UNDERSTANDING THE INFLUENCE OF NOISE, SAMPLING DENSITY AND DATA DISTRIBUTION ON SPATIAL PREDICTION QUALITY THROUGH THE USE OF SIMULATED DATA*

Dragoljub Pokrajac
School of Electrical Engineering and Computer Science, Washington State University,
Pullman WA 99164 USA
E-mail: dpokraja@eecs.wsu.edu

Zoran Obradovic
School of Electrical Engineering and Computer Science, Washington State University,
Pullman WA 99164 USA
E-mail: zoran@eecs.wsu.edu

Tim Fiez
Department of Crop and Soil Sciences, Washington State University,
Pullman WA 99164 USA
E-mail: tfiez@wsu.edu

KEYWORDS
Decision-support systems, AI-supported simulation, Regression analysis, Least-squares methods, Agriculture

ABSTRACT

The influence of data parameters (sensor error, unexplained variance, sampling density and data distribution) on spatial data prediction quality is considered through the use of a spatial data simulator. Performance of linear and non-linear regression models (feedforward neural networks) is compared on simulated agricultural data, but the results can be generalized to geological, oceanographic and other spatial domains. For a highly non-linear response variable, non-linear models are shown to perform better regardless of unexplained variance and sensor error, but linear models outperform non-linear models when the sampling density of spatial data is not sufficient to produce accurate interpolated values. In the presence of non-homogenous data distributions, a significant prediction quality improvement can be achieved by using specialized local models assuming that distributions are properly discovered.

INTRODUCTION

Precision agriculture combines agronomy, sensors, and geospatial technologies to vary crop production practices within fields instead of treating them as homogeneous units, in order to increase agricultural profitability and environmental stewardship. One approach to precision agriculture is yield prediction based on relevant features (e.g. land topography, soil test analyses, in order to optimize field treatments (e.g. fertilizer, pesticide and irrigation rates). The development of reliable prediction methods suitable for spatial regression is crucial for such an approach. However, the data typically available for spatial regression exhibit many properties (e.g. variable sampling density and data quality, field heterogeneity, missing features and unexplained variance) that make this a complex process.

Therefore, the determination of the influence of data quality and the choice of prediction method on spatial regression are important goals among precision agriculture researchers. However, a limited amount of appropriate data has prevented systematic experimentation towards this goal. The reasons for this are twofold. First, it is impossible to vary field and sensor parameters (statistical properties, unexplained variance and sensor accuracy). Second, it is very expensive to sample a large number of features at a high enough resolution so that one can determine the influence of sampling density on data quality.

To overcome these problems, we have developed a spatial data simulator (Pokrajac et al. 2000) that can provide large quantities of data with controlled statistical properties, yield influence and noise levels so the effects of many scenarios on crop yield prediction and production input optimization can be characterized. Using data from the simulator, exploratory work on a limited number of data sets has shown that the choice of type and number of simulation models, as well as feature selection has a significant impact on prediction results. Also, follow-up analysis showed that the introduction of spatial heterogeneity decreased the ability of regression models to successfully generalize on unknown test data.

In this paper, we perform systematic experiments on simulated data to determine the influence of various parameters on yield prediction accuracy. These parameters include: regression model type and methods, the amount of unexplained variance and error in yield and feature measurement, sampling density and procedures and field heterogeneity.

In addition, through the evaluation process of different types of yield simulation models, we tried to determine if the yield simulation process in the data simulator biased our analyses. Note that the results obtained here are applicable to other areas where spatial data are considered (e.g. geology, oceanography, and forestry).

METHODOLOGY

Experiment Outline

Agricultural fields were simulated using method and software described in (Pokrajac et al. 2000). Generated fields were divided into two spatially disjoint equal size subfields. Each subfield in turn was used for model training using the other subfield for model testing and the results were averaged. Linear regression was performed using standard statistical approaches (Devore 1995). Non-linear modeling was performed using back-propagation neural networks (NN) with 1 hidden layer having 4 neurons. Experiments were repeated 10 times each. Prediction accuracy was measured using R^2 values. (R^2 is a measure of the explained variability of the response variable. In the case of useful prediction models it ranges from 0 to 1 where 0 results from using a trivial mean predictor and 1 represents the ideal case of no prediction error). A one-sided t-test was used to compare linear and non-linear model results (Devore 1995).

Description of Generated Data Sets

To explore the influence of regression methods, sampling procedures, unexplained variance, and field heterogeneity, four simulated data sets-fields with different statistical properties were generated. Each field consisted of the spatial coordinates (latitude and longitude), features simulating nitrogen (N), phosphorus (P), potassium (K), terrain profile curvature and slope and simulated crop yield. Features were generated as samples on a 10m*10m grid from an approximately normal distribution.

