Publicaciones en Ciencias y Tecnología
Vol 9, N^{o}2, Jul–Dic 2015, pp.139–158, ISSN:1856-8890, EISSN: 2477-9660

PREDICCIÓN DEL RENDIMIENTO DE UN CULTIVO DE NARANJA "VALENCIA" MEDIANTE REDES NEURONALES DE REGRESIÓN GENERALIZADA

[*] Edwin Hernández-Caraballo

Recibido: 05/10/2015 Aprobado: 09/05/2016

Resumen

El rendimiento de un cultivo es el resultado de múltiples variables, cuyas complejas interacciones hacen que sea difícil de predecir por medios convencionales. Las redes neuronales de regresión generalizada constituyen una alternativa prometedora para dicha tarea, gracias a su capacidad para modelar relaciones desconocidas, y de naturaleza no lineal. Este trabajo tuvo como propósito evaluar dicha aproximación, en la predicción del rendimiento de un cultivo de naranja "Valencia" (Citrus sinensis L. Osbeck), a partir de una base de datos real, contentiva de los rendimientos de 78 árboles de naranja, y las concentraciones de nitrógeno, fósforo, potasio, calcio, y magnesio, en el tejido foliar. Especial énfasis se hizo en el tratamiento de los datos de entrada/salida, utilizando técnicas convencionales (normalización, estandarización, y componentes principales), y otras no convencionales (cálculos de los log-cociente centrado, e índices nutricionales individuales y globales, a partir del Sistema de Diagnóstico de la Composición Nutricional). Los resultados mostraron que la combinación índices nutricionales individuales/rendimiento normalizado (Error de predicción= $0{,}98\, kg \cdot árbol^{-1}$), y componentes principales no rotados/rendimiento normalizado (Error de predicción= $0{,}51\, kg \cdot árbol^{-1}$) condujeron al desarrollo de las redes neuronales con las mejores capacidades de predicción del rendimiento, evidenciadas por los menores errores de predicción anteriormente indicados.

Palabras clave: Citrus sinensis; rendimiento; composición foliar; sistema de diagnóstico de la composición nutricional.

[*] *Universidad Centroccidental "Lisandro Alvarado", Decanato de Agronomía, Barquisimeto, Venezuela, ehernandez@ucla.edu.ve Doctor en Química Analítica*

FORECASTING THE YIELD OF A "VALENCIA" ORANGE'S CROP, BY MEANS OF A GENERALIZED-REGRESSION NEURAL NETWORKS

Abstract

The yield of a given crop is the result of multiple variables whose complex interactions make its prediction difficult to achieve by regular means. Generalized regression artificial neural networks represent a promising alternative for such a task, due to its ability to model non-linear relationships, without the need of knowing its explicit nature. The present work aimed at assessing such approximation for predicting the potential yield of a crop of 'Valencia' orange (Citrus sinensis L. Osbeck), using a real database containing the yield of 78 orange trees, and the concentration of nitrogen, phosphorus, potassium, calcium, and magnesium in their foliar tissue. Special emphasis was placed in the mathematical treatment of the input/output data, resorting to conventional (normalization, standardization, and principal components) as well as other less common techniques (row-centered log ratios, and individual and global nutritional indices from the Compositional Nutrient Diagnosis System). The results showed that the individual nutrient indices/normalized yield combination (Prediction error= $0{,}98\,kg \cdot \text{tree}^{-1}$), and the unrotated principal components/normalized yield combination (Prediction error= $0{,}51\,kg \cdot \text{tree}^{-1}$) resulted in the development of the neural networks with the highest yield prediction capabilities, as evidenced by the previously indicated prediction errors.

Keywords: Citrus sinensis; yield; foliar composition; compositional nutrient diagnosis system.

Introducción

La naranja "Valencia" (Citrus sinensis L. Osbeck) representa, junto a las musáceas, el grupo de frutas más importante -en términos de producción- en Venezuela. Sin embargo, en términos del mercado mundial, el rendimiento de dicho cultivo en nuestro país se encuentran muy por debajo de aquellos de los principales productores de América (Food y Organization, 2008).

Ciertamente, el rendimiento de un cultivo es el resultado de una gran cantidad de variables (Wallace y Wallace, 1993), el conocimiento de las cuales permitiría su optimización. No obstante, hay que destacar que modelar las relaciones entre dichas variables es una tarea que puede ser muy difícil, debido a la complejidad (o al desconoci-

miento de la totalidad) de las mismas. Lo anterior es particularmente cierto, si ello se hace a través de modelos matemáticos convencionales. Esto no ha impedido la evaluación con varios grados de éxito de diversos modelos predictivos, por ejemplo, a partir de datos climáticos (Fornaciari, Orlandi, y Romano, 2005), propiedades de los suelos (Kitchen, Drummond, Lund, Sudduth, y Buchleiter, 2003), índices nutricionales (Arizaleta, Rodríguez, y Rodríguez, 2002), datos morfológicos (Mourtzinis, Arriaga, Balkcom, y Ortiz, 2013), entre otros.

Las redes neuronales artificiales, en general, representan una vía para salvar los problemas anteriormente mencionados, por cuanto las mismas son capaces de desarrollar los modelos a partir de las relaciones intrínsecas entre las variables, sin el conocimiento a priori de las relaciones funcionales (Zupan y Gasteiger, 1999). Estas herramientas han sido ampliamente empleadas en el desarrollo de modelos para la predicción del rendimiento de diversos cultivos, entre los cuales se pueden destacar, plátano (Ávila de Hernández, Rodríguez-Pérez, y Hernández-Caraballo, 2012), trigo (Naderloo y cols., 2012), albahaca (Parent y Dafir, 1992), maíz (Matsumura, Gaitan, Sugimoto, Cannon, y Hsieh, 2015), y tomate (Salazar, López, Rojano, Schmidt, y Dannehl, 2015), entre otros.

