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

Mustafa Erbilen

Submitted to the
Institute of Graduate Studies and Research
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
in
Mathematics

Eastern Mediterranean University
July 2019
Gazimağusa, North Cyprus

Approval of the Institute of Graduate Studies and Research

Prof. Dr. Ali Hakan Ulusoy
Acting Director

I certify that this thesis satisfies all the requirements as a thesis for the degree of Doctor of Philosophy in Mathematics.

Prof. Dr. Nazım Mahmudov
Chair, Department of Mathematics

We certify that we have read this thesis and that in our opinion it is fully adequate in scope and quality as a thesis for the degree of Doctor of Philosophy in Mathematics.

Asst. Prof. Dr. Yücel Tandoğdu
Supervisor

Examining Committee

1. Prof. Dr. Cherkez Agayeva
2. Prof. Dr. Rashad Aliyev
3. Prof. Dr. Agamirza Bashirov
4. Prof. Dr. Emine Mısırlı
5. Asst. Prof. Dr. Yücel Tandoğdu

ABSTRACT

In any data collection process, regardless of the sampling method, missing data values are encountered due to many different reasons. Depending on the amount of missing data the results to be obtained from the analysis of such data will somehow be affected. Therefore, starting from 1950s an increasing interest is shown by statisticians on one hand how to minimize the missing data values and also how to impute the missing values.

In this thesis the theory and methods employed so far for the imputation of missing values in a data set are studied in detail. This is followed by the introduction of a new concept in the imputation of missing data using the support variables as part of multivariate regression process. Conversion of the units of support variables to that of the response variable is very important and is studied in detail via the imputation of missing values in a barley grain yield data set. Application results of the support variable concept is comepared with the results obtained from Markov Chain Monte Carlo (MCMC), Gaussian and Epanechnikov Kernel regression and found to be better performer in terms of lower error levels and in terms of robustness. The robustness of the results of all methods are checked using the Relative Aitchison Distance (RDA) concept.

Keywords: Missing Value, Imputation, Support Variables, Mean Squared Error (MSE), Regression, Correlation, Kernel Regression.

ÖZ

Veri toplama yöntemine bakılmaksızın, herhangi bir veri toplama işleminde çok değişik nedenlerden kaynaklanan veri eksiklikleri oluşmaktadır. Eksik verilerin az veya çok oluşuna göre, böyle bir veri tabanını kullanarak yapılacak herhangi bir veri analizinin sonuçlarıda etkilenecektir. Bu nedenle, veri toplama işleminde eksik verilerin minimuma indirgenmesi veya eksik verilerin tahmin edilmesi konularında istatistikciler 1950li yıllardan bu yana giderek artan oranda konuyla ilgili araştırmalarına devam etmektedir.

Bu tez çalışmasında bugüne kadar konuyla ilgili yapılan birçok teorik ve pratik çalışma detaylı olark incelenmiştir. Bunu takip eden aşamada eksik verilerin tahmin işleminde, desetek değişkenlerinin çok değişkenli regresyonda kullanımı önerilmiştir. Destek değişkenlerine ait birimlerin bağımlı değişken birimine dönüştürülmesi çok önemli olduğundan, detaylı olarak incelenmiş ve arpa verimliliği verisi kullanılarak uygulaması yapılmıştır. Destek verileri kullanılarak yapılan uygulamadan elde edilen sonuçlar, Markov Chain Monte Carlo (MCMC), Gaussian ve Epanechnikov Kernel regresyon metodlarından elde edilen sonuçlarla, tahmin hataları, ve tahminlerin güçlülüğü açısından kıyaslanmıştır. Elde edilen sonuçlara göre önerilen destek verileri ile tahmin yöntemi daha düşük hatalı va daha güçlü tahminler vermiştir. Tahminlerin gücü Relative Aitchison Distance (RDA) yöntemi ile hesaplanmıştır.

Anahtar kelimeler: Eksik Değer, Veri Atama, Destek Değişkeni, Hata Karelerinin Ortalaması (HKO), Regresyon, Korelasyon, Kernel Regresyonu.

DEDICATION

To the Memory of My Dear Parents: Dudu and Hüseyin

ACKNOWLEDGEMENT

I am deeply indebted to my supervisor, Asst. Prof. Dr. Yücel Tandoğdu, who has long inspired me to pursue my dissertation from a wide range of perspectives. Through this research, I always felt his support and confidence.

I am also grateful to my committee members Prof. Dr. Rashad Aliyev, Prof. Dr. Sonuç Zorlu Oğurlu, Prof. Dr. Emine Mısırlı and Prof. Dr. Cherkez Agayeva for their critical and encouraging comments, and feedbacks which provided different perspectives to look at my dissertation that refined it.

I am extremely thankful to the Chair of the Mathematics department Prof. Dr. Nazim Mahmudov for kind guidance and encouragement and all department members of Mathematics for their coordination and cooperation.

I would thank to my late advisor Prof. Dr. Gündüz İkeda and Prof. Dr. Cemal Koç that we lost a few years ago. Also, I am grateful to Akil Çelebi and Emirali Erek, who I have taken as a model to study mathematics.

And finally, I am deeply grateful to my wife, Assoc. Prof. Dr. Süheyla Üçışık Erbilen. Her emotional support, her incredible patience and understanding, were invaluable during the ups and downs of my dissertation project. I am also deeply indebted to my daughters, Yasemin Erbilen and Merve Uysal, whose support and understanding to me and for being by my side.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS

$\mathbf{e}$	Eigenvector
$\mathcal{N}(\mu,\sigma^2)$	Normal distribution with mean μ and variance σ^2
$\mathbf{X}'$	Transpose of a vector $\mathbf{X}$
$N(0,1)$	Standard normal distribution
σ	Standard deviation
$\overline{\mathbf{x}}$	Sample mean vector
$\overline{x}$	Sample mean
$\mathbf{S}$	Sample covarince matrix
$\mathbf{R}$	Sample correlaiton coefficient matrix
Y	Response variable
X	Predictor variable
$\boldsymbol{\mu}$	Mean vector
μ	Mean
$\boldsymbol{\Sigma}$	Covariance matrix
$\boldsymbol{\rho}$	Correlation
$\hat{h}_0$	Optimum *h* value
$\hat{m}_h(x)$	Nadaraya - Watson Estimator
f	Kernel function
K	Gaussian Kernel
$\hat{f}$	Estimator of f
ε	Error variable

h	Bandwidth
$\mathbf{X}$	Data vector matrix
$\boldsymbol{\sigma}^2$	Variance vector
$\mathbf{X}_{n\times p}$	Data matrix with some missing values
$f(x_1, x_2)$	Normally distributed bivariate p.d.f.
θ	Population parameter
$f(\mathbf{R}\|\mathbf{X})$	Conditional distribution
$\mathbf{X}^*_{n\times p}$	Matrix with missing values
X_j	Missing pattern
$\mathbf{y}_{n\times 1}$	Response vector matrix
$\boldsymbol{\varepsilon}$	Error vector
$\mathbf{W}$	Random matrix
R_k	Roughness
w_h	Weight function
$m_h(x)$	Average function
$\hat{m}_h(x)$	Estimator
$\hat{\mu}$	MC approximation of μ
$\hat{\sigma}/\sqrt{n}$	Monte Carlo Standard Error (MCSE)
S^d	Whole sample space
$d_A(x_i, \hat{x}_i)$	Aitchison distance
E	Expectation

t/ha	Grain production in tons/hectare
X_1 (mm/m^2)	Monthly average rain
X_2 ($^{\circ}C$)	Monthly average temperature
X_3 (unitless)	Soil organic matter ratio
$\overline{t}$	The overall average temperature

Chapter 1

INTRODUCTION

Data analysis is one of the corner stones in statistical science as it enables the testing of many statistical theory developed for various purposes or aims. Analysing large data volumes is becoming more efficient in terms of processing time parallel to the developments in computer technology. On the other hand, data itself is extremely important in the sense that any errors in the data will undoubtedly lead to incorrect analysis results. This in turn will most of the time cause grave consequences while using such results in decision making.

One other frequently encountered problem is in the collection of data. In a data set with time or space coordinates, collection of certain data values at certain locations may not be possible due to many different reasons. Such cases will result in incomplete data sets, sometimes rendering the data unusable, or severely reducing the use of such data depending on the proportion of missing values. The missing data problem has drawn the attention of researchers and serious research results started appearing from the middle of 20^{th} century onwards. Increasing amount of research has gone into this topic in order to alleviate the impact of missing values in the analysis results. Earlier work dated to 1956 Edgett [14] attempted to estimate population parameters via multivariate regression when missing observations exist in the independent variables. Anderson (1957) [6] proposed an estimation method for population means $\boldsymbol{\mu}$,

variances $\boldsymbol{\sigma}^2$, and correlations $\boldsymbol{\rho}$ for the *k* variate normal distribution via Maximum Likelihood Estimation (MLE), when missing values exists in the response variable. His work can be summarized in the bivariate case where $f(x_1, x_2)$ being normally distributed with means μ_1, μ_2, variances σ_1^2, σ_2^2, and correlation $\rho_{x_1 x_2}$. Out of *N* total data *n* are tuples belonging to (x_1, x_2) pairs and *N-n* observations belonging to x_1 only, meaning *N-n* missing observations in x_2. Maximum likelihood estimates are obtained for the distribution parameters by employing the *n* tuples with complete data and ignoring the tuples with missing values. Then goes on to generalize the concept to multivariate case. The works of Edgett and Anderson correspond to what is now known as "missing values at random". Trawiski and Bergmann (1964) [52] worked on estimation and testing methods on hypothesis testing in the multivariate case. In their study missing values are systematically introduced into the system. In the absence of current computation power of computers, they attempted to show their case by some artificial data and manual computations. Obtained results were claimed to be satisfactory. Afifi and Elanshoff (1966) [2] have worked on the simplification of estimation problems when the missing data fits into certain patterns, as well as the statistical properties of such estimators. In their work they employed multivariate regression, and maximum likelihood methods to estimate the population parameters.

Rubin (1976) [40] examined the case when data is missing at random and observations are observed at random by making direct likelihood or Bayesian inferences about the population parameter θ, ignoring the process that caused the missing data. It is pointed out that inferences made about population parameters are conditioned on the pattern of missing data. This is possible when the parameter of missing data process is not the

same as θ. Under such conditions correct inference is possible by ignoring the process that caused missing data. In his work Little (1992) [30] reviewed six different procedures that are used in imputation process. Namely these are complete case, available case, least squares, maximum likelihood, Bayesian and multiple imputation methods. The methods are compared for the missing data one independent variable, and the logic developed is extended to more general patterns. The likelihood methods are preferred over the least square methods, as they utilize both the dependent and independent variables. Copt and Feser (2003) [10] compared different robust estimators using a simulation study to determine their efficiency when missing data exists. They proposed faster algorithms to compute robust estimators with missing data and compared obtained results. The Orthogonalized Gnanadesikan-Kettenring estimator was applied in the case of missing data Zamar and Maronna (2002) [57]. Toutenburg et. al. (2005) [51] offered some modification to the linear regression model when missing observations exist in the independent variables. The standard first order regression is modified to enable the imputation of missing values. Under the proposed modification asymptotic properties of the estimators for the regression coefficients are derived. Zhang et. al. (2008) [58] proposed a sequential local least squares imputation approach to deal with missing values in the gene microarray data. Imputation proceeds sequentially starting with the gene with least missing values. Imputed values are then utilized in the estimation of subsequent missing values together with existing values. Also, an automatic parameter selection and estimation algorithm is introduced. Robbins et. al (2013) [41] in their study of the high dimensional data of the US Agricultural Resource Management Survey, they transformed the data using skewed marginal densities, under the assumption that they can be linked using a Gaussian copula. This in turn facilitate the obtaining a joint model. Imputation is based on these

joint models. Parameter estimation and imputation are estimated using the Markov Chain Monte Carlo sampling approach. Yozgatlıgil et. al. (2012) [56] compare six imputation methods and together with their proposed algorithm applied the methodologies to impute missing values in a spatio – temporal meteorological time series. Criteria such as accuracy, robustness, precision, and efficiency are considered in the comparison process. In a more recent research Jinubala and Lawrance (2016) [25] an analysis of Predictive Mean Matching Method has been used to determine and impute missing values for crop pest data. This method is similar to the regression method with the exception that missing value, it imputes a value randomly from a set of observed values are estimated and the estimate is compared with the estimates for the same data obtained from a simulated regression model.

In this study the barley grain yield data in t/ha from Northern Cyprus covering the years 1996 – 2012 and split into 17 production areas is taken as an example for the application of the proposed "Use of Support Variables in Imputing Missing Data". From the obtained complete data set two new data sets are generated with 10% and 40% missing values. Values are deleted on a random basis to obtain the new data sets satisfying the "completely missing at random" process. 3 support variables are selected from amongst many variables affecting the barley grain yield, such that they have high or significant effect on grain yield. Namely, these are Rain (mm/m^2), temperature (^0C), and soil organic matter ratio. Efficient use of these variables in estimating the missing values for the grain yield data, their units were converted into the same unit as grain yield (t/ha). For this purpose, an extensive literature survey was undertaken in developing the suitable algorithm for the conversion of each support variable's unit into t/ha.

Most useful research work that contributed in converting rain figures from mm/m^2 to t/ha are briefly given below.

Cantero et. al. (1995) [8] studied the barley grain yield performance of two barley cultivars with different phenology for 4 years, in the Ebro valley of Spain which has similar semi-arid conditions as Cyprus. Factors such as growth, yield and yield components, water use and root development were taken into consideration. For the period of study with rainfall below average, yields ranged from 1.2 to 3.0 t/ha. It is determined that there is 0.75 correlation between the total water use by the crop and evapotranspiration during the flowering (pre-anthesis) period. Number of ears per square meter is the determining factor for final grain yield. It was found that the Dobla cultivar's grain filling period was less adversely affected compared with the Tina cultivar, with respective yields 3.0 and 2.3 t/ha. Water use and water use efficiency figures given in the study were useful in the conversion of rainfall figures to grain yield in t/ha. Lopez and Arrue (1997) [31] studied the efficiency of different tillage methods in terms of barley and wheat production in the semi – arid climate of Aragon in NE Spain. In order to determine the feasibility of conservation tillage 3 locations with soil type loam to silt loam soils (Xerollic Calciorthid) and at one location with silty clay loam (Fluventic Ustochrept) were utilized, where annual average rainfall ranged between 300 to 600 mm. Grain yield under both continuous cropping and cereal-fallow rotation, when conventional tillage (mouldboard plough), reduced tillage (chisel plough) and no-tillage were implemented were studied for winter barley (Hordeum vulgare). Growth, grain yield and water use efficiency were taken into consideration. No tillage treatment was found to exhibit a poor performance, with 53% loss in grain yield. Under no tillage condition the water usage was lower and

evapotranspiration at 69% was much higher than the tilled treatments being at 50%. On the other hand, water use efficiency for grain production was 0.7-17.0 kg/ha/mm, and transpiration efficiency was 7.4 to 23.8 kg/ha/mm, being typical values for the semi-arid regions of Mediterranean environments. Samarah (2005) [42] 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. It was determined that grain dry weight under severe drought conditions reached their maximum value earlier than those subjected to mild drought stress or well-watered plants. It was also observed that sever drought-stress resulted in shorter duration for grain filling. This in turn caused low grain yield for plants subjected to sever stress conditions. Ebrahimian and Playan (2014) [13] studied the efficiency and uniformity of water and fertilizer application for optimum management of irrigation and fertigation systems. Variables monitored for decision making were inflow discharge, irrigation cut off, start times and duration of fertilizer injection. During the experiments soil water content, soil nitrate concentration, discharge and nitrate concentration in runoff, advance and recession times were recorded in order to enable the calibration of the models. Observed water and nitrate application efficiencies ranged from 72% to 88%.

In transforming the effect of heat to barley grain yield, research carried out by Nahar et. al. (2010) [13] and Hossein et. al. (2012) [23] have contributed to the development of the transformation algorithm. In their study Nahar et. al. considered 5 varieties of wheat grown under normal and under heat stress conditions in order to assess the effect of heat on grain yield. Those subjected to heat stress were significantly affected in terms of days required to germination, booting, anthesis, maturity and grain yield

compared to cultivars treated with normal conditions from sowing to harvest. For the normal seeding case temperature during the grain filling or maturing period was around 23°C for the late seeding case temperature in the range 28°C to 30°C or even above this range in the later occasions. Hence it was observed that grain yield was significantly lowered in all 5 varieties of wheat as compared to those grown under normal conditions. Average loss in grain yield for all 5 varieties subjected to heat stress was about 62%.

