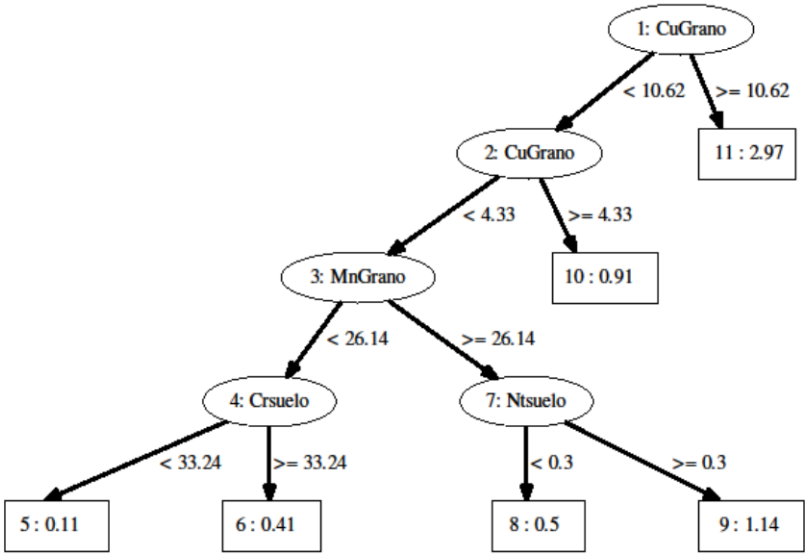


Figura 9-49 Árbol de regresión obtenido de níquel en el grano.

Parámetros del árbol. Profundidad máxima=4, peso mínimo=8, datos para la poda=9.

Las variables presentes en la representación de níquel en grano son el cobre y manganeso en grano y el cromo y el nitrógeno en el suelo. Una gran cantidad de cobre refleja un aumento de níquel en grano. El menor valor de níquel se encuentra cuando la cantidad de cobre en grano es inferior a 4,33 mg Kg^{-1}, manganeso en grano por debajo de 26,14 mg Kg^{-1} y cromo en suelo con menos de 33,24 mg Kg^{-1}.

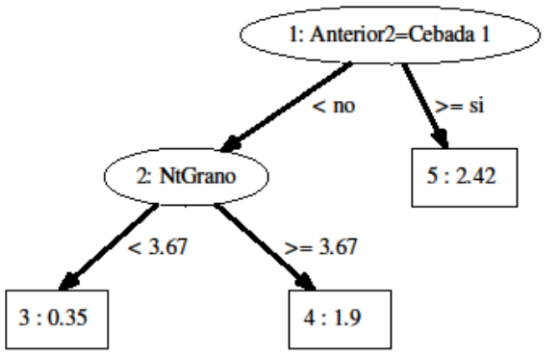


Figura 9-50 Árbol de regresión obtenido de plomo en el grano.

Parámetros del árbol. Profundidad máxima=4, peso mínimo=8, datos para la poda=3.

Las variables que intervienen en la representación de este árbol son el cultivo de cebada por primera vez dos años antes. Las cantidades de nitrógeno en grano por debajo de 3,67 % da los menores valores de plomo en grano (0,35 mg Kg^{-1}).

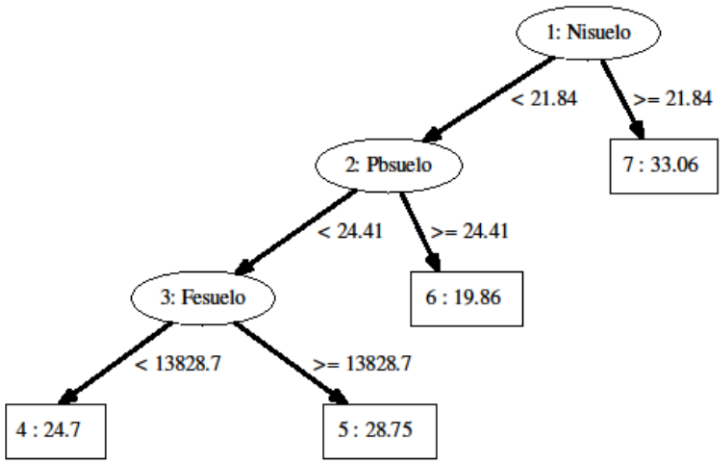


Figura 9-51 Árbol de regresión obtenido de zinc en el grano.

Parámetros del árbol. Profundidad máxima=4, peso mínimo=8, datos para la poda=3.

El árbol de regresión obtenido para el zinc en grano tiene un *r* = 0,5985 y muestra una dependencia de variables como son el níquel, el plomo y el hierro en el suelo. El valor más alto de zinc (33,05 mg Kg^{-1}) está determinado por el níquel en suelo por encima de 21,84 mg Kg^{-1}. Cuando el plomo en el suelo es de 24,41 se obtiene menor cantidad de zinc en el grano (19,86 mg Kg^{-1}).

Serrana

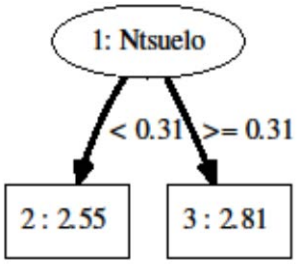


Figura 9-52 Árbol de regresión obtenido de materia orgánica en el suelo.

Parámetros del árbol. Profundidad máxima=4, peso mínimo=8, datos para la poda=3.

La única variable que influye en la cantidad de materia orgánica presente en el suelo es la cantidad de nitrógeno. Con valores superiores a 0,31% de nitrógeno se encuentra el máximo de materia orgánica con 2,81 %.

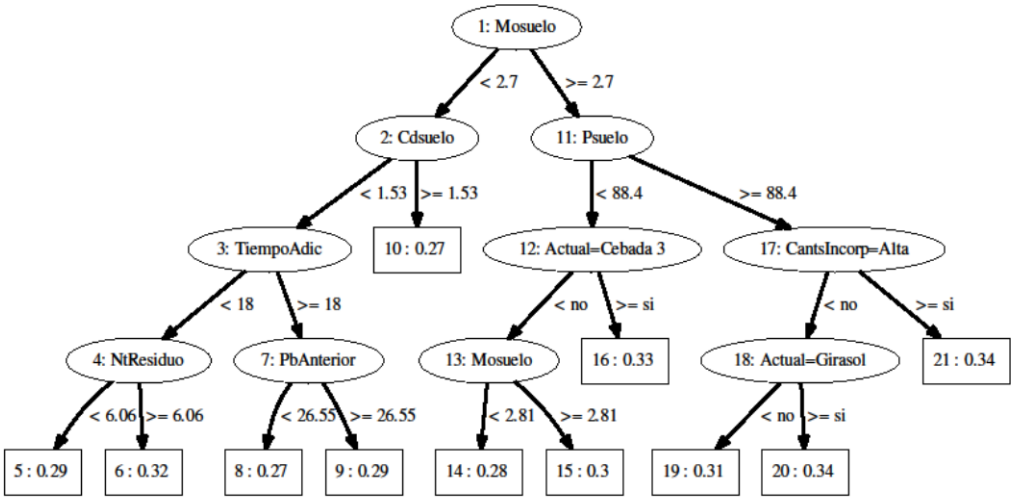


Figura 9-53 Árbol de regresión obtenido de nitrógeno en el suelo.

Parámetros del árbol. Profundidad máxima=4, peso mínimo=8, datos para la poda=3.

El árbol de regresión obtenido para el contenido de nitrógeno en el suelo presenta múltiples variables que participan en este modelo algunas relativas a los tipos de cultivos presentes, en el caso de la cebada por tercera vez o del girasol, otras relacionadas con las características de residuo incorporado, la cantidad de nitrógeno del residuo o una incorporación en dosis altas. También vemos la influencia del tiempo transcurrido desde la última adición y algunas características del suelo como son la cantidad de materia orgánica, el fósforo y el cadmio en el momento de la cosecha y la cantidad de plomo presente en el suelo el año anterior. Los valores más altos se encuentran con cantidades altas incorporadas, con cantidades de materia orgánica y de fósforo en el suelo de 2,7 % y 88,4 mg Kg^{-1} respectivamente. Los menores valores se dan con valores de plomo, cadmio y materia orgánica menores a 26,55 mg Kg^{-1}, 1,53 mg Kg^{-1} y 2,7% así como un tiempo de adición de 24 meses.

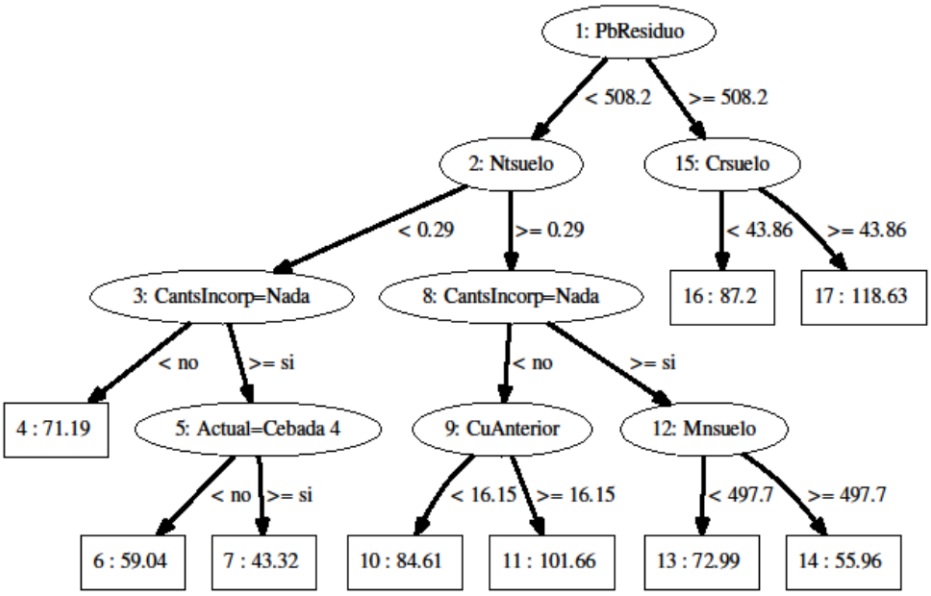


Figura 9-54 Árbol de regresión obtenido de fósforo en el suelo.

Parámetros del árbol. Profundidad máxima=4, peso mínimo=8, datos para la poda=9.

Los valores más bajos de fósforo en el suelo se dan cuando no se han añadido biosólidos, cuando el cultivo ha sido cebada 4, las cantidades de nitrógeno en el suelo son menores a 0,29 y las cantidades de plomo en el residuo son inferiores a 508,2 mg Kg^{-1}. Las mayores cantidades de fósforo aparecen relacionadas con dos metales pesados como son el plomo y el cromo, en el residuo 508,2 mg Kg^{-1}, el primero, y en el suelo 43,86 mg Kg^{-1}, el segundo.

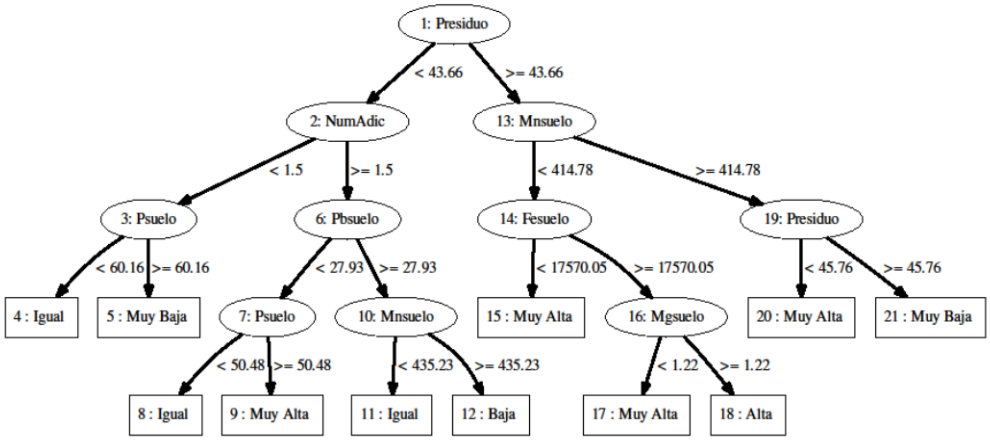


Figura 9-55 Árbol de regresión obtenido de productividad de los cultivos.

Parámetros del árbol. Profundidad máxima=4, peso mínimo=8, datos para la poda=3. Instancias clasificadas correctamente = 0,6215.

La productividad está directamente relacionada con la cantidad de nutrientes incorporados en el compost principalmente por el contenido de fósforo del residuo, el número de adiciones, con metales como el plomo o el hierro y manganeso.

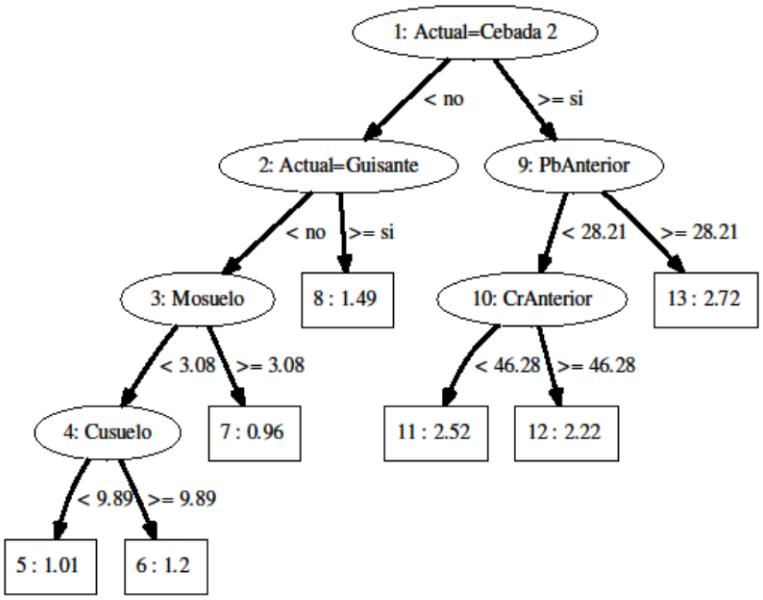


Figura 9-56 Árbol de regresión obtenido de cadmio en el suelo.

Parámetros del árbol. Profundidad máxima=4, peso mínimo=8, datos para la poda=7.

Siendo este árbol el que mejor coeficiente de correlación de los parámetros del suelo (r= 0,9037) de los estudiados en cuanto a la presencia de metales se observa cómo el cultivo de la cebada, por segunda vez en la experiencia realizada, tiene una gran influencia en la cantidad de cadmio en el suelo.

Cuando se cultivó la cebada por segunda vez tuvo lugar la adición de enmienda orgánica en el suelo. Los valores más altos se dan con el cultivo de cebada 2 y con cantidades de plomo, anterior en el suelo, superiores a 28,21 mg Kg^{-1}. El valor más bajo se observa cuando no se ha cultivado cebada 2 ni guisante y con un valor de materia orgánica en el suelo de 3,08 %. Algunas de las variables como la materia orgánica, citada anteriormente y cromo anterior tienen efectos antagónicos sobre las cantidades de cadmio presentes en el suelo.

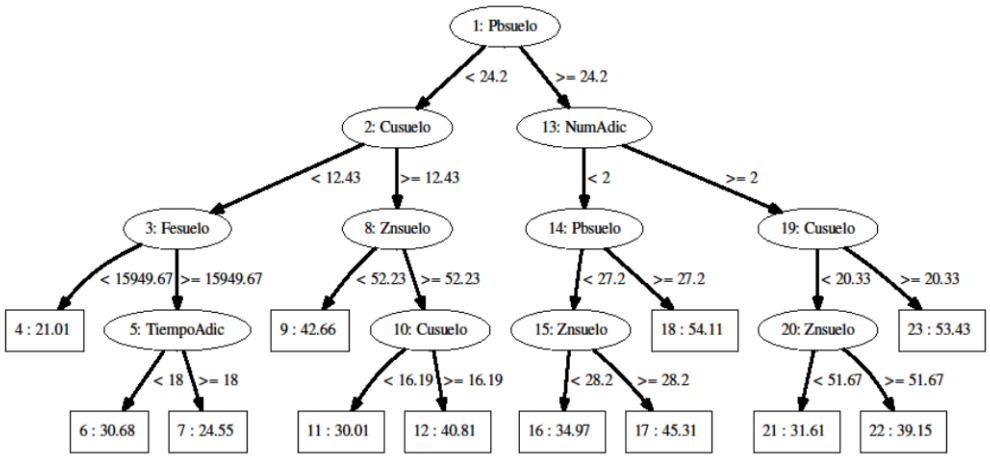


Figura 9-57 Árbol de regresión obtenido de cromo en el suelo.

Parámetros del árbol. Profundidad máxima=4, peso mínimo=8, datos para la poda=9.

En este árbol podemos ver las relaciones entre las cantidades de cromo con el plomo en el suelo, el zinc, el cobre, el hierro, el tiempo transcurrido de la adición y el número de las mismas. Todas las variables dependientes de los metales tienen una influencia positiva con la cantidad de cromo en el suelo excepto el tiempo de adición y el número de estas últimas que tienen correlación negativa con la cantidad de cromo en el suelo. El valor más elevado se da con valores de plomo en el suelo de 27,2 mg Kg^{-1} y con una adición de enmienda orgánica en el suelo. Con cantidades de hierro, cobre y plomo inferiores a 15949,67 mg Kg-1, 12,43 mg Kg^{-1} y 24,2 mg Kg^{-1}, respectivamente, hallamos el valor más bajo de cromo en el suelo (21,01 mg Kg^{-1}).

