Sixth Conference on Artificial Intelligence Applications to Environmental Science
88th AMS Annual Meeting, New Orleans, LA
20-24 January 2008

J2.6

Imputation of missing data with nonlinear relationships

Michael B. Richman[1], Indra Adrianto[2], and Theodore B. Trafalis[2]

[1]School of Meteorology, University of Oklahoma, 120 David L. Boren Blvd, Suite 5900, Norman, OK 73072, USA. Phone: 405-325-1853; Fax: 405-325-7689.
E-mail: mrichman@ou.edu

[2]School of Industrial Engineering, University of Oklahoma, 202 West Boyd St, Room 124, Norman, OK 73019, USA.
E-mails: ttrafalis@ou.edu, adrianto@ou.edu

Introduction

- A problem common to meteorological and climatological datasets is missing data.
- In cases where the individual observations are thought not important, or there are few missing data, deletion of every observation missing one or more pieces of data (complete case deletion) is common. It has been assumed this is innocuous.
- As the amount of missing data increases, tacit deletion has been shown to lead to bias in the remaining data and in subsequent analyses, such as data mining.

Introduction (Cont.)

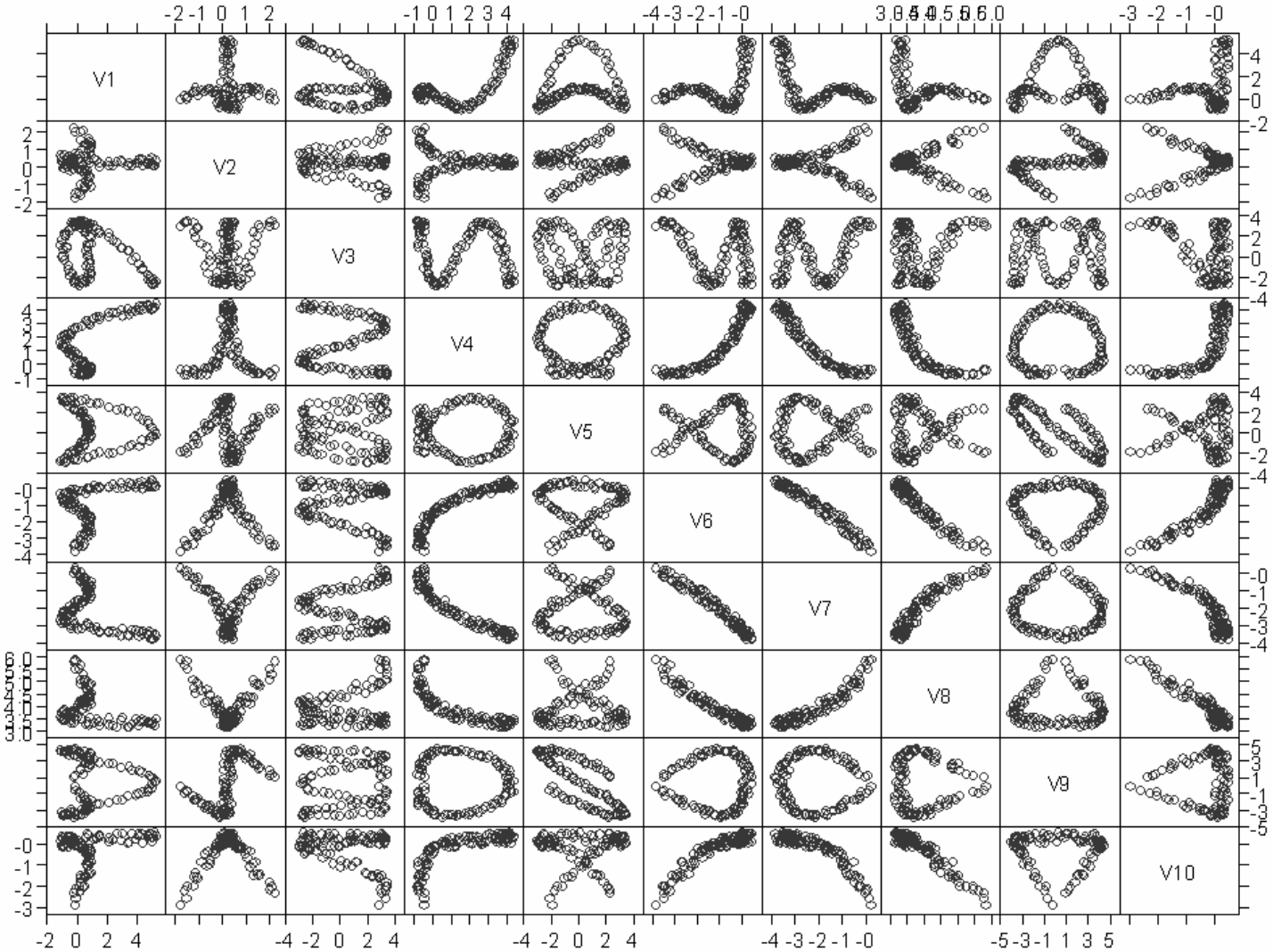
- If the data are deemed important to preserve, some method of imputing the missing values may be used.
- The present analysis seeks to examine how a number of techniques used to estimated missing data perform when missing data exist for configurations where the relationships are nonlinear. Previous work tested algorithms on nearly linear relationships.
- In this work, different types of machine learning techniques, such as support vector machines (SVMs) and, artificial neural networks (ANNs) are tested against standard imputation methods (regression based EM algorithm).

