ENVIRONMETRICS
Environmetrics 2006; **17**: 339–349
Published online 7 November 2005 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/env.773

Comparison of missing value imputation methods for crop yield data

Ravindra S. Lokupitiya[1]*,†, Erandathie Lokupitiya[2] and Keith Paustian[2]

[1]*Department of Atmospheric Science, Colorado State University, CO 80523, USA*
[2]*Department of Soil and Crop Sciences and Natural Resource Ecology Laboratory, Colorado State University, CO 80523, USA*

SUMMARY

Most ecological data sets contain missing values, a fact which can cause problems in the analysis and limit the utility of resulting inference. However, ecological data also tend to be spatially correlated, which can aid in estimating and imputing missing values. We compared four existing methods of estimating missing values: regression, kernel smoothing, universal kriging, and multiple imputation. Data on crop yields from the National Agricultural Statistical Survey (NASS) and the Census of Agriculture (Ag Census) were the basis for our analysis. Our goal was to find the best method to impute missing values in the NASS datasets. For this comparison, we selected the NASS data for barley crop yield in 1997 as our reference dataset. We found in this case that multiple imputation and regression were superior to methods based on spatial correlation. Universal kriging was found to be the third best method. Kernel smoothing seemed to perform very poorly. Copyright © 2005 John Wiley & Sons, Ltd.

KEY WORDS: missing values; regression; kriging; kernel smoothing; multiple imputation

1. INTRODUCTION

Missing values are common when working with large data sets. During the last few decades, researchers have applied several methods for imputing missing values, including various ad hoc methods as well as advanced model-based approaches. One of the most primitive techniques is to fill in the missing value with the mean of non-missing values. While this technique is simple and easy to apply, it causes underestimation of standard deviations and standard errors, since there is no variation in the imputed values. Also, it ignores correlations that often occur in spatially and temporally varying data.

We compared four methods—regression, universal kriging, kernel smoothing, and multiple imputation—which are commonly used by researchers. Data on crop yields from national agricultural

*Correspondence to: R. S. Lokupitiya, Department of Atmospheric Science, Colorado State University, CO, 80523, USA.
†E-mail: ravi@atmos.colostate.edu

Contract/grant sponsor: Consortium for Agricultural Mitigation of Greenhouse Gases (CASMGS); contract/grant number: 2001-38700-11092.

Received 28 February 2004
Accepted 15 September 2005

databases were used for the analysis. These data sets were spatially correlated and contain large numbers of missing values. We conclude with a discussion of the strengths and weaknesses of these four methods in the present setting.

2. DATA

Crop yields, aggregated at the county level, are reported in two main agricultural databases maintained by the US Department of Agriculture (USDA): the National Agricultural Statistical Survey (NASS) and the Census of Agriculture (Ag Census). The NASS crop yield data are produced annually using a statistical sampling approach and surveys done on selected farms within a county (Karkosh, 2002), whereas Ag Census crop yield estimates are produced every 5 years, based on a survey covering almost the total number of farms in a county. There are gaps (missing data) in both databases, but our aim was to fill the gaps in the NASS database, since it reports yields every year.

Several reasons exist for the presence of missing values in the NASS even when Ag Census has recorded data. A major source of missingness is that NASS only surveys those states that produce 90%–95% of the US total for each crop, whereas Ag Census tries to cover every acre for each crop. We also found that for certain states, NASS does not have county yield records because the state offices of NASS may not report certain crops to the national database. For example, NASS does not report county-level data for alfalfa hay in California, although California is known to be a state that produces significant quantities of alfalfa. In contrast, Ag Census does report county-level alfalfa hay information for California. Since this type of underreporting appears to be mainly due to policy-related considerations for some individual states, it does not represent a systematic bias with respect to crop yield potential. However, if such a state is far and away from the greatest crop yield potential, and if all the county level data is missing for the particular state, it might lead to a bias. Also NASS surveys use the location of the household as the location of the actual farm, and Ag Census uses the county in which the most of the income of the farmer is produced, again causing some discrepancies between the recorded crop yields. For this particular study, we selected barley crop yields for 1997, and only the counties that have yields recorded by both NASS and Ag Census were used for the purpose of comparison of imputation methodologies. County locations (i.e., 355) for NASS barley yields in 1997 are given in Figure 1.