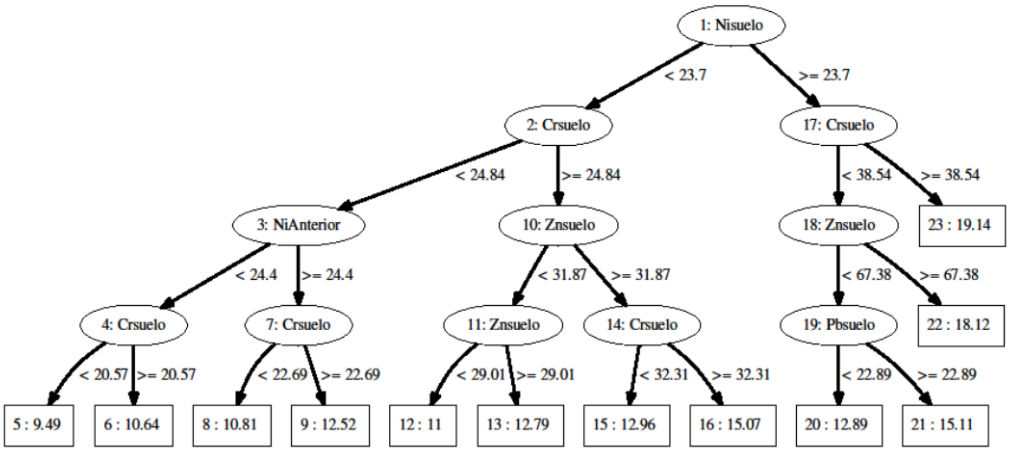


Figura 9-58 Árbol de regresión obtenido de cobre en el suelo.

Parámetros del árbol. Profundidad máxima=4, peso mínimo=8, datos para la poda=11.

El árbol de regresión obtenido para el contenido de cobre en el suelo muestra como principales variables que participan en este modelo a las concentraciones de níquel, cromo, zinc, plomo y las cantidades de níquel presentes en el suelo en el año anterior. De todas estas variables las que determinan los valores máximos y mínimos son aquellas que dependen de las cantidades de níquel y de cromo presentes en el suelo. Se trata de variables con correlaciones positivas con el metal estudiado.

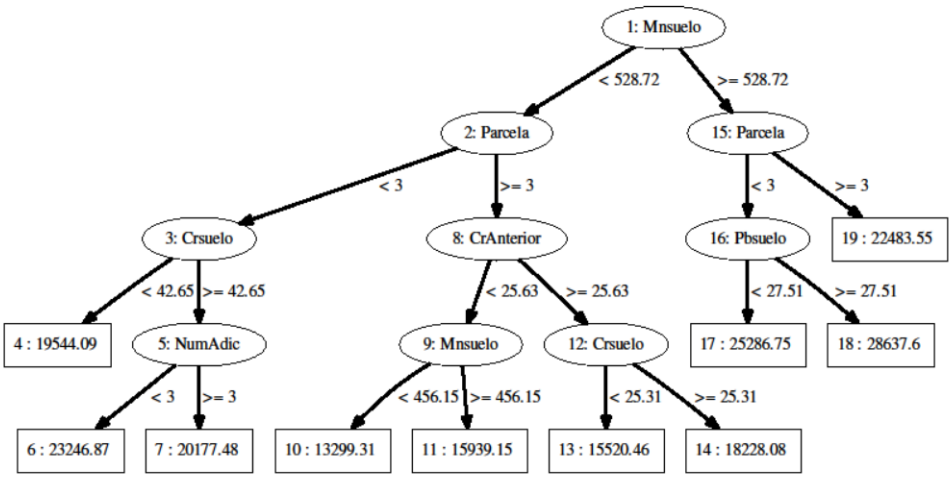


Figura 9-59 Árbol de regresión obtenido de hierro en el suelo.

Parámetros del árbol. Profundidad máxima=4, peso mínimo=8, datos para la poda=5.

El árbol de regresión obtenido para el contenido de hierro en el suelo, muestra como principales variables que participan en este modelo a las concentraciones de Mn, plomo del suelo, la cantidad de cromo del suelo en el año y en el anterior, el número de aplicaciones de enmienda y la parcela. Se observa un claro efecto determinado por la parcela, donde las parcelas 3,4 o 5 tienen

menores cantidades de hierro en sus suelos, lo cual se comprueba en ambas ramas del árbol tanto a la izquierda como a la derecha del valor del manganeso.

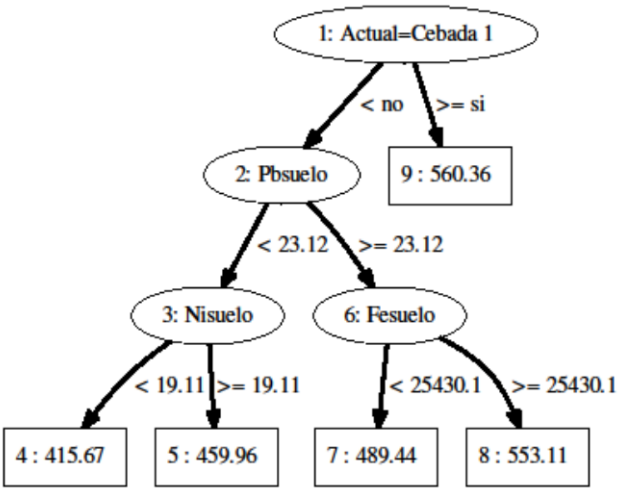


Figura 9-60 Árbol de regresión obtenido de manganeso en el suelo.

Parámetros del árbol. Profundidad máxima=4, peso mínimo=8, datos para la poda=7.

El árbol de regresión obtenido para el contenido de manganeso en el suelo (*r* = 0,7521), muestra como principales variables que participan en este modelo a las concentraciones de Fe, Ni, Pb y tener como cultivo actual cebada 1. Cuando se cultivó cebada el primer año se dan los mayores valores de manganeso en el suelo, por otro lado cuando no ha sido así y los valores de plomo y níquel en suelo son menores a 23,12 y 19,11, respectivamente el valor de manganeso en suelo baja hasta 415,67 mg Kg^{-1}.

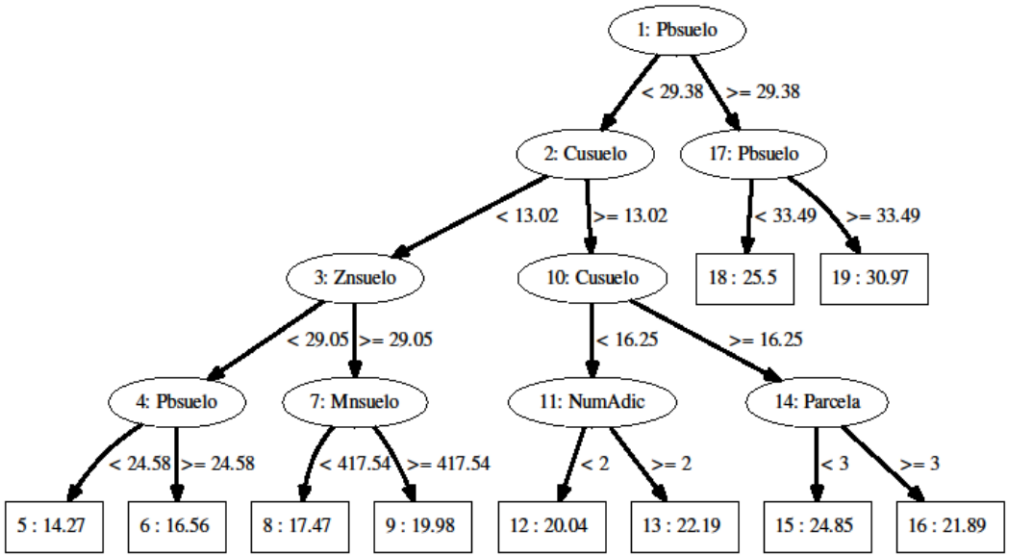


Figura 9-61 Árbol de regresión obtenido de níquel en el suelo.

Parámetros del árbol. Profundidad máxima=4, peso mínimo=8, datos para la poda=5.

En las medidas de níquel en el suelo las variables influyentes han sido el plomo, el cobre, el zinc y el manganeso presentes en el suelo y el número de adiciones y la parcela de cultivo. De nuevo se observa diferencias entre las parcelas 1 y 2 y el resto (3,4 y 5) dando mayores valores las primeras. Cuando la cantidad de plomo en el suelo es superior a 33.49 se encuentra el mayor valor de níquel, por otro lado cuando es inferior a 24.58 y los valores de zinc son menores de 29,05 mg Kg^{-1} y de cobre menor a 13,02, ambos en el suelo, la cantidad presente en el suelo de níquel disminuye hasta 14.27 mg Kg^{-1}.

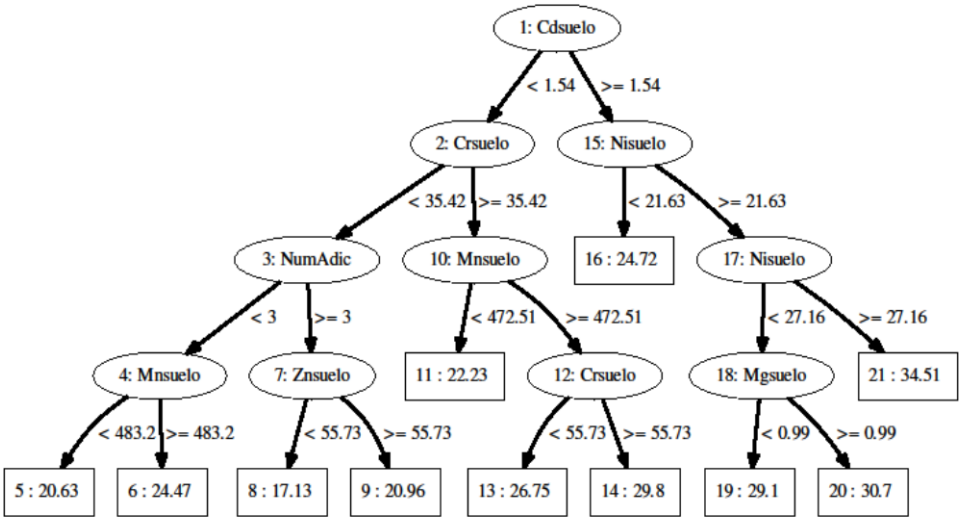


Figura 9-62 Árbol de regresión obtenido de plomo en el suelo.

Parámetros del árbol. Profundidad máxima=4, peso mínimo=8, datos para la poda=9.

Se observa un árbol con una repetida influencia de la cantidad de cromo, níquel y manganeso en el suelo y también de cadmio, magnesio y el número de adiciones y donde los valores más altos de plomo se encuentran con cantidades superiores de níquel en el suelo a 27,16 mg Kg^{-1} junto con valores de cadmio de 1,54 mg Kg^{-1} o superiores. Cuando el número de adiciones es tres y el zinc en el suelo es menor a 55,73 mg Kg^{-1}, además de cromo y cadmio con cantidades inferiores a 35,42 y 1,54 mg Kg^{-1}, respectivamente, se obtiene el valor más bajo.

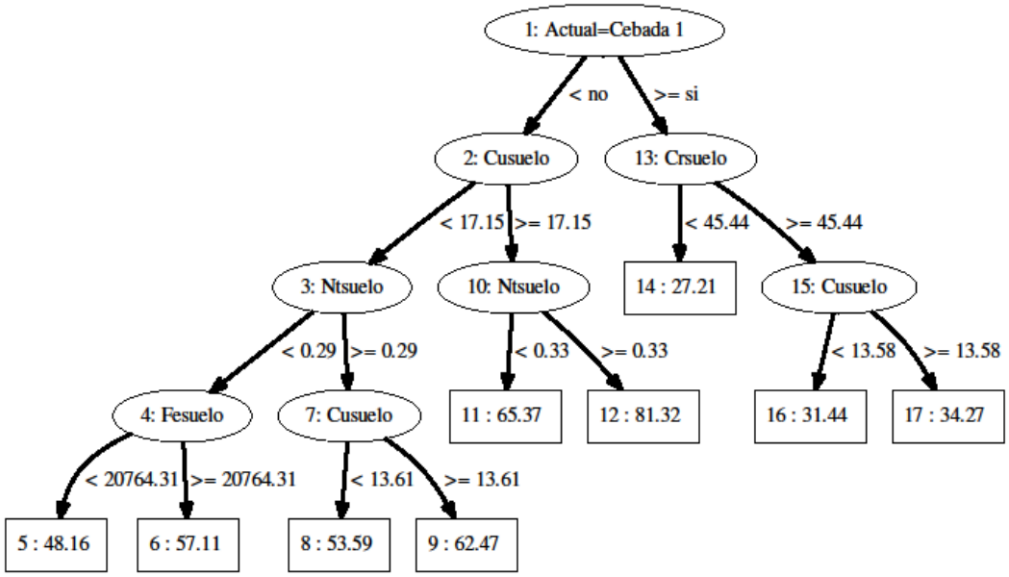


Figura 9-63 Árbol de regresión obtenido de zinc en el suelo.

Parámetros del árbol. Profundidad máxima=4, peso mínimo=8, datos para la poda=7.

Las variables que describen este árbol son el cultivo actual siendo cebada 1, la cantidad de cromo, cobre, nitrógeno y hierro en el suelo. Se observan los mayores valores de zinc en el suelo cuando el cultivo no es la cebada de primer año, que es el primer año de adición de enmiendas, y los menores cuando las cantidades de cromo en el suelo son inferiores a 45,44 mg Kg^{-1}. Todas las variables tienen correlaciones positivas con la cantidad de zinc en el suelo.

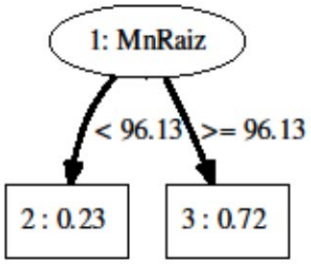


Figura 9-64 Árbol de regresión obtenido de cadmio en la raíz.

Parámetros del árbol. Profundidad máxima=4, peso mínimo=8, datos para la poda=3.

La cantidad de cadmio en la raíz se ve influida la cantidad de manganeso presente en la raíz.

Cr 4,8,11 No existe representación con la combinación propuesta de las cantidades de cromo en raíz.

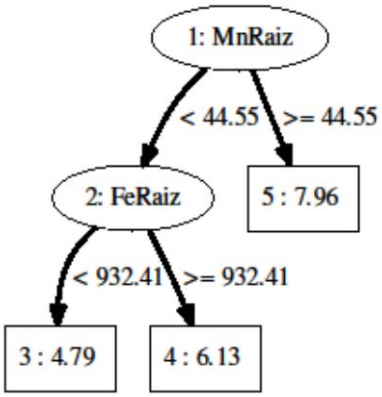


Figura 9-65 Árbol de regresión obtenido de cobre en la raíz.

Parámetros del árbol. Profundidad máxima=2, peso mínimo=8, datos para la poda=9.

El árbol de regresión obtenido para el contenido de Cu en la raíz presenta como principales variables explicativas el manganeso y el hierro en la raíz. Se

puede observar una disminución de la cantidad de cobre en la raíz cuando el manganeso es menor de 44,55 mg Kg^{-1} y el hierro está por debajo de 932,41 mg Kg^{-1}, ambos en la raíz.

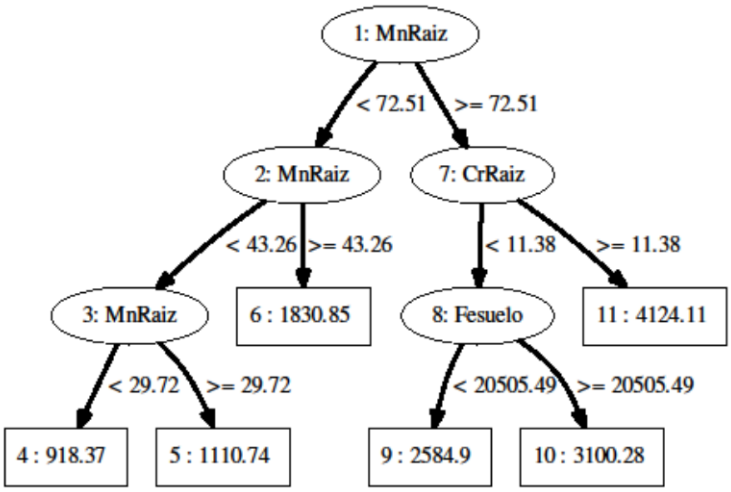


Figura 9-66 Árbol de regresión obtenido de hierro en la raíz.

Parámetros del árbol. Profundidad máxima=4, peso mínimo=8, datos para la poda=7.