Simulated Fields 1a and 1b were 800*1600m in size with uncorrelated features whose influence on simulated yield was modeled using both linear (Field 1a) and exponential plateau models (Field 1b). For the linear plateau models, the parameters for N, P and K were based on fertilizer recommendation guides while parameters for the rest of the features were based on expert knowledge. For the exponential plateau models, parameters were derived from the linear plateau parameters such that the exponential curves reached 95% of their maximal values at the linear plateau thresholds. Means and variances of the generated features were chosen so the resulting yield was approximately normally distributed and that 50% of all samples had nitrogen values above the plateau threshold. Spatial statistical parameters and the correlation matrix for the simulated features were estimated from Idaho and Washington potato and wheat fields. The variance of the simulated yield (standard deviation divided by sample mean) was similar to that of real-world data.

Simulated Field 2 was designed to examine the influence of interpolation error. The simulation parameters and the size were the same as for Field 1a except variograms with zero nuggets were used and, since the primary intention was to obtain data with controllable spatial behavior, features were not de-correlated.

Simulated Field 3 of size 1600*1600m was generated to explore the influence of multiple models that were associated with distinct feature values. The features of Field 3 were not de-correlated and had the same statistics as for Field 2.Five feature clusters were identified based on slope and profile curvature, and a separate linear plateau simulation model was assigned to each cluster, using overlapping

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

cluster model assignment (Pokrajac et al. 2000). The yield generated for each data point was a weighted combination of the yield generated by the specific models (weights for a data point were proportional to the number of neighboring points that belonged to the corresponding clusters). Since the intention was to examine the influence of highly non-homogeneous distributions, model parameters were chosen such that in each cluster the influence of particular features on yield was different.

RESULTS

Our results focus on investigating the combined influence of feature and yield sensor noise, data sampling density, and data distribution heterogeneity on the prediction of crop yield.

The Influence of Unexplained Variance and Sensor Noise

The influence of sensor and measurements errors are inevitable and usually in practice it is not possible to separate these forms of variability from the actual variability of the response variable. Field 1a and 1b data were used to determine how linear and non-linear models responded to these influences. In addition, the effects of using particular neural network (NN) learning algorithms (Polak-Ribiere conjugate-gradient, Quasi-Newton, resilient backpropagation and Levenberg-Marquardt (LM), (Demuth and Beale 1998)) were examined. Since the LM algorithm consistently performed the best, only results obtained with this algorithm are reported.

Prediction accuracy was measured in response to controlled additions of white noise to the simulated yield (Table 1). For all levels of unexplained variance, the NN outperformed the linear model with 99.99% significance. This was expected, since simulation model for yield was non-linear. Increasing the level of unexplained variance decreased the performance of both NN and linear models. As the unexplained variance was increased to 50%, the performance of the linear models decreased by about 30% while the performance of the NN decreased by about 45%. Observe that for the 30% and 50% unexplained variance levels, the R^2 values for the NN approached the theoretical maximums of 70% and 50%.

Although exponential plateau models are more natural, linear plateau models offer greater computational simplicity and might be preferred for simulation studies if their use over exponential models did not prejudice conclusions. The results shown in Table 1 indicate that there was no substantial difference between the two models. Therefore, in further experiments, we used the linear plateau model.

Unexplained variance (%)	Plateau model	Linear model R^2	Neural networks R^2	
			Mean	Std
None	Linear	0.58	0.91	0.01
	Exponential	0.66	0.94	0.02
30 %	Linear	0.38	0.62	0.01
	Exponential	0.45	0.65	0.01
50 %	Linear	0.29	0.45	0.01
	Exponential	0.32	0.46	0.01

Table 1: The influence of unexlained variance and simulation model type on yield prediction for fields 1a and 1b by linear and non-linear models

3-σ error of yield sensor (%)	Linear model R^2	Neural networks R^2	
		Mean	Std
5	0.58	0.91	0.01
10	0.56	0.89	0.01
15	0.55	0.88	<0.01
15 + 30% unexplained variance	0.38	0.62	<0.01

Table 2: The influence of yield sensor error on yield prediction for Field 1a by linear and non-linear models.

The influence of yield sensor error was modeled as multiplicative noise with a unit mean and a 3-σ interval equal to the specified sensor error. The effect of increasing yield sensor error on prediction accuracy for Field 1a is shown in Table 2. Comparison of these results to the results shown in Table1 indicates that the typical values of sensor error have little effect on yield prediction. There was no significant drop in accuracy due to 5% sensor error. The same was true when a 15% sensor error occurred with a 30% level of unexplained variance. This suggests that reasonable levels of sensor accuracy have little impact on overall prediction accuracy when the percentage of unexplainable variance is small. Again, neural networks outperformed the linear model with 99.99% significance.

