Artículo de Revisión

Métodos de análisis estadístico con datos dispersos

Statistical analysis methods with scattered data

Jaime Izquierdo Bautista
https://orcid.org/0000-0002-3464-4483
Doctorado en Planificación y manejo ambiental de cuencas hidrográficas.
Profesor Asociado de la Universidad Surcolombiana.
Email: jaimeizquierdo@usco.edu.co

Fecha de envío: 01/06/2020 *Fecha de revisión: 21/07/2020* *Fecha de aprobación: 10/11/2020*

DOI: 10.25054/22161325.2537

Resumen

En este artículo se presentan las principales características de algunas metodologías para derivar modelos de predicción o clasificación cuando se tienen muchos datos. Dentro de estos se describen las redes neuronales, bosques aleatorios, árboles de decisión, árboles altamente aleatorizados (extra trees), método de regresión asistida por patrón de contraste (CPXR), enfoque Bayesiano y análisis multivariado. Técnicas que han sido utilizadas para derivar modelos de predicción de retención de agua, clasificación de imágenes, clasificación de suelos, análisis en sicología educativa y social, al igual que la determinación de las principales variables que intervienen en un proceso.

Palabras clave: Redes neuronales; bosques aleatorios; árboles de decisión; árboles altamente aleatorizados (extra trees) y método de regresión asistida por patrón de contraste (CPXR).

Abstract

This paper presents the main features of some methodologies for deriving prediction or classification models when you have a lot of data. These include neural networks, random forests, decision trees, highly randomized trees (extra trees), contrast pattern-assisted regression (CPXR), Bayesian approach and multivariate analysis. Techniques that have been used to derive models of prediction of water retention, classification of images, classification of soils, analysis in educational and social psychology, as well as the determination of the main variables involved in a process.

Keywords: Neural networks; random forests; decision trees; highly randomized trees (extra trees) and contrast pattern assisted regression (CPXR) method.

1. Introducción

Cuando se quiere predecir variables con datos obtenidos de una experimentación o el acopio de muestras, se hace para facilitar la obtención de dicho valor. Lo anterior puede deberse a la dificultad para obtener el valor de la variable, la utilización de recursos económicos que son escasos, la imposibilidad de acceder a un laboratorio y el costo de los insumos que allí se utilizan, también se puede pensar en el personal que se necesita y el tiempo que se empleará para hallar la variable.

Generalmente se recurre a las técnicas de regresiones estadísticas, donde se puede obtener una o varias ecuaciones que permiten la predicción de las variables. Lo que sucede muchas veces, es que los datos no se ajustan a los modelos estadísticos conocidos, por lo tanto, hay que recurrir a técnicas de análisis de datos que permitan acceder a dichas variables, planteando soluciones con metodologías poco difundidas, pero igualmente efectivas (Pachepsky y Rawls 2004).

Los investigadores en este tipo de análisis, al no obtener ajustes buenos ante un modelo, retiran datos extremos, que causan desviaciones, hacen transformaciones de las variables teles como sacar raíz cuadrada, tomar el logaritmo de la variable, elevarlo a una constante, hacer el inverso de la variable, entre otros procedimientos buscando el mejor ajuste para derivar un modelo (Obiero, *et al.*, 2013).

El adelanto de los sistemas computacionales ha posibilitado el acceso a información y su procesamiento. En áreas como la agricultura, hidrología, ciencias de la tierra, ingeniería ambiental, ingeniería del petróleo etc. Se generan muchos datos que pueden predecir el comportamiento de algunas variables, haciendo más rápida la investigación o más económica la evaluación de datos que se necesitan para determinar la humedad del suelo como ejemplo. (Jain et al 2004, Minasny, *et al*., 1999)

La necesidad de encontrar las variables involucradas en algún fenómeno, puede llevar a que se utilicen infinidad de datos que tienen una gran precisión, pero con poca economía de consecución de datos. Se debe buscar que las ecuaciones predictoras sean de gran precisión, pero que además involucren el menor número de variables (Patil y Singh 2016, Schaap, *et al.*, 2001).

Tener datos de otras latitudes induce a errores, es indispensable realizar la colección de datos de las regiones donde se encuentran los desarrollos económicos, agrícola o de conservación ambiental. Datos como el contenido de materia orgánica, el contenido de humedad del suelo, la densidad aparente, son indispensables para la proyección de obras hidráulicas, producción agrícola, almacenamiento de agua, cambios en las condiciones de humedad, pero de lugares específicos y no extrapolarlos de estudios realizados en otros lugares (Patil, *et al.*, 2010).

Las investigaciones a nivel mundial se presentan principalmente en el hemisferio norte, Estados Unidos y Europa. Para las zonas tropicales como Colombia están en mora de hacer colección de datos de todo tipo, ambientales, de suelos, económicos, sociales para proyectar el desarrollo o identificar cuáles son las variables relevantes que influyen en un fenómeno. Con algunos datos que se colecten en campo, se puede inferir el comportamiento de las situaciones que se presentan en cierta región, la cual se puede monitorear por satélite, observando los cambios que se presentan en tiempo real y prediciendo los valores de variables relevantes (Minasny y Harteminck 2011, García y Medina 2005, Wosten, *et al.*, 2001).