Este árbol es el que segundo mejor coeficiente de correlación muestra de los estudiados en cuanto a la presencia de metales en la raíz. El hierro en la raíz tiene como variables más relevantes, en la formación del gráfico del árbol, al manganeso y el cromo en la raíz y el hierro presente en el suelo. Todas las variables tienen correlaciones positivas y los valores más bajos se dan cuando la cantidad de manganeso en la raíz es inferior a 29,72 mg Kg^{-1} y cuando el manganeso es superior a 72,51 y el cromo mayor de 11,38 mg Kg^{-1} se obtiene un valor de hierro en la raíz de 4124,11 mg Kg^{-1}.

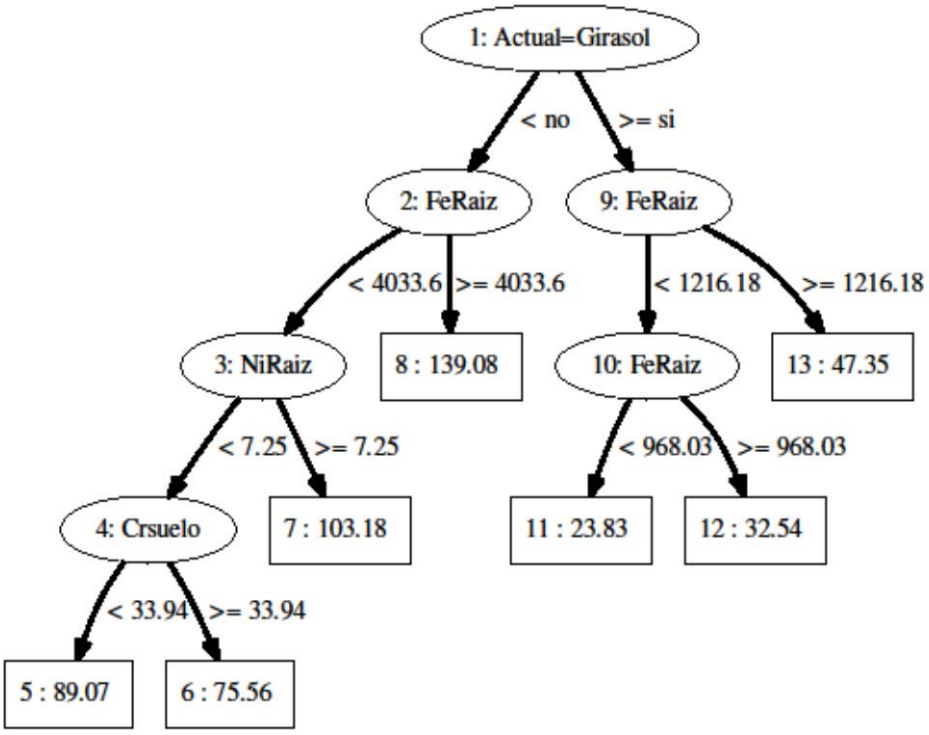


Figura 9-67 Árbol de regresión obtenido de manganeso en la raíz.

Parámetros del árbol. Profundidad máxima=4, peso mínimo=8, datos para la poda=7.

Este árbol es el que mejor coeficiente de correlación muestra de los estudiados en cuanto a la presencia de metales en la raíz. Vemos la gran relación entre el hierro y el manganeso como aparecía en la representación anterior. Las variables que aparecen en el árbol son el hierro y níquel en la raíz, el cromo en el suelo y el cultivo de girasol en el año. Se observa que los valores más bajos de manganeso en la raíz se dan cuando los valores de hierro en la raíz son menores 968,03 mg Kg^{-1} y el cultivo ha sido girasol. La mayor cantidad de manganeso aparece con valores de hierro en raíz de 4033,6 mg Kg^{-1} y cuando no se ha cultivado girasol.

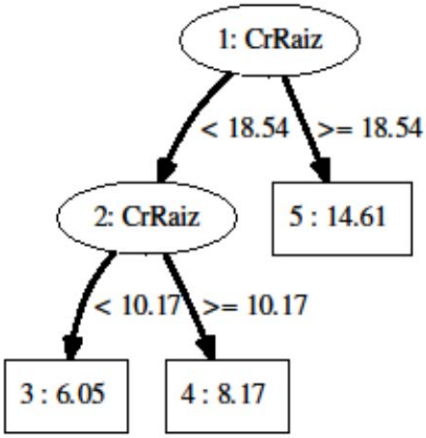


Figura 9-68 Árbol de regresión obtenido de níquel en la raíz.

Parámetros del árbol. Profundidad máxima=4, peso mínimo=8, datos para la poda=5.

El níquel en la raíz aparece correlacionado con el cromo en la raíz. La rama de mayores valores de níquel se presenta con valores superiores a 18,54 mg Kg^{-1} de cromo en la raíz. El menor valor de níquel 6,05 mg Kg^{-1} se da con valores inferiores a 10,17 mg Kg^{-1} de cromo en la raíz.

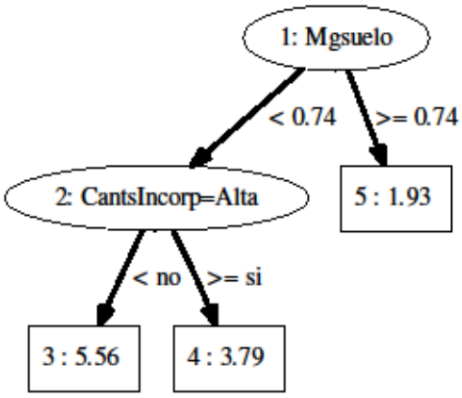


Figura 9-69 Árbol de regresión obtenido de plomo en la raíz.

Parámetros del árbol. Profundidad máxima=4, peso mínimo=8, datos para la poda=3.

El plomo presente en la raíz está relacionado con dos variables que son el magnesio en el suelo y la incorporación de enmiendas en cantidades altas. Los valores más bajos de plomo en la raíz se obtienen cuando el magnesio en el suelo es superior a 0,74 mg Kg^{-1}. Los valores más altos, 5,56 mg Kg^{-1}, tienen como factor principal la disminución del magnesio en el suelo y cuando no se han incorporado cantidades altas de enmienda.

Zn 4.8.11

No hay representación gráfica para este metal en la raíz.

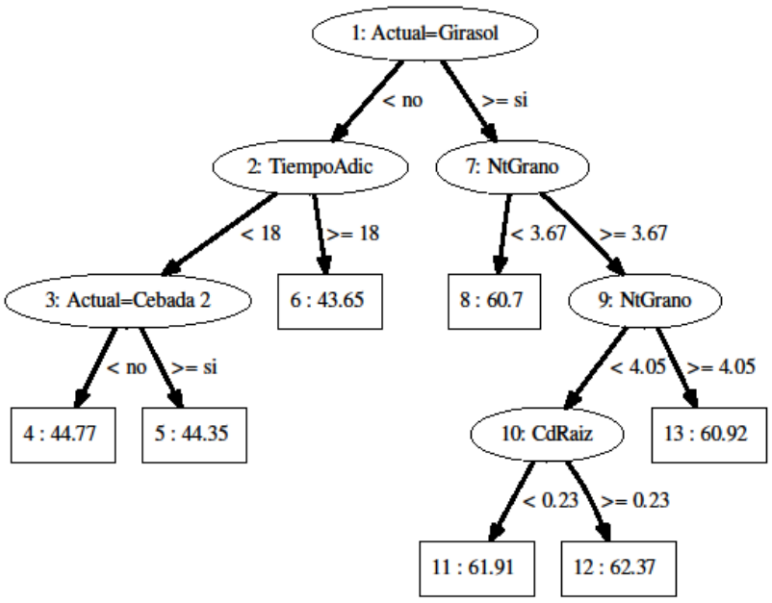


Figura 9-70 Árbol de regresión obtenido de carbono en el grano.

Parámetros del árbol. Profundidad máxima=4, peso mínimo=8, datos para la poda=9.

El cultivo de girasol da mayor valor de carbono en el grano y mientras que la cebada un menor valor de carbono en grano, por otro lado un aumento del rango del nitrógeno entre 3,67- 4,05 en el grano también permite algún incremento del carbono. Cuando el tiempo de adición es de 24 meses hay menos carbono en el grano.

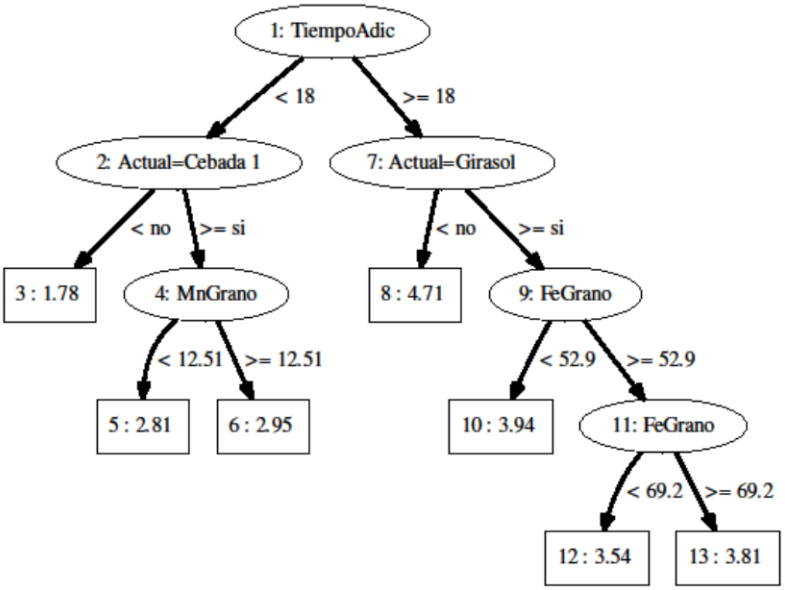


Figura 9-71 Árbol de regresión obtenido de nitrógeno en el grano.

Parámetros del árbol. Profundidad máxima=4, peso mínimo=8, datos para la poda=11.

El árbol de regresión obtenido para el contenido de nitrógeno total del grano presenta como principales variables explicativas, el tiempo desde la última adición, el cultivo de girasol y de cebada 1 en el año actual y las cantidades de manganeso y hierro presentes en el grano. Se muestra un valor mínimo de la cantidad de nitrógeno en el grano (1,78) cuando no se ha cultivado cebada 1 y han transcurrido 12 meses desde la adición. Esto corresponde con los cultivos siguientes de cebada. El cultivo de de girasol proporciona valores más altos de nitrógeno en grano cuando se tienen cantidades de hierro en el grano inferiores a 52,9 mg Kg^{-1}.

Cd Grano 1.8.3 0,3087

No hay representación.

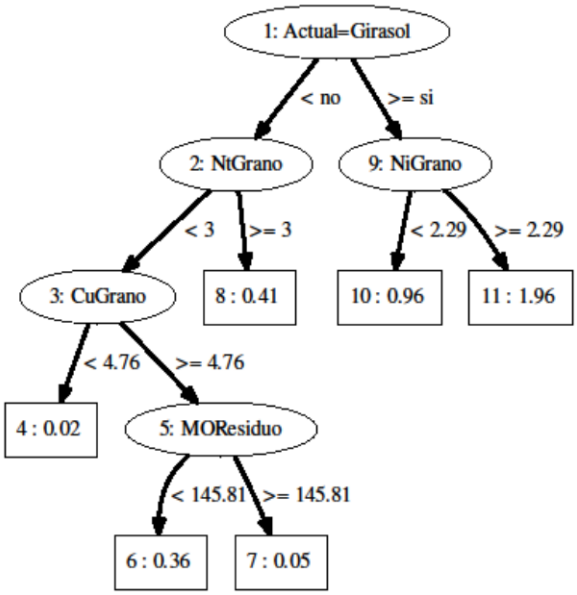


Figura 9-72 Árbol de regresión obtenido de cromo en el grano.

Parámetros del árbol. Profundidad máxima=4, peso mínimo=8, datos para la poda=11.

El cultivo de girasol y una presencia de níquel en el grano de, al menos, 2,29 mg Kg^{-1} establecen un valor máximo de cromo en el grano de 1,96 mg Kg^{-1}. Cuando no se cultiva girasol, el nitrógeno en el grano es inferior a 3 % y el cobre es menor de 4,76 mg Kg^{-1} se obtiene un valor mínimo de cromo en el grano, con valor de 0,02 mg Kg^{-1}.

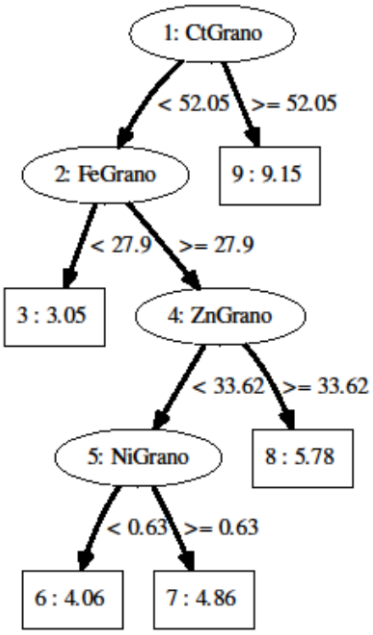


Figura 9-73 Árbol de regresión obtenido de cobre en el grano.

Parámetros del árbol. Profundidad máxima=4, peso mínimo=8, datos para la poda=7.

La cantidad de cobre presente en el grano establece correlaciones con distintos metales como son el zinc, el níquel y el hierro, así como también con la cantidad de carbono total presente en el grano. Se observa que el aumento del carbono total aumenta el cobre presente en el grano. Una menor cantidad de carbono, de 52,05 y de hierro de 27,9 mg Kg^{-1} nos ofrece la menor cantidad de cobre en grano con un valor 3,05 mg Kg^{-1}.

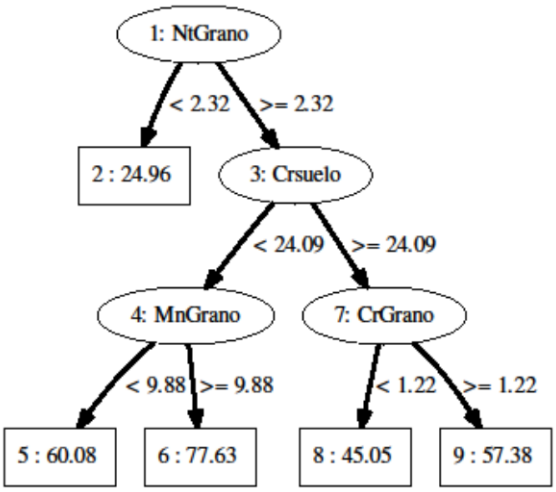


Figura 9-74 Árbol de regresión obtenido de hierro en el grano.

Parámetros del árbol. Profundidad máxima=4, peso mínimo=8, datos para la poda=5.

La cantidad de hierro presente en el grano establece correlaciones con el nitrógeno, el cromo y el manganeso de grano y el cromo presente en el suelo. Se muestra un efecto antagónico con el cromo en el suelo. Los mayores valores se dan con manganeso en grano con valores de 9,88, con cromo en suelo con cantidades inferiores a 24,09 mg Kg^{-1} y con nitrógeno en grano con valores de 2,32 % o superiores. Cuando este último es inferior a 2,32 aparecen las menores cantidades de hierro en el grano (con un valor de 24.96 mg Kg^{-1}).

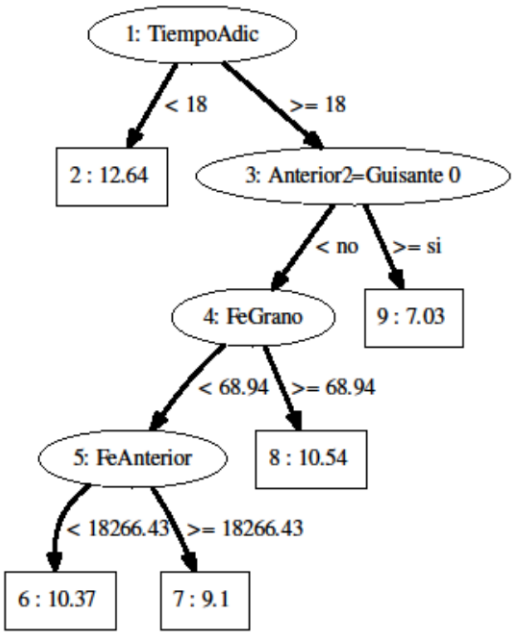


Figura 9-75 Árbol de regresión obtenido de manganeso en el grano.

Parámetros del árbol. Profundidad máxima=4, peso mínimo=8, datos para la poda=9.

El árbol de regresión obtenido para el manganeso en grano tiene una dependencia de variables como son el hierro en grano y en el suelo, el tiempo de adición y el cultivo inicial de guisante. El valor más alto de manganeso (12,64 mg Kg^{-1}) está determinado por el tiempo de adición de 12 meses. La presencia de hierro en el suelo muestra una disminución de la cantidad de manganeso en grano.

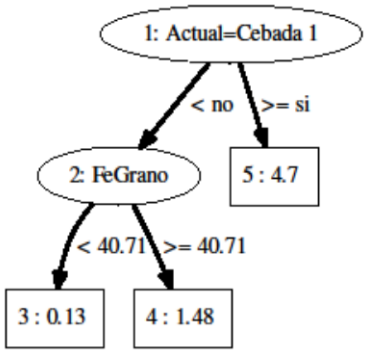


Figura 9-76 Árbol de regresión obtenido de níquel en el grano.

Parámetros del árbol. Profundidad máxima=4, peso mínimo=32, datos para la poda=11.

