Predicción de las propiedades químicas del suelo Calcio y Magnesio, a través de una relación suelo-paisaje

Y. Labrador[1], C. Chang[1] y J. Viloria[2]

[1]Grupo de Inteligencia Artificial, Departamento de Computación y Tecnología de la Información, Universidad Simón Bolívar, Caracas, Venezuela
[2]Instituto de Edafología, Facultad de Agronomía, Universidad Central de Venezuela, Caracas, Venezuela

RESUMEN

Los usuarios necesitan saber cómo son las características del suelo en sitios de su interés; pero las propiedades del suelo pueden ser determinadas solo en un número pequeño de puntos de muestreo. Como consecuencia, es necesario predecir cómo es el suelo en los puntos no muestreados. En este estudio se propone un sistema de predicción de valores de propiedades del suelo, basado en redes neuronales de regresión generalizada. Este tipo de red es particularmente útil cuando se dispone de una cantidad pequeña de datos, como ocurre comúnmente en los inventarios de suelo. El sistema propuesto calcula el error medio cuadrático, el error absoluto medio y el coeficiente de determinación como indicadores del error de predicción. También calcula la proporción de puntos sin predicción. Esta información ayuda al usuario a seleccionar la combinación óptima de variables de entrada y parámetros del sistema, de acuerdo a sus necesidades. El sistema permitió generar mapas de concentración de calcio y magnesio en el suelo, a partir de un modelo digital de elevación, una imagen satelital y los valores medidos en un número limitado de puntos de muestreo, en un sector representativo de la cuenca del río Caramacate (estado Aragua, Venezuela). La selección de las variables de entrada a la red y el valor del parámetro σ permitió minimizar el error de predicción y el porcentaje de puntos sin clasificar. Los resultados revelan que la selección de las variables de entrada a la red es crucial para garantizar el éxito de la predicción.

ABSTRACT

Users need to know the soil properties at sites of their interest, but soil properties can be determined only in a small number of sampling points. Consequently, it is necessary to predict soil conditions at not sampled points. This study proposes a system to predict values of soil properties, based on generalized regression neural networks. This type of network is particularly useful when the amount of available data is small. This is usually the case in soil inventories. The proposed system calculates the mean square error, the mean absolute error and the determination coefficient as indicators of the prediction error. It also calculates the proportion of points with no prediction. This information helps the users to select the optimal combination of input variables and system parameters, according to their needs. The system produced maps of soil available calcium and magnesium, from a digital elevation model, a satellite image and the measured values at a limited number of sampling points, in the Caramacate river basin (Aragua State, Venezuela). The selection of input variables to the network and the parameter σ allowed minimizing the prediction error and the percentage of items rated. The results show that the selection of the input variables is crucial for the success of predictions.

Keywords: Red neuronal de regresión generalizada (GRNN), relación suelo-paisaje, variabilidad de suelos.

1. Introducción

El suelo es la capa superficial que cubre la corteza terrestre en forma de un manto continuo, con excepción de aquellas áreas donde existen cuerpos de agua, afloramientos rocosos u otros elementos del paisaje que no son suelo [Vil06]. A finales del siglo XIX se plantea que la variación espacial del suelo no es totalmente aleatoria, sino que sigue patrones relacionados con el clima, la vegetación, el material parental, el relieve y la edad formadora del terreno, todos ellos denominados factores formadores del suelo [Vil06]. Jenny propone la ecuación de los factores de estado de suelos y ecosistemas [Jen61], denominada así debido a que estos factores por sí solos no forman suelo, pero su intervención ayuda a explicar el estado de las propiedades del mismo.

$$s = f(cl, b, mp, r, t, \ldots) \quad (1)$$

Dada la ecuación 1, podemos concluir que una característica del suelo s en un momento y lugar determinado, es el resultado de una función en donde interviene el clima (cl), la biota (b), el material parental (mp), el relieve (r), la edad del paisaje (t) y en casos particulares factores adicionales.

Para conocer el valor exacto de diversas propiedades químicas y físicas del suelo, se hace necesario tomar puntos de muestreo en el área que se desea estudiar. En estos puntos se toman porciones de suelo a profundidades específicas para analizarlas y determinar los valores de las propiedades del suelo en ese sitio.

Los planificadores, agrotécnicos, inversionistas y otros usuarios, comúnmente necesitan saber cómo son las características del suelo en lugares o áreas de su interés, para así tomar decisiones sobre el uso apropiado del mismo. Sin embargo, por razones prácticas, las propiedades del suelo pueden ser determinadas solo en un número limitado de puntos de muestreo. Este número está determinado por factores como la morfología y accesibilidad del terreno, el costo monetario asociado a la recolección y análisis de las muestras y el tiempo disponible para realizar el estudio.

Generalmente se tienen pocos puntos de muestreo en la zona de estudio. Como consecuencia, se hace necesario aplicar métodos de predicción para determinar el valor de las propiedades del suelo entre estos puntos y conocer cómo es su comportamiento en toda la zona. Estos métodos de predicción hoy en día son variados, y se fundamentan en alguno de los siguientes modelos o hipótesis de variación espacial del suelo [Vil06]:

- El suelo es un mosaico de unidades discretas: este modelo considera que el suelo está formado por un conjunto de cuerpos naturales, representados en un mapa como unidades geográficas individuales separadas entre sí por límites abruptos.
- El suelo es un continuo: este modelo supone que los procesos que generan la variabilidad del suelo tienden a originar cambios graduales a lo largo de un continuo, en lugar de cuerpos discretos con límites abruptos.
- El suelo es un continuo con unidades discretas: este modelo supone que el suelo se comporta como un continuo dentro de distancias cortas, pero en un ámbito más regional frecuentemente se caracteriza por discontinuidades abruptas [Hud92].

