J. Agr. Sci. Tech. (2018) Vol. 20: 829-839

Imputing Missing Values Using Support Variables with Application to Barley Grain Yield

Y. Tandoğdu[1*], and M. Erbilen[1]

ABSTRACT

Missing values in a data set is a widely investigated problem. In this study, we propose the use of support variables that are closely associated with the variable of interest for the imputation of missing values. Level of association or relationship between the variable of interest and support variables is determined before they are included in the imputation process. In this study, the barley (*Hordeum vulgare*) grain yield in the semi-arid conditions of Cyprus was used as a case study. Monthly rain, monthly average temperature, and soil organic matter ratio were selected as support variables to be used. Multivariate regression employing support variables, bivariate, kernel regression and Markov Chain Monte Carlo techniques were employed for the imputation of missing values. Obtained results indicated a better performance using multivariate regression with support variables, compared with those obtained from other methods.

Keywords: Imputing missing data, Incomplete data, Rain equivalent grain yield, Regression techniques.

INTRODUCTION

Despite various safeguards, missing values are frequently encountered in many fields during data collection. Therefore, an increasing interest amongst researchers to the topic has led to the development of different methodologies contributing to more accurate imputation of missing values. Edgett (1956) considered the estimation of population parameters using multiple regression, when missing observations exist among the independent variables. Anderson (1957) used the Maximum Likelihood Estimates (MLE) for a multivariate normal distribution in the presence of missing data and proposed an approach that minimizes mathematical manipulation. Trawinski and Bargmann (1964) explored the MLE of population parameters in the multivariate case with incomplete data. Afifi and Elashoff (1966) in their study attempted to show how the estimation of missing data is simplified, when the missing data follows a certain pattern. Rubbin (1976) examined the case of making direct likelihood or Bayesian inferences about the population parameter θ, ignoring the process that caused the missing data. In his work, Little (1992) compared six classes of procedures used in the imputation process of missing values, with a view to Bayesian simulation methods. Copt and Feser (2003) presented a general class of estimators adapted to the case of missing data. Toutenburg *et al.* (2005) offered some modification to the linear regression model when missing observations exist in the independent variables. Zhang *et al.* (2008) proposed a sequential local least squares imputation approach to deal with missing values in the gene microarray data. Qin *et al.* (2009) proposed a unified empirical likelihood

[1] Department of Mathematics, Eastern Mediterranean University, Mağusa, North Cyprus.
[*]Corresponding author; e-mail: yucel.tandogdu@emu.edu.tr

approach to missing data problems. Robbins *et al.* (2013) proposed a variable transformation using marginal density model to be used in imputation of missing data. Yozgatlıgil *et al.* (2013) examined six methods for the imputation of missing values of spatio–temporal meteorological time series data. Jinubala and Lawrance (2016) used predictive mean matching method for identifying and replacing missing values for crop pest data.

Since Cyprus has a semi-arid climate, a review of research related with this topic from agricultural science mostly pertaining to semi-arid regions is undertaken to determine the relationship between the Support Variables (SV) and the dependent variable (barley grain yield). Cantero-Martinez *et al.* (1995) studied two contrasting barley cultivars for grain yield in Ebro valley of northern Spain where Mediterranean climate is dominant. Samarah (2005) compared barley grain yield under well irrigated conditions at 100% field capacity, mildly stressed at 60% field capacity, and severely stressed at 20% field capacity conditions for semi-arid climate in Jordan. In a recent study, Ebrahimian and Playan (2014) focused on efficiency and uniformity of water and fertilizer application in the management of irrigation and fertigation systems. Lopez and Arrue (2005) studied the efficiency of different tillage methods in terms of barley and wheat production in the semi–arid climate of Aragon in NE Spain. Quiroga *et al.* (2005) investigated the relationship between barley grain yield and soil organic matter (g kg^{-1}) ratio to clay+silt content (g kg^{-1}), in the semi-arid to arid regions of Pampas in Argentina. Experimentally, they observed that $R^2 = 0.51$ or 51% of barley grain yield can be explained by this coefficient of determination obtained through the linear regression of barley grain yield on soil organic matter ratio. Nahar *et al.* (2010) observed the phonological variation and its relation with grain yield in Bangladesh and determined that there was an average of 62% loss in yield due to heat stress. Johnston (2011) explored the role of soil organic matter in crop production in relation to various Nitrate (N) and potassium (K) application levels. Hossain *et al.* (2012) examined the optimum sowing time for two genotypes of barley and wheat under two drought stress conditions in an arid region in Russia. Adekanmbi and Olugbarab (2015) focused on a constrained multi-objective optimization of mixed cropping pattern, by maximizing profit and crop production in minimized planting area. However, in many such studies the possibility of missing data is not catered for.