Entre el abanico de redes neuronales existentes, se encuentran las empleadas en este trabajo, a saber, las de regresión generalizada (GRNN, por sus siglas en inglés). La selección de las mismas se fundamentó en que, contrario a las versiones más populares (las entrenadas con el algoritmo de retropropagación), las GRNN pueden ser desarrolladas en un menor tiempo, y con mayor facilidad, debido al menor número de variables que han de ser optimizadas (Hernández-Caraballo, Rivas, y Ávila de Hernández, 2005).

El objetivo del presente trabajo es desarrollar las redes neuronales de regresión generalizada en la predicción del rendimiento de un cultivo de naranja 'Valencia', empleando las concentraciones foliares de nitrógeno, fósforo, potasio, calcio, y magnesio, como variables explicatorias. Esta aproximación ha sido escasamente estudiada en nuestro país, habiendo sido sólo utilizada, hasta donde alcanza el conocimiento del autor, para la predicción del cultivo de plátano (*Musa* AAB Subgrupo plátano cv. Hárton) (Ávila de Hernández y cols., 2012).

Especial énfasis se hizo en este trabajo en la transformación de las variables originales (concentraciones foliares de nutrimentos), a través de estrategias convencionales (normalización, estandarización, y análisis por componentes principales), y no convencionales (Sistema de Diagnóstico de la Composición Nutricional), a fin de mejorar la capacidad de predicción de las redes neuronales de regresión generalizada.

Desarrollo

Metodología

Base de datos: características y procesado

La base de datos estuvo conformada por las concentraciones foliares de nitrógeno, fósforo, potasio, calcio, y magnesio, así como los rendimientos de 78 árboles de naranja 'Valencia' (Citrus sinensis L. Osbeck), injertados en patrones de limón Volkameriano (Citrus volkameriana Pasq.) (Rodríguez, Rojas, y Sumner, 1997). La determinación de las concentraciones de dichos nutrimentos se llevó a cabo empleando protocolos de análisis convencionales, y que el lector interesado puede consultar en la literatura correspondiente, por ejemplo en Rodríguez y cols. (1997). El Cuadro 1 resume la estadística básica de la base de datos empleada en el presente trabajo.

Cuadro 1: Estadística básica (Promedio, %RSD, porcentaje de la desviación estándar relativa, Valores Mínimo y Máximo) del cultivo de naranja 'Valencia' (edad y rendimiento de los árboles), y de las concentraciones foliares de nutrimentos mayoritarios.

			Concentraciones foliares ($g \cdot kg^{-1}$)				
Estadístico	Edad (Años)	Redimiento ($kg \cdot árbol^{-1}$)	N	P	K	Ca	Mg
Promedio	6,3	120,38	29,4	2,0	13,0	34,1	2,6
%RSD	27,0	17,1	10,5	33,3	28,3	36,6	43,6
Mínimo	5,0	93,25	21,9	0,9	6,9	10,8	0,8
Máximo	9,0	196,65	38,1	3,8	26,8	65,5	6,3

Redes neuronales: descripción y estrategia de entrenamiento

La arquitectura general de una red neuronal de regresión generalizada se muestra en la Figura 1. Las variables descriptivas del sistema (e.g., concentraciones de nutrientes en tejido foliar), son transferidas hacia todas las neuronas de la primera de las dos capas escondidas, la primera de las cuales posee un número de neuronas determinado por la cantidad de casos que conforman la base de datos. Allí, son sometidos a una transformación dada por la función de base radial Ecuación 1:

$$RBF = \left[\frac{\left(\| c_i - p_i \| \right)^2}{\sigma_i} \right] \tag{1}$$

Donde c_i representa al centroide (valor fijo correspondiente a los valores de los vectores utilizados para generar un modelo); p_i al valor de un dato de prueba determinado; y σ_i a la amplitud de la función de base radial. La optimización de esta última variable, también conocida como factor de suavizado, determina la eficiencia de la red neuronal. A continuación, los datos son transferidos a la segunda capa escondida, en donde una neurona estima la sumatoria ponderada de los centroides de las neuronas de la capa anterior (distancia entre c_i y p_i en la ecuación 1); en tanto que la otra neurona determina la sumatoria de los factores de ponderación (suma de los pesos asignados a las neuronas de acuerdo con el grado de similitud). Finalmente, la capa de salida genera el valor de la variable de predicción (rendimiento), el cual es el cociente entre los valores generados por las dos neuronas de la capa precedente (Bauer, 1995).