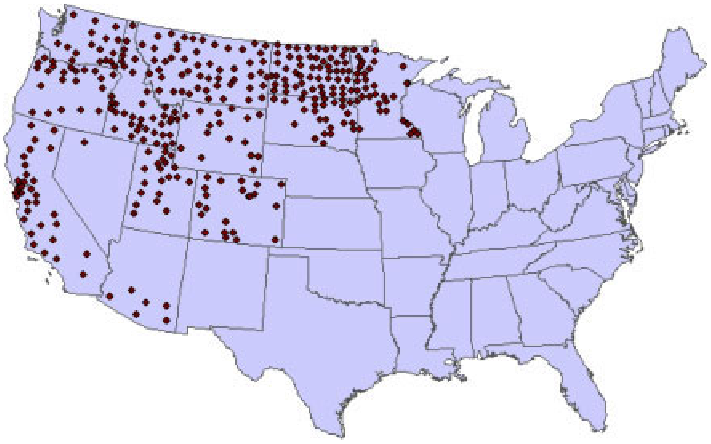

Figure 1. County locations for barley yields in the US in 1997

3. METHODS

We compared four methods for accounting for missing data: regression, kernel smoothing, universal kriging, and multiple imputation. A brief explanation of each approach is given below.

3.1. Regression

Regression analysis can be used to estimate the missing values when both NASS and Ag Census data sets are available. Variables in the Ag Census database can be regressed on the variables in the NASS database to obtain the model

$$\text{NASS} = \beta_0 + \beta_1 \cdot \text{Ag Census}$$

where β_0 and β_1 are intercept and slope parameters, respectively. Then this model is used to impute the missing NASS values when Ag Census data is available.

3.2. Kernel smoothing

Kernel smoothing is a model-based approach, which utilizes the spatial variation in data. Here a missing value is estimated as a weighted average of available data. The weights depend on the distances from location of the missing value to the locations of available data. Data locations closer to the missing location get more weight than the locations further away from the missing location. In this analysis, we have used a Gaussian distribution as the kernel.

If we are given a data set y_i with locations s_i, where $i = 1, 2, \ldots, n$, then the estimated missing value $\hat{y}$ at location s is given by

$$\hat{y} = \frac{\sum_{i=1}^{n} W(s - s_i) y_i}{\sum_{i=1}^{n} W(s - s_i)}$$

with

$$W(x) = \mathrm{e}^{-\frac{x^2}{2c^2}}$$

where c is a scaling parameter controlling the size of the neighborhood, which corresponds to the standard deviation of the Gaussian distribution. Here, we choose J neighboring points around location s and compute the distances from s to these points and parameter c is computed as twice the average of these neighboring distances. The choice of c depends on the choice of the neighborhood size (J), or vise versa. In this experiment, we have kept c constant while selecting J as the tuning parameter.

The problem of selecting J, the neighborhood size, is analogous to the problem of selecting the degree for a polynomial regression or selecting variables in multiple regression. Choosing a small neighborhood will produce an estimate that is close to the original data. Overfitting can produce a nearly unbiased estimate, but smoothing will create a large variance under repeated sampling. A large neighborhood will produce a very smooth estimate. A popular method for selecting neighborhood size is leave-one-out cross validation, or prediction sum of squares (Altman, 1992).

