STATISTICAL AND NEURAL METHODS FOR SITE–SPECIFIC YIELD PREDICTION

S. T. Drummond, K. A. Sudduth, A. Joshi, S. J. Birrell, N. R. Kitchen

ABSTRACT. *Understanding the relationships between yield and soil properties and topographic characteristics is of critical importance in precision agriculture. A necessary first step is to identify techniques to reliably quantify the relationships between soil and topographic characteristics and crop yield. Stepwise multiple linear regression (SMLR), projection pursuit regression (PPR), and several types of supervised feed–forward neural networks were investigated in an attempt to identify methods able to relate soil properties and grain yields on a point–by–point basis within ten individual site–years. To avoid overfitting, evaluations were based on predictive ability using a 5–fold cross–validation technique. The neural techniques consistently outperformed both SMLR and PPR and provided minimal prediction errors in every site–year. However, in site–years with relatively fewer observations and in site–years where a single, overriding factor was not apparent, the improvements achieved by neural networks over both SMLR and PPR were small. A second phase of the experiment involved estimation of crop yield across multiple site–years by including climatological data. The ten site–years of data were appended with climatological variables, and prediction errors were computed. The results showed that significant overfitting had occurred and indicated that a much larger number of climatologically unique site–years would be required in this type of analysis.*

Keywords. *Neural networks, Precision agriculture, Prediction, Regression analysis.*

It has long been known that spatial variability exists in agricultural fields in soil properties, topographic characteristics, and crop yields. Precision agriculture aims to improve production efficiency by adjusting crop treatments to these localized conditions within the field. Therefore, the success of precision agriculture depends on accurate and detailed knowledge of yield potential and crop response to specific conditions.

Several means exist to investigate these relationships. Agronomic methods, involving numerous small plot trials over multiple site–years, are the most traditional and possibly the best means to compile the necessary data. However, they are extremely time consuming and labor intensive and seem unrealistic for the near future. Another approach that may hold promise for understanding these response relationships involves the use of mechanistic crop growth models (i.e., Mathews and Blackmore, 1997; Fraisse et al., 2001). However, the current usefulness of crop models is somewhat limited due to their extreme development expense, both in terms of time and money, and the fact that yield simulations from these models can be unreliable (Varcoe, 1990).

A third method of investigating yield response consists of empirical analysis of large, spatial, multivariate data sets — just the type of data sets collected in precision agriculture. Linear analyses of such data sets have often been reported in the literature. Several authors have found that linear correlations between yield and soil properties, or between two soil properties, vary greatly both within and between fields (Pierce et al., 1994; Drummond et al., 1995; Khakural et al., 1999) and can also exhibit strong temporal variability (Lamb et al., 1997). Most authors have not found correlation analyses to be very useful in understanding and quantifying yield response.

More complex linear methods, including various forms of multiple linear regression, have been widely considered (i.e., Kravchenko and Bullock, 2000; Khakural et al., 1999; Kitchen et al., 1999; Drummond et al., 1995), with generally poor results. Sudduth et al. (1996) used linear techniques on a data set consisting of several site–years of topographic, soil, and yield data and found that linear methods generally failed to produce good approximations of spatial yield variability, even within sub–field regions thought to be reasonably homogenous. Other researchers (i.e., Khakural et al., 1999) have found linear techniques able to perform reasonably well on some data sets. Multiple linear regression techniques using polynomial and interaction terms have also been considered (Kitchen et al., 1999) with some improvement over strictly linear models.

A variety of nonlinear techniques for investigating yield response have also been investigated, including boundary

Article was submitted for review in February 2002; approved for publication by the Power & Machinery Division of ASAE in November 2002.

Mention of trade names or commercial products is solely for the purpose of providing specific information and does not imply recommendation or endorsement by the USDA, University of Maryland, or Iowa State University.

The authors are **Scott T. Drummond,** Computer Specialist, **Kenneth A. Sudduth, ASAE Member Engineer,** Agricultural Engineer, and **Newell R. Kitchen,** Soil Scientist, USDA–ARS Cropping Systems and Water Quality Research Unit, Columbia, Missouri; **Anupam Joshi,** Associate Professor, Computer Science and Electrical Engineering Department, University of Maryland – Baltimore County, Baltimore, Maryland; and **Stuart J. Birrell, ASAE Member Engineer,** Assistant Professor, Agricultural Engineering Department, Iowa State University, Ames, Iowa. **Corresponding author:** Scott T. Drummond, 269 Agricultural Engineering Building, University of Missouri, Columbia, MO 65211; phone: 573–882–1146; fax: 573–882–1115; e–mail: drummonds@missouri.edu.

line analysis (Kitchen et al., 1999), state–space analysis (Wendroth et al., 1999), Bayesian networks, and regression trees (Adams et al., 1999). However, many nonlinear methods can be difficult to implement, and comparison of the results from these vastly different methods is problematic. Clearly, nonlinear methods that are relatively easy to implement and can be readily compared to one another would be highly desirable.

A relatively new branch of nonlinear techniques, artificial neural networks (ANN or NN), has been applied not only to artificial intelligence (Rumelhart and McClelland, 1986) and classification applications (Burks et al., 2000) but also as general, non–parametric "regression" tools. A neural network consists of layers of highly interconnected processing units, each containing a small amount of local "memory." The network is trained using an iterative method to adjust the weights of connections between these units. Network types, topologies, and training techniques vary considerably, but a rudimentary explanation of the critical aspects of backpropagation neural networks is contained in Burks et al. (2000). A pedagogy of neural networks can be found in Rumelhart and McClelland (1986).

Liu et al. (2001) used a standard backpropagation neural network to estimate corn yields over a number of years of small plot data, which included soil, weather, and management factors. Their results were promising, with predictive errors reported to be approximately 20% of the actual yield, although only a single validation set was used. Shearer et al. (1999) investigated a relatively large number of variables, including fertility, satellite imagery, and soil conductivity, for a relatively small number of observations in one site–year of data with limited success. Sudduth et al. (1996) applied a number of analysis methods, including backpropagation neural networks, on several large data sets and achieved higher training accuracies than linear methods could provide, but lower training accuracies than with projection pursuit regression (PPR), another nonlinear non–parametric technique (Friedman and Stuetzle, 1981). In our initial work (Drummond et al., 1998), we investigated a number of different supervised feed–forward neural methods on one site–year of the data set used by Sudduth et al. (1996) and found that several of these neural methods could provide training accuracies nearly as high as those produced by PPR.

A critical consideration in the evaluation of regression techniques is the selection of a fair means of comparing the methods chosen for analysis. The most commonly used criterion of model performance has been the coefficient of determination (R^2), which is the ratio of the variance explained by the model to the total variance in the data. However, a number of authors have concluded that R^2 is not a good means of comparison between models representing yield response (i.e., Cerrato and Blackmer, 1990). Numerous reasons have been cited as to why R^2 is not a suitable measure; however, a primary consideration is the fact that it gives no indication of how well a model performs when applied to data that were not used to create the model. Overfitting, best described as "good performance on calibration data, but poor performance on test data," is often the result.

