J3.9

MULTIPLE IMPUTATION THROUGH MACHINE LEARNING ALGORITHMS

Michael B. Richman[1*], Theodore B. Trafalis[2], and Indra Adrianto[2]
[1]School of Meteorology, [2] School of Industrial Engineering, University of Oklahoma, Norman, OK

A problem common to meteorological and climatological datasets is how to address missing data. The majority of multivariate analysis techniques require that all variables be represented for each observation; hence, some action is required in the presence of missing data. In cases where the individual observations are thought not important, deletion of every observation missing one or more pieces of data (complete case deletion) is common. As the amount of missing data increases, tacit deletion can lead to bias in the first two statistical moments of the remaining data as population estimators and inaccuracies in subsequent analyses. What is desired is a principled method that uses information available in the remaining data to predict the missing values. Such techniques include substituting nearby data, interpolation techniques and linear regression using nearby sites as predictors. One class of technique that uses the information available in an iterative manner is known as multiple imputation.

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 (e.g., multiple regression), simple regression, mean substitution, and casewise deletion. All methods are used to predict the known values of climatological data which have been altered to produce missing data. These data sets are on the order of 400 variables (data station sites) and a large number of observations. Both precipitation and air temperature data are used to provide a range of inherent spatial coherence seen by analysts.

The MSE of the prediction and the MAE of the variance are presented to assess the efficacy of each technique. Results indicate that the non-iterative methods, such as casewise deletion and mean substitution, lead to the largest errors and iterative imputation has considerably lower errors. Within the iterative techniques, SVMs are most promising in reducing error.

1. INTRODUCTION

How to address missing data in meteorological and climatological datasets is an issue most researchers face. The decisions made can have a profound impact on subsequent analyses (e.g., Kidson and Trenberth, 1988 and Duffy et al. 2001 summarize the importance of this issue). The majority of multivariate analysis techniques require that all variables be represented for each observation; hence, some action is required in the presence of missing data. Additionally, proxy-based reconstruction methods are sensitive to the technique used to relate the data that are present to those that are missing (Rutherford et al., 2005; Mann et al., 2005).

In cases where the individual observations are thought not important, deletion of every observation missing one or more pieces of data (complete case deletion) is common. As the amount of missing data increases, tacit deletion can lead to bias in the first two statistical moments of the remaining data and inaccuracies in subsequent analyses. In datasets, where extreme values are of importance, extremes in wind speed and rainfall may be associated with meteorological conditions that lead to instrument failure and loss of data. Significantly, it is those extreme values that are of interest. If the data are deemed important to preserve, some method of imputing the missing values may be used.

Historically, the statistical mean has been used most often as it was thought to minimize perturbations. Despite that, the use of the mean injects the same value into every instance of missing data and has been shown to create artificially low variation (Roth et al., 2005). What is desired is a principled method that uses information available in the remaining data to predict the missing values. Such techniques include substituting nearby data, interpolation techniques and linear regression using nearby sites as predictors. One class of technique that uses the information available in an iterative manner is known as multiple imputation.

The results from any technique used to estimate missing data depend, to a large extent, on the patterns of interrelated data (the degree of oversampling) and the manner in which the data are missing. The mechanism responsible for missing data should be assessed as random or systematic. In many cases, a few consecutive missing observations can be estimated with little error; however, if a large amount of data is missing, the results would be different. Motivated by such design questions, the present analysis seeks to examine how a number of techniques used to estimate missing data perform when various types and amounts of missing data exist.

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 (e.g., multiple regression). All methods are used to predict the known values of climatological data which have been altered to produce missing data. These data sets are on the order of 400 variables (data station sites) and a large number of

**Corresponding author address*: Michael B. Richman, University of Oklahoma, School of Meteorology,
120 David L. Boren Blvd, Norman, OK 73072;
Email: mrichman@ou.edu

observations. Both precipitation and air temperature data are used to provide a range of inherent spatial coherence seen by analysts as the former is known to have a small correlation scale whereas the latter has a much larger spatial scale).

Many research studies have investigated multiple imputation (e.g., Rubin, 1988; Wayman, 2003). Rubin (1988) showed the remarkable improvements when using multiple imputation rather than single imputation. Wayman (2003) discussed some missing data issues and explained the basic process of multiple imputation. Variations of regression forms of imputation techniques have been applied to climate data (e.g., EM algorithm, Schneider, 2001) with promising results. We implement imputation techniques herein with the newest learning methods, such as ANN and SVR. These are compared to older techniques to document improvement over older techniques.

The data used in the analyses are described in Section 2. A brief overview of the methodology and experiments is provided in Sections 3 and 4. The results are summarized in Section 5 and conclusions presented in Section 6.