In this study, we aimed to develop a proposal for the imputation of missing values, investigating missing at random, by application of the Multivariate Linear Regression employing Support Variables (MRSV) concept to a data set related to barley grain yield.

MATERIALS AND METHODS

Here, utilization of variables that are part of the process and highly correlated with the variable of interest with missing data is explored. These variables are named as Support Variables (SV). Positive contribution of SVs to the imputation process became evident based on the proposed methodology. When the measurement units of the SVs are different, a suitable way of converting them to the same unit as the variable with missing data is necessary. Multivariate Linear Regression Employing Support Variables (MRSV), bivariate linear regression, kernel regression, and Markov Chain Monte Carlo (MCMC) methods using available data only to predict missing values are used and results compared.

Application of the MRSV concept to a data set related with barley grain yield, produced more accurate and robust imputation results with low errors levels compared with bivariate linear regression, kernel regression, and MCMC methods. Data set used is barley grain yield in tons per hectare over 17 years from 17

production areas in Northern part of the island of Cyprus with no missing values. Average monthly rain and monthly temperature for the months from sowing to harvest, and soil organic matter ratio that have significant influence on yield are taken as SVs.

Use of Support Variables in the Imputation Process

Statistical analysis of homogenous multivariate and complete data tends to reflect the message inherent to the process where data comes from. Absence of some data will inevitably lead to certain misinterpretations following the statistical analysis of such data. The pattern of missing values provide some guidance in the selection of the imputation method. Imputation of missing values improves the robustness of the results to be obtained via statistical analysis. In this study, an attempt is made to impute missing values of a certain variable, by utilizing existing data belonging to this variable, and closely related SVs. Interpretation of the results of analysis will be prone to error if SVs have different units. Conversion of units requires careful study of the process that relates the dependent and SVs used as independent variables.

For 17 farming areas in North Cyprus, barley grain yield (X) (t ha^{-1}) over a period of 17 years, with no missing values, was used as a data set. Missing values were artificially created from this data by deleting values at random, forming two new data sets, one having 10%, and the other having 40% missing values. Monthly average rain X_1 (mm/m^2), monthly average temperature X_2 ($^{\circ}C$), and soil organic matter ratio X_3 were used as SVs. Relevant research from agronomy was taken into account for the conversion of the units of SVs to that of the dependent variable. Linear correlation coefficient between the dependent variable X and the SVs were computed as $\rho_{XX_1} = 0.55,\ \rho_{XX_2} = 0.94,\ \rho_{XX_3} = 0.74$, indicating that the use of the support variables in the estimation process will be beneficial.

Rain Equivalent Grain Yield

Rain is one of the main parameters that influence the grain yield. Here, an attempt was made to establish a relationship between the water used from germination to harvest, based on research undertaken by various researchers, to enable the conversion of average rain data into equivalent grain yield.

Water use efficiency was defined by Cantero *et al.* (1995) as grain yield produced per unit area per unit of water evapotranspired by the crop (kg ha^{-1} mm^{-1}). Evapotranspired water is the amount of water used by the plant for the period from germination to harvest. We assumed conventional tillage cereal–fallow rotation, as it is the practice supported by state subsidies. Here, the aim was to convert rain data (mm m^{-2}) of October to April i.e. the period from germination to harvest, to grain yield in t ha^{-1}. Water Use Efficiency for grain (WUE_g) given by Cantero *et al.* (1995) and Lopez (2005) was considered to be the most representative of local conditions, with an average value of WUE_g= 8 kg ha^{-1} mm^{-1}. Average evapotranspiration for north Cyprus, during the months November to April, is given as 188 *mm* by Tandoğdu and Camgöz (1999). This corresponds to 55% of Annual Average Precipitation (AAP) for the study area $(E_{tr} = 0.55)$. Then, the Rain Equivalent Yield (REY) in tons per hectare was computed by $REY = (AAP \times E_{tr} \times WUE_g)/1000$.