Tratamiento de datos

Los valores de las variables explicatorias se encuentran en distintos órdenes de magnitud y ello pudiera afectar el desempeño de los modelos desarrollados por las redes neuronales. Para evitar este problema, los datos se sometieron a diversas transformaciones, las mismas se describen a continuación:

(a) *Normalización*:

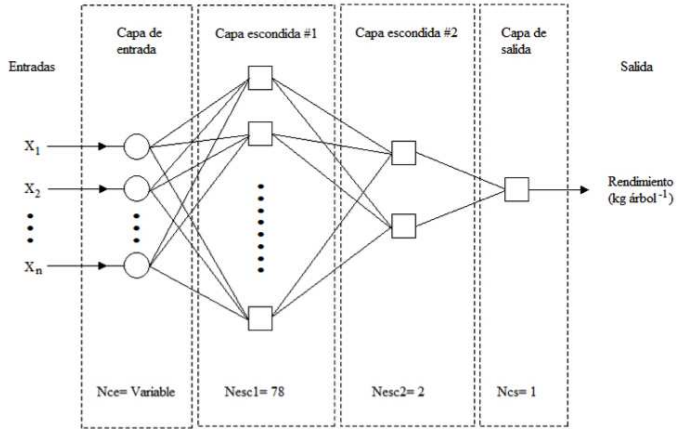

Figura 1: Representación esquemática de una red neuronal de regresión generalizada. Nce: número de neuronas en la capa de entrada; Nesc1: número de neuronas en la capa escondida #1; Nesc2: número de neuronas en la capa escondida #2; Ncs: número de neuronas en la capa de salida

$$C_i^N = \frac{C_i}{C_{Max}} \tag{2}$$

Donde C_i^N y C_i corresponden a la concentración normalizada y original en el tejido foliar de un nutriente determinado, en el ísimo individuo; y C_{Max} la concentración máxima de dicho nutriente.

(b) *Estandarización*:

$$C_i^S = \frac{\left(C_i - \overline{C}\right)}{SD} \tag{3}$$

Donde C_i^S y C_i corresponden a las concentraciones estandarizada y original en el tejido foliar de un nutriente determinado, en el

í-ésimo individuo; en tanto que $\overline{C}$ y SD, representan la concentración promedio de dicho nutriente, y la desviación estándar.

(c) *Log-cociente centrado*: Esta transformación fue desarrollada por (Aitchison, 1982), a fin de superar la llamada "restricción de la suma constante", propia de los datos composicionales; a saber, que los valores no pueden variar de manera irrestricta, sino que están confinados a que la suma de conocidos y desconocidos (denominados estos últimos globalmente como el "valor de relleno", Rd) sea igual al 100 %. El log-cociente centrado se estima como:

$$V_i^X = \ln\left[\frac{C_i^X}{\left(C_i^N C_i^P C_i^K ... R_d\right)^{\frac{1}{d+1}}}\right] \tag{4}$$

En donde V_i^X y C_i^Xcorresponden al log-cociente centrado y a la concentración de un nutriente determinado, en el í-ésimo individuo. En el denominador se encuentra representado el cálculo de la media geométrica (d corresponde al número de nutrientes).

(d) *Índice nutricional individual*: Esta transformación forma parte del Sistema de Diagnóstico de la Composición Nutricional, desarrollado por Parent y Dafir (1992). La misma es, en realidad, una estandarización empleando los log-cocientes centrados, tal y como se muestra:

$$I_i^X = \frac{\left(V_i^X\right) - \left(V_{Ref}^X\right)}{SD_{Ref}^X} \tag{5}$$

En donde I_i^Xy V_i^X corresponden al índice nutricional y al log-cociente centrado de un nutriente dado, para el í-ésimo individuo, y V_{Ref}^X SD_{Ref}^X al log-cociente centrado y a la desviación estándar del mismo nutriente, para una muestra de alto rendimiento (Hernández-Caraballo, Rodríguez-Rodríguez, y Rodríguez-Pérez, 2009).

(e) *Índice nutricional global*: El índice nutricional global es, como su nombre lo indica, una medida del estatus global de la planta. Desde el punto de vista de este trabajo, su importancia reside en la reducción de la dimensionalidad que implica su implementación. El mismo se calcula de la siguiente manera:

$$CND - r^2 = \sum_{X=1}^{d+1} \left(I_i^X\right)^2 \tag{6}$$

En donde I_i^X representan a los índices nutricionales individuales previamente definidos (Ec. 5). La sumatoria se ejecuta hasta el nutriente *d+1*, por cuanto incluye al índice nutricional individual estimado a partir del valor de relleno (R_d).

(f) *Componentes principales*: Los datos originales (concentraciones en tejido foliar) fueron sometidos a la transformación log-cociente centrado, y posteriormente procesados a través de un análisis por componentes principales, a fin de reducir la dimensionalidad de la base de datos. Los factores correspondientes fueron obtenidos con/sin aplicación de la rotación Varimax, y los puntajes mediante el método de Anderson-Rubin para garantizar la ortogonalidad de los mismos (Jackson, 1991).

Redes neuronales: arquitectura y estrategia de evaluación.

En el Cuadro 2 se muestran las características de las redes neuronales evaluadas en el presente estudio, en términos del número de nodos de entrada, y las características de los pares entrada-salida empleados en el desarrollo de los modelos. Las redes neuronales se construyeron con el software Statistica®, en ambiente Windows®, llevándose a cabo bajo la modalidad "*leave-one-out*" (deja uno afuera), según la cual *n-1* datos son usados para el entrenamiento, y el faltante se emplea para verificar la exactitud de la predicción (Elisseeff y Pontil, 2002), y repitiendo el proceso hasta que todos los datos son empleados en la comprobación (de manera automática) del modelo. El desempeño de las redes se monitorizó, por una parte, mediante

la raíz del error cuadrático medio (RMSE, por sus siglas en inglés), mostrada en la ecuación (7)

$$RMSE = \sqrt{\frac{\sum_i^N \left[\left(Rend_i^{Pred}\right) - \left(Rend_i^{Real}\right)\right]^2}{N}} \tag{7}$$

Por otra parte, también se recurrió al uso del criterio visual (elaboración de gráficos de correlación) y del criterio matemático (cálculo de la curva de correlación y del coeficiente de correlación, r^2), como medios adicionales de la estimación del desempeño de los modelos generados con las redes neuronales de regresión generalizada.

