An Application into Using Artificial Intelligence for Estimating Organic Carbon

[1]Spencer, M., [1]McCullagh, J., [1]Whitfort, T. and [2]Reynard, K.

[1]La Trobe University, [2]Department of Primary Industries, E-Mail: j.mccullagh@latrobe.edu.au

Keywords: *Organic carbon; Artificial neural networks; Artificial intelligence.*

EXTENDED ABSTRACT

Soil is an important resource that is vital for the sustainability of plants, animals and humans. Organic Carbon is an important soil property that describes the amount of carbon broken down from plants or life forms to the soil below. Traditionally to quantify soils organic carbon content, a soil sample is collected and analysed in a laboratory. The collection of the soil sample is a time consuming and expensive process. A promising alternative is to estimate organic carbon based solely upon remotely sensed data. The advantage of using remotely sensed data is that organic carbon can be estimated without going out into the field.

Researchers have investigated the association between organic carbon and climatic factors as well as soil site attributes. Factors such as precipitation, temperature, clay content and soil texture have been found to relate to organic carbon content.

Artificial Intelligence (AI) and statistical techniques have been successfully used in various soil studies. Artificial Neural Networks and Decision Trees are examples of AI techniques that have been applied to many prediction and classification problems. These AI approaches use a supervised learning approach, that uses historical data to determine the relationship between the influencing factors (e.g. rainfall, elevation), and the output (e.g. organic carbon). Once trained, these techniques can be applied to previously unseen input data and used to produce an estimate of organic carbon content.

Victorian soil data was provided by the Department of Primary Industries (DPI) to assist in the machine learning process. The factors provided in each set are climatic (i.e. rainfall, temperature and evaporation), landform (i.e. elevation and slope) and soil physical and chemical attributes (i.e. pH, sand, silt and clay). From the DPI dataset two datasets were produced. The first dataset (*Vic-All)* contained all factors related to the presence of organic carbon. The second dataset (*Vic-Remote*) contained landscape data predominantly derived from Digital Elevation Models (DEMs) as well as climatic data. The first dataset was used to determine if chemical analysis of soil samples could be replaced by using soil survey and automated factors to estimate organic carbon. The second dataset was used to investigate the potential of organic carbon mapping only using remotely sensed factors.

Experiments have been undertaken using Decision Trees, Artificial Neural Networks and Multiple Linear Regression. Models were developed using Artificial Neural Networks and Multiple Linear Regression to predict the percentage of organic carbon in the topsoil. Neural Networks out performed Multiple Linear Regression when predicting the percentage of organic carbon. Correlations of 0.82 and 0.59 were obtained on the *Vic-All* and *Vic-Remote* testing datasets using Neural Networks.

A classification study was also undertaken, using Decision Trees and Artificial Neural Networks, to determine how well each model could classify organic carbon into classes (i.e. Low, Moderate or High). Neural Networks showed better performance compared to Decision Trees on three of the four experiments. Testing performances of 80.2% and 62.3% were achieved on the *Vic-All* and *Vic-Remote* datasets respectively when classifying to three classes.

The results indicate that Neural Networks have promise to be a cost effective method of estimating organic carbon. In particular, the use of Neural Networks combined with remotely sensed data has the potential to make it feasible to map organic carbon over broader areas with limited soil analytical data.

Potential research extensions are discussed, including possible directions for developing enhanced models for improving performance, and demonstrating the applicability of the model to broader data sets.

1. INTRODUCTION

Soil is an important resource that is vital for the sustainability of plants, animals and humans. Soil is an entity composed of sand, silt, clay and organic matter. It is required for agricultural activities such as growth of plants and crops, for the support of flora and fauna, and for the construction of buildings. It is critical that its quality remains adequate for the sustainability of human life.

Carbon is an important element for the support of all life. It exists within plants and animals, is stored within geology and organic matter, and stored within the world's oceans (McVay & Rice 2002). Carbon can be stored as part of one of four reservoirs (Post *et al.* 1990):

- Oceanic carbon,
- Geological carbon,
- Atmospheric carbon, and
- Terrestrial carbon.

Figure 1 shows the organic carbon cycle and the exchange processes between each of the reservoirs.