Another commonly used measure of accuracy is the root mean square error (RMSE), more commonly referred to as the standard error (SE). A major advantage of using SE over R^2 for model evaluation is that it can be calculated not only on the calibration data (termed the standard error of calibration, or SEC) but also on any "new" data set not used in developing the model (termed the standard error of prediction, or SEP) to estimate true predictive ability. While training accuracies in terms of R^2 and/or SEC have almost always been reported, predictive statistics (such as SEP) are often not reported nor even considered. When they are considered, various rules–of–thumb are applied in an attempt to avoid overfitting. A common means of measuring prediction accuracy is to use a "split–sample" approach, in which a subset of the data is withheld from training. A measure of the accuracy of prediction on this validation set is then reported. In small data sets, or those that include outliers, this single measure can be quite misleading.

Cross–validation is a more robust, reliable method of measuring prediction accuracy (Stone, 1973). In *k*–fold cross–validation, the data are divided into *k* subsets of equal size. The regression technique is then applied *k* times, each time leaving out one of the subsets and using only that subset to compute the generalization statistic (e.g., SEP). Cross–validation has generally been accepted as superior to split–sample techniques, particularly on small data sets (Goutte, 1997). However, in practice the selection of *k* is a difficult matter. An overly large choice of *k* can make the computational problem prohibitive. Too small a choice of *k* may cause the prediction error estimate to be unreliable. In our initial work (Drummond et al., 1998), we were able to achieve acceptable results using a five–fold cross–validation technique on a single site–year of data, while keeping the computational complexity of the problem manageable.

Objectives

The objectives of this study were to evaluate the predictive ability of representative linear, nonlinear, and neural network techniques on a multiple site–year data set of grain yield and site and soil characteristics in order to: (1) identify those techniques that provided the most accurate predictions within each individual site–year and establish reasonable estimates of prediction accuracy for those techniques, and (2) analyze those same data sets, concatenated with appropriate climatological data, to evaluate the ability of the methods to estimate yield across multiple site–years.

Materials and Methods

Data were collected on three fields located near Centralia in central Missouri. These fields (hereafter referred to as fields 1, 2, and 3) were 36, 28, and 13 ha in size, respectively. The soils found on these fields are characterized as claypan soils of the Mexico–Putnam association. The predominant soils were Mexico (Aeric Vertic Epiaqualfs), Adco (Vertic Albaqualfs), and Leonard (Aeric Vertic Epiaqualfs). Surface textures ranged from silt loam to silty clay loam. The subsoil claypan horizon(s) were silty clay loam, silty clay, or clay. Within each study field, topsoil depth above the claypan ranged from less than 10 cm (highly eroded) to greater than 100 cm (depositional), with a small area of field 2 having as much as 250 cm of topsoil. Because of the high–clay subsurface horizons and their effect on soil water–holding capacity, infiltration, and rooting, topsoil depth above the claypan is often correlated with spatial variation in crop productivity (Kitchen et al., 1999).

Composite soil samples (eight 6–in. deep cores per sampling point) were taken on a 30 m grid in the spring of 1995 for field 1, on a 25 m grid in the spring of 1996 for field 2, and on a 25 m grid in the spring of 1997 for field 3. The samples were analyzed by the University of Missouri Soil and Plant Testing Services Laboratory, using the methods prescribed by Brown and Rodiguez (1983). Soil properties measured included soil pH (salt method), organic matter, phosphorus, calcium, magnesium, potassium, and cation exchange capacity (CEC).

In addition to the nutrient and soil property data, detailed topographic data were obtained for each field using a Nikon Topgun A200LG total station surveying instrument (vertical accuracy <1 cm). Data were sampled on a rectangular 25 m grid, with additional densification on and around breaklines. Between 500 and 1000 data points were used to characterize each field. The elevation data were then interpolated to 10 m grids using block kriging with appropriate semivariograms (GS+ v5.1, Gamma Software Design, Plainwell, Mich.). Slope was calculated from these 10 m grids with Surfer v7.0 (Golden Software, Golden, Colo.) using terrain modeling methods outlined by Moore et al. (1993). Elevations across the three fields were quite similar, falling between 257 m and 266 m above sea level, with elevation ranges of 3.4, 3.3, and 6.9 m for fields 1, 2, and 3, respectively. Maximum computed slopes within the three fields were 1.6%, 1.5%, and 2.8%, respectively.

Topsoil depth above the claypan was estimated for each field from soil electrical conductivity, using a mobile electromagnetic induction sensing system and methods described by Sudduth et al. (2001). Grain yield measurements were obtained using a full–size combine equipped with a commercial yield sensing system and global positioning system (GPS) receiver, using data processing techniques described by Birrell et al. (1996). Yield and topsoil depth data were interpolated to the same 10 m grid as used for the topographic data, using block kriging with appropriate semivariograms. The "point" soil data were then merged with the gridded topographic, topsoil depth, and yield data by selecting the 10 m cell whose center was nearest the location where the soil sample was taken. Figure 1 shows an example of one complete dataset (site–year 6), demonstrating the degree of variability present in soil properties, topography, and yield.

Yield data were collected on these fields between years 1993 and 1997. Table 1 shows the fields and years for which yield data were available, the crop harvested, descriptive yield statistics, and the assignment of identification numbers to individual site–years. These identification numbers will be used hereafter for reference. Five–fold cross–validation sets were created for each site–year of data, creating a total of 50 training/validation pairs for individual site–year analysis.

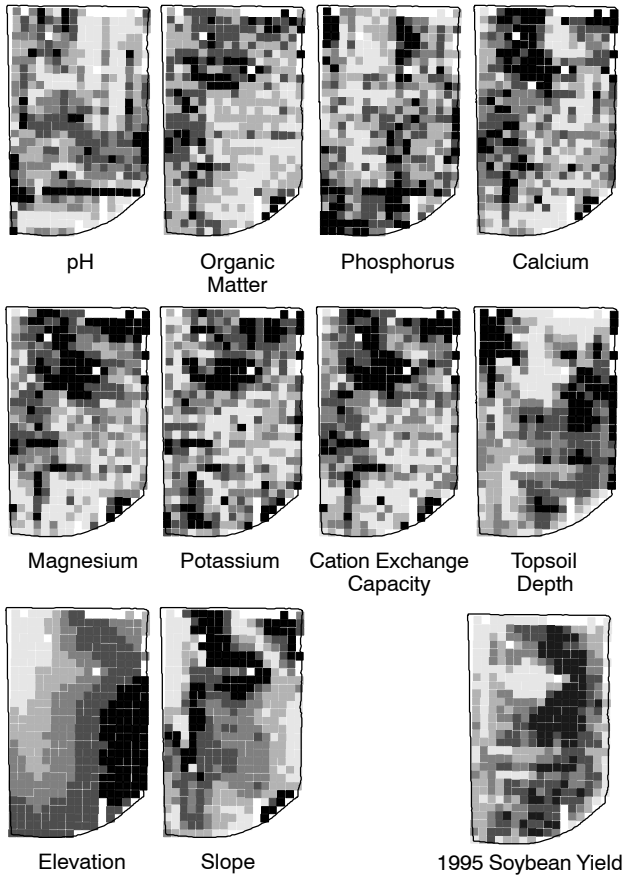

Figure 1. Sample data set (site–year 6) showing all soil/topography parameters and crop yield.