Para el planteamiento de modelos predictivos de alguna variable se hace una planificación de la identificación de los datos, el planteamiento del problema y alguna depuración de la información colectada. Seguidamente se desarrollan los modelos y se observa su ajuste o nivel de predicción haciendo pruebas para determinar su desempeño con relación a los datos tomados de los casos reales. Finalmente se validan los datos calculando los resultados obtenidos con las condiciones reales que se dan. El problema viene cuando los modelos de regresión tradicionales

no explican muy bien los datos estimados, por lo tanto, hay la necesidad de utilizar otras metodologías que presenten mejor comportamiento ante los eventos reales.

En este documento se tratará de explicar la aplicación de algunas metodologías para el análisis de datos y la proposición de ecuaciones para la predicción de las variables. Entre estos métodos se encuentran las redes neuronales, manejo de datos en grupos, los árboles de regresión, el algoritmo del vecino más cercano y los árboles altamente aleatorizados conocidos como extra trees, el análisis multivariante y el enfoque bayesiano.

2. Materiales y métodos

2.1 Red Neuronal

Las redes neuronales fueron creadas para ofrecer solución a problemas no lineales, donde se tengan muchas variables y su solución se torne dificultosa. Una red neuronal artificial trata de imitar al cerebro al tener neuronas y estas interconectadas con otras hacen diversas tareas.

Son redes interconectadas masivamente en paralelo de elementos simples y en una jerarquía, interactuando en casos reales, con objetos reales, tal como lo hace el sistema nervioso central. En esta interacción aprende de la experiencia, encuentran relación con acciones realizadas anteriormente, obtienen características de los datos iniciales para ofrecer una respuesta a partir de estos, por lo anterior las redes neuronales presentan varias ventajas.

Entre las ventajas se puede mencionar el aprendizaje de tareas con un entrenamiento previo o experiencias iniciales; una red neuronal puede organizarse así misma creando su propia organización; se pueden presentar fallos durante la ejecución de la red neuronal con degradación de su estructura, pero por su bondad, puede retener algunas capacidades de la red; se pueden usar en tiempo real, con las tareas que se presentan, tomar decisiones de ejecución inmediata (Matich, 2001).

2.2 Componentes básicos de una red neuronal

Una red neuronal artificial puede estar compuesta básicamente por tres partes. Una capa inicial, que son las entradas o datos iniciales, una capa intermedia donde se realizan los procesos de cálculo y una salida o respuesta.

El grupo de neuronas en un mismo nivel forman una capa, las cuales se conectan con las capas adyacentes que pueden tener igual o diferente número de neuronas. La conexión de dos neuronas adyacentes en diferentes capas tiene una fuerza de conexión o peso (Schaap, *et al*., 1998).

De acuerdo con lo anterior los datos ingresan por una capa de entrada, pasan a una zona intermedia, llamada capa oculta, que puede estar constituidas por una o varias capas y finalmente salen por la capa de salida.

La capa de entrada está controlada por una función de entrada. Aquí las neuronas reciben los datos como si fueran homogéneos, la función posee un operador apropiado que junto a los pesos puede combinar estos datos. Los valores de entrada se multiplican por los pesos, cambiando los pesos de acuerdo con las influencias que estos ejercen, entonces un valor grande puede no tener gran influencia en el proceso debida a que el peso es pequeño.

2.2.1 Las funciones de entrada más comunes son:

Sumatoria de las entradas, consiste en sumar todos los valores de entrada y multiplicarlos por sus correspondientes pesos.

Producto de las entradas, es la multiplicación de todos los valores de entrada en la neurona, este producto es multiplicado por los pesos correspondientes.

Máximos de las entradas, aquí se analiza el mayor valor o más fuerte y se multiplica con su peso adecuado.

2.3 Tipos de redes neuronales

2.3.1 Redes unidireccionales (feedforward)

Son caracterizadas porque la información circula en único sentido, comenzando en las neuronas de entradas y siguiendo los caminos en la red, hasta alcanzar la salida.

2.3.2 Redes recurrentes (feedback)

En este tipo de redes la información puede fluir en cualquier dirección entre las capas que posee la red, incluso desde la salida hacia la entrada.

2.3.3 Redes auto asociativas

Cuando se presenta cierto estimulo o información de entrada la red hace una interpretación con relación al mismo patrón.

2.3.4 Redes hetero asociativas

Aquí la red se entrena para que ante un estímulo la respuesta de salida sea opuesta o diferente ante el patrón que se le presenta (Ortiz y Socha, 2005).

2.3.5 Redes neuronales como aproximación estadística

Las redes neuronales se desarrollan mediante un esquema computacional, utilizando muchas de las funciones clásicas de la estadística. Para alcanzar lo anterior simplemente se varían el número de nodos ocultos y las funciones de activación. Frente a los modelos estadísticos clásicos presenta algunas ventajas significativas, tales como la implementación mucho más flexible, no necesita de cumplir supuestos paramétricos estadísticos (normalidad, independencia, linealidad, etc.) y la extensión más sencilla a casos multivariados (García, 2005).

La red neuronal artificial más utilizada en el ámbito científico es la feedforwar, lo anterior basado en estudios en la configuración de dos capas que contengan suficientes neuronas se puede llegar a una función continua con un grado de precisión arbitrario (Cybenko, 1989, Vásquez, 2014). También se sabe que cada día se buscan soluciones a problemas complejos de predicción e identificación de patrones, con las redes multicapa como las feedforward presentan similitud en el análisis generalizado de las regresiones (Warner y Misra, 1996, Vásquez, 2014).

