SPATIAL-TEMPORAL TECHNIQUES FOR PREDICTION AND COMPRESSION OF SOIL FERTILITY DATA

D. Pokrajac

Center for Information Science and Technology
Temple University
Philadelphia, Pennsylvania

A. Lazarevic

Computer Science Department
University of Minnesota
Minneapolis, Minnesota

R. L. Hoskinson

Idaho National Engineering and Environmental Laboratory
Department of Energy
Idaho Falls, Idaho

Z. Obradovic

Center for Information Science and Technology
Temple University
Philadelphia, Pennsylvania

ABSTRACT

In this paper, we propose several techniques for data reduction and spatial-temporal prediction in precision agriculture databases. The proposed methods are based on various statistical and machine learning techniques including sensitivity based analysis, spatial-temporal autoregression, multiple time series and response modeling with spatially-correlated residuals. The considered techniques are implemented in described a prototype software and applied for analysis and compression of multi-temporal precision agriculture data. The spatial-temporal prediction on real-life soil fertility data using the proposed spatial-temporal autoregression method is discussed in an accompanying paper.

Keywords: data compression, machine learning, geostatistics, temporal, spatial

INTRODUCTION

Advances in spatial databases have allowed for the collection of huge amount of data in various GIS applications ranging from remote sensing and satellite telemetry systems, to computer cartography and environmental planning. In addition, majority of such collected data may also change through time, so not only the spatial but also temporal dimension should be considered in these databases. An area of data mining aimed to extraction of knowledge and spatial relationships not explicitly stored in spatial-temporal databases is called spatial-temporal knowledge discovery.

In many real life applications, both the number and the size of spatial-temporal databases are rapidly growing, and therefore the need for data reduction of very large spatial databases is of fundamental importance for efficient spatial-temporal data analysis. The main purpose of this paper is to discuss various techniques that could reduce the size of spatial-temporal database without loosing much information. The considered techniques include spatial statistical analysis, spatial-temporal modeling and sensitivity analysis as well as identification of data subsets that can be reduced. After the brief overview of the considered methods, we proceed with the structure of the developed software for spatial-temporal data-reduction with particular emphasis on the user interface and its functionality.

METHODS

To perform data reduction on spatial data, we apply estimation of spatial data statistics at preprocessing stage by estimating spatial variograms. Variograms are standard descriptive statistics for spatial data (Cressie, 1993; Deutsch and Journel, 1998; Chilès and Delfiner, 1999; Olea, 1999). Basically, variograms depict spatial (or-spatial-temporal) dissimilarity of data samples on a given distance and can be applied for data interpolation by kriging techniques. In kriging, the data value is interpolated as a linear combination of the known data values (Cressie, 1993). In addition, data interpolation can be performed through the inverse distance technique, where the weights in the interpolation function are inversely proportional to the distance from the sample point to the points where the attribute value is known (Isaaks and Srivastava, 1990). Regardless of the applied interpolation technique, we propose to gradually vary the sampling density and to measure the interpolation error such that the maximal sampling distance that still could provide pre-specified interpolation accuracy is applied to reduce the size of original dataset by sampling.

When performing data reduction on spatial-temporal databases, we apply two different families of techniques. In the first family, our goal is to reduce storage requirements for the *response variable* (e.g. crop yield), while in the second group of techniques, we aim to reduce required storage space for the driving attributes (e.g. concentrations of nutrients, terrain attributes, etc.).

One of possibilities to reduce memory requirements necessary for storing redundant response variable information is to first apply response modeling based on response values in previous time instants and on current attribute values, and then to keep in storage only those values that cannot be predicted within specified tolerance. Here, we apply modeling with spatially correlated lagged residuals (Pokrajac and Obradovic, 2001). In this model, for each point of each temporal data layer, the response is predicted as a function of attributes and the prediction residuals in neighboring points at previous time instant on uniform grid.

A similar procedure is applied to reduce the need for storing attribute values. Using **S**patial-**T**emporal autoregressive modeling on **U**niform **G**rid (STUG) (Pokrajac et al., 2002), we model values of an attribute according to the values of the same attribute in historical data. Based on a recent history of each observed attribute, a spatial-temporal model for predicting future attribute values is constructed, stored and then evaluated using data from the next temporal layer. Coordinates of points where the predictive model performs within pre-specified accuracy boundaries are stored and the corresponding data values are removed from the main data set. In the data set reconstruction phase, at each spatial-temporal location an attribute value that is not stored at the data set is reconstructed using appropriate historical data, previous reconstructions and their corresponding prediction model.