3.3. Universal kriging

Universal kriging is also a model-based approach, which accounts for the existing spatial correlation of the data. Suppose that $Z(s)$ is a real-valued Gaussian random field on R at location s with mean

$$E[Z(s)] = x(s)'\beta$$

where $x(s) = [x_1(s), x_2(s), \ldots, x_p(s)]'$ is a known $p \times 1$ vector valued function and β is a $p \times 1$ vector of unknown regression coefficients. The covariance function is represented by

$$\mathrm{Cov}[Z(s_i), Z(s_j)] = \alpha C(s_i, s_j) \quad \text{for all} \quad s_i, s_j \in \mathrm{R}$$

where $\alpha > 0$ is a scale parameter. Let $z = [z(s_1), z(s_2), \ldots, z(s_n)]'$ be a vector of observed values at locations $s_1, s_2, \ldots, s_n$. Then $E[z] = X\beta$, where $X = [x(s_1), x(s_2), \ldots, x(s_n)]'$ and $\mathrm{Cov}[z] = \alpha\{\mathrm{C}(s_i, s_j)\}_{\mathrm{n}\times\mathrm{n}}$.

If we need to predict a value $z(s)$ at the location s, then the kriging predictor, which is the best linear unbiased predictor that minimizes the variance of the prediction error, is given by

$$\hat{z}(s) = \{[x(s) - X'C^{-1}c(s)]'(X'C^{-1}X)^{-1}X'C^{-1} + c'(s)C^{-1}\}z$$

where $c(s) = C(s, s_i)$.

Spatial correlation of the data is examined on the basis of semivariogram analysis. The semivariogram is estimated by the equation

$$\hat{\gamma}(h) = \frac{1}{2|N(h)|}\sum_{N(h)} [z(s_i) - z(s_j)]^2$$

where $N(h)$ denotes the set of pairs of locations at distance h and $|N(h)|$ denotes the number of corresponding pairs of locations (Cressie, 1993). The estimated semivariogram can be modeled by various models such as the spherical, Gaussian, or exponential models (Cressie, 1993). For example, the exponential model with a nugget effect is given by

$$\gamma(h) = \begin{cases} 0, & \text{if } |h| = 0 \\ c_0 + c_1\left[1 - \mathrm{e}^{-\frac{|h|}{a}}\right], & \text{if } |h| > 0 \end{cases}$$

where c_0 is the *nugget effect*, which provides a discontinuity at the origin, a is related to the *range* (3a = range), which provides a distance beyond which the semivariogram value remains essentially constant, and $c_0 + c_1$ is called the *sill*, which is the semivariogram value for very large distances (Isaaks and Srivastava, 1989).

3.4. Multiple imputation

Multiple imputation (MI) was originally proposed by Rubin as a three-step process (Little and Rubin, 1987; Rubin, 1987, 1996). First, a set of plausible values is estimated for each missing value, which reflect uncertainty about the non-response model. By filling the missing values with these imputations, complete data sets are created. Second, each complete data set can be analyzed using standard statistical analyses. Finally, the results are combined such that the uncertainties of imputations have been taken into account (Horton and Lipsitz, 2001).

Two major assumptions are made regarding the data. First, it is assumed that the missingness is missing at random (MAR), that is, the probability that an observation is missing may depend on the observed values but not the missing values. Second, multivariate normality is assumed for the data.

A Markov Chain Monte Carlo (MCMC) method is used to impute the missing values. Mean vector and the covariance matrix for the data that do not have missing values are computed as starting values. These estimates are considered as the prior distribution. Filling missing values with the random numbers, which are drawn from the available distribution, creates a complete data set. The mean vector and covariance matrix are recomputed for the complete data set. This is the posterior distribution. Then the missing values are imputed again by generating random numbers from the posterior distribution. This procedure is iterated until the mean vector and covariance matrix are unchanging as we iterate. Imputations from the final iteration are taken to form a data set with no missing values.

SAS/STAT software, Version 8.2, introduces the experimental versions of MI and MIANALYZE procedures for imputing and analyzing the incomplete multivariate data sets (Horton and Lipsitz, 2001).

4. ANALYSIS

Cross-validation methods have been widely used for selecting the best models (Efron, 1982; Efron and Tibshirani, 1993; Hjorth, 1994; Libiseller and Grimvall, 2003; Shao, 1993; Shao and Tu, 1995; Stone, 1974; Wahba and Wald, 1975; Zhang, 1993). In traditional cross-validation methods, the data set is randomly divided into two halves, where the first half is used to fit the model. The model fitted to the first half is used to predict the second half. However this method has a disadvantage because it uses only half of the data for model fitting.