2. DATA SETS

There are two data sets used in this study based on the Lamb/Richman climate datasets (Richman and Lamb, 1985). The first data set is the monthly precipitation data set, where the values are reported in units of inches (to the hundredth of an inch). This data set consists of 528 monthly observations (1949 – 1992) with 400 variables or stations. Precipitation data are included to test how well the techniques work on data with a small spatial scale. The second data set is monthly average temperature measured in degrees Celsius. This data set consists of 528 observations with 400 variables and is included to assess the techniques for data with a large spatial scale. Since the spatial scale of the temperature covariation is much larger than the station spacing, analysis of these data should provide insight into situations that are similar to superobing, where data thinning techniques are desirable. Each data set is altered to produce missing data by randomly removing three different percentages (5%, 10%, and 20%) of the observations. Since the data removed are known and retained for comparison to the estimated values, information of the error in prediction and the changes in the variance structure are calculated.

3. METHODOLOGY

The support vector machines (SVMs) and artificial neural networks (ANNs) are machine learning algorithms used in this paper to predict missing data. Several standard methods such as casewise deletion, mean substitution, simple linear regression, and stepwise multiple regression, are employed for comparison. The SVM algorithm was initially developed by Vapnik and has become a favored method in machine learning (Boser et al., 1992). The version of SVMs for regression called support vector regression (SVR) is used in this study. Trafalis et al. (2003) applied SVR for prediction of rainfall from WSR-88D Radar and showed that SVR is better, in terms of generalization error, than traditional regression.

The SVR formulation by Vapnik (1998) can be described as follows. Given a training set $\{(\mathbf{x}_i, y_i)\}_{i=1}^{\ell}$ of ℓ observations, our objective is to construct a function for approximating expected values *y*: $f(\mathbf{x}) = \langle \mathbf{w} \cdot \mathbf{x} \rangle + b$ where **w** is the weight vector and *b* is a bias. Vapnik (1998) proposed the linear *ε*-insensitive loss function in the support vector regression (SVR) formulation (Fig. 1). The linear *ε*-insensitive loss function is defined by:

$$L_\varepsilon(\mathbf{x}, y, f) = \begin{cases} 0 & \text{if } |y - f(\mathbf{x})| \le \varepsilon \\ |y - f(\mathbf{x})| - \varepsilon & \text{otherwise} \end{cases} \qquad (1)$$

The SVR formulation can be represented as follows (Vapnik, 1998):

$$\begin{aligned} \text{min } & \phi(\mathbf{w}, \xi, \xi') = \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{l}(\xi_i + \xi_i') \\ \text{subject to } & (\langle \mathbf{w} \cdot \mathbf{x}_i \rangle + b) - y_i \le \varepsilon + \xi_i, \\ & y_i - (\langle \mathbf{w} \cdot \mathbf{x}_i \rangle + b) \le \varepsilon + \xi_i', \\ & \xi_i, \xi_i' \ge 0, \quad i = 1, \ldots, l \end{aligned} \qquad (2)$$

where **w** is the weight vector, *b* is a bias, *C* is a user-specified parameter, and ξ_i, ξ_i' are slack variables representing the deviations from the constraints of the *ε*-tube.

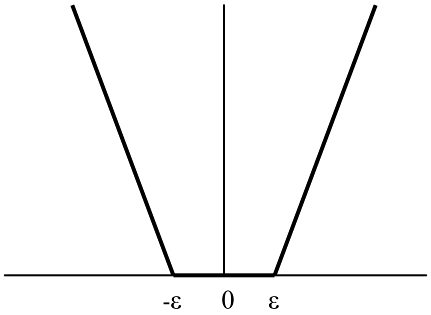

Figure 1. The *ε*-insensitive loss function.

The SVR formulation in Eq. 2 can be solved in the dual formulation using Lagrange multipliers α_i, α_i' where: $\mathbf{w} = \sum_{i=1}^{l}(\alpha_i' - \alpha_i)\mathbf{x}_i$ .Using the linear *ε*-insensitive loss function, the dual formulation becomes (Vapnik, 1998):

$$\begin{aligned} \text{max } Q(\alpha, \alpha') = & \sum_{i=1}^{l} y_i(\alpha_i' - \alpha_i) - \varepsilon\sum_{i=1}^{l}(\alpha_i' + \alpha_i) \\ & - \frac{1}{2}\sum_{i=1}^{l}\sum_{j=i}^{l}(\alpha_i' - \alpha_i)(\alpha_j' - \alpha_j)\langle \mathbf{x}_i \cdot \mathbf{x}_j \rangle \end{aligned}$$

subject to

$$\sum_{i=1}^{l}(\alpha_i' - \alpha_i) = 0, \quad 0 \le \alpha_i, \alpha_i' \le C, \quad i = 1,...,l \qquad (3)$$

In the case of nonlinear problems (Fig. 2), given a function $\phi : \mathbf{x} \rightarrow \phi(\mathbf{x})$ which maps $\mathbf{x}$ from the input space into a higher dimensional feature space, the inner product $\langle \mathbf{x}_i \cdot \mathbf{x}_j \rangle$ in Eq. 3 can be replaced by a kernel function $k(\mathbf{x}_i, \mathbf{x}_j)$. The following kernel functions are used in this study:

1. Linear: $k(\mathbf{x}_i, \mathbf{x}_j) = \langle \mathbf{x}_i \cdot \mathbf{x}_j \rangle$.
2. Polynomial: $k(\mathbf{x}_i, \mathbf{x}_j) = (\langle \mathbf{x}_i \cdot \mathbf{x}_j \rangle + 1)^p$, *p* is the degree of polynomial.
3. Radial basis function (RBF): $k(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\gamma \|\mathbf{x}_i - \mathbf{x}_j\|^2\right)$, *γ* is the parameter that controls the width of RBF.

