A Data Partitioning Scheme for Spatial Regression*

Slobodan Vucetic[1], Tim Fiez[2] and Zoran Obradovic[1],
svucetic@eecs.wsu.edu, fiez@wsu.edu, zoran@eecs.wsu.edu

[1] School of Electrical Engineering and Computer Science
[2] Department of Crop and Soil Sciences
Washington State University, Pullman, WA 99164-2752, USA

Abstract

Precision agriculture data consisting of crop yield and topographic features are examined with the objective of explaining yield variability as a function of topographic attributes in order to extrapolate this knowledge to unseen agricultural sites. It is demonstrated that random data partitioning into training, validation and test subsets is not appropriate when dealing with agricultural problems characterized with strong spatial data correlation. A simple spatial data partitioning scheme that leads to significantly faster neural network training and slightly better generalization is proposed. Also, integration of predictors formed from spatially partitioned data led to improved generalization over a bagging integration procedure in experiments. The margin between the best spatial model and a trivial predictor for our precision agriculture problem was small indicating that topographic features alone could explain only a small amount of the yield variability.

Purpose

Data sets in typical machine learning problems [8] are selected as a random sample drawn from an underlying data generating process that is to be modeled. A typical first step in estimating data generating process by neural networks and similar methods is to partition available data randomly into training, validation and test subsets. The validation data are used to prevent over-training and the testing data are used to provide a fair assessment of a model's prediction ability. The application of neural networks and other prediction methods to spatial data sets may require different partitioning schemes than simple random selection, however. Spatial data are a collection of variables whose dependence is strongly tied to a spatial location where observations close to each other are more likely to be similar than observations widely separated in space. Spatial data exhibit spatial continuity that can be quantified by calculating correlation, covariance, or moment of inertia versus separation distance between points [7].

One source of spatial data is what is called precision agriculture where agricultural producers are collecting large amounts of spatial data using global positioning systems to georeference sensor readings and sampling locations. Based on the interpretation of spatial data sets that include features such as topography, soil type, soil fertility levels, remotely sensed crop and soil reflectance, and previous crop yields, management decisions can be varied within fields instead of keeping them constant across an entire field area. Neural networks offer the potential to develop site-specific regression functions from spatial agricultural data that, given the ability to predict yield response, would allow calculation of optimum levels of production inputs. In this case, the objective is to explain yield variability as a function of the site-specific driving variables in order to *extrapolate* this knowledge to different agricultural sites, or to the same sites but to different years. This objective differs from the majority of research encountered in geostatistics [4] or spatial econometrics [1] where the goal is spatial interpolation.

Important questions for achieving this objective are how to design an appropriate procedure for fitting spatial extrapolation models such as neural networks, and how to estimate their generalization properties. This paper compares a spatial data partitioning scheme based on spatial blocking of data to simple random partitioning for deriving

* Partial support by the INEEL University Research Consortium project No. C94-175936 to T. Fiez and Z. Obradovic is gratefully acknowledged.

training, validation and test subsets. In addition, a modification of the non-spatial bagging procedure is presented to integrate predictors fitted on spatially partitioned data to further improve model generalization capabilities.

Method

Spatial data properties must be considered when training predictors and testing their generalization capabilities. Explanatory variables, as well as dependent variable in spatial data sets are usually highly spatially correlated. As a consequence, applying least squares regression techniques (often used for neural network training) on such data is likely to produce errors that are also spatially correlated. A similar phenomenon occurs when forecasting time-series [5].

When randomly partitioning data into model-fitting and testing subsets, it is possible that one might obtain good prediction results on the test data because of the spatial proximity between test and model-fitting samples. In a sense, the training and testing data are the same. This would not be useful for estimating true generalization properties of a predictor. Therefore, for spatial regression experiments, the test subset should be spatially separated from the model-fitting data employed by the learning algorithm. In the data-partitioning phase, the area containing the data (a field in an agricultural example) should be split into two spatially disjoint sub-areas (sub-fields) used for model fitting and testing (east and west fields shown in Figure 1).