Las variables presentes en la representación de níquel en grano son el hierro en grano y el cultivo de cebada el primer año. Cultivando por primera vez, en la experiencia, la cebada se obtienen los valores más elevados de níquel en

grano, mientras que con otros cultivos y con cantidades de hierro en grano por debajo de 40,71 mg Kg^{-1} obtenemos los valores más bajos de níquel en grano (0,13 mg Kg^{-1}).

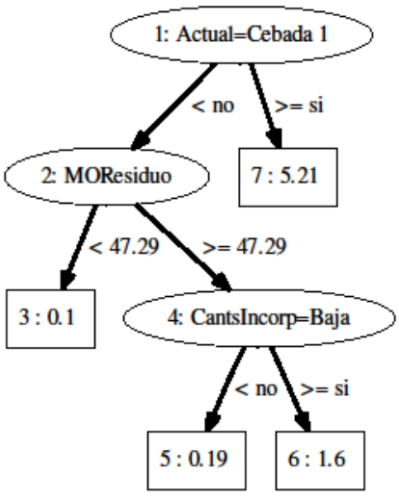


Figura 9-77 Árbol de regresión obtenido de plomo en el grano.

Parámetros del árbol. Profundidad máxima=4, peso mínimo=8, datos para la poda=3.

Las variables que intervienen en la representación de este árbol son las cantidades incorporadas de biosólido, las cantidades de materia orgánica del residuo y el cultivo de cebada en el año. Cuando se ha cultivado este cereal se dan los mayores valores de plomo en grano (5,21 mg Kg^{-1}). Por el contrario cuando no es así y la cantidad de materia orgánica del residuo es menor a 47,29 mg Kg^{-1} se obtienen los menores valores.

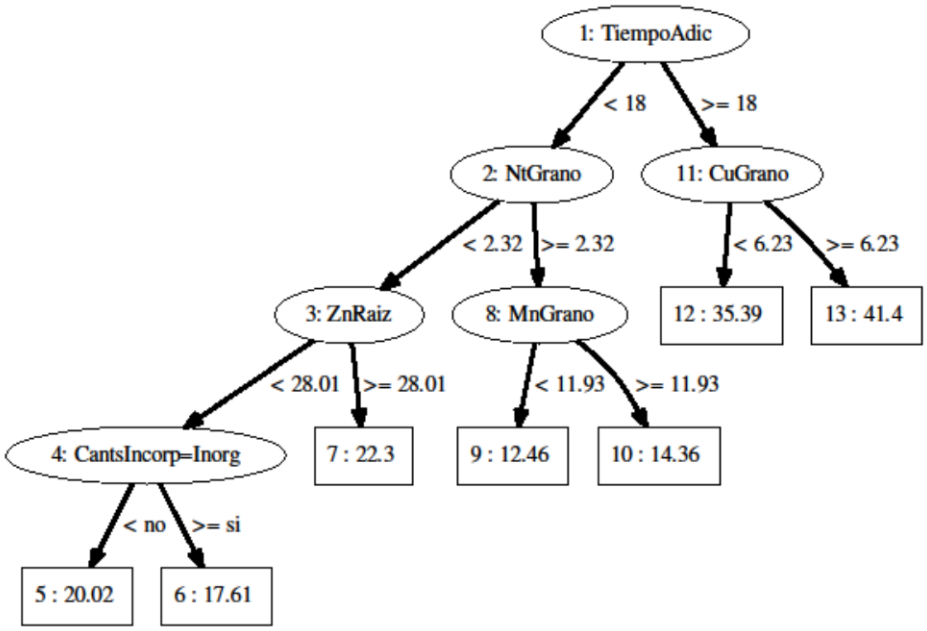


Figura 9-78 Árbol de regresión obtenido de zinc en el grano.

Parámetros del árbol. Profundidad máxima=4, peso mínimo=8, datos para la poda=11.

El árbol de regresión obtenido para el zinc en grano muestra una dependencia de variables como son el tiempo de adición, las cantidades cobre, nitrógeno y manganeso en el grano, el zinc presente en la raíz y si ha habido fertilización inorgánica. Esta última variable muestra un efecto antagonista con la cantidad de zinc presente en el grano.

Tres Fincas

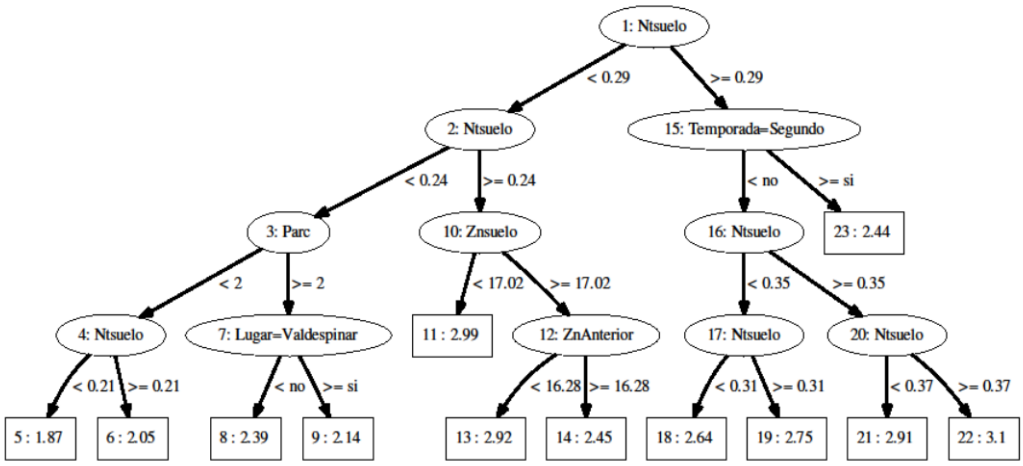


Figura 9-79 Árbol de regresión obtenido de materia orgánica en el suelo.

Parámetros del árbol. Profundidad máxima=4, peso mínimo=8, datos para la poda=9.

Entre los principales factores que se observa afectan a esta variable se encuentra la cantidad de nitrógeno en el suelo. Cuando las cantidades de nitrógeno en el suelo son superiores a 0,37 % se muestran los valores más altos de materia orgánica mientras que cuando la cantidad es inferior a 0,21% aparecen los valores más bajos de materia orgánica con valores de 1,87 %.

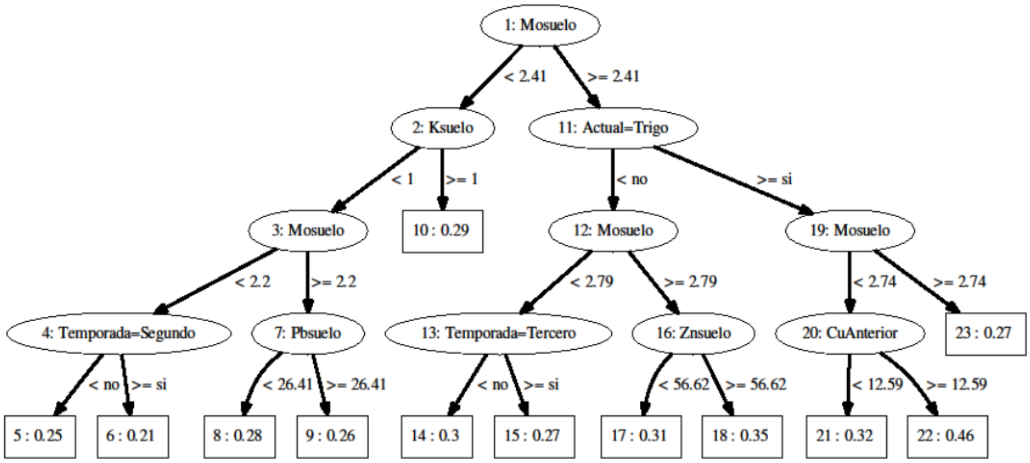


Figura 9-80 Árbol de regresión obtenido de nitrógeno en el suelo.

Parámetros del árbol. Profundidad máxima=4, peso mínimo=8, datos para la poda=5.

La principal variable que se muestra en este árbol es la cantidad de materia orgánica presente en el suelo. Se observa valores más bajos en el segundo año de experiencia y los valores más elevados (0,46 %) se ven con el cultivo

de trigo, cantidades de materia orgánica por debajo de 2,74 % y cantidades de cobre en el suelo en el año anterior de 12,59 mg Kg^{-1}.

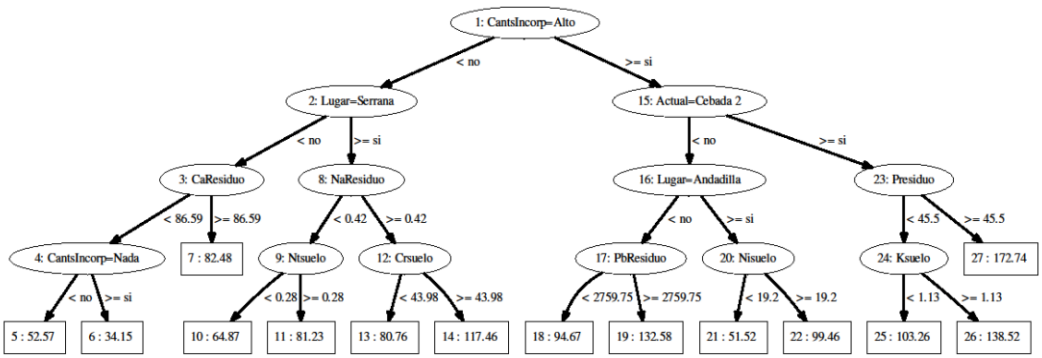


Figura 9-81 Árbol de regresión obtenido de fósforo en el suelo.

Parámetros del árbol. Profundidad máxima=4, peso mínimo=8, datos para la poda=5.

Los tratamientos de biosólidos de dosis definidas como altas y la cantidad de fósforo en el residuo muestran las mayores cantidades de fósforo en el suelo. Los menores valores se dan en el tratamiento control (cantidades incorporadas nada).

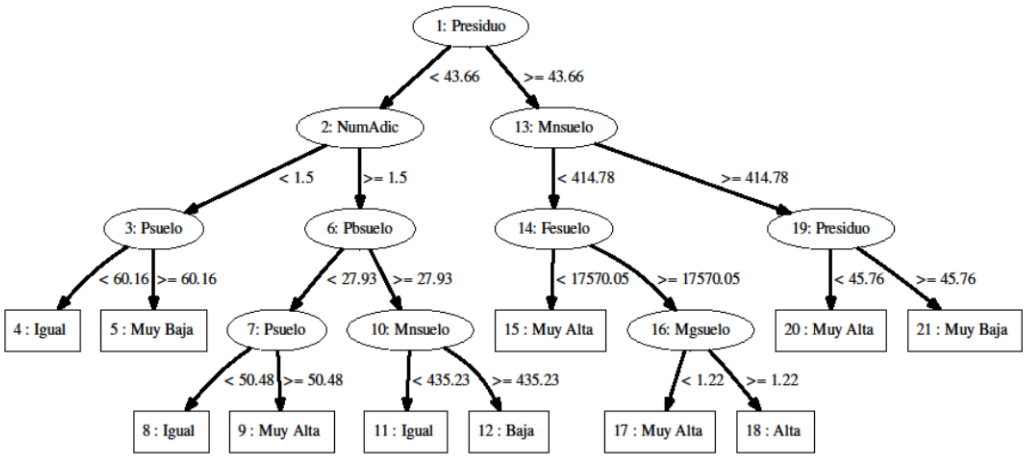


Figura 9-82 Árbol de regresión obtenido de productividad de los cultivos.

Parámetros del árbol. Profundidad máxima=4, peso mínimo=8, datos para la poda=3. Instancias clasificadas correctamente = 0,6215.

La productividad está directamente relacionada con la cantidad de nutrientes incorporados en el compost principalmente por el contenido de fósforo del residuo, el número de adiciones, con metales como el plomo o el hierro y manganeso.

Cadmio

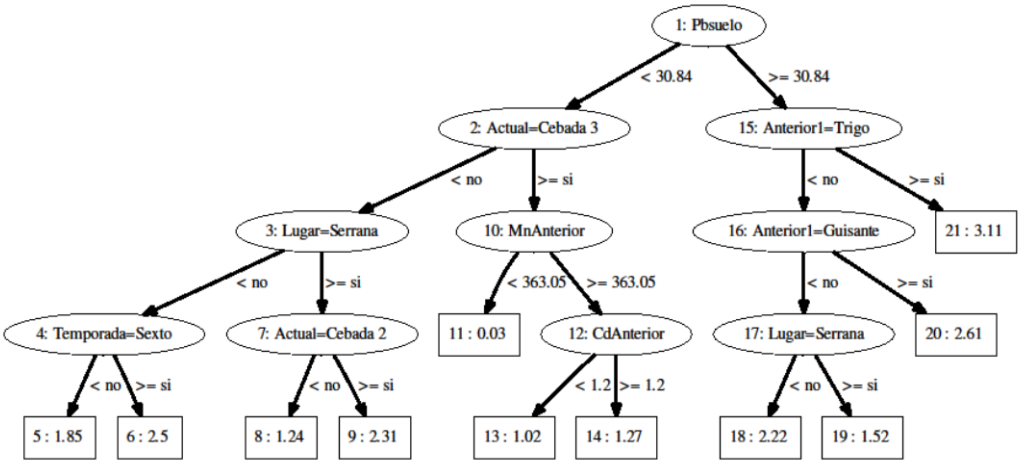


Figura 9-83 Árbol de regresión obtenido de cadmio en el suelo.

Parámetros del árbol. Profundidad máxima=4, peso mínimo=8, datos para la poda=9.

Las cantidades de cadmio en el suelo son mayores cuando el plomo en el suelo se encuentra en valores de 30,84 mg Kg^{-1} y el cultivo del año anterior resulta ser trigo. Por otro lado las menores cantidades de cadmio aparecen con el cultivo actual de cebada por tercera vez y la cantidad de manganeso en el suelo por debajo de 363,05 mg Kg^{-1}.

Cromo

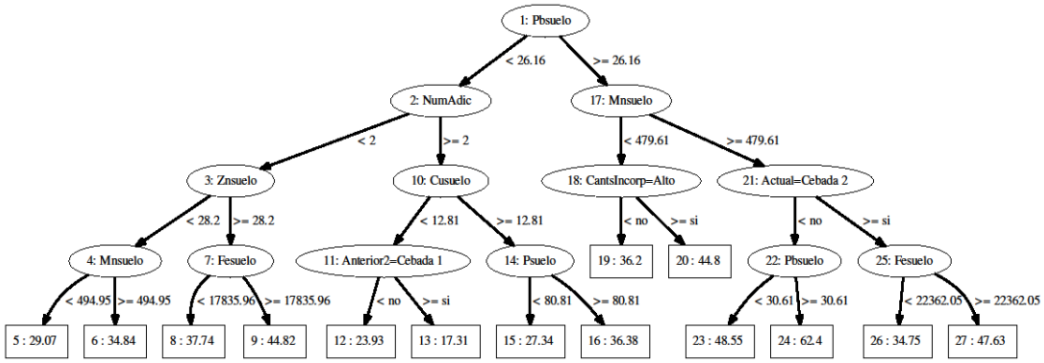


Figura 9-84 Árbol de regresión obtenido de cromo en el suelo.

Parámetros del árbol. Profundidad máxima=4, peso mínimo=8, datos para la poda=9.

Las mayores cantidades de cromo en el suelo (62,4 mg Kg^{-1}) se observan con cantidades de plomo en el suelo de 30,61 mg Kg^{-1} y las menores cantidades se encuentran ligadas al cultivo de cebada por primera vez dos años antes.

Cobre

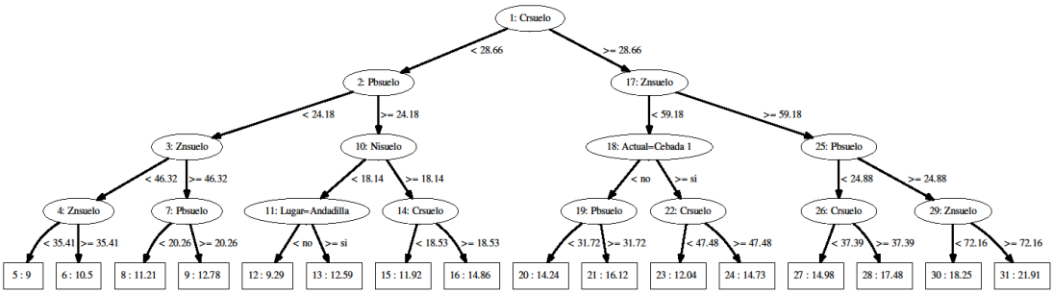


Figura 9-85 Árbol de regresión obtenido de cobre en el suelo.

Parámetros del árbol. Profundidad máxima=4, peso mínimo=8, datos para la poda=3.

En el caso de cobre en suelo se observan unas variables que determinan tanto el valor máximo como el mínimo. Junto con el cromo en el suelo como raíz del árbol las otras dos variables que ofrecen estos valores extremos son la cantidad de plomo y de zinc en el suelo.