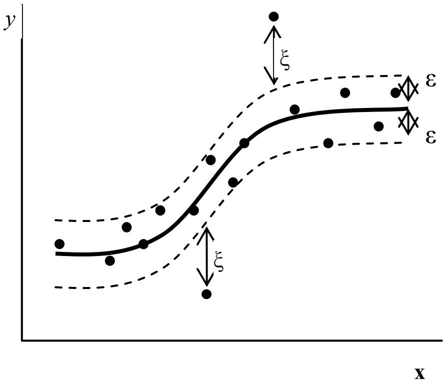

Figure 2. Nonlinear regression problem.

The type of ANNs used herein is feedforward ANNs (Fig. 3). The network consists of a set of information-processing units called neurons that constitute an input layer, one or more hidden layers, and an output layer of computational nodes (Haykin, 1999). The formulation of feedforward ANNs is well explained by Haykin (1999).

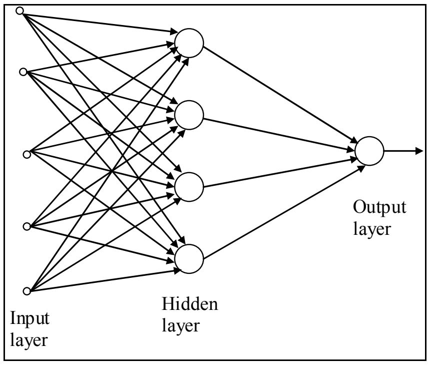

Figure 3. A feedforward neural network with one hidden layer and one output layer.

The multiple imputation scheme used in this study can be described as follows:

Step 1. Given a data set of multiple observations (rows) and variables (columns). Identify which rows and columns that have missing data.

Step 2. Separate the observations that do not contain any missing data (*set 1*) with the ones that have missing data (*set 2*).

Step 3. *Iteration 1.* For each column in *set 2* that has missing data, construct regression functions using *set 1*. The dependent or response variable is the column that has missing data and the independent or predictor variables are the other columns. Predict the missing data for each column in *set 2* using those regression functions. Therefore, we have created values (imputes) to be substituted the missing data in *set 2*.

Step 4. *Iteration 2.* Merge the imputed set from previous step with *set 1*. For each column in *set 2* that has missing data, construct again regression functions using this merged set. Predict the missing data for each column in *set 2* using the regression functions from previous. Therefore, we have created again values (imputes) to be substituted the missing data in *set 2*.

Step 5. *Iteration 3.* The same as *Iteration 2*.

Several iterations can be applied to construct imputed data sets. Our experiments show that 3 iterations should be adequate to substitute missing data. In this study, we apply SVR, ANNs, and stepwise-regression to construct the regression functions for multiple imputation methods. For single imputation methods, we perform mean substitution and simple linear regression.

In order to measure performance of our methods, we use the mean squared error (MSE) to show the difference between the original data set and the imputed data set. For *N* observations, the MSE is the average squared error between the predictions *y* and the target outputs *t*,

$$MSE = \frac{1}{N}\sum_{j=1}^{N}(y_j - t_j)^2 \qquad (4)$$

The analysis on the difference of variance between the original data set and the imputed data set is performed. At this point, we add another standard method namely casewise or listwise deletion where the observations or rows that contain missing data are removed or not included in the data analysis. We cannot use the MSE to measure the performance of casewise deletion because there is not any imputation to replace missing data. The difference of variance between the original data set and the imputed data set is measured using the mean absolute error (MAE). For *N* observations, the MAE is the average absolute error between the variance of variables in the original data set

T and the variance of variables in the imputed data set *P*,

$$MAE = \frac{1}{N}\sum_{j=1}^{N}\left|T_j - P_j\right| \qquad (5)$$

4. EXPERIMENTS

For each data set, we create randomly 10 different seed data sets for 5%, 10%, and 20% of the observations that have missing data. Then we applied the multiple imputation methodology as described in Section 3 to predict missing data where SVR, ANNs, and stepwise-regression are used to construct the regression functions. In addition, stepwise-regression is combined with ANNs as a regressor to examine the potential gained by predictor thinning. Using stepwise regression reduces the number of variables as predictors and these reduced set of variables are used to construct a regression function using ANNs. For multiple iteration experiments, 5 iterations are applied for each method. Additionally, these data sets are used for single imputation methods using mean substitution and simple linear regression to substitute missing data.

The experiments are performed in the Matlab® environment. The SVM experiments use LIBSVM toolbox (Chang and Lin, 2001) whereas the ANN, simple linear and stepwise regression experiments utilize the neural network and statistics toolboxes, respectively.