Climatological data were available from an onsite weather station, located within 8 km of all three fields. Considering the limited number of site–years available, a parsimonious selection of climatological variables was required. The plant reproductive phase (defined as days 51–110 after planting) was divided into four 15–day intervals, and rainfall during each period was summed. In addition, average daily high and low temperatures during the entire plant reproductive phase were included. Table 2 summarizes this climatological information for each site–year.

Since we reasoned that significant overfitting was likely, even with these few climatological variables, a second analysis was run with only one climatological variable —

Table 1. Identification and yield information for site–years used in this study.

Site–year	Field	Year	Crop	Minimum Yield (kg/ha)	Maximum Yield (kg/ha)	Yield SD (kg/ha)	Number of Observations
1	1	1993	Corn	5042	9008	683	315
2	1	1994	Soybean	1044	3008	289	340
3	1	1995	Sorghum	2910	6835	642	300
4	1	1996	Soybean	2019	3510	221	355
5	1	1997	Corn	3888	9612	987	335
6	2	1995	Soybean	452	3209	542	435
7	2	1996	Corn	3767	12678	1127	265
8	2	1997	Soybean	1244	3060	319	420
9	3	1996	Soybean	1537	4119	440	185
10	3	1997	Corn	1197	10872	1895	170

Table 2. Climatological data from the plant reproductive phase for all site–years.

Site–year	Planting Date	Temperatures (°C) Mean[a] Highs	Temperatures (°C) Mean[a] Lows	Rainfall (mm) Days[a] 51–65	Days[a] 66–80	Days[a] 81–95	Days[a] 96–110	Days[b] 51–110
1	15 May 1993	30.3	20.1	102	42	99	33	275
2	21 May 1994	29.1	16.7	5	13	0	64	82
3	15 June 1995	26.3	15.6	141	6	38	32	217
4	14 June 1996	25.1	14.0	41	81	16	70	208
5	6 May 1997	28.8	17.9	10	6	8	94	119
6	1 June 1995	28.3	18.2	133	69	6	38	246
7	14 May 1996	27.6	17.4	5	65	36	86	192
8	14 May 1997	29.1	17.9	3	9	51	47	111
9	23 May 1996	27.5	17.3	61	33	96	5	195
10	6 May 1997	28.8	17.9	10	6	8	94	119

[a] Six climatological variables used in initial multiple site–year analysis.
[b] Single climatological variable used in subsequent multiple site–year analyses.

total rainfall during the plant reproductive phase. For the multiple site–year analysis, five–fold cross–validation sets were created from the soybean site–years (five) and corn site–years (four), with each training set including 80% of the data, randomly selected from each site–year, and each validation set containing the remaining 20% from each site–year. To evaluate the severity of overfitting on the relatively sparse climatological data, leave one site–year out (LOO) cross–validation sets were also created by holding each site–year out in turn, with all remaining site–years for that crop included in the training set.

In summary, the data sets created for analysis consisted of: (1) 50 individual site–year training/validation sets including 10 soil/topographic predictor variables, (2) 19 multiple site–year training/validation sets including 16 soil/topographic/climatological predictor variables, and (3) 19 multiple site–year training/validation sets including 11 soil/topographic/climatological predictor variables. Drummond (1998) provides more detailed descriptions of these data sets and compilation procedures.

Analysis Methods

Drummond et al. (1995) used several different linear methods on a subset of these individual site–years, including multiple linear regression, stepwise multiple linear regression, and partial least squares regression. Results among these linear methods were similar and were inferior to those from nonlinear methods. On all three fields, significant correlations were observed among most of the independent variables. Although we realized that the colinearity present in these variables could present difficulties (Neter et al., 1985), we proceeded with a representative linear method to provide comparison with the two nonlinear methods. Stepwise multiple linear regression (SMLR) was chosen to provide an estimate of the prediction accuracy of linear methods, since it performed well in the previous study, and since it is a well documented and easily implemented method (SAS, 1989).

The SMLR algorithm was applied to all 88 training data sets, and in each run all predictor variables were available for entry into the model. For each model size, the regression parameters were compiled and applied to the corresponding validation data set to calculate cross–validation error results. Since there was a significant amount of scatter between the results from the five cross–validation sets, we averaged the error results from each validation set at each model size to get a less noisy representation of generalization error, the cross–validated mean squared error (CVMSE). Within each site–year, the model size that provided the lowest CVMSE was determined, and the associated standard error of prediction (SEP) was calculated. A single SMLR model was then created for each site–year by training on all observations for that site–year and limiting the model to the size that produced the lowest CVMSE for that site–year. The associated standard error of calibration (SEC) was then calculated.

Projection pursuit regression (PPR) (Friedman and Stuetzle, 1981), a nonlinear, non–parametric regression technique, was also investigated. This was done due to the results reported by Sudduth et al. (1996), which showed that PPR was able to consistently produce the highest training accuracies on their data sets. Additionally, PPR was easy to implement and required that only a few parameters be set. In PPR, the regression response surface is modeled as the sum of a set of general smooth functions of linear combinations of all of the predictor variables. The process continues iteratively, finding and adding smooth terms until some user–defined threshold is reached. In the S–PLUS implementation of PPR (Mathsoft, 1997), several parameters affect the development of these smooth functions, including the width of the data span to be used in creating the smooth functions, the amount of high–frequency variation allowed on the smooth function, and an optimization level that determines the thoroughness of the optimization routine. Additionally, the number of predictor variables used in the creation of these latent variables can be manually adjusted. Initial investigations into the use of PPR on a test data set (site–year 2) (Drummond, 1998) provided some insight into parameter selection. As a result, the analysis for each data set was performed with three different optimization levels, using either four, seven, or ten predictor variables and with up to nine latent variables allowed. The PPR algorithm was applied to the same 88 training data sets, and the validation sets were used to calculate CVMSE and SEP results.