Different values of feature sensor error (5, 10, 15, and 20%) were introduced into the Field 1a data. The effects on yield prediction were tested with linear and NN regression models as in the previous experiments. Results are shown in Table 3 where the average relative loss of explained variance (the relative decrement of explained variance due to error, $\mathbf{R^2_{loss}=(R^2_{without\ error}- R^2_{with\ error})/ R^2_{without\ error}}$) is given versus % of feature sensor error and % of unexplainable variance.

The presence of 20% feature sensor error caused drops in explained variance of 29% and 18% for the NN and linear model, respectively. This suggests that linear models are more resistant to data acquisition errors. Similar results were obtained when feature sensor error was considered in conjunction with added unexplained yield variance (experiment was performed for low and high error levels). However, the results of t-tests verify that in the cases examined in Table 3 the NN still outperformed the linear model with high (99.9%) significance.

Feature sensor error (%)	Unexplainable variance (%)	Average R^2_{loss} (%)	
		NN	Linear model
5	0	3	3
10	0	15	10
15	0	23	16
20	0	29	18
5	20	8	5
20	20	29	19

Table 3: Average relative loss of explained variance on Field 1a test data using neural networks and linear regression models, when feature sensor error is varied

Sampling resolution	Linear model R^2	Neural networks R^2	
		Mean	Std
10m*10m (no interpolation)	0.52	0.92	<0.01
50m*50m	0.33	0.26	0.05
100m*100m	0.29	0.07	0.11
200m*200m	0.13	-0.07	0.09

Table 4. The influence of sampling density on prediction accuracy for Field 2

The Influence of Sampling Density

The sampling density and interpolation experiments were performed on Field 2 data, with more closely controlled the feature spatial statistics. Data from the artificial field were sampled at different resolutions. These samples were then used to interpolate data back to the original 10m*10m grid and the interpolated values were used for yield prediction. This approach is a standard procedure for combining low-resolution sampled data with higher resolution data (simulated yield in our case) (Cressie 1993). Since we operated with simulated data, we were able to show that as sampling resolution decreased, interpolation error increased.

Using these interpolated data, the performance of non-linear and linear prediction models was again compared (Table 4). While the NN model outperformed the linear model when no interpolation was performed, the introduction of interpolation error dramatically decreased NN prediction accuracy. Using data interpolated from samples collected on a 50m*50m grid, the linear model outperformed the NN although the NN performed well on training data. When data were sampled on a 100*100m grid and then interpolated, NN regression was practically useless, with only 7% explained variance. For the 200*200m sampling grid, the linear regression model was still able to explain 13% of the variance while the neural network

predictions were worse than just using the mean yield value of the training set as the prediction (negative R^2). Therefore, linear models are recommended if sparse data must be interpolated prior to regression.

To try to determine optimal sampling frequency, we constructed a series of linear predictors using a method that takes into account the spatial correlation of the data (Judge at al. 1988). We varied the sampling density and used the original samples instead of working with interpolated values to compute regression coefficient t-statistic confidence (Devore 1995) versus sampling distance. The maximal confidence in the obtained coefficients showed a large drop around d=50m suggesting that the optimal sampling density might be around 50m. This coincides with our results from the interpolated data, where the linear model started to outperform the non-linear model at the 50m sampling distance. Therefore, it appears possible to determine whether the sampling density for features is optimal or should be increased by comparing linear and non-linear model behavior and by computing linear model coefficient significance.

The Influence of Interpolation

The expected interpolation error of spatially interpolated data increases as the distance to known data points, the sample points, becomes greater. Thus, by taking only a subset of the interpolated data that are near to the sample points one can control the level of interpolation error in the data used for learning. To test the benefits of such sub-sampling, interpolated data within a given radius of the sampling points in Field 2 were used to develop yield prediction models (Fig. 1). As this sub-sampling radius was increased, more points were available for training but the average accuracy of the training data decreased. As before, all points in a spatially disjoint test part of the field were used to test the models.

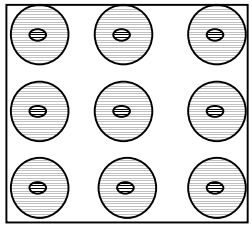

Fig.1. Experiments with the influence of interpolation on prediction accuracy

Prediction error was measured as the circle radius was increased, for both linear and NN regressors and for grid spacing of 50, 100 and 200m. For each radius, experiments were repeated 10 times and 2-cross validation was performed. The impact of changing subsampling radius was analyzed by ANOVA methods (Devore 1995). For the NN, regardless of sampling density, training error decreased as the radius was increased. However, there was no significant improvement in prediction accuracy except for the case of the 50*50m sampling grid where (with 99.99% significance) the benefit from increasing the subsampling radius and thus using more interpolated data points was evident.