Data Set

- A nonlinear synthetic data set is generated using the following function with 10 variables:

$$y = \frac{5\sin 10x_1}{10x_1} + 2{x_2}^3 - 3\cos 10x_3 + \frac{4\sin 5x_4}{5x_4} - 3\sin 4x_5 - 4{x_6}^2 - 9\frac{\sin 3x_7}{7x_7} + \frac{3}{\cos x_8}$$
$$+ 4\sin 3x_9 - 3{x_{10}}^4 + \zeta$$

- where $-1 \leq x_i < 1$ for $i = 1,\ldots, 10$, with a 0.02 increment, and ζ is the uniformly distributed noise between 0 and 0.5.
- The data set consists of 100 rows and 10 columns.



- **Figure 1**. A scatter plot matrix for the nonlinear synthetic data set. Each row or column represents a variable.

Data Set (Cont.)

- The data set is altered to produce missing data by randomly removing one or more data in four different percentages (1%, 2%, 5% and 10%) of missing data (see Table 1).
- For each percentage of missing data, we repeat the alteration process 100 times (replicates) in order to obtain stable statistics and confidence intervals.
- Since the data removed are known and retained for comparison to the estimated values, information on the error in prediction and the changes in the variance structure are calculated.

Data Set (Cont.)

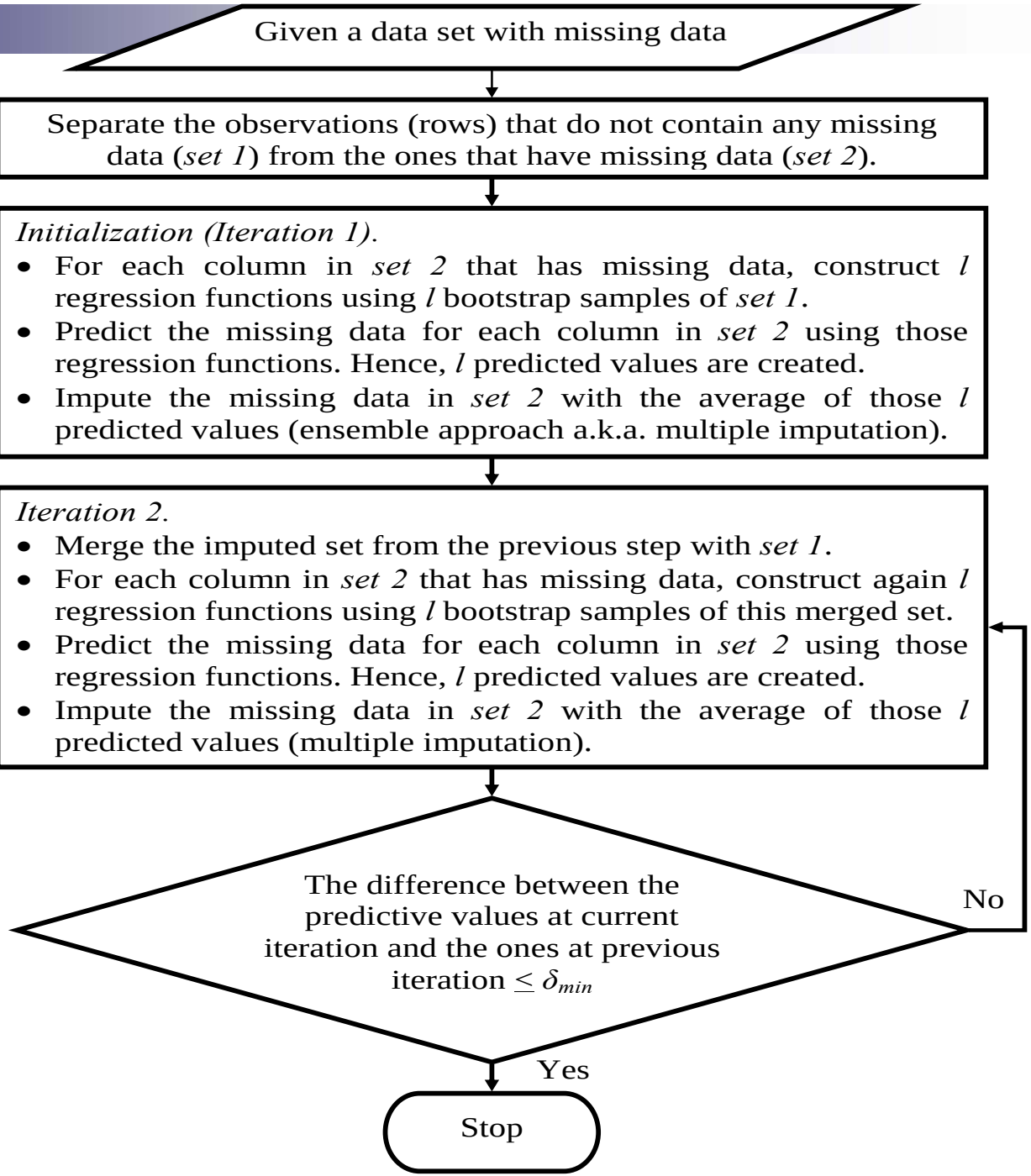
- **Table 1**. Four different percentages of missing data with the corresponding number of rows with missing data.