5. RESULTS

Tables 1-3 and Figures 4-6 show the results for each method with three different percentages of the observations that have missing data for the precipitation data set. The average MSE from 10 different randomly seeded data sets is reported. For SVR experiments, the different combinations of kernel functions (linear, polynomial, radial basis function) and penalty cost (*C*) values are applied to determine the parameters that give the lowest MSE. After experimentation, the SVR parameters used the radial basis function kernel and *ε*-insensitive loss function with *γ* = 0.000006, *C* = 100, and *ε* = 0.3. For ANNs, we train several feed-forward neural networks with different number of hidden nodes with the tangent-sigmoid activation function for the hidden layer and a linear activation function for the output layer. The scaled conjugate gradient backpropagation network is used for the training function. Training stopped when 50 epochs is reached. Based on this testing, the neural network that gives the lowest MSE has 4 hidden nodes. 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. The number of predictor variables used in stepwise regression for the precipitation data set is between 13 and 85. For the method that combines stepwise regression and ANNs, the best neural network has 5 hidden nodes and training stopped when 40 epochs is reached. For mean substitution, the missing data in a variable are replaced with the mean of the same monthly observations are used for that variable. This monthly stratification accounts for cyclostationarity by preventing inclusion of data into the mean from months where the temperature or precipitation might be much different from the month with missing data. Simple linear regression uses only one independent variable that has the highest correlation with the response variable to predict missing data. There is no iteration used for simple regression. By doing so, the simple regression acts similar to using the nearest station to the missing data as a proxy.

For the precipitation data, with 5% missing data, some interesting results emerge (Table 1 and Fig. 4) in prediction of the missing values. The most noteworthy finding is that the technique most commonly employed in the literature, mean substitution, results in an average error of 4.72 in^2 for the missing value (2.17 inches / month). Application of simple linear regression with a single predictor reduces the error considerably to 2.58 in^2 $^{(}$1.61 in. /mo.). Every iterated technique works considerably better. Stepwise regression, ANN and the combination are nearly tied in their MSE. A typical value for MSE is about 1.6 in^2 (1.26 in./mo.). The SVR performed best with a MSE of approximately 1.36 in^2 (1.17 in. /mo.).

Table 1. The average MSE for six methods with 5% of the observations missing for the precipitation data set.

Iteration	SVR	Stepwise Reg.	ANN	Stepwise Reg.&ANN	Mean Subst.	Simple Lin. Reg.
1	1.390	1.677	1.829	1.596	4.716	2.582
2	1.351	1.634	1.780	1.677		
3	1.352	1.605	1.571	1.615		
4	1.352	1.637	1.827	1.488		
5	1.352	1.556	1.693	1.684		

As the percentage of missing data was set at 10 % (Table 2, Fig. 5), the size of the error for mean substitution grows, whereas that for simple regression remains essentially constant. The iterative techniques begin to show more variation with ANN consistently the worst, though better than for simple regression. The stepwise regression and the stepwise regression followed by ANN gave similar results. As for the previous case, the SVM gave the lowest errors with an average mean square of 1.53 in^2 (1.24 in. /mo.).

Table 2. The average MSE for six methods with 10% of the observations missing for the precipitation data set.

Iteration	SVR	Stepwise Reg.	ANN	Stepwise Reg.&ANN	Mean Subst.	Simple Lin. Reg.
1	1.559	2.205	2.384	2.059	4.975	2.508
2	1.524	1.897	2.160	1.939		
3	1.524	1.949	2.219	1.969		
4	1.524	1.925	2.167	2.066		
5	1.524	1.965	2.120	1.905		

As the percentage of missing data was set at 20 % (Table 3, Fig. 6), the size of the error for mean substitution grows again, whereas that for simple regression grows at a slower rate. The iterative techniques continue to show that ANN consistently the worst, though it continues to perform better than for simple regression. The stepwise regression and the stepwise regression followed by ANN gave similar results. As for the previous case, the SVM gave the lowest errors with an average mean square of 1.83 in^2 (1.35 in./mo.).

Table 3. The average MSE for six methods with 20% of the observations missing for the precipitation data set.

Iteration	SVR	Stepwise Reg.	ANN	Stepwise Reg.&ANN	Mean Subst.	Simple Lin. Reg.
1	1.897	2.481	2.880	2.435	5.410	2.811
2	1.819	2.240	2.608	2.245		
3	1.811	2.287	2.608	2.218		
4	1.812	2.279	2.548	2.118		
5	1.815	2.292	2.581	2.255		

As the number of iterations increases, the behavior of each of the imputation techniques can be compared. All of the iterative imputation methods have a lower MSE in the second iteration, compared to the first. In many instances a third iteration leads to a small improvement over the second. Beyond three iterations, the techniques show no systematic reduction in MSE.