Debido a la necesidad de predecir cómo es el comportamiento del suelo entre los puntos de muestreo, diversos investigadores han intentado establecer o modelar relaciones entre el suelo y el paisaje [MV04, Vil07, ZBVD97, Cru09, Que08]. Técnicas tales como: métodos estadísticos, redes neuronales artificiales, árboles de decisión, entre otras, han sido utilizadas para identificar y definir relaciones suelo-paisaje y para determinar la factibilidad de extrapolar los valores de los puntos de muestreo, con base en estas relaciones. El análisis de estas relaciones también permite determinar cuáles factores intervienen en la variación de las propiedades del suelo de manera individual [SHS05].

Por otra parte, los métodos clásicos usados para extrapolar los datos de los puntos de muestreo, agrupan estos puntos en clases de suelo (cuerpos naturales), y representan la distribución geográfica de estas clases en forma de unidades cartográficas. Una de las principales desventajas de estos métodos

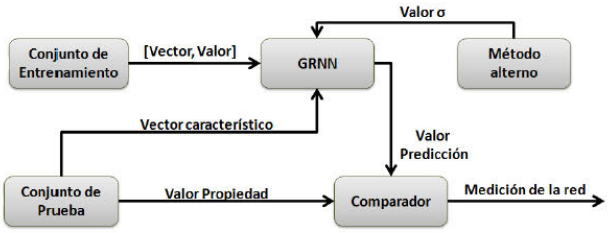

Figura 1: *Sistema propuesto para la obtención y creación de mapas de predicción por propiedades del suelo, teniendo como entrada un vector de características del paisaje y un conjunto de puntos de muestreo*

consiste en que los mapas resultantes representan la variabilidad del suelo en forma escalonada. Esto es, los cambios ocurren en los límites entre unidades cartográficas, dentro de cada unidad el suelo se considera uniforme. En adición a esto, las clases de suelos con poca superficie pueden ser ignoradas o varias clases de suelo pueden ser incluidas en una sola unidad cartográfica [Vil06, ZHBL01].

Por otro lado, existen métodos de interpolación para la predicción de propiedades del suelo, siendo kriging uno de las más destacados [JLT05, MM03, DKOB97, MSW98]. Sin embargo, kriging requiere la estimación de un variograma, para lo cual se requiere un número relativamente alto de datos porque esta función es sensible a valores extremos. Asimismo, la aplicación de kriging exige que la variación de los datos se ajuste a determinados supuestos teóricos de estacionaridad estadística. De esta manera, kriging no es un método adecuado para conjuntos de datos que tienen una variabilidad extrema o presentan cambios bruscos [Hu95].

Debido a las desventajas de los métodos de estimación convencionales y de interpolación, en investigaciones recientes [MV01, Vil07, ZBVD97, ZHBL01, ZQMB10, Cru09] se ha planteado utilizar paradigmas de inteligencia artificial, los cuales permiten extraer conocimientos sobre la relación entre las características del paisaje, para luego correlacionarlas con las propiedades del suelo.

En diversas investigaciones [Vil07, Cru09, SK10] se ha evaluado el desempeño de técnicas de inteligencia artificial para dar solución al problema planteado, sin embargo, estas técnicas pudieran resultar inadecuadas cuando la cantidad de datos es escasa. Es por esto que se debe determinar un modelo que realice una buena predicción con un conjunto pequeño de puntos de muestreo.

Las redes neuronales de regresión generalizada (GRNN), son redes neuronales probabilísticas [Spe91] que pueden predecir resultados continuos. Por estar basadas sobre métodos estadísticos colocan más énfasis en la estructura de los datos. Como consecuencia, el número de datos necesario para poder predecir resultados es mucho más pequeño que el requerido por otros modelos de redes neuronales. Esto es una gran ventaja al momento de resolver problemas donde no se cuenta con una gran cantidad de datos y se trabaja solo con algunos ejemplos [SHZN08, BMAP05, PNB*09, DTZ*08].

Teniendo esto en cuenta, se propone construir un sistema de predicción de valores de propiedades químicas del suelo, como se muestra en la figura 1. Las variables de entrada

consisten en valores medidos en puntos de muestreo de suelo y variables del paisaje, derivadas de un modelo digital de elevación (MDE) y de una imagen satelital. El sistema establece una relación entre el suelo y las características del paisaje a través de una red neuronal de regresión generalizada (GRNN). Esto permite predecir valores de la disponibilidad de calcio (*Ca*) y magnesio (*Mg*) en el suelo, en otros puntos del área de estudio. Asimismo, se propone utilizar la red GRNN para generar mapas predictivos de las propiedades del suelo mencionadas, que muestren a los usuarios la variabilidad espacial de las mismas en la zona de estudio.

2. Materiales y Métodos

La zona de estudio es un sector representativo de la cuenca del río Caramacate, la cual se encuentra ubicada en los municipios Santos Michelena y San Sebastián de los Reyes del Estado Aragua, entre las coordenadas UTM (huso 19) 1.098.310 - 1.123.583 (norte) y 696.879 - 712.415 (este). El clima de la zona varía desde Bosque Seco Tropical hasta Bosque Húmedo Premontano con el aumento de la altitud [MV04].

A continuación se presentan algunas definiciones asociadas con el contexto del problema, las variables utilizadas y el método de predicción propuesto.

2.1. Recolección y análisis de puntos de muestreo

Los puntos de muestreo son puntos en donde toman muestras del perfil del suelo a una profundidad especifica, los cuales permiten determinar el valor exacto de diversas propiedades químicas y físicas del suelo, a través de un análisis posterior realizado sobre la muestra en un laboratorio.