% of missing data (# of missing elements)	Average # of rows with missing data
1% (10 elements)	10
2% (20 elements)	15
5% (50 elements)	25
10% (100 elements)	50

Methodology

- Support vector machines (SVMs) and artificial neural networks (ANNs) are machine learning algorithms used in this research to predict missing data.
- Several standard methods such as casewise deletion, mean substitution, simple linear regression, and stepwise multiple regression, are employed for comparison.

- **Figure 2**. Flowchart of multiple imputation.



Methodology (Cont.)

- We use the mean squared error (MSE) to measure the difference between the actual values from the original data set and the corresponding imputed values.
- The difference of variance and covariance between the original data set and the imputed data set is measured using the mean absolute error (MAE).

Experiments

- We apply the multiple imputation methodology as described in Figure 2 to predict missing data with 10 bootstrap samples to obtain the ensemble mean.
- 5 iterations are performed for SVR, ANN, stepwise regression, and simple linear regression experiments.
- Additionally, mean substitution and casewise deletion are used in the experiments.
- The experiments are performed in the MATLAB environment using a Pentium M Centrino 1.8 GHz laptop with 1.23 GB RAM.
- The SVR experiments use LIBSVM toolbox (Chang and Lin, 2001) whereas the ANN, stepwise and simple linear regression experiments utilize the neural network and statistics toolboxes, respectively.

Experiments (Cont.)

- For SVR experiments, different combinations of kernel functions (linear, polynomial, radial basis function) and *C* values (the tradeoff between the flatness of a regression function and the amount up to which the deviations larger than ε are tolerated) are applied to determine the parameters that give the lowest MSE.
- More flatness means we try to find the small weight of a regression function.
- The “best” SVR parameters use the radial basis function (RBF) kernel and ε-insensitive loss function with $\varepsilon = 0.07$, $\gamma = 0.09$ (the parameter that controls the RBF width) and $C = 10$.

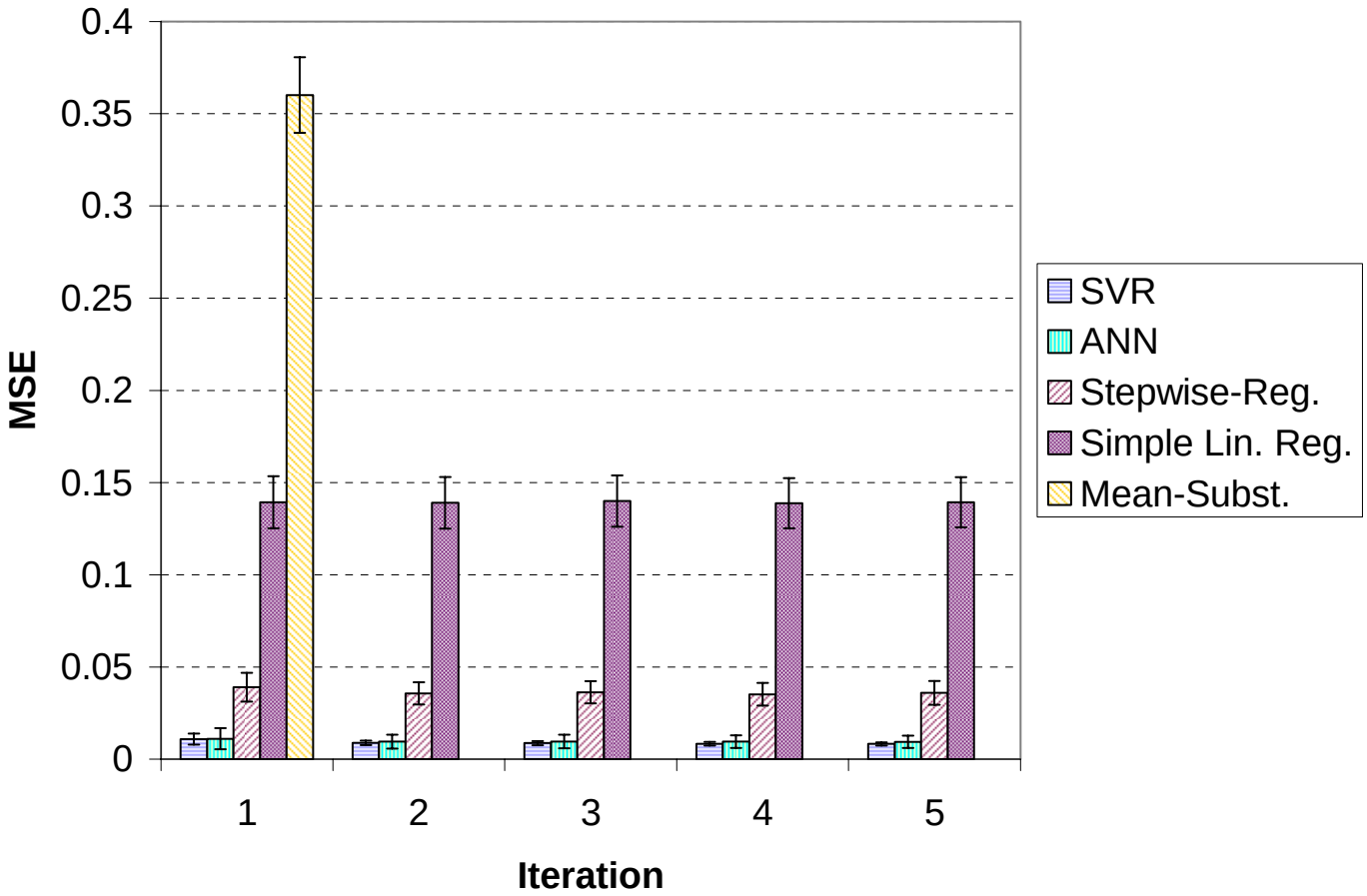
Experiments (Cont.)