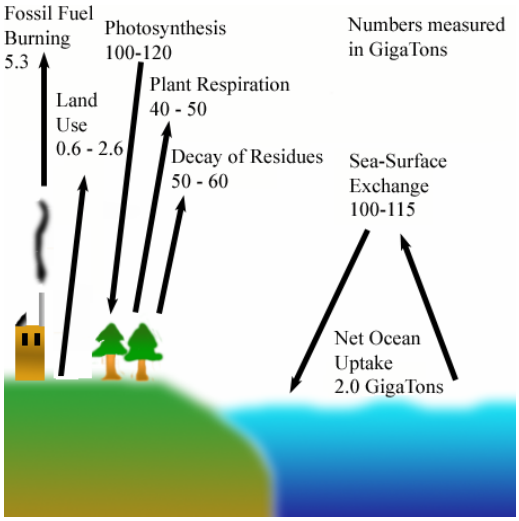

Figure 1. The Global Carbon Cycle.

Oceanic carbon is either dissolved in the water or present as live organisms or dead animals or plants. This is the largest of all carbon reservoirs, containing an estimated 38,000 gigatons. Geological carbon is stored within rocks or materials under the earth, with an estimated 4,000 gigatons of carbon stored in this reservoir. Atmospheric carbon is stored in the air or gaseous regions of the earth. In 1990, there was an estimated 748 gigatons of carbon stored within the earth's atmosphere (Post *et al.* 1990). The burning of fossil fuels has resulted in the amount of atmospheric carbon increasing exponentially at a rate of 1.5% per year (McVay & Rice 2002, Climate Change Science Program and Global Change Research 2003).

The terrestrial carbon reservoir includes carbon stored in plants and animals, and the carbon in soil from the breakdown of carbon materials into the soil. There is estimated to be between 2,000 to 2,400 gigatons of carbon in the terrestrial reservoir (Post *et al.* 1990, Batjes 1996). The most common process of storage into this reservoir is through plant photosynthesis, where plants accept CO_2 and create organic compounds. Plant breakdown, respiration and fires also result in carbon being released into the atmosphere (Post *et al.* 1990).

As litter falls from vegetation to the ground, organic content is placed on top of the soil. This organic matter is rich in carbon, and is returned to either the atmospheric or terrestrial reservoirs. The decomposition of the organic matter releases some carbon to the air. Some of the carbon is stored deeper into the ground, maintaining soil fertility. This research focuses on the terrestrial reservoir, and the quantification of organic carbon in the topsoil.

2. ARTIFICIAL INTELLIGENCE

Artificial Intelligence (AI) has been successfully used in classification, prediction, and recognition problems (Kim & Kim 2002, McCullagh 2005, Zhang 2004, Zhu 1995, Kohonen *et al.* 1988). The potential benefits of AI include greater prediction reliability and cost-efficient estimation.

Artificial Intelligence has been utilised in various soil studies. Examples of such research include the prediction of pH content and clay in soil (Henderson *et al.* 2004), mapping of dryland salinity (Spencer *et al.* 2004), and predicting river salinity (Maier & Dandy 1998).

Two models that have been investigated in soil science are:

- Decision Trees, and
- Artificial Neural Networks

2.1. Decision Trees

Decision Trees use a supervised learning approach that uses a 'divide and conquer' approach for classification problems (Quinlan 1993). In this technique, nodes are used to split training examples into classes based on their input values. A collection of these nodes are then used to

generate rules that can be used to classify unseen examples. An example of a decision tree is shown in Figure 2.

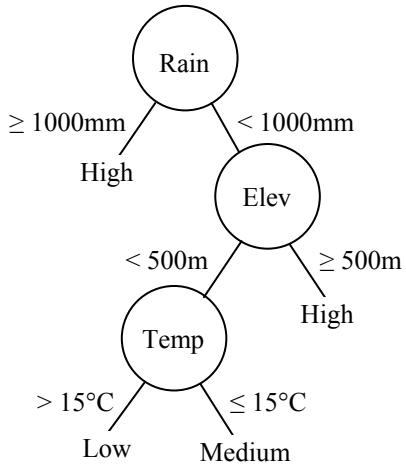

Figure 2. Example of a Decision Tree structure when classifying organic carbon into three classes (High, Medium and Low)

2.2. Artificial Neural Networks

Artificial Neural Networks (ANNs) are inspired by biological neural networks. An ANN can be visualised as a black-box, non-linear approach, where a collection of inputs are presented to the network and one or more results (outputs) are produced. ANNs use a supervised learning approach, learning from training examples, adjusting weights to reduce the error between the correct result and the result produced by the network. ANNs endeavour to develop a general relationship between the inputs and outputs provided.

An example of an Artificial Neural Network is shown in Figure 3.