Temperature Equivalent Grain Yield

Heat is another important stress factor influencing the grain yield. During the grain filling period temperatures above the long term monthly average will adversely affect the grain yield. Referring to the studies of

Nahar *et al.* (2010) and Hossain *et al.* (2012), and given the local semi-arid climatic conditions, it is anticipated that for every $1^{\circ}C$ increase in the long term monthly average temperature, an average of 3 to 6% loss in grain yield will occur. Let $x_i;\ i=1,\ldots,n$ be the annual grain yield (t ha^{-1}), n the number of production years, $\overline{t}_i$ the average temperature for the grain maturing period (March, April) for the i^{th} year, $\overline{t}$ the long term average temperature for the periods from germination to harvest (October to April) for all n years. Minor fall of average temperature in the order of several $^{\circ}C$ during the maturing period to below the average of germination to harvest period will not have a major effect on the grain yield. However, an average maturing period temperature above the long term average will have adverse effect on the grain yield. Therefore, the temperature equivalent grain yield for a certain year, is proposed to be the grain yield for that year multiplied by the factor $\overline{t}/\overline{t}_i$. If $\overline{t}/\overline{t}_i<1$, means a higher than long term average temperature for the maturing period, resulting in a loss in grain yield. If $\overline{t}/\overline{t}_i\geq 1$, it is assumed that grain yield will not be affected.

Soil Organic Matter Equivalent Grain Yield

Soil Organic Matter Ratio (SOMR) is another important factor taken as a support variable. In their study in the semi-arid Pampa region in Argentina with varying soil organic matter to clay and silt ratio, Quiroga *et al.* (2005) determined that 51% of grain yield can be attributed to *SOMR*. Johnston (2011) presented results from experiments carried out with spring and winter barley as a function of Nitrogen (N) applied and the level of organic matter. It is observed that with no N application, the spring and winter barley combined with low and high organic matter ratio shows an average grain yield of 2.78 t ha^{-1}. Combined low and high organic matter ratio average is 2.35%. According to Stine and Weil (2002), there is a linear relationship between soil organic matter ratio and grain yield. Then, for 1% soil organic matter ratio there corresponds a 1.18 t ha^{-1}grain yield, leading to the conclusion that for any soil organic matter ratio *p*, the corresponding grain yield becomes 1.18 *p*. Derici *et al.* (2000) reports that the sub areas under study are considered homogenous in terms of *SOMR*. For each sub area, it is assumed that *SOMR* values are constant for the period under study. Due to the nature of the factors affecting *SOMR*, wild fluctuations are not expected in a matter of few years or a few decades.

Imputing Missing Values Using Regression

Missing values encountered in every field of endeavour have detrimental and adverse effects on statistical estimation. A review of research across the spectrum contributed to the development of the current proposal.

Data set represented by a $n\times p$ matrix **X**, n being the number of rows or observations and p the number of variables involved in the process of concern. In the example used in this study, rows represent production areas, and columns represent years. Proposed methodology to impute the missing values either row by row or column by column; include the utilization of data from SVs and the variable with missing values.

Theoretically, regression is the expected value of a response variable *Y* conditioned on some regressor variables subject to the existence of a joint probability distribution $f(y,x_i);\ i=1,\ldots,p$, and $\mu_Y=E(y|X_1=x_1,\ldots,X_p=x_p)$.

Statistically, each X_i is a vector of n data values observed for the i^{th} variable.

Estimation of the regression relation from available data is possible using various regression techniques.

In this study, MRSV with back substitution of the estimated values, Kernel regression using Epanechnikov and Gaussian kernel functions, and MCMC methods are used in estimating the missing values.

Methodology to be followed in the imputation of missing values is independent of the application field. Data used in this forms the $17' 17$ complete data matrix **X**. Let the $17' 17$ matrix with 10% missing values be $\mathbf{X}_1^m$, and the one with 40% missing values be $\mathbf{X}_2^m$. For the 10% missing case 5 splits were used to test the validity of MRSV. Following the imputation Mean Square Error (MSE) and robustness for each method are computed and used as measures to determine the performance of each method.

Imputation Using Multivariate Regression

Given *p* response variables Y_i depending on *k* predictor variables $X_1,\ldots,X_k$, the multivariate regression for the i^{th} variable is $y_i = b_0 + b_1x_{1i} + \cdots + b_kx_{ki} + e_i$, where $b_i;\ i = 0,\ldots,k$ are regression constants and $e_i;\ i = 1,\ldots,p$ are the random error terms assumed to be identically distributed with mean zero and variance σ^2. Minimizing Σe_i^2 yields the set of equations enabling the computation of constants $b_0,\ b_1,\cdots,\ b_k$.