Figure 1. A 220 ha wheat field near Pullman, WA separated into east and west sub-fields used for fitting a model (model-fitting sub-field) and testing its accuracy (test sub field*)*

An important part of the neural network design process is deciding when to stop training to avoid overfitting. One popular approach is to use part of the model-fitting data as a training set for designing the model, and use the rest as validation data for stopping the training process. Training is halted when the mean squared error (MSE) for the validation data starts to increase. For spatially correlated training and validation sets, minimizing the error on the training subset would likely minimize the error on a randomly chosen validation subset, since each sample in the validation subset would have samples in the training subset as its spatial neighbors. Therefore, it could be expected that the training of a neural network with a randomly selected validation subset would continue to the point of gross overfitting resulting in increased training time and lower generalization accuracy.

To address this problem, we propose a procedure that increases the separation distance between the data points of the training and validation subsets. The model-fitting portion of the field is partitioned into squares of size M×M, and half of these squares are randomly assigned for use in training and the rest for validation. A possible partitioning of the east subfield in Figure 1 into squares of size 100 x 100 m is shown in Figure 2. One way to assign squares to the training and validation subsets is to use a regular checkerboard-like partitioning, assigning neighboring squares to different subsets. A checkerboard-like assignment has desirable packing properties maximizing the distance between the points in the two subsets for a given size of squares. However, in this study the squares were assigned randomly such that an ensemble of models could be constructed and integrated for possible accuracy improvements.

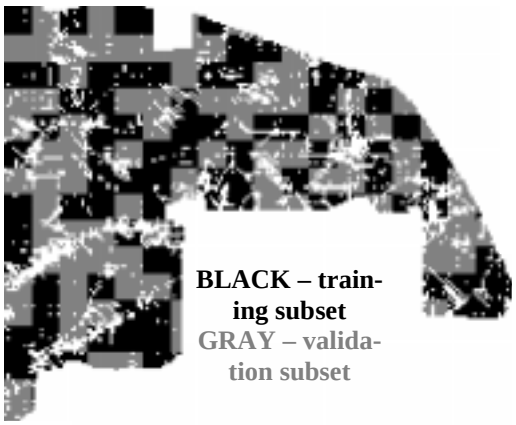

Figure 2. Spatial partitioning of east half of wheat field from Figure 1 into squares of size M=100m. Black and gray squares represent training and validation data while white represents missing data

The size M of each square should be selected such that the squares are sufficiently large to minimize the influence of spatial correlation between training and validation data, and still small enough to provide a training set representative of the variability of the model-fitting part of the field. One useful tool for describing the spatial variation of data

is a correlogram, which is a plot of the correlation coefficient as a function of the separation distance between data points. We have explored if an optimum size M could be selected by analyzing correlograms. The proposed method selects M to be within a range where correlograms of all topographic features start to approach zero. This minimizes the spatial dependence between training and validation samples, and hopefully allows the validation set to better track neural network generalization capabilities during the training process.

Neural networks fitted on the obtained spatial partitions are going to be unstable for two reasons. First, training of feedforward multilayer neural networks, as powerful nonlinear models, is very dependent on weight initialization. Second, it is influenced by the training set choice with small changes in the training set often causing larger changes in the predictor. Instability of neural network models can in principle be addressed through multiple model averaging. In one of the more successful techniques called bagging [3] each predictor is independently trained on N data points sampled with replacement from the N original data points of the training set and the ensemble prediction is obtained by averaging all individual predictors.

In this study *spatial bagging* using the proposed spatial partitioning scheme instead of sampling with replacement is considered as a possibly more appropriate choice for spatial data like ours. More precisely, we propose training a number of neural networks for different random assignments of squares into training and validation subsets followed by averaging the predictions of all such neural networks. This procedure allows combining desirable properties of spatial partitioning and ensemble predictors into a more powerful prediction method.

Experimental results

A precision agriculture database containing a grid of 8 topographic attributes and winter wheat yield from a 220 ha field located near Pullman, WA was used for experiments (Figure 1). In this case, we desired to predict wheat yield as a function of the terrain attributes. The eight terrain features were used as the input to the neural networks and wheat yield was the dependent variable at the output of the neural networks. The terrain attributes were derived from USGS 30 m DEM data using the software package TAPES-G [6]. The terrain attributes were: (f1) compound topographic index; (f2) aspect east-west (0 to 180°, 0 = east); (f3) aspect north-south (0 to 180°, 0 = north); (f4) distance to flow paths > 232 m; (f5) flow direction; (f6) slope; (f7) profile curvature; and (f8) average upslope slope. All 8 features except for flow direction were continuous and mostly non-normally distributed (Figure 3). The crop yield values were collected with a combine