If an attribute value can be properly predicted using historical values of *other* attributes, there is again no need to store the value in the database. To determine locations where this prediction and corresponding data compression is possible, we apply multiple time series approach (Lütkepohl, 1991).

In data reduction for knowledge discovery, one of the goals is to maintain fidelity of the attributes sufficiently high to preserve attribute-response relationship (Vucetic and Obradovic, 2000). When applying this technique, we build neural network prediction models (Haykin, 1999) and estimate the influence of attributes quantization on models prediction accuracy. Finally, we compress their driving attributes based on the results of a sensitivity analysis of the applied neural network model.

SOFTWARE ORGANIZATION

We developed software package for spatial-temporal data reduction based on VisualBasic® and Matlab® functionality. The software is organized into following software modules (tools) as shown in Figure 1:

- Data Manipulation and Loading
- Spatial Data Reduction
- Reduction of response data through spatial-temporal modeling
- Data Compression using model sensitivity analysis
- Partial Spatial-Temporal Data Reduction

All implemented functions (operations) are organized into standard menu items corresponding to existing software modules. When some of the menu items are opened, available operations within that corresponding module can be selected from a drop-down menu.

For every function to be executed within the developed software system, the appropriate dialog box is opened, and the user may choose various parameters in order to optimize results of the corresponding function. In general when some default parameter values are offered the user does not need to specify any parameters. In contrast, when a parameter is not pre-specified, the user **must** assign its value explicitly.

In order to perform any implemented operation, the data have to be loaded first. If one wants to accomplish some operations on a different data set, there are two possibilities. The better option, which is always available, is to load the new data set. In this case, this data become active and all future operations are performed on this new data. The second option, not always available, is related to loading the data sets when some operation has already been started. In this case, the user can choose to perform the operation on already loaded data set or to load new data set. However, this new data set is active only for this operation, so when the user close the dialog box related to this operation the old data is active again.

One of common approaches in machine learning is to test model (e.g. some relationship or function describing the association between the driving variables and the response variable) by examining its generalization capabilities on new unseen data set called **test data set**. The data used for determining the relationships or models is called **train data**. For some operations user must specify both the train and test data.