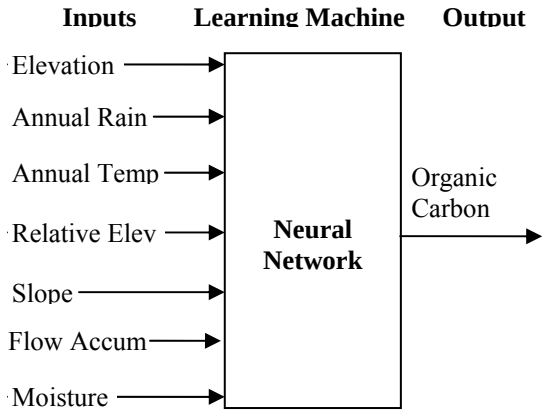

Figure 3. Neural Network as a black-box approach.

3. RESEARCH INTO ORGANIC CARBON ESTIMATION

To determine the quantity of organic carbon in a soil sample, a sample needs to be collected then analysed in a laboratory. Unfortunately, collecting the sample is a time-consuming and expensive process. Researchers are investigating cheaper alternatives for estimating organic carbon rather than quantifying it directly from soil samples.

Research has focused on understanding the relationships between organic carbon and climatic factors as well as soil site variables. Factors such as precipitation, temperature, clay content and soil texture have been found to relate to organic carbon content (Hontoria *et al.* 1999, Parton *et al.* 1987). The purpose of this research is to use automated and soil survey data to provide a time and cost efficient method for estimating organic carbon.

Jenny's equation (Jenny 1941) states that there are five major factors associated with the formation of soil and its properties. The formula is shown below:

$$s = f(cl, o, r, p, t, ...) \quad (1)$$

where s is soil, cl is climate, o are organisms, r is relief, p is parent material, and t is time. This formula states that any soil property or taxonomy is a result of the weathering processes undertaken.

Models have been developed to estimate organic carbon based on known relationships between organic carbon and its related factors. So far, research efforts have focused on linear models for the estimation of organic carbon.

Linear regression was used to estimate organic carbon for a study in America (Nichols 1984). Due to the high correlation between clay content and soil organic carbon, it was believed that a regression model could be used with annual precipitation to estimate organic carbon. Results demonstrated that organic carbon could be predicted well under certain conditions. Annual precipitation was also used as a factor to estimate the maximum and minimum boundaries of organic carbon content in another study in America (Parton *et al.* 1987). Hontoria *et al.* (1999) used a linear regression model to estimate organic carbon in Spain. It was discovered that by using annual precipitation in addition to temperature, a general indication of organic carbon could be estimated. Other linear regression models have been adapted, using soil taxonomy, texture, drainage, slope and elevation for the estimation of organic carbon (Tan *et al.* 2004). It was discovered that a relationship

could be established between organic carbon and these factors, especially for cropland.

Research has been conducted in Australia to estimate a range of soil properties, including organic carbon (Henderson *et al.* 2001). The nation-wide database had 11,483 soil points available to predict organic carbon in the soil. An enhanced decision trees tool (Cubist), catering for continuous outputs was used for this study. A correlation of up to 0.64 was obtained between the predicted and actual organic carbon levels.

Decision Trees have been used in other environmental programs such as salinity mapping (Walklate 2002). An advantage of using decision trees is that a rule base can be extracted from the tree and used to provide an explanation for decisions. Typically decision trees are used to classify examples into classes.

Artificial Neural Networks have been successful in the classification of other soil properties, such as dryland salinity (Spencer *et al.* 2004). Due to their ability to solve complex or noisy problems, Artificial Neural Networks are considered to be a suitable tool for a difficult problem such as the estimation of organic carbon in soil.

There are three benefits of modelling organic carbon. Firstly, this may lead to the identification and greater understanding of the factors that influence the amount of organic carbon in soil. Secondly, an accurate model using all available factors including the relationship with soil physical attributes will reduce the need for laboratory analysis of organic carbon and therefore reduce the costs associated with this analysis. Thirdly, models using only remotely sensed factors permits organic carbon estimates to be made anywhere the remote data is available rather than only at points where suitable soil profile descriptions are available This allows for relatively inexpensive estimation of organic carbon and the extrapolation and estimation of organic carbon over large areas.

4. EXPERIMENTATION

4.1. Data

A Victoria-wide soil database was provided by the Department of Primary Industries (DPI), Victoria. This database contains about 1,600 soil profiles with organic carbon readings. A map showing the organic carbon observations for the dataset is shown in Figure 4.