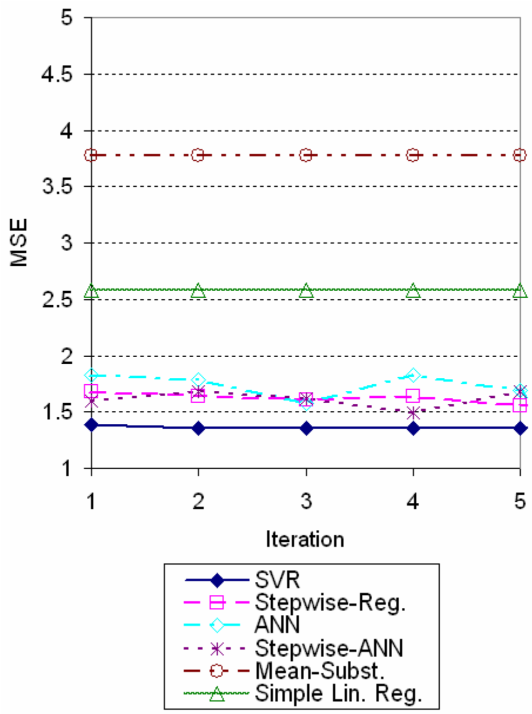

Figure 4. Average MSE for six methods for 5 iterations when 5% of the observations have missing data for the precipitation data set.

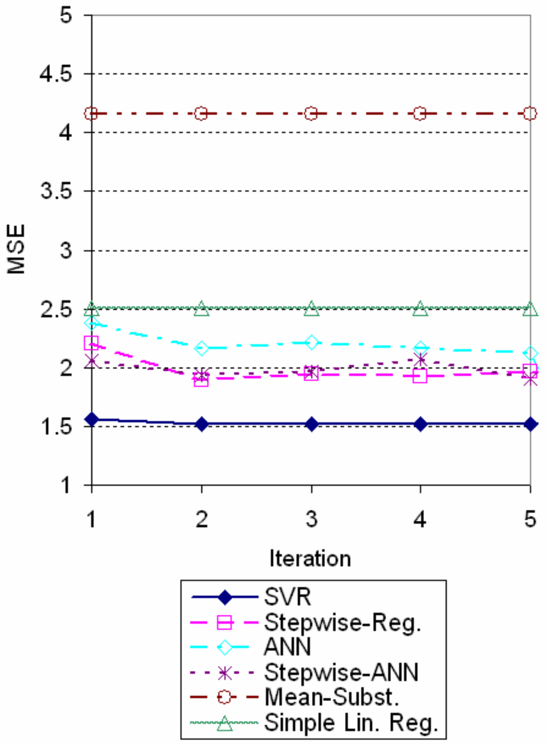

Figure 5. Average MSE for six methods for 5 iterations when 10% of the observations have missing data for the precipitation data set.

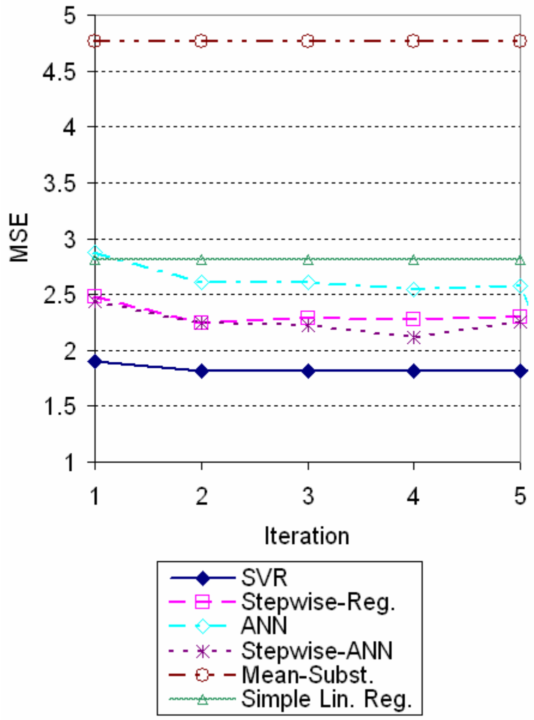

Figure 6. Average MSE for six methods for 5 iterations when 20% of the observations have missing data for the precipitation data set.

Table 4 and Figure 7 show the MAE of the difference of precipitation variance between the original and imputed data sets for each method with three different percentages of the observations that have missing data. The most obvious result is the deleterious effect of casewise deletion at all percentages of missing data. The mean substitution has the second worst ability to recreate the original data variances whereas the remaining techniques are tightly clustered.

Table 4. The MAE for each method to illustrate the difference of variance between the original and imputed data sets using the precipitation data set.