Cuadro 2: Número de nodos de entrada, y características de los pares entrada-salida, para los modelos de redes neuronales de regresión generalizada evaluadas.

Modelo	Nodos de entrada	Descripción de los pares entrada-salida
1	5	Concentraciones normalizadas – Rendimiento bruto
2	5	Concentraciones normalizadas – Rendimiento normalizado
3	5	Concentraciones estandarizadas – Rendimiento bruto
4	5	Concentraciones estandarizadas – Rendimiento estandarizado
5	6	Índices nutricionales individuales (I_x) – Rendimiento bruto
6	6	Índices nutricionales individuales (I_x) – Rendimiento normalizado
7	6	Índices nutricionales individuales (I_x) – Rendimiento estandarizado
8	2	Componentes principales sin rotación – Rendimiento bruto
9	2	Componentes principalessin rotación – Rendimiento normalizado
10	2	Componentes principales sin rotación – Rendimiento estandarizado
11	4	Componentes principales rotados – Rendimiento bruto
12	4	Componentes principales rotados – Rendimiento normalizado
13	4	Componentes principales rotados – Rendimiento estandarizado
14	1	Índice nutricional global normalizado ($CND - r^2$) – Rendimiento bruto
15	1	Índice nutricional global normalizado ($CND - r^2$) – Rendimiento normalizado
16	1	Índice nutricional global normalizado ($CND - r^2$) – Rendimiento estandarizado

Resultados y discusión

Desarrollo de modelos de predicción del rendimiento de un cultivo de naranja "Valencia", mediante GRNN.

Las características de las variables explicatorias, a saber número y magnitud, son factores determinantes en el desempeño de las redes neuronales en general. Por una parte, es necesario disponer de suficientes descriptores para que la red pueda aprender las relaciones intrínsecas entre las variables dependiente(s) e independiente(s). Sin

embargo, un número innecesario de variables pueden hacer que la red memorice una relación determinada, y que con ello sea incapaz de funcionar apropiadamente ante datos nuevos (capacidad de predicción). Por otra parte, la magnitud de la variables debe ser vigilada igualmente, a fin de que la red no le asigne una falsa importancia en el modelo, sólo por esa razón. En el Cuadro 1 se puede apreciar una situación relacionada con lo dicho previamente. La concentraciones foliares de nitrógeno ($29{,}4\, g \cdot kg^{-1}$) son, en promedio, un orden de magnitud mayores que las concentraciones foliares de fósforo ($2{,}0\, g \cdot kg^{-1}$), pero ello no indica que, por esa razón, la primera variable explicativa sea intrínsecamente más importante en el desarrollo del modelo, que la segunda.

La selección de las transformaciones mostradas en el Cuadro 2 es el resultado de las consideraciones anteriormente descritas. En todas ellas se lleva a cabo un escalado de los datos, mientras que algunas de ellas también reducen la dimensionalidad de la base de datos original. En este sentido, el uso del análisis por componentes principales requiere un pequeño aparte. Al aplicar el análisis por componentes principales se busca reducir el número de variables originales, al tiempo de conservar la mayor parte de la información original (Jackson, 1991).

La Figura 2 muestra los gráficos de dispersión a través de los cuales se llevó a cabo la selección del número de componentes principales a usar como variables explicatorias en algunas de las redes neuronales de regresión generalizada. En la Figura 2(a) se puede apreciar que es posible reducir el número de variables originales (seis concentraciones de nutrientes en el tejido foliar) a sólo dos (número de componentes principales), si se aplica el criterio de Kaiser-Guttman (Raîche, Riopel, y Blais, 2006). De acuerdo con dicho criterio, se conservan aquellos componentes principales cuyos autovalores sean mayores a 1 (ver el primer eje de ordenadas a la izquierda de la Figura 2 (a)). Esta reducción del número de variables permite aún conservar suficiente información (72 % de la varianza original), tal y como se aprecia en el segundo eje de ordenadas, a la derecha de la figura en cuestión.

En la Figura 2(b) se puede observar que la reducción en el número de variables explicatorias no es tan marcado (de las concentraciones

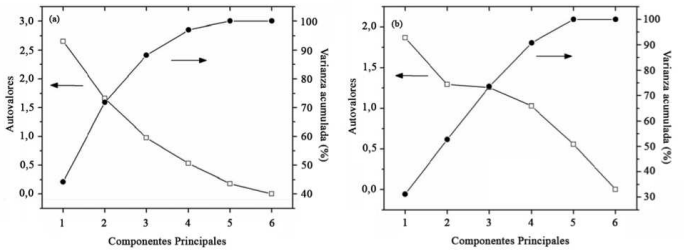

Figura 2: Gráfico de dispersión combinado para la determinación del número de componentes principales para su uso como datos de entrada para las redes neuronales: (a) sin rotación; (b) con rotación Varimax de las soluciones finales.

foliares de seis nutrientes, a cuatro componentes principales), y al mismo tiempo, se conservó un porcentaje mayor de la varianza original (91 %). Ya que en ambos casos se logró una reducción del número de variables de entrada a las redes neuronales, aunque en diferente medida, se decidió evaluar el comportamiento de los correspondientes modelos basados en redes neuronales. A continuación, se optimizaron las redes neuronales descritas en el Cuadro 1, hallando el factor de suavizado (amplitud de la función de base radial) que minimiza el criterio de desempeño seleccionado en este trabajo, a saber, la raíz cuadrada del error cuadrático medio (Ecuación 7).