A study carried out in the southern arid region of Russia by Hossein et. al. assessed the performance of four spring barley and two spring wheat genotypes under two stress (early and late) conditions. In order to determine the optimum sowing time for specific genotypes of crops, it is observed that late sown crops were affected by high temperature and a deficit of soil moisture in all stages from germination through to harvest substantially reducing the grain yield. Early sown crops were affected by low temperature, resulting in deficiencies in germination and stand establishment of crop resulting in lower grain yield. From this study it is estimated that for every increase in the long term monthly average temperature, an average of 3% to 6% loss in grain yield will occur.

The Soil Organic Matter Ratio (SOMR) is also a very important variable that has significant contribution to grain yield. SOMR is a unitless quantity and its conversion into grain yield in t/ha necessitates careful consideration of the factors involved in its determination. For this purpose, the research work undertaken by different researcher have been very useful.

In his study Tiessen, et. al. (1994) [50] point out to the fact that the effect of fertilization is not inconsistent due to leaching or fixation of inorganic nutrients. An attempt is made to quantify the effects of organic matter on the fertility of soils in temperate prairie, and tropical semi-arid climate zones. Carbon turnover was estimated from ecological measurements and 14°C dating and determine Relation between the soil carbon and nutrient budgets is also determined. It is also found that on temperate prairie, without supplementary fertilization agriculture was economical for 65 years, but only for six years in a tropical semi-arid thorn forest.

In a study Reeves (1997) [39] took the Soil organic carbon (SOC) as the attribute to investigate as it is an indicator of soil quality and agronomic sustainability. It is a known fact based on long term experience manures, adequate fertilization, and crop rotation, can increase SOC when coupled with intensive cropping. Good rotation practice is necessary to achieve agronomic productivity and economic sustainability. Stine & Weil (2002) [45] investigated the relation between SOM and grain yield under 3 tillage systems. They determined that there is a linear relation soil organic matter ratio and grain yield. This is a useful tool in converting SOMR to grain yield in t/ha. Quiroga et. al. (2005) [36] defines the SOMR as he ratio of SOM (g/kg) to clay + silt content (g/kg) and uses it as an indicator for soil quality. They determined that 51% of grain yield is attributable to SOMR. Almost 68% grain yield prediction was due to combining the SOM to clay + silt indicator and initial nitrate (N) content of the soil at seeding. High proportion of water use efficiency can be explained by this indicator. They concluded that SOMR is a better tool for estimating grain yield compared to nutrient availability or SOM alone.

Based on a very long term study that dates back to 1852 which started by enriching loam type soils with Nitrate (N), Phosphorus (P), and Potassium (K) as well as farmyard manure Johnston (2011) [27] studied the role of soil organic matter (SOM) on grain yield. Soils following such a long-term treatment with fertilizers are now showing SOM values ranging between 1.74% to 6.16%. Interaction between N and P with SOM and its effects on crop yield are illustrated, at different SOM levels, over 4 different harvesting periods during the 1980s and 1990s.

The main objective of this thesis is the study of imputing missing values in a given data set. Along this line a new imputation method is proposed where the estimation of missing values in a certain variable (response variable) will utilize the existing data from that variable, as well as data from other variables that have a strong influence in the realization of the values of the response variable. This approach is named as "Imputation Using Support Variables". The idea is applied to the barley grain yield data from North Cyprus for the years 1996 – 2012. This is a complete data set representing 17 production areas. Data values were deleted at random from this data set to obtain two new ones, one with 10% and the other 40% missing values. In general support variables do not have the same unit as the repose variable. Therefore, an in-depth study is necessary to establish a relationship that will transform the unit and quantity of the data from support variables to the same unit of the response variable. As a result, extensive research is conducted to develop the necessary algorithms for transforming the support variables' (rain, heat, and SOMR) units into barley grain yield unit (t/ha). Upon successful transformation of the SV units into t/ha, they were employed in multivariate regression for the imputation process of missing values for the 10% and 40% missing cases. For the comparison of the performance of the

proposed SV method, other well-known estimation methods were also used. Namely, these were simple linear regression, multivariate regression, univariate kernel regression, and Markov Chain Monte Carlo (MCMC) methods.

In Chapter 2 concepts and theory related with the mechanism of missing data and its imputation are explained. Important theory and theorems related with multivariate regression are explained in detail. In kernel regression the most important point is the determination of band width, and the proper use of Nadaraya Watson estimator. Theory and some relevant theorems are explained in detail. Similarly, a fair explanation of the theory related with Chain Monte Carlo (MCMC) method is given. In this chapter an interesting method called the Relative Atchison distance is also explained, as it provides some idea about the robustness of estimated made by any estimation method. Analysis of variance (ANOVA) related theory is also summarized, as it is employed in testing the means of the different estimation methods used in the study.

Chapter 3 is devoted to the proposed idea of using support variables for imputation of missing data. Determination of SVs is explained and their transformation to the same unit as the response variable is given in detail for the SVs used for barley grain yield.

In Chapter 4 the necessary background i.e. climatic conditions, mainly rain and heat variables soil related conditions (SOMR) are examined for the area of study from where barley grain yield data is obtained. Algorithms developed in Chapter 3 for the transformation of SV units into the same unit as the response variable are employed to convert rain, temperature, and SOMR data into t/ha. Then all methods mentioned in

Chapter 3 are used for imputation. For each method error levels computed and compared to determine the best performing imputation method. The proposed multivariate regression using support variables was the best performer in terms of errors committed. Table 3 in Chapter 4 summarizes the mean square error expressed as a percentage of observed and imputed values (MSE%). Robustness of the estimates are also determined using the Relative Aitchison Distance method, where the proposed method turned out to be the most robust one.

Chapter 2

INCOMPLETE DATA SETS AND IMPUTATION OF MISSING DATA

Missing values are frequently encountered in many fields during data collection, resulting in incomplete data sets. In many applications missing data is encountered due to many different reasons. This may lead to difficulties in the estimation process or results that are not reliable. 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.

Conceptually it is assumed that missing values can occur as *completely at random* and *at random.* It is obvious that any missing value occurs as a function of some unknown process. Given a data matrix $\mathbf{X}_{n\times p}$ with some missing values, the mechanism that results in missing data can be defined as follows. Define matrix $\mathbf{R}_{n\times p}$ as an indicator matrix where $r_{ij}=1$, if the corresponding x_{ij} element from the $\mathbf{X}_{n\times p}$ matrix is observed, and $r_{ij}=0$ when the x_{ij} element is missing. Then the conditional distribution $f(\mathbf{R}|\mathbf{X})$ can be used to distinguish between the mechanisms leading to missing values.

If $f(\mathbf{R}|\mathbf{X}) = f(\mathbf{R}),\ \forall \mathbf{X}$, mechanism is called missing completely at random (MCAR).

If $f(\mathbf{R}|\mathbf{X}) = f(\mathbf{R}|\mathbf{X}^*),\ \forall \mathbf{X}^*$, mechanism is missing at random (MAR).

$\mathbf{X}^*_{n\times p}$ is the matrix with missing values.

The pattern of missing data is also important and can be categorised as Rao [26]

- Monotone missing pattern. If rows and columns of the data matrix $\mathbf{X}_{n\times p}$ can be rearranged such that $X_{j+1}; j = 1,\ldots,p-1$ exists for all cases where X_j exists. A special case of this pattern is the when missing values exist in any one variable X_j only.

- Special missing pattern. When any two variables X_j and X_k are never observed together. This pattern may be encountered when two data sets are merged.

- General missing pattern. This is the case where any specific pattern cannot be observed or attributed to the variables in terms of missing values.

2.1 When missing values are in the response variable

Missing values can in general exist both in the dependent (response) and independent (predictor) variables. In this study as multivariate regression is extensively used, it is deemed appropriate to mention some of the important theoretical background related to this topic.

2.1.1 Multivariate linear regression model

Given a data set with p variables also considered as predictor variables and n observations is to be used in estimating a response variable Y that also has n observations. In general, assume Y is a function of p variables $X_i;\ i=1,\ldots,p$ plus an error term given as

$$Y = f(X_1,\ldots,X_p)+e$$

f may be a linear or non-linear function. In the linear case we can write

$$Y=\beta_0+X_1\beta_1+\cdots+X_p\beta_p+e$$

called the linear regression model. In the multivariate case let the response vector be $\mathbf{y}_{n\times 1}$ and predictor matrix $\mathbf{X}_{n\times p}$. Then the multivariate regression model can explicitly be written as Johnson [15]

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix} \rightarrow \mathbf{y}_{n\times 1} = \mathbf{X}_{n\times(p+1)}\boldsymbol{\beta}_{(p+1)\times 1} + \boldsymbol{\varepsilon}_{n\times 1}$$

Here $\boldsymbol{\varepsilon}$ is the error vector with $E(\boldsymbol{\varepsilon}) = \mathbf{0}$ and $Cov(\boldsymbol{\varepsilon}) = E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}') = \sigma^2\mathbf{I}$. Information regarding the unknown parameter σ^2 is contained in $\hat{\boldsymbol{\varepsilon}} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}$. $\boldsymbol{\beta}$ and $\boldsymbol{\varepsilon}$ are unknown parameters. The estimator of the coefficient vector $\boldsymbol{\beta}$ is $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}\mathbf{y}$. This is given in detail in Theorem 1.

Theorem 1: The least squares estimate of $\boldsymbol{\beta}$ can be written as $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$, when $\mathbf{X}$ has full rank $p+1 \le n$.

Proof: To show that $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$;

$$\boldsymbol{\varepsilon} = \mathbf{y} - \hat{\mathbf{y}} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}} = \left[\mathbf{I} - \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\right]\mathbf{y} .$$

The matrix $\left[\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right]$ satisfies the following three conditions;

i) It is symmetric: $\left[\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right]' = \left[\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right]$,

ii) Satisfies the idempotent condition:
$\left[\mathbf{I} - \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\right]\left[\mathbf{I} - \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\right] = \mathbf{I} - 2\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'$
$+\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}' = \left[\mathbf{I} - \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\right]$,

iii) Uniqueness condition:

$$\mathbf{X}'\left[\mathbf{I} - \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\right] = \mathbf{X}' - \mathbf{X}' = \mathbf{0} .$$

As a result

$$\mathbf{X}'\boldsymbol{\varepsilon} = \mathbf{X}'(\mathbf{y} - \hat{\mathbf{y}}) = \mathbf{X}'\left[\mathbf{I} - \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\right]\mathbf{y} = \mathbf{0}$$

can be written. Hence,

$$\hat{\mathbf{y}}'\hat{\boldsymbol{\varepsilon}} = \hat{\boldsymbol{\beta}}'\mathbf{X}'\hat{\boldsymbol{\varepsilon}} = \mathbf{0}.$$

Further, $\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}} = \mathbf{y}\left[\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right]\left[\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right]\mathbf{y}$

$= \mathbf{y}'\left[\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right]\mathbf{y} = \mathbf{y}'\mathbf{y} - \mathbf{y}'\mathbf{X}\hat{\boldsymbol{\beta}}$.

Let $\mathbf{y} - \mathbf{Xb} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{X}\hat{\boldsymbol{\beta}} - \mathbf{Xb} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{X}(\hat{\boldsymbol{\beta}} - \mathbf{b})$, and $S(\mathbf{b})$ the sum of the squares of the differences $\sum_{j=1}^{n}(y_j - b_0 - b_1 x_{j1} - \cdots - b_p x_{jp})$.

Then

$$\begin{aligned} S(\mathbf{b}) &= \left(\mathbf{y} - \mathbf{Xb}\right)'\left(\mathbf{y} - \mathbf{Xb}\right) \\ &= \left(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\right)'\left(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\right) + \left(\hat{\boldsymbol{\beta}} - \mathbf{b}\right)'\mathbf{X}'\mathbf{X}\left(\hat{\boldsymbol{\beta}} - \mathbf{b}\right) + 2\left(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\right)'\mathbf{X}(\hat{\boldsymbol{\beta}} - \mathbf{b}) \\ &= \left(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\right)'\left(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\right) + \left(\hat{\boldsymbol{\beta}} - \mathbf{b}\right)'\mathbf{X}'\mathbf{X}\left(\hat{\boldsymbol{\beta}} - \mathbf{b}\right) \end{aligned}.$$

As $\left(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\right)'\mathbf{X} = \boldsymbol{\varepsilon}\mathbf{X} = \mathbf{0}'$.

Note the following points

i. The first term in $S(\mathbf{b})$ does not depend on $\mathbf{b}$. Second is the squared length of $\mathbf{X}\left(\hat{\boldsymbol{\beta}} - \mathbf{b}\right)$.

ii. As $\mathbf{X}$ is full rank, $\mathbf{X}\left(\hat{\boldsymbol{\beta}} - \mathbf{b}\right) \neq \mathbf{0}$ if $\hat{\boldsymbol{\beta}} \neq \mathbf{b}$, meaning the minimum sum of squares is unique occurring when $\mathbf{b} = \hat{\boldsymbol{\beta}} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y}$.

iii. $\mathbf{X}'\mathbf{X}$ has rank $p+1 \le n$ means $(\mathbf{X}'\mathbf{X})^{-1}$ exists. Assuming $\mathbf{X}'\mathbf{X}$ is not full rank, $\mathbf{X}'\mathbf{X}\mathbf{a} = \mathbf{0}$ for some $\mathbf{a} \neq \mathbf{0}$. But $\mathbf{a}'\mathbf{X}'\mathbf{X}\mathbf{a} = \mathbf{0}$ or $\mathbf{X}\mathbf{a} = \mathbf{0}$, contradicts $\mathbf{X}$ having full rank $p+1$. *Q. E. D.*

Relationship between theoretical least squares and classical least squares estimators should be mentioned to appreciate the meaning of the application of regression theory. Sampling properties of least squares estimators for $\hat{\boldsymbol{\beta}}$ and for the residuals $\hat{\boldsymbol{\varepsilon}}$ are given in the following theorem.

Theorem 2: The least squares estimator of $\boldsymbol{\beta}$, in the general linear regression model $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ has expectation $E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$ and $Cov(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$.

Similarly, the residuals $\hat{\boldsymbol{\varepsilon}}$ have $E(\hat{\boldsymbol{\varepsilon}}) = \mathbf{0}$ and $Cov(\hat{\boldsymbol{\varepsilon}}) = \sigma^2\left[\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right]$.

Then

$$s^2 = \frac{\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}}{n-(p+1)} = \frac{\mathbf{y}'\left[\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right]\mathbf{y}}{n-p-1}$$

leading to $E(s^2) = \sigma^2$. Further $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\varepsilon}}$ are uncorrelated.