Elements of the matrices $\mathbf{X}_1^m$ and $\mathbf{X}_2^m$ with an existing value are denoted by $x_{ij}: i = 1,\ldots,n \quad j = 1,\ldots,p$ and the elements with missing values are denoted by $x_{ij}^m: i = 1,\ldots,n \quad j = 1,\ldots,p$. In both cases, the indices *i* and *j* have the same range, meaning observed values are x, missing values are x^m, and their locations in a matrix are defined by *i* and *j*. Missing values can be estimated either on column by column or row by row basis. It is wise to start with the row or column with least number of missing values. In this study each estimated missing value is included in the computations for the estimation of the subsequent missing values. Imputation of column wise missing values is preferred, since the selected support variables, especially rain and temperature, are susceptible to annual climatic changes.

Steps in column by column imputation process using regression are as follows:

1. Determine the l^{th} column ($1 \le l \le p$) with minimum number of missing values (k_l).
2. Let the missing values of the l^{th} column be $x_i^m;\ i = 1,\ldots,k_l$ where $k_l < n$.
3. Compute the $n-k_l$ row averages ($\bar{x}_j^m;\ j = 1,\ldots,n-k_l$) for the rows corresponding to existing data in column *l*, but excluding the values of the l^{th} column. These averages are elements of the first independent random variable X_{1a}. Let the l^{th} column be the dependent variable X_l. Form the $n-k_l$ tuples between the existing values of the dependent variable X_l and the corresponding values of the independent variable $X_{1a} = \{\bar{x}_1^m,\ldots\bar{x}_{n-k_l}^m\}$.
4. Similarly, define the variables $X_{2a}, X_{3a}, X_{4a}, X_5, X_6$ to have the same

number of rows ($n-k_l$) as the dependent variable X_l, as below.

4.1. X_{2a}: Average of the $n-k_l$ rows of rain equivalent data excluding the l^{th} column.

4.2. X_{3a}: Average of the $n-k_l$ rows temperature equivalent data excluding the l^{th} column.

4.3. X_{4a}: Average of the $n-k_l$ rows soil organic matter equivalent data excluding the l^{th} column.

4.4. X_5: Rain equivalent data of the l^{th} column.

4.5. X_6: Temperature equivalent data of the l^{th} column.

X_5 and X_6 are specifically included in the regression process as they represent available data for the year when imputation is undertaken.

5. Carry out multivariate regression of column l

$$x_l = b_0 + b_1x_{1a} + b_2x_{2a} + b_3x_{3a} + b_4x_{4a} + b_5x_5 + b_6x_6$$

.

6. Estimate the first missing value and substitute in the l^{th} column, leaving $k_l - 1$ missing values in the l^{th} column.

7. Repeat the process until all missing values are imputed in the l^{th} column.

8. Repeat steps 1 to 7, till all missing values in all columns are imputed.

Mean Square Error (MSE), and Mean Absolute Error (MAE) committed during the imputation process were computed and expressed as a percentage of averages from the complete data for easy comparison (MSE%, MAE%). For the 10% missing data, estimation errors were very low. An overall average *MSE*% obtained from 5 splits applied to 10% missing case using MRSV was 1.24% supporting the validity of MRSV. Subsequently, work concentrated on the 40% missing case to assess the performance of the proposed imputation method MRSV.

For the 40% missing case, MRSV performed much better than the bivariate regression. As seen in Table 1, *MSE*% is significantly less in MRSV compared with bivariate case at 0.05 significance level using the t test. Similarly, *MAE*% are 11 and 29% for MRSV and bivariate regression, respectively, favouring MRSV.

Performance of MRSV and bivariate regression methods are also tested against the use of kernel regression in imputation.

Kernel Regression in the Imputation Process

Kernel estimator employs local weights in the estimation process. An estimate at a certain point is the linear combination of neighbouring observations. The Nadaraya–Watson estimator of the regression function is $\hat{g}(x) = \Sigma_{i=1}^{n} w_i k(u) / \Sigma_{i=1}^{n} k(u)$, where w_i represents local weights and k is a kernel function. Here, $u = (x_i - X)/h$, h being the band width. If an observation at a given point is x_i and its estimated value $\hat{x}_i$, then the weights w_i are determined such that $SSE = \Sigma(x_i - \hat{x}_i)^2$ is minimum. Size of the band width h is important since it determines the level of smoothing. It is also necessary to compute the kernel values within the neighbourhood of X. Neighbouring values

Table 1. *MSE*% values obtained in different methods.