En la Figura 3 se puede apreciar que: (1) las redes neuronales con salidas transformadas muestran un mejor desempeño cuando el factor de suavizado es pequeño (Figuras a-f); y, (2) no parece haber una diferencia en el desempeño en función del tipo de transformación (normalización vs. estandarización) de la variable de salida (Figuras c-f).

Efectivamente, el desempeño de las GRNNs cuyas salidas no han sido transformadas (representadas con "•") es en general menor, esto es, el RMSE es más elevado, que en aquellas redes cuyas salidas han sido normalizadas ("△"), o estandarizadas ("▼"). Ello es particularmente cierto cuando el logaritmo del factor de suavizado es pequeño (varía de acuerdo con el tipo de red neuronal). Sin embargo, los desempeños tienden a hacerse independientes de la transformación,

a medida que el ancho de la función de base radial se hace mayor. Lo anterior contraste de manera interesante con los resultados obtenidos por (Ávila de Hernández y cols., 2012). En su estudio, los autores hallaron que el desempeño de las redes de regresión generalizada desarrollados para la predicción del rendimiento del cultivo de plátano desmejoraba significativamente a medida que el ancho de la función de base radial se hacía mayor. Es evidente que la optimización de los modelos debe llevarse a cabo de manera individual para cada caso de estudio.

Una vez optimizados los anchos de las funciones de base radial para cada una de las dieciséis redes neuronales, se procedió a evaluar su capacidad para predecir el rendimiento del cultivo de naranja 'Valencia', haciendo uso criterios adicionales, a saber, criterios visuales (gráficos de correlación), y matemáticos (estimadores de bondad de ajuste) (Asuero, Sayago, y González, 2006). Ello se debe a que el uso exclusivo de estos últimos, muy frecuente en la literatura especializada, puede conducir a conclusiones erróneas sobre la capacidad de predicción de los modelos.

Predicción del rendimiento de un cultivo de naranja "Valencia" mediante GRNN.

Las Figuras 4a y 4b muestran que, tanto la normalización como la estandarización de los datos de salida provocan una mejora del desempeño de las redes neuronales, siendo ésta más pronunciada en el primer caso (mayor cambio en la pendiente de la curva de regresión). Sin embargo, aunque en el Cuadro 3 se aprecia que la correlación entre el rendimiento predicho y el real es perfecta ($r^2 = 1{,}000$), resulta evidente un sesgo en el modelo hacia la sobre-estimación de rendimientos por encima de ca. $150\,kg \cdot árbol^{-1}$ (la curva de correlación está por encima de la línea punteada de referencia). Está claro que es fundamental transformar los datos de salida, independiente del tratamiento que se le aplique a los datos de entrada (con la excepción que se menciona más adelante). No hacerlo genera los modelos con los peores desempeños (mayores valores de RMSE).

Los modelos en los que se usan los índices nutricionales individuales como variables de entrada tienen, tanto desde el punto de vista

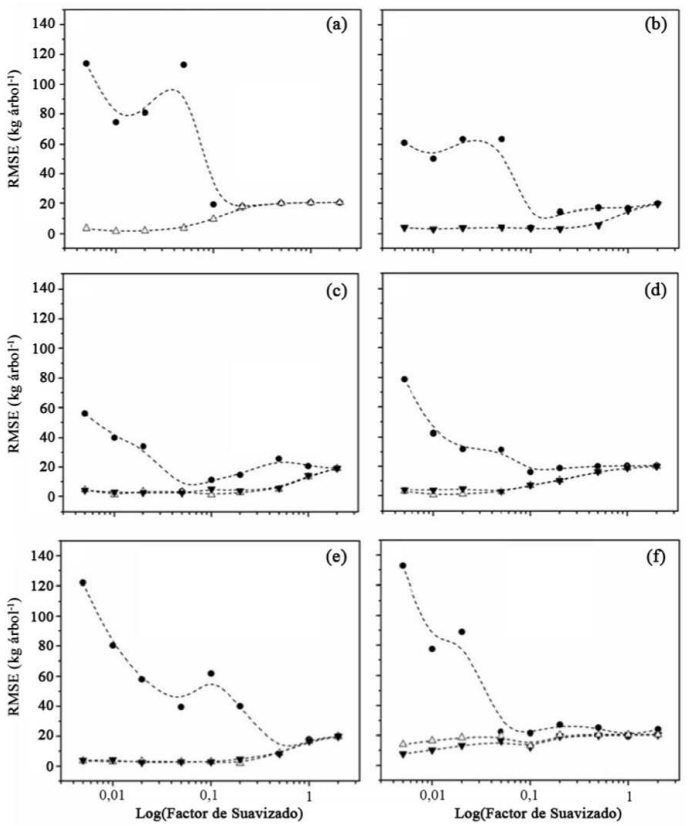

Figura 3: Variación del error cuadrático medio ($RMSE$) en función del logaritmo del factor de suavizado. Datos de entrada: (a) concentraciones normalizadas; (b) concentraciones estandarizadas; (c) ííndices nutricionales individuales; (d) puntuaciones de los primeros dos componentes principales no rotados; (e) puntuaciones de los primeros cuatro componentes principales rotados; (f) índice nutricional global. Datos de salida: (●) rendimiento sin transformación; (△) rendimiento normalizado; (▼) rendimiento estandarizado.

gráfico (Figura 4c), como estadístico (Cuadro 3), desempeños muy similares. Sin embargo, una mirada detallada muestra la importancia de emplear otro criterio de comparación, como es la raíz cuadrada

del error cuadrático medio ($RMSE$). Con éste es posible seleccionar la red "Índices nutricionales individuales/Rendimiento normalizado" como una de las mejores ($RMSE = 0{,}98\, kg \cdot árbol^{-1}$), tal y como se muestra en el Cuadro 3. Otro aspecto importante que destacar, es que la combinación "Índices nutricionales individuales/Rendimiento sin transformar" conduce a la red neuronal con el mejor desempeño de todas las redes neuronales con salidas no transformadas. Efectivamente, mientras que el RMSE de ésta es de $3{,}46\, kg \cdot árbol^{-1}$ los RMSE de las otras redes neuronales con salidas no transformadas oscila entre 3,74 y $21{,}53 kg \cdot árbol^{-1}$.