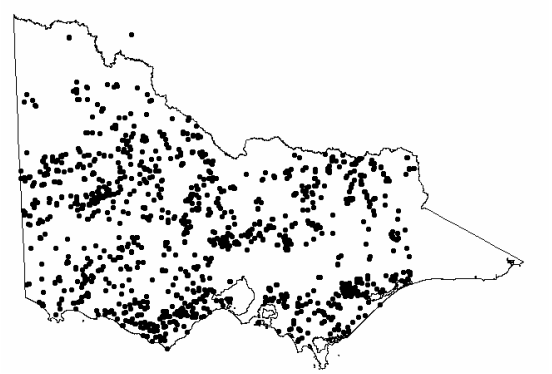

Figure 4. Soil profiles and organic carbon readings recorded in Victoria for the DPI dataset.

4.2. Data Selection

Two different sets were created from the original dataset. The first of these new datasets (*Vic-All*) contained all site survey factors, soil physical and chemical attributes, climate and landform data. The second dataset (*Vic-Remote*) contained landscape data, predominantly derived from Digital Elevation Models (DEMs), as well as climatic data. This was used to evaluate how well a model could estimate organic carbon without utilising soil survey data, and to determine the potential for mapping organic carbon over large study areas.

The input factors used in both the Vic-All and Vic-Remote datasets were: elevation, annual rainfall, annual temperature, relative elevation, slope, flow accumulation, topographical wetness, evaporation, and cation exchange capacity. The additional factors only used in the Vic-All dataset were: fine sand, coarse sand, silt, clay and pH.

The available examples were distributed into training and testing sets. Two-thirds of the available examples were placed in the training set and a third in the testing set.

4.3. Models

The three techniques used in this study are:

- Decision Trees (DTs),
- Multiple Linear Regression (MLR),
- Artificial Neural Networks (ANNs)

The first series of experiments used ANNs and MLR to predict a continuous output indicating the percentage of organic carbon. The results were evaluated using measures such as relative error and correlation between the predicted and correct output.

The second series of experiments used DTs and ANNs to classify the data into classes. The organic carbon readings were grouped into classes as follows:

- 3-Class organic carbon classification (low, moderate and high carbon content)
- 5-Class organic carbon classification (very low, low, moderate, high, very high)

Examples in the Decision Tree experiments were allocated to classes based on the percentile of the output class. For the 3-class study, 'low', 'medium' and 'high' were determined based on percentiles of the output class, with approximately one-third of examples in each class. Similarly in the 5-class study, 'very low', 'low', 'moderate', 'high' and 'very high' were based on percentiles with approximately one-fifth of the examples in each category.

In the Artificial Neural Network experiments the same approach was used to allocate examples to classes as for Decision Trees. A single output neuron was used for all of the classes. In the 3-class study, the 'low' examples were trained to a value of 0.1, 'moderate' as 0.5 and 'high' as 0.9. Estimated results within 0.2 of the correct result were considered correct. In the 5-class study, the 'very low', 'low', 'moderate', 'high' and 'very high' classes were trained to values of 0.1, 0.3, 0.5, 0.7, and 0.9 respectively. Estimated results within 0.1 of the correct result were considered correct.

4.4. Evaluation

The testing sets were used to evaluate the effectiveness of techniques for classifying or predicting organic carbon. The percentage of examples correctly estimated was used as the performance measure for the experiments classifying examples into 3 or 5 classes.

For the prediction of the percentage of organic carbon using a continuous output, statistics were calculated based on the difference between the correct and estimated output. In order to evaluate the models ability to predict organic carbon, the average error, relative error, correlation, R^2 and RMSE were calculated to assist in the interpretation of results.

Root Mean Squared Error (RMSE) estimates the standard deviation of errors within the testing set. The average error is the average absolute error between the model's estimated output and its correct (target) output for all of the samples in the testing set. The relative error is calculated relative to the testing set's mean. Equations for the root mean squared error (RMSE) (Equation 2), average error (Equation 3) and relative error (Equation 4) are shown below.

$$RMSE = \sqrt{\frac{1}{m}\sum_{j=1}^{m}\left(y_j - \hat{y}_j\right)^2} \quad (2)$$

$$average\ \ error = \frac{1}{m}\sum_{j=1}^{m}\left|y_j - \hat{y}_j\right| \quad (3)$$

$$relative\ \ error = \frac{\sum_{j=1}^{m}\left|y_j - \hat{y}_j\right|}{\sum_{j=1}^{m}\left|y_j - \bar{y}_j\right|} \quad (4)$$