2.4 Bosques aleatorios

Los bosques aleatorios o Random Forest es una técnica de clasificación de datos a través de árboles de decisión. Estos modelos estadísticos son para trabajar cuando se tienen gran cantidad de datos y muchas variables. Del total de datos que se tienen se pueden crear árboles que tratan conjuntos de datos o sub muestras más homogéneos que se procesan en cada árbol.

Esta técnica hace una clasificación supervisada de los datos y tienen una organización que se asemeja a la de árbol con raíz, ramas y hojas. Cada una de sus partes que generalmente son representadas por un círculo, son los nodos, los cuales se conectan con otros nodos. El nodo inicial es la raíz, desde el cual se extienden las ramas hasta llegar a los extremos de la cadena donde se encuentran las hojas (Medina y Ñique, 2017).

Cuando se tienen arboles predictivos, pero con clasificadores débiles, se trabaja en conjunto, con los árboles interconectados, sus resultados se promedian para dar una respuesta a su clasificación.

Algunas ventajas de los árboles es que se asemejan a la forma intuitiva en que los humanos clasifican y predicen el comportamiento de un sistema, además su forma es relativamente fácil de interpretar. Como se comporta sin parámetros establecidos no tiene que cumplir ninguna distribución específica. La preparación de los datos es mucho menos exigente que otros métodos de aprendizaje estadístico debido a que no se ve muy influenciado por datos

atípicos. Para la exploración de datos permite una identificación rápida de las variables influyentes en una predicción. Si la predicción no llega hasta el nodo final, la información obtenida hasta el nivel que llega se puede interpretar (Rodrigo, 2017).

Algunas desventajas que presenta esta técnica de predicción se presentan cuando se tiene un solo árbol de decisión, pues su capacidad predictiva se ve disminuida frente a otros modelos. Cuando se trabaja con variables continuas es posible que se pierda parte de la precisión porque trata de agrupar los datos en conjuntos (Rodrigo, 2017).

2.5 Árboles de decisión

Los árboles de decisión son una herramienta poderosa y sencilla, que se implementa fácilmente. Se puede tomar como un modelo de predicción, tarea que hace aprendiendo a partir de las observaciones de la realidad. Parte una serie de objetos que tienen a su vez una serie de atributos. Los atributos caracterizan al objeto y toma una serie de valores que se pueden excluir unos a otros, se excluyen, por lo tanto, en este punto la decisión sigue un único camino.

Como los árboles principalmente lo que hacen es clasificaciones de grupos, esto inicia en el nodo principal, desde allí se va extendiendo a nodos secundarios donde se responden preguntas del atributo, que pueden ser valores o características. Al final del proceso se llega a una decisión que corresponde a una variable del problema planteado (Barrientos, *et al.,* 2009).

Si el árbol es creado para tomar una decisión esta puede ser positiva o negativa. El árbol puede ser entrenado para la toma de decisiones, con una serie de atributos de los objetos puede tomar varias rutas y con suficiente información llegar a una respuesta correcta. Sin embargo, en la medida que se incrementan los atributos y sus características su tamaño puede crecer exponencialmente, haciéndose un poco más difícil de comprender.

Con varios conjuntos de datos se puede usar una parte para entrenamiento y otra para validar las salidas que está calculando el árbol. Con lo anterior se pueden hacer pruebas estadísticas y determinar que tan buen desempeño tiene el árbol en las decisiones o clasificaciones que arroja.

Cuando los atributos se van haciendo grandes, se puede partir las tareas para hacer el árbol más comprensible. Al final el clasificará los objetos de acuerdo con los atributos propuestos en subsecciones para unirlos al gran árbol (Barrientos, *et al.,* 2009, Quinlan 1986, 1993).

Los árboles de decisión se construyen por intermedio de algoritmos, los cuales están divididos en dos partes. La primera se llama inducción y se hace con datos de entrenamiento, que generalmente corresponde a una buena proporción de los datos observados en la realidad. Como se parte del nodo raíz, el árbol crece de acuerdo con las necesidades, pues depende de las características que tiene en cada nodo, si estas pertenecen a más de dos clases de clasificación, se genera un nuevo nodo, pero si es simple allí termina y se tiene una respuesta.

En la etapa de clasificación o validación de los datos, los datos nuevos son clasificados de acuerdo con la red del árbol que se construyó previamente, siguiendo los caminos y asignándole alguna clase o generando una respuesta a las variables de entrada.

Algunos algoritmos que se utilizan usualmente en los árboles de decisión sin ID3, J48, Naive Bayes. Estos son sencillos, de rápida ejecución con bajo consumo de máquina y son precisos. El ID3 (Quinlan, 1993) utiliza el total de los datos inicialmente y posteriormente los va dividiendo hasta encontrar homogeneidad en ellos. Con esto se logra obtener el mejor atributo que agrupe los datos en clases parecidas. J48 (Quinlan, 1993) hace un proceso iterativo agregando nodos o ramas y determinando la menor diferencia entre los datos, así de esta manera como de ensayo y error va encontrando un camino que tenga el menor error entre los datos observados y los calculados o clasificados por el árbol.