Visualmente, es fácil apreciar que el uso de los componentes principales no rotados (Figura 4d) conduce a la obtención de modelos con un mejor desempeño que el de sus análogos con componentes principales rotados y salidas transformadas (Figura 4e). El Cuadro 3 confirma lo anterior: el desempeño de las GRNNs que emplean el primer tipo de datos es muy superior ($RMSE = 0{,}51\, kg \cdot árbol^{-1}$.) al del segundo tipo ($RMSE = 1{,}89\, kg \cdot árbol^{-1}$), cuando la salida es normalizada. Sin embargo, también es justo destacar que ambos modelos muestran un mejor desempeño que el del resto de las redes neuronales, con la excepción de las destacados en párrafos previos.

Finalmente, el uso de una única variable de entrada (índice nutricional global), produce un modelo cuyo desempeño comparativo es insatisfactorio (véase la Figura 4f y el Cuadro 3). Resulta razonable pensar que, tal reducción de información dificulta el establecimiento de relaciones apropiadas entre la variable descriptiva y la predicha, por parte de la red neuronal. Por esa razón, se desaconseja este tipo de transformación. En definitiva, los mejores modelos son aquellos en los que se emplean las combinaciones: (a) Índices nutricionales individuales/Rendimiento normalizado ($RMSE = 0{,}98\, kg \cdot árbol^{-1}$); y, (b) Componentes principales no rotados/Rendimiento normalizado ($RMSE = 0{,}51\, kg \cdot árbol^{-1}$). El usuario interesado puede emplear cualquiera de las dos transformaciones indicadas; sin embargo, es necesario destacar lo siguiente. Los índices nutricionales individuales son variables que se estiman siempre que se desea evaluar el estatus nutricional de una planta, mediante sistemas de diagnóstico como el CND. En consecuencia, su utilización como variables de entrada para las

redes neuronales formaría parte de un esquema de trabajo integral, gracias al cual sería posible para el productor agrícola, primeramente, el estatus nutricional del cultivo, y segundo, una estimación del rendimiento en las condiciones de trabajo. Dicha información le permitiría tomar decisiones relacionadas con la gestión de la unidad de producción.

Por su parte, el análisis por componentes principales es una herramienta más específica, cuya utilización suele estar más dirigida hacia la exploración no supervisada de datos. Evidentemente, la misma puede emplearse para la reducción del espacio de datos, como en este trabajo; sin embargo, su implementación en un flujo de trabajo como el planteado en el párrafo anterior, sería menos fluido, y menos informativo para el productor agrícola.

Por lo dicho en las líneas previas, aunque no se descarta la red neuronal que emplea los componentes principales no rotados como datos de entrada, se favorece aquella que utiliza los índices nutricionales individuales para el mismo fin.

En un trabajo realizado anteriormente con el cultivo de plátano (Ávila de Hernández y cols., 2012), los autores encontraron que el modelo de red neuronal de regresión generalizada que empleaba las concentraciones foliares de nutrientes como datos descriptivos, resultó ser el segundo modelo más efectivo para tal fin. Sin embargo, no es posible hacer una comparacióón directa entre los resultados mostrados en este trabajo, y los hallados por los autores en cuestión. Para empezar, (Ávila de Hernández y cols., 2012) emplearon una base de datos con 10 nutrientes (macro y micronutrientes), en oposición a la base de datos empleada en este trabajo, que sólo incluyó las concentraciones foliares de seis macronutrientes. Otro aspecto que dificulta una comparación, es la utilización de las estrategias de normalización y estandarización como únicos tratamientos de los datos. Sin embargo, en lo que sí existe una similitud entre los dos trabajos, es en el éxito de las redes neuronales de regresión generalizada para el modelado de las relaciones desconocidas entre las variables dependiente e independientes, y en la utilidad de dichos modelos para la predicción exitosa del rendimiento de los cultivos.