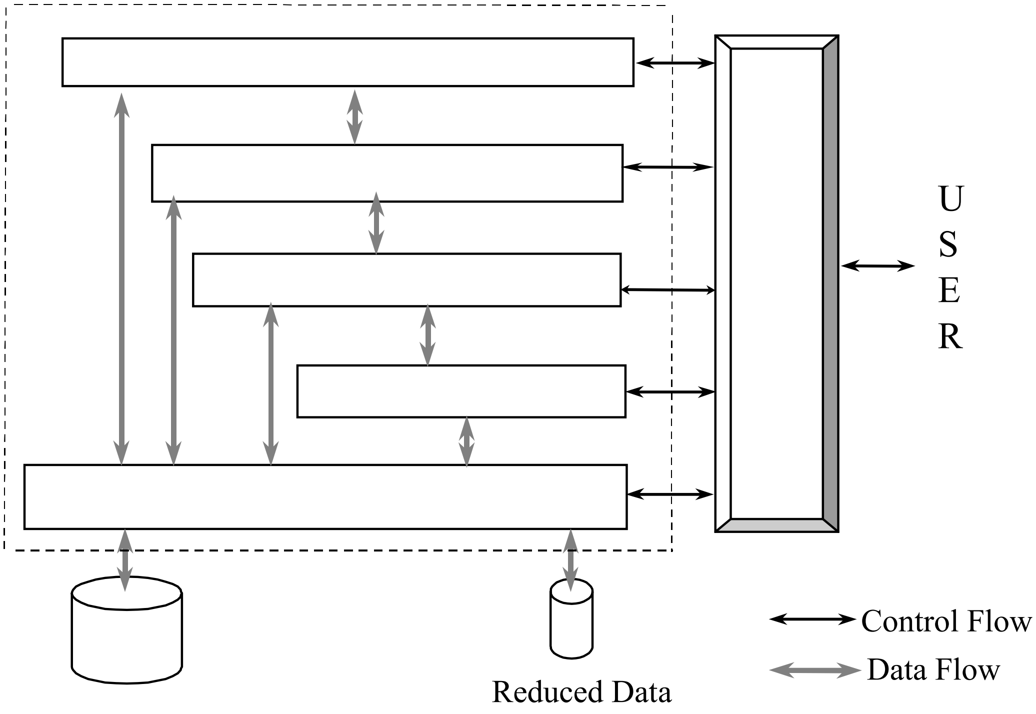


Figure 1. Data reduction tools organization

SPATIAL DATA REDUCTION

Spatial data reduction module provides several techniques for reduction of spatial layers. In spatial data, explanatory attribute values as well as the target attributes are strongly related to a spatial location where observations close to each other are more likely to be similar than observations widely separated in space. Using the tools of spatial statistics, ‘spatial continuity’ can be quantified by measuring the spatial covariance of samples as a function of their distance. Our preliminary results indicated that for the known spatial characteristics of data, one could determine the grid distance satisfying a predetermined accuracy loss in the spatial reconstruction of compressed data. In this module we have implemented the following operations for spatial data reduction:

- Spatial Statistics Estimation
- Sampling Grid Estimation
- Data Sampling
- Data Interpolation

Spatial Statistics Estimation

This option is employed to estimate the basic spatial characteristics of the data, producing spatial variograms. Variogram itself can be estimated using *all* samples from original dataset (which is prohibitive for large data sets due to high requirements for memory and computational time) or from a random subset of a data set with a pre-specified size. In order to estimate the variograms the user has to set the following parameters at the software (Deutsch and Journel, 1998):

- sampling flag determines whether the variogram is estimated on original data or on a random data subset
- minimal and maximal lag distance for which the variograms are to be computed
- number of equidistant lags on which the variogram is estimated
- minimal number of points in a bin necessary to estimate variograms for a particular lag
- percentage that specifies the effective bin size (ratio of the bin size and the lag difference)
- the estimation variogram will be estimated. In addition to original Matheron’s method (e.g. Cressie, 1993) we provide Cressie’s robust estimation (Hawkins and Cressie, 1984)
- the name of the file where the parameters of the estimated variogram will be saved.

For each attribute, variograms estimation is an iterative process during wherein the user may vary parameters of the theoretical variogram (range, nugget, sill and variogram type) that serves as an approximation to the estimated one. At each iteration, the user observes the goodness of variograms fit, based on MSE (Mean Squared Error) criterion as well as Cressie weighted MSE (Cressie, 1985)(Figure 2).

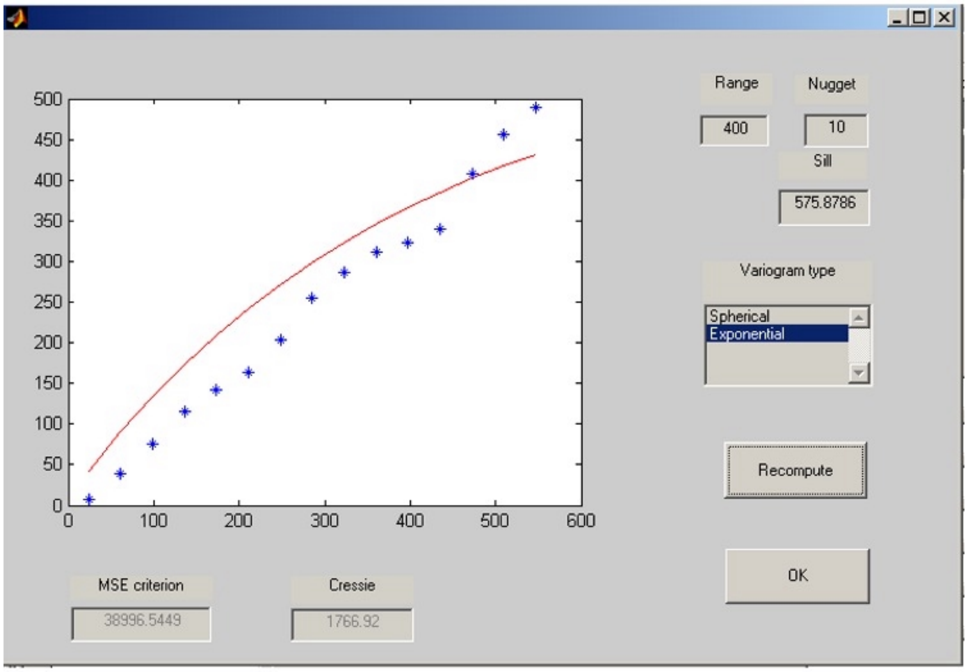
Sampling Grid Estimation

In this operation the maximal sampling distance that satisfies the pre-specified loss is determined. The user has to select the following parameters:

- Minimal interpolation error for each feature separately. Here, the interpolation error is squared difference between predicted and true sample value normalized with the attribute variance. The interpolation error for all features is initially set to 0.5.