Para la zona de estudio, expertos del Instituto de Edafología de la Facultad de Agronomía, de la Universidad Central de Venezuela, se han desplazado al sitio y han realizado la obtención de 76 puntos de muestreo, ubicandose en lugares en donde los puntos seleccionados pudiesen representar la mayor variabilidad de los factores formadores del suelo, previamente seleccionados de un mapa resultante de una clasificación borrosa del paisaje a través de una herramienta que implementa un algoritmo de inteligencia artificial, llamado Fuzzy Kohonen Clustering Networks (FKCN) [Vil07]. La ubicación de los 76 puntos de muestreo que actualmente se tienen sobre la zona de estudio se muestran en la figura 2.

Es importante resaltar que para cada uno de los puntos de muestreo se tiene la ubicación del punto en donde se tomó la muestra, coordenadas (x,y), y los valores reales de las propiedades químicas del suelo para el calcio (Ca), magnesio (Mg), potasio (K), sodio (Na), capacidad de intercambio catiónico (CIC), suma de bases, entre otros.

2.2. Definición de variables de entrada y propiedades a inferir

Para la investigación se tienen definidas diversas variables de entrada que nos aportan información acerca de cómo es el área de estudio. Estas variables en conjunto con los valores de las propiedades químicas tomadas de los puntos de muestreo son las entradas a la red neuronal.

Las variables utilizadas para el estudio se presentan a continuación.

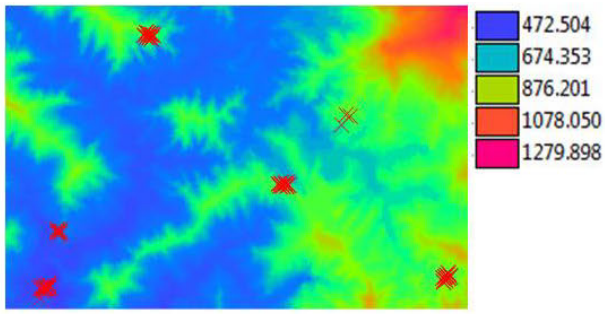

Figura 2: *Modelo digital de Elevación y ubicación de Puntos de Muestro. Subcuenca del río Caramacate*

Modelo Digital de Elevación

Un modelo digital de elevación (MDE) es una estructura numérica de datos que representa la distribución espacial de la altitud de la superficie del terreno [Fel94]. Para el área seleccionada se utilizó un modelo MDE tipo raster el cual consiste en una matriz dividida en celdas, en donde el valor de la altitud para cada una de estas celdas es la media o promedio de los valores no nulos de la altitud en cada uno de los puntos que conforman la misma.

En este tipo de estructura, la localización de cada dato se encuentra determinada por su situación en la matriz de datos, y la resolución o tamaño del modelo está dado por la cantidad de celdas y el tamaño del lado de cada una de las mismas.

La figura 2 muestra el MDE de la zona de estudio, subcuenca del río Caramacate, el cual es un raster (matriz) de 240 filas y 380 columnas, donde cada celda o pixel representa un cuadrado de 20mts, es decir, el área de estudio tiene un tamaño de 4800 x 7600 mts.

Modelos derivados del modelo digital de elevación

A partir del MDE y mediante el análisis de vecindad entre las celdas del mismo, es posible inferir nueva información, tal como el grado de la pendiente, la orientación del terreno, la curvatura, la delimitación de cuencas de drenaje, entre otras propiedades del paisaje, y con ello poder construir un conjunto de modelos derivados a partir de esta información implícita o explícita que ofrece el modelo.

Actualmente, se cuenta con un conjunto de herramientas y/o aplicaciones que permiten realizar este proceso de obtención de modelos derivados de una manera automática. Para esta investigación se utiliza ILWIS v.3.4 Open y DiGeM v.2.0 como herramientas de manipulación de modelos y variables del terreno.

Los modelos derivados del MDE que son utilizados en esta investigación son los siguientes:

Gradiente de la pendiente (slope): en un punto del terreno se define como el ángulo existente entre el vector normal a la superficie en ese punto y la vertical, siendo la tasa de cambio de la altitud obtenida al realizar un desplazamiento horizontal [Vil07].

Orientación de la pendiente (aspect): la orientación de la pendiente es la dirección donde se produce el máximo grado de cambio en la altitud de cada celda con respecto a sus 8 vecinos.

Perfil de curvatura (profile): es la curvatura de la superficie del terreno en la dirección de la pendiente, mide el grado de cambio de la pendiente que afecta a la aceleración o desaceleración del flujo del agua e influencia la erosión y deposición de las partículas del suelo. Las áreas con un perfil convexo indican mayor potencial para la erosión y áreas con perfil cóncavo indican mayor potencial para la deposición.

Plano de curvatura (plang): la curvatura en planta, como también es llamada, es la curvatura en la dirección perpendicular a la pendiente, mide la divergencia o convergencia del flujo del agua y por tanto de la concentración del agua en el paisaje. Representa la curvatura de las curvas de nivel de un mapa topográfico.

Área de captación (catchment): es el área de drenaje contribuyente a un punto específico de la cuenca.