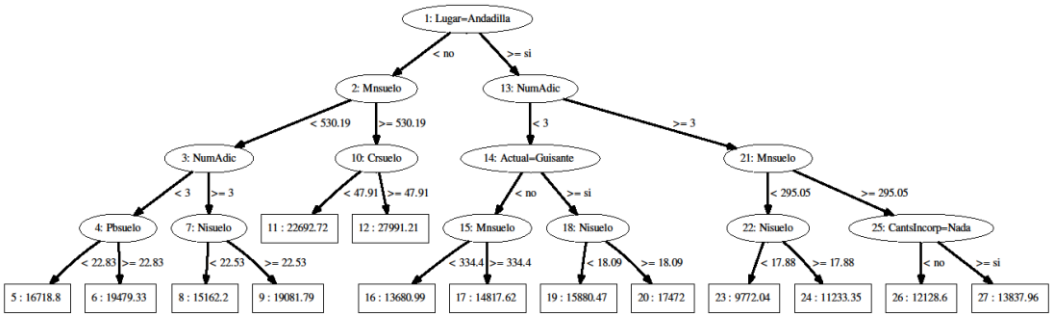


Figura 9-86 Árbol de regresión obtenido de hierro en el suelo.

Parámetros del árbol. Profundidad máxima=4, peso mínimo=8, datos para la poda=3.

El valor más bajo se da cuando la parcela se encuentra en Andadilla , el número de adiciones es menor de 3 , se ha cultivado guisante y la cantidad de níquel en el suelo es de 18,09 mg Kg^{-1} y el más alto cuando no es Andadilla y la cantidad de cromo en el suelo es de 47,91 mg Kg^{-1}

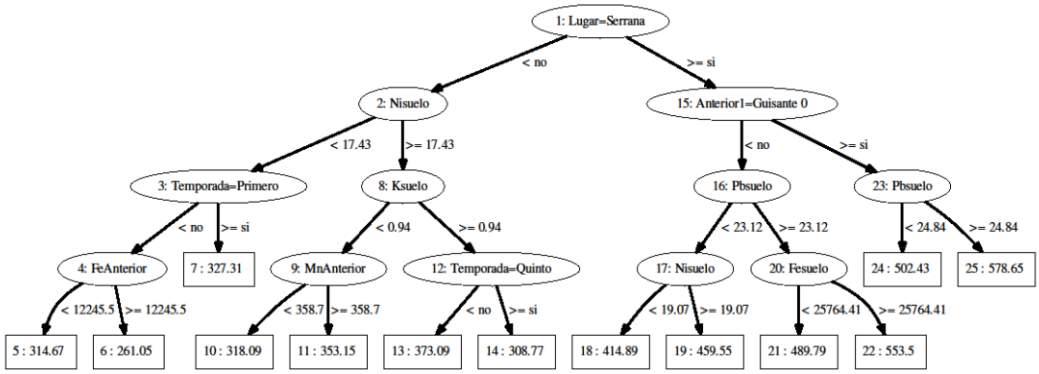


Figura 9-87 Árbol de regresión obtenido de manganeso en el suelo.

Parámetros del árbol. Profundidad máxima=4, peso mínimo=8, datos para la poda=3.

Cuando la parcela de estudio está en la Serrana y la cantidad de plomo en el suelo es de 24,84 mg Kg^{-1} se muestran los mayores valores de manganeso mientras que cuando no es ese el lugar de las parcelas y las cantidades de hierro el año anterior en el suelo son de 12245,5 mg Kg^{-1} aparecen los valores más bajos en manganeso.

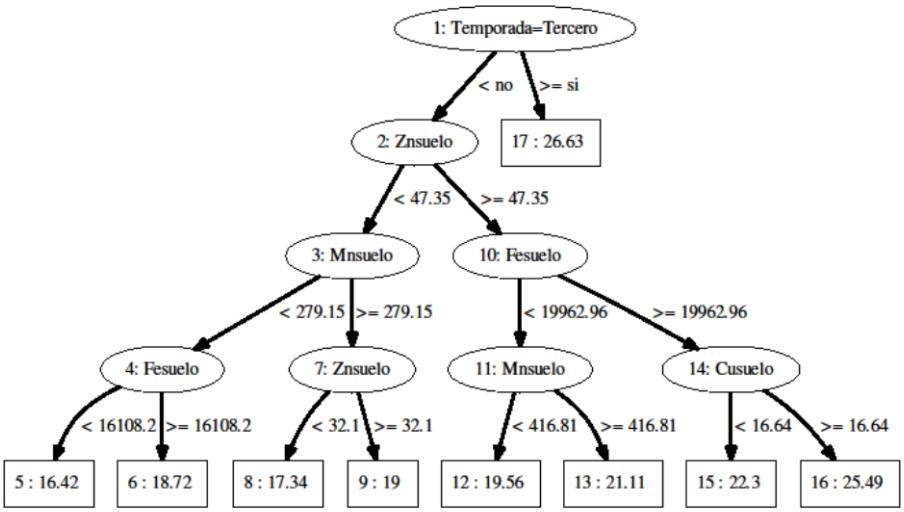


Figura 9-88 Árbol de regresión obtenido de níquel en el suelo.

Parámetros del árbol. Profundidad máxima=4, peso mínimo=8, datos para la poda=5.

Los mayores valores de níquel en el suelo se dan con el tercer año de experiencia y los menores cuando no es ese año y las cantidades hierro, manganeso y zinc en el suelo están por debajo de 16108,2, 279,15 y 47,35 mg Kg^{-1} respectivamente.

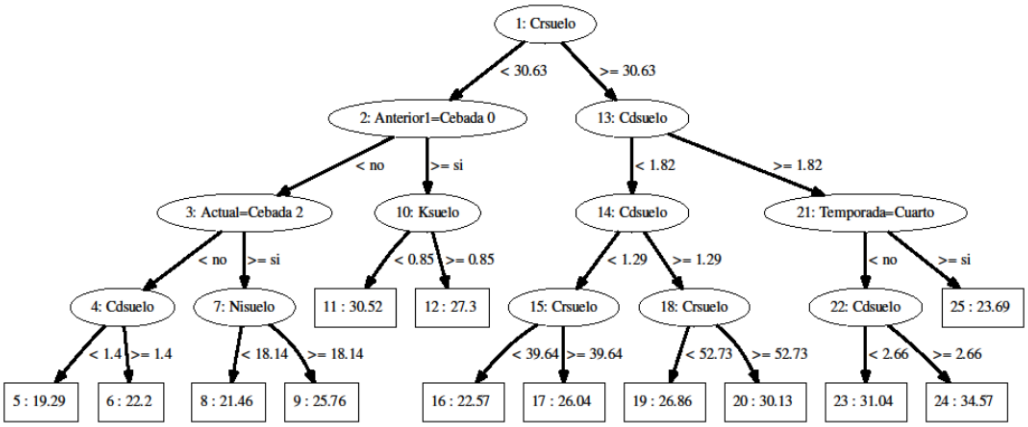


Figura 9-89 Árbol de regresión obtenido de plomo en el suelo.

Parámetros del árbol. Profundidad máxima=4, peso mínimo=8, datos para la poda=3.

El plomo en el suelo aparece ligado a la cantidad de cromo en la raíz del árbol. Para los valores más bajos tiene importancia cadmio en el suelo. Y las mayores cantidades de plomo en el suelo aparecen con cantidades de cadmio en el suelo superiores a 2,66 mg Kg^{-1}.

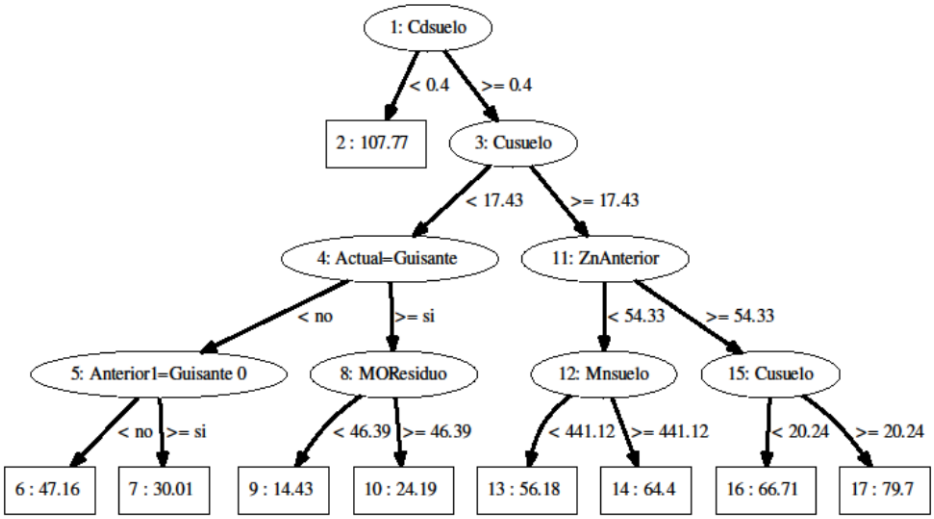


Figura 9-90 Árbol de regresión obtenido de zinc en el suelo.

Parámetros del árbol. Profundidad máxima=4, peso mínimo=8, datos para la poda=3.

Los mayores valores se dan con cantidades de cadmio en el suelo por debajo de 0,4 mg Kg^{-1} y los menores con cantidades por encima de 0,4 de cadmio y con cultivo de guisante en el año y cantidades de materia orgánica de los residuos inferiores a 46,39 %.

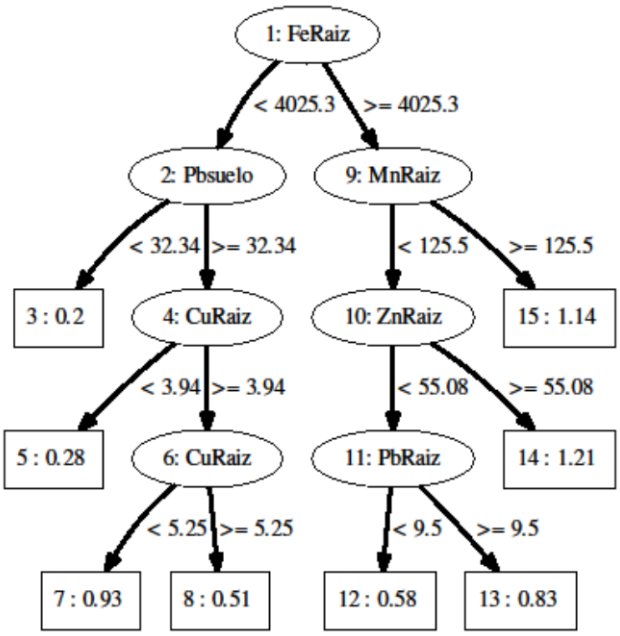


Figura 9-91 Árbol de regresión obtenido de cadmio en la raíz.

Parámetros del árbol. Profundidad máxima=4, peso mínimo=8, datos para la poda=7.

Los menores valores se encuentran definidos por variables como el hierro de la raíz y el plomo presente en el suelo, en cuanto a los valores máximos además del hierro, el manganeso y el zinc en la raíz lo definen.

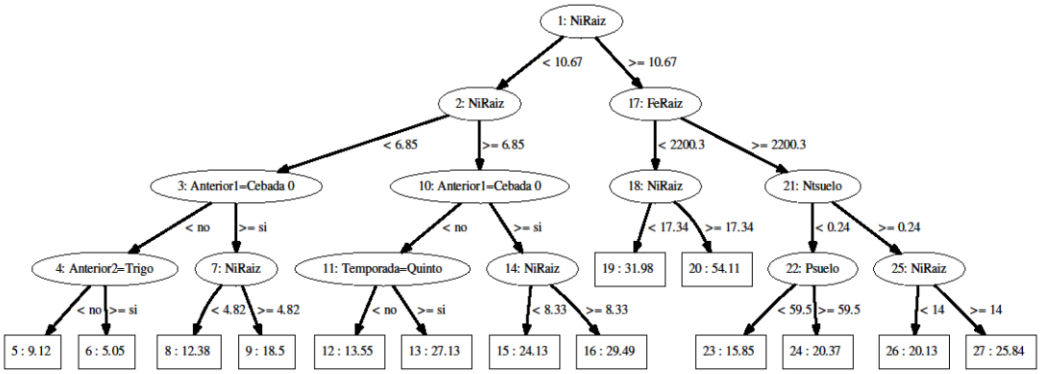


Figura 9-92 Árbol de regresión obtenido de cromo en la raíz.

Parámetros del árbol. Profundidad máxima=4, peso mínimo=8, datos para la poda=7.

El valor máximo se muestra cuando las cantidades de níquel en la raíz se encuentran por encima de 17,34 mg Kg^{-1}. Las variables que determinan los mínimos son el níquel en la raíz con valores menores a 6,85 mg Kg^{-1}.

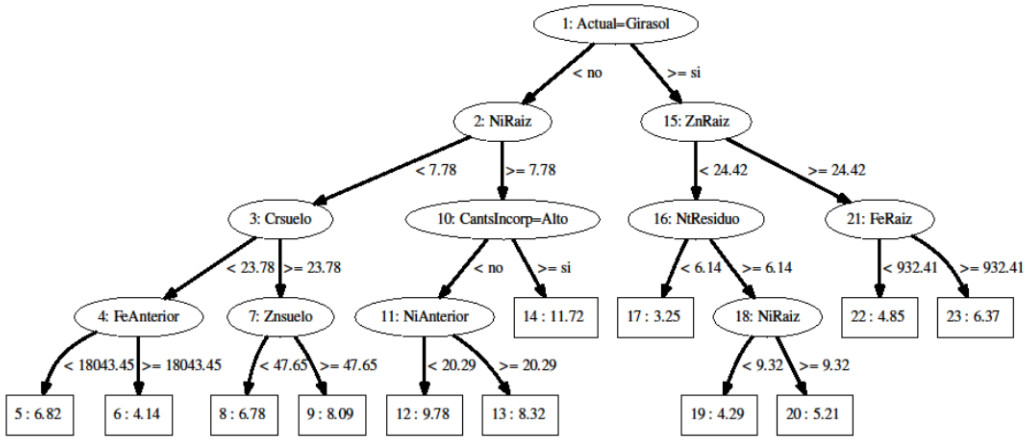


Figura 9-93 Árbol de regresión obtenido de cobre en la raíz.

Parámetros del árbol. Profundidad máxima=4, peso mínimo=8, datos para la poda=9.

Numerosas variables participan en la conformación del árbol con valores máximos y mínimos, como principal vemos el cultivo de girasol en el año presente, cuando se ha tenido este cultivo el valor mínimo se da con valores de zinc en raíz inferiores a 24,42 mg Kg^{-1} y cantidad de nitrógeno en el residuo menor a 6,14. Los valores más altos aparecen cuando no se cultiva girasol, con níquel en raíz por encima de 7,78 mg Kg^{-1} y dosis incorporadas altas.

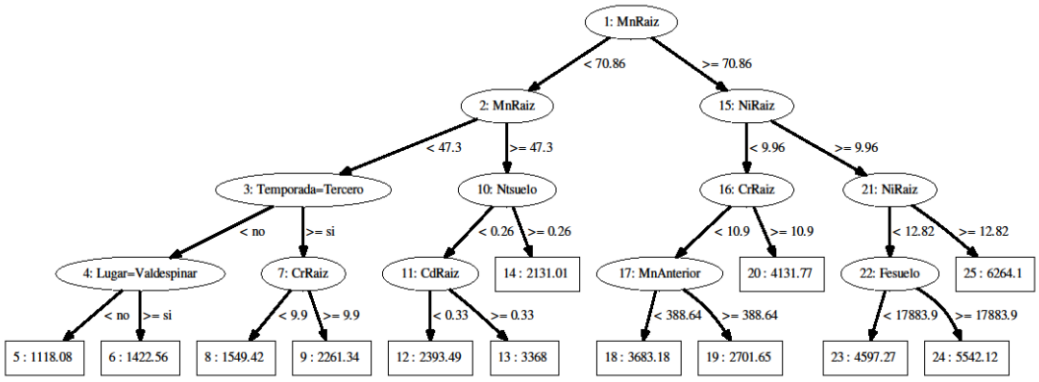


Figura 9-94 Árbol de regresión obtenido de hierro en la raíz.

Parámetros del árbol. Profundidad máxima=4, peso mínimo=8, datos para la poda=7.

La variable principal del hierro es el manganeso en raíz. Los mayores valores se encuentran relacionados con valores de níquel en raíz de 12,82 mg Kg^{-1} y los menores se dan cuando el manganeso en raíz es inferior a 47,3 mg Kg^{-1}.

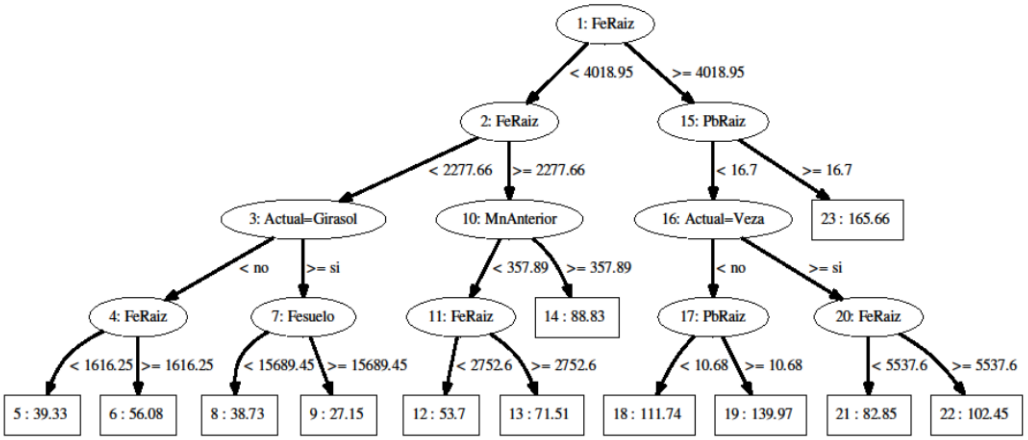


Figura 9-95 Árbol de regresión obtenido de manganeso en la raíz.