- Minimal and maximal sampling distance
- The decremental value that specifies how fast we decrease the sampling distance from maximum to minimum.
- The method for data interpolation, which can be Kriging or inverse distance method.

After finishing this operation, software provides all the errors and obtained sampling distances in a table (Figure 3). It is important to note, that if the kriging method is used for interpolation, variograms have to be computed beforehand.



Range 400

Nugget 10

Sill 575.8786

Variogram type

Spherical
Exponential

MSE criterion 38996.5449

Cressie 1766.92

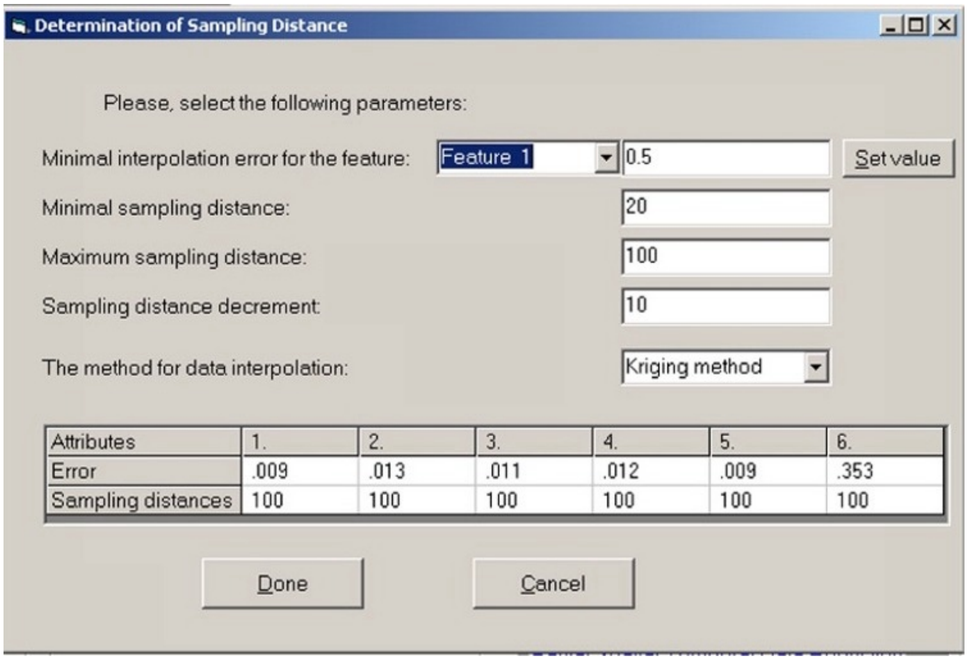
Figure 2. Variogram estimation

Data Sampling

The purpose of this function is to interpolate the dense raw data to a proper sparse regular grid and obtain a reduced data layer. The required step for this operation is to first determine the sampling density in the Sampling Grid Estimation module. The only parameter that the user has to specify is the name of the file where the sampled data will be saved.

Data Interpolation

This operation is used to reconstruct the original layer from a reduced data layer using an appropriate spatial interpolation method. In order to perform this operation, the reduced data file created in the Data Sampling module has to exist on the disk and to be loaded into the software. The user specifies only the file name where the interpolated data will be saved.



Please, select the following parameters:

Minimal interpolation error for the feature: Feature 1 0.5

Minimal sampling distance: 20

Maximum sampling distance: 100

Sampling distance decrement: 10

The method for data interpolation: Kriging method

Attributes	1.	2.	3.	4.	5.	6.
Error	.009	.013	.011	.012	.009	.353
Sampling distances	100	100	100	100	100	100

Figure 3. Determination of sampling distance in spatial data reduction

REDUCTION OF RESPONSE DATA

In this module we have implemented the following operations for spatial-temporal data reduction based on modeling with spatially correlated lagged residuals:

- Reduction of Response Variable
- Reconstruction of Response Variable

Reduction of Response Variable

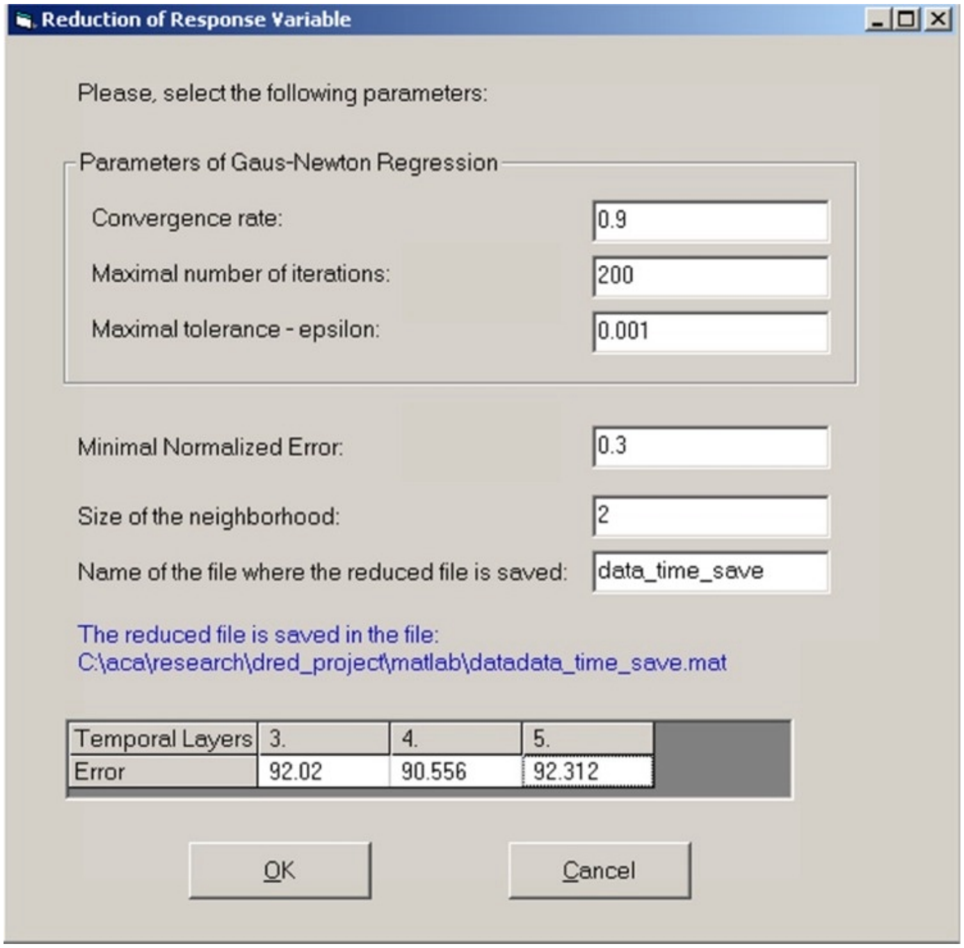
In this module, the user has to specify the following parameters:

- parameters of a regression model (convergence rate, maximal number of iterations and the maximal tolerance (ε - epsilon))
- minimal normalized error (squared difference between predicted and true sample response value normalized with the response variance).
- size of the neighborhood that has influence on the prediction model
- name of the file where the reduced file is saved

and as a result the information about mean-squared error on each temporal layer (first two layers are used for model learning and cannot be compressed) is displayed (Figure 4).

Reconstruction of Response Variable

In order to reconstruct the values of the response variable that have not been stored, the user first has to load the reduced data set obtained by previous operation (Reduction of Response Variable). In addition to specifying the name of the saved reduced data set, the user has also to specify the name of the file where the reconstructed data will be saved. Finally, the user is prompted when the reconstruction phase is finished (Figure 5).