Índice topográfico de humedad (ind hum): es una función que permite inferir la cantidad de agua que puede llegar a un punto dado, influenciado por el área de captación y la pendiente. Se calcula por medio de la siguiente fórmula matemática [Vil07]:

$$indhum = Ln(\frac{(catchment)}{(Tag(slope))}) \quad (2)$$

Imagen Satelital

En la imagen satelital que se tiene de la zona, se muestra por cada pixel el promedio de brillo o radiación de la medición electrónica del pixel al área correspondiente de la tierra [Vil07]. De esta imagen satelital podemos obtener las bandas roja e infrarroja necesarias para generar el índice de vegetación de diferencia normalizada (NDVI) de un área.

$$NDVI = \frac{IRCercano - ROJO}{IRCercano + ROJO} \quad (3)$$

El NDVI se utiliza para calcular la calidad y cantidad de vegetación de un punto determinado, mostrando la respuesta de la vegetación al grado de variación del clima de una determinada zona. Es el resultado de la división pixel a pixel de los valores de reflectividad de dos bandas de la imagen, rojo e infrarrojo cercano (Ecuación 3). El resultado de esta división toma valores entre $-1{,}0$ y $+1{,}0$.

Propiedades químicas del suelo

Basados en estudios realizados por el Instituto de Edafología de la Facultad de Agronomía sobre la zona de estudio de la cuenca del Río Camaracate, en donde se investiga y mide a través de métodos estadísticos la variabilidad de las propiedades del suelo a diferentes escalas: 8, 45, 90, 700 y 1500 metros, se han seleccionado las propiedades químicas calcio (Ca) y magnesio (Mg), como propiedades a estudiar en la predicción de valores sobre el área de estudio, en una fase inicial.

En recientes investigaciones del instituto sobre la zona, se ha podido determinar que los niveles de Ca y Mg tienden a presentar mayor variabilidad a una escala de 700 metros lo que nos indica que las características del paisaje influyen sobre estas propiedades en una alta escala, lo cual puede conllevar a mejores resultados y por lo cual las hemos escogido como variables para ser estudiadas.

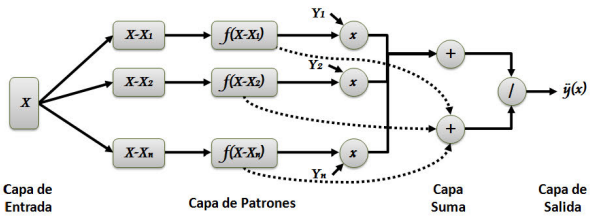

Figura 3: *Arquitectura Red Neuronal de Regresión Generalizada*

2.3. Redes Neuronales de Regresión Generalizada (General Regression Neural Network, GRNN)

Las redes neuronales de regresión generalizada (GRNN) fueron propuestas en 1991 por Specht [Spe91], son redes neuronales probabilísticas las cuales se encuentran basadas en la teoría de regresión no lineal, en donde se realiza la predicción de una variable de salida sobre el conjunto de datos de entrenamiento. La figura 3 muestra la arquitectura de la red, la cual consta de cuatro capas: una capa de entrada, una capa de patrones unitarios, una capa de suma y una capa de salida.

Definamos X como el conjunto independiente de patrones de entrada y la variable y como un patrón de salida el cual es una variable dependiente de x, es ampliamente aceptado que el mejor valor predicho para y (en sentido del mínimo error cuadrático esperado) es su esperanza condicional dado x [BMAP05]. Así que definiendo $f(x,y)$ como la función de densidad probabilística conjunta continua de un vector cualquiera, la esperanza de y dado un vector x viene dado por la siguiente ecuación [Spe91]:

$$E[y|x] = \frac{\int_{-\infty}^{\infty} y f(x,y)dy}{\int_{-\infty}^{\infty} f(x,y)dy} \quad (4)$$

Debido a que en la práctica no se conoce la función de densidad probabilística conjunta, es posible utilizar el estimador de Parzen multivariado para aproximar a la función. Specht [Spe91] demuestra que utilizando este estimador el valor esperado de y puede ser computado utilizando la siguiente fórmula:

$$\dot{Y}(X) = \frac{\sum_{i=1}^{n} Y_i exp(-\frac{D_i^t}{2\sigma^2})}{\sum_{i=1}^{n} exp(-\frac{D_i^t}{2\sigma^2})} \quad (5)$$

Donde

$$D_i^t = (X - X_i)^t (X - X_i) \quad (6)$$

Estas fórmulas permiten que la salida de la red $\ddot{Y}(X)$ sea un promedio pesado de todos los valores de salida Y_i, donde cada valor es influenciado por el peso exponencial de la distancia euclidiana del patrón de predicción X al patrón de entrenamiento X_i. Si la distancia entre un patrón de entrenamiento y el patrón de predicción es pequeña, $exp(-\frac{D_i^t}{2\sigma^2})$ será un número grande y el punto de evaluación Y_i para este patrón representa mucho mejor el punto $\dot{Y}(X)$. Si por el contrario la distancia entre un patrón de entrenamiento y el

patrón de predicción es grande, $exp(-\frac{D_i}{2\sigma^2})$ será un número pequeño y su contribución a la predicción será muy poca.

Para esta red existe solo un parámetro de configuración el cual se denomina smoothing parameter o parámetro de suavidad σ, el cual delimita que tan cerrada puede ser la influencia de un patrón de entrenamiento al patrón de predicción. Cuando σ es un número muy pequeño, tiende a cero, solo los valores Y_i muy cercanos al punto de predicción son tomados en cuenta, cuando σ es un número muy grande el resultado tiende a ser el promedio de todas las observaciones Y_i, para valores intermedios de σ solo son tomados en cuenta los Y_i cuyo vector de entrenamiento se encuentre cercanos al vector de predicción, este es el σ que se debe encontrar.