Proof: Theoretically $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ is a random vector. Then,

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) = \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon}$$

$$\hat{\boldsymbol{\varepsilon}} = \left[\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right]\mathbf{y} = \left[\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right](\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) = \left[\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right]\boldsymbol{\varepsilon},$$

Since

$$\left[\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right]\mathbf{X} = \mathbf{X} - \mathbf{X} = \mathbf{0}.$$

$E\left(\hat{\boldsymbol{\beta}}\right) = \boldsymbol{\beta} + \mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E(\boldsymbol{\varepsilon}) = \boldsymbol{\beta}$ as $E(\boldsymbol{\varepsilon}) = \mathbf{0}$

$$\begin{aligned} Cov\left(\hat{\boldsymbol{\beta}}\right) &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'Cov(\boldsymbol{\varepsilon})\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= \sigma^2\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} = \sigma^2(\mathbf{X}'\mathbf{X})^{-1} \end{aligned}$$

$E(\hat{\boldsymbol{\varepsilon}}) = \left[\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right]E(\boldsymbol{\varepsilon}) = \mathbf{0}$,

$$\begin{aligned} \mathbf{Cov}(\hat{\boldsymbol{\varepsilon}}) &= \left[\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right]\mathbf{Cov}(\boldsymbol{\varepsilon})\left[\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right]' \\ &= \sigma^2\left[\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right]. \end{aligned}$$

This equality follows the idempotent property of $\left[\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right]$. Also,

$$\begin{aligned} \mathbf{Cov}\left(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\varepsilon}}\right) &= \mathbf{E}\left[\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)\hat{\boldsymbol{\varepsilon}}'\right] = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}')\left[\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right] \\ &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\left[\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right] = \mathbf{0}. \end{aligned}$$

Since

$$\mathbf{X}'\left[\mathbf{I}-\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right]=\mathbf{0}\ .$$

From the definition of $\hat{\boldsymbol{\varepsilon}}$ and remembering given a square matrix $\mathbf{A}_{k\times k}$

and vector $\mathbf{x}_{k\times 1}$,

$$\mathbf{x}'\mathbf{A}\mathbf{x}=tr(\mathbf{x}'\mathbf{A}\mathbf{x})=tr(\mathbf{A}\mathbf{x}\mathbf{x}')$$

holds. Then,

$$\begin{aligned}\hat{\boldsymbol{\varepsilon}}'\boldsymbol{\varepsilon}&=\boldsymbol{\varepsilon}'\left[\mathbf{I}-\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right]\left[\mathbf{I}-\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right]\boldsymbol{\varepsilon}=\boldsymbol{\varepsilon}'\left[\mathbf{I}-\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right]\boldsymbol{\varepsilon}\\&=\mathbf{tr}\left(\boldsymbol{\varepsilon}'\left[\mathbf{I}-\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right]\boldsymbol{\varepsilon}\right)=\mathbf{tr}\left(\left[\mathbf{I}-\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right]\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\right).\end{aligned}$$

For any arbitrary $n\times n$ random matrix $\mathbf{W}$,

$$E\left(tr\left(\mathbf{W}\right)\right)=E\left(\mathbf{W}_{11}+\mathbf{W}_{22}+\cdots+\mathbf{W}_{nn}\right)=E\left(\mathbf{W}_{11}\right)+E\left(\mathbf{W}_{22}\right)+\ldots+E\left(\mathbf{W}_{nn}\right)=tr\left(\mathbf{E}\left(\mathbf{W}\right)\right).$$

Keeping in mind that for a square matrix $\mathbf{A}$, $tr(c\mathbf{A}) = ctr(\mathbf{A})$ we have

$$E\left(\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}\right) = tr\left(\left[\mathbf{I} - \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\right]E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}')\right) = \sigma^2 tr\left(\left[\mathbf{I} - \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\right]\right)$$
$$= \sigma^2 tr\left(\mathbf{I}\right) - \sigma^2 tr\left(\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{X}\right) = n\sigma^2 - \sigma^2 tr\left(\mathbf{I}_{(p+1)\times(p+1)}\right) = \sigma^2\left(n - p - 1\right)$$

leading to the result

$$s^2 = \frac{\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}}{\left(n - p - 1\right)}. \quad Q.E.D.$$

Since the least square technique plays a major role in linear regression, it is considered necessary to mention the theorem that explains the concept of the technique.

Theorem 3: (*Gauss least square theorem*.) In the multivariate regression model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, with $E\left(\boldsymbol{\varepsilon}\right) = \mathbf{0}$, $\mathbf{Cov}\left(\boldsymbol{\varepsilon}\right) = \sigma^2\mathbf{I}$, and $\mathbf{X}$ having full rank $p+1$, for any ***c***, the estimator

$$\mathbf{c}'\hat{\boldsymbol{\beta}} = \mathbf{c}_0'\hat{\boldsymbol{\beta}}_0 + \mathbf{c}_1'\hat{\boldsymbol{\beta}}_1 + \cdots + \mathbf{c}_p'\hat{\boldsymbol{\beta}}_p$$

of $\mathbf{c}'\boldsymbol{\beta}$ has the minimum variance among the linear estimators $\mathbf{a}'\mathbf{Y} = a_1Y_1 + \cdots + a_nY_n$ that are unbiased for $\mathbf{c}'\boldsymbol{\beta}$.

Proof: Let any unbiased estimator of $\mathbf{c}'\boldsymbol{\beta}$ be $\mathbf{a}'\mathbf{Y}$, where **a** and **c** are the same size. Then $E\left(\mathbf{a}'\mathbf{Y}\right) = \mathbf{c}'\boldsymbol{\beta}$, independent of $\boldsymbol{\beta}$.

$$E(\mathbf{a}'\mathbf{Y}) = E(\mathbf{a}'\mathbf{X}\boldsymbol{\beta} + \mathbf{a}'\boldsymbol{\varepsilon}) = \mathbf{a}'\mathbf{X}\boldsymbol{\beta} \cdot$$

Then

$$\mathbf{a}'\mathbf{X}\boldsymbol{\beta} = \mathbf{c}'\boldsymbol{\beta} \rightarrow (\mathbf{c}' - \mathbf{a}'\mathbf{X})\boldsymbol{\beta} = \mathbf{0} \text{ for all } \boldsymbol{\beta} .$$

This includes the choice $\boldsymbol{\beta} = (\mathbf{c}' - \mathbf{a}'\mathbf{X})'$, implying that $\mathbf{c}' = \mathbf{a}'\mathbf{X}$ for all unbiased estimators.

Now $\mathbf{c}'\hat{\boldsymbol{\beta}} = \mathbf{c}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = \mathbf{a}^{*'}\mathbf{Y}$ with $\mathbf{a}^* = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{c}$. From Theorem 2, $E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$ so $\mathbf{c}'\hat{\boldsymbol{\beta}} = \mathbf{a}^{*'}\mathbf{Y}$ becomes an unbiased estimator of $\mathbf{c}'\boldsymbol{\beta}$. Thus, for any $\mathbf{a}$ satisfying the unbiased requirement $\mathbf{c}' = \mathbf{a}'\mathbf{X}$,

$$Var(\mathbf{a}'\mathbf{Y}) = Var(\mathbf{a}'\mathbf{X}\boldsymbol{\beta} + \mathbf{a}'\boldsymbol{\varepsilon}) = Var(\mathbf{a}\boldsymbol{\varepsilon})$$

$$= \mathbf{a}\mathbf{I}\sigma^2\mathbf{a} = \sigma^2(\mathbf{a} - \mathbf{a}^* + \mathbf{a}^*)'(\mathbf{a} - \mathbf{a} + \mathbf{a}^*)$$

$$= \sigma^2\left[(\mathbf{a} - \mathbf{a}^*)'(\mathbf{a} - \mathbf{a}^*) + \mathbf{a}^{*\prime}\mathbf{a}^*\right].$$

Since

$$(\mathbf{a} - \mathbf{a}^*)'\mathbf{a}^* = (\mathbf{a} - \mathbf{a}^*)'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{c} = \mathbf{0}$$

based on

$$\left(\mathbf{a}-\mathbf{a}^{*}\right)' \mathbf{X}=\mathbf{a}'\mathbf{X}-\mathbf{a}^{*\prime}\mathbf{X}=\mathbf{c}'-\mathbf{c}'=\mathbf{0}'.$$

Because $\mathbf{a}^{*}$ is fixed and $\left(\mathbf{a}-\mathbf{a}^{*}\right)'\left(\mathbf{a}-\mathbf{a}^{*}\right)$ is positive, $Var\left(\mathbf{a}'\mathbf{Y}\right)$ is minimized by choosing

$$\mathbf{a}^{*\prime}\mathbf{Y}=\mathbf{c}'\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{Y}=\mathbf{c}'\hat{\boldsymbol{\beta}}\ .\ Q.E.D.$$

According to Theorem 3, substitution of $\hat{\boldsymbol{\beta}}$ for $\boldsymbol{\beta}$ leads to the best estimator of $\mathbf{c}'\boldsymbol{\beta}$ for any $\mathbf{c}$. In statistical terminology, $\mathbf{c}'\hat{\boldsymbol{\beta}}$ is called the best (minimum-variance) linear unbiased estimator (BLUE) of $\mathbf{c}'\boldsymbol{\beta}$.

2.1.2 Kernel Regression

In multivariate data when missing values for each observation of $X_{i1}, X_{i2}, \ldots X_{ip}; i=1, 2,\ldots, n$ occur at random, kernel regression appears as an optimal technique to help impute the missing values. In the univariate case the use of kernel smoothing assumes that the random sample $X_1, X_2, \ldots X_n$ consisting of independent and have identical distribution (i.i.d) random variables with a certain p.d.f. The kernel estimator of the unknown p.d.f f is $\hat{f}$. $\hat{f}$ is obtained using available data that utilizes the kernel function *K*. As $\hat{f}(x)$ depends of $X_1, X_2, \ldots X_n$ it can be considered as a random variable. Difference between $\hat{f}$ and f forms the error and Mean Squared Error (MSE) or the Mean Integrated Squared Error (MISE) can be used to measure this error. MSE is given as

$$MSE(\hat{f}) = E\{[\hat{f}(x) - f(x)]^2\}.$$

MSE consists of two components, namely Bias and Variance. Decomposing MSE into Bias and Variance is essential to be able to maintain the balance between the two components. Bandwidth that is the basis of the kernel function is very sensitive in determining the bias – variance balance.

Decomposing MSE into bias - variance components is explained below.

Let $x_1, x_2, \cdots x_n$ be the data. Based on this data $y = f(x) + \varepsilon$ will be estimated. Here ε is random error or noise. It has $E(\varepsilon) = 0$ and $Var(\varepsilon) = \sigma^2$). *f(x)* will be estimated using the $\hat{f}(x)$. Obviously the smaller the difference between $\hat{f}(x)$ and *f(x),* the more accurate the estimation will be. Achieving a minimum MSE using the sample $x_1, x_2, \cdots x_n$, is aimed and this will also be valid for observations that are not part of the sample. Expectation of the squared error is made up of the 3 components as given below

$$E\{[y - \hat{f}(x)]^2\} = Bias[\hat{f}(x)]^2 + Var[\hat{f}(x)] + \sigma^2 \qquad (2.1)$$

where

$$Bias[\hat{f}(x)] = E[\hat{f}(x) - f(x)] \text{ , and } \mathrm{Var}[\hat{f}(x)] = E[\hat{f}(x)^2] - f[(x)]^2.$$

Approximating *f(x)* by its estimator $\hat{f}(x)$ obviously have functional effect on the square of the bias element. The Var($\hat{f}(x)$) is also important as it has a major effect on bias – variance relation. Therefore, the more sophistication put into $\hat{f}(x)$ towards reducing the bias, will lead to higher $Var[\hat{f}(x)]$.

Equation (2.1) is to be obtained.

It is worth remembering, given a random variable its variance is

1. $\sigma^2 = E[(X-\mu)^2] \Rightarrow Var(X) = E(X^2) - E(X)^2 \Rightarrow E(X^2) = Var(X) + E(X)^2$.

   Then

2. $E[f(x)+\varepsilon] = E[f(x)] = f(x)$ but *f(x)* being deterministic and $E(\varepsilon) = 0$.

3. $Var(y) = E[(y-E(y))^2] = E[(y-f(x))^2] = E[(f(x)+\varepsilon-f(x))^2]$
   $= E(\varepsilon^2) = Var(\varepsilon) + E(\varepsilon)^2 = \sigma^2.$

Based on the independence of ε and $\hat{f}(x)$

$$
\begin{aligned}
&E[(y-\hat{f}(x))^2] = E[y^2 + \hat{f}(x)^2 + 2y\hat{f}(x)] \\
&= Var(y) + \mathrm{E}(y)^2 + Var(\hat{f}(x)) + E(\hat{f})^2 - 2f(x)E(\hat{f}(x)) \\
&= Var(y) + Var(\hat{f}(x)) + (f(x)^2 - 2f(x)E(\hat{f}(x)) + E(\hat{f}(x))^2) \\
&= Var(y) + Var(\hat{f}(x)) + (f(x) - E(\hat{f}(x)))^2 \\
&= \sigma^2 + Var(\hat{f}(x)) + Bias(\hat{f}(x))^2.
\end{aligned}
$$

It should be remembered that in $y = f(x) + \varepsilon$, the error term ε has $E(\varepsilon) = 0$. Hence,

$$E[(y - \hat{f}(x))^2] = bias(\hat{f}(x)^2) + Var(\hat{f}(x)^2)$$

can be shown to be valid as below.

For brevity $\hat{f}$ will be used in place of $\hat{f}(x)$

$$\begin{aligned} &E\{[y - \hat{f}]^2\} = E\{[y - E(\hat{f}) + E(\hat{f}) - \hat{f}]^2\} \\ &= E\{[y - E(\hat{f})]^2\} + E\{[E(\hat{f}) - \hat{f})]^2\} + 2E\{[E(\hat{f}) - \hat{f})][y - E(\hat{f})]\} \\ &= \text{bias}(\hat{f})^2 + \text{var } \hat{f} \end{aligned}$$

As it can be shown that $2E\{[E(\hat{f}) - \hat{f})][y - E(\hat{f})]\} = 0$ as follows.

$$2E\{[E(\hat{f}) - \hat{f})][y - E(\hat{f})]\} = 2\{E(yE(\hat{f})) - E(E(\hat{f})^2 - E(\hat{f}y) + E\hat{f}(E(\hat{f}))\}.$$

Each term inside the brace bracket can be expressed as

$$E(yE(\hat{f})) = yE(\hat{f}),$$

since *y* is deterministic.

$$E(E(\hat{f})^2) = E(\hat{f})^2 \text{ , since } E(E(\bullet)) = E(\bullet).$$

$$E(\hat{f}y) = yE(\hat{f}) \text{ , and } E(\hat{f}E(\hat{f})) = E(\hat{f})^2.$$

Then

$$2\{E(yE(\hat{f}))-E(E(\hat{f})^2-E(\hat{f}y)+E\hat{f}(E(\hat{f}))\}=2[yE(\hat{f})-E(\hat{f})^2-yE(\hat{f})+E(\hat{f})^2]=0.$$

Various kernel functions are available for the computation of kernel values. Properties such as variance, roughness, and efficiency associated with each kernel function are also given, where variance

$$\sigma_k^2=\int_{-\infty}^{\infty}u^2k(u)du\ ,$$

roughness

$$R_k=\int_{-\infty}^{\infty}\left[k(u)\right]^2du\ ,$$

and efficiency

$$K_E=\frac{\sqrt{\int_{-\infty}^{\infty}u^2k(u)du}}{\int_{-\infty}^{\infty}k(u)^2du}\ .$$

Variance, roughness, and efficiency for some commonly used kernel functions are Hansen [20].

Table 2.1: Roughness, variance and efficiency values for different kernel functions

Kernel		Roughness (R_k)	Variance (σ_k^2)	Efficiency
Uniform	$k(u) = \frac{1}{2} 1(\lvert u \rvert \le 1)$	½	1/3	1.155
Epanechnikov	$k(u) = \frac{3}{4}(1-u^2)1(\lvert u \rvert \le 1)$	3/5	1/5	0.745
Biweight	$k(u) = \frac{15}{16}(1-u^2)^2 1(\lvert u \rvert \le 1)$	5/7	1/7	0.529
Triweight	$k(u) = \frac{35}{32}(1-u^2)^3 1(\lvert u \rvert \le 1)$	350/429	1/9	0.408
Gaussian	$k(u) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{u^2}{2})$	$1/2\sqrt{\pi}$	1	3.545

2.1.2.1 Bandwidth Selection

Level of smoothing in a kernel smoother is governed by the bandwidth *h.* Various methods used in theoretically determining the bandwidth are available, but none is providing the optimum value in terms of the sensitive issue of bias – variance balance. Hence the establishment of the Kernel density relies on the choice of *h*. The Gaussian Kernel is given by

$$K(u)=\varphi(u)=\frac{1}{\sqrt{2\pi}}e^{-u^2/2}.$$

This is based on the normal density expressed as

$$\mathcal{N}(\mu,\sigma^2)\to f(x)=\frac{1}{\sigma}\varphi\left(\frac{x-\mu}{\sigma}\right).$$

If h_o is the optimum, for its determination, let

$$\int K^2(x)dx=\int\frac{1}{2\pi}e^{-u^2}du=\int\frac{1}{\sqrt{2\pi}}e^{-u^2/2}du.\frac{1}{2\sqrt{\pi}}=\frac{1}{2\sqrt{\pi}};$$

and τ^2 be the second moment of the standard normal distribution $\mathcal{N}(0,1)$. Then

$$\tau^2=1.$$

$$f''(x)=\frac{1}{\sigma^3}\varphi''\left(\frac{x-\mu}{\sigma}\right).$$

Hence,

$$\int(f''(x))^2dx=\frac{1}{\sigma^6}\int\left(\varphi''\left(\frac{x-\mu}{\sigma}\right)\right)^2dx=\frac{1}{\sigma^5}\int(\varphi''(y))^2dy=\frac{3}{8\sqrt{\pi}\sigma^5}$$

therefore,

$$h_0 = \left(\frac{4}{3n}\right)^{1/5} \bullet \hat{\sigma} \cong 1.06\hat{\sigma}n^{-1/5}. \qquad (2.2)$$

Outliers are easily detected from Equation (2.2) Silverman [44], but not a desired case. Instead the interquartile range of the data can be substituted in place of $\hat{\sigma}^2$. Interquartile range is defined as

$$R = X_{0.75} - X_{0.25}.$$

Substituting into equation (2.2),

$$\hat{h}_0 = 0.79\hat{\mathrm{R}}\mathrm{n}^{-1/5} \qquad (2.3)$$

is obtained.