- For ANNs, we train several feed-forward neural networks using one hidden layer with different number of hidden nodes (from 1 to 10) and different activation functions (linear and tangent-sigmoid) for the hidden and output layers.
- The scaled conjugate gradient backpropagation network is used for the training function.
- To avoid overfitting, the training stops if the number of iterations reaches 100 epochs or if the magnitude of the gradient is less than 10^{-6}.
- The neural network that gives the lowest MSE has 4 hidden nodes with the tangent-sigmoid for the hidden layer and linear activation functions for the output layer.

Experiments (Cont.)

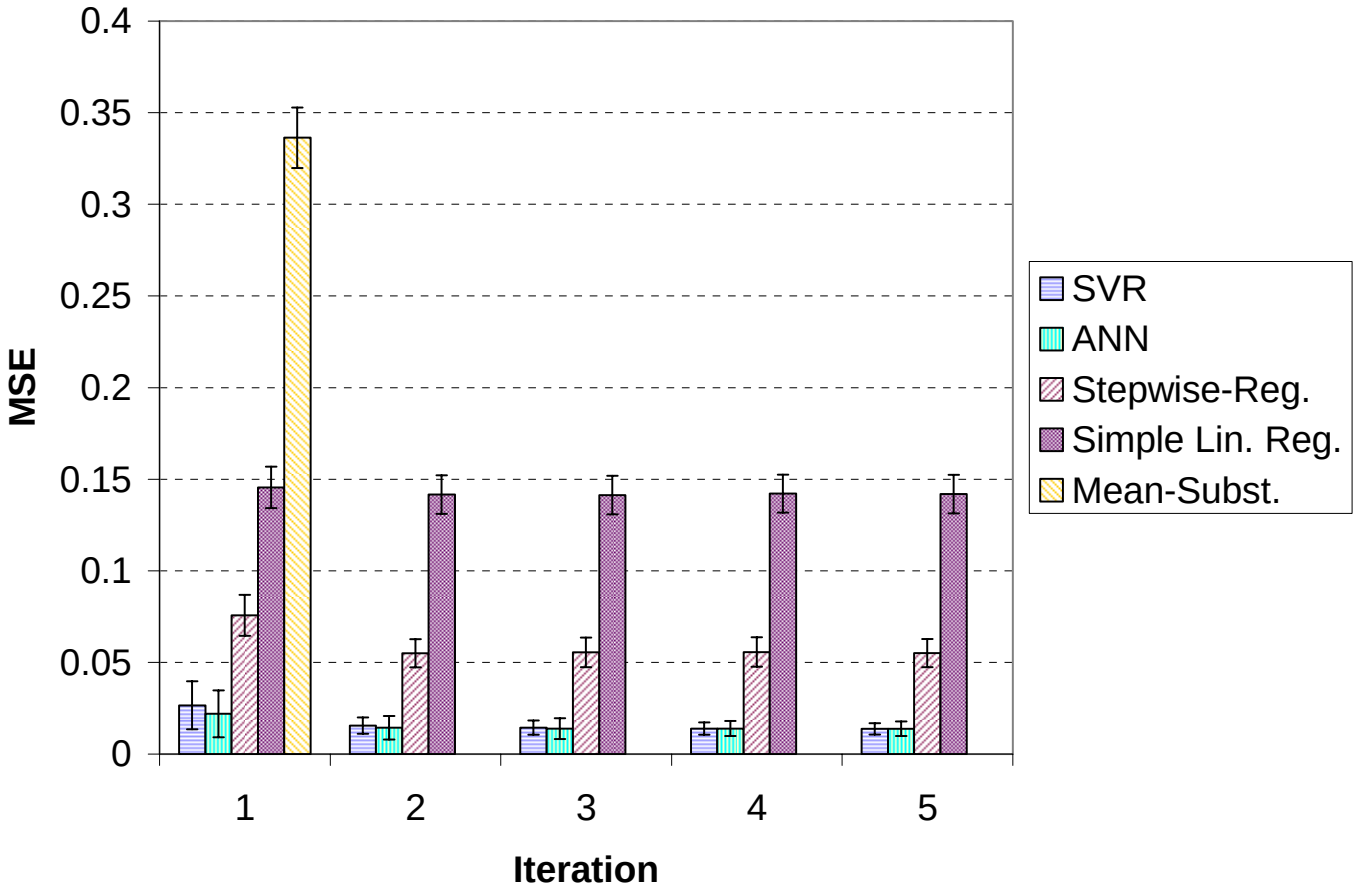
- For stepwise regression, the maximum *p*-value that a predictor can be added to the model is 0.05 whereas the minimum *p*-value that a predictor should be removed from the model is 0.10.
- Simple linear regression uses only one independent variable that has the highest correlation with the response variable to predict missing data.
- For mean substitution, the missing data in a variable are replaced with the mean of its variable.