2.4. Indicadores de desempeño del modelo

El error medio cuadrático (Root Mean Squared Error, RMSE), el error absoluto medio (Mean Absolute Error, MAE) y el coeficiente de determinación (R^2), entre la salida del modelo y la medida de los datos de entrenamiento y pruebas, son indicadores comunes para establecer e indicar una estimación de la exactitud de los modelos estimados. Estos indicadores son estimados y presentados en las ecuaciones 7, 8 y 9 respectivamente.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}[A_i - T_i]^2} \quad (7)$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|A_i - T_i| \quad (8)$$

$$R^2 = \frac{\sum_{i=1}^{n}[A_i - \bar{T}]^2}{\sum_{i=1}^{n}[T_i - \bar{T}]^2} \quad (9)$$

Donde
n = número de observaciones.
T_i = valor observado.
A_i = valor predicho.
$\bar{T}$ = valor promedio de la variable sobre n observaciones.

Cuando el RMSE y el MAE son mínimos y $R^2 >= 0{,}80$, se puede decir que un modelo es bueno [Kas98].

La validación de los datos se hará a través de una validación cruzada de tres particiones, en donde dos de ellas seran utilizadas como conjunto de entrenamiento y una como conjunto de prueba.

Se utilizó Java 1.5 para el desarrollo de la aplicación.

3. Experimentos y Resultados

Debido a que el valor de la salida de la red y el valor esperado son valores reales, un primer punto a tener en cuenta es establecer un valor de diferencia ε que indique la diferencia máxima permitida entre la salida de la red y el valor esperado. Esto permite contar y establecer el número de aciertos o fallas obtenidas por la red para el método de validación cruzada. Definiendo esta constante se tiene la siguiente regla de clasificación:

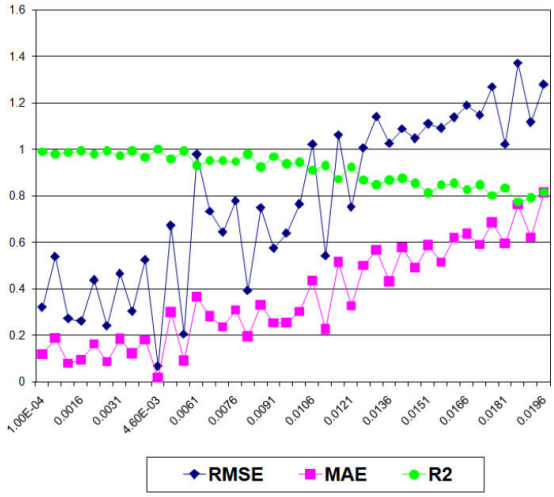

Figura 4: *Variación de indicadores RMSE, MAE y R^2 en la predicción del calcio, con vector de características: aspect, catchment y ndvi*

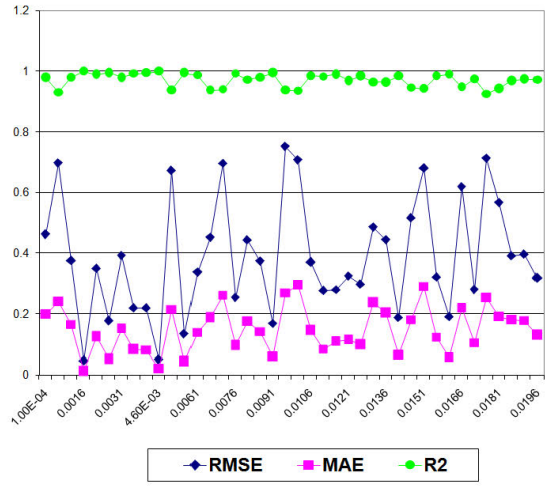

Figura 5: *Variación de indicadores RMSE, MAE y R^2 en la predicción del magnesio, con vector de características: aspect, catchment, slope, indHum y profile*

$$|prediccion - valorreal| > \varepsilon \rightarrow error \quad (10)$$

Un parámetro importante, el cual debemos ajustar y que influye notablemente en la obtención de la solución es el parámetro σ. Para esta primera fase de la investigación el valor de σ es establecido de manera manual, inicializando la variable para cada una de las pruebas en 0,0001 e incrementando su valor en un factor de 0,0005 para cada iteración. Para las pruebas realizadas con valores de σ mayores a 0,02 el error de predicción obtenido es muy alto, por lo cual se decidió realizar las pruebas con valores de σ entre 0,0001 y 0,02.

El objetivo principal de variar el valor de σ, con intervalos cortos entre valores consecutivos, es obtener el comportamiento de la solución generada por el sistema para diversos valores de este parámetro. De esta manera se podrá determinar el valor de σ que genera el mejor resultado.

Para cada una de las pruebas realizadas, se calculan los errores obtenidos en el proceso de validación cruzada, el valor de los indicadores RMSE, MAE y R^2, y la cantidad de puntos que la red no logra predecir sobre la zona de estudio. Estas pruebas se realizan, a su vez, variando el vector de características de entrada a la red, entre las variables aspect, catchment, slope, profile, indHum, ndvi, definidas anteriormente.

Las diversas pruebas realizadas para predecir los valores de la disponibilidad de calcio y magnesio en el suelo, variando el valor del parámetro σ, revelan que el error de predicción es muy bajo cuando σ es pequeño y crece a medida que σ aumenta. En efecto, las figuras 4 y 5 muestran que los indicadores RMSE y MAE alcanzan sus valores mínimos en valores de σ pequeños y se alejan de ellos a medida que el valor de σ va aumentando.

Las figuras 4 y 5 también muestran la variación del indicador R^2 para la predicción del calcio y magnesio, respectivamente, en donde para valores de σ cercanos a 0.02 el indicador R^2 es menor o se hace cercano a 0.80, lo cual nos hace concluir que estas soluciones no son una buena predicción.

Analizando lo anterior, se podría pensar que el valor deseable de σ es aquel que simplemente genera el menor error. Sin embargo, otro factor importante que se debe medir y considerar en la solución encontrada, es la cantidad de puntos sin clasificar en el mapa resultante.