A better approach to overcome this deficiency is to use the ''leave-one-out'' cross validation, which uses all but one observation in each subsample (Efron, 1982; Hjorth, 1994; Libiseller and Grimvall, 2003). The omitted observation changes with each subsample so that every observation is held out exactly once. Each time the subsample is used to estimate the left-out observation and compare the estimated value with the left-out observation. For example, in the regression approach, the data consist of pairs $\{(X_1, Y_1), \ldots., (X_n, Y_n)\}$, where Y_i is the response variable and X_i is the predictor variable. A subsample $\{(X_1, Y_1), \ldots, (X_{i-1}, Y_{i-1}), (X_{i+1}, Y_{i+1}), \ldots, (X_n, Y_n)\}$ is created by deleting the ith data point (X_i, Y_i) for each $i = 1, 2, \ldots, n$. A model is fit to the subsample and used to re-estimate the deleted observation, say $\hat{y}_{(i)}(x_i)$. Then the deleted residuals can be computed as $r_{(i)} = [y_i - \hat{y}_{(i)}(x_i)]$. Best models were selected based on the mean absolute prediction error (MAPE), which was calculated as

$$\text{MAPE} = \frac{\sum_{i=1}^{n} |r_{(i)}|}{n}$$

Efron (1986) showed that the leave-one-out cross validation could produce unsatisfactory prediction errors due to overfitting and suggested that some form of bootstrap method would be more appropriate. Another problem with leave-one-out cross validation is lack of continuity—a small change in the data can make a large impact on the selected model. Geisser (1975) introduced multifold cross validation, where several $(k > 1)$ observations were deleted instead of a single observation. However this method is computationally intensive because there are ${}^{n}C_k$ possible subsets are involved. A nice application was given in Wahba and Wald (1975), who used k-fold cross-validation mean square errors to determine the correct degree of smoothing in fitting smoothing splines.

Breiman *et al.* (1984) introduced a less expensive deleting-k multifold cross validation method, where m mutually exclusive subsamples $s_1, s_2, \ldots, s_m$ were selected from the total data set $\{1, \ldots, n\}$

such that $n = k \times m$. Partition was done randomly to avoid possible biases. MAPE was computed for each model as

$$\text{MAPE} = \frac{1}{n}\sum_{i=1}^{m} |\boldsymbol{Y}_{s_i} - \hat{\boldsymbol{Y}}_{(-s_i)}|$$

where $\boldsymbol{Y}_{s_i}$ is a vector of observed values in subsample s_i and $\hat{\boldsymbol{Y}}_{(-s_i)}$ is the vector of predicted values computed after deleting the same subsample (Zhang, 1993).

The above four methods used for imputing missing data were compared using the leave-one-out cross validation procedure and deleting-5 multifold cross validation. Relative errors for each observation were computed using the formula

$$\text{Relative error} = \frac{\text{True value} - \text{Estimated value}}{\text{True value}}, \quad \text{for each method.}$$

If the true value is zero, the relative error is undefined. In this particular example, we have only selected the counties, which produce the given crop. Hence the true value cannot be zero.

We compared the correlation coefficients of the results of different imputation methods (Table 1). Regression and multiple imputation showed high correlations with the actual observations. Kriging showed a relatively high correlation (0.8094) with the true values whereas kernel smoothing showed the lowest correlation (0.4609). There is a very high correlation (0.9998) between the regression and the multiple imputation method. Hence it seems that both methods perform equally well in this particular application. This fact was further confirmed by the MAPE values given in Table 2.