Results



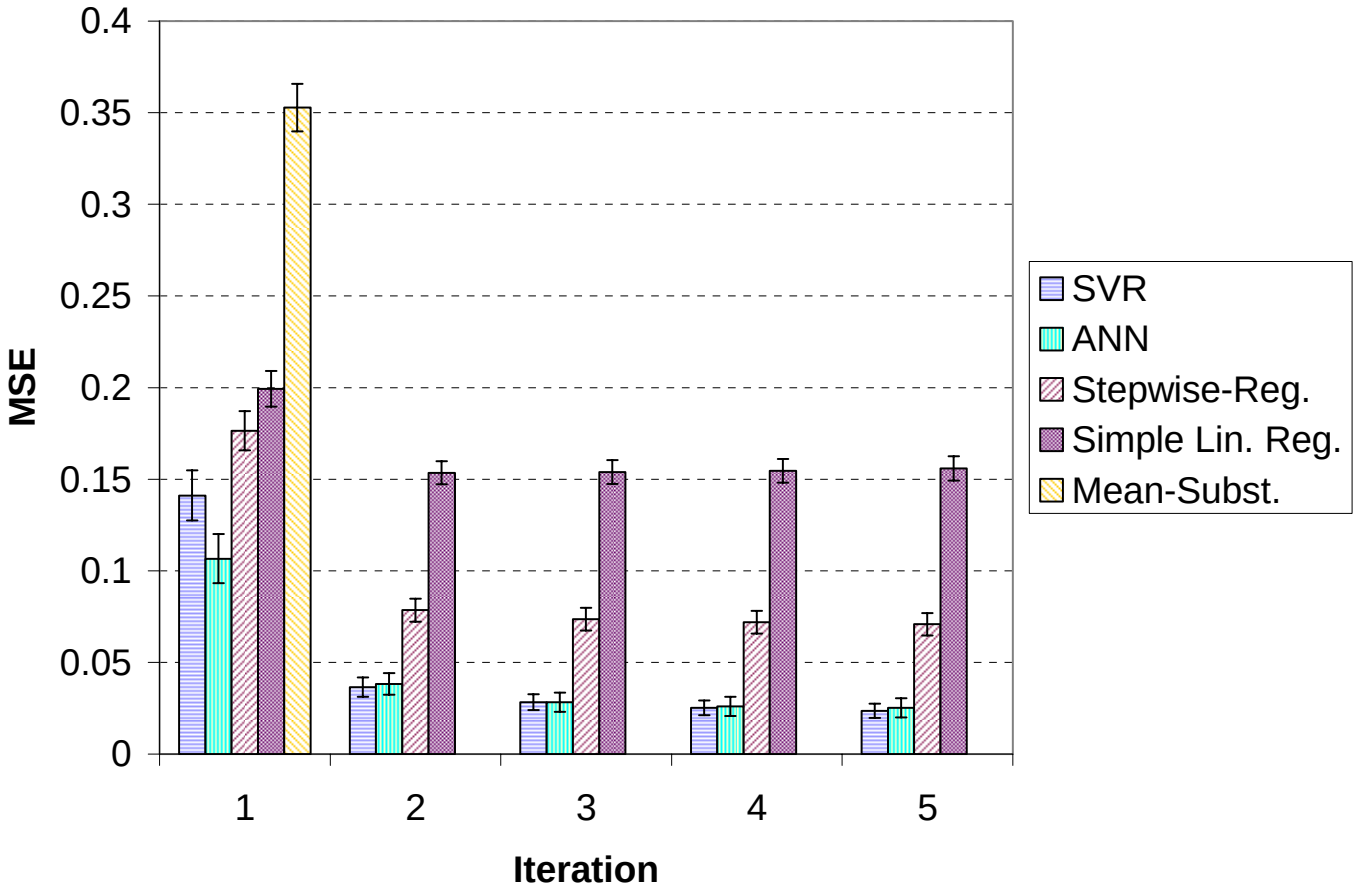
- **Figure 3.** The average MSE for all methods after 100 replicates with 95% confidence intervals for 1% missing data.

Results (Cont.)