Reduction of Response Variable

Please, select the following parameters:

Parameters of Gaus-Newton Regression

Convergence rate: 0.9

Maximal number of iterations: 200

Maximal tolerance - epsilon: 0.001

Minimal Normalized Error: 0.3

Size of the neighborhood: 2

Name of the file where the reduced file is saved: data_time_save

The reduced file is saved in the file:
C:\aca\research\dred_project\matlab\datadata_time_save.mat

Temporal Layers	3.	4.	5.
Error	92.02	90.556	92.312

Figure 4. Spatial-temporal response reduction

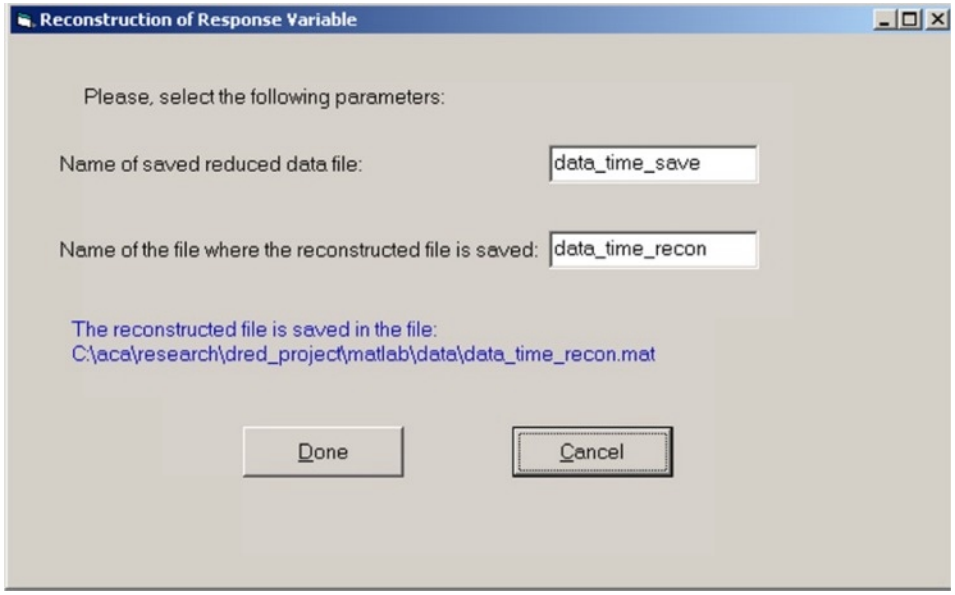
PARTIAL SPATIAL TEMPORAL DATA REDUCTION

The objective of this module is to determine if subsets of the spatial-temporal attributes can be determined (predicted) by other data, such that the predictable attributes or attribute values would not have to be collected, analyzed, compressed and stored.

In this module we have implemented the following operations for partial spatial-temporal data reduction:

- Reduction and reconstruction of attributes using spatial-temporal autoregression

Reduction and data reconstruction using multiple time series.



Reconstruction of Response Variable

Please, select the following parameters:

Name of saved reduced data file: data_time_save

Name of the file where the reconstructed file is saved: data_time_recon

The reconstructed file is saved in the file:
C:\aca\research\dred_project\matlab\data\data_time_recon.mat

Figure 5. Screen for data reconstruction in spatial-temporal response reduction

Reduction of Attributes using Spatial-Temporal Autoregression

This operation is performed in order to reduce memory requirements for attribute data using spatial-temporal autoregressive models on uniform grid. Here, the attribute value at sampling location is predicted using the values of the same attribute at the same location and its neighborhood taken in previous time intervals.

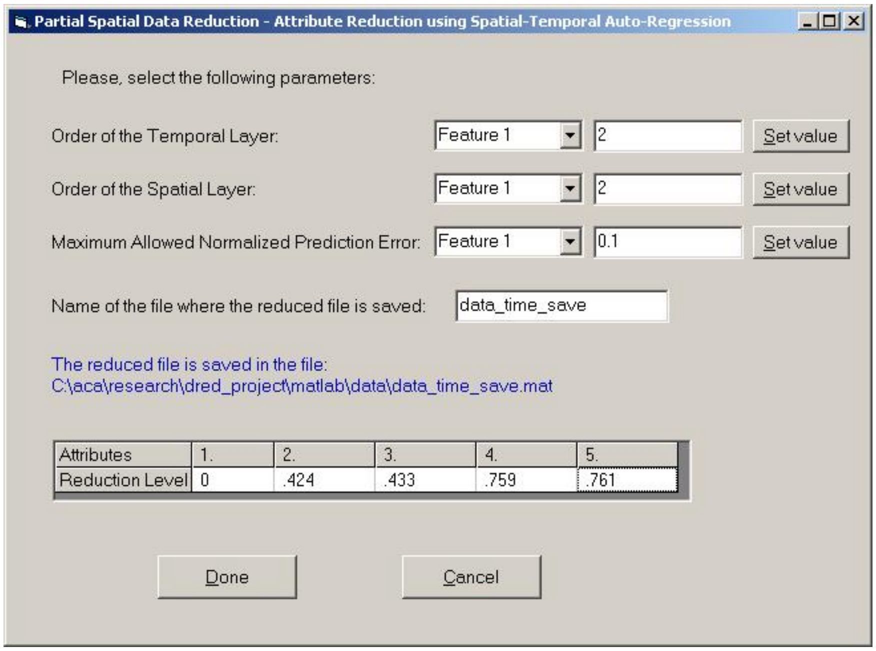
After user has specified the following parameters:

- the order of the temporal layer specifies the number of temporal layers the observed sample depends on
- the order of the spatial layer specifies spatial neighborhood size
- maximum allowed normalized prediction error; for each sample, we do not save the actual attribute value if squared difference between predicted and true value, normalized with the attribute variance, does not exceed this maximum
- the name of the file where the reduced file will be saved,

the software shows achieved compression levels for each attribute (Figure 6).

Reconstruction of Attributes Compressed using Spatial-Temporal Autoregression

In order to perform this operation the user first has to load the reduced data set obtained by previous operation. In addition to specifying the saved reduced data set, the user has also to specify the name of the file where the reconstructed data will be saved. Finally, the information that the reconstruction phase is completed and that the reconstructed file is saved under the specified name is given (Figure 7).



Partial Spatial Data Reduction - Attribute Reduction using Spatial-Temporal Auto-Regression

Please, select the following parameters:

Order of the Temporal Layer: Feature 1 2

Order of the Spatial Layer: Feature 1 2

Maximum Allowed Normalized Prediction Error: Feature 1 0.1

Name of the file where the reduced file is saved: data_time_save

The reduced file is saved in the file:
C:\aca\research\dred_project\matlab\data\data_time_save.mat

Attributes	1.	2.	3.	4.	5.
Reduction Level	0	.424	.433	.759	.761

Figure 6. Screen for attribute reduction using spatial-temporal auto-regression

Reduction of Attributes Using Multiple Time Series

This operation is performed in order to reduce memory requirements for attribute data using multiple time-series. Here, the attribute value at sampling location is predicted using the values of the same attribute as well as other attributes at the same location taken in specified past intervals.

The user has to specify the following parameters:

- the order of the temporal layer specifies the number of temporal layers the observed sample depends on
- maximum allowed normalized prediction error. For each sample, we do not save the actual attribute value if squared difference between predicted and true value, normalized with the attribute variance, does not exceed this maximum
- the name of the file where the reduced file will be saved

Reconstruction of Attributes Compressed Using Multiple Time Series

In order to perform this operation the user first has to load the reduced data set obtained by previous operation. In addition to specifying the saved reduced data set, the user has again to specify the name of the file where the reconstructed data will be saved. Finally, the information that the reconstruction phase is completed and that the reconstructed file is saved under the specified name.

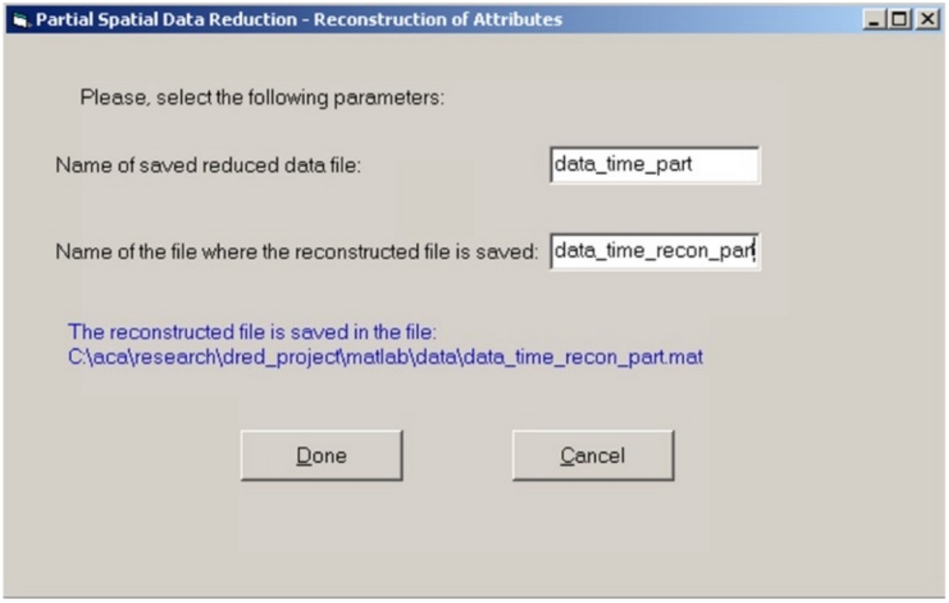
SENSITIVITY BASED ANALYSIS

This module contains operation which purpose is to perform data reduction based on non-uniform sensitivity based attribute quantization.