El algoritmo Naive Bayes tiene previamente un clasificado bayesiano, que clasifica cada clase haciendo que los atributos de esta sean independientes de otras, solo se usan esos atributos para determinada clase. Lo anterior hace que los caminos sean fijos y solo se aprenden los parámetros. (Dunham, 2003)

2.6 Extra Trees – Arboles Altamente Aleatorizados

Es una variante de los árboles de decisión, pues lleva mucho más allá la decorrelación en cada nodo, haciendo que no se presente dependencia entre los datos que se están analizando. En cada nodo, evalúa solo un subconjunto de los predictores y de estos elije un grupo de puntos que corte el eje. Lo anterior hace que el valor de la varianza se haga más pequeño (Rodrigo, 2017).

Este método se desempeña bien con valores que presentan una tendencia no lineal, además de ser flexibles y escalabilidad que otras metodologías como las redes neuronales no lo expresan bien. El proceso es progresivo entre los diferentes modelos que se van evaluando, pero esto es apoyado en la validación cruzada, donde se escogen grupos de datos para el entrenamiento y se comparan con el resto en la validación, estos grupos seleccionados se rotan hasta copar todo el grupo de muestras que se tengan.

El proceso de los árboles aleatorios se hace con tres pasos básicos. El primero consiste en hacer una clasificación de los datos de entrada utilizado alguna medida estadística. De estos datos de entrada se tienen unas respuestas, las cuales se reservan las de mejor respuesta. En un segundo paso, los datos obtenidos anteriormente se comparan con los observados y se dejan los de mejor desempeño. En el paso final se identifica el modelo que mejor representa los datos calculados con respecto a los observados, el proceso se repite hasta que una medida estadística evalúe que la diferencia sea la menor (Galelli y Castelletti, 2013).

2.7 Método de regresión asistida por patrón de contraste (CPXR)

Este método ha sido difundido por Taslimitehrani y Dong (2014), Dong y Taslimitehrani (2015) se presenta como una alternativa para derivar funciones de edafotransferencia con robustez para modelos de predicción. Se presenta como un método alternativo donde otros modelos no presentan una respuesta aceptada de predicción. De acuerdo con Ghanbarian, *et al*., (2015) los modelos desarrollados por este método son más sencillos, comprensibles, comparados con los modelos de regresión lineal y las redes neuronales. Las variables utilizadas en los diferentes patrones no están restringidas a un solo camino y por mínimo que sean los datos que se tienen siempre trata de ponderar los resultados.

Este método busca tener conjunto de datos que se ajusten a ciertos rangos de datos, que tengan un patrón conocido. Con este grupo se puede hacer modelos de regresión local para grupo de datos coincidentes.

Cuando una muestra de datos se ajusta a un patrón determinado se le aplica la función lineal local para dicho grupo. Cuando los datos muestran coincidencia con varios grupos de datos, entonces se aplican las funciones lineales y se ponderan para dar un solo valor de respuesta, sin embargo, cuando algunos datos no se ajustan a ningún patrón, se tiene una función auxiliar que se les puede aplicar.

Ghanbarian, *et al*., (2015) presenta 5 pasos para calcular los patrones y llevar el error de predicción al mínimo.

1. Del total del conjunto de datos que se poseen, se puede hacer una regresión lineal múltiple, con los datos de entrenamiento, el cual lo llama *f0*.

2. Del total de los datos de entrenamiento se dividen en dos grupos, el primero llamado de grandes errores y el segundo de pequeños errores. Para lograr lo anterior se toma un valor c, de tal manera que los valores del error de predicción f0 que superen a c son los grandes errores. Ghanbarian, *et al*., (2015) afirma que la suma del error de predicción absoluto alcanza el 45% del total de la suma de los errores absolutos. Lo que indica que el resto son los errores pequeños de los datos del entrenamiento.

3. Para definir los grupos en los cuales van a estar los límites de aplicabilidad de las variables, así como discretizar las variables de entrada se utiliza un modelo de entropía propuesto por Fayyad y Irani (1993). Luego el modelo de regresión por contraste de patrones clasifica las variables de acuerdo con el patrón al cual ajustan. Como es posible que algunos subgrupos sean bastante parecidos se pueden utilizar filtros para que se asignen a un solo grupo de ajuste.

4. Con los datos agrupados se procede a realizar una regresión múltiple local, para cada grupo. Los modelos de regresión local que no mejoren la predicción de todo el modelo se pueden eliminar.

5. Al final se aplica un doble proceso de optimización con el fin de hallar los mejores patrones, funciones y pesos que mejor representan el modelo de predicción. Así se van aplicando los patrones a los conjuntos de datos clasificados anteriormente, el proceso termina cuando al probar las funciones en los conjuntos la precisión del modelo no mejora.