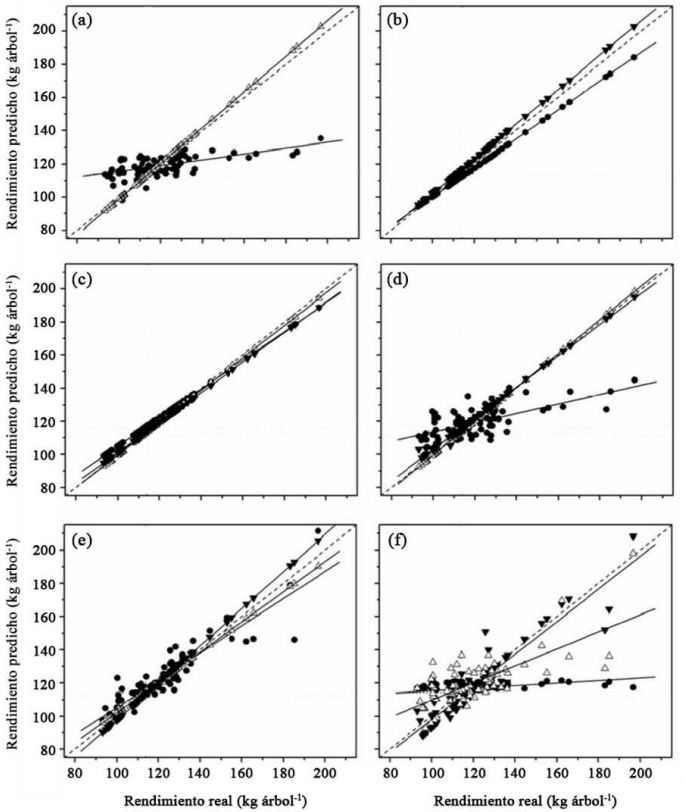

Figura 4: Gráficos de correlación del rendimiento predicho versus el rendimiento real para los modelos optimizados basados en las redes neuronales de regresión generalizada. Datos de entrada: (a) concentraciones normalizadas; (b) concentraciones estandarizadas; (c) índices nutricionales individuales; (d) puntuaciones de los primeros cuatro componentes principales no rotados; (e) puntuaciones de los primeros cuatro componentes principales rotados; (f) índice nutricional global. Los símbolos corresponden a la salida de las redes neuronales: (•) rendimiento sin transformación; (△) rendimiento normalizado; (▼) rendimiento estandarizado.

Cuadro 3: Desempeño de las GRNN desarrolladas para la predicción del rendimiento del cultivo de naranja 'Valencia'.

			Análisis de regresión	
Datos $E/S^{(a)}$	Factor de suavizado	RMSE $(kg \cdot árbol^{-1})$	Ecuación	r^2
$NorC/Rend$	0,2	17,61	$Rend_{Pred} = 98,3581 + 0,1730 \cdot Rend_{Real}$	0,3475
$NorC/NorRend$	0,01	1,57	$Rend_{Pred} = -8,6684 + 1,0750 \cdot Rend_{Real}$	1,0000
$EstC/Rend$	0,1	3,74	$Rend_{Pred} = 12,9390 + 0,8706 \cdot Rend_{Real}$	1,0000
$EstC/EstRend$	0,01	3,03	$Rend_{Pred} = -1,7699 + 1,0380 \cdot Rend_{Real}$	1,0000
$I_x/Rend$	0,05	3,46	$Rend_{Pred} = 18,1857 + 0,8666 \cdot Rend_{Real}$	1,0000
$I_x/NorRend$	0,01	0,98	$Rend_{Pred} = 1,6222 + 0,9792 \cdot Rend_{Real}$	1,0000
$I_x/EstRend$	0,05	2,06	$Rend_{Pred} = 10,2775 + 0,9078 \cdot Rend_{Real}$	1,0000
$NRotPC/Rend$	0,1	16,03	$Rend_{Pred} = 84,8582 + 0,2850 \cdot Rend_{Real}$	0,4462
$NRotPC/NorRend$	0,01	0,51	$Rend_{Pred} = -3,0389 + 1,0244 \cdot Rend_{Real}$	1,0000
$NrotPC/EstRend$	0,05	2,68	$Rend_{Pred} = 8,4794 + 0,9440 \cdot Rend_{Real}$	0,9925
$RotPC/Rend$	0,5	8,36	$Rend_{Pred} = 23,2924 + 0,8194 \cdot Rend_{Real}$	0,8386
$RotPC/NorRend$	0,2	1,89	$Rend_{Pred} = 11,0922 + 0,9101 \cdot Rend_{Real}$	0,9996
$RotPC/EstRend$	0,02	2,38	$Rend_{Pred} = -13,9106 + 1,1162 \cdot Rend_{Real}$	1,0000
$CND/Rend$	0,1	21,53	$Rend_{Pred} = 107,5460 + 0,0768 \cdot Rend_{Real}$	0,0253
$NorCND/NorRend$	0,005	14,05	$Rend_{Pred} = 58,3238 + 0,5133 \cdot Rend_{Real}$	0,5273
$EstCND/EstRend$	0,005	7,76	$Rend_{Pred} = -0,6797 + 0,9847 \cdot Rend_{Real}$	0,8828

(a) Datos E/S: Concentraciones normalizadas (NorC); Concentraciones estandarizadas (EstC); Índices nutricionales individuales (I_x); Componentes principales no rotados (NrotPC); Componentes principales rotados (RotPC); Índice nutricional global (CND); Índice nutricional global normalizado (NorCND); Índice nutricional global estandarizado (EstCND); Rendimiento (Rend); Rendimiento normalizado (NorRend); Rendimiento estandarizado (EstRend).

Conclusiones

Las redes neuronales de regresión generalizada representan una opción atractiva para el modelado de relaciones complejas, como las existentes entre las concentraciones de nutrientes en tejido foliar y el rendimiento de un cultivo. Sin embargo, las variables explicativas que se empleen en cualquier modelo predictivo basado en redes neuronales deben estar apropiadamente escaladas, a fin de que el desempeño del modelo esté determinado por las relaciones entre las variables, y no por la magnitud relativa de las mismas. En este estudio se evaluaron diversas transformaciones convencionales (normalización, estandarización y Análisis por Componentes Principales) y no convencionales

(Sistema de Diagnóstico de la Composición Nutricional), siendo dos variantes de estas últimas estrategias las que condujeron a la obtención de los modelos predictivos más efectivos. El uso de los componentes principales condujo a la generación de un modelo de predicción del rendimiento más exacto ($0,51kg \cdot árbol^{-1}$) que el obtenido con el sistema CND ($0,98kg \cdot árbol^{-1}$. Sin embargo, esta última aproximación es más apropiada si se contempla un esquema de trabajo integral, en el que, en primer lugar, se lleve a cabo el diagnóstico nutricional del cultivo, y en segundo lugar, se prediga el rendimiento del mismo en esas condiciones. La información obtenida por el productor en ambas etapas, le permitiría tomar decisiones que mejoren el desempeño del sistema de producción.