Utilizing equations (2.2) and (2.3) leads to a better estimate for *h*.

$$h_0 = 1.06 \min\left(\hat{\sigma}, \frac{\hat{R}}{1.34}\right) n^{-1/5}.$$

Bandwidth obtained via theoretical approach does not lead to the smoothing. Very small *h* value reduces the bias, and a large *h* value leads to an increase of the variance. Then a trial and error approch is recommended for achieving the optimum ballance between bias – variance. Hence, the minimization of the MSE or Average of the Sum

Squared Errors (ASSE) nececitates a simulation process where the variation of MSA and ASSE values are monitored based on changes given to *h* values. pp 58 – 86 [53].

2.1.2.2 Nadaraya–Watson Estimator

Nonparametric regression smooths a process based on some weighting system. The estimator $\hat{m}_h(x)$ of the smooth average function $m_h(x)$ is estimated

$$\hat{m}_h(x) = n^{-1}\sum_{i=1}^{n} w_{hi}(x)\,\mathrm{y}_i$$

Here y is the response variable, w_h is the weight function determined as a function of distance between *x* and X_i the i^{th} observed value within the bandwidth *h.*

A weighting system is employed in Nadaraya – Watson estimator.

$m_h(x)$ can be expressed as the conditional expectation based on *n* observations coming from i.i.d. r.v.s $\{(X_i, Y_i)\}_{i=1}^{n}$, $X_i \in \Re,\ Y_i \in \Re$, as

$$m(x) = E(Y \mid X = x) = \int yf(x, y)dy \,/\, f_x(x) \qquad (2.4)$$

Kernel density estimator is used to estimate $f(x)$. Multiplicative kernel can be used to estimate the joint density $f(x, y)$ in the numerator of equation (2.4) as

$$\hat{f}_{h_1,h_2}(x, y) = n^{-1}\sum_{i=1}^{n} K_{h_1}(x - X_i)K_{h_2}(y - Y_i)$$

The numerator of equation (2.4) is estimated as follows,

$$\int y\hat{f}_{h_1,h_2}(x,y)dy = n^{-1}\sum_{i=1}^{n} K_{h_1}(x-X_i)\int yK_{h_2}(y-Y_i)dy$$

$$= n^{-1}\sum_{i=1}^{n} K_{h_1}(x-X_i)\int \frac{y}{h_2}K_{h_2}\left(\frac{y-Y_i}{h_2}\right)dy$$

$$= n^{-1}\sum_{i=1}^{n} K_{h_1}(x-X_i)\int (sh_2+Y_i)K(s)ds$$

$$= n^{-1}\sum_{i=1}^{n} K_{h_1}(x-X_i)Y_i \qquad (2.5)$$

pp 120 – 140 [52].

Ratio of the result obtained in equation (2.5) and the kernel estimate of *f(x)* are used to obtain an estimate for the conditional expectation $m(x)$ (2.4). This gives the Nadaraya-Watson estimator as follow;

$$m_h(x) = \frac{\sum_{i=1}^{n} K_h(x-X_i)Y_i}{\sum_{j=1}^{n} K_h(x-X_j)}. \qquad (2.6)$$

The non-parametric regression smoother can be expressed as

$$\hat{m}_h(x) = n^{-1}\sum_{i=1}^{n} W_{hi}(x)Y_i.$$

Where the weights $W_{hi}(x)$ are

$$W_{hi}(x) = \frac{h^{-1}K\left(\frac{x - X_i}{h}\right)}{\hat{f}_h(x)} \qquad (2.7).$$

The whole sample $\{X_j\}_{j=1}^{n}$ plays a major role in determining the weights in equation (2.7) via $\hat{f}_h(x)$.

For sparse X_i higher weights are assigned to Y_i are assigned.

When the denominator is zero, so is the numerator leading to an estimated value zero.

When $h \to 0$, $W_{hi}(x) \to n$ if $x = X_i$. Estimated value X_i converges to Y_i.

When $h \to \infty$, $W_{hi}(x) \to 1 \; \forall x$. Therefore, $m(x) \to \bar{Y}$

Once again it must be stressed that bandwidth *h* plays a major role on the level of smoothness in the estimation process.

2.1.2.3 Mean and Variance of the Nadaraya – Watson Estimator

Numerator and denominator of Nadaraya – Watson estimator can be considered as r.v.s. Hence, for each separate calculations can be done. this statistic are both random variables, they can be dealt with separately. The numerator is,

$$r(x) = \int yf(x, y)dy = m(x)f(x) \text{ and } \hat{r}_h(x) = n^{-1}\sum_{i=1}^{n} K_h(x - X_i)Y_i .$$

Regression estimate is

$$\hat{m}_h(x) = \frac{\hat{r}_h(x)}{\hat{f}_h(x)}.$$

Theorem 2.2: Nadaraya-Watson smoother's numerator $\hat{r}_h(x)$ is asymptotically unbiased.

Proof: $E(\hat{r}_h)$ is

$$E[\hat{r}_h(X)] = E[n^{-1}\sum_{i=1}^{n} K_h(x - X_i)Y_i] = E[K_h(x - X)]Y$$

$$= \int\int yK_h(x-u)f(y|u)f(u)dydu = \int K_h(x-u)f(u)\left(\int yf(y|u)dy\right)du$$

$$= \int K_h(x-u)f(u)(E[Y|X = u])du = \int K_h(x-u)f(u)m(u)du = \int K_h(x-u)\,\mathrm{r}(u)du\ . \quad (2.8)$$

Analogically to the density estimate using kernels, if $r \in C^2$, then

$$E[\hat{r}_h(x)] = r(x) + \frac{h^2}{2} r''(x)\mu_2(YK) + o(h^2)$$

meanig $r_h(x)$ is asymptotically unbiased as $h \to 0$.

Theorem 2.3: Nadaraya-Watson smoother's denominator $\hat{r}_h(x)$ is asymptotically consistent. This can be shown using its variance.

Proof: Let $s^2(x) = E\left[Y^2 \middle| X = x\right]$, then

$$Var\left[\hat{r}_h(x)\right] = Var\left[n^{-1} \sum_{i=1}^{n} K_h(x - X_i) Y_i \right]$$

$$= n^{-1} Var\left[K_h(x - X)Y\right]$$

$$= n^{-1} \left\{ \int K_h^2(x-u) s^2(u) f(u) du - \left(\int K_h(x-u) r(u) du \right)^2 \right\}$$

$$= n^{-1} h^{-1} \int K_h^2 s^2(x+uh) f(x+uh) du + o((nh)^{-1})$$

$$= n^{-1} h^{-1} f(x) s^2(x) \|K\|_2^2 + o((nh)^{-1}) \qquad (nh \to \infty). \qquad (2.9)$$

Putting equations (2.8), (2.9) together when $h \to 0, nh \to \infty$ the MSE of $\hat{r}_h(x)$ becomes

$$MSE\left[\hat{r}_h(x)\right] = \frac{1}{nh} f(x)s^2(x)\left\|K\right\|_2^2 + \frac{h^4}{4}\left(r''(x)\mu_2(K)\right)^2 + o(h^4) + o\left((nh)^{-1}\right).$$

If $nh \to \infty$, $\text{MSE}\left[\hat{r}_h(x)\right] \to 0$. It means $\hat{r}_h(x)$ is a consistent estimator of $r(x)$. That is for any $c > 0$ and $c \to 0$,

$$\lim_{\substack{h \to 0 \\ nh \to \infty}} P\left[\left|\hat{r}_h(x) - r(x)\right| < c\right] = 1.$$

Briefly, $\hat{r}_h(x) \xrightarrow{P} r(x)$.

Theorem 2.4: Nadaraya-Watson smoother's denominator $\hat{f}_h(x)$ is asymptotically unbiased.

Proof: Since $X_i;\ i = 1,\ldots,n$ are i.i.d.

$$E\left[\hat{f}_h(x)\right] = \frac{1}{n}\sum_{i=1}^{n} E\left[K_h(x - X_i)\right]$$

$$= E\left[K_h(x - X)\right]$$

$$= \int K_h(x - u) f(u) du$$

$$= \int K(s) f(x+sh) ds.$$

When $h \to 0$ gives

$$E\left[\hat{f}_h(x) \right] \to f(x) \int K(s) ds = f(x) .$$

For $h \to 0\, h$, $E\left[\hat{f}_h(x) \right]$ is asymptotically unbiased. *Q.E.D.*

Taylor expansion of $f(x+sh)$ in x based on the assumption that f is twice continuously differentiable $\left(f \in C^2 \right)$ can be used to analize the bias..

$$\text{Bias}\left[\hat{f}_h(x) \right] = \int K(s) f(x+sh) ds - f(x)$$

$$= \int K(s) \left[f(x) + shf'(x) + \frac{h^2 s^2}{2} f''(x) + o(h^2) \right] ds - f(x)$$

$$= f(x) + \frac{h^2}{2} f''(x) \mu_2(K) + o(h^2) - f(x). \qquad (2.10)$$

Proof of equation (2.10) see [14].

Because of the symmetry property of K around 0, the term $\int sK(s) hf'(x) ds = 0$. Then the bias of kernel density becomes

$$\text{Bias}\left[\hat{f}_h(x)\right] = \frac{h^2}{2} f''(x)\mu_2(K) + o(h^2)\ ,\ h \to 0 \qquad (2.11).$$

Due care must be given to the bandwidth to avoid large bias values as equation (2.11) contains h^2. It is worth mentioning that bias is proportional to f'' in *x*. As a result $E\left[\hat{f}_h(\bullet)\right] < f(x)$ wile estimating points close to local minimum ($f''(x) > 0$) , and $E\left[\hat{f}_h(\bullet)\right] > f(x)$ when estimated points are around a local maximum ($f''(x) < 0$).

Theorem 2.5: The variance of Nadaraya-Watson smother's denominator $\hat{f}_h(x)$ is used to show that $\hat{f}_h(x)$ asymptotically consistent.

Proof: As $X_i;\ i = 1,\ldots,n$ are i.i.d.

$$\text{Var}\left[\hat{f}_h(x)\right] = n^{-2}\text{Var}\left[\sum_{i=1}^{n} K_h(x - X_i)\right]$$

$$= n^{-2}\sum_{i=1}^{n}\text{Var}\left[K_h(x - X_i)\right]$$

$$= n^{-1}\text{Var}\left[K_h(x - X)\right]$$

$$= n^{-1}\left\{E\left[K_h^2(x - X)\right] - \left(E\left[K_h(x - X)\right]\right)^2\right\}$$

$$= n^{-1}\left\{h^{-2}\int K^2\left(\frac{x-u}{h}\right) f(u)du - \left(f(x)+o(h)\right)^2\right\}$$

$$= n^{-1}\left\{h^{-1}\int K^2(s) f(x+sh)ds - \left(f(x)+o(h)\right)^2\right\}$$

$$= n^{-1}\left\{h^{-1}\|K\|_2^2\left(f(x)+o(h)\right) - \left(f(x)+o(h)\right)^2\right\}.$$

Equation (2.13) leads to $E\left[K_h(x-X)\right] = f(x)+o(h)$ and

$$\int K^2(s) f(x+sh)ds = \int K^2(s)ds\left(f(x)+o(h)\right) = \|K\|_2^2\left(f(x)+o(h)\right)$$

Then

$$\text{Var}\left[\hat{f}_h(x)\right] = (nh)^{-1}\|K\|_2^2 f(x) + o\left((nh)^{-1}\right),\ nh \to \infty. \tag{2.12}$$

Clearly $(nh)^{-1}$ exerts significant influence on variance, resulting in bigger values of *h* and reduced variance. But small *h* value is preferred for lower bias. Combining MSE, the variance and square of the bias of $\hat{f}_h(x)$, as $h \to 0$ and $nh \to \infty$ gives

$$\text{MSE}\left[\hat{f}_h(x)\right] = \frac{1}{nh} f(x)\|K\|_2^2 + \frac{h^4}{4}\left(f''(x)\mu_2(K)\right)^2 + o\left((nh)^{-1}\right) + o(h^4)$$

This means the kernel density estimate is consistent and satisfies $\hat{f}_h(x) \xrightarrow{P} f(x)$. *Q.E.D.*

MSE plays a significant role in balancing variance and bias such that

i. Decreased variance leads to under smoothing.

ii. Decreased bias gives way to over smoothing.

It should be mentioned that using the MSE optimal bandwidth for kernel density can be written as

$$h_0 = \left(\frac{f(x)\|K\|_2^2}{\left(f''(x)\right)^2 \left(\mu_2(K)\right)^2 n} \right)^{1/5}.$$

See reference [22], p 59 [7].

2.1.3 Markov Chain Monte Carlo (MCMC)

In very simple terms Markov chain is some set of random elements $X_1, X_2, \ldots$, where the conditional distribution of X_{n+1} given $X_1, \ldots, X_n$ depends on X_n only. That is

$$f(X_{n+1} | X_1, \ldots, X_n) = f(X_{n+1} | X_n).$$

The *state space* of the Markov chain is defined as the set in which X_i take values. If the conditional distribution of X_{n+1} given X_n does not depend on *n*, then the Markov chain has stationery transition probabilities, on which MCMC concepts are based.

When the state space is finite $\{x_1, x_2, \ldots, x_n\}$, the initial distribution obtains values defined by

$$P(X_1 = x_i) = \lambda_i, \quad i = 1, \ldots, n \rightarrow \boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_n).$$

Transition probabilities forms a matrix **T**, with elements given by

$$P(X_{n+1} = x_j \,|\, X_n = x_i) = t_{ij}, \quad i, j = 1, \ldots, n.$$

A stochastic or random process is defined as a collection $[X_t : t \in T]$ of random variables on a probability space (Ω, F, P). T is time, can be discrete T=[1,2,…] or continuous $T = [0, \infty)$. A stochastic process is said to be *stationary* if for any positive integer *k* the joint probability distribution of *k* random variables $(X_{n+1}, \ldots, X_{n+k})$ does not depend on *n* or does not change under any time shift. That is the stochastic process $\{X_t\}$ and $F_X(x_{t_1+\tau}, \ldots, x_{t_n+\tau})$ be the cumulative distribution of $\{X_t\}$, then $\{X_t\}$ is strictly or strongly stationary if

$$F_X(x_{t_1+\tau}, \ldots, x_{t_n+\tau}) = F_X(x_{t_1}, \ldots, x_{t_n}) \quad \forall \tau, t_1, \ldots, t_n \in \mathbb{R} \text{ and } \forall n \in \mathbb{N}.$$

A Markov chain is also stationary, if it fits into the definition of a stationary stochastic process. That is in a Markov chain $f(X_{n+2}, \ldots X_{n+k} \,|\, X_{n+1})$ does not depend on *n.* It becomes evident that a Markov chain is stationary if $f_{x_n}(x_n)$ does not depend on *n*.

Stationarity of a Markov chain implies stationary transition probabilities, but vice versa is not true.

Ordinary Monte Carlo is a special case of MCMC when $X_1, X_2, \ldots$ are i.i.d., leading to a stationary and reversible MCMC. In the following section the Monte Carlo method is explained in more detail [48], pp 8 – 25 [7].

2.1.3.1 Monte Carlo Method

The Simple Monte Carlo (SMC) method is a special case of Markov chain with $\mathbf{X} = \{X_1, X_2, \ldots\}$ being i.i.d. If $\mathbf{X} = \{X_1, X_2, \ldots\}$ is a stochastic process and g a real valued function on the state space of $\mathbf{X} = \{X_1, X_2, \ldots\}$, then $g(X_1), g(X_2), \ldots$ is a functional of $X_1, X_2, \ldots$ with state space $\mathbb{R}$. The mean and variance of this functional $\mathbf{Y} = g(\mathbf{X})$ are $\mu_{g(\mathbf{X})} = E(g(\mathbf{X}))$ and $\sigma^2_{g(\mathbf{X})} = Var[(g(\mathbf{X}) - \mu_{g(\mathbf{X})})^2]$ respectively. As these are theoretical definition, their computation is not possible when the distribution functions are not available. Instead a random sample from the stochastic process $\mathbf{X} = \{X_1, X_2, \ldots\}$ can be obtained. Then the sample mean of the functional $g(X_1), g(X_2), \ldots$ or the sample mean of Y_i is defined as

$$\hat{\mu} = n^{-1} \sum_{i=1}^{n} g(X_i)$$

$\hat{\mu}$ is called the Monte Carlo approximation of μ .