2.8 Enfoque Bayesiano

El análisis Bayesiano no es nuevo, sin embargo, en épocas actuales ha vuelto a ser relevante, se basa en tener estudios anteriores sobre un tema específico, utilizados como base para probar las hipótesis que se plantean. Con los datos anteriores y los actuales se modifica el comportamiento o no de una o varias variables. (Rendón, *et al*., 2018)

Bayes se base en que los datos u observaciones que se toman no son del todo probalísticos o inciertos, sino que se va aprendiendo a través de la experiencia, entre más datos se toman u observaciones se hacen, se supone se acerca a una certeza mayor. Lo anterior se puede expresar mediante una ecuación sencilla (Ecuación 1).

$$P(H/D) = \frac{P(D/H)\,P(H)}{P(D)} \quad (1)$$

Donde P (H/D) es el resultado de haber utilizado los datos anteriores; P (D/H) son los datos del análisis y representa la posibilidad de que una hipótesis H dados los datos D va a tener una proporcionalidad a la probabilidad de hallar D conociendo a H de antemano; P (H) es el conocimiento previo que se tiene de los datos u observaciones, lo que se sabe sobre los valores que puede tomar la hipótesis; P (D) son los promedios de las observaciones probables de D sobre las posibles hipótesis H.

Su utilidad se base en que las investigaciones generalmente buscan es la probabilidad de las causas dados unos efectos y no al contrario. Con el uso del teorema de Bayes no se garantiza una respuesta correcta, pues el mundo está lleno de incertidumbre, sin embargo, de una gama de posibilidades permite escoger cual podría ser la respuesta más acertada. (Etz, *et al*., 2018)

2**.9 Análisis multivariado**

El análisis multivariado de datos se hace cuando las investigaciones quieren saber la relación simultánea entre tres o más variables, de igual manera, en la medida que se realiza una investigación se pueden ir analizando los datos u observaciones obtenidas, lo que puede permitir que se eliminen variables redundantes o que no tienen relevancia en los procesos, además el desarrollo tecnológico de los sistemas computacionales permite manjar mayor volumen de datos y sus relaciones para sacar conclusiones. (Dillon y Goldstein, 1984)

Esta técnica de análisis de datos se creó debido a que los métodos univariantes y bivariantes no los cubrían. También ayuda al investigador en la toma de decisiones, aunque no se cumplan estrictamente las hipótesis de normalidad y homocedasticidad multivariantes, sin embargo, hay que fijarse en los resultados y la relación que tienen con dichas hipótesis, pues con suficientes datos el método se muestra robusto y resiste las desviaciones de los supuestos estadísticos.

El análisis multivariante se puede agrupar en dos grupos, el primero denominado técnicas explicativas o de dependencia, aquí se busca establecer como dos grupos de variables, uno dependiente y el otro independiente, se relacionan entre sí, ya sea en conjunto o con algunas variables en particular. En el segundo grupo llamado técnica descriptiva o de interdepencia, como en algunos casos no es fácil distinguir cuales son las variables dependientes y las independientes se utiliza esta técnica. Entonces el análisis busca determinar cómo se relacionan las variables, que las está relacionando y porqué lo están. (Uriel y Aldás, 2005, Benzécri 1973)

Hair *et al.,* (1995), propone algunos pasos para la correcta aplicación de esta técnica y poder llegar a unas conclusiones confiables. Lo primero es definir claramente el problema, cuáles son los objetivos generales y específicos, en lo posible establecer las relaciones entre las diferentes variables que se involucran. Segundo es aplicar la técnica elegida, se debe tener el suficiente número de datos y su característica, si son numérico o de apreciación. Tercero es que la técnica cumpla con los parámetros estadísticos, pues si esto no se cumple se debe tratar de probar con otra técnica que obtenga mejores ajustes o eliminar variables que redunden o no sean relevantes. Al final hay que interpretar los resultados que pueden conducir a nuevas hipótesis y validar el modelo de tal manera que se pueda aplicar en otros lugares. (Closas, *et al*., 2013)

3. Resultados y discusión

Para determinar si los resultados calculados por las metodologías propuestas por los investigadores se ajustan a los resultados encontrados en campo y en laboratorio se verifica su ajuste calculando el error cuadrático medio (MSE), la raíz del error cuadrático medio (RMSE) y la estimación o subestimación de los valores medios de cada variable calculada (Nemes, *et al*., 2003). Su ecuación se expresa de la siguiente manera:

$$ME = \frac{1}{n}\sum_{i=1}^{n}(Y_{estimado} - Y_{observado}) \quad (2)$$

El error cuadrático medio (MSE) (Ecuación 2), indica el ajuste de los datos calculados con los esperados, es así que lo ideal sería que este fuera cero, pero en estos casos con valores cercanos a cero indica un buen desempeño del modelo.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_{observado} - Y_{calculado})^2 \quad (3)$$

Donde Y_{obser} representa los datos observados o medidos y Y_{calc} los datos calculados por el modelo, n representa las parejas de datos comparados o diferenciados (Ecuación 4).

La raíz del error cuadrático medio (RMSE) (Ecuación 4), indica la dispersión de los datos, entre los encontrados directos en laboratorio o campo y los calculados de los modelos propuestos por los investigadores (Jana and Mohanty, 2011; Patil *et al.,* 2011; Vereecken, *et al.,* 2010; Tomasella, *et al.,* 2003).

$$\text{RMSE} = \left[\frac{1}{n}\sum_{i=1}^{n}(Z_{0-} Z_p)^2\right]^{1/2} (4)$$

El coeficiente de determinación (R2) (Ecuación 5), que representa una relación entre la covarianza de las variables y las desviaciones típicas de cada variable, tiende a uno para un ajuste lineal perfecto, así que entre más cercano a la unidad representa un ajuste mucho mejor, una explicación mejor entre las variables independientes y dependientes. (Stumpp, *et al*., 2009, Min-Fei, *et al*., 2020):

$$R^2 = \frac{\sum_{i=1}^{n}(Y_{estimado} - Y_{medio})^2}{\sum_{i=1}^{n}(Y - Y_{medio})^2} \quad (5)$$

3.1 Discusión

Los computadores dan gran ventaja con estos métodos. Lo anterior ofrece unas nuevas herramientas que permiten encontrar modelos que no están o no caben en la estadística tradicional, haciendo que se generen ecuaciones o se definan variables relevantes de un número de datos que puede ser muy grande (Wehenket, *et al.,* 2006).