Los mapas predictivos generados por el sistema para las variables calcio y magnesio, con valores diferentes del parámetro σ, muestran que para valores muy pequeños de σ se generan muchos puntos del mapa sin un valor de predicción. Esto se debe a que la red se encuentra muy ajustada a los puntos y a las variables de entrada, por consiguiente predice solo los puntos cuyos vectores de características son muy similares a algún vector de características del conjunto de entrenamiento. Si el valor de σ es muy grande, la red tiende a predecir todos los puntos pero de una manera muy general, debido a que el valor de salida tiende a ser el promedio de todas las observaciones. Es por ello que se debe realizar un análisis comparativo de los resultados, para determinar el valor del parámetro σ que genera un porcentaje de error aceptable y un porcentaje pequeño de puntos sin clasificar.

3.1. Predicción del Calcio

Para la predicción de los valores de calcio disponible en el suelo, se ejecutaron diversas pruebas, en donde para cada una de ellas se realizó una variación en el vector de características de entrada a la red, estimando en cada prueba los errores generados y la cantidad de puntos en el mapa sin clasificar. La tabla 1 nos muestra los dos mejores valores del parámetro σ para diferentes combinaciones del vector de variables de entrada.

La figura 4 muestra una gráfica comparativa de la variación de los indicadores RMSE, MAE y R2, cuando el vector de características de entrada para la predicción del calcio disponible en el suelo está conformado por las variables aspect, catchment y ndvi. El mejor valor obtenido para el parámetro σ parece ser 0.0046, debido a que los indicadores RMSE y MAE son mínimos, R2 es mayor que 0,80 y el porcentaje de valores no clasificados en el mapa es menor que 5%.

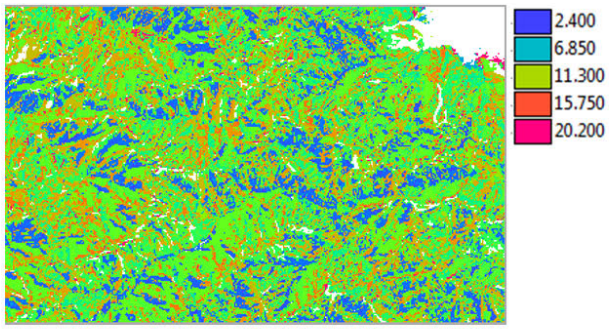

Figura 6: *Mapa predictivo generado por el sistema para la propiedad del calcio,* $\sigma = 0{,}0046$ *y vector de características: aspect, catchment, ndvi*

Carac	σ	No Clsf	RMSE	MAE	R^2
a,c	0.0016	3.55%	0.330	0.127	0.971
a,c	0.0036	9.12%	0.573	0.235	0.960
a,c,n	0.0046	4.87%	0.066	0.018	0.999
a,c,n	0.0056	3.38%	0.208	0.090	0.994
a,c,s, p,ih,n	0.0071	9.97%	0.132	0.039	0.998
a,c,s, p,ih,n	0.0136	0.10%	0.104	0.037	0.999

Tabla 1: *Error promedio para los mejores valores de* σ*, en la predicción del calcio, utilizando las variables aspect (a), catchment (c), slope (s), profile (p), indHum (ih), ndvi (n)*

En la tabla 1 se puede observar que la mejor solución obtenida se encuentra con el vector conformado por las variables aspect, catchment y ndvi, con $\sigma = 0{,}0046$. Esta solución tiene el mejor error encontrado (valores mínimos de los indicadores RMSE y MAE y valor máximo de R2). Sin embargo, el porcentaje de puntos sin clasificar no es el mínimo, aunque es menor al 10%. La figura 6 muestra el mapa de disponibilidad de calcio en el suelo, generado con esta solución. Se puede observar que gran parte de los puntos no clasificados por la red, se encuentran ubicados sobre una zona montañosa más alta del área de estudio, donde no se tienen puntos de muestreo.

Como una solución alternativa, el usuario podría preferir aquella correspondiente a las variables aspect, catchment, slope, profile, indHum, ndvi como vector de características de entrada, con $\sigma = 0{,}0136$. Con esta solución los valores de los indicadores del error de predicción RMSE y MAE son mayores, pero el porcentaje de puntos no clasificados es muy pequeño (0.1%).

3.2. Predicción del Magnesio

Para la propiedad del magnesio se realizaron varios ensayos, que al igual que para la predicción de la propiedad del calcio, consistían en realizar pruebas con varias combinaciones del vector de características de entrada y analizar los resultados obtenidos para cada una de ellas, eligiendo así la mejor solución.

La tabla 2 muestra un resumen de los resultados obteni-

Carac	σ	No Clsf	RMSE	MAE	R^2
a,c	0.0006	21.59 %	0.261	0.073	0.989
a,c	0.0021	1.92 %	0.180	0.070	0.992
a,c,n	0.0001	99.69 %	0.139	0.038	0.996
a,c,n	0.0031	9.65 %	0.254	0.102	0.993
a,c,s, p,ih	0.0016	93.67 %	0.047	0.013	0.999
a,c,s, p,ih	0.0046	22.22 %	0.050	0.021	0.999
a,c,s, p,ih,n	0.0031	82.33 %	0.094	0.028	0.999
a,c,s, p,ih,n	0.0176	0.02 %	0.072	0.028	0.994

Tabla 2: *Error promedio para los mejores valores del parámetro σ, en la predicción del magnesio, utlizando las variables aspect (a), catchment (c), slope (s), profile (p), indHum (ih), ndvi (n)*

dos al variar el vector de características de entrada para la predicción de la propiedad del magnesio.