Method	5% obs. of missing data	10% obs. of missing data	20% obs. of missing data
SVR	0.005	0.005	0.007
Stepwise Reg.	0.006	0.005	0.008
ANN	0.004	0.005	0.010
Stepwise-ANN	0.006	0.004	0.008
Mean Subst.	0.007	0.008	0.011
Simple Lin. Reg.	0.006	0.006	0.009
Casewise Deletion	0.082	0.124	0.203

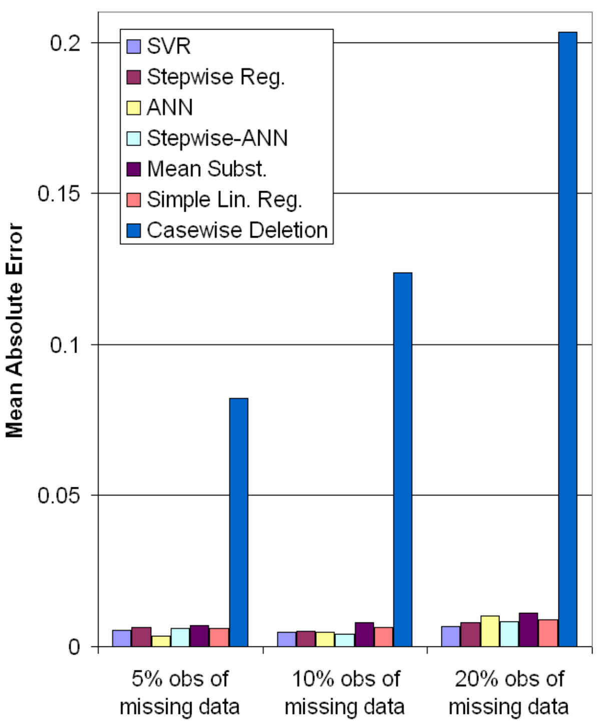

Figure 7. A bar chart illustrating the difference of variance between the original and imputed data sets using the precipitation data set. The MAE for each method is reported.

The same experiments are applied for the temperature data set. Tables 4-6 and Figures 8-10 illustrate the results for each method with three different percentages of the observations that have missing data for the temperature data set. For SVR, the best parameters used the radial basis function kernel and *ε*-insensitive loss function with *γ* = 0.000003, *C* = 1000, and *ε* = 0.1. For ANNs, the training activation and network functions for the precipitation data set are also used for the temperature data set. Training stopped when 300 epochs is reached. The best neural network has 5 hidden nodes. For stepwise regression, the same step in and step out p-values as used for the precipitation data set are employed. The predictor variables used in stepwise regression the temperature data set are between 19 and 83. For the method that uses both stepwise regression and ANNs, the best neural network has 10 hidden nodes and training stopped when 700 epochs is reached.

For 5% missing data, the use of the mean value for the same month (Table 5, Fig. 8) resulted in a large error (over 3.5 $^0C^2$ or almost 1.9 ^0C for the month). The simple linear regression is somewhat more accurate and reduces the error by close to 80%. Application of the iterative results led to a notable further decrease in the variance errors (approximately an additional 75% decrease in the error). Again, the SVR led to the minimum error in reconstructing the variance field.

Table 5. The average MSE for six methods with 5% of the observations missing for the temperature data set.

Iteration	SVR	Stepwise Reg.	ANN	Stepwise Reg.&ANN	Mean Subst.	Simple Lin. Reg.
1	0.144	0.179	0.237	0.179	3.670	0.751
2	0.142	0.156	0.206	0.164		
3	0.140	0.154	0.197	0.169		
4	0.140	0.159	0.211	0.167		
5	0.140	0.159	0.188	0.163		

For 10% missing data, the use of the mean value for the same month (Table 6, Fig. 9) resulted in a growth of error for the substitution (over 3.98 $^0C^2$ or almost 2 ^0C for the month). The simple linear regression results in some improvement, reducing the error by over 80%. Application of the iterative results led to a further decrease in the variance errors (approximately an additional 75%). As in the earlier experiment, the SVR led to the minimum error in reconstructing the variance field.

Table 6. The average MSE for six methods with 10% of the observations missing for the temperature data set.

Iteration	SVR	Stepwise Reg.	ANN	Stepwise Reg.&ANN	Mean Subst.	Simple Lin. Reg.
1	0.155	0.182	0.288	0.207	3.976	0.717
2	0.148	0.169	0.223	0.178		
3	0.147	0.173	0.214	0.177		
4	0.147	0.173	0.197	0.171		
5	0.147	0.178	0.208	0.164		

For 20% missing data, the use of the mean value for the same month (Table 7, Fig. 10) resulted similar error for the substitution (over 3.8 $^0C^2$ or almost 2 ^0C for the month). The simple linear regression gives about half the error reduction as the iterated techniques and reduces the error by over 80%. As in both previous experiments, the SVR gave the minimum error in the variance field reconstruction.

Table 7. The average MSE for six methods with 20% of the observations missing for the temperature data set.