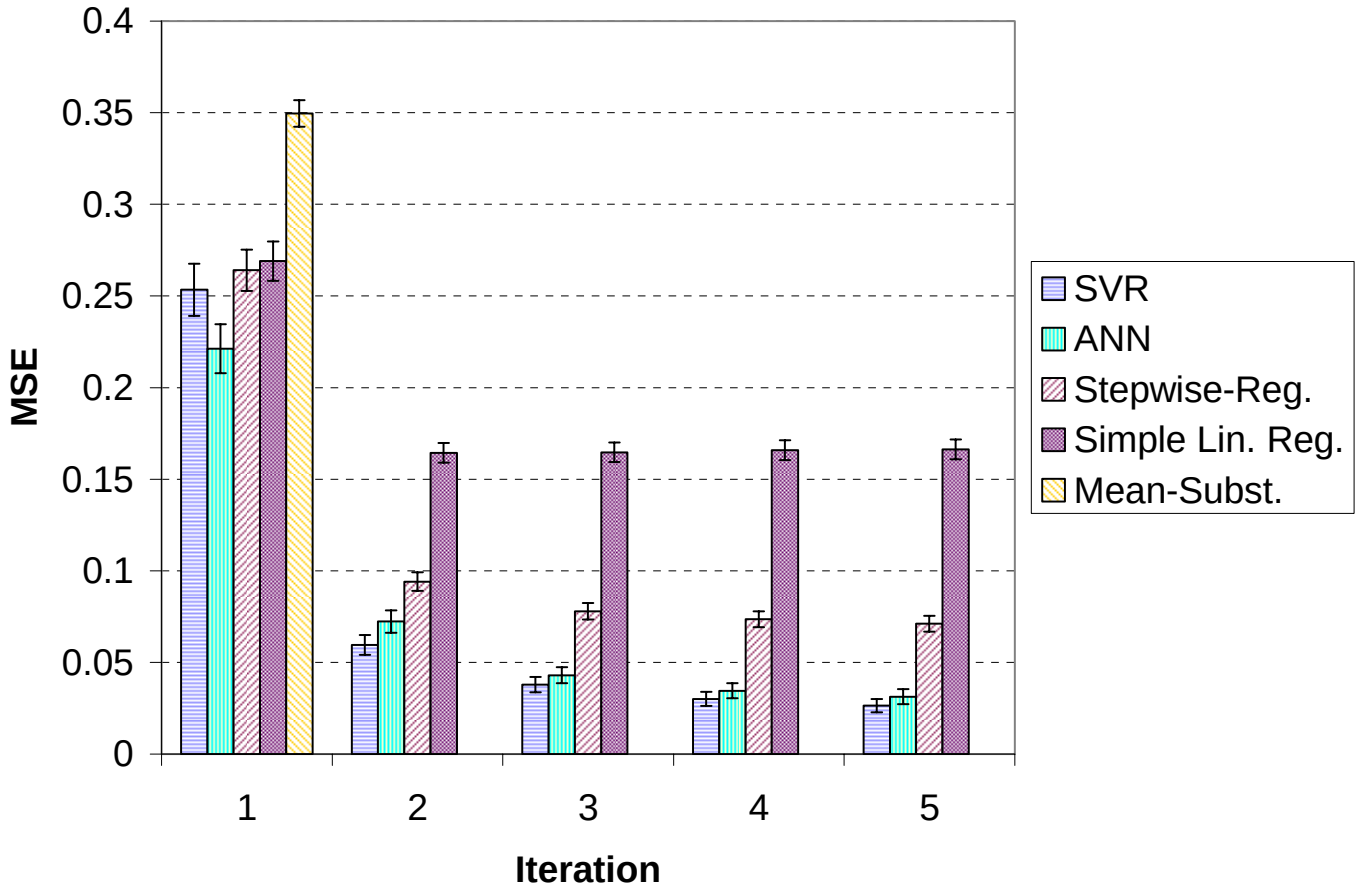
- **Figure 4.** The average MSE for all methods after 100 replicates with 95% confidence intervals for 2% missing data.

Results (Cont.)



- **Figure 5.** The average MSE for all methods after 100 replicates with 95% confidence intervals for 5% missing data.

Results (Cont.)



- **Figure 6.** The average MSE for all methods after 100 replicates with 95% confidence intervals for 10% missing data.

Results (Cont.)

- **Table 2.** Reduction in MSE from Iteration 1 (initial guess) to other iterations for each method.