According to Central Limit Theorem (CLT) the sample mean $\hat{\mu}$ is approximately normally distributed with mean μ and variance σ^2 / n , $\hat{\mu} \approx \mathcal{N}(\mu, \frac{\sigma^2}{n})$.

Similarly, the variance can be estimated by

$$\hat{\sigma}^2 = n^{-1} \sum_{i=1}^{n} (g(X_i) - \hat{\mu})^2.$$

Then $\hat{\sigma} / \sqrt{n}$ is the Monte Carlo Standard Error (MCSE).

Based on the simple statistics it is obvious that the accuracy of estimating μ is inversely proportional to the square root of the sample size. This leads to the difficulty of necessitating very large sample sizes for improving the accuracy of the Monte Carlo method. For instance; ten times increase in the accuracy of the estimate will require an increase of 100 times in the sample size.

The Markov Chain Monte Carlo concept is similar to SMC but the variables involved in the stochastic process $\mathbf{X} = \{X_1, X_2, \ldots\}$ are dependent.

Assume $\mathbf{X} = \{X_1, X_2, \ldots\}$ is a stationary Markov chain, with initial distribution being the same as that of the state space *X*. According to the Markov chain CLT [29] MCMC will be $\hat{\mu} \approx \mathcal{N}(\mu, \frac{\sigma^2}{n})$. , and variance will be

$$\sigma^2 = Var[g(X_i)] + 2 \sum_{k=1}^{\infty} \text{cov}[g(X_i), g(X_{i+k})] \qquad (2.13).$$

For example; let us consider a toy problem where an autoregressive process or Markov chain is given by $X_{n+1} = bX_n + Y_n$ where Y_n are normally distribute with mean 0 and variance θ^2 . r.v. X_1 also should have a finite variance. Then

$$\operatorname{cov}(X_{n+k}, X_n) = b \operatorname{cov}(X_{n+k-1}, X_n) = \cdots$$

$$= b^{k-1} \operatorname{cov}(X_{n+1}, X_n) = b^k \operatorname{cov}(X_n, X_n) = b^n \operatorname{var}(X_n),$$

Under stationary condition

$$\operatorname{var}(X_n) = \operatorname{var}(X_{n+1}) = \operatorname{var}(bX_n + Y_n) = b^2 \operatorname{var}(X_n) + \operatorname{var}(Y_n) = b^2 \operatorname{var}(X_n) + \theta^2$$

$$\operatorname{var}(X_n) - b^2 \operatorname{var}(X_n) = \theta^2 \rightarrow \operatorname{var}(X_n) = \frac{\theta^2}{1-b^2} . \qquad (2.14)$$

Here b^2 must be less than 1 ($b^2 < 1$) as variance cannot be negative. Since the linear combination of independently distributed normal r.v.s is also normal, here we have a linear combination with mean zero and variance given by equation (2.14).

Let $\operatorname{var}(X_n) = \frac{\theta^2}{1-b^2} = \omega^2$. Then the invariant distribution is $\mathcal{N}(0, \omega^2)$.

Then Markov chain follows the CLT as shown below,

$$\sigma^2 = Var[g(X_i)] + 2\sum_{k=1}^{\infty} \text{cov}[g(X_i), g(X_{i+k})] = \frac{\theta^2}{1-b^2}\left(1 + 2\sum_{k=1}^{\infty} b^k\right)$$
$$= \frac{\theta^2}{1-b^2}\left(1 + \frac{2b}{1-b}\right) = \frac{\theta^2}{1-b^2}\left(\frac{1+b}{1-b}\right) = \omega^2 \cdot \frac{1+b}{1-b}. \qquad (2.15)$$

Equation (2.15) shows that the Markov chain given by $X_{n+1} = bX_n + Y_n$ is normally distributed with parameters $\mathcal{N}(0, \omega^2)$ with finite variance.

Here it must be pointed out that the obtained expression for the variance is not a general one, but it is specific to the given Markov chain as a specific example.

It is also visible that as $b \to 1$ equation (2.15) goes to infinity. This is meaningless as σ^2 in general is unknown. But it is a fact that obtaining a close approximation of $\hat{\mu}$ to μ with an error level $\sigma / \sqrt{n} = \varepsilon$ means increasing the sample size. However, there is a limit to which sample size can be increased in a real-life problem.

In the multivariate case the mean vector $\boldsymbol{\mu}$ is approximated by

$$\hat{\boldsymbol{\mu}} \approx \mathcal{N}(\boldsymbol{\mu}, n^{-1}\boldsymbol{\Sigma}).$$

The functional *g(X)* becomes a vector $g(\mathbf{x})$ with components $g_l(\mathbf{x})$. $\boldsymbol{\Sigma}$ is the covariance matrix given as

$$\boldsymbol{\Sigma} = \text{var}\{\text{g}(X_i)\} + 2\sum_{k=1}^{\infty}\text{cov}\{g(X_i, g(X_{i+k})\}. \qquad (2.16).$$

The difference between the equations (2.13) and (2.16) is that, in equation (2.16) $\text{var}(g(X_i)) = \text{cov}\{g_l(X_i), g_m(X_i)\}$ makes up the components of a variance matrix, while $\text{cov}\{g(X_i), g(X_{i+k})\}$ forms a matrix with components $\text{cov}\{g_l(X_i), g_m(X_{i+k})\}$ pp 8 – 11 [7].

2.1.4 Relative Atchison Distance

One other way of assessing the quality of the imputed values or the loss of information is to compute the differences between the imputed and the observed data. This gives a measure of robustness of the imputation process and is named as the Relative Atchison Distance (RDA). It can be used for the comparison of the accuracy of estimates between different imputation methods. Main contribution to the topic comes from Atchison. J (1982). This mainly deals with the Simplex sample space where positive simplex covers the major component or the whole sample space can be defined as

$$S^d = \{(x_1, \ldots, x_d) : x_i > 0 (i = 1, \ldots, d), x_1 + \cdots x_d < 1\} \qquad (2.17)$$

Based on the limitations of S^d , data types that are ratio of two components or data that can be expressed between zero and one are suitable.

Concepts of independence and scarcity of parametric class of distributions on the simplex are still an area of research. However, Atchison has made major contributions towards this goal. Especially in the transformation of normal classes from $\mathbb{R}^d$ to S^d via an appropriate transformation $\mathcal{N}^d(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \rightarrow \mathcal{fN}^d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Some additive logistic transformations are given below.

Table 2.2: Logistic transformations from $\mathbb{R}^d$ to S^d

Type	Specification	Inverse
Additive a_d	$x_i\left\{1+\sum_{i=1}^{d} e^{y_i}\right\} = \begin{cases} e^{y_i} & i=1,\ldots,d \\ 1 & i=d+1 \end{cases}$	$y_i = \ln \dfrac{x_i}{x_d+1}$
Multiplicative m_d	$x_i \prod_{j=1}^{i}\left\{1+e^{y_i}\right\} = \begin{cases} e^{y_i} & i=1,\ldots,d \\ 1 & i=d+1 \end{cases}$	$y_i = \ln \dfrac{x_i}{1-\sum_{j=1}^{i} x_j}$
Hybrid h_d	$x_1 = e^{y_1}/(1+e^{y_1})$ $x_i\left\{1+\sum_{j=1}^{i-1} e^{y_i}\right\}\left\{1+\sum_{j=1}^{i} e^{y_i}\right\} = e^{y_i}, \quad i=2,\ldots,d$ $x_{d+1} = 1 \Big/ \left\{1+\sum_{j=1}^{d} e^{y_j}\right\}$	$y_i = \ln \dfrac{x_1}{1-x_1}$ $y_i = \ln \dfrac{x_i}{\left(1-\sum_{j=1}^{i-1} x_j\right)\left(1-\sum_{j=1}^{i} x_j\right)}, \quad i=1,\ldots,d$

Compositional invariance, and sub - compositional independence concepts are also investigated in this matter [3].

Based on the fundamentals of data compositions [49] the computation of RDA is

$$RDA = \frac{1}{n_M} \sum_{i \in M} d_A(x_i, \hat{x}_i),$$

where $M \subset \{1, \ldots, n\}$, and n_M is the number of cells with missing values in a variable, $d_A(x_i, \hat{x}_i)$ is the Aitchison distance. In this sense the quality or statistical robustness of estimation can be measured by the magnitude of RDA [49], [4].

2.1.5 Analysis of Variance

ANOVA is used in cases when the number of populations to be compared is more than 2 (k>2). That is *k* samples of size *n* will be selected at random, one from each population to be used in testing the hypothesis

$$\begin{aligned} &H_0 : \mu_1 = \mu_2 = \cdots = \mu_k \\ &H_1 : \text{ 2 or more of means not equal} \end{aligned} \qquad (2.18)$$

Then the variation within each sample, and variation between samples play a major role on the results to be obtained from the analysis. Within sample variation depends on chance or it is considered random. Variation between sample means may be due to chance or may also depend on the characteristics of the populations.

Points such as the sample sizes to be used, the size of the variation within samples being large enough to obscure systematic differences are points to be addressed.

Classification of *k* different populations using a single criterion is undertaken in a one-way analysis of variance. Treatment is used to express the various classifications criteria implemented in classifying the populations.

Assumption is that the *k* populations are independent with means $\mu_1, \mu_2, \ldots, \mu_k$ and variance σ^2.

Design of the model to be used in testing the hypothesis is given in Table 2.3

Table 2.3: *k* Random Samples

	Treatment						
	1	2	…	*i*	…	*k*	
	y_{11}	y_{21}	…	y_{i1}	…	y_{k1}	
	y_{12}	y_{22}	…	y_{i2}	…	y_{k2}	
	⋮	⋮	⋮	⋮	⋮	⋮	
	y_{1n}	y_{2n}	…	y_{in}	…	y_{kn}	
Total	$Y_{1.}$	$Y_{2.}$	…	$Y_{i.}$	…	$Y_{k.}$	$Y_{..}$
Mean	$\bar{y}_{1.}$	$\bar{y}_{2.}$	…	$\bar{y}_{i.}$	…	$\bar{y}_{k.}$	$\bar{y}_{..}$

In the table

y_{ij} : Denotes the j^{th} observation from the i^{th} treatment.

$Y_{i\bullet}$: Sum of the data values from the i^{th} treatment.

$\bar{y}_{i\bullet}$: Mean of all observations in the sample from the i^{th} treatment,

$Y_{\bullet\bullet}$: Total of all *nk* observations,

$\bar{y}_{\bullet\bullet}$: mean of all *nk* observations.

2.1.5.1 One-Way ANOVA Model

Each observation can be expressed as

$$Y_{ij} = \mu_i + \varepsilon_{ij}, \qquad (2.19)$$

here ε_{ij} is the deviation of the j^{th} observation of the i^{th} sample from the corresponding treatment mean. Then ε_{ij} is called the random error.

Alternately substituting $\mu_i = \mu + \alpha_i,$ into Eq. (2.19) subject to the constraint

$$\sum_{i=1}^{k} \alpha_i = 0.$$

Gives

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij},$$

where μ is *mean* of all the μ_i, that is, $\mu = \frac{1}{k}\sum_{i=1}^{k}\mu_i$, and α_i is the effect of the i^{th} treatment or the deviation of μ_i from the overall mean μ.

Then the null hypothesis of the *k* population means can be written

$$H_0 : \alpha_1 = \alpha_2 = \ldots = \alpha_k = 0, \qquad (2.20).$$
$$H_1 : \text{Minimum one } \alpha_i \text{ is } \textit{not} \text{ zero.}$$

2.1.5.2 Splitting the Total Variability into Components

The test is based on the comparing of the two estimates of the common variance σ^2. These estimates are required to be independent of each other and obtainable by dividing the total variation into two sections as given below

$$\text{Total variability} = \sum_{i=1}^{k}\sum_{j=1}^{n}\left(y_{ij} - \bar{y}_{..}\right)^2$$

is split into two components as

$$\sum_{i=1}^{k}\sum_{j=1}^{n}\left(y_{ij} - \bar{y}_{..}\right)^2 = n\sum_{i=1}^{k}\left(\bar{y}_{i.} - \bar{y}_{..}\right)^2 + \sum_{i=1}^{k}\sum_{j=1}^{n}\left(y_{ij} - \bar{y}_{i.}\right)^2.$$

For convenience the Sum of Squares can be expressed as:

i) The total sum of squares. $SST = \sum_{i=1}^{k}\sum_{j=1}^{n}\left(y_{ij} - \bar{y}_{..}\right)^2$,

ii) The treatment sum of square (between treatment variation). $SSA = \sum_{i=1}^{k}\left(\bar{y}_{i.} - \bar{y}_{..}\right)^2$,

iii) The error sum of squares (within treatment variation).

$$SSE = \sum_{i=1}^{k}\sum_{j=1}^{n}\left(y_{ij} - \bar{y}_{i\bullet}\right)^2 .$$

Then

$$SST = SSA + SSE .$$

Expected values E(SSA) and E(SSE) are worth investigating.

Theorem 1: $E\left(SSE\right) = \left(k-1\right)\sigma^2$.

Proof: SSE is given as,

$$SSE = \sum_{i=1}^{k}\sum_{j=1}^{n}\left(y_{ij} - \bar{y}_{i\bullet}\right)^2 = \sum_{i=1}^{k}\sum_{j=1}^{n}\left(\varepsilon_{ij} - \bar{\varepsilon}_{i\bullet}\right)^2 = \sum_{i=1}^{k}\left[\sum_{j=1}^{n}\varepsilon_{ij}^2 - n\bar{\varepsilon}_{i\bullet}^2\right].$$

Hence

$$E\left(SSE\right) = \sum_{i=1}^{k}\left[\sum_{j=1}^{n}E(\varepsilon_{ij}^2) - nE(\bar{\varepsilon}_{i\bullet}^2)\right] = \sum_{i=1}^{k}\left[n\sigma^2 - n\frac{\sigma^2}{n}\right] = k\left(n-1\right)\sigma^2 .\ QED.$$

Division by $k(n-1)$ yields

$$E\left[\frac{SSE}{k\left(n-1\right)}\right] = \frac{k\left(n-1\right)\sigma^2}{k\left(n-1\right)} = \sigma^2 .$$

Theorem 2: Show that $E(SSA) = (k-1)\sigma^2 + n\sum_{i=1}^{k}\alpha_i^2$.

$$SSA = \sum_{i=1}^{k}\left(\bar{y}_{i.} - \bar{y}_{..}\right)^2 = n\sum_{i=1}^{k}\bar{y}_{i.}^2 - kn\bar{y}_{..}^2,$$

$$y_{ij} \sim n\left(y; \mu + \alpha_i, \sigma^2\right),$$

hence

$$\bar{y}_{i.} \sim n\left(y; \mu + \alpha_i, \frac{\sigma}{\sqrt{n}}\right) \text{ and } \bar{y}_{..} \sim n\left(y; \mu + \bar{\alpha}, \frac{\sigma}{\sqrt{kn}}\right),$$

then

$$E\left(\bar{y}_{i.}^2\right) = Var\left(\bar{y}_{i.}\right) + \left[E\left(\bar{y}_{i.}\right)\right]^2 = \frac{\sigma^2}{n} + \left(\mu + \alpha_i\right)^2,$$

And because of constraints on α's.

$$E\left(y_{..}^2\right) = \frac{\sigma^2}{kn} + \left(\mu + \bar{\alpha}\right)^2 = \frac{\sigma^2}{kn} + \mu^2,$$

can be written. Therefore,

$$E(SSA) = n\sum_{i=1}^{k} E\left(\bar{y}_{i.}^2\right) - knE\left(\bar{y}_{..}^2\right) = k\sigma^2 + n\sum_{i=1}^{k}\left(\mu + \sigma_i\right)^2 - \left(\sigma^2 + kn\mu^2\right)$$

$$= (k-1)\sigma^2 + n\sum_{i=1}^{k}\sigma_i^2 .\ \textit{Q.E.D.}$$

When H_0 is true

$$s_1^2 = \frac{SSA}{k-1}$$

becomes an estimate of σ^2 with $k-1$ degrees of freedom. This is called the *Treatment Mean Square.*

In equation (2.20) if H_0 is true and then α_is are equal to zero leads to

$$E\left(\frac{SSA}{k-1}\right) = \sigma^2 .$$

Then s_1^2 is an unbiased estimate of σ^2. However, when H_1 is true

$$E\left(\frac{SSA}{k-1}\right) = \sigma^2 + \frac{n}{k-1}\sum_{i=1}^{k}\alpha_i^2,$$

meaning s_1^2 estimates σ^2 and the additional term measures the variation resulting from systematic effects.