Como estos son modelos de aprendizaje que requieren datos para su entrenamiento y validación, le da una nueva dimensión al tratamiento de estos. Se pueden usar parte de los datos para evaluar su evolución o tomar todo el conjunto y analizarlo enfrentando problemas estocásticos o no lineales (Geurts, *et al.,* 2006; Wehenket, *et al.,* 2006).

Actualmente con la disponibilidad de sensores remotos se pueden tener variables en tiempo real, lo cual hace que se pueden tener muchos datos en un corto periodo de tiempo. Para el procesamiento de estos datos se pueden usar las redes neuronales, los árboles de decisión o los árboles altamente aleatorizados. (De León Mata, *et al*., 2014, Llamas, *et al*., 2013)

Estos métodos combinados con técnicas como los sistemas de información geográfica pueden hacer clasificación de características similares para suelos con variables como la altitud, la latitud, la cobertura, la pendiente o la curvatura del terreno que junto con datos tomados de campo pueden inferir otras características físicas y químicas (Minasny y Hartemink 2011, Jalmacín, *et al*., 2017, Kebede, *et al*., 2021).

Un aspecto muy importante de estos métodos es que no son paramétricos, por lo tanto, no se deben ajustar a ninguna distribución específica. Estos simplemente con sus algoritmos buscan el modelo que mejor se ajuste y a partir de este demuestran que tienen los mejores resultados. Sin embargo, se han encontrado diferencias al comparar los modelos, que analizan los mismos datos, pero con algunos se obtienen mejores desempeños comparados con otros, por ejemplo, al comprar los coeficientes de determinación (Mao y Wang 2012, Rodrigo, 2017, Nicolau, 2017, Castillo-Paez, *et al.,* 2019, Kalli 2020).

4. Conclusiones

Con los avances tecnológicos de diferentes procesos como son características de suelos, clima, comportamientos sociales, producen muchos datos que no son fáciles de analizar. Con los datos anteriores se pueden plantear modelos de comportamiento o de predicción de las variables involucradas. Estos modelos pueden ser paramétricos y se tienen que ajustar a sus valores establecidos, pues fuera de estos parámetros se puede estar subestimando o sobrestimando.

Los modelos como las redes neuronales, los árboles aleatorios o los árboles altamente aleatorizados permiten tomar todo el conjunto de datos, que, a partir de un algoritmo, aprender como es el comportamiento de los datos y sin atender ninguna restricción se pueden plantear varios modelos que cumplan con los parámetros escogidos. Al final se puede evaluar el desempeño de los diferentes modelos para escoger el mejor, que además puede involucrar el menor número de variables, haciendo el modelo económico o parsimonioso.

Con la selección del menor número de variables en un fenómeno o proceso estudiado se direccionan mejor los recursos para la determinación de las variables relevantes. Esto puede ayudar a los investigadores a aprovechar mejor sus recursos económicos, técnicos, humanos.

5. Referencias bibliográficas

Barrientos MRE., Cruz RN., Acosta MHG., Rabatte SI., Gogeascoechea TMC., Pavón LP., Blázquez Morales SL., 2009. Árboles de decisión como herramienta en el diagnóstico médico. Revista Médica de la Universidad Veracruzana. N. 2.

Benzécri, J. L., 1973. L'Analyse des Données. París: Duno

Castillo-Páez, S., Fernández-Casal, R., García-Soidán, P., 2019. A nonparametric bootstrap method for spatial data, Computational Statistics & Data Analysis, Volume 137, Pages 1-15, https://doi.org/10.1016/j.csda.2019.01.017.

Closas, A., Arriola, E., Kuc, C. I., Amarilla, M. R., Jovanovich, E., 2013. Análisis multivariante, conceptos y aplicaciones en Psicología Educativa y Psicometría. Efoques XXV. P. 65 – 92.

Cybenko, George., 1989. Approximation by Superpositions of a Sigmoidal Function. Mathematics of control, signals, and systems, 2, 303–314. https://doi.org/10.1007/bf02551274

De León Mata, G., Pinedo Álvarez, A., Martínez Guerrero, J., 2014. Aplicación de sensores remotos en el análisis de la fragmentación del paisaje en Cuchillas de la Zarca, México, Investigaciones Geográficas, Boletín del Instituto de Geografía, Volume 2014, Issue 84, Pages 42-53, https://doi.org/10.14350/rig.36568.

Dillon, W. R. y M. Goldstein., 1984. Multivariate Analysis. Methods and Applications. New York: Wiley & Sons.

Dong,G. & Taslimitehrani, V. Pattern-aided regression modeling and prediction model analysis. IEEE Transactions on Knowledge and Data Engineering. P. 2452-2465. 2015. https://doi.org/10.1109/tkde.2015.2411609

Dunham, M., 2003. Data Mining Introductory and Advanced Topics. Prentice Hall.