The user has to specify the following parameters:

- allowed percentage loss in prediction accuracy
- number of hidden neurons in a neural network model
- number of epochs for training a neural network model
- the name of the file where the reduced data set will be saved

and at the end, the result is achieved data compression and the achieved loss in prediction accuracy (in percents).



Partial Spatial Data Reduction - Reconstruction of Attributes

Please, select the following parameters:

Name of saved reduced data file: data_time_part

Name of the file where the reconstructed file is saved: data_time_recon_par

The reconstructed file is saved in the file:
C:\aca\research\dred_project\matlab\data\data_time_recon_part.mat

Figure 7. Reconstruction of attributes compressed using spatial-temporal autoregression

CONCLUSIONS

This paper proposes several techniques for data reduction and spatial-temporal prediction in precision agriculture databases. The techniques have been implemented in a prototype software, operational at Idaho National Engineering and Environmental Laboratory. Although the software has been preliminary tested on analyses of agricultural spatial-temporal data (Hoskinson et al., 2002) we emphasize the need to evaluate software performance on the broader spectrum of datasets, including data from various spatial-temporal domains. In addition, our future work will focus on providing automatic procedures for parameter setting of the implemented techniques as well as on exploring other, alternative data compression techniques and their applications in spatial-temporal domains.

ACKNOWLEDGEMENTS

Work supported through the INEEL Laboratory Directed Research & Development (LDRD) Program under DOE Idaho Operations Office Contract DE-AC07-99ID13727.

REFERENCES

Chilès, J., and P. Delfiner. 1999. Geostatistics-Modeling Spatial Uncertainty, John Wiley & Sons, New York.

Cressie, N. 1985. Fitting variogram models by weighted least square. *Math. Geology*, 17 (5), pp. 563-586.

Davidson, R., and J. G. MacKinnon. 1993. Estimation and Inference in Econometrics, Oxford Univ. Press, New York.

Deutsch, C.V., and A. G. Journel. 1998. GSLIB: Geostatistical Software Library and Users Guide, 2nd edn, Oxford Univ. Press, New York.

Hawkins, D.M., and N. Cressie. 1984. Robust kriging-a proposal. *J. International Assoc. Math. Geology*, 16 (1), pp. 3-18.

Haykin, S. 1999. Neural Networks: A Comprehensive Foundation, 2nd edn, Prentice Hall, Englewood Cliffs, NJ.

Hoskinson, R.L., D. Pokrajac, Z. Obradovic, and A. Lazarevic. 2002. The unpredictability of soil fertility across space and time. In: Proc. Sixth International Conference on Precision Agriculture and other Precision Resource Management. Minneapolis, MN, *in press.*

Isaaks, E. H., and R. M. Srivastava. 1990. Applied Geostatistics Oxford Univ. Press, New York.

Lütkepohl, H. 1991. Introduction to Multiple Time Series Analysis, Springer Verlag, Berlin.

Olea, R.A.. 1999. Geostatistics for Engineers and Earth Scientists, Kluwer Academic Publishers, Boston. MA.

Pokrajac, D., and Z. Obradovic. 2001. Improved spatial-temporal forecasting through modeling of spatial residuals in recent history. In: Proc. First SIAM International Conference on Data Mining, Chicago, paper No. 9, CD-ROM, ISBN 0-89871-495-8.

Pokrajac, D., R. L. Hoskinson, and Z. Obradovic. 2002. Modeling spatial-temporal data with a short observation history. *Knowledge and Information Systems: An International Journal*, in press.

Vucetic, S., and Z. Obradovic. 2000. Performance controlled data reduction for knowledge discovery in distributed databases. In: Proc. Pacific-Asia Knowledge Discovery in Databases Conf. 2000, Kyoto, Japan, pp. 29-39.