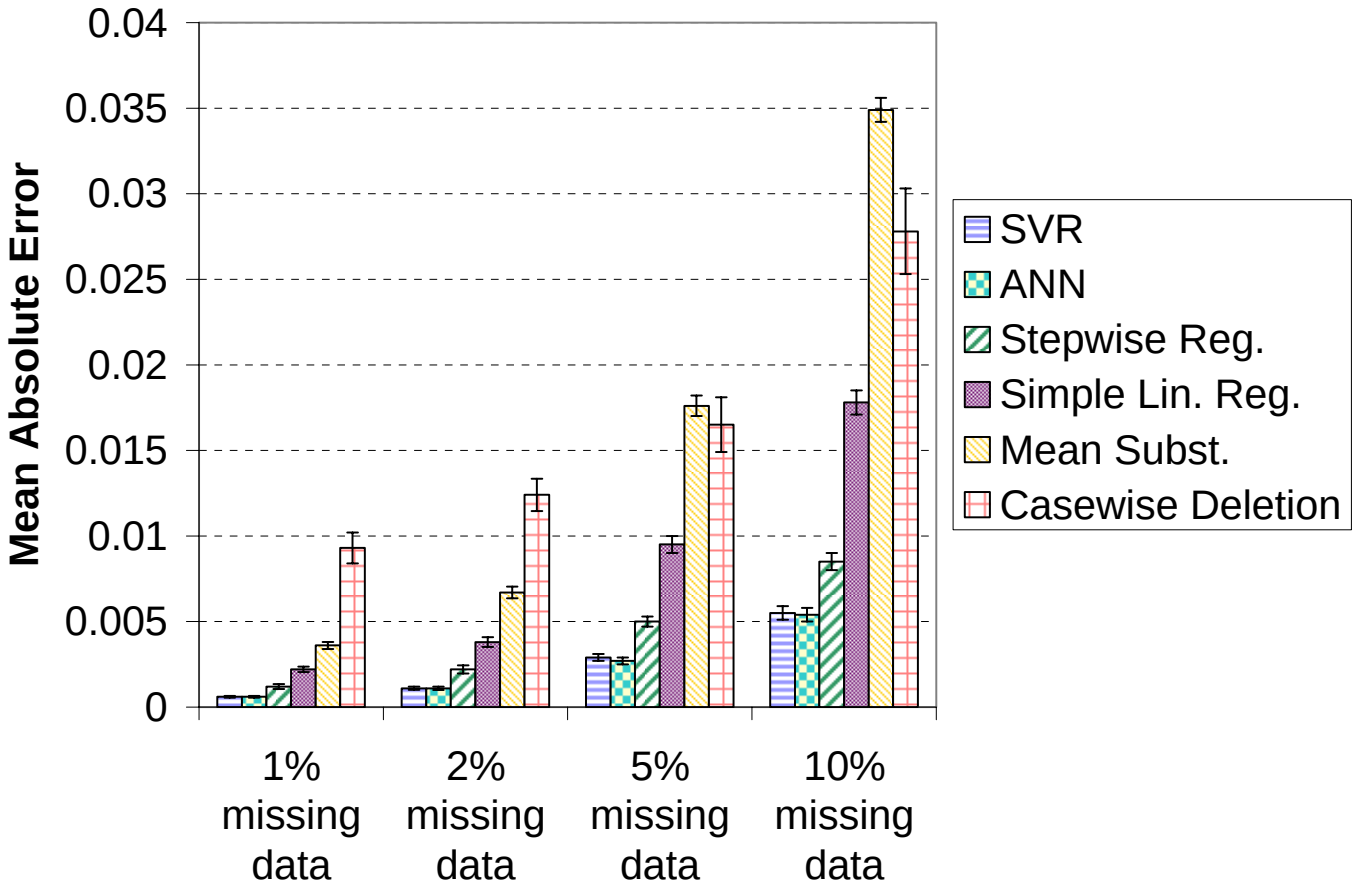
% missing data	Iteration	SVR	ANN	Stepwise Regression	Simple Linear Regression
1% missing data	Iter. 2	18.3%	13.5%	8.7%	0.2%
	Iter. 3	20.2%	13.5%	7.2%	-0.5%
	Iter. 4	22.9%	14.4%	10.0%	0.4%
	Iter. 5	23.9%	15.3%	7.9%	0.0%
2% missing data	Iter. 2	41.7%	34.2%	27.3%	23.0%
	Iter. 3	45.9%	36.5%	26.7%	22.8%
	Iter. 4	47.7%	36.5%	26.4%	22.5%
	Iter. 5	48.5%	37.0%	27.2%	21.8%
5% missing data	Iter. 2	74.1%	64.1%	55.5%	23.0%
	Iter. 3	79.9%	73.5%	58.3%	22.8%
	Iter. 4	82.1%	75.6%	59.2%	22.5%
	Iter. 5	83.3%	76.4%	59.9%	21.8%
10% missing data	Iter. 2	76.5%	67.3%	64.4%	38.9%
	Iter. 3	85.0%	80.6%	70.5%	38.8%
	Iter. 4	88.1%	84.4%	72.1%	38.4%
	Iter. 5	89.6%	85.8%	73.1%	38.2%

Results (Cont.)

- **Table 3.** Reduction in MSE from Stepwise Regression to SVR and ANN for each iteration.

% missing data	Iteration	SVR	ANN
1% missing data	Iter. 1	72.1%	71.6%
	Iter. 2	75.1%	73.1%
	Iter. 3	76.0%	73.6%
	Iter. 4	76.1%	73.0%
	Iter. 5	76.9%	73.9%
2% missing data	Iter. 1	64.9%	71.1%
	Iter. 2	71.8%	73.8%
	Iter. 3	74.1%	75.0%
	Iter. 4	75.0%	75.0%
	Iter. 5	75.1%	75.0%
5% missing data	Iter. 1	20.0%	39.6%
	Iter. 2	53.5%	51.2%
	Iter. 3	61.5%	61.5%
	Iter. 4	65.0%	63.9%
	Iter. 5	66.7%	64.4%
10% missing data	Iter. 1	4.0%	16.2%
	Iter. 2	36.7%	23.2%
	Iter. 3	51.3%	44.8%
	Iter. 4	59.1%	53.1%
	Iter. 5	62.9%	56.0%

Results (Cont.)