Based on $k(n-1)$ degrees of freedom a second estimate of σ^2 is obtained as

$$s^2 = \frac{SSE}{k(n-1)}.$$

This is called the *Error Mean Square*.

2.1.5.3 Use of *F*-Test in ANOVA

The expectations $E(s_1^2)$ and $E(s^2)$ worth commenting on. When H_1 is true and $E(s_1^2) > E(s^2)$ is satisfied a right tailed F test can be used. The error mean square (sample variance) s^2 is an unbiased estimator of the population variance σ^2 and this is not affected by the validity or invalidity of the null hypothesis given in equation (2.19). It can be shown that for ANOVA the mean square error $s^2 = \frac{SSE}{k(n-1)}$ in a one-way classification, is an unbiased estimate of σ^2. It is also worth noting that while the partitioning of SST as $SST = SSA + SSE$ leads to the partitioning of the degrees of freedom as $nk - 1 = k - 1 + k(n-1)$.

When H_0 is true, the ratio $f = s_1^2 / s^2$ is the F value of the random variable with the F distribution having $k-1$ and $k(n-1)$ degrees of freedom.

The P value defined as $P = P\{f[k-1, k(n-1)] > f\}$ can also be used for the decision making about H_0.

Table 2.4: Shows how the analysis of variance computations are summarized

Source of Variation	**Sum of Squares**	**Degrees of Freedom**	**Mean Square**	**Computed**
Treatments	SSA	$k-1$	$s_1^2 = \frac{SSA}{k-1}$	$\frac{s_1^2}{s^2}$
Error	SSE	$k(n-1)$	$s^2 = \frac{SSE}{k(n-1)}$	
Total	SST	$kn-1$		

For the clarification of the point the following example is given.

Example: In order to determine how the mean absorption of moisture in concrete varies, 5 different concrete aggregates were exposed to moisture for 48 hours. 6 samples are tested from each aggregate. Obtained results are given in Table 2.5.

The purpose is to see whether there is a significant difference between the population means based on the sample data.

Table 2.5: Aggregate types to determine how the mean absorption of moisture in concrete varies

Aggregate type						
	1	2	3	4	5	
	551	595	639	417	563	
	457	580	615	449	631	
	450	508	511	517	522	
	731	583	573	438	613	
	499	633	648	415	656	
	632	517	677	555	679	
Total	3320	3416	3663	2791	3664	16854

$\bar{x}_1 = 553.33,\ \bar{x}_2 = 569.33,\ \bar{x}_3 = 610.5,\ \bar{x}_4 = 465.17,\ \bar{x}_5 = 610.67.$

The null and alternative hypothesis are

$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$
$H_1 : 2$ or more of means not equal

Using a significance level $\alpha = 0.05$ the hypothesis will be tested.

For the degrees of freedom $v_1 = k - 1 = 4,\ v_2 = k(n-1) = 25$, the critical f value is 2.76. Using the model sum of squares $s_1^2 = \frac{SSA}{k-1} = \frac{85356}{4} = 21399.1$ and error sum of squares $s^2 = \frac{SSE}{k(n-1)} = \frac{124021}{25} = 4960.8$ are obtained. Then the test statistics is $f = \frac{s_1^2}{s^2} = \frac{21339.1}{4960.8} = 4.3$ leading to the rejection of the null hypothesis. This indicates the sample statistics does not provide sufficient evidence to support the equality of population means. It means at least two of the population means are not equal. To determine which aggregate has the lowest mean, the box and whisker diagram is a good tool.

ANOVA is also used in Chapter 4 to show that the mean of the different estimation methods can not be considered to be the same. Once this fact is accepted based on ANOVA results, then the box and whisker diagrams are produced and given in Figure 4.8. It is evident that the proposed MRSV yields the lowest RDA statistics indicating the methods robustness over the other methods employed.

Chapter 3

SUPPORT VARIABLES

The concept of support variable (SV) is proposed by [47] as part of the research undertaken during the course of this PhD study. The concept involves the inclusion of different variables that are part of the process where missing values are encountered in one variable. The SVs are employed in determining the multivariate regression equation to be used for the imputation of missing values. However, the inclusion of such variables into the estimation process will result in incorrect estimated values due to inhomogeneity of the data coming from SVs that are part of the process. This may be due to different units of different variables, or significant difference in the magnitude of values taken by different variables. Interpretation of the results of analysis obtained from the inclusion of such variables will lead to unforeseen error levels.

To overcome this problem the following should be adhered to:

1. Determine the variables that are closely correlated to the variable with missing values, and ensure that they are part of the same process under study. Use the scatter diagram between the variable with missing values and the presumed SV, compute correlation coefficient between the variable with missing values and each of the SVs to help in deciding the acceptability of a variable as SV.

2. Establish a logical model that will convert the unit of each SV, to the unit of the variable with missing values. This requires a detailed research based on past work, as well as novice proposals with proofs to reach at an acceptable model.

Following these two steps, the SVs are ready to be included into the multivariate regression process.

The concept of SVs is a general one and their determination will entirely depend on the process under study.

In this study the imputation of missing values was considered for the barley (Hordeum vulgare) grain production in tons/hectare (t/ha) over 17 years from 17 different regions of North Cyprus. Grain yield is a complicated multivariate process that starts with germination and ends with harvest. Its accurate modelling is well beyond the scope of the research undertaken in a PhD study, and requires a multidisciplinary approach. However, in this study only an attempt is made to introduce the concept of SV, and in the specific case of barley grain yield only the most obvious variables that have undeniably strong influence on barley grain yield are considered. These are monthly average rain X_1 (mm/m^2), monthly average temperature X_2 ($^{\circ}C$), and soil organic matter ratio X_3 (unitless). Since the units of the SVs are not in t/ha, a careful study of the literature related with this topic is undertaken to determine a conversion algorithm for each SV into t/ha.

3.1 Converting Rain to Grain Equivalent Yield

Without rain life in the way that it exists on Earth would be impossible. Therefore, rain is one of the essential components for the existence of plant and animal life forms. Hence, one of the main factors to be considered for the determination of grain yield is rain. It has a direct and major influence on the yield expected from any plant type.

Here, an attempt is 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 average grain yield in t/ha.

A significant contribution comes from Cantero et. al. [8] from their study of two types of barley cultivars. They compared their yields under prevailing soil structure, and climatic conditions by considering as many variables as possible that influence the yield. 180 kg/ha of barley was used in sowing, 45 kg/ha P_2O_2 and K_2O, and 100 kg/ha nitrogen (N) fertilizer was applied during sowing. Following careful assessment of input water (rain water) during the period sowing to harvesting, water use considered as equivalent to evapotranspiration, is formulated as

$$ET = (w_n - w_{n-i}) + P - D \pm R \qquad (3.1)$$

where *ET*: evapotranspiration, *w:* volumetric water content, *P*: rainfall, *D*: Drainage below 120 cm, *R*: Surface run off. According to Eq. (3.1) any drop of the *ET* value below the wilting point till a few weeks before the crop is ready for harvesting, or at least till the end of the grain filling period, is undesirable. If this happens plants will lose water to the point that even extra rain will not result in rejuvenation of the plant, meaning loss of production. Therefore, the concept "*water use efficiency* (WUE)"

becomes important. This is defined by Cantero et. al. [8] as *grain yield produced per unit area per unit of water evapotranspiration by the crop (kg/ha/mm)*. Water used by the plant for the period from germination to harvest is considered to be equal to the evapotranspiration during this period. Based on research carried out by Lopez and Arrue [32] conventional tillage ploughing 30 to 40 cm depth and cereal – fallow rotation type cropping tends to maximize the water retained in soil as moisture, hence contributing in achieving higher grain yield. Traditionally this is the kind of cereal production method (cereal – fallow rotation) used in Cyprus for decades, and it is the current practice also supported by state subsidies.

For the purpose of this study the conversion of rain data (mm/m^2) to grain yield in t/ha is computed by applying the proposed average values to the rain data, since the climatic conditions where Cantero et. al. [8], and Lopez and Arrue [32] carried out their research are very similar to those in Cyprus. Time from October to April, the period from germination to harvest is taken as basis for the computations. Water use efficiency figures for grain (WUE_g) given by Cantero et. al. [8] and Lopez [32] are considered to be highly representative of conditions in Cyprus, having an average value $WUE_g = 8$ kg/ha/mm. Average evapotranspiration for the study area for the months November to April is taken as 188 *mm.* Tandoğdu & Camgöz [46]. This is equivalent to 55% of annual average precipitation (*AAP*) for the area $(E_{tr} = 0.55)$. Then the rain equivalent yield (*REY*) in t/ha can be computed by $REY = AAP \cdot E_{tr} \cdot WUE_g / 1000$.

3.2 Temperature Equivalent Grain Yield

Among the main factors adversely affecting grain yield are drought combined with heat stress.

Hossain et.al. [23], carried out a set of experiments with four spring barley and two spring wheat genotypes under early, optimal and late sowing conditions in a southern arid region of Russia to determine wheat and barley genotypes suitable for prevailing climatic conditions as well as optimum sowing time. It is reported that this part of Russia has a semi-arid climate similar to the Mediterranean climate. High temperature results in deficiency of soil moisture resulting in the loss of grain yield. Similarly, low temperature also affects germination and stand establishment of crop sown early resulting in lower grain yield [34].

In a study to assess the grain yield the following important factors are generally considering.

i. Adjusted moisture content. In general, an ideal moisture content of 12% is assumed for grain yield assessment. For prevailing different moisture contents, the following formula is used for adjustment [24];

$$y(m_2) = \frac{100 - m_1}{100 - m_2} y(m_1) \qquad (3.2)$$

where $y(m_1)$ grain weight at current moisture level. $y(m_2)$ grain weight at ideal moisture level. m_1 current moisture level. m_2 ideal moisture level.

Equation (3.2) gives an idea on how close the current moisture level is to the ideal or expected moisture level.

ii. Harvest index. It is defined as the ratio of grain yield to grain yield plus straw yield [12],

$$HI = \frac{Grain\ yield}{Grain\ yield + Straw\ yield} \qquad (3.3)$$

Harvest index given in Eq. (3.3) can be used in comparing the grain yield of different genotypes.

iii. Relative Yield Performance is a useful tool to compare the heat performance of different genotypes [5]. It is given by

$$RYP = \frac{Heat\ stress\ performance}{Optimum\ performance} \qquad (3.4)$$

RYP can be used as a measure of grain yield to be expected, depending on the heat stress performance of a particular genotype.

iv. Stress susceptibility index is defined as the ratio of minimum yield under tolerable stress conditions to yield under favourable conditions [18].

$$SSI = \frac{1 - y / y_s}{1 - x / x_s} \qquad (3.5)$$

where y_s is mean yield under stressed conditions for one genotype, y is mean yield under stress free conditions for the same genotype, x_s mean yield for all studied genotypes under stressed conditions, and x mean yield for all studied genotypes under stress free conditions. The tolerance of a genotype to stress (high temperature and drought) is measured by its SSI. If SSI<0.5 very tolerant, 0.5<SSI<1 moderately tolerant, and SSI>1 poorly tolerant. Hence, Eq. (3.5) gives an idea about the heat stress tolerance of particular genotype.

Hossain et.al. (2012) reports that grain yield loss ranges between 43 to 82% for crops subjected to high heat and drought stress compared to crop grown under optimum conditions.

According to a study by Nahar et.al. [34] carried out in Bangladesh, heat stress caused between 53% to 73% loss in 5 different types of wheat grain. This is a clear indication of the importance of heat stress on grain yield.

On the other hand, Hasan [21] reported that about 2.6% to 5.8% reduction in grain yield was observed in heat tolerant and 7.2% reduction in heat sensitive genotype for every 1°C temperature above the average air temperature assumed for normal growing conditions during anthesis to maturity. An analysis of the findings of a study by Karahan. T., and Sabancı. C. O. (2010) on 8 different barley genotypes in Diyarbakır and Ceylanpınarı areas in the South Eastern part of Turkey, with similar climatic conditions to those in the study area of North Cyprus, has also led to around 11% reduction in grain yield per 1°C increase in air temperature.

Based on research done on the matter and given the local semi-arid climatic conditions prevailing in the study area, it is estimated 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. To develop a factor to be used in converting heat stress into equivalent yield the following logic is proposed.

Let $x_i;\ i=1,\ldots,n$ be the annual grain yield (t/ha), n the number of production years.

$\bar{t}_i$ the average temperature for the grain maturing period (March, April) for the i^{th} year. For the area of study this is shown in Figure 3.1.

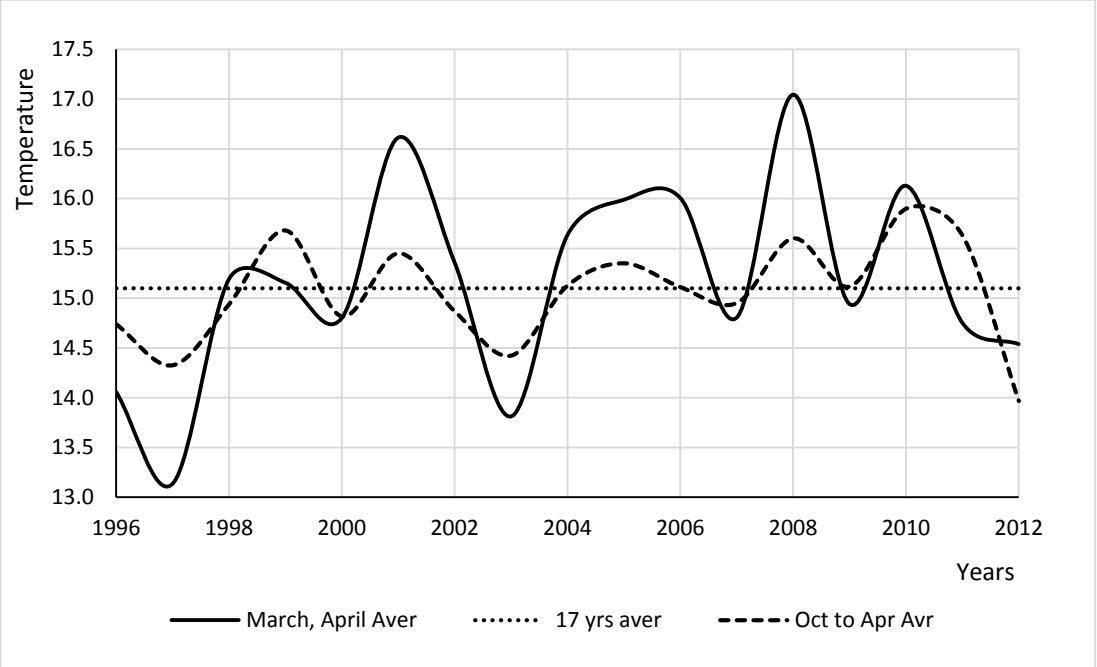

Figure 3.1: Average temperature for the grain maturing period (March, April), and for the period October to March the period from germination to maturing in the area of study.

Annual averages from October to April, and $\bar{t}$ the overall average temperature for the 17 years from October to April (the period from germination to harvest) are also seen in Figure 3.1. When the barley grain yield versus March – April average temperatures is examined, it turned out that for temperature above the long-term average of 15 ^0C a negative correlation of -0.52 is observed. This confirmed the idea of grin maturing period high temperature has a negative effect on grain yield. Figure 3.2 shows this clearly.

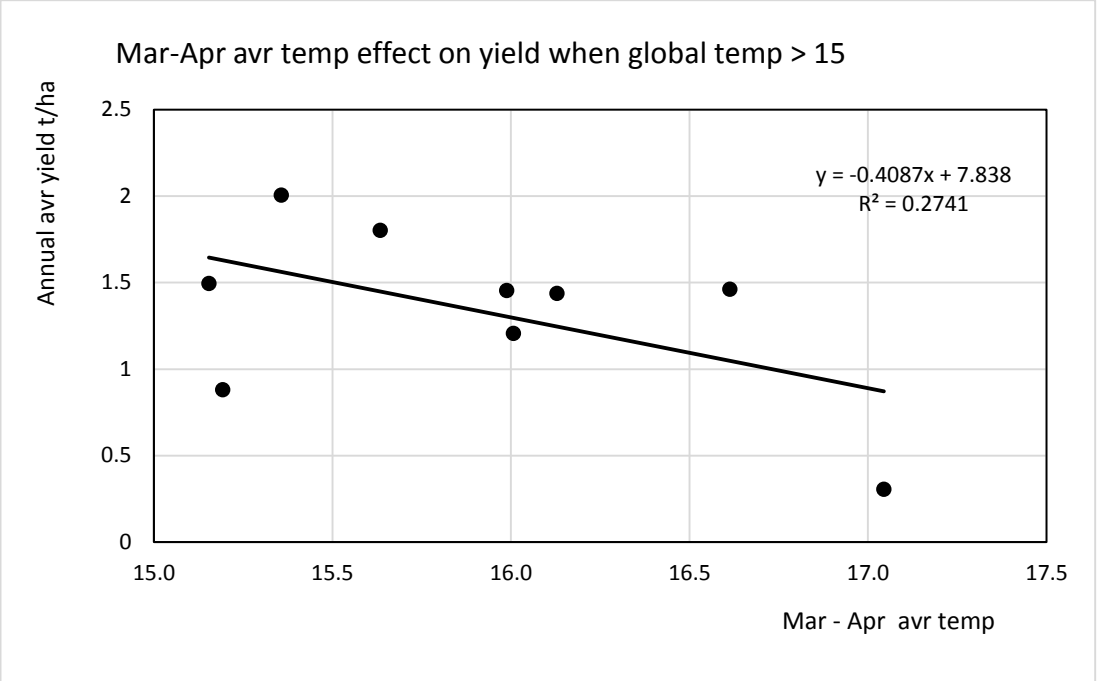

Figure 3.2: Adverse effect of high temperature during grain maturing period that is above the long-term global average from germination to harvest.