Similar experiments with linear models did show some improvement from using a subset of the interpolated points when data used for interpolation were sampled at 50 and 100m grid spacings. Thus, the selection of an appropriate interpolation range might benefit linear predictors more than nonlinear ones.

The Influence of Yield Heterogeneity

One approach to yield prediction for heterogeneous fields is to discover regions of homogeneity and to build models specific to these regions. To test best case results, models were trained on points from simple feature-space regions of the training part of Field 3, and tested on corresponding points in the test subfield. When trained on these regions, neural networks achieved average accuracy of R^2=0.90 which was an improvement compared to the accuracy of a single network trained on all training data (R^2=0.67). In both cases, neural networks outperformed linear models with 99.9% significance (accuracy of R^2=0.85 and R^2=0.37, respectively).

The results imply that proper detection of homogeneous regions for model development (training) and for model application (testing), using regression models that recognize distinct data distributions (Lazarevic et al. 1999), can result in better accuracy than the use of single global models. Note that due to heterogeneity average explained variance by the global model was 20 to 30% less then that observed in the Field 1a and Field 1b experiments (Table 1).

It is important to emphasize that if regions of homogeneity are wrongly detected when applying region specific models, results can be worse than those from using a single global model. To demonstrate this, we trained local models on each of the distinct data distributions in the training subset of Field 3 and applied these models to the entire test set instead of applying them to points from the data distribution matching their training data. The maximum global accuracy of these local models was only R^2=0.16 and in most cases, the global accuracy of the local models was worse than what could be obtained by simply using the mean value of all the training data as the global prediction.

DISCUSSION

In this paper, using simulated data, we considered the appropriateness of linear and non-linear models (feed-forward neural networks) and the effect of key data parameters on crop yield prediction for precision agriculture management. For highly non-linear relationships between features and simulated yield, non-linear models outperformed linear models when data sampling was appropriate.

Additions of unexplained variance resulted in significant drops in prediction accuracy. Additions of error to feature values also caused drops in prediction accuracy although linear models seem to be more resistant to this kind of error.

The common practice of using interpolation procedures to estimate observations at locations that are not physically sampled drastically affected non-linear model performance. When the sampling density of the features was low and there were large interpolation errors, linear models outperformed non-linear models. Our results suggest that there is no significant benefit from using interpolated data over the original feature values obtained by sampling, when non-linear models are applied. Even though the interpolation process results in larger data sets and allows the use of all of the response variable data (collected at a high resolution), there was no benefit in terms of prediction accuracy.

On the other hand, initial results for linear models show that they can benefit from using interpolated data instead of just using sampled data. These results coincide with those obtained on real-life agricultural data and suggest that users must assess data sampling density and the number of sampled points in selecting the best modeling approach. Comparisons of linear and non-linear model performance along with the statistical confidence of linear regression coefficients appear to be useful for testing sampling grid optimality.

The detection of yield distribution heterogeneity is another important aspect in obtaining maximum prediction accuracy. Using a global model when there are multiple yield generation functions leads to markedly lower performance. By using appropriately specialized local models, one can achieve a significant improvement in prediction quality if accurate methods for identifying the data distribution and selecting the appropriate model are applied. Otherwise, performance is poorer than when global models are used.

REFERENCES

Cressie, N. 1993. *Statistics for spatial data. Rev. ed*. John Wiley & Sons, inc., New York Chichester Toronto Brisbane Singapore.

Demuth, H. and M. Beale. 1998. *Neural network toolbox for use with MATLAB, Users Guide, Version 3*. The MathWorks, Inc.

Devore, J.L. 1995. *Probability and statistics for engineering and the sciences. 4th ed*. Int'l Thomson Publishing Company, Belmont.

Judge, G.G., R.C. Hill, W. Griffiths, H. Lutkepohl and T.C. Lee. 1988. *Introduction to the theory and practice of econometrics. 2nd edn*. John Wiley & Sons, inc., New York.

Lazarevic, A.; X. Xu; T. Fiez; and Z. Obradovic. 1999. "Clustering-regression-ordering steps for knowledge discovery in spatial databases." In *Proc. IEEE/INNS Int'l Joint Conf. on Neural Networks*, ISBN 0-7803-5532-6, Washington D.C., No. 346, Session 10.9.

Pokrajac, D.; T. Fiez; and Z. Obradovic. 2000. "A Tool for Controlled Knowledge Discovery in Spatial Domains." In *Proc. of 14. ESM Conference* (in press).