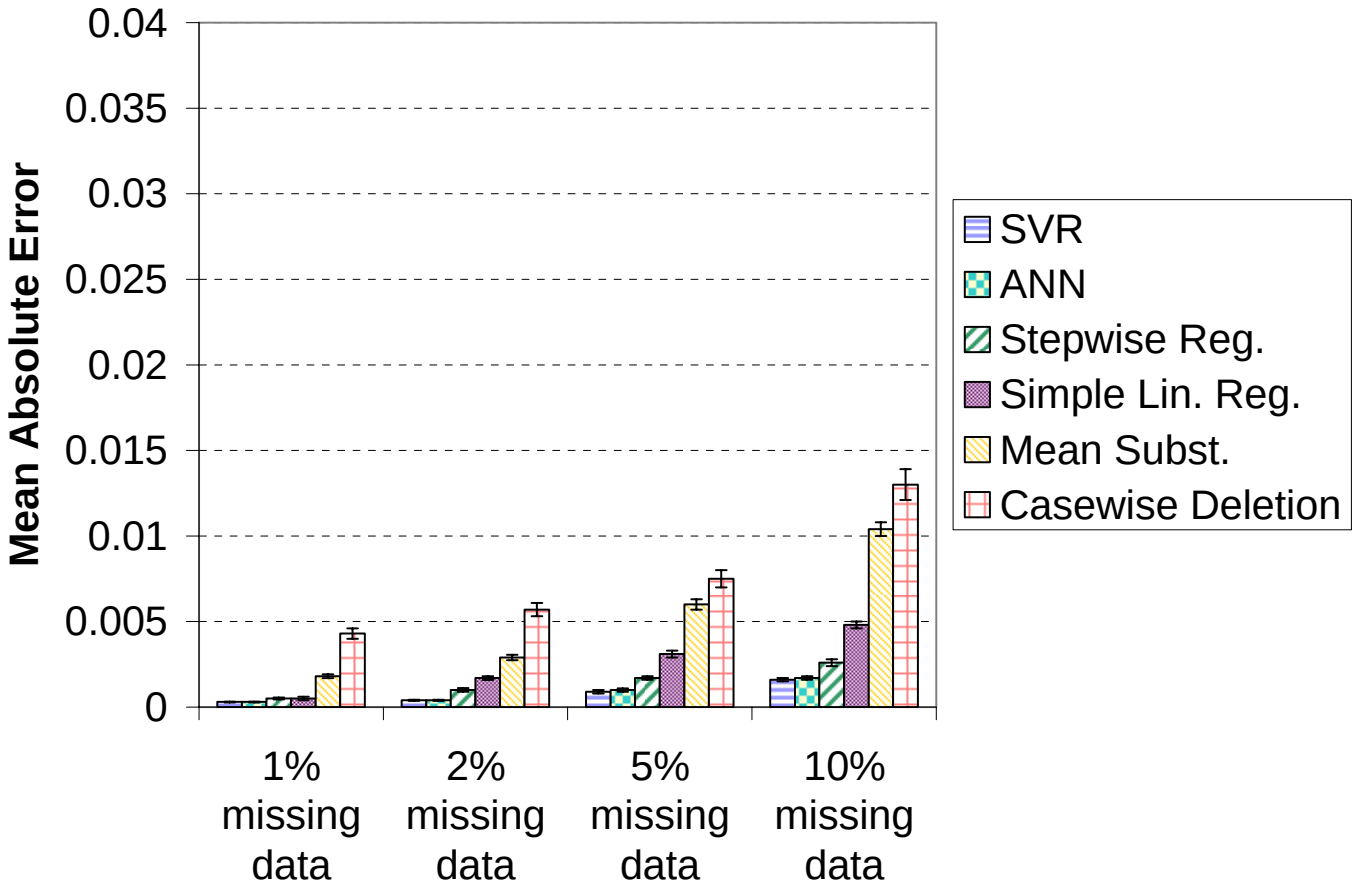
- **Figure 7.** A Bar chart with 95% confidence intervals illustrating the difference of **variance** between the original and imputed data set at Iteration 5.

Results (Cont.)

- **Table 4.** Reduction in MAE from Stepwise Regression to SVR and ANN at Iteration 5 for the difference of **variance** between the original and imputed data set.

% missing data	SVR	ANN
1% missing data	50.0%	50.0%
2% missing data	50.0%	50.0%
5% missing data	42.0%	46.0%
10% missing data	35.3%	36.5%

Results (Cont.)



- **Figure 8.** A Bar chart with 95% confidence intervals illustrating the difference between the lower triangle of the **covariance** matrix of the original data set and the one of the imputed data set at Iteration 5.

Results (Cont.)

- **Table 5.** Reduction in MAE from Stepwise Regression to SVR and ANN at Iteration 5 for the difference between the lower triangle of the **covariance** matrix of the original data set and the one of the imputed data set.

% missing data	SVR	ANN
1% missing data	40.0%	40.0%
2% missing data	60.0%	60.0%
5% missing data	47.1%	41.2%
10% missing data	38.5%	34.6%

Results (Cont.)

- **Table 6**. Computation time for each method substituting missing data.

Method	Computation time for 1 run in sec.			
	1% missing data	2% missing data	5% missing data	10% missing data
SVR	1.51	1.73	2.07	2.50
ANN	152.95	174.73	216.10	229.63
Stepwise Regression	2.89	3.86	4.08	4.86
Simple Linear Regression	1.01	1.36	1.59	1.85
Mean Substitution	0.25	0.26	0.32	0.33

Conclusions

- Most obvious: multiple imputation using machine learning algorithms (SVR and ANN) is superior to all other methods. In general, we find a 60 – 75% improvement over the traditional use of an iterated stepwise linear regression imputation in ***reducing MSE of the estimate***.
- Compared to analysis shown last year, the use of an ensemble mean, via 10 replicates, makes the results of multiple iterations clear: for very small amounts of missing data (1%) a single iteration is sufficient. ***As the amount of missing data increases, additional iterations are required for the results to stabilize.***

- The analysis of the MAE introduced into the variance-covariance matrix shows improvements with machine learning methods. The insertion of the mean value has a negative impact on the ***variance estimates*** in excess of casewise deletion when > 2% of the data are missing. This arises as the same value is getting inserted in many instances, so the variance is underestimated.
- Beyond that, SVR and ANN lead to a 35 – 50% improvement in MAE over the traditional method.
- For the ***covariances***, casewise deletion is always worst, followed by mean substitution, simple regression, multiple stepwise regression, ANN and SVR. However, ANN and SVR are statistically indistinguishable.
- For the modest number of inputs in this experiment, SVR is more efficient computationally by two orders of magnitude. One experiment (100 replicates) takes ~ 3½ min. for SVR and ~ 330 min. (5½ hrs) for ANN.