A variety of supervised feed–forward neural network training techniques were selected for investigation, including several variations of backpropagation — standard, batch, momentum (Rumelhart and McClelland, 1986), and weight decay (Werbos, 1988) — as well as quickprop (Fahlman, 1988) and resilient backpropagation (rprop) (Riedmiller and Braun, 1993). These methods were selected primarily due to

their abundant descriptions in the literature and/or due to their reported effectiveness on a variety of problems. The analysis for each of the chosen methods was carried out with the aid of the Stuttgart Neural Network Simulator (SNNS) software package (SNNS Group, 1997). Each technique investigated had a few (2 to 4) model–specific training parameters defined in the SNNS implementation. The purpose and suitable ranges of each of these parameters were quite variable; however, basic descriptions of these parameters and reasonable starting ranges have been suggested (SNNS Group, 1997). In our previous research (Drummond et al., 1998), a coarse parameter–space search for suitable model–specific parameters was performed on the site–year 2 data set. For this work, similar values for these parameters were selected. While parameter ranges were carefully and systematically explored, the parameters used in this study were not necessarily optimal. Therefore, methods that required fine–tuning of several critical parameters might have been at a disadvantage. In general, we were trying to locate training parameters that would begin to overfit the data within the length of training time allowed, since we were interested in finding the minimum prediction error for each of these methods as opposed to optimizing training accuracies.

Network topology was limited to a single hidden layer, since networks of this complexity can emulate any arbitrary function (Hornik et al., 1989). Within this hidden layer, it was necessary to determine the appropriate number of processing units. Five reasonable hidden layer sizes ranging from 4 to 25 hidden units were selected, and initial networks were created with randomized weights for each of these network sizes. Each network created included 10 to 16 input nodes (depending on whether and which climatological variables were included) and one output node. All neural methods were analyzed using the same network files, including the same initialization weights. The effect of scaling of the input parameters was also investigated. All network sizes were analyzed with all observation variables scaled to a [0–1] range. In addition to this, the network with 10 hidden units was analyzed on the same data sets but compressed to [0.1–0.9] and [0.2–0.8] ranges.

For multiple site–year data sets, the data were scaled only to a [0–1] range, representing the maximum range of variability present for each variable over all site–years. Each of these combinations was applied to the single and multiple site–year training data sets. Individual observations were presented to the network in a random order within each epoch, with an epoch consisting of a single pass through each observation in the training set. Training on each data set continued until 5000 epochs had elapsed, with both training and validation set mean squared error (MSE) calculated and reported every 10 epochs. As with the results for SMLR and PPR, the validation MSE results from each set at each training stage were averaged to get a less noisy representation of generalization error, the CVMSE. Within each site–year, the model that provided the minimal CVMSE was determined, and the associated standard error of prediction (SEP) was calculated.

Results and Discussion

Statistical Methods

Figure 2 shows CVMSE results for SMLR analysis of site–years 1 to 5, where a relatively large number of terms could be included in the regression model without fear of overfitting. In every site–year, the minimal CVMSE occurred with at least four predictor variables included, while minimal CVMSE results for most site–years indicated that at least eight terms could be included. The SEP values achieved by the optimal SMLR model in each site–year are listed in table 3. Predictive variables that entered into the models were quite variable across site–years, with only magnesium entering into the model in every site–year. Topsoil depth

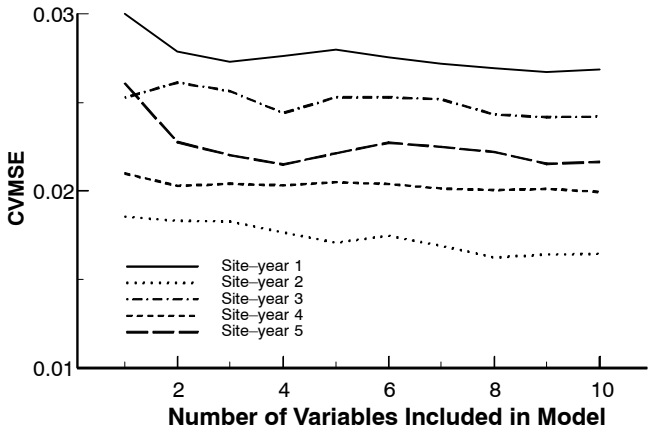

Figure 2. Results from stepwise multiple linear regression (SMLR) analysis, showing cross–validated mean square error (CVMSE) as a function of model size for each of site–years 1 through 5.

Table 3. Comparison of prediction results for neural network (NN), projection pursuit regression (PPR), and stepwise multiple linear regression (SMLR) methods.

Site–year	Yield SD (kg/ha)	SMLR SEP[a] (kg/ha)	PPR SEP[a] (kg/ha)	NN SEP[a] (kg/ha)	Best Neural Method[b]	Number of Epochs
1	683	648	626	602	quickprop	2420
2	289	250	199	196	rprop	3650
3	642	610	577	555	rprop	1910
4	221	211	209	205	weight decay	1760
5	987	840	704	693	rprop	3360
6	542	406	345	334	standard	3850
7	1127	1033	1034	1024	rprop	750
8	319	281	268	257	momentum	2790
9	440	366	366	365	rprop	70
10	1895	1194	1114	1088	rprop	240

[a] Standard error of prediction.
[b] Neural network training method that provided the smallest SEP for that site–year.

entered in nine of the ten site–years, with elevation, slope, and phosphorus occurring in eight of the models. The least common entry into the models was CEC, which was included in only six of the ten site–years.

The SEP results achieved with this cross–validation analysis were almost identical to the SEC results reported by Sudduth et al. (1996) for the SMLR analysis of the subset of site–years that they investigated. This may indicate that the methods they used to avoid overfitting using SMLR were successful. Nevertheless, the results achieved by SMLR were somewhat disappointing, with most site–years showing SEP results only slightly better than the standard deviation of crop yield.

PPR analysis results are also included in table 3. In certain site–years (2, 5, and 6), there was a significant reduction in SEP compared to the SMLR analysis. However, there were also site–years (4, 7, and 9) in which no improvement, or even a slight increase in SEP, was seen with PPR. All three of these site–years (4, 7, and 9) occurred during the 1996 cropping season, a year in which crop water stress was minimal. The PPR parameters that produced minimum SEP values were highly different among the site–years, and little information could be gleaned about optimal tuning of the parameters. While the nonlinearity introduced by PPR allowed some improvement in SEP compared to SMLR analysis, the improvement was not seen for all site–years. The SEP results for PPR were considerably poorer than the SEC results reported by Sudduth et al. (1996) on the subset of site–years that they investigated. These two points make it clear that overfitting is a concern when using PPR, and that some method of avoiding or at least measuring this overfitting is critical.