En este caso, el vector que genera el menor porcentaje de error es aquel conformado por las variables aspect, catchment, slope, profile e indHum, con $\sigma = 0,0046$. Sin embargo, el porcentaje de puntos en los cuales no se genera un valor de predicción es alto (>20 %). Esto da lugar a una gran cantidad de espacios en blanco, en el mapa de predicción de valores de esta propiedad del suelo generado con base en el modelo antes mencionado (figura 7).

No obstante, al agregar al vector de características la propiedad ndvi, con un valor de $\sigma = 0,0176$, el porcentaje de puntos sin predicción disminuye drásticamente a 0,02 %. Como consecuencia, el mapa de predicción de esta propiedad del suelo exhibe una cobertura total del área de estudio (figura 8). En adición a esto, el error de predicción obtenido con este modelo es bajo, en comparación con las otras combinaciones de σ y de variables de entrada mostradas en la tabla 2.

Estos resultados demuestran que la variable ndvi (indicadora de la densidad y vigor de la cobertura vegetal del terreno) brinda información valiosa para predecir la distribución espacial de la concentración del magnesio en el suelo, dentro de la zona estudiada. Sin embargo, se debe depurar el vector de características de entrada, para eliminar variables que quizás introduzcan ruido a la solución e incrementen el porcentaje de error, cuando se incluye la variable ndvi.

Los resultados obtenidos aportan información nueva sobre la relación entre propiedades del paisaje y del suelo. Este tipo de información es de gran utilidad para entender los procesos que determinan la variación espacial de este recurso natural y para predecir los valores de propiedades específicas del suelo, en sitios no muestreados. En adición a esto, los expertos pueden utilizar los mapas que muestran los sitios donde el sistema no generó valores de predicción (figuras 6 y 7 por ejemplo), como base para planificar nuevas campañas de muestreo.

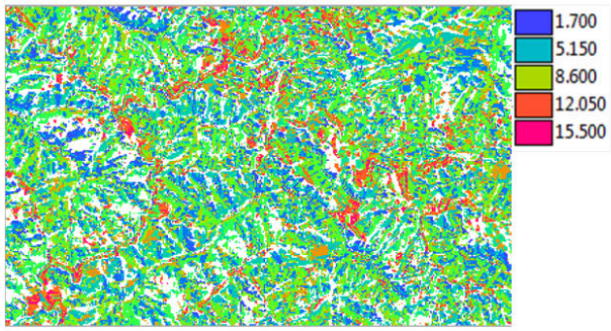

Figura 7: *Mapa predictivo generado por el sistema para la propiedad del magnesio, $\sigma = 0,0046$ y vector de características: aspect, catchment, slope, profile, indHum*

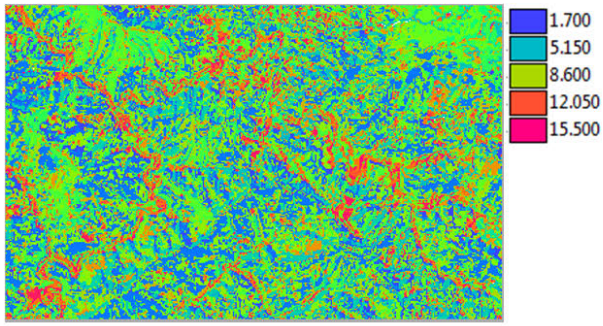

Figura 8: *Mapa predictivo generado por el sistema para la propiedad del magnesio, $\sigma = 0,0176$ y vector de características: aspect, catchment, slope, profile, indHum, ndvi*

4. Conclusiones

Este estudio muestra resultados iniciales enmarcados en la búsqueda de ofrecer una herramienta de apoyo a expertos y usuarios, que permita establecer relaciones entre las propiedades del suelo y el paisaje, de una forma sencilla y automática, reduciendo así los costos y tiempos de investigación.

El sistema propuesto, se basa en redes neuronales de regresión generalizada. Este tipo de red neuronal es particularmente útil cuando se dispone de una cantidad pequeña de datos, como ocurre comúnmente en los inventarios del recurso suelo.

El sistema permitió generar mapas de concentración de calcio y magnesio en el suelo, en la zona de estudio, a partir de un MDE, una imagen satelital y los valores medidos en un número limitado de puntos de muestreo. Adicionalmente, el sistema calcula el error de predicción, medido por medio de indicadores bien definidos, y la proporción de puntos sin predicción. Esta información ayuda al usuario a seleccionar la combinación de variables de entrada y parámetros del sistema que mejor se ajuste a sus necesidades.

Este proceso de generación de mapas de propiedades químicas del suelo es rápido y simple. Reduce el número de horas hombres necesarias para la generación de mapas de propiedades del suelo. Los mapas generados por el sistema, para la predicción de las propiedades químicas estudiadas, son mapas que reflejan la continuidad del suelo y pueden brindar mucha más información a los usuarios para tomar

sus decisiones en cuanto al mejor uso y aprovechamiento del mismo.

El sistema no sustituye la competencia del experto para la elaboración de los mapas. En efecto, la experiencia y habilidad del este último es de suma importancia, para escoger las variables de entrada que recibe la red. Al mismo tiempo, el sistema le permite al experto descubrir nueva información sobre las relaciones entre variables del paisaje y propiedades especificas del suelo.

Para próximos trabajos, enmarcados en este tema, se tiene previsto establecer un mecanismo de obtención del valor del parámetro σ de manera automática, para obtener un resultado mucho más exacto sin necesidad de realizar múltiples ensayos y analizar los resultados de manera manual. Así mismo, se quiere llevar este análisis y pruebas sobre otras propiedades químicas del suelo.