Parámetros del árbol. Profundidad máxima=4, peso mínimo=8, datos para la poda=5.

En la parte más alta del árbol vemos como la cantidad de hierro en raíz es la variable más importante, para los valores más bajos cantidades de hierro en la raíz inferiores a 15689,45 mg Kg^{-1} ofrecen las menores cantidades de manganeso. En cuanto a los mayores valores se dan cuando la cantidad de plomo en raíz es mayor a 16,7 mg Kg^{-1}.

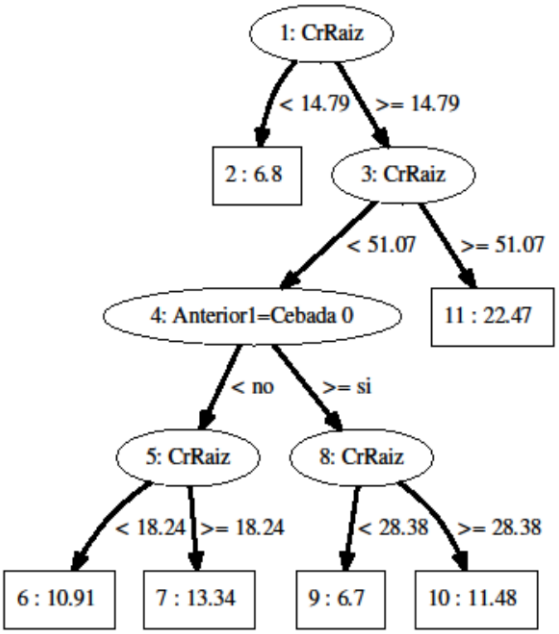


Figura 9-96 Árbol de regresión obtenido de níquel en la raíz.

Parámetros del árbol. Profundidad máxima=4, peso mínimo=8, datos para la poda=5.

El níquel aparece relacionado con el cromo en la raíz como variable principal, en el valor más bajo se muestra el cultivo anterior de cebada 0 (corresponde

con el inicio de las experiencias) como variable que lo determina junto con cantidades de cromo en la raíz inferiores a 28,38 mg Kg^{-1}. Por otro lado el valor máximo está determinado por valores de cromo en la raíz de 51,07 mg Kg^{-1}.

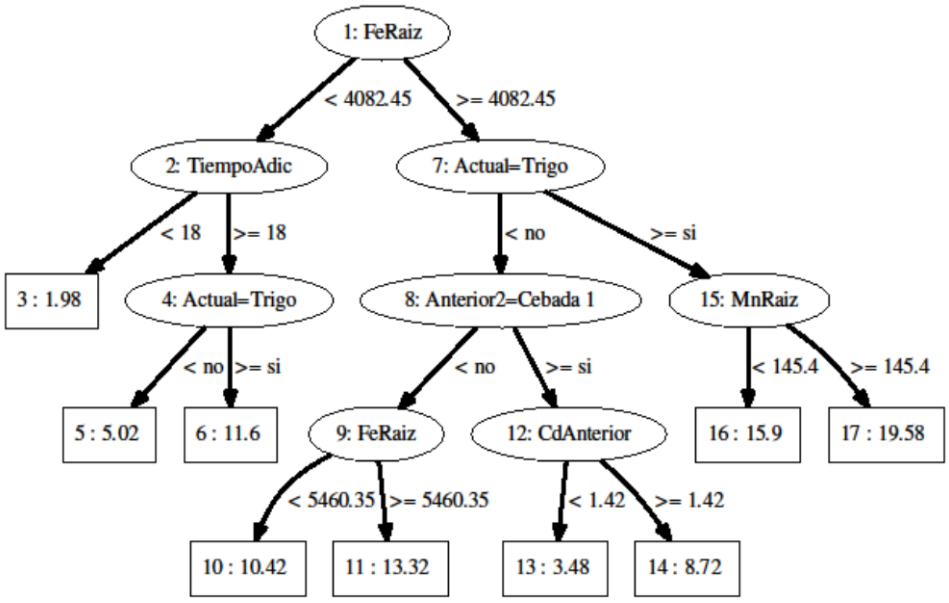


Figura 9-97 Árbol de regresión obtenido de plomo en la raíz.

Parámetros del árbol. Profundidad máxima=4, peso mínimo=8, datos para la poda=7.

Los valores máximos de este metal se dan con el cultivo del trigo, la cantidad de manganeso de 145,4 y de hierro en la raíz de 4082,45 mg Kg^{-1}. Uno de los valores más bajos aparece ligado al tiempo de adición de 12 meses.

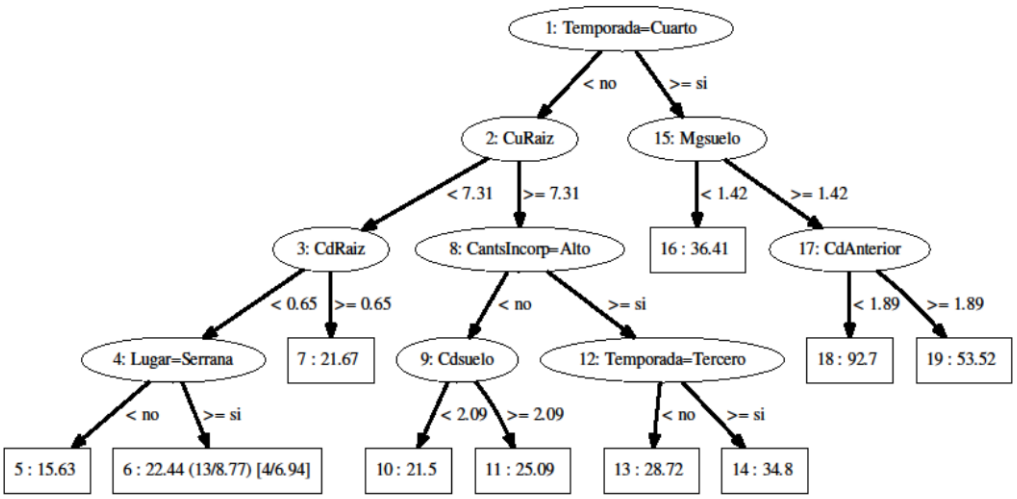


Figura 9-98 Árbol de regresión obtenido de zinc en la raíz.

Parámetros del árbol. Profundidad máxima=4, peso mínimo=8, datos para la poda=7.

Los valores más altos de zinc en raíz aparecen en la temporada cuarta del experimento, cuando la cantidad de magnesio en el suelo es de 1,42 mg Kg^{-1} y la cantidad de cadmio en el suelo el año anterior es de 1,89. Los valores más bajos se encuentran ligados a cantidades de cadmio y cobre en la raíz inferior a 0,65 y 7,31 mg Kg^{-1} respectivamente así como que no se trate de la finca de la Serrana ni sea la cuarta temporada de la experiencia.

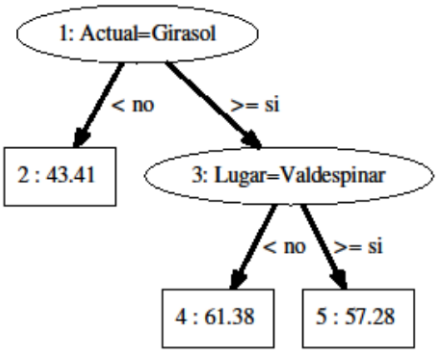


Figura 9-99 Árbol de regresión obtenido de carbono en el grano.

Parámetros del árbol. Profundidad máxima=4, peso mínimo=8, datos para la poda=3.

Se observa una cantidad menor de carbono en grano cuando el cultivo no es el girasol aunque cuando si se tiene este cultivo la parcela de Valdespinar tiene resultados más bajos que otras parcelas en cuanto a esta variable.

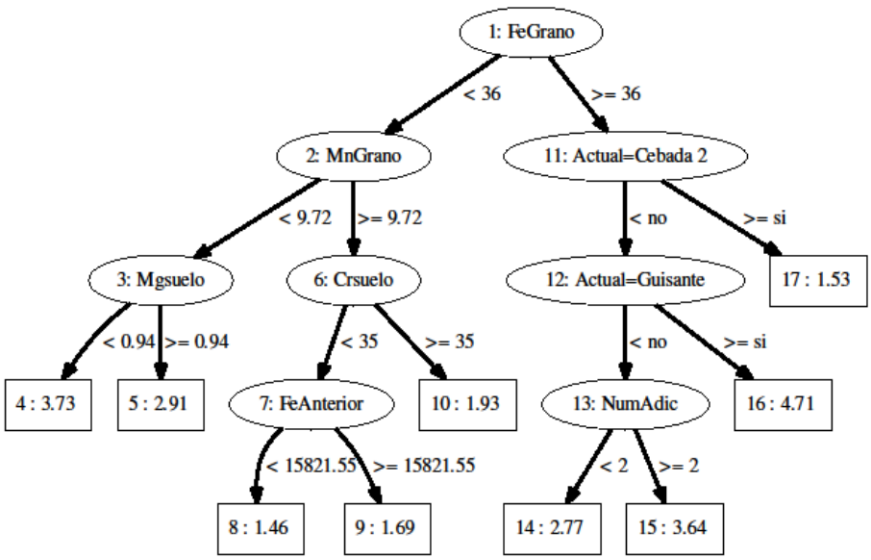


Figura 9-100 Árbol de regresión obtenido de nitrógeno en el grano.

Parámetros del árbol. Profundidad máxima=4, peso mínimo=8, datos para la poda=7.

Las mayores cantidades de nitrógeno en el grano se obtienen con el cultivo en el año del guisante y las menores se dan con cantidades de hierro inferiores a 36 y manganeso de 9,72 mg Kg^{-1} en grano y cantidades de cromo por debajo de 35 y de hierro anterior de 15821,55 mg Kg^{-1} ambas en el suelo.

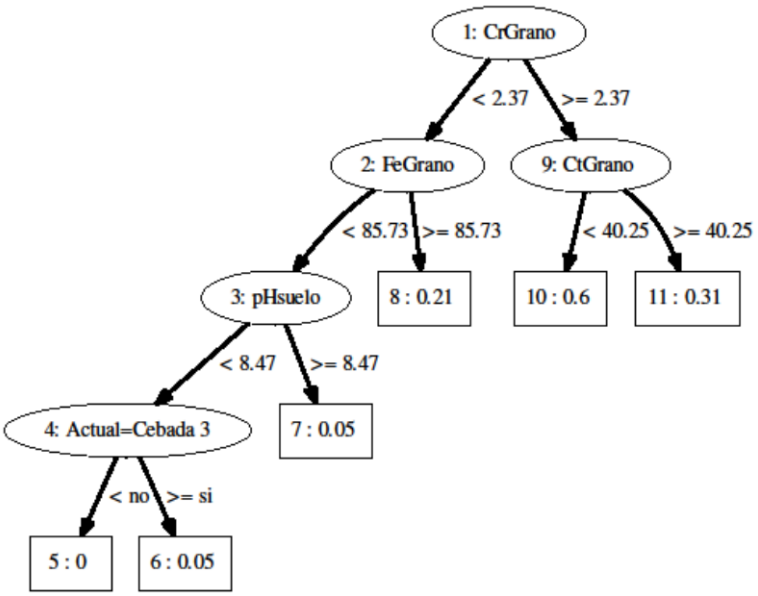


Figura 9-101 Árbol de regresión obtenido de cadmio en el grano.

Parámetros del árbol. Profundidad máxima=4, peso mínimo=8, datos para la poda=5.

El cadmio en el grano aparece para sus valores más elevados correlacionado con cantidades de cromo de 2,37 y cantidades de carbono, ambas en grano, inferiores a 40,25 mg Kg^{-1}. En cuanto a las menores cantidades de cadmio se dan cuando el cultivo actual no ha sido cebada 3, las cantidades de hierro en grano son menores a 85,73 y el cromo en grano es menor a 2,37 mg Kg^{-1}.

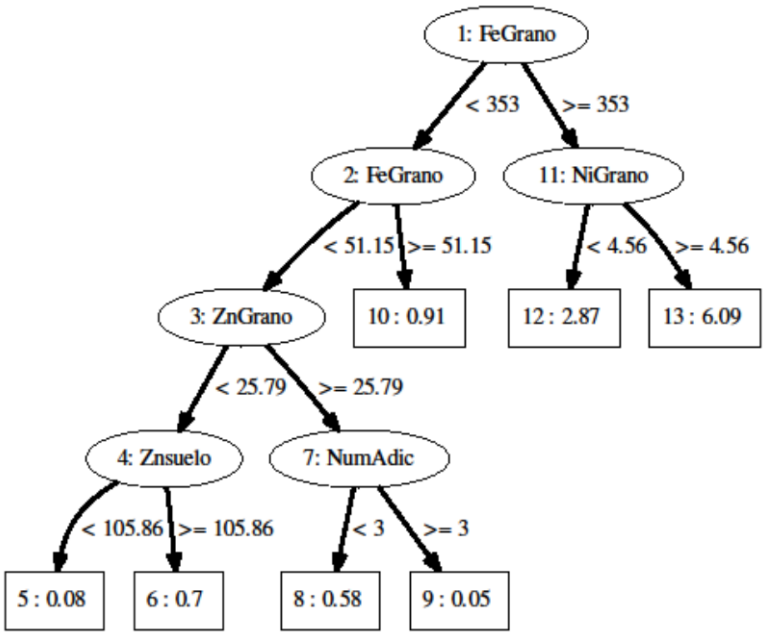


Figura 9-102 Árbol de regresión obtenido de cromo en el grano.

Parámetros del árbol. Profundidad máxima=4, peso mínimo=8, datos para la poda=7.

La mayor cantidad de cromo en grano se correlaciona con cantidades de hierro de 353 y níquel de 4,56 mg Kg^{-1} ambas en el grano, cuando las cantidades son inferiores a 51,15 y de 25,79 de hierro y zinc en grano respectivamente y el número de adiciones ha sido de 3 vemos la cantidad más baja de cromo en el grano.

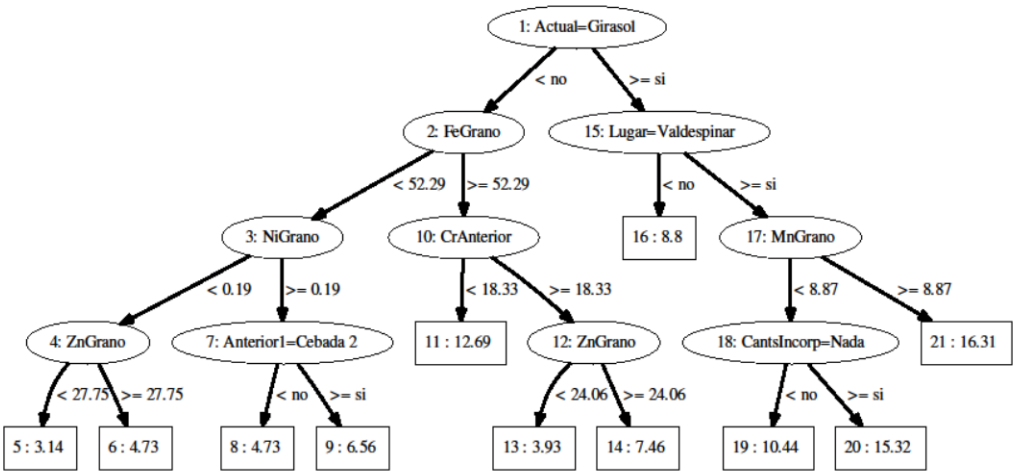


Figura 9-103 Árbol de regresión obtenido de cobre en el grano.

Parámetros del árbol. Profundidad máxima=4, peso mínimo=8, datos para la poda=7.

El cultivo de girasol en el año aparece en la raíz del árbol que representa el cobre en el grano. Cuando hay cultivo de girasol, el lugar es Valdespinar y el manganeso en el grano es de 8,87 mg Kg^{-1} tenemos el valor máximo. Si por el contrario no se ha cultivado girasol el valor inferior está determinado por el hierro, níquel y zinc en el grano.

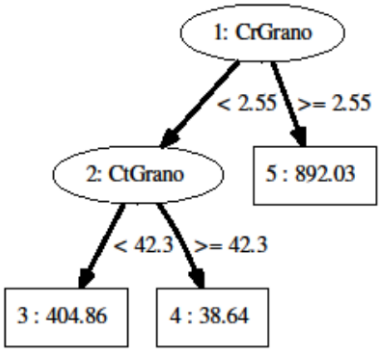


Figura 9-104 Árbol de regresión obtenido de hierro en el grano.

Parámetros del árbol. Profundidad máxima=4, peso mínimo=8, datos para la poda=5.

El hierro en grano aparece con valores máximos con cantidades de cromo en grano de 2,55 mg Kg^{-1} y mínimos con cantidades de carbono en grano de 42,3 % si las cantidades han sido inferiores de cromo en grano a lo referido previamente.