Neural Network Methods

Curves of the neural network CVMSE for each method and site–year showed some obvious differences (see fig. 3 for two such site–years). The CVMSE values for several variations of backpropagation (standard, momentum, and weight decay) were quite noisy, with relatively small but rapid changes over the entire training period. Batch backpropagation tended to provide a relatively smooth curve, but often appeared to train so slowly that a minimum validation error was not reached within the allotted training time, particularly on larger data sets, such as site–year 5. Quickprop and rprop trained more rapidly and provided smooth error curves, with results generally better than the average of the other methods. The data from all 60 method/site–year combinations was further summarized by finding the minimum SEP values for each of these curves (table 3). In six of the ten site–years, rprop produced the overall minimum SEP of the methods investigated, and it was very close to the minimal value in three other site–years. In the four site–years where rprop was not minimal, four different neural methods provided the optimal solution. Only batch backpropagation was unable to provide the best solution in any site–year. The best scaling range overall was [0–1]. Of the 60 minimal method/site–year combinations, only five used a scaling range other than [0–1], none of which provided the overall best solution in any site–year. The effect of hidden layer size was more complex, as sizes ranging from 4 to 25 hidden units were able to provide reasonable results in at least some cases. Networks with 4 hidden units were too parsimonious, providing minimal solutions only about 10% of the time.

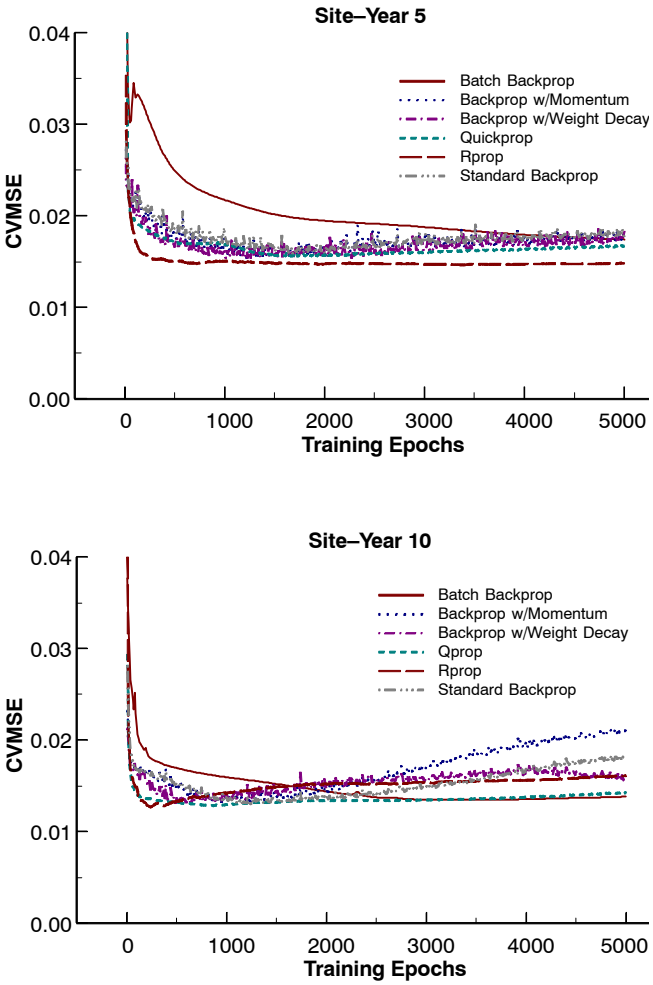

Figure 3. Cross–validation mean squared error (CVMSE) results from neural networks (NN) created with optimal network sizes and scaling parameters for each NN training algorithm, for two site–years.

Networks with 25 hidden units provided too much opportunity for overfitting, as they were only able to provide minimal solutions 12% of the time. Networks containing either 7 or 10 hidden units accounted for about 62% of the minimal solutions, with 7 hidden units being selected most often.

The amount of training time (measured in epochs) required to reach optimal solutions varied considerably, both between methods and across site–years. The variation in training time between neural methods could be attributed to the specifics of that learning algorithm or to the training parameters used, or to both, and are of little importance to our study. However, there was significant variation between the training times for different site–years. Although most site–years (1 through 6 and 8) produced minimal solutions at training times measured in thousands of epochs, site–years 7, 9, and 10 trained in very few epochs (table 3). These three site–years contained the smallest data sets, ranging from 170 to 265 observations each. The other seven site–years ranged from 300 to 435 observations (table 1). This indicated that the neural methods selected may tend to overfit small data sets rapidly. Figure 3 shows the results for two cornfields planted on the same date during the water–stressed 1997 crop year. CVMSE results for site–year 5 (335 observations) seem to be very stable and well behaved over a long period of training time, with overfitting occurring very gradually after the

optimal value was reached at 3360 epochs. However, for site–year 10 (170 observations), the optimal solution is achieved very rapidly (240 epochs), with evidence of significant overfitting after that point.

Comparison of Methods

The minimal neural network results, by site–year, compared favorably with results from SMLR and PPR (table 3). SEP values were significantly reduced in certain site–years by the use of NN. In fact, in every site–year, the minimal NN technique provided at least some improvement in terms of cross–validated SEP over both SMLR and PPR. In site–years 4, 7, and 9, there was very little difference in terms of SEP between the minimal SMLR, PPR, and NN results. All three of these site–years occurred in 1996, a year in which the crops received almost ideal weather conditions throughout the growing season. It is possible that the predictor variables in these data sets produced either no crop response or extremely linear crop responses in 1996. Another possibility is that much of the yield variation seen in this year on these fields was simply experimental error (noise) in the measurement of crop yield (or in the predictor variables) or possibly significant but unmeasured yield–limiting factors (e.g., weeds). A third possibility is that the data in one or more predictor dimensions were sparse enough that no advantage was realized by the introduction of nonlinearity.

Figure 4 demonstrates why data set size can be so critical for nonlinear techniques like NN and PPR. Consider a Gaussian function of the form $f(x) = a^*\exp(-b^*x^2)$ defined over the range [0–1]. The left graph of figure 4 shows a very sparse data set, consisting of only ten observations randomly distributed over the range of x. The right graph shows 100 randomly distributed observations. A linear method used to estimate the underlying function for each of these data sets will provide similar and reasonable results with no major changes seen in terms of prediction error. It is clear that overfitting is not a major concern for linear techniques on these two data sets. This is not the case for nonlinear techniques. To demonstrate, simple polynomial curves were developed for both data sets that fit the calibration data to a very high degree of accuracy ($R^2 > 0.99$). For the dense data set, the fit is quite good throughout the entire range of the predictor variable. However, overfitting is a clear and significant danger on the sparse data set, especially in regions where the curve is poorly defined, as can be seen near the origin in this case. Therefore, the nonlinear technique is likely to show a significant advantage over the linear technique, with respect to prediction error, only for the denser data set, as shown in the right graph of figure 4. Although this example used a simple, smooth curve, with no noise on the signal, the effects of data density are even more critical for complex relationships (i.e., periodic functions, functions with multiple local optima, multiple predictors, inherent measurement errors, etc.).