	MRSV	Bivariate regression	Epanechnikov kernel	Gaussian kernel	MCMC
MSE%	3	16	16	30	27

x_i should be at equal distance increments (*dx*) on the left and right of the point *X*. Hence, a large *h* value will result in over smoothing, shadowing the underlying structure represented by the data, while a too small *h* will not give any desired smoothing effectively, meaning no idea about the underlying function representing the variable or process under study. Hence, determination of the optimum size of bandwidth is crucial. For the Gaussian kernel function, the following is suggested by Silverman (1998) $h = \left(\frac{4s^5}{3n} \right)^{1/5}$. In general, determination of the size of *h* is via the cross validation method of Chiu (1991), Silverman (1998), Ramsey, J. O. and Silverman, (2006), and Haerdle (2004), which is an iteration method to determine the *h* value that minimize the average error.

Two commonly used kernel functions are employed in this study: Epanechnikov kernel function: $k(u) = 0.75(1-u^2)\ ;\ |u| \le 1$, and Gaussian kernel function: $k(u) = (2\pi)^{-1/2} \exp(-\mathrm{u}^2/2)$.

Using *h* = 2, 4, and 6, imputed values were computed for 10 and 40% missing cases using Gaussian and Epanechnikov kernel regression. For each case, *MSE*% values are determined. These are summarized in Table 2, from where it is evident that bandwidth *h*= 2 results in the lowest *MSE*% values for 10 and 40% missing data cases, with *MSE*% values being lower for the 10% missing data case. Also, Epanechnikov kernel performs better than Gauss kernel in terms of *MSE*% values.

However, results obtained from kernel regression are not better than those obtained from MRSV. See Table 1.

Imputation with Markov Chain Monte Carlo (MCMC) Method

MCMC is based on an iterative simulation concept for imputation. Parameters are estimated from conditioning information.

Let data set *x* with missing values be denoted as x^m, observed values as x^o, and θ set of parameter values to be estimated. Then, the posterior distribution of the parameter values conditioned on the observed data is given by Schunk (2008) as:

$$f(\theta|x^0) = \int_{X^m} f(\theta|x^0, x^m)\, f(x^m|x^0)\, dx^m$$

Where, $f(\theta|x^o, x^m)$ is the conditional density of θ conditioned on the complete data *X*, and $f(x^m|x^o)$ is the predictive density of missing values conditioned on observed data.

Tanner (1987) proposed an iterative approach for use in the imputation of missing values and updating the distribution parameter θ taking into account the imputed values together with the observed ones. Initial approximation of the posterior values to be imputed are estimated from observed data by a suitable algorithm.

The following algorithmic approach was used in imputing missing values.

i. A suitable *dx* value to be used in increasing or decreasing an observed data value x^o_{i-1} neighboring a cell x^m_i with missing value till the other neighboring value x^o_{i+1} is reached.

Table 2. A summary of average *MSE*% values for various bandwidths with *dx=2*.

10% Missing at random			40% Missing at random	
h	Epanechnikov	Gauss	Epanechnikov	Gauss
2	7.2	15.7	16.1	30
4	17.6	25.2	35.8	26.5
6	20.8	30.2	40.9	42.2

ii. For a row or column, all missing values are subjected to this process. Since the difference between two cells neighboring a x_i^m cell will not be the same for all x_i^m cells, the process will stop at different iterations for different x_i^m cells.
iii. Following the completion of all iterations in a row or column, the average of this row or column is computed and compared with the average of the corresponding row or column from the complete data. Imputed row or column with average closest to the average of the complete data is selected.

The process given in steps i to iii was repeated to complete the imputation process.

Algorithm was applied to data set with 40% missing values. Obtained *MSE*% results indicate a better performance for the MRSV as seen in Figure 1.

The Relative Aitchison Distance (RDA) criteria defined as a measure of robustness is also used for the comparison of the accuracy of estimates. This is defined by Templ *et al.* (2009) as $RDA = \frac{1}{n_M} \sum_{i \in M} d_A(x_i, \hat{x}_i)$, where $M \subseteq \{1,\ldots,n\}$, n_M is the number of cells with missing values in a variable, $d_A(x_i, \hat{x}_i)$ is the Aitchison distance. Magnitude of *RDA* is used as an indicator of the quality or statistical robustness of estimation.