Table 1. Correlations between results of different imputation methods (based on the leave-one-out cross validation)

	True value	Regression	Kernel smoothing	Universal kriging	Multiple imputation
True value	1.0000	0.9825	0.4609	0.8094	0.9823
Regression		1.0000	0.4206	0.7744	0.9998
Kernel smoothing			1.0000	0.5768	0.4205
Universal kriging				1.0000	0.7742
Multiple imputation					1.0000

Table 2. Mean absolute prediction errors for each method

Method	MAPE	
	Leave-one-out CV	Deleting-5 multifold CV
Regression	2.8272	2.8370
Multiple imputation	2.8003	3.0450
Universal kriging	8.8162	9.2356
Kernel smoothing	25.1922	25.5886

5. RESULTS AND DISCUSSION

All four methods were compared using barley crop yield data for year 1997, which consisted of both Ag Census and NASS data. Crop yield data from NASS were regressed against the data from Ag Census. According to the regression analysis, there exists a high correlation between the NASS and the Ag Census databases. Hence this correlation can be utilized in filling the missing values in the NASS database. The regression equation obtained for this data set was

$$\text{NASS} = 0.80753 + 1.00923 \text{ Ag Census}, \quad \text{with } R^2 = 0.97$$

According to the p-values of the intercept parameter (p-value $= 0.2017$) and slope parameter (p-value < 0.001), the intercept parameter was statistically insignificant and the estimated slope parameter was approximately equal to 1, which agrees with our prior knowledge about the data. Hence it is logical to force the regression line to go through the origin. The regression equation after forcing through the origin was found as

$$\text{NASS} = 1.02143 \text{ Ag Census}$$

Considering the spatial locations of the NASS data, we performed kernel smoothing and universal kriging. For kernel smoothing, according to the plot of neighborhood size (J) versus cross validation, we selected 150 as the suitable size for J, which corresponds to a relatively low cross validation value (see Figure 2).

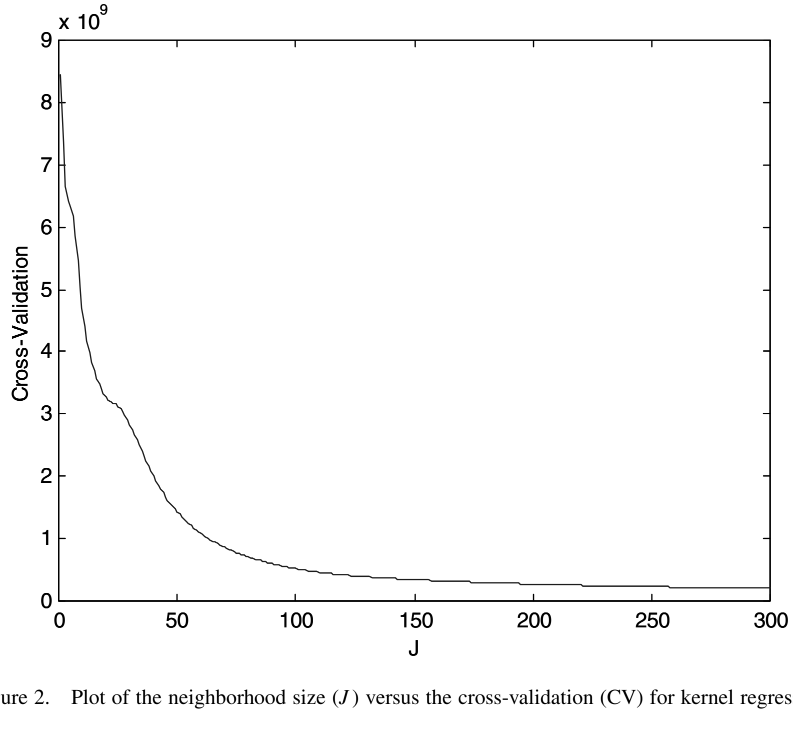

Figure 2. Plot of the neighborhood size (J) versus the cross-validation (CV) for kernel regression

In universal kriging analysis, we considered a third-order polynomial of northing and easting of the survey locations as the trend surface for the data. Here the distances were measured in 10-km increments. The assumption of stationarity and isotropy can be checked approximately, by observing the directional semivariograms for small distances. The semivariograms were estimated from the residual observations, $z - X\hat{\beta}$, with respect to the trend surface model. The directional semivariograms provided little evidence of anisotropy (see Figure 3). Hence, we assumed isotropy in this analysis. The isotropic semivariogram given in Figure 4 is fitted well by the exponential model with range parameter 39.29, sill 267.38, and nugget effect 86.05, according to the AIC values.