A NN model was created from the entire data set for each individual site–year using the parameters found to produce the minimal SEP results from the cross–validation analysis, including optimal training method, network size, scaling parameters, and length of training time. Our goal was to produce a model that matched the data as well as possible without overfitting. For most site–years, the results from the NN cross–validation analysis provided CVMSE curves that were extremely stable over a long period of training time, similar to those in figure 3. It seemed likely that a network trained on the entire data set, using the same parameters that produced minimal SEP results, would be unlikely to experience a high degree of overfitting. Likewise, SMLR models were created by selecting the optimal model size from the cross–validation analysis and performing SMLR on the entire data set. The dangers of significant overfitting for the SMLR method were also considered to be quite small. Due to the previous results, which indicated that the PPR method was very likely to overfit these data sets, no PPR models were created for this comparison. Figure 5 shows an example of the estimated yields from the trained NN and SMLR models versus the actual yield for site–year 6. The map from the NN model appears much more similar to the actual yield than that produced by SMLR and does not show any visual indications of overfitting.

Statistics were compiled for each site–year for both the NN and SMLR models (table 4). In general, the NN models were able to explain the largest percentage of variation in actual yield, with an average of 45% of the variation in yield explained by the model (range 21% to 74%, table 4). SMLR models were able to explain an average of 31% of the yield variation (range 14% to 65%). For the larger data sets

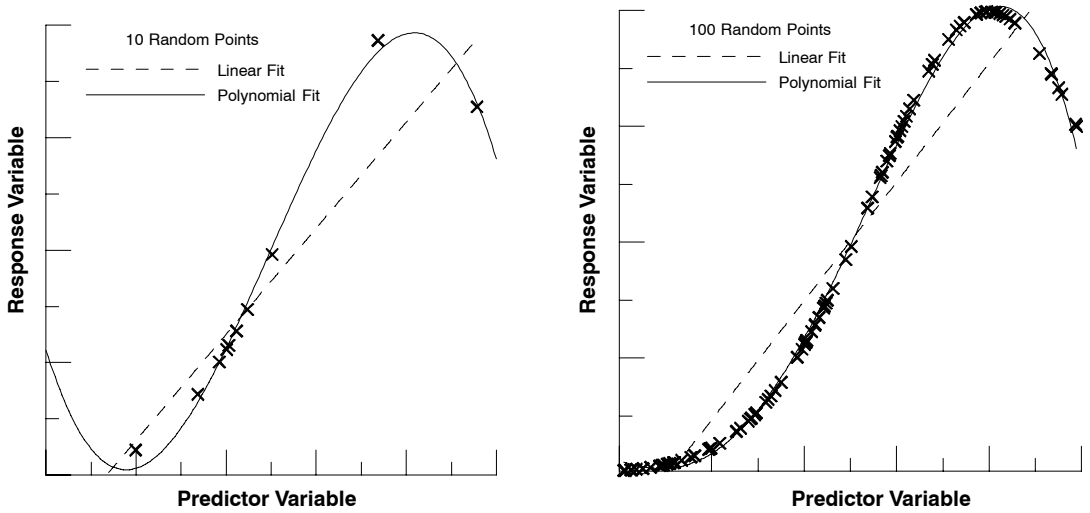

Figure 4. Effect of data density on nonlinear estimation techniques.

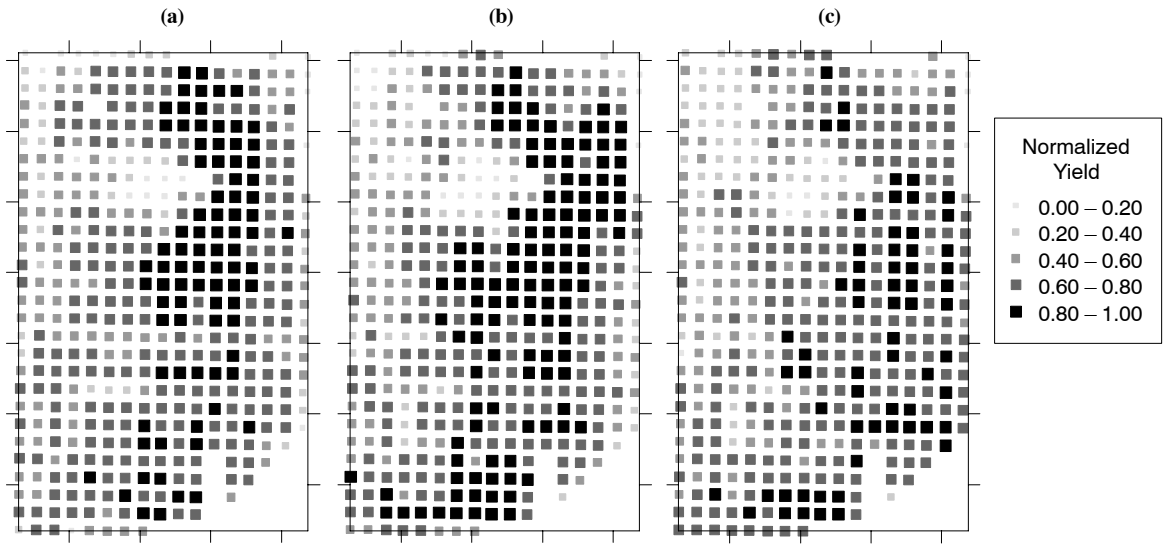

Figure 5. Visual comparison of (a) actual yield, (b) neural network predicted yield, and (c) stepwise multiple linear regression predicted yield for site–year 6 data.

(site–years 1, 2, 4, 5, 6, and 8), NN models provided significantly better fit, in terms of R^2, than did SMLR models. However, for smaller data sets (site–years 3, 7, 9, and 10), the results for NN were almost identical to those for SMLR. The SEC values for SMLR models were slightly lower (1% to 5%) than SEP results from the cross–validation analysis. However, SEC results for the NN models were much more variable, with values ranging from 4% higher to 19% lower than the SEP results. The three largest data sets (site–years 4, 6, and 8) provided the greatest improvement in SEC compared to SEP (13% to 19%). Although overfitting during training may have been responsible for these large discrepancies, the fact that the largest data sets provided large improvements over linear methods, and that small data sets provided little or no improvement, is clearly in line with the previous discussion of data density (fig. 4).