RDA values were computed for all methods used in the imputation process for the 40% missing case and results are given in Figure 2. MRSV still exhibiting the best performance.

DISCUSSION

Comparison of Multivariate Bivariate Kernel Regression and MCMC Results

Imputation of missing values using the proposed MRSV method is compared with the results obtained from 4 other known methods. Based on *MSE*% criteria, it was observed that the 10% missing case for 5 splits produced estimates fairly close to the true values in all methods with low error levels. Hence, 40% missing case was taken as a basis for comparison of the results obtained.

Using the *t* test for the difference between the average *MSE*% of MRSV and other methods has shown strong significance with a *P* value less than 0.001 in favour of MRSV, indicating the better performance of the proposed method. Table 1 summarizes, Figures 1 and 2 signify this fact.

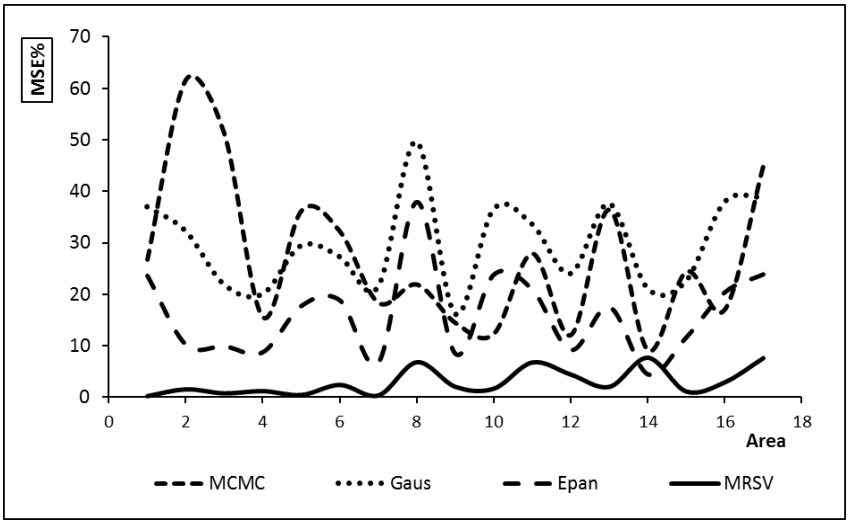

Figure 1. *MSE* as a percentage of row averages for different imputation methods with 40% missing values.

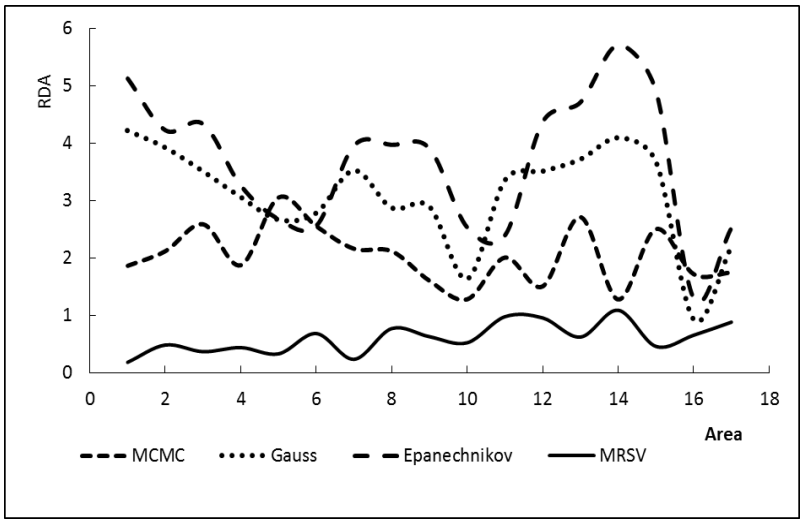

Figure 2. *RDA* measures for different imputation methods with 40% missing data.

In view of this study for the imputation of missing values when data for a certain variable are missing at random, the following conclusions are drawn.

i. Careful determination of the relatedness of the support variables to the variable with missing values is very important. Further, if the units of the SVs are not the same as that of the dependent variable, it is vital to convert them to that of the dependent variable.
ii. MRSV method is recommended as it produced better results compared with other studied imputation methods included in this study.
iii. MRSV can be used with fair confidence when up to 40% of data is missing at random.

Further work is needed to assess the performance of the MRSV method when missing values is higher than 40%.