Iteration	SVR	Stepwise Reg.	ANN	Stepwise Reg.&ANN	Mean Subst.	Simple Lin. Reg.
1	0.180	0.229	0.287	0.230	3.804	0.665
2	0.169	0.190	0.245	0.200		
3	0.165	0.183	0.213	0.200		
4	0.164	0.187	0.222	0.196		
5	0.164	0.194	0.213	0.206		

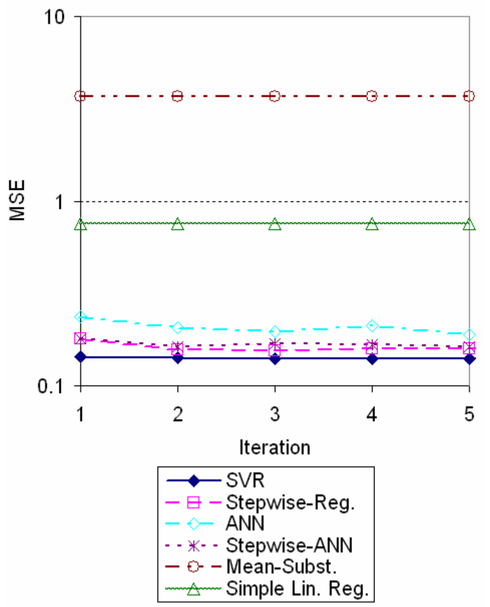

Figure 8. Average MSE for six methods for 5 iterations when 5% of the observations have missing data for the temperature data set.

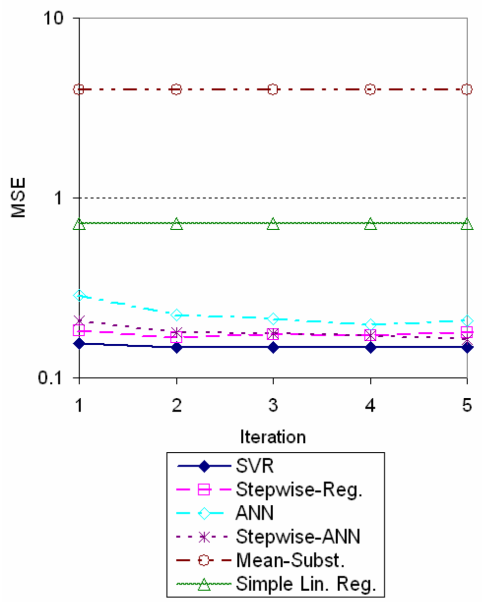

Figure 9. Average MSE for six methods for 5 iterations when 10% of the observations have missing data for the temperature data set.

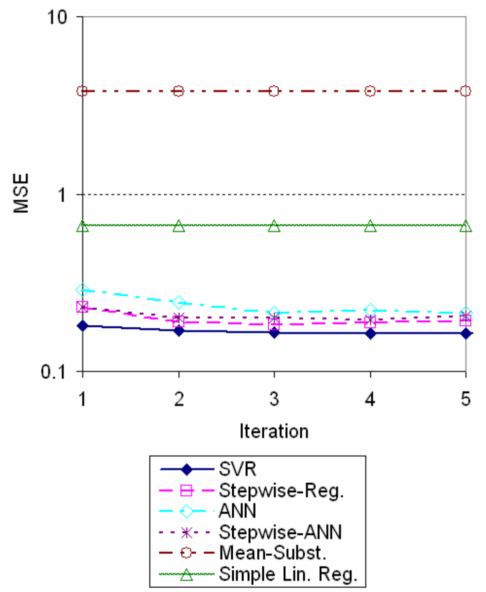

Figure 10. Average MSE for six methods for 5 iterations when 20% of the observations have missing data for the temperature data set.

As the number of iterations increases, the behavior of each of the imputation techniques can be compared. All of the iterative imputation methods have a lower MSE in the second iteration, compared to the first. In many instances a third iteration leads to a small improvement over the second. Beyond three iterations, the techniques show no systematic reduction in MSE. This behavior for the temperature data was consistent with that for the precipitation data.

In Table 8 and Fig. 11, the MAE of the difference of variance between the original and imputed data sets for each method with 5%, 10%, and 20% of the observations that have missing data using the temperature data set are reported. The variance analysis for the temperature data show interesting results (Table 8). Casewise deletion causes variance errors of over an order of magnitude more than the next closet method, mean substitution. Simple linear regression has errors that are larger than the iterated techniques. The results are shown graphically in Fig. 11.

Table 8. The MAE for each method to illustrate the difference of variance between the original and imputed data sets using the temperature data set.

Method	5% obs. of missing data	10% obs. of missing data	20% obs. of missing data
SVR	0.009	0.008	0.009
Stepwise Reg.	0.008	0.008	0.009
ANN	0.006	0.008	0.010
Stepwise-ANN	0.008	0.009	0.011
Mean Subst.	0.045	0.039	0.044
Simple Lin. Reg.	0.017	0.020	0.022
Casewise Deletion	0.232	0.311	1.014