Agradecimientos

Al Prof. Orlando Rodríguez Rodríguez por suministrar la base de datos empleada en la realización de este estudio, y a la Dra. Rita M. Ávila de Hernández, por las valiosas observaciones realizadas al manuscrito.

Referencias

Aitchison, J. (1982). The statistical analysis of compositional data. *J. Royal Stat. Soc. Series B*, *44*, pp.139-177.

Arizaleta, M., Rodríguez, O., y Rodríguez, V. (2002). Relación de los índices DRIS, índices de balance de nutrientes, contenido foliar de nutrientes y el rendimiento del cafeto en venezuela. *Bioagro*, *14*, pp.153-159.

Asuero, A. G., Sayago, A., y González, A. G. (2006). The correlation coefficient: an overview. *Ccrit. Rev. Anal. Chem*, *36*, pp.41-59.

Ávila de Hernández, R., Rodríguez-Pérez, V., y Hernández-Caraballo, E. (2012). Predicción del rendimiento de un cultivo de plátano mediante redes neuronales artificiales de regresión generalizada. *Publ. Cien. Tec*, *6*, pp.31-40.

Bauer, M. M. (1995). Generalized regression neural network for technical use. *Master's Thesis, University of Wisonsin-Madison, USA*.

Elisseeff, A., y Pontil, M. (2002). Leave-one-out error and stability of learning algorithms with applications. *J. Mach. Learn. Reas*, *1*, pp.6-21.

Food, y Organization, A. (2008). *Static database (faostat.* Disponible en : http://faostat.fao.org consultado: 13/04/2014).

Fornaciari, M., Orlandi, F., y Romano, B. (2005). Yield forecasting for olive trees: a new approach in a historical series (Umbria, Central Italy). *Agron. J*, *97*, pp.1537-1542.

Hernández-Caraballo, E. A., Rivas, G., y Ávila de Hernández, R. M. (2005). Evaluation of a generalized regression artificial neural network for extending cadmium's working calibration range in graphite furnace atomic absorption spectrometry. *Anal. Bioanal. Chem*, *381*, pp.788-794.

Hernández-Caraballo, E. A., Rodríguez-Rodríguez, O., y Rodríguez-Pérez, V. (2009). Corrigendum to "evaluation of the boltzmann equation as an alternative model in the selection of the high-yield subsample within the framework of the compositional nutrient diagnosis system". *Environ. Exp. Bot*, *65*, p. 91.

Jackson, J. E. (1991). A user's guide to principal components. *Wiley Series in Probability and Mathematical Statistics. USA*.

Kitchen, N. R., Drummond, S. T., Lund, E. D., Sudduth, K. A., y Buchleiter, G. W. (2003). Soil electrical conductivity and topography related to yield for three contrasting coil-crop systems. *Agron. J*, *95*, pp. 483-495.

Lobell, D. B., y Burke, M. B. (2010). On the use of statistical models to predict crop yield responses to climate change. *Agricult. For. Meteo*, *150*, pp.1443-1452.

Matsumura, K., Gaitan, C. F., Sugimoto, K., Cannon, A. J., y Hsieh, W. (2015). Maize yield forecasting by linear regression and artificial neural networks in Jilin, China. *J. Agric. Sci*, *153*, pp.399-410.

Mourtzinis, S., Arriaga, F. J., Balkcom, K. S., y Ortiz, B. V. (2013). Corn grain and stover yield prediction at R1 growth stage.

Agron. J, *105*, pp.1045-1050.

Naderloo, L., Alimardani, R., Omid, M., Sarmadian, F., Javadikia, P., Torabi, M. Y., y Alimardani, F. (2012). Application of ANFIS to predict crop yield based on different energy inputs. *Measurements*, *45*, pp.1406-1413.

Pahlavan, R., Omid, M., y Akram, A. (2012). Energy input-output analysis and application of artificial neural networks for predicting greenhouse basil production. *Energy*, *37*, pp.171-176.

Parent, L. E., y Dafir, M. (1992). A theoretical concepto of compositional nutrient diagnosis. *J. Am. Soc. Hort. Sci*, *117*, pp.239-242.

Raîche, W., Riopel, M., y Blais, J. G. (2006). Non graphical solutions for the cattell´s scree test. *Trabajo presentado en el International Meeting of the Psychometric Society, Montréal, Junio 16, 2006*.

Rodríguez, O., Rojas, G., y Sumner, M. (1997). Valencia orange DRIS norms for Venezuela. *Commun. Soil Sci. Plant Anal*, *28*, pp.1461-1468.

Salazar, M. R., López, C. I., Rojano, A., Schmidt, U., y Dannehl, D. (2015). Tomato Yield Prediction in a Semi-Closed Greenhouse. *Acta Hort*, *1107*, pp.263-269.

Wallace, A., y Wallace, G. A. (1993). 10. limiting factors, high yields, and law of the maximum. *En: Janick, J. (Editor). Horticultural Reviews*, *15*, pp.409-448.

Zupan, J., y Gasteiger, J. (1999). *Neural networks in chemistry and drug design* (2nd Ed. ed.). Wiley-VCH, Weinheim.