REFERENCES

1. Adekanmbi, O. and Olugbarab, O. 2015. Multiobjective Optimization of Crop-Mix Planning Using Generalized Differential Evolution Algorithm. *J. Agr. Sci. Tech.*, **17**: 1103–1114.
2. Afifi, A. A. and Elashoff, R. M. 1966. Missing Observations in Multivariate Statistics. I: Review of the Literature. *J. Amer. Stat. Assoc.*, **61**: 595-604.
3. Anderson, T. W. 1957. Maximum Likelihood Estimates for a Multivariate Normal Distribution When some Observations Are Missing. *J. Amer. Stat. Assoc.*, **52**: 200-203.
4. Cantero-Martinez, C., Villar, J. M., Romagosa, I. and Fereres, E. 1995. Growth and Yield Responses of Two Contrasting Barley Cultivars in a Mediterranean Environment. *Eur. J. Agron.*, **4(3):** 317-326.
5. Chiu, S. T. 1991. Bandwidth Selection for Kernel Density Estimation. *Annal. Stat.*, **19(4):** 1883 – 1905.
6. Copt, S. and Feser, M. V. 2003. *Fast Algorithms for Computing High Breakdown Covariance Matrices with Missing Data.* Report No 2003.04. Cahiers du Département d'Econométrie Faculté des Sciences Economiques et Sociales Université de Genève.
7. Derici, M. R., Kapur, S.A., Kaya, Z., Gök, M. and Ortas, İ. 2000. *Kuzey Kıbrıs Türk Cumhuriyeti Detaylı Toprak Etüd ve Haritalama Projesi.* North Cyprus State Printing House.
8. Ebrahimian, H. and Playan, E. 2014. Optimum Management of Furrow Fertigation to Maximize Water and Fertilizer Application Efficiency and

Uniformity. *J. Ag. Sci. Tech.,* **16**: 591 – 607.

9. Edgett, G. L. 1956. Multiple Regression with Missing Observations among the Independent Variables. *J. Amer. Stat. Ass.,* **51**: 122-131.
10. Haerdle, W. 2004. *Applied Nonparametric Regression. Economic Society Monographs.* Institut für Statistik und Ökonometrie, Wirtschaftswissenschaftliche Fakultat, Humboldt-Universitat zu Berlin, Spandauer Str. 1, D-10178 Berlin.
11. Hossain, A., Teixeira da Silvia, J. A., Lozovskaya, M. V., Zvolinsky, V. P. and Mukhortov, V. I. 2012. High Temperature Combined with Drought Affect Rainfed Spring Wheat and Barley in South-Eastern Russia: Yield, Relative Performance and Heat Susceptibility Index. *J. Plant Breed. Crop Sci.,* **4(11):** 184 -196.
12. Jinubala, V. and Lawrance, R. 2016. Analysis of Missing Data and Imputation on Agriculture Data Using Predictive Mean Matching Method. *Int. J. Sci. App. Info. Tech.,* **5(1):** 1-4.
13. Johnston, J. 2011. The Essential Role of Soil Organic Matter in Crop Production and the Efficient Use of Nitrogen and Phosphorus. Better Crops with Plant Food. *Int. Plant Nutr. Inst.,* **95(4):** 9 -11.
14. Little, J. A. 1992. Regression with Missing X's: A Review. *J. Amer. Stat. Assoc.,* **87(420):** 1227 – 1237.
15. Lopez, M. V. and Arrue, J. L. 2005. *Growth, Yield and Water Use Efficiency of Winter Barley in Response to Conservation Tillage in Semi-Arid Region of Spain*. Spanish National Research Council, Departamento de Edafología, Estación Experimental de Aula Dei, Consejo Superior de Investigaciones Científicas (CSIC), POB 202, 50080-Zaragoza (Spain)
16. Nahar, K., Ahamed, K. U. and Fujita, M. 2010. Phenological Variation and Its Relation with Yield in Several Wheat (*Triticum aestivum* L.) Cultivars under Normal and Heat Stress Condition. *Notulae Scientia Biologicae,* **2(3):** 51 - 56.
17. Qin, J., Zhang B. and Leung H. Y. 2009. Empirical Likelihood in Missing Data Problems. *J. Amer. Stat. Assoc.*, **104(488):** 1492 – 1502.
18. Quiroga, A., Funaro, D., Noellemeyer, E. and Peinemann, N. 2005. Barley Yield Response to Soli Organic Matter and Texture in the Pampas of Argentina. *Soil Till. Res.,* **90**: 63-68.
19. Ramsey, J. O. and Silverman, B. W. 2006. *Functional Data Analysis*. 2nd Edition, Springer.
20. Rubbin, D. B. 1976. Inference and Missing Data. *Biometrica,* **63**: 581 – 592.
21. Robbins, M. W., Ghosh S. K. and Habiger J. D. 2013. Imputation in High Dimensional Economic Data as Applied to the Agricultural Resource Management Survey. *J. Amer. Stat. Assoc.,* **108(501):** 81 – 95.
22. Samarah, N. H. 2005. Effects of Drought Stress on Growth and Yield of Barley. *Agro. fSust. Dev.,* **25**: 145-149.
23. Schunk, D. 2008. A Markov Chain Monte Carlo Algorithm for Multiple Imputation in Large Surveys. American Statistical Association.. *Adv. Stat. Analysis,* **92**: 101 – 114.
24. Silverman, B. W. 1998. *Density Estimation for Statistics and Data Analysis.* Monographs on Statistics and Applied Probability, Chapman & Hall, Kernel Regression in the Imputation process.
25. Stine, M. A. and Weil, R. R. 2002. The Relationship between Soil Quality and Crop Productivity across Three Tillage Systems in South Central Honduras. *Am. J. Alter. Agri.,* **17 (**1): 2 – 8.
26. Tandoğdu, Y. and Camgöz, T. O. 1999. An Experimental Approach for Estimating Evapotranspiration. *CIM Bull*., **92**: 55-60.
27. Tanner, M. A. and Wong, W. H. 1987. The Calculation of Posterior Distributions by Data Augmentation. *J. Amer. Stat. Assoc.,* **82(398):** 528 – 540.
28. Templ, M., Filzmoser, P. and Horn, K. 2009. Robust Imputation of Missing Values in Compositional Data Using the R Package. http://cran.salud.gob.sv/web/packages/robCompositions/vignettes/imputation.pdf
29. Toutenburg, H., Srivastava, V. K., Shalabh, and Heumann, C. 2005. Estimation of Parameters in Multiple Regression with Missing Covariates Using a Modified First Order Regression Procedure. *Annal. Econ. Fin.,* **6:** 289-301.
30. Trawinski, I. M. and Bargmann, R. E. 1964. Maximum Likelihood Estimation with Incomplete Multivariate Data. *Annal. Math. Stat.,* **35:** 647-657