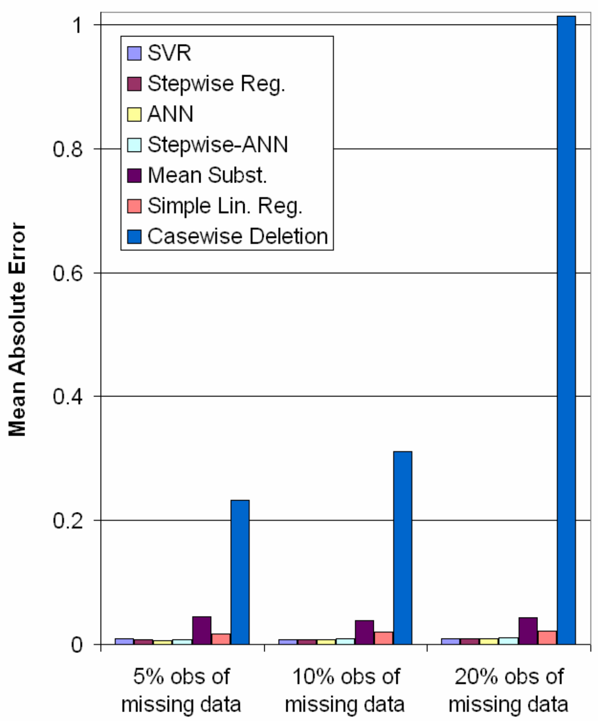

Figure 11. A bar chart illustrating the difference of variance between the original and imputed data sets using the temperature data set. The MAE for each method is reported.

6. CONCLUSIONS

Data sets of monthly total precipitation and monthly mean temperature are tested to determine the impact of removing these data on the mean and variance error. Elements are removed from these data matrices randomly in increments of 5, 10 and 20%. Casewise deletion, mean substitution, simple regression, and imputation, with stepwise linear regression, ANN and SVM, are tested to determine how well the techniques can reproduce the variance structure and estimate the missing values (except for the casewise deletion, which can not estimate the missing value).

Results of extensive experimentation with the aforementioned methods provide interesting findings and implications. By estimating the missing data and then comparing these estimates with the known values, the amount of signal that can be recovered is identified. In all experiments, the use of casewise deletion causes large errors in the variance of the estimates. The use of mean substitution leads to large errors in the mean and moderate errors in the variance estimates. Simple linear regression is a minor improvement over use of the mean. The lowest errors are found for the multiple imputation methods. Among the imputation techniques tested, SVM is ranked lowest in data error reported. Hence, more widespread use of this technique is warranted in situations when it is important to obtain accurate estimates of missing data.

ACKNOWLEDGEMENTS

Funding for this research was provided under the National Science Foundation Grant EIA-0205628 and NOAA Cooperative Agreement NA17RJ1227.

REFERENCES

Boser, B. E., Guyon, I. M., and Vapnik, V. N., 1992: A training algorithm for optimal margin classifiers. In D. Haussler, editor, *5th Annual ACM Workshop on COLT*, ACM Press, Pittsburgh, PA, 144-152.

Chang, C. and Lin, C., 2001: LIBSVM: a library for support vector machines. Software available at <http://www.csie.ntu.edu.tw/~cjlin/libsvm>.

Duffy, P.B., Doutriaux, C. Santer, B.D. and Fodor, I.K., 2001: Effect of missing data estimates of near-surface temperature change since 1900. *J. Climate*, **14**, 2809-2814.

Haykin, S., 1999: *Neural Networks: A Comprehensive Foundation*. 2nd Edition, Prentice Hall, New Jersey.

Kidson, J.W. and Trenberth, K.E., 1988: Effects of missing data on estimates of monthly mean general circulation statistics. *J. Climate*, **1**, 1261-1275.

Mann M. E., Rutherford, S. Wahl, E. and Ammann, C., 2005: Testing the fidelity of methods used in proxy-based reconstructions of past climate. *J. Climate*, **18**, 4097-4107.

Richman, M.B. and Lamb, P.J., 1985: Climatic pattern analysis of three- and seven-day rainfall in the central United States: some methodological considerations and a regionalization. *J. Climate and Applied Meteorology*, **24**, 1325 – 1343.

Roth, P.L., Campion, J.E. and Jones, S.D., 1996: The impact of four missing data techniques on validity estimates in human resource management. *J. of Business and Psychology*, **11**, 101-112.

Rubin, D.B. 1988: An Overview of Multiple Imputation. *Proceedings of the Survey Research Methods Section of the American Statistical Association*, 79-84

Rutherford, S., Mann, M.E., Osborn, T.J., Bradley, R.S., Briffa, K.R., Hughes, M.K. and Jones, P.D., Proxy-based Northern Hemisphere surface temperature reconstructions: sensitivity to method, predictor network, target solution and target domain. *J. Climate*, **18**, 2308-2329.

Schneider, T., 2001: Analysis of incomplete climate data: estimation of mean values and covariance matrices and imputation of missing values. *J. Climate*, **14**, 853-871.

Trafalis, T., Santosa, B., and Richman, M., 2003: Prediction of rainfall from WSR-88D radar using kernel-based methods", *International Journal of Smart Engineering System Design*, **5**, 429-438.

Vapnik, V. N., 1998: *Statistical Learning Theory*. Springer Verlag. New York.

Wayman, J.C. 2003: Multiple Imputation For Missing Data: What Is It And How Can I Use It? *Annual Meeting of the American Educational Research Association*, Chicago, IL.