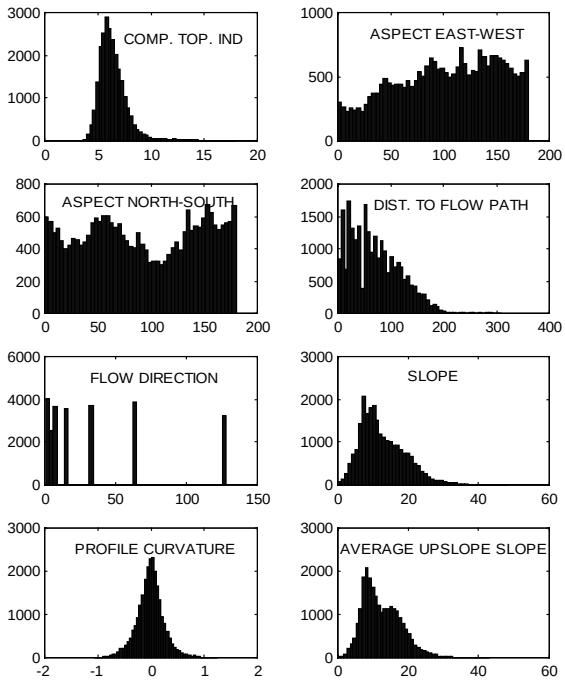

Figure 3. Histograms of 8 topographic features

mounted yield monitor and a global positioning system. All data were gridded to a 10 x 10 m grid. There were 24,592 patterns in the entire data set.

Correlation between the terrain attributes and wheat yield is shown in Table 1. Based on previous experience [9], the wheat yield data was considered to be very noisy. Correlograms for the topographic features are shown in Figure 4.

FEATURE	CORRELATION WITH YIELD
Comp. topog. index	-0.06
Aspect east-west	-0.09
Aspect north-south	0.22
Distance to flow path	0.16
Flow direction	-0.19
Slope	-0.09
Profile curvature	0.08
Average upslope slope	-0.16

Table 1. Correlation coefficients between yield and topographical features

The neural networks used in our experiments were feedforward neural networks with one hidden layer. Before training, both the input terrain features and wheat yield were normalized to mean zero and standard deviation one to allow for more efficient training procedures. The resilient backpropagation algorithm [10] was used for neural network training.

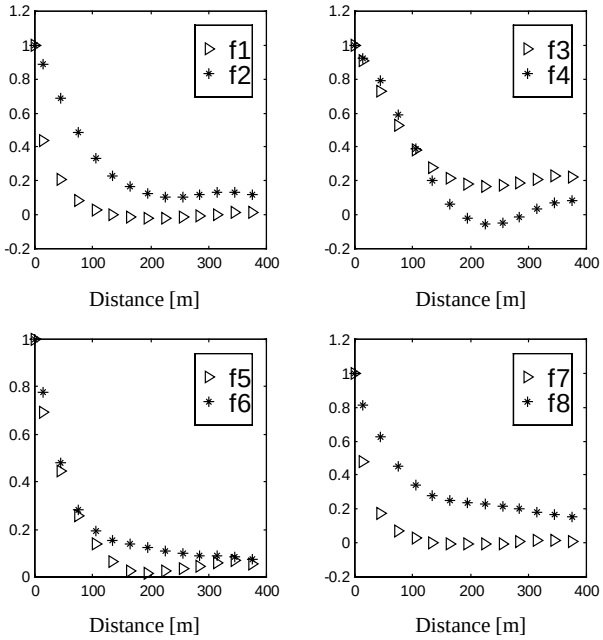

Figure 4. Correlograms for variables whose histograms are shown in Figure 3

The wheat field was separated into east and west halves (Figure 1) and a two-fold cross validation procedure was performed with the test set at the east and at the west half of the field to allow for generalization testing. Forty predictors were trained for each position of a test set, and an average MSE of the 80 fitted predictors is reported.

Selection of testing sets for spatial data

In the first experiment, models were fit to either the east or west sub-fields. Data from the chosen sub-field was randomly partitioned into two equal sets. One of these sets was used to train neural networks with 5 and 15 hidden nodes (100 training epochs), to fit a linear model, and to fit a trivial predictor that uses the mean yield from training data as a prediction for any test data point. These models were tested with a random test set (RTS) and a spatially disjoint test set (SDTS). The data partition of the modeled sub-field that was not used for training was used for the RTS; this corresponds to the classical random choice of test set. Data from the sub-field that was not modeled was used for the SDTS. Training, RTS and SDTS results averaged over 80 experiments are reported in Table 2.