31. Yozgatlıgil, C., Aslan, S., Iyigun, C. and Batmaz, I. 2013. Comparison of Missing Value Imputation Methods in Time Series: The Case of Turkish Meteorological Data. *Theor. App. Clim.*, **112**: 143–167.

32. Zhang, X., Song, X., Wang, H. and Zhang, H. 2008. Sequential Local Least Squares Imputation Estimating Missing Value of Microarray Data. *Comp. Biol. Med.*, **38**: 1112–1120.

جای‌گذاری داده های گمشده با متغیرهای حمایتی و کاربرد آن در عملکرد دانه جو

ی. تاندوقدو، م. اربیلن

چکیده

عددهای گمشده در یک مجموعه از داده ها مسله ای است که به طور گسترده مطالعه شده است. در پژوهش حاضر، برای استفاده از متغیر های حمایتی (support variables) که رابطه نزدیکی با متغییر مورد نظر دارند و جاگذاری آنها به جای داده های گمشده پیشنهادی ارایه شده است. پیش از این که متغیرهای حمایتی جاگذاری شود، سطح همراهی یا رابطه آنها با متغیر مورد نظر تعیین شد. در این پژوهش، از عملکرد دانه جو (*Hordeum vulgare*) در منطقه نیمه خشک قبرس به صورت مطالعه موردی استفاده شد. در این زمینه، باران ماهانه، میانگین ماهانه درجه حرارت، و نسبت مواد آلی خاک به عنوان متغیر های حمایتی در نظر گرفته شد. برای جاگذاری داده های گمشده از تکنیک های رگرسیون چند متغییره با متغییر های حمایتی، متغیر دو گانه (bivariate)، رگرسیون هسته ای (kernel regression)، و زنجیره مونت کارلو مارکوف(Markov Chain Monte Carlo) استفاده شد. نتایج به دست آمده حاکی از کارکرد بهتر رگرسیون چند متغیره با متغیر های حمایتی در مقایسه با دیگر روش ها بود.