Multiple Site–Year Data Sets, Including Climatological Data

Results for the climatological study were mixed. In the first experiment, where 80% of the observations from each site–year were included in each training set with the other 20% going into the validation sets, the results were very positive. Both the PPR and NN methods had SEP results very similar in magnitude and, in the case of the corn site–years, even smaller than the weighted average SEP from the individual site–year study, indicating little or no loss in predictive ability due to the inclusion of multiple fields and climatological data (table 5). This suggests that the nonlinear methods had the capability to fit the additional complexities introduced by the inclusion of climatological data. On these large, multi–year cross–validation data sets, PPR was able to produce results that were nearly identical to those for NN but at a greatly reduced computational cost. By contrast, SEP results from SMLR were significantly higher than the weighted average SEP results for the single site–year experiments. Linear methods did not provide enough complexity to model the additional climatological variables and achieve a quality fit between crop yield, topography, and soil parameters. The left graph of figure 6 shows the strong relationship between the NN–predicted and actual yields for all validation observations, for all soybean site–years, with all six climatological variables included.

Table 4. Comparison between neural network (NN) and stepwise multiple linear regression (SMLR) model calibration results (R^2 and SEC) versus prediction results (SEP) compiled using cross–validation.

						SEP[b] (kg/ha)			
		R^2		SEC[a] (kg/ha)		NN		SMLR	
Site–year	N	NN	SMLR	NN	SMLR	Mean	Range[c]	Mean	Range[c]
1	315	0.31	0.16	567	637	602	532–732	648	620–786
2	340	0.59	0.29	187	246	196	161–222	250	207–290
3	300	0.21	0.2	573	583	555	430–700	610	535–767
4	355	0.35	0.14	179	207	205	168–248	211	186–267
5	335	0.57	0.3	647	832	693	627–791	840	781–977
6	435	0.74	0.47	278	399	334	298–383	406	363–501
7	265	0.23	0.22	989	1003	1024	807–1371	1033	843–1433
8	420	0.49	0.27	227	275	257	224–289	281	268–326
9	185	0.39	0.38	351	358	365	311–443	366	350–427
10	170	0.66	0.65	1132	1144	1088	824–1265	1194	863–1368

[a] Standard error of calibration.
[b] Standard error of prediction.
[c] Range of SEP values achieved across the five cross–validation sets.

Table 5. Cross–validation results from multiple site–year climatological data sets.

Crop	Method	Yield SD (kg/ha)	Mean SEP[a] (kg/ha)	Multi–year SEP[b] (kg/ha)	Multi–year LOO SEP[c] (kg/ha)
Soybean	SMLR	639	301	360	987
	PPR	639	272	281	902
	NN	639	265	281	519
Corn	SMLR	1574	887	936	4160
	PPR	1574	826	797	2583
	NN	1574	809	790	1949

[a] Mean SEP results from individual site–year analysis.
[b] SEP calculated from CVMSE for multi–year data sets including six climatological variables.
[c] Mean SEP results from leave one site–year out (LOO) analysis, with one climatological variable.

The results from the leave one site–year out (LOO) study were less promising. As expected, extreme overfitting was evident when all six climatological variables (table 2) were included in the model. Therefore, for the LOO analysis, data sets with only one climatological variable, total rainfall during the plant reproductive phase (days 51–110 after planting), were considered. While this adjustment improved the results slightly, overfitting was still obvious and extensive (table 5). Observations from the NN analysis with each site–year held out as a validation set were clustered, with little relationship apparent either within or between the clusters (fig. 6, right graph). In other words, when the single climatological observation representing one site–year was removed from the training data, the predictive ability of the method on that particular site–year was dramatically reduced. While the cross–validation experiment indicated that the nonlinear methods had enough complexity to successfully use the additional information contained in the climatological variables, the leave one site–year out (LOO) study clearly demonstrated severe overfitting on even a single climatological variable. This suggested that a much larger set of climatologically unique site–years would be required for these models to be used in a predictive manner.

Summary and Conclusions

The goal of this study was to evaluate and compare the predictive accuracies of various function approximation techniques, including supervised feed–forward neural networks (NN), projection pursuit regression (PPR), and stepwise multiple linear regression (SMLR), in relating crop yields to topography and soil parameters. Yield estimation within individual site–years was carried out through the use of a 5–fold cross–validation technique. SMLR, PPR, and NN methods were each investigated on ten individual site–year data sets (objective 1) and on multiple site–year data sets including climatological variables (objective 2).

With respect to objective 1, NN methods produced the minimal SEP results of all the methods investigated in every site–year. The rprop NN technique was consistently superior to the other techniques, producing minimal SEP results in 6 out of the 10 site–years. Nonlinear techniques, both NN and PPR, showed only small gains over SMLR in site–years with small data sets and in site–years when water stress was minimal. A likely explanation was that the data in one or more predictor dimensions was sparse enough in these data sets that no advantage in terms of reduced SEP was realized by the introduction of nonlinearity. Additional evidence suggested that data set size was an important factor in the predictive accuracy of nonlinear methods, and that the results could have been improved with larger initial data sets.

With respect to objective 2, cross–validation experiments across multiple site–years showed both PPR and NN SEP results that were equal to or better than those achieved on the individual site–year experiments. Investigation and interpretation of these cross–validated models may help improve our understanding of the complex relationships between soil properties, topography, and crop yields. However, leave one site–year out (LOO) experiments showed clear signs of overfitting, even when climatological conditions were reduced to a single variable. A large number of additional site–years of data representing a range of climatological conditions would be required for these models to become useful as predictive tools.

References

Adams, M. L., S. E. Cook, P. A. Caccetta, and M. J. Pringle. 1999. Machine learning methods in site–specific management research: An Australian case study. In *Proc. 4th Int. Conf. on Precision Agriculture*, 1321–1333. P. C. Robert, R. H. Rust, and W. E. Larson, eds. Madison, Wisc.: ASA–CSSA–SSSA.

Birrell, S. J., K. A. Sudduth, and S. C. Borgelt. 1996. Crop yield mapping: Comparison of yield monitors and mapping techniques. *Computers and Electronics in Agric*. 14(2): 215–233.

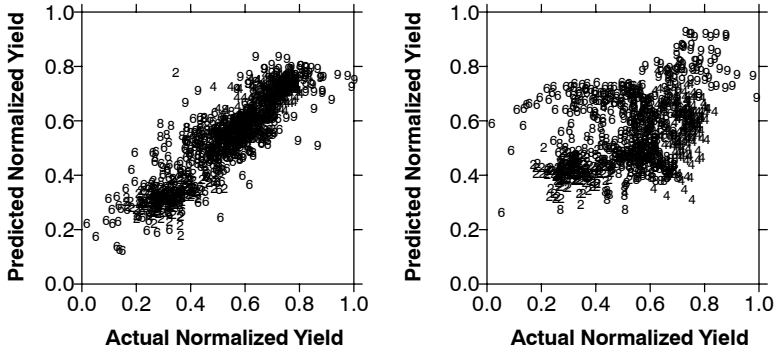

Figure 6. Neural network prediction results for the multiple site–year cross–validation study (left) versus the leave one site–year out (LOO) study (right) for all soybean site–years. Numbers denote the site–year to which the observation belongs.