While fluctuations in daily weekly or monthly temperature has some effect on grain yield, in the local conditions it is expected that a drop of a few degrees $^{\circ}C$ in the average temperature during the grain maturing period to below the long-term average from germination to harvest period does not have a major effect on the grain yield under optimum conditions (GYOC). On the other hand, if the average temperature for grain maturing period is above the long-term average will result in some loss on the grain yield. Hence, the temperature equivalent grain yield (TEGY) for a certain year, is proposed to be

$$TEGY = \frac{\overline{t}}{\overline{t_i}}(GYOC)$$

A higher than long term average temperature for the maturing period ($\bar{t} / \bar{t}_i < 1$) will result in some loss in grain yield. The amount of loss obviously depends on $\bar{t} / \bar{t}_i$ ratio. On the other hand, for $\bar{t} / \bar{t}_i \geq 1$ it is assumed that grain yield will not be affected when the average temperature during the maturing period is below the long-term average temperature for the periods from germination to harvest. The scatter diagram, in Figure 3.3 supports this idea, where annual average temperature during the maturing period is drawn against the average annual grain yield. Linear correlation coefficient is also between the two variables at 0.14 is very low indicating no correlation between the two variables. That is grain yield is not affected by temperatures during grain maturing period when it is below the germination to harvest average temperature.

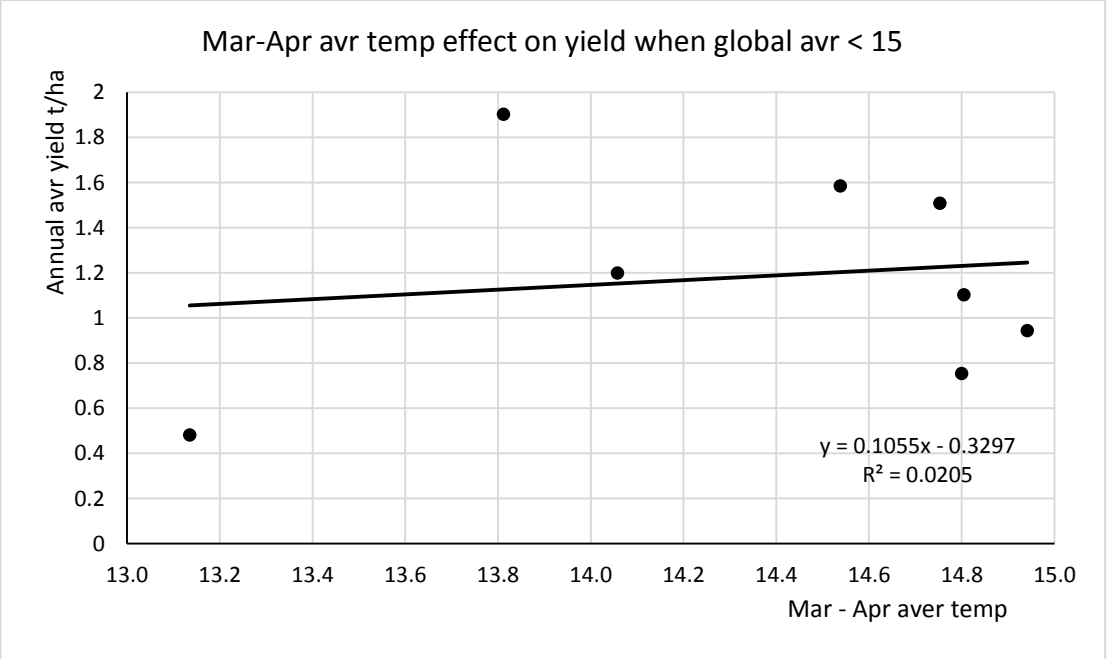

Figure 3.3: No significant effect on grain yield when the temperature during grain maturing period is below the germination to harvest average temperature.

3.3 Equivalent grain yield as a function of soil organic matter

Fertility of any soil type is depending on organic matter contained in the soil [50], [39]. This is measured by the Soil Organic Matter (SOM) which is defined as the fraction of the soil consisting of plant and/or animal tissue in various stages of breakdown or decomposition levels. Most productive agricultural soils have a SOM value between 3 and 6% organic matter.

One way of measuring the SOM value of soil is the method called Loss on Ignition (LOI). In this method the soil is exposed to 105 ^0C for 1.5 hours to remove soil moisture. The weight of the de-moisturized soil is recorded and then the temperature is increased 500 ^0C for 2 hours. Difference between the final weight and the de-moisturized weights is used to determine LOI giving the SOM value [17].

SOM has significant benefits for grain yield, which can be classified into Physical, Chemical, and Biological categories. Physical benefits include factors that improve water infiltration, water holding capacity, reducing run off, soil aeration, reducing surface crusting.

Some of the chemical benefits can be listed as increased ability to hold essential nutrients such as calcium, magnesium, and potassium. It also includes the creation of resistance to changes in the pH level of soil, contributing to the decomposition of nutrients in minerals.

Biological benefits are providing food for living organisms in the soil, suppressing disease and pests via enhanced soil microbial biodiversity, increasing the pore space thus contributing to increased infiltration.

Soil Organic Matter Ratio (SOMR) is defined as the ratio of SOM to clay + silt content. SOMR can be used as an indicator for soil conditions that limits grain yield. Based on a study carried out in the semi-arid Pampa region in Argentina with varying SOMR values over 3 years Quiroga et. al. [36] found that the contribution of SOMR to grain yield can be as high as 51%. However, when only SOM is considered, its contribution to grain yield is about 32%. This indicates that SOMR is a better indictor to be used in measuring the grain yield.

Stine. M. A and Weil R. R. [45] studied the relation between grain yield and soil productivity indicators under 3 different tillage methods. Through standard soil tests the pH level of soil, availability of Phosphorus (P), Potassium (K), Magnesium (Mg), Calcium (Ca), total nitrogen (N), and active carbon (C) levels, as well as porosity, aggregate stability of the soil was determined. Chemical elements that are useful part of the SOM were found to be in a linear relation with productivity or grain yield. Based on the data and graphs presented in their studies it is estimated that 1% increase in SOM value results in between 1.7% and 2.8% increase in grain yield.

According to Johnston [27] a long-term program started in 1852 to determine SOM values as part of the Hoosfield Continuous Barley experiment at Rothamsted USA on a silty clay loam soil applying Nitrogen Phosphorus Potassium (NPK) fertilizer or Farmyard manure (FYM) at 35t/ha resulted in SOM values of 1.74% and 6.16%

respectively. Starting from 1968 different plots were treated with 0, 48, 96, and 144 kg/ha N to determine the effect of N on grain yield. Monitoring grain yield results at 1.74% and 6.16% SOM levels for 3 different barley genotypes for the periods 1976 – 1979, 1988 – 1991, and 1996 – 1999 periods have shown consistent increase in yield and always higher in the high SOM level. On average soils with high SOM value have an average yield of 2.5 t/ha more than that of the soil with lower SOM value.

It is also observed that in the soils where FYM is used, the addition of N above 96 kg/ha did not result in an increase in the grain yield. However, the use of some N is necessary to enable the roots to absorb the nutrients more efficiently. This means some N has to be added to compensate for the loss of N from soil due many different factors, a topic still open for investigation.

In a different experiment Johnston [27] highlights the relation of SOM level and Phosphorus (P) to grain yield was studied. Following 12 years of preparation on a silty clay loam soil to achieve 1.5% and 2.4% SOM levels, for two years for each SOM level 24 different P levels were used and grain yield determined. The study has shown that with no N application, the spring and winter barley, coupled with low and high organic matter ratio resulted in an average grain yield of 2.78 t/ha. Combined low and high organic matter ratio average is recorded as 2.35%. Using linear interpolation, this corresponds to 1.18 t/ha increase in grain yield, for a 1% increase in organic matter ratio. Then for any SOMR value *p*, corresponding grain yield can simply be computed as

$$Gran\ yield = 1.18p.$$

In a study by Derici et. al. [11] covering all areas cultivated by barley or wheat genotypes, the sub areas can be considered as homogenous in terms of SOMR. Considering the fact that a significant change in the SOMAR values in an area may take decades or even centuries, it is assumed that SOMR values for the study area and for the study period are constant. Then major changes in the SOMAR values over short periods of a few years or even a few decades are not expected.

Chapter 4

APPLICATION

4.1 Geography and Climate of Cyprus

Geographically Cyprus is situated in the North-Eastern corner of the Mediterranean Sea. It lies between $32^{\circ}\ 15' - 34^{\circ}\ 35'$ east longitudes and $34^{\circ}\ 33' - 35^{\circ}\ 41'$ north latitudes. Total area of Cyprus 9251 km^2. Out of the total area of the island of Cyprus, 3298 km^2 forms the Northern Cyprus (TRNC). Of the total area of North Cyprus, 20% is declared as forest areas and not used for agricultural activity of any kind, with some exceptions subject to special permission of the Department of Agriculture. Area where agricultural activity of all kinds takes place is about 57% of the area. Remaining 23% is mostly residential, industrial, and other public facilities, as well as areas that are currently not used or not suitable to be used for any kind of activity.

Along the northern part of the country runs the Pentadactylos Mountain Range (Five Fingers Range) rising to an altitude of just over 1024 m at Selvili Tepe. The mountain range loses altitude beyond the Kantara Castle but extends eastward to from the backbone of the Pan-Handle shaped peninsula called Karpas. In the central and southern part of the island lies the Trodos Range rising to an altitude of 1951 m at Mount Olympus. Between the two mountain ranges lies the Mesarya (Mesaoria) plain of which together with narrow alluvial plains along the coast form up the total bulk of the agricultural land of the country. The northern part of the island of Cyprus in which

the area of study is located, includes the Pentadactylos Mountain Range, the Mesarya plain, the Northern coastal plain and the Karpas peninsula.

Cyprus has a warm and dry Mediterranean climate with the main rainy period being November to March, during which barley and wheat agriculture takes place. Long dry summers from mid–May to mid–September, preceded by a short spring and followed by a short autumn seasons in climatic sense. Amount of rain and temperature varies with altitude and distance from the coast [19].

Summers are generally under the influence of low-pressure trough extending from western Asia leading to high temperatures. In winter the island is on the path of fairly frequent depressions crossing the Mediterranean from the west mainly emanating from the continental anticyclone of Eurasia and low pressure from Africa. These atmospheric activities result in short periods of a few days of rainy spells resulting an overall annual average of 500 mm for the whole of the island. However, amount of rain varies significantly depending on altitude and distance from coast. Annual average rain is above 1000 mm/m^2 in the Trodos mountains, around 650 mm/m^2 in the Beşparmak (Pentadaktilos) range, and a mere 350 mm/m^2 in the Mesarya (Mesaoria) plains. The overall annual average rain profile for the whole area of study is given in Figure 4.1.

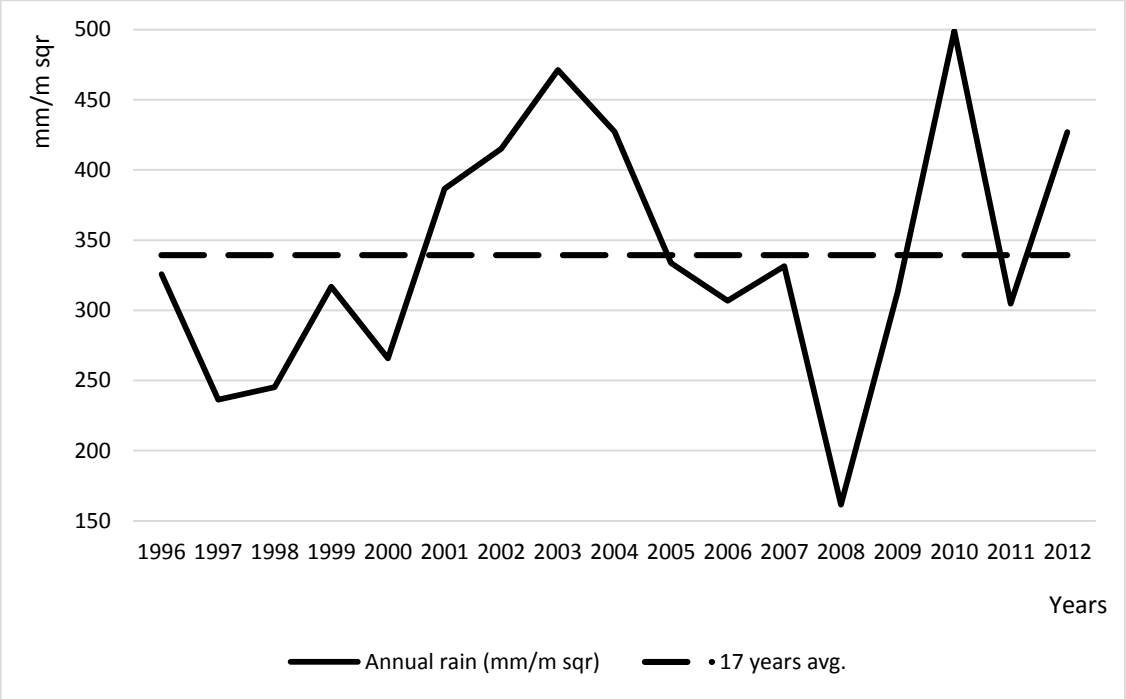

Figure 4.1: Annual average rain and overall average for the 17 years in the area of study.

Annual average temperature for the area of study is given in Figure 4.2, where years of high temperature has a negative effect of barley grain yield.

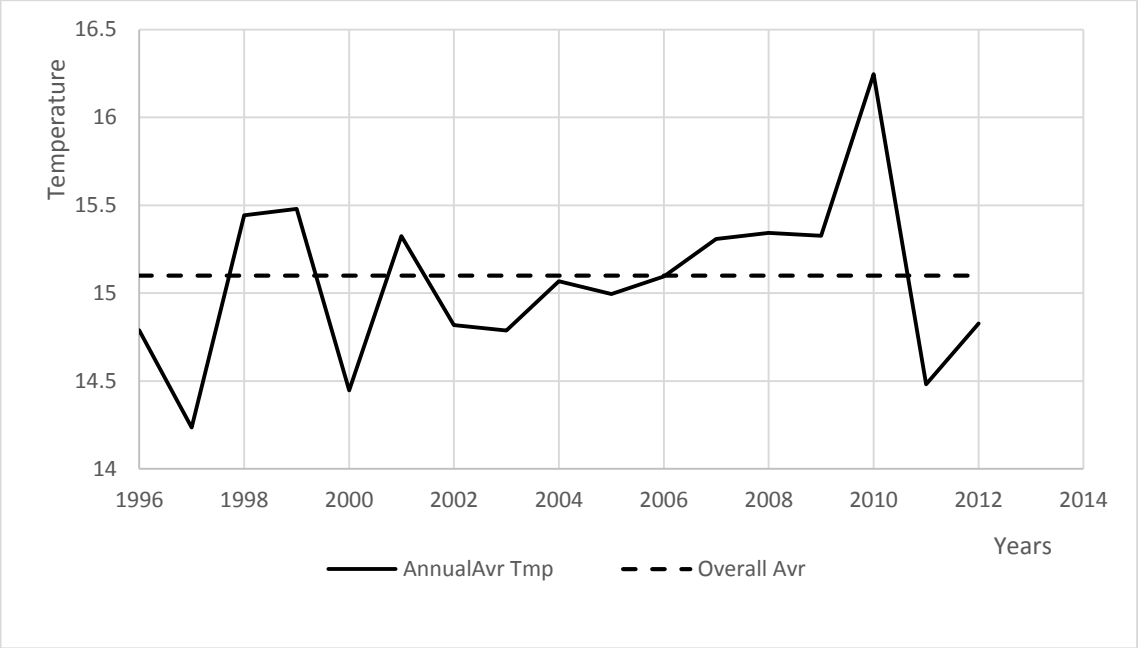

Figure 4.2: Annual average temperature and overall average for the 17 years in the area of study.