Increasing the number of hidden nodes decreased the MSE on both the training subset and the RTS. Also, neural networks were superior to linear models and the trivial predictor on the training and RTS subsets. However, these results from the training and RTS data sets are overly optimistic about the ability to learn useful relationships that can be extrapolated to different data sites. Prediction results on the SDTS are much worse than on the RTS, and there was little improvement, if any at all, from increasing the complexity of predictors.

MODEL	MSE		
	Train	RTS	SDTS
Trivial	369	366	390
Linear	322	317	371
NN 5 hidden nodes	253	267	370
NN 15 hidden nodes	218	238	375

Table 2. MSE averaged over 80 experiments on training set, random and spatially disjoint test sets for several models.

Prediction error within a randomly chosen 15 ha region of a model-fitting field obtained by using a neural network with 5 hidden nodes is shown on Figure 5. As can be seen, errors are spatially correlated. This explains the 'success' in predicting the yield on RTS; decreasing MSE on training subset causes decreasing MSE on RTS because of spatial correlation in prediction error. This does not guarantee good generalization properties, as is obvious from the prediction results on SDTS. Therefore, using test data spatially separated from training data is necessary for a better assessment of generalization capabilities of fitted predictors.

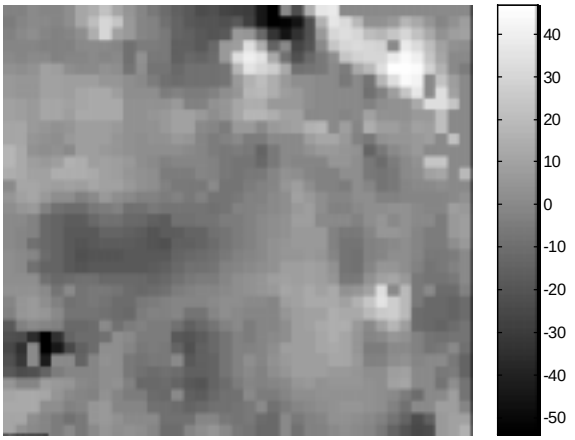

Figure 5. Training error for randomly chosen 15 ha region of the wheat field

Spatial partitioning of training and validation subsets

The proposed spatial partitioning of training and validation sets on neural network training was examined next. Correlogram analyses for all attributes showed a weakening of spatial dependencies between 40 and 200 meters as can be seen in Figure 4. Thus, we decided to experiment with squares sizes of M = 40, = 100, and = 200 m. For each of these values and for M = 10 m which corresponds to a typical random partitioning of data into training and vali-

dation sets, we used linear models and neural networks with 5 and 15 hidden nodes to perform experiments summarized in Table 3. Reported are average MSE together with the standard deviation of MSE for 80 trained neural networks when tested on spatially disjoint sets. Also shown is the average number of epochs before early stopping, together with its standard deviation over 80 experiments. For comparison, average MSE over 80 linear models, as well as the MSE of the trivial predictor are included in the table.

MODEL	MSE	EPOCHS
Trivial	390	-
Linear	370 ± 4	-
NN 5 hidden nodes		
M=10m	371±14	147±31
M=40m	365±17	43±19
M=100m	365±15	30±14
M=200m	364±21	25±10
NN 15 hidden nodes		
M=10m	377±13	190±38
M=40m	369±15	57±21
M=100m	361±16	34±13
M=200m	371±23	29±13

Table 3. Results in the table are means and standard deviations averaged over 80 experiments. MSE is reported on spatially disjoint test sets.

The MSE of the linear model was 370, and the MSE of the trivial predictor was 390. Average MSE for the neural networks ranged from 361 to 377. With our experimental agricultural data set, there was little decrease in MSE from that obtained with the trivial predictor as model complexity was increased. These results indicate that the topographic features in our data set explained only a small part of the yield variability. While topography can influence yield through associations with soil characteristics, water availability, and microclimate [11], yield can be strongly influenced by many factors such as soil fertility, weeds, and diseases that were not measured or included in our data set. Including such features would probably improve predictability. Second, it is yet to be determined how much knowledge can be extrapolated from one site to different sites, or from the same sites but to different years. Further research is needed to obtain an answer to this question.