4.5. Experimental Results

Two series of experiments were conducted. The first series used Multiple Linear Regression and Artificial Neural Networks to predict the percentage of carbon. Two datasets were used for these experiments, Vic-All and Vic-Remote. The second series of experiments used Decision Trees and ANNs to classify the output into 3 or 5 classes using the Vic-All and Vic-Remote datasets

Estimating the Percentage of Organic Carbon

The 2 datasets described in Section 4.2 were used to evaluate how well each model could estimate organic carbon content. Tables 1 and 2 show the experimental results for the Victoria-All (Vic-All) and Victoria-Remotely-Sensed (Vic-Remote) datasets using Multiple Linear Regression (MLR) and Artificial Neural Networks (ANNs).

Table 1. Results using Multiple Linear Regression (MLR) and Artificial Neural Networks (ANNs) for the Vic-All testing data.

	RMSE	**Av. Err**	**R**	**R^2**	**Rel. Err**
MLR	2.19	1.29	0.73	0.54	0.60
ANN	2.17	1.25	0.82	0.67	0.56

Table 2. Results using Multiple Linear Regression (MLR) and Artificial Neural Networks (ANNs) for the Vic-Remote testing data

	RMSE	**Av. Err**	**R**	**R^2**	**Rel. Err**
MLR	2.09	1.44	0.54	0.29	0.88
ANN	1.94	1.23	0.59	0.35	0.75

Classifying Organic Carbon into 3 and 5 Classes

Experiments were carried out to investigate the potential for Decision Trees and Neural Networks for the classification of organic carbon. The experimental results for the Vic-All dataset are shown in Table 3, and the Vic-Remote set in Table 4.

Table 3. Percentage of testing examples correctly classified for the Vic-All set using Decision Trees (DT), and Artificial Neural Networks (ANN).

	3-Class	5-Class
DT	70.8%	54.7%
ANN	80.2%	67.9%

Table 4. Percentage of testing examples correctly classified for the Vic-Remote set using Decision Trees (DT), and Artificial Neural Networks (ANN).

	3-Class	5-Class
DT	57.3%	44.1%
ANN	62.3%	41.6%

4.6. Discussion

Multiple Linear Regression and Artificial Neural Networks were used for the prediction of the percentage of organic carbon in the topsoil. Both techniques performed well indicating that they were able to establish an association between the input factors and the level of organic carbon. In all of the experiments the ANN approach outperformed MLR. This improvement is encouraging for further research in organic carbon estimation using Artificial Intelligence. The results shown in Table 1 demonstrate that with soil survey and chemical data, high correlations and a low error can be achieved. Such a model could replace chemical analysis if small error tolerances were permitted.

Two techniques were employed (Decision Trees and Artificial Neural Networks) for classifying organic carbon. Neural Networks performed best in three of the four classification experiments. Decision trees performed better than ANNs on the Vic-Remote data using 5 output classes.

In all of the experiments, better results were obtained using the Vic-All dataset than the Vic-Remote dataset This was expected, as the additional soil survey and chemical analysis factors present in the Vic-All dataset provide additional information for the techniques to make associations between the input factors and the organic carbon levels. When the soil survey and chemical factors were removed, and only remotely sensed factors were used, determining the organic carbon content was more difficult.

When using only remotely sensed data to estimate organic carbon, a moderate relationship can also be established. When using automated and complete data such as DEM and climate data, an indication of general organic carbon trends can be mapped. The accuracy of such maps would be less than the model that uses the soil survey factors, however the cost of production would be less as automated data is solely used in this model.

Due to the different evaluation techniques, it is difficult to compare between the classification and prediction approaches. The prediction approach provides a quantitative measure of organic carbon content, whereas the classification provides a less precise, more descriptive measure. The high correlations in some of the studies show that it is possible to estimate the percentage of carbon within topsoil.

5. CONCLUSION

Experiments have been carried out to evaluate the performance of Artificial Intelligence and statistical techniques to estimate organic carbon in topsoil. A study was conducted using data collected from across Victoria. Remotely-sensed data, as well as data obtained from soil-surveys and chemical analysis were used to evaluate the performance of the techniques.

When predicting the percentage of organic carbon, ANNs outperformed MLR in all experiments. ANNs were used for classification of organic carbon into three or five classes and out performed DTs in three out of the four experiments.