5. Agradecimientos

Los autores agradecen al proyecto LOCTI de la Facultad de Agronomía de la Universidad Central de Venezuela titulado "Sistema de información geográfica de la cuenca alta del río Guárico" y al Consejo de Desarrollo Científico y Humanístico de esa universidad, por su apoyo al desarrollo de esta investigación.

Referencias

[BMAP05] BALLARIN V., MESCHINO G., ABRAS G., PASSONI L.: Segmentación de imágenes cerebrales de resonancia magnética basada en redes neuronales de regresión generalizada. In *XV Congreso Argentino de Bioingenieria* (2005).

[Cru09] CRUZ G.: *Evaluación de métodos para la cartografía digital de clases de tierras campesinas*. PhD thesis, Institución de enseñanza e investigación en ciencias agrícolas. Colegio de postgraduados, Montecillo, Texcoco, Estado de México, 2009.

[DKOB97] DURRANI S., KHAYRAT A., OLIVER M., BADR I.: Estimating soil radon concentration by kriging in the biggin area of derbyshire (uk). *Radiation Measurements 28*, 1-6 (1997), 633–639.

[DTZ*08] DU C., TANG D., ZHOU J., WANG H., A. S.: Prediction of nitrate release from polymer–coated fertilizers using an artificial neural network model. *Biosystems Engineering 99* (2008), 478–486.

[Fel94] FELICÍSIMO A.: *Modelos Digitales del Terreno. Introducción y aplicaciones en las ciencias ambientales*. Pentalfa ediciones, Oviedo, España., 1994.

[Hu95] HU J.: Methods of generating surfaces in environmental gis applications. In *International Users Conference* (1995).

[Hud92] HUDSON B.: The soil survey as paradigm based science. *Soil Science Society of America JOURNAL 56* (1992), 836–841.

[Jen61] JENNY H.: Derivation of state factor equation of soils and ecosystems. *Soil Science Society of America Procedings 25* (1961), 385–388.

[JLT05] JUANGA K., LEEB D., TENG Y.: Adaptive sampling based on the cumulative distribution function of order statistics to delineate heavy-metal contaminated soils using kriging. *Environmental Pollution 138*, 2 (2005), 268–277.

[Kas98] KASABOV N.: *Foundations of neural networks, fuzzy systems, and knowledge engineering*. MIT Press, Cambridge, Massachussets., 1998.

[MM03] MEUL M., MEIRVENNE M.: Kriging soil texture under different types of nonstationarity. *Geoderma 111238*, 3-4 (2003), 217–233.

[MSW98] MEULI R., SCHULIN R., WEBSTER R.: Experience with the replication of regional survey of soil pollution. *Environmental Pollution 101*, 3 (1998), 311–320.

[MV01] MORALES A., VILORIA J.: Aplicabilidad del enfoque de conjuntos borrosos a la clasificación de suelos de la depresión del lago de valencia, venezuela. *Interciencia 38* (2001), 598–604.

[MV04] MORALES A., VILORIA J.: *Clasificación borrosa de puntos del terreno y su relación con las propiedades de los suelos en la subcuenca del río Caramacate, estado Aragua*. Tech. rep., Facultad de Agronomía. Universidad Central de Venezuela., Maracay, Venezuela, 2004.

[PNB*09] PASTORE J., NOCERA M., BALLARIN V., MESCHINO G., ADRIANI J.: Segmentación de la cavidad endocraneana en imágenes de rmn para diagnóstico de hidrocefalia a presión normal (hpn). In *XVII Congreso Argentino de Bioingenieria* (2009).

[Que08] QUEZADA C.: Aplicación de la espectroscopía de reflectancia infrarojo cercano (nirs) en el análisis de suelos. *Ciencia Ahora 21* (2008), 75–85.

[SHS05] SULAEMAN Y., HIKMATULLAH, SUBAGYO H.: Modeling soil–landscape relationships. *Jurnal Ilmu Tanah dan Lingkungan 5*, 2 (2005), 1–14.

[SHZN08] SUN G., HOFF S., ZELLE B., NELSON M.: Development and comparison of backpropagation and generalized regression neural network models to predict diurnal and seasonal gas and pm10 concentrations and emissions from swine buildings. *Transactions of the ASABE 51*, 2 (2008), 685–694.

[SK10] SARMADIAN F., KESHAVARZI A.: Developing pedotransfer functions for estimating some soil properties using artificial neural network and multivariate regression approaches. *International Journal of Environmental and Earth Sciences 1* (2010), 31Ű–37.

[Spe91] SPECHT D.: A general regression neural network. *IEEE Transactions on Neural Networks 2*, 6 (1991), 588–576.

[Vil06] VILORIA J.: *Predicción espacial de propiedades del suelo por medio de modelos discretos y modelos continuos de variación*. Tech. rep., Facultad de Agronomía. Universidad Central de Venezuela., Maracay, Venezuela, 2006.

[Vil07] VILORIA A.: *Estimación de modelos de clasificación de paisaje y predicción de atributos de suelos a partir de imágenes satelitales y Modelos Digitales de Elevación*. Master's thesis, Facultad de Ciencias, Universidad Central de Venezuela, Caracas, Venezuela, 2007.

[ZBVD97] ZHU A., BAND L., VERTESSY R., DUTTON B.: Derivation of soil properties using a soil land inference model (solim). *Soil Science Society of America JOURNAL 61*, 2 (1997), 523–533.

[ZHBL01] ZHU A., HUDSON B., BURT J., LUBICH K.: Soil mapping using gis, expert knowledge, and fuzzy logic. *Soil Science Society of America JOURNAL 65* (2001), 1463–1472.

[ZQMB10] ZHU A., QI F., MOORE A., BURT J.: Prediction of soil properties using fuzzy membership values. *Geoderma 158* (2010), 199–206.