While overall predictability appears limited by the features available to us, the experimental results suggest that significant improvements can be achieved by the proposed data partition scheme. As can be seen, the proposed spatial partitioning leads to significantly faster training and somewhat improved prediction as compared to using the typical random partitioning of training and validation data (Table 3, column for M = 10 m). Benefits of spatial data partitioning are more evident for more complex models than for simpler ones, since the danger of overfitting increases with the model complexity [2]. Best overall prediction results were obtained for M = 100 m, which is near the middle of the range suggested by the data correlograms.

Spatial bagging

In Table 4, we present comparative results between bagging and spatial bagging proposed in the Method section. Forty neural networks, the same ones used for Table 3, were used as an ensemble in spatial bagging. Forty neural networks were trained using the original bagging procedure with half of the model-fitting data randomly assigned to the validation subset. Experiments were performed for two positions of the test field (east or west) and reported MSE on spatially disjoint test sets are average values for both ensembles. As can be seen, original bagging improved prediction capabilities over single models from Table 3. Further, spatial bagging led to even better performance than original bagging. Thus, it seems that the proposed spatial partitioning incorporated into a spatial bagging procedure has advantage over bagging. Finally, both bagging and spatial bagging lead to better prediction for more complex neural networks with 15 hidden nodes, as compared to simpler ones with 5 hidden nodes.

METHOD	MSE	
	NN 5 hidden nodes	NN 15 hidden nodes
Bagging	357	351
Spatial Bagging		
M=40m	351	344
M=100m	348	342
M=200m	348	339

Table 4. Bagging vs. spatial bagging comparison for ensembles of 40 neural networks. MSE is reported on spatially disjoint test sets.

New aspect of work

It has been shown that, for spatial data, spatial partitioning of data into training and validation subsets can lead to substantial improvements in learning speed, and increased predictability, as compared to the traditional random partitioning. Experimental results indicate that correlograms can be used to determine the parameters of such spatial data partitioning. In addition, a spatial bagging procedure has been proposed, and experiments indicate that it can lead to improved generalization capabilities as compared to bagging.

Conclusions

The proposed method for dealing with spatial dependencies between neighboring data points appears to be superior to the usual random split approach to spatial regression. Spatial statistics were helpful in determining the parameters of such partitions. Integration methods, such as spatial bagging, are shown to further improve predictability of neural networks for the very noisy spatial agricultural domain.

For our experimental data set, the margin between the best models and a trivial predictor was fairly small, indicating that topographic features alone were able to explain only a small amount of the yield variability.

References

[1] Anselin, L., *Spatial Econometrics: Methods and Models*, Kluwer, Dordrecht, 1988.

[2] Bishop, C., *Neural Networks for Pattern Recognition*, Oxford University Press, Oxford, 1995.

[3] Breiman, L., "Bagging Predictors," *Machine Learning*, vol 26, no 2, pp 123-140, 1996.

[4] Cressie, N.A.C., *Statistics for Spatial Data*, John Wiley & Sons, Inc., New York, 1993.

[5] Davidson, R., MacKinnon, J., *Estimation and Inference in Econometrics*, Oxford University Press, New York, 1993.

[6] Gallant, J.C., Wilson, J.P., "TAPES-G: A Grid-Based Terrain Analysis Program for the Environmental Sciences," *Computers and Geosciences*, vol. 22, no. 7, pp 713-722, 1996.

[7] Isaaks, E.H., Srivastava, R.H., *An Introduction to Applied Geostatistics*, Oxford University Press, New York, 1989.

[8] Murphy, P.M., Aha, D.W., *UCI Repository of Machine Learning Databases*, Department of Information and Computer Science, University of California, Irvine, CA, 1999.

[9] Pierce, F.J., Anderson, N.W., Colvin, T.S., Schueller, J.K., Humburg, D.S., McLaughlin, N.B., "Yield Mapping," In The *State of Site-Specific Management for Agriculture* (ed.) F.J. Pierce and E.J. Sadler. ASA, CSSA, SSSA, Madison, WI, 1997.

[10] Riedmiller, M., Braun, T., "A Direst Adaptive Method for Faster Backpropagation Learning: The RPROP Algorithm," *Proc. IEEE Int'l. Conf. Neural Networks*, San Francisco, *1993*.

[11] Timlin, D.J., Pachepsky, Y., Snyder, V.A., Bryant, R.B., "Spatial and Temporal Variability of Corn Grain Yield on a Hillslope," *Soil Science Society of America Journal*, vol. 62, no. 3, pp 764-773, 1998.