In multiple imputation, we considered a bivariate distribution of NASS and Ag Census data. For each iteration, a single observation of NASS data set was set to missing and imputed using the proc MI procedure. Five imputations were drawn for each missing value and the average of them was taken as the estimate for the missing point. Usually in multiple imputation, final results are combined using the MIANALYZE procedure. However, in this analysis, since we did not perform any analysis on data and our primary goal was to fill the missing values, the average of imputations were taken as the final estimate.

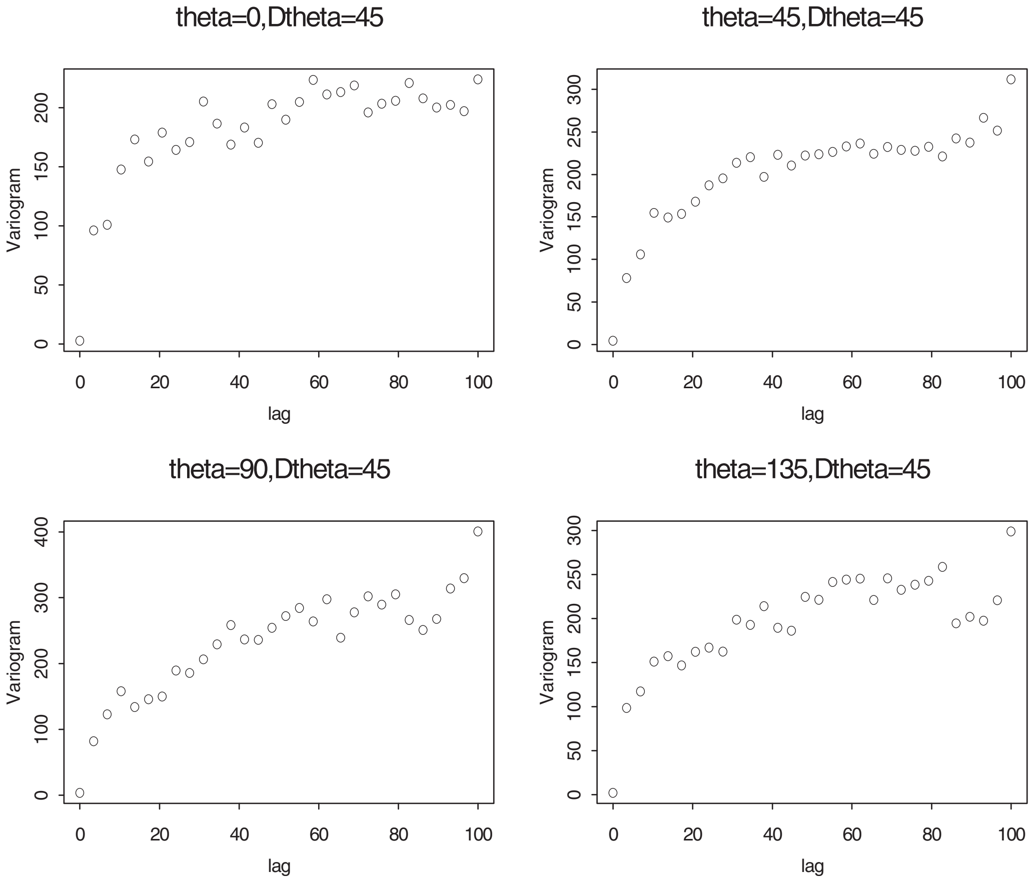

Figure 3. Plots of the directional variograms

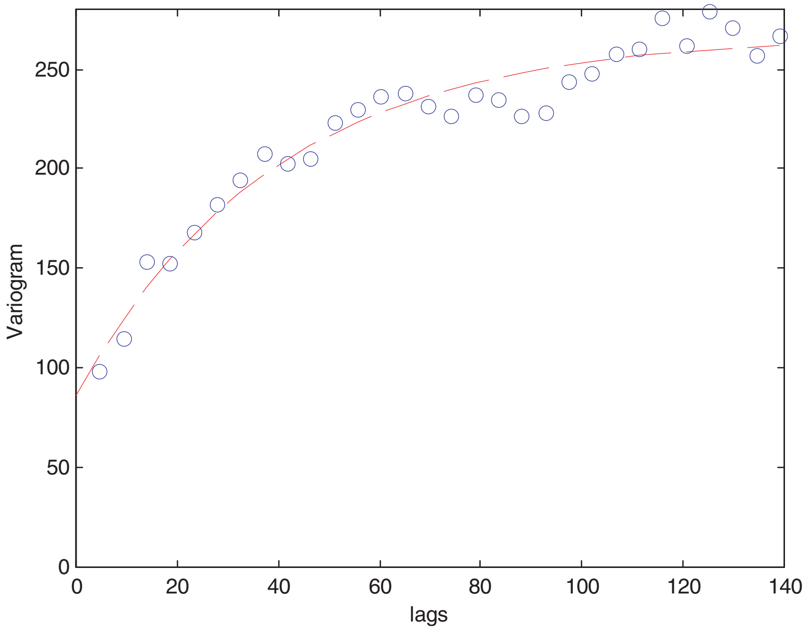

Figure 4. Plot of the isotropic variogram. The lags are expressed in 10 km

Each method was compared using the leave-one-out cross validation and deleting-*k* multifold cross validation methods. Based on the leave-one-out cross validation, we plotted observation number versus relative error for each method (Figure 5). Plots of the relative error for deleting-*k* multifold cross validation were similar (results not shown here). MAPE values computed using both cross validation methods are shown in the Table 2. The MAPE values for the multiple imputation and regression were very small and similar. Kernel smoothing gave the largest MAPE value and universal kriging was moderately acceptable. At some