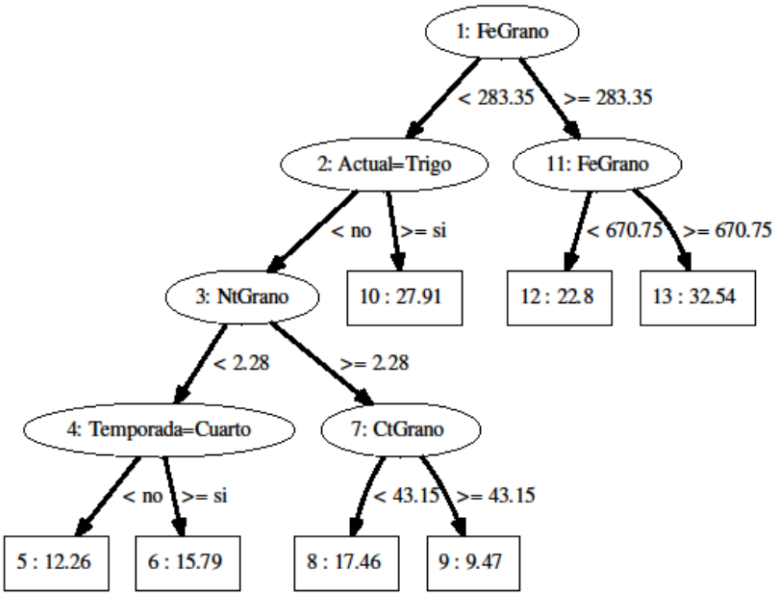


Figura 9-105 Árbol de regresión obtenido de manganeso en el grano.

Parámetros del árbol. Profundidad máxima=4, peso mínimo=8, datos para la poda=9.

El manganeso en el grano está correlacionado con el hierro en el grano Si el valor de este último es de 670,15 mg Kg^{-1} tenemos el máximo de manganeso en grano (32,54 mg Kg^{-1}). Los valores inferiores aparecen cuando no se cultiva trigo y las cantidades de nitrógeno y de carbono en el grano son de 2,28% y 43,15%, respectivamente.

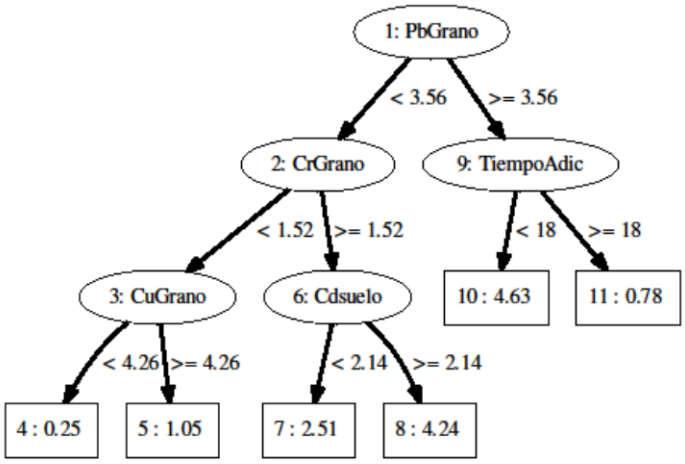


Figura 9-106- Árbol de regresión obtenido de níquel en el grano.

Parámetros del árbol. Profundidad máxima=4, peso mínimo=8, datos para la poda=5.

El níquel en el grano está correlacionado con la cantidad de plomo en el grano y para el valor máximo con tiempo de adición de biosólidos y para el mínimo con cantidades de cobre en el grano inferiores a 4,26 mg Kg^{-1}.

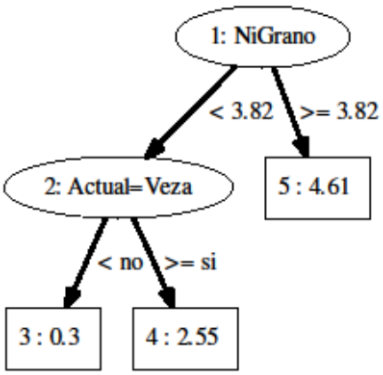


Figura 9-107 Árbol de regresión obtenido de plomo en el grano.

Parámetros del árbol. Profundidad máxima=4, peso mínimo=8, datos para la poda=9.

Las variables que conforman el árbol son el níquel en el grano y el cultivo actual de veza. El níquel en grano con valores de 3,82 mg Kg^{-1} muestra las mayores acumulaciones de plomo en el grano, mientras que cuando no se ha cultivado veza se ven las menores cantidades de plomo en grano.

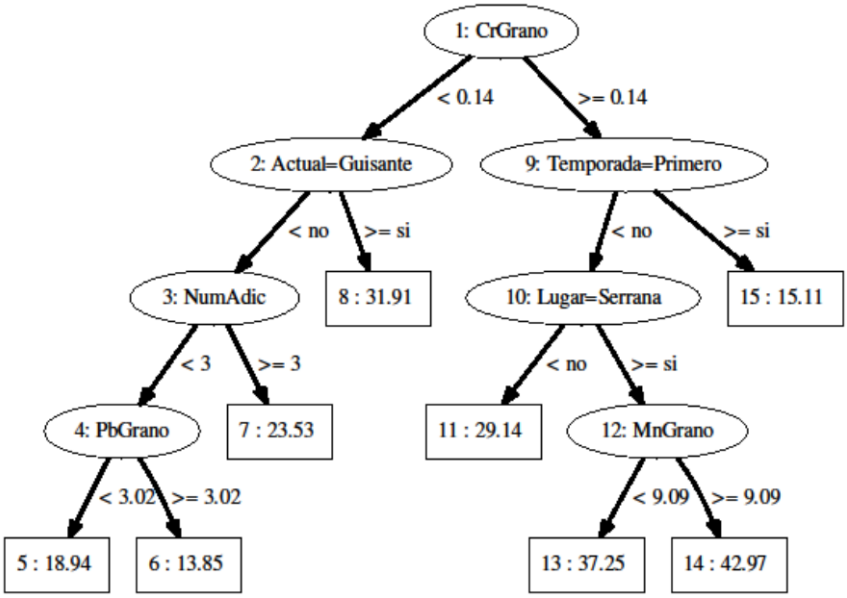


Figura 9-108 Árbol de regresión obtenido de zinc en el grano.

Parámetros del árbol. Profundidad máxima=4, peso mínimo=8, datos para la poda=11.

El zinc en el grano está correlacionado con el cromo presente en el grano, cuando las cantidades de este metal son menores de 0,14 mg Kg^{-1} y no hay cultivo actual del guisante, las adiciones son inferiores a 2 y el plomo en el grano es de 3,02 tenemos los menores valores de zinc, por otro lado si el cromo en grano es de 0,14 y no nos encontramos en la primera temporada, el lugar es Valdespinar, con valores de manganeso en grano de 9,09 mg Kg^{-1} se muestran los valores más altos de zinc en el grano.

10.Anexo II

Todos los análisis han sido realizados en el marco del convenio firmado entre la Consejería de Medio Ambiente de Castilla y León y el Grupo de Investigación en Compostaje de la Universidad de Burgos (UBUCOMP). Gracias a los datos obtenidos de esas experiencias y al trabajo de laboratorio realizado por la Doctora Susana Peña se ha podido realizar este estudio.

Los datos parciales de algunas de las fincas experimentales sirvieron para las siguientes publicaciones:

Estudio de la transferencia metálica suelo-planta mediante técnicas de Minería de Datos

ISBN: 978-84-606-9409-0

Uso de la Minería de Datos como herramienta en la predicción la calidad agronómica y el riesgo metálico de un compost de lodo EDAR.

De Residuo a Recurso: Estrategias de Gestión, Tratamiento y Valorización

ISBN: 978-84-617-2429-1

Medida del pH del suelo.

Material.

- pH-metro (Crison).

- Agitador magnético.

- Vasos de precipitados.

Procedimiento.

Colocar 10 g de suelo tamizado y seco en un vaso de precipitados de 100 ml, añadir 25 ml de agua destilada. Agitar durante 5 min. Repetir varias veces la operación y a la media hora, efectuar la medida en el pH-metro, agitando mecánicamente durante la misma.

Medida de la Conductividad eléctrica

Material:

- Conductivímetro (Crison).

- Centrífuga (Kubota, mod. Hertz)

- Botes de centrífuga

- Embudo

- Tubo

- Filtros

Procedimiento.

Pesar 10 g de muestra en un frasco de 100 ml, añadir 50 ml de agua destilada y tapar. Agitar mecánicamente durante 30 min para que se equilibre la solución. Dejar en reposo durante unos minutos, centrifugar a 10000 g y filtrar (se puede añadir una gota de hexametafosfato sódico al 0,1 %). Leer la conductividad.

Para la calibración del aparato se precisa de una solución estándar de conductividad conocida.

Cálculo de la Materia orgánica

Métodos por vía húmeda, basados en una oxidación parcial con un agente oxidante. El grado de oxidación logrado dependerá de las condiciones en que tenga lugar la reacción, con aporte de calor, o sin él. Este hecho hace necesaria la utilización de un factor estadístico, que correlacione el C oxidable determinado con la técnica de oxidación seguida y el C oxidable por vía seca. Este factor no tiene un carácter universal.

El método por vía húmeda se fundamenta en las siguientes reacciones:

a) Reducción del Cr^{6+}

$$Cr_2O_7^{2-} + 14\,H^+ + 12\,e^- \rightarrow 2\,Cr^{3+} + 7\,H_2O$$

b) Oxidación de la materia orgánica.

Se considera que la materia orgánica del suelo se comporta como un hidrato de carbono (glucosa):

$$C_6H_{12}O_6 + 6\,O_2 \rightarrow 6\,CO_2 + 6\,H_2O$$

c) Valoración del exceso de oxidante con una sal ferrosa.

$$Fe^{2+} \rightarrow Fe^{3+}$$

Este método es aplicable a tomas de muestra que contengan menos de 20 mg de C en valor absoluto. La rentabilidad de los resultados es dudosa si el suelo contiene un porcentaje de Materia Orgánica superior al 15 %. La mínima cantidad de muestra pulverizada a tomar es de 0,0625 g, por debajo de este peso la

repetitividad de los resultados es baja. En el caso de muestras muy ricas en materia orgánica es preferible duplicar la cantidad de dicromato potásico y ácido sulfúrico a disminuir excesivamente el peso de muestra. La sensibilidad del método es de 0,3 a 0,8 % de materia orgánica.

Material.

- Mortero de ágata (homogeniezer 2094 Foss)
- Balanza (Sartorius BP 121 S)
- Bureta.
- Pipeta
- Matraces erlenmeyer de 500 ml de vidrio Pyrex.
- Matraces aforados de 1000 ml.
- Probeta de 25 ml y 100 ml
- Vasos de precipitado de 100 ml y 1000 ml.

Reactivos.

R-1. Dicromato potásico 1 N, preparado en matraz aforado de 1 l:

1. Desecar el dicromato potásico en estufa a 105 °C una noche.
2. Disolver 49,05 g de dicromato potásico en 900 ml de agua
3. Enrasar con agua desmineralizada a 1 l.

R-2. Ácido sulfúrico concentrado, 96 %, d = 1,84 $g \cdot cm^{-3}$.

R-3. Sal de Mohr 0,5 N, preparada en matraz aforado de 1000 ml en el siguiente orden:

1. 196,1 g de $Fe(NH_4)_2(SO_4)_2.6H_2O$ (sal de Mohr).
2. 800 ml de agua desmineralizada que contenga 20 ml de H_2SO_4, disolver la sal.
3. Enrasar con agua desmineralizada a 1000 ml.

R-4. Solución de difenilamina, 2,50 g. del producto y se disuelven en 20 ml. de agua destilada y 100 ml. de H_2SO_4 conc.

R-5. Acido ortofosfórico concentrado H_3PO_4.

Procedimiento.

Pulverizar en mortero unos 10 g de muestra seca, de forma que el polvo resultante pase por un tamiz de 0,2 mm. Pesar entre 0,5 y 1 g de muestra (que contenga menos de 20 mg de C). Añadir 20 ml de solución 1 N de dicromato potásico (R-1), imprimiendo un movimiento de giro al erlenmeyer para asegurar una mezcla íntima. Añadir lentamente y agitando, 20 ml de ácido sulfúrico concentrado (R-2). Agitar suavemente para asegurar el contacto íntimo de los reactivos con la muestra. Evitar que se adhieran partículas en las paredes del erlenmeyer fuera del contacto de la solución. Dejar el erlenmeyer en reposo durante 20 minutos. Añadir unos 100 ml de agua desmineralizada y dejar enfriar. Añadir 10 ml de ácido fosfórico concentrado (R-5). Añadir 1 ml. del indicador (difenilamina) (R-4). Valorar con la solución ferrosa (R-3), anotando el volumen gastado. (Viraje pardo rojizo a verde). Hacer paralelamente un ensayo en blanco.

Cálculos y expresión de resultados.

Para pasar de carbono orgánico a materia orgánica, se debe multiplicar por el factor de Van Bemmelen, deducido estadísticamente y que supone que la M.O. del suelo contiene un 58 % de carbono orgánico.

$$\% \text{ M.O.} = \% \text{ C} \times 1{,}72$$

Otra forma de expresar el % de materia orgánica sería:

$$\%\text{M.O.} = \left(1 - \frac{V_m}{V_b}\right) \times \frac{6{,}7}{p} \times \frac{100}{(100 - h)}$$

Donde:

V_m: volumen gastado en la muestra

V_b: volumen gastado en el blanco

p: peso de la muestra

h: humedad de la muestra

Los resultados se expresan como porcentaje de materia orgánica y se realizaron por triplicado.

Determinación de la materia orgánica en mufla.

Este método se usa en muestras con una cantidad de materia orgánica superior al 15 %, con lo que el método descrito anteriormente no sería el adecuado.

Material.

- Cápsulas de porcelana.

- Horno mufla (Selecta, Select Horn).

- Desecador.

Procedimiento.

El procedimiento para hallar el % de materia orgánica sobre suelos o residuos es similar al cálculo de la humedad, con la diferencia, que las muestras secas a 110 ºC se pesan (alrededor de 1 g) en cápsulas de incineración redondas de fondo plano de porcelana y la estufa se sustituye por el horno de mufla eléctrico a 550 ºC, durante 4 horas.

Cálculos.

$$\%\text{M.O.} = 100 \times \frac{P_2 - P_3}{P_2 - P_1}$$

Donde:

P_1: peso en g de la cápsula vacía.

P_2: peso en g de la cápsula con la muestra sin tratar

P_3: peso en g de la cápsula con la muestra a 550 ºC

Evaluación del N total.

Material.

- Digestor (Selecta, Selecta block digest 20).
- Destilador por arrastre de vapor (2100 Kjeltec Distillation Unit, Foss Tecator).
- Matraces Kjeldahl de 250 ml
- Erlenmeyers de 250 ml
- Probeta de 25 ml
- Bureta
- Agitador

Reactivos.

R-1. Ácido sulfúrico (H_2SO_4) concentrado.

R-2. Ácido clorhídrico 0,05 N.

R-3. Ácido bórico (H_3BO_3) con indicador: Se disuelven 20 g de H_3BO_3 en 700 ml de agua caliente. La solución enfriada se vierte en un matraz aforado de 1 l que contiene 200 ml de etanol y 20 ml de la mezcla de indicadores (R-5), después de mezclar se añade lentamente NaOH 0,05 N hasta tener pH de 5 entonces se enrasa a un litro con agua destilada.

R-4. Hidróxido sódico (NaOH) al 40 %.

R-5. Mezcla de indicadores: 0,099 g de verde de bromocresol y 0,066 g de rojo de metilo disueltos en 100 ml de etanol.

Procedimiento.

Se introducen 2 g de suelo seco y tamizado en un matraz Kjeldahl de 250 ml, y se añaden 20 ml de ácido sulfúrico concentrado.

Se agita circularmente y se coloca en el digestor a temperatura de 350 ºC durante unas 3 horas hasta que la mezcla se vea clarificada.

Se deja enfriar y se diluye con unos 50 ml de agua destilada. El matraz se conecta al destilador, se añade suficiente cantidad de NaOH al 40 % para desplazar todo el amoníaco y se recoge el destilado en un matraz de 250 ml sobre una solución de ácido bórico (25 ml); y se valora con HCl 0,05 N. La valoración nos indica directamente los meq de N-NH_4^+ por gramo de suelo. Se debe realizar un blanco de reactivos para eliminar las posibles interferencias.

Cálculos:

$$g\ \ N\text{-}NH_4^+ \cdot kg_{SS}^{-1} = (V_m - V_n) \times N \times \frac{14}{g_{suelo}} \times \frac{100}{(100 - h)}$$

Donde:

V_m: volumen gastado en la muestra

V_b: volumen gastado en el blanco

N: Normalidad del HCl (tener en cuenta si se usa diferente concentración para la muestra y para el blanco)

g : gramos de suelo pesados en la digestión

h: humedad de la muestra

El resultado se expresa en % y se realiza el análisis por triplicado.

El laboratorio dispone de un equipo más moderno que realiza el análisis del nitrógeno total por medio de la combustión de la muestra sin necesidad de añadir ningún tipo de reactivo. Este equipo es el LECO modelo TruSpec CN Carbon/Nitrogen determinator.