Brown, J. R., and R. R. Rodiguez. 1983. *Soil Testing: A Guide for Conducting Soil Tests in Missouri*. Columbia, Mo.: University of Missouri, College of Agriculture, Department of Agronomy.

Burks, T. F., S. A. Shearer, R. S. Gates, and K. D. Donohue. 2000. Backpropagation neural network design and evaluation for classifying weed species using color image texture. *Trans. ASAE* 43(4): 1029–1037.

Cerrato, M. E., and A. M. Blackmer. 1990. Comparison of models for describing corn yield response to nitrogen fertilizer. *Agron. J*. 82(1): 138–143.

Drummond, S. T. 1998. Application of neural techniques for spatial yield estimation. MS thesis. Columbia, Mo.: University of Missouri.

Drummond, S. T., K. A. Sudduth, and S. J. Birrell. 1995. Analysis and correlation methods for spatial data. ASAE Paper No. 951335. St. Joseph, Mich.: ASAE.

Drummond, S. T., A. Joshi, and K. A. Sudduth. 1998. Application of neural networks: Precision farming. In *Proc. IEEE World Congress on Computational Intelligence* (CD–ROM). Piscataway, N.J.: IEEE.

Fahlman, S. E. 1988. An empirical study of learning speed in back–propagation networks. Technical Report. Pittsburgh, Pa.: Carnegie–Mellon University.

Fraisse, C. W., K. A. Sudduth, and N. R. Kitchen. 2001. Calibration of the CERES–MAIZE model for simulating site–specific crop development and yield on claypan soils. *Applied Eng. in Agric*. 17(4): 547–556.

Friedman, J. H., and W. Stuetzle. 1981. Projection pursuit regression. *J. American Statistical Assoc*. 76(376): 817–823.

Goutte, C. 1997. Note on free lunches and cross–validation. *Neural Computation* 9(6): 1211–1215.

Hornik, J., M. Stinchcombe, and H. White. 1989. Multilayer feedforward networks are universal approximators. *Neural Net*. 2: 359–366.

Khakural, B. R., P. C. Robert, and D. R. Huggins. 1999. Variability of corn/soybean yield and soil/landscape properties across a southwestern Minnesota landscape. In *Proc. 4th Int. Conf. on Precision Agriculture*, 573–579. P. C. Robert, R. H. Rust, and W. E. Larson, eds. Madison, Wisc.: ASA–CSSA–SSSA.

Kitchen, N. R., K. A. Sudduth, and S. T. Drummond. 1999. Electrical conductivity as a crop productivity measure for claypan soils. *J. Prod. Agric*. 12(4): 607–617.

Kravchenko, A. N., and D. G. Bullock. 2000. Correlation of corn and soybean grain yield with topography and soil properties. *Agron. J*. 92(1): 75–83.

Lamb, J. A., R. H. Dowdy, J. L. Anderson, and G. W. Rehm. 1997. Spatial and temporal stability of corn grain yield. *J. Prod. Agric*. 10: 410–414.

Liu, J., C. E. Goering, and L. Tian. 2001. A neural network for setting target yields. *Trans. ASAE* 44(3): 705–713.

Mathews, R., and S. Blackmore. 1997. Using crop simulation models to determine optimum management practices in precision agriculture. In *Precision Agriculture '97*, 413–420. J. V. Stafford, ed. Oxford, U.K.: BIOS Scientific Publishers.

Mathsoft. 1997. *S–Plus 4 Guide to Statistics.* Seattle, Wash.: Data Analysis Products Division.

Moore, I. D., A. Lewis, and J. C. Gallant. 1993. Terrain attributes: Estimation methods and scale effects. In *Modeling Change in Environmental Systems*, 198–214. A. J. Jakeman et al., eds. New York, N.Y.: John Wiley and Sons.

Neter, J., W. Wasserman, and M. H. Kutner. 1985. Multicollinearity, influential observations, and other topics in regression analysis–II. Chapter 11 in *Applied Linear Statistical Models*, 377–416. 2nd ed. Homewood, Ill.: Richard D. Irwin.

Pierce, F. J., D. D. Warncke, and M. W. Everett. 1994. Yield and nutrient variability in glacial soils of Michigan. In *Proc. 2nd Int. Conf. on Site–Specific Management for Agricultural Systems*, 133–152. P. C. Robert, R. H. Rust, and W. E. Larson, eds. Madison, Wisc.: ASA–CSSA–SSSA.

Riedmiller, M., and H. Braun. 1993. A direct adaptive method for faster backpropagation learning: The RPROP algorithm. In *Proc. IEEE International Conference on Neural Networks*, 586–591. Piscataway, N.J.: IEEE.

Rumelhart, D. E., and J. L. McClelland. 1986. *Parallel Distributed Processing.* Vol. 1. Boston, Mass.: MIT Press.

SAS. 1989. *SAS/STAT User's Guide*. Version 6, 4th ed. Cary, N.C.: SAS Institute, Inc.

Shearer, S. A., J. A. Thomasson, T. G. Mueller, J. P. Fulton, S. F. Higgins, and S. Samson. 1999. Yield prediction using a neural network classifier trained using soil landscape features and soil fertility data. ASAE Paper No. 993042. St. Joseph, Mich.: ASAE.

SNNS Group. 1997. SNNS v4.1. Stuttgart, Germany: University of Stuttgart, Institute for Parallel and Distributed High–Performance Systems (IPVR).

Stone, M. 1973. Cross–validatory choice and assessment of statistical predictions. *J. Royal Stat. Soc. Ser.* B 36: 111–147.

Sudduth, K. A., S. T. Drummond, S. J. Birrell, and N. R. Kitchen,. 1996. Analysis of spatial factors influencing crop yield. In *Proc. 3rd Int. Conf. on Precision Agriculture*, 129–140. P. C. Robert, R. H. Rust, and W. E. Larson, eds. Madison, Wisc.: ASA–CSSA–SSSA.

Sudduth, K. A., S. T. Drummond, and N. R. Kitchen. 2001. Accuracy issues in electromagnetic induction sensing of soil electrical conductivity for precision agriculture. *Computers and Electronics in Agric*. 31(3): 239–264.

Varcoe, V. J. 1990. A note on the computer simulation of crop growth in agricultural land evaluation. *Soil Use and Management* 6(3): 157–160.

Wendroth, O., P. Jurschik, and D. R. Nielsen. 1999. Spatial crop yield prediction from soil and land surface state variables using an autoregressive state–space approach. In *Precision Agriculture '99*, 419–428. J. V. Stafford, ed. Sheffield, U.K.: Sheffield Academic Press.

Werbos, P. 1988. Backpropagation: Past and future. In *Proc. IEEE International Conference on Neural Networks*, 343–353. Piscataway, N.J.: IEEE.