locations, kernel smoothing appears to have over estimated the missing values. Kernel smoothing is a distance-based method. Hence overestimates of missing values could occur when estimating a value that belongs to a state with a low crop yield that is surrounded by states with high yields.

Even though we assumed isotropy in universal kriging, the directional semivariograms show some evidence of anisotropy. Estimation can be improved by correcting for anisotropy. However, since several crops and several years of yields have been considered in this study, spatial covariance structure could be different for each case. Because universal kriging involves a great deal of cost in its implementation without significant benefit it was not judged to be practical to use as a missing value imputation tool in this case.

6. CONCLUSION

This study compared four major methods of estimating missing values. According to the MAPE values, regressing Ag Census on NASS and multiple imputation performed equally well in filling the missing values (Table 2). Based on our results, and the simplicity of the method, we suggest regressing Ag Census on NASS data, as the most appropriate method to fill missing values in NASS database (for

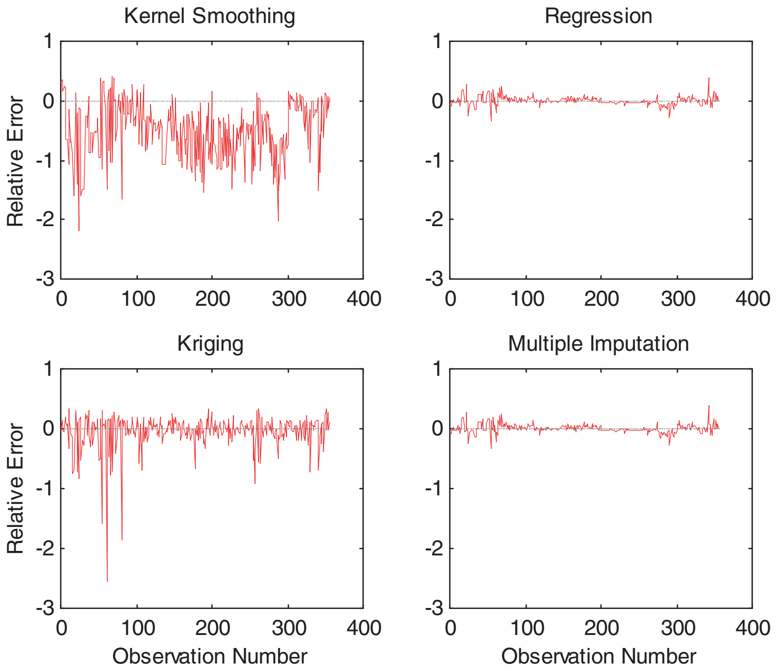

Figure 5. Plots of observation number versus the relative error for kernel smoothing, regression, universal kriging, and multiple imputation (based on leave-one-out cross validation)