Etz A, Gronau QF, Dablander F, Edelsbrunner PA, Baribault B., 2018 How to become a Bayesian in eight easy steps: an annotated reading list. Psychon Bull Rev. ; 25(1):219-234. DOI: 10.3758/s13423-017- 1317-5

Fayyad, U.M. and Irani, K.B., 1993. Multi-Interval Discretization of Continuous-Valued Attributes for Classification Learning. Proceedings of the 13th International Joint Conference on Artificial Intelligence (IJCAI-93), Chambèry, 28 August-3 September p1022-1027.

Galelli, S. & Castelletti, A. Tree based iterative input variable selection hydrological modeling. Water Resourses Research. Vol 49 16 p. 2013. https://doi.org/10.1002/wrcr.20339

Garcia, J. & Medina, H., 2005. Una revisión sobre las funciones de pedotransferencia en la determinación de las propiedades hidráulicas del suelo. Revista Ciencias Técnicas Agropecuarias. Vol 18 N. 3 https://doi.org/10.35537/10915/39548

Geurts, P., Ernst, D., y Wehenkel, L. (2006). Extremely Randomized Trees. Mach Learn. Springer Science and Business Media, Inc. DOI 10.1007/s10994-006-6226-1

Ghanbarian, B., Taslimitehrani, V., Dong, G., & Pachepsky, Y. A., 2015. Sample dimensions effect on prediction of soil water retention curve and saturated hydraulic conductivity.Journal of Hydrology,528, 127-137. https://doi.org/10.1016/j.jhydrol.2015.06.024

Hair, J. F., Anderson, R. E., Tatham, R. L. y Black, W., 1995. Multivariate Data Analysis. Englewood Cliffs, NJ: Prentice-Hall, 1995, 4ª ed.

Jain, S. K., Singh, V. P. and Van Genuchten., M. Th. 2004. Analysis of soil water retention data using artificial neural networks. J. Hydro. Engg., 9: 415-420.

Jalmacin-Nené-Preciado, A., González G., Mendoza, M., Silva Bátiz, M., 2017. Cambio de cobertura y uso de suelo en cuencas tropicales costeras del Pacífico central mexicano, Investigaciones Geográficas, Boletín del Instituto de Geografía, Volume 2017, Issue 94, Pages 64-81, https://doi.org/10.14350/rig.56770.

Jana R B, & Mohanty B P. 2011. Enhancing PTFs with remotely sensed data for multi-scale soil water retention estimation. J Hydrol. 399: 201–211. https://doi.org/10.1016/j.jhydrol.2010.12.043.

Kalli, M., 2020.Chapter 4- Bayesian nonparametric methods for financial and macroeconomic time series analysis, Editor(s): Yanan Fan, David Nott, Michael S. Smith, Jean-Luc Dortet-Bernadet, Flexible Bayesian Regression Modelling, Academic Press, Pages 91-119, https://doi.org/10.1016/B978-0-12-815862-3.00012-3.

Kebede, Y., Endalamaw, N., Sinshaw, B., Atinkut, H., 2021. Modeling soil erosion using RUSLE and GIS at watershed level in the upper beles, Ethiopia, Environmental Challenges, Volume 2, https://doi.org/10.1016/j.envc.2020.100009.

Llamas, R., Bonifaz, R., Valdés, M., Riveros-Rosas, D., LeyvaContreras, A., 2013. Spatial and temporal variations of atmospheric aerosol optical thickness in northwestern Mexico, Geofísica Internacional, Volume 52, Issue 4, Pages 321-341, https://doi.org/10.1016/S0016-7169(13)71480-6.

Mao W and Wang F., 2012. Advances in intelligence and security informatics. Elsevier. Oxford UK. 103 p.

Matich, D. J., 2001. Redes Neuronales: Conceptos Básicos y Aplicaciones. Libro de clase. Informatica aplicada la ingeniería de procesos. Universidad Tecnológica Nacional facultad Regional Rosario Argentina.

Medina, R. & Ñique, C., 2017. Bosques aleatorios como extensión de los árboles de clasificación con los programas R y Python. Interfasses. Ed N. 10. P 165-189. https://doi.org/10.26439/interfases2017.n10.1775

Min-fei Sun, Jing-yi Yang, Wen Cao, Jing-yuan Shao, Guo-xiang Wang, Hai-bin Qu, Wen-hua Huang, Xing-chu Gong, 2020. Critical process parameter identification of manufacturing processes of Astragali Radix extract with a weighted determination coefficient method, Chinese Herbal Medicines, Volume 12, Issue 2, Pages 125-132,. https://doi.org/10.1016/j.chmed.2019.11.001.

Minasny, B., McBratney, A.B. and Bristow, K.L. 1999. Comparison of different approaches to the development of pedotransfer functions for water retention curves. Geoderma, 93: 225-253.

Minasny B., & Hartemink A E., 2011. Predicting soil properties in the tropics. Earth-Sci Rev. 106: 52–62. https://doi.org/10.1016/j.earscirev.2011.01.005

Nemes A, Schaap M G, & W¨osten J H M. 2003. Functional evaluation of edofotransfer functions derived from different scales of data collection. Soil Sci Soc Am J. 67: 1093–1102. https://doi.org/10.2136/sssaj2003.1093.