The use of soil survey and chemical analysis factors have been shown to improve performance in estimating organic carbon. However this approach requires that soil samples be taken and analysed, which may not be feasible. A more cost-effective approach is to use remotely sensed data for predicting organic carbon. This approach can be applied where suitable remotely sensed data exists, providing greater flexibility such as the ability to estimate organic carbon over larger areas.

The results of the Multiple Linear Regression experiments demonstrate that organic carbon can be estimated well using linear modelling. This is supported in the literature (Nichols 1984, Parton *et al.* 1987, Hontoria *et al.* 1999, Tan *et al.* 2003).

Further research could be conducted to assist in optimising the models for predicting organic carbon. In developing the dataset used in experiments, many examples had to be discarded due to missing values, reducing the size of the resultant dataset. Another problem with organic carbon data is that there is a dominance of soils that contain a low percentage of organic carbon making it is difficult for the technique to learn about examples containing high quantities of organic carbon.

Potential areas for future work are listed below:

- Implementing an approach that uses an ensemble of ANNs to assist prediction in different soil conditions,
- Improving data selection techniques to improve the generalisation,
- Applying the technique to data from a wider area to determine how the technique generalises
- Evaluating methods for replacing missing values to improve prediction,

6. REFERENCES

Batjes, N. (1996), Total carbon and nitrogen in the soils of the world, *European Journal of Soil Science,* 47, 151–163.

Bui, E., B. Henderson, and K. Viergever. (in Press), Knowledge discovery from models of soil properties developed through data mining, *Ecological Modelling*.

Climate Change Science Program and Global Change Research (2003), Strategic plan for the U.S. climate change science program, Technical report.

Henderson, B., E. Bui, C. Moran, and D. Simon (2004), Australia-wide predictions of soil properties using decision trees, *Geoderma,* 124, 383–398.

Henderson, B., E. Bui, C. Moran, D. Simon, and P. Carlile (2001), Asris: Continental-scale soil property predictions from point data, *Technical Report 28/01*, CSIRO, Canberra.

Hontoria, C., J. Rodriguez-Murillo, and A. Saa (1999), Relationships between soil organic carbon and site characteristics in peninsular Spain, *Soil Science Society of America Journal,* 63, 614–621.

Jenny, H. (1941), *Factors of Soil Formation*, McGraw-Hill, New York.

Kim, M., and T. Kim (2002), A neural classifier with fraud density map for effective credit card fraud detection, IDEAL, 378–383.

Kohonen, T., K. Torkkola, M. Shozakai, J. Kangas, and O. Venta (1988), Phonetic typewriter for Finnish and Japanese, *International Conference on Acoustics, Speech, and Signal Processing,* 1, 607–610.

Maier, H., and G. Dandy (1998), Understanding the behaviour and optimising the performance of backpropagation neural networks: an empirical study, *Environmental Modelling and Software,* 13(2), 179–191.

McCullagh, J. (2005), A modular neural network architecture for rainfall estimation, *Artificial Intelligence and Applications, Innsbruck, Austria*, 767–772.

McVay, K., and C. Rice (2002), Soil organic carbon and the global carbon cycle, *Technical Report MF-2548*, Kansas State University.

Nichols, J. (1984), Relation of organic carbon to soil properties and climate in the southern great plains, *Soil Science Society of America Journal,* 48, 1382–1384.

Parton, W., D. Schimel, C. Cole, and D. Ojima (1987), Analysis of factors controlling soil organic matter levels in great plains grassland, *Soil Science Society of America Journal,* 51, 1173–1179.

Post, W., T. Peng, W. Emanuel, A. King, V. Dale, and D. DeAngelis. (1990), The global carbon cycle, *American Scientist,* 78, 310–326.

Quinlan, J. (1993), *C4.5: Programs for Machine Learning*, Morgan Kaufmann, San Mateo, CA.

Spencer, M., T. Whitfort, and J. McCullagh (2004), Mapping dryland salinity using neural networks, *AI 2004*, Springer-Verlag, 1233–1238.

Tan, Z., R. Lal, N. Smeck, and F. Calhoun (2004), Relationships between surface soil organic carbon pool and site variables, *Geoderma,* 121, 187–195.

Walklate, J. (2002), Machine learning using AI techniques, BComp(Hons) thesis, School of Business and Technology, La Trobe University.

Zhang, G. (2004), *Neural Networks in Business Forecasting*, IRM Press, Hershey, PA.

Zhu, X. (1995), Multiple neural networks model and its application in pattern recognition, *IEEE International Conference on Neural Information Processing, Beijing,* 969–996.