the years when both NASS and Ag Census datasets are available). However, Ag Census data are produced only every 5 years. Hence, in the absence of Ag Census data, multiple imputation can be used to impute the missing values, by considering the multivariate normal distribution of yearly crop yields of NASS. Multiple imputation can be effectively done by lumping the years with high correlations. It is noted that most years tend to be highly correlated for the crop yield data sets. Universal kriging was found to be the third best method. One of the drawbacks of the universal kriging is that some observations have been overestimated by the method (Figure 5). Kernel smoothing performed very poorly among all four methods. Most of the observations were overestimated by kernel smoothing, and gave the largest MAPE value (Table 2 and Figure 5).

ACKNOWLEDGEMENTS

We thank Richard Anderson-Sprecher and anonymous reviewers for their valuable comments. Support for this work was provided by the Consortium for Agricultural Mitigation of Greenhouse Gases (CASMGS) through the Cooperative State Research, Education, and Extension Service, U.S. Department of Agriculture, under Agreement No. 2001-38700-11092.

REFERENCES

Altman NS. 1992. An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician* **46**: 175–185.

Breiman L, Friedman JH, Olshen RA, Stone C. 1984. *Classification and Regression Trees*. Wadsworth: Belmont, CA.

Cressie NAC. 1993. *Statistics for Spatial Data* (rev. ed.). Wiley: New York.

Efron B. 1982. The jackknife, the bootstrap, and other resampling plans. In *Regional Conference Series in Applied Mathematics No. 38*; Society for Industrial and Applied Mathematics. Philadelphia, Pennsylvania, 49–59.

Efron B. 1986. How biased is the apparent error rate of a prediction rule? *Journal of the American Statistical Association* **81**: 461–470.

Efron B, Tibshirani RJ. 1993. *An Introduction to the Bootstrap*. Chapman & Hall: New York; 237–257.

Geisser S. 1975. The predictive sample reuse method with applications. *Journal of the American Statistical Association* **70**: 320–328.

Hjorth JSU. 1994. *Computer Intensive Statistical Methods*. Chapman & Hall: New York; 24–56.

Horton NJ, Lipsitz SR. 2001. Multiple imputation in practice: comparison of software packages for regression models with missing variables. *American Statistical Association* **55**: 244–254.

Isaaks EH, Srivastava RM. 1989. *An Introduction to Applied Geostatistics*. Oxford University Press, Inc.: New York.

Karoush X. 2002. Personal communication from representative at USDA's National Agricultural Statistics Service.

Libiseller C, Grimvall A. 2003. Model selection for local and regional meteorological normalization of background concentrations of tropospheric ozone. *Atmospheric Environment* **37**: 3923–3931.

Little RJA, Rubin DB. 1987. *Statistical Analysis with Missing Data*. Wiley: New York.

Rubin DB. 1987. *Multiple Imputation for Nonresponse in Surveys*. Wiley: New York.

Rubin DB. 1996. Multiple imputation after 18 + years. *Journal of the American Statistical Association* **91**: 473–489.

Shao J. 1993. Linear model selection by cross-validation. *Journal of the American Statistical Association* **88**: 486–494.

Shao J, Tu D. 1995. *The Jackknife and Bootstrap*. Springer: New York; 306–311.

Stone M. 1974. Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society Series B* **36**: 111–147.

Wahba G, Wald S. 1975. A completely automatic French curve. *Communications in Statistics* **4**: 1–17.

Zhang P. 1993. Model selection via multifold cross validation. *The Annals of Statistics* **21**: 299–313.