Nicolau, J., 2017. A simple nonparametric method to estimate the expected time to cross a threshold, Statistics & Probability Letters, Volume 123, Pages 146-152, https://doi.org/10.1016/j.spl.2016.12.011.

Obiero, J., Gumbe, L., Omuto, C., Hassan, M., and Agullo, J., 2013. "Development of Pedotransfer Functions for Saturated Hydraulic Conductivity," Open Journal of Modern Hydrology, Vol. 3 No. 3, pp. 154-164. https://doi.org/10.4236/ojmh.2013.33019

Ortiz, A. & Socha D., 2005. Aplicación de las redes neuronales MLP a la predicción de un paso en series de tiempo. Proyecto de grado. Fundación universitaria Konrad Lorenz. Facultad de ingeniería de sistemas. https://doi.org/10.4995/thesis/10251/64909

Pachepsky, Y., & Rawls, W.J., 2004. Development of pedotransfer functions in soil hydrology. Developments in Soil Science Vol. 30. Editors book. Elsevier Science. https://doi.org/10.1016/s0166-2481(04)30023-1

Patil, N G., Rajput, G S., Nema, R K., Singh, R B., 2010. Predicting hydraulic properties of seasonally impounded soils. J Agr Sci Cambridge. 148: 159–170. https://doi.org/10.1017/s002185960999030x

Patil, N G., Pal, D K., Mandal, C., Mandal, D K., 2011. Soil water retention characteristics of Vertisols and pedotransfer functions based on nearest neighbor and neural networks approach to estimate AWC. J Irrig Drain Eng. 138: 177–184. https://doi.org/10.1061/(asce)ir.1943-4774.0000375

Patil, N G., Singh, S K., 2016. Pedotransfer functions for estimating soil hydraulic properties: A review. Pedosphere. 26(4): 417–430. https://doi.org/10.1016/s1002-0160(15)60054-6

Quinlan, J. R., March 1986 Induction of decision trees. Mach. Learn., 1(1):81–106.

Quinlan, J. R., 1993. C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

Rendón-Macías ME, Riojas-Garza A, Contreras-Estrada D, Martínez-Ezquerro JD, 2018. Análisis bayesiano. Conceptos básicos y prácticos para su interpretación y uso. Rev Alerg Mex. 65(3):285-298. DOI: 10.29262/ram.v65i3.512

Rodrigo, J. A., 2017. Árboles de predicción: bagging, random forest, boosting y C 5.0. Estadística con R. p 97.

Schaap, M. G., Leij, F. J., 1998. Database related accuracy and uncertainty of pedotransfer funtions. Soil Science 163. p 765 -–769. https://doi.org/10.1097/00010694-199810000-00001

Schaap, M.G., Leij, F. J., & Van Genuchten, M. T. 2001. ROSETTA: acomputer program for estimating soil hydraulic parameters with hierarchical pedotransfer functions. Journal of Hydrology, 251, 163–176. https://doi.org/10.1016/s0022-1694(01)00466-8

Stumpp, C., Engelhardt, S., Hofmann, M., & Huwe, B., 2009. Evaluation of pedotransfer functions for estimating soil hydraulic properties of prevalent soils in a catchment of the Bavarian Alps. Eur J For Res. 128: 609–620. https://doi.org/10.1007/s10342-008-0241-7

Taslimitehrani,, V., Guozhu Dong., 2014. A new CPXR based logistic regression method and clinical prognostic modeling results using the method on traumatic brain injury. IEEE International Conference on Bioinformatics and Bioengineering. P 283-290. https://doi.org/10.1109/bibe.2014.16

Tomasella, J., Pachepsky, Y., Crestana, S., & Rawls, W J., 2003. Comparison of two techniques to develop pedotransfer functions for water retention. Soil Sci Soc Am J. 67: 1085–1092. https://doi.org/10.2136/sssaj2003.1085

Uriel, E. y Aldás, J., 2005. Análisis Multivariante Aplicado. Madrid: Thomson.

Vásquez,, J. P., 2014. Red Neuronal feedforward como estimador de patrones de corrientes en el interior del puerto de Manzanillo sujeto a la acción de tsunamis. Instituto Mexicano del Transporte. Publicación técnica 406. 63 p.

Vereecken, H., Weynants, M., Javaux, M., Pachepsky, Y., Schaap, M G., & van Genuchten, M T., 2010. Using pedotransfer functions to estimate the van Genuchten-Mualem soil hydraulic properties: a review. Vadose Zone J. 9: 795–820. https://doi.org/10.2136/vzj2010.0045

Warner, Brad, & Misra, Manavendra. 1996. Understanding Neural Networks as Statistical Tools. The american statistician, 50(4), 284–293. https://doi.org/10.1080/00031305.1996.10473554

Wehenket, L,. Ernst, D., and Geurts, P., 2006. Ensembles of extremely randomized trees and some generic applications. RTE-VT workshop, paris.

Wosten, J H M., Pachepsky, Y A., Rawls, W J., 2001. Pedotransfer functions: bridging the gap between available basic soil data and missing soil hydraulic characteristics. J Hydrol. 251: 123–150. https://doi.org/10.1016/s0022-1694(01)00464-4