Previo a la utilización de este equipo se comprobó que los valores obtenidos por ambos métodos eran iguales.

Determinación de nitrógeno total con LECO TruSpec CN

El equipo Leco modelo TruSPec CN es un analizador elemental macromuestra para la determinación de carbono y nitrógeno basado en la técnica Dumas de combustión total de la muestra y posterior determinación del contenido total de cada elemento.

Procedimiento:

La muestra, tanto de suelo como de planta, tiene que estar finamente pulverizada y ser homogénea.

Para pesar la muestra se necesitan unas finas hojas de estaño. La cantidad de muestra es de 0,5 g y se realiza por triplicado.

1. Preparar el instrumento para la operación tal como se indica en el manual de instrucciones del operador.
2. Fijar el valor del blanco como se indica en el manual de instrucciones del operador.
3. Calibrar el equipo como se describe en el manual de instrucciones del operador.
4. Realizar la corrección de la deriva como se indica en el manual de instrucciones del operador. Nota: La deriva se debe realizar al comienzo de cada día o cuando el patrón de control no devuelve el resultado correcto (s).
5. Pesar la muestra en una lámina de estaño e introducirla en la posición apropiada del carrusel de muestra y proceder con el análisis.

Para calibrar el equipo se necesitan unos patrones con contenido en nitrógeno y carbono semejante al de la muestra a analizar.

Fósforo asimilable

Esta técnica pretende determinar una cantidad de fósforo o de fosfato que esté directamente relacionada con la que extraen las plantas durante la estación de crecimiento: fósforo "asimilable" o "fácilmente soluble", que no es sino un índice de asimilabilidad. Dada la dinámica del fósforo y la influencia que sobre ella tienen las características de cada suelo por la presencia de carbonatos en unos casos, por la de hierro y aluminio solubles en otros, se hace difícil alcanzar el objetivo indicado.

Para la determinación del fósforo asimilable vamos a utilizar el método de Olsen-Watanabe, debido a que el pH de nuestros suelos es básico. Sin embargo, existen

otras técnicas propuestas por otros autores para suelos con distintos pH. Por ejemplo: el método propuesto por Bray-Kurtz extrae el fósforo soluble en ácido clorhídrico y fluoruro amónico diluidos, concretamente los fosfatos cálcicos y parte de los alumínicos y férricos. Se suele emplear en suelos ácidos. El método Burriel-Hernando, realiza la extracción del fósforo asimilable con una solución de carbonato cálcico y carbonato magnésico solubilizados con ácido sulfúrico y ácido acético. Se recomienda para suelos de pH superior a 6,5.

Método Olsen-Watanabe

Materiales.

- Balanza (Sartorius BP 121 S).
- pH-metro (Crison).
- Estufa (Selecta).
- Desecador.
- Agitador magnético.
- Espectrofotómetro visible y ultravioleta (Milton Roy, Spectronic genesys 2).
- Centrífuga (Kubota, mod Hertz)
- Botes de centrífuga.
- Material de vidrio.
- Embudos

Reactivos:

R-1. Bicarbonato sódico ($NaHCO_3$) 0,5 M a pH = 8,5, preparado en el siguiente orden:

1. Pesar 42 g de $NaHCO_3$ en un vaso de precipitado de 100 ml.
2. Disolver en aproximadamente 800 ml de agua desmineralizada.
3. Ajustar el pH a 8,5 con NaOH.
4. Pasar cuantitativamente a matraz aforado de 1000 ml.
5. Añadir agua desmineralizada hasta 1000 ml. Homogeneizar.

R-2. Solución madre de 100 $mg \cdot l^{-1}$ de P, preparado de la siguiente manera:

1. Pesar del orden de 1 g de fosfato monopotásico KH_2PO_4.
2. Secar en estufa a 105 °C durante 1 h.

3. Enfriar en desecador.

4. Pesar en un vaso de precipitado de 500 ml 0,4392 g de KH_2PO_4 seco.

5. Disolver en agua desmineralizada.

6. Pasar cuantitativamente a matraz aforado de 1000 ml.

7. Añadir agua desmineralizada hasta 1000 ml. Homogeneizar.

R-3.Solución de trabajo de 10 $mg \cdot l^{-1}$ de P, preparada de la siguiente manera:

1. Pipetear 10 ml de la solución de 100 $mg \cdot l^{-1}$ de P y llevar a un matraz aforado de 100 ml.

2. Añadir agua desmineralizada hasta 100 ml. Homogeneizar.

R-4. Reactivo A, preparado de la siguiente manera:

1. Disolver 6 g de molibdato amónico en 125 ml de agua desmineralizada.

2. Disolver 0,1454 g de tartrato antimónico potásico en 50 ml de agua desmineralizada.

3. Mezclar las dos soluciones anteriores con 500 ml de H_2SO_4 5N.

4. Pasar cuantitativamente a matraz aforado de 1000 ml.

5. Añadir agua desmineralizada hasta 1000 ml. Homogeneizar.

R-5. Reactivo B, preparado de la siguiente manera:

1. Disolver 1,056 g de ácido ascórbico en 200 ml del reactivo A.

2. Homogeneizar.

R-6. Soluciones diluidas de H_2SO_4.

Procedimiento:

Extracción del P-asimilable.

1. Pesar 5 g de muestra en un bote inerte al fósforo de 200 ml de capacidad.

2. Añadir 100 ml de la solución extractante (bicarbonato sódico).

3. Agitar durante 30 min.

4. Filtrar la suspensión.

5. Realizar un ensayo en blanco

Preparación de la curva patrón. Se prepara una escala de patrones de 0; 0,2; 0,4; 0,6; 0,8 y 1,0 $mg \cdot l^{-1}$ de P de la siguiente manera:

1. Llevar a vasos de precipitado de 100 ml las siguientes cantidades de la solución de trabajo de 10 $mg \cdot l^{-1}$ de P: 0, 1, 2, 3, 4 y 5 ml.

2. Añadir 10 ml de la solución extractante (bicarbonato sódico).

3. Añadir H_2SO_4 para conseguir pH = 5 (medir con pH-metro.).

4. Pasar cuantitativamente a un matraz aforado de 50 ml.

5. Añadir 8 ml de reactivo B.

6. Añadir agua desmineralizada hasta 50 ml. Homogeneizar.

Preparación del extracto de suelo.

1. Llevar a un vaso de precipitado de 50 ml una alícuota de 5 ml de extracto de suelo.

2. Proceder de forma análoga a los puntos 3, 4, 5 y 6 del apartado anterior.

Determinación del contenido de P-asimilable. Efectuar las lecturas, comenzando por los patrones, pasados 15 min y antes de 24 h de haber añadido el reactivo B.

Las lecturas se realizan a una longitud de onda de 882 nm.

<u>Cálculos.</u>

$$mg \cdot kg^{-1}_{ss}\ P = c \frac{100}{p} \frac{50}{a} \frac{100}{(100 - h)}$$

Donde:

c: lectura del espectrofotómetro en $mg \cdot l^{-1}$

p: peso de la muestra

a: ml de alícuota tomada

h: humedad de la muestra.

El laboratorio dispone de un equipo automatizado para la medición de este parámetro llamado SKALAR SAN^{plus} ANALYZER. Al igual que en el caso del nitrógeno se comprobó previamente que los resultados eran los mismos por ambos métodos.

Determinación del fosfato por colorimetría. Método automatizado

Principio:

El procedimiento automatizado para la determinación de fosfato se basa en la siguiente reacción; después de la diálisis con agua destilada, el heptamolibdato de amonio y el potasio de antimonio (III) óxido de tartrato, reaccionan en un medio ácido con soluciones diluidas de fosfato para formar un complejo antimonio-fosfo-molibdato. Este complejo es reducido a un complejo de color azul intenso por L(+) ácido ascórbico. El complejo se mide a 880 nm

Procedimiento para la preparación de la muestra:

-Se utiliza el mismo procedimiento que en el método anterior.

Reactivos del equipo:

R1. Solución de ácido sulfúrico: Diluir 40 ml de H_2SO_4 (96%) en ± 800 ml de agua destilada. Llevar a 1 l con agua destilada, añadir 2 ml de FFD6 y mezclar. Solución estable 1 semana.

R2. Agua destilada + FFD6: Diluir 2 ml de FFD6 en 1 litro de agua destilada y mezclar. Solución estable 1 semana.

R3. Solución de heptamolibdato de amonio: Diluir 40 ml de H_2SO_4 (96%) en ± 800 ml de agua destilada. Añadir 4,8 g de heptamolibdato de amonio y disolver. Llevar a 1 l con agua destilada y añadir el FFD6 y mezclar. Solución estable 1 semana.

R4. Solución de potasio de antimonio (III) óxido de tartrato [$K(SbO)C_4H_4O_6 . H_2O$]: Disolver 300 mg de potasio de antimonio (III) óxido de tartrato en ± 80 ml de agua destilada. Llevar a 100 ml con agua destilada y mezclar. Solución estable 4 semanas. Conservar a 4ºC.

R5. Solución de .L (+) ácido ascórbico: Disolver 18 g de L (+) ácido ascórbico ($C_6H_8O_6$) en ± 800 ml de agua destilada. Añadir 20 ml de solución madre de potasio de antimonio (III) óxido de tartrato. Llevar a 1 l con agua destilada y mezclar. Solución estable 5 días. Conservar a 4ºC.

El lavado de las muestras con agua destilada

Blancos:

Solución madre 100 $mg \cdot l^{-1}$: Disolver 439,4 mg de fosfato de dihidrógeno potásico (KH_2PO_4) en ± 800 ml de agua destilada. Llevar a 1 l con agua destilada y mezclar. Solución estable 1 mes. Conservar a 4ºC.

Normas de trabajo para 5, 4, 3, 2 y 1 $mg \cdot l^{-1}$ P: Diluir 5, 4, 3, 2 y 1 ml respectivamente de solución madre 100 $mg \cdot l^{-1}$ P a 100 ml con lavado de muestra líquida.

Los reactivos son suministrados por MERCK. En particular ácido sulfúrico (95-97%) (Merck 100731), FFD6 (Skalar SC 13908), heptamolibdato de amonio tetrahidratado (Merck 101182), potasio de antimonio (III) óxido de tartrato hemihidratado (Merck 108092), L (+) ácido ascórbico (Merck VWR PRO 20150.231), fosfato de dihidrógeno potásico (Merck 104873), solución de hipoclorito de sodio (6-14% cloro activo) (Merck 105614).

Cationes de cambio

Los cationes intercambiables del suelo se desplazan mediante extracciones sucesivas con una solución 1 N de acetato amónico a pH = 7,0 y se determinan en dicho extracto.

Material:

- Agitador orbital (C25 incubator shaker, New Brunswick scientific CO., IN).

- Centrífuga

- Tubos de centrífuga de 50 ml.

- Espectrofotómetro de absorción-emisión atómica (Perkin Elmer, mod 3100)

Reactivos:

R.1. Ácido acético glacial (d =1,05 $mg \cdot l^{-1}$)

R.2. Hidróxido amónico concentrado (d = 0,90 $mg \cdot l^{-1}$)

R.3. Solución de acetato amónico, 1 N, ajustada a pH = 7,0; por cada litro de solución que se prepare añadir 57 ml de ácido acético glacial a unos 600 ml de agua y 68 ml de hidróxido amónico concentrado. El hidróxido debe incorporarse en una vitrina de gases a través de un embudo de cuello largo de tal manera que llegue al fondo de la solución del ácido. Dejar enfriar y ajustar a pH 7,0 con ácido acético o hidróxido amónico. Diluir la solución al volumen convenido.

Procedimiento

1.- Extracción de cationes intercambiables: Añadir al tubo que contiene el suelo 25 ml de acetato amónico 1N, tapar y agitar durante 20 minutos. Centrifugar 12 minutos a 2500 rpm. Decantar el líquido en un matraz aforado de 100ml. Repetir la extracción y decantación dos veces más, teniendo siempre en cuenta que durante las agitaciones el suelo debe volver a suspenderse. Enrasar con la solución de acetato amónico.

2.- Determinación de Na y K por emisión atómica.

3.- Determinación de Ca y Mg por absorción atómica.

Cálculos

$$mg \cdot kg_{ss}^{-1}\ Catión = c \cdot \frac{100}{g} \cdot \frac{100}{100 - h}$$

$$cmol_c \cdot kg_{ss}^{-1}\ Catión = mg \cdot kg_{ss}^{-1}\ \frac{100}{M \cdot 1000} \cdot v$$

Siendo:

c = Concentración ($mg \cdot l^{-1}$) de Na^+, K^+, Ca^{2+} o Mg^{2+} del espectrofotómetro de absorción o emisión.

g = peso de la muestra.

M = masa molecular del elemento

v = valencia del catión.

Digestión de muestras

A efectos de asegurarse una evaluación completa de los elementos traza presentes se realizó una digestión en microondas, (modelo Anton Paar Multiwave 3000).

Material:

- Estufa mod. 204 P. Selecta
- Molino de bolas (Fritsch Pulverisette mod. 6)
- molino de cuchillas tipo SM 100 “Restsh”
- Balanza de precisión mod. 405M-200PA
- Triturador “Homogenieer 2094 Foss”
- Tubos de teflon
- Pipetas
- Matraces aforados de 25 ml
- Rotor de 16 posiciones media presión
- Rotor de 8 posiciones de alta presión
- Filtros
-Embudos

- Tubos de vidrio de diferentes volúmenes

Reactivos:

R.1. Ácido nítrico P.A. 69 %

R.2. Agua oxigenada al 33 %

Procedimiento:

a) Suelos y compost de lodo

Estas muestras recibieron un tratamiento de presecado a 105 ºC en estufa, (mod. 204 P. Selecta). Posteriormente se homogeniza la muestra con ayuda de un molino de bolas (Fritsch Pulverisette mod. 6) incrementando así la superficie de ataque oxidante. Del polvo obtenido, se pesan 0,5 g dentro de los reactores del microondas (botes de teflón) usando una microbalanza (mod. 405M-200PA Precisa, de sensibilidad 0,0001 mg), y se añaden 10 ml de HNO_3 69 %, introduciéndose en el microondas con el rotor de 8 posiciones para su digestión.

Paralelamente se realizan blancos en los que no se añadió ni suelo ni compost. Muestras y blancos se digieren por triplicado (con fines estadísticos). Una vez finalizado el proceso de oxidación, el digerido se filtra y se lleva a matraces aforados que se enrasan a 25 ml con agua ultrapura.

b) Planta

Estas muestras recibieron un tratamiento de presecado a 60 ºC en estufa, (mod. 204 P. Selecta). Posteriormente se homogeniza la muestra con ayuda de un molino de bolas (Fritsch Pulverisette), un molino de cuchillas tipo SM 100 “Restsh” o un triturador “Homogenieer 2094 Foss”, incrementando así la superficie de ataque oxidante. Del polvo obtenido, se pesan 0,5 g dentro de los reactores del microondas (botes de teflón) usando una microbalanza (mod. 405M-200PA Precisa, de sensibilidad 0,0001 mg), y se añaden 5 ml de HNO_3 69 % y 2 ml de peróxido de oxígeno 33 %, introduciéndose en el microondas para su digestión.

Paralelamente se realizan blancos en los que no se añadió planta. Muestras y blancos se digieren por triplicado (con fines estadísticos). Una vez finalizado el proceso de oxidación, el digerido se lleva a matraces aforados que se enrasan a 25 ml con agua ultrapura.

Medida de elementos mediante espectrofotometría de absorción atómica

Esta técnica utiliza los espectros de absorción, que junto con los de emisión y fluorescencia coexisten en la llama. La ley de Lambert-Beer relaciona linealmente la absorbancia medida con la concentración de partículas metálicas de la muestra (Koirtyohann, 1991; Kolodko y Fedoseenko, 1991).

Material:

- Espectrofotómetro de absorción atómica de lla Perkin Elmer mod 3100
- Matraces aforados de distintas medidas
- Pipetas

Reactivos:

R.1. Soluciones patrón de los diferentes metales. Uvisol Merck de 1000 $mg \cdot l^{-1}$

Procedimiento:

Las soluciones de plantas, compost de lodos o suelos obtenidas en la digestión con microondas se someten a medida directa de Cd, Cu, Fe, Mn, Ni, Pb y Zn en "llama fría" aire-acetileno, con un espectrofotómetro de A.A. Perkin Elmer mod. 3100. La determinación de Cr se hizo también por absorción, pero con una llama reductora. Para la determinación de Na y K se empleó la técnica de emisión, y para Ca y Mg absorción atómica en llama aire-acetileno en las muestras extraídas con acetato amónico.

En el caso de la necesidad de diluir, derivada de una concentración que supera a la curva patrón, se realiza su dilución con agua ultrapura.

La curva patrón se prepara a partir de disoluciones acuosas preparadas (Uvisol Merck de 1000 $mg \cdot l^{-1}$) de cada elemento, a partir de la cual se prepararon diluciones sucesivas de los distintos patrones de calibración.

En las digestiones se utiliza HNO_3 calidad Merck, este reactivo se emplea como estándar interno en los patrones de calibración. Todos los patrones de calibración fueron preparados con agua ultrapura (Milli-Q50).

A efectos estadísticos, blancos y digeridos se preparan por triplicado; de cada muestra, sometida a control, se realizan tres medidas.