With hot summers and mild winter, temperatures are rather variable depending on elevation and distance from the coast. Temperature difference between mid-summer and mid-winter is in the order of 18 ^0C inland and about 14 ^0C on the coasts. Daily difference between maximum and minimum temperature in the agricultural areas in winter is around 9^0C and in summer about 16^0C. The average minimum temperature in December, the coldest month of the year is 11.4 ^0C, while the average maximum temperature 30.7^0C in July.

Capital city Nicosia geographically is almost in the center of the Mesarya plain. Hence, for the period of study rain values are graphed as seen in Figures 4.3. This is on average representative for the main central part of the Mesarya plain where more than half of barley and wheat production of North Cyprus takes place.

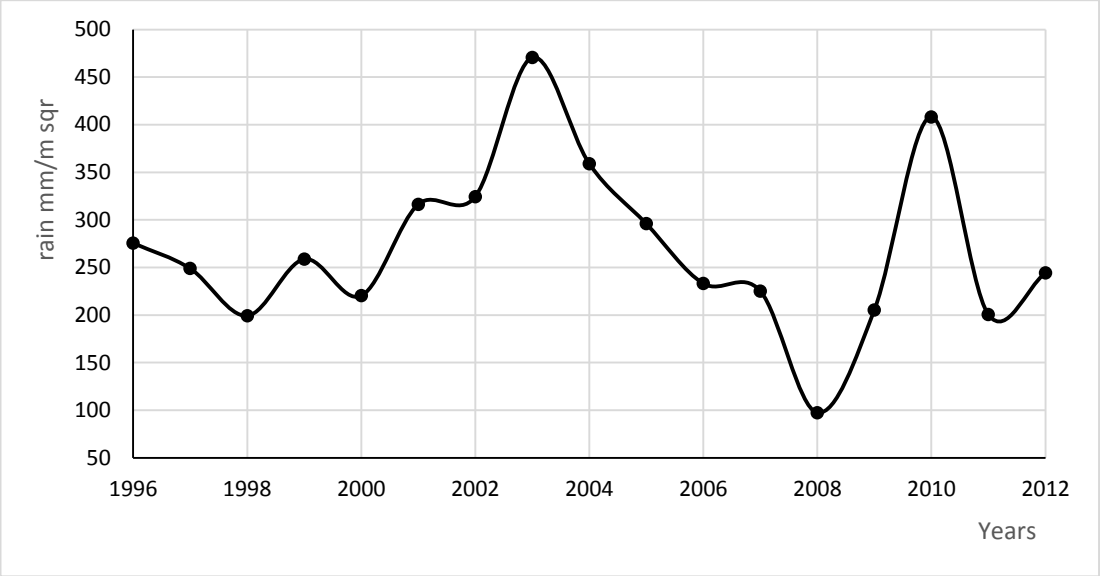

Figure 4.3: Annual average rain profile from October to April for Nicosia area from 1996 to 2012.

Annual average temperature for the Nicosia area representing central Mesarya for the study period is shown in Figure 4.4.

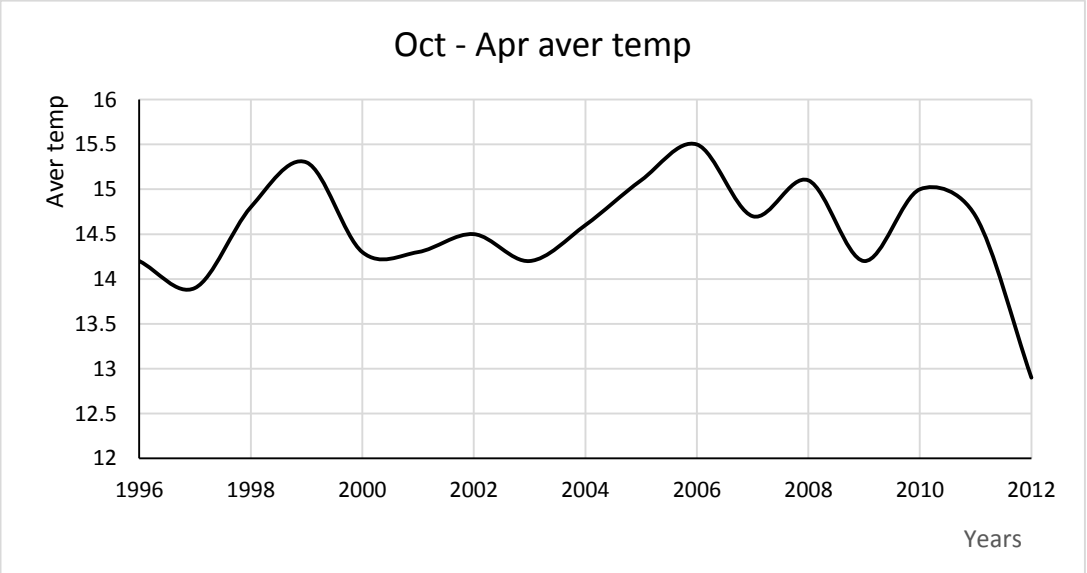

Figure 4.4: Average temperature for central Mesarya for the 7 months period from germination to harvest.

4.2 Review of the data

The use of support variables (SV) in the imputation of missing values concept is explained in Chapter 3. For the application of SV concept, it was considered necessary to first find a data set that is complete, and randomly cancel certain percentage of the data to obtain a new data set with missing values. This will enable the determination of the error element following the imputation process carried out based on different imputation methods. Records of the Department of Agriculture were considered a good source of data for this purpose, as they are keeping crop production records on annual basis.

One agricultural product that has significant impact on the economy of North Cyprus is barley. Whole production is utilized as animal feed in livestock or husbandry sector. Considering the average annual value of barley being around 11 million USD, is a good indication why barley is selected for this study. Therefore, it was decided to use the annual barley production figures in t/ha as raw data.

Agricultural land in Northern part of Cyprus from where the data is collected is divided into 17 production areas by the DA, based on administrative structure of DA.

Production areas with their geographic name are given in Table 4.1. Boundaries of the areas are shown in Figure 4.5.

Table 4.1: Geographic names of production areas

1	Lefkoşa Merkez	10	Girne Batı
2	Değirmenlik	11	Boğaz
3	Ercan	12	Çamlıbel
4	Gazi Mağusa A	13	Güzelyurt
5	Gazi Mağusa B	14	Lefke
6	Akdoğan	15	YeniİskeIe
7	Geçitkale	16	Mehmetçik
8	Gönendere	17	Yeni Erenköy
9	Girne Doğu		

Overall area of the 17 regions amounts to a total of 1880 km^2 used for agricultural activities. Average area where barley is grown is 592313 donum that is equivalent to 78825 hectares, using the conversion factor 1 Donum =0.13308 hectare.

Barley yield data in t/ha for the 17 regions covering the years from 1996 to 2012 are given in Appendix A. The 17×17 square data matrix represents a total of 289 data values, where rows represents production areas and columns are the years.

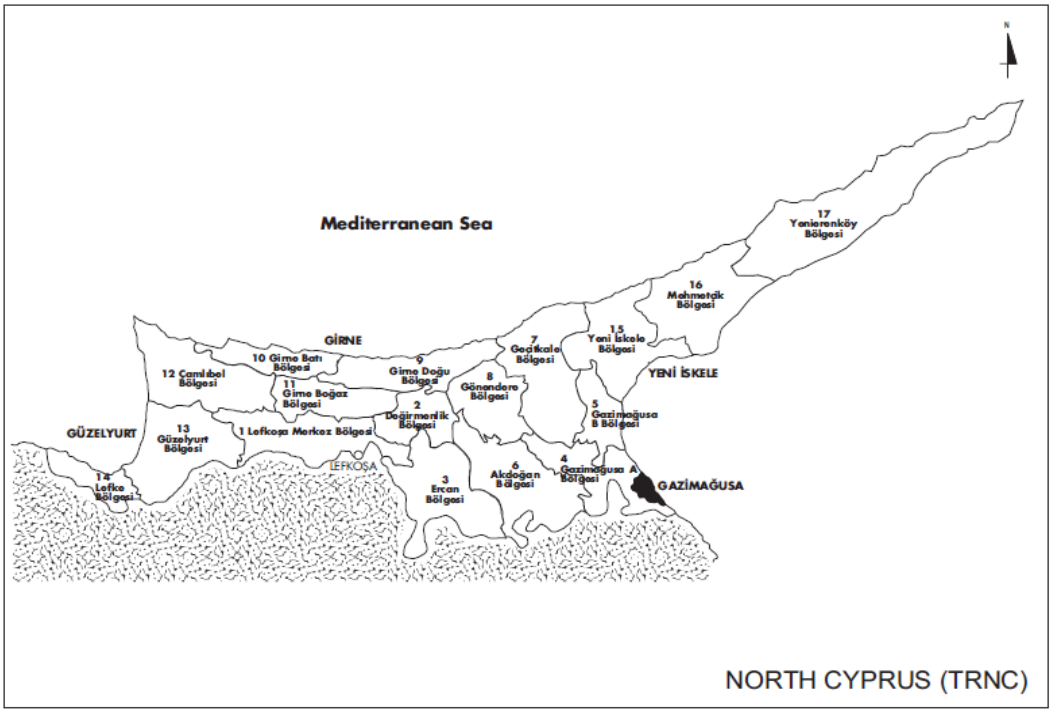

Figure 4.5: Map of North Cyprus showing the boundaries of 17 production areas or regions.

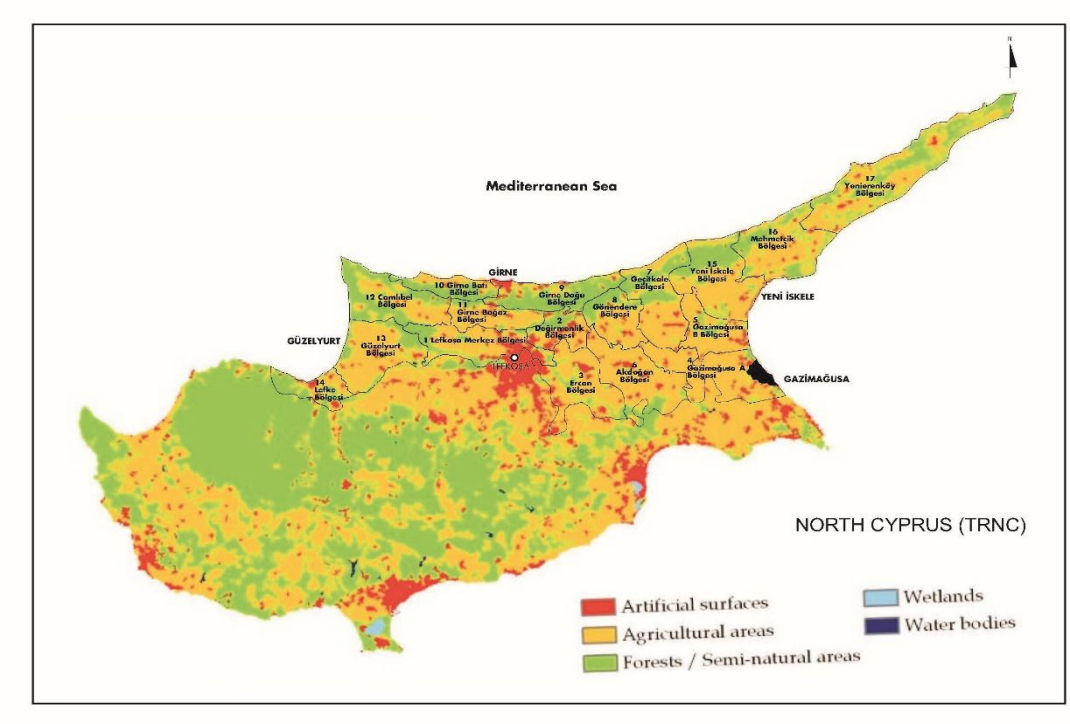

Figure 4.6: Land cover map of Cyprus showing the boundaries of 17 production areas or regions.

In addition to the barley data set, data related with rain, temperature and soil organic matter for the same period and production areas had to be collected, since these were selected to be used as support variables in the imputation process.

Annual average rainfall figures obtained from the department of Meteorology [33]. In Appendix C annual average temperature data for the last 17 years is presented, that is obtained from the Statistical Information Office. Appendix D shows the soil organic matter ratio data according to the production areas and years [11].

The rain between November and April (germination to harvest period) has a major impact on barley yield, as explained under section 3.1. Hence the average rainfall values for November to April are given in Appendix B.

4.3 Computations

In this study the imputation of missing data values using some existing methods are compared with the proposed support variables concept. For this end the complete data set on barley production given in Appendix A consisting a total of 289 data values was used to generate two new tables with missing values. Data values from Appendix A were deleted at random to create two new data sets with 10% and 40% missing values. That is 260 data values in the 10% missing case represented by the matrix $\mathbf{X}_1^m$, and 173 data values in the 40% missing case denoted by $\mathbf{X}_2^m$ were retained for analysis. See Appendices H and I.

As part of the proposed support variables, in the process of barley production rain (mm/m^2), temperature in ${}^0C$, and soil organic matter ratio (unitless), were considered as the most important variables and used as SVs. Raw data related with these variables are presented in Appendix B, Appendix C, and Appendix D respectively.

4.3.1 Converting units of the support variables to that of the raw data (t/ha)

Conversion of data on S.Vs to the same unit as the data with missing values was undertaken, as explained in Chapter 3 for each S.V. Raw rain data converted to equivalent t/ha is given in Appendix E. Computation of t/ha values is done according to logic explained under section 3.1 and using the proposed conversion equation.

Temperature related raw data from Appendix C is converted into t/ha following the logic given in section 3.2 employing the proposed formula $TEGY = \frac{\bar{t}}{\bar{t}_i}(GYOC)$.

Converted temperature data from ${}^0C$ to t/ha are shown in Appendix F. Raw data

related to soil organic matter (see Appendix D) also converted to t/ha according to the logic given under section 3.3, and given in Appendix G.

4.3.2 General approach followed in the imputation process

Following the formation of data files with 10% and 40% missing values, and the conversion of support variables data into t/ha in units according to methodologies explained in Chapter 3. Bivariate linear regression, Multivariate Linear Regression Employing Support Variables (MRSV), kernel regression, and Markov Chain Monte Carlo (MCMC) methods using available data only to predict missing values are used and results compared. In each case error levels were computed by using the complete data set and the imputed values. Mean absolute error, mean square error and root mean square deviation values are used as measures to compare the performance of each imputation method.

In principal the column with minimum number of missing data values is taken as dependent variable, and imputation is carried out based on the equation determined by the method used for this purpose. Then, the next column with minimum number of missing data values is processed accordingly. Once the imputation process is completed with the individual method in application, determination of error levels was carried out.

4.3.3 Simple Linear Regression.

Average figures for the 17 years and for each production area was taken as independent (X) and each year's figures where missing values exists taken as dependent or response variable (Y). Obtained simple linear regression equation is used to estimate one of the missing values for that year and production area, then the corresponding average for that area recalculated. If there are more missing values under that year, the process is

repeated until all missing values for that year are imputed. Then same steps followed for the imputation of missing values for all years and production areas.

4.3.4 Multivariate Linear Regression

Multivariate linear regression is applied when p response variables Y_i depend on k predictor variables $X_1,\ldots,X_k$. The multivariate regression for the i^{th} variable is then expressed as $y_i = b_0 + b_1 x_{1i} + \cdots + b_k x_{ki} + e_i$, where $b_i;\ i = 0,\ldots,k$ are regression constants and $e_i;\ i = 1,\ldots,p$ are the random errors that are identically distributed with mean zero and variance σ^2. Minimization of Σe_i^2 results in a set of equations from where of constants $b_0,\ b_1,\cdots,\ b_k$ can be computed.

In $\mathbf{X}_1^m$ and $\mathbf{X}_2^m$ the matrices representing the 10% and 40% missing values, the existing value are denoted by x_{ij}, $i = 1,\ldots,n\ \ j = 1,\ldots,p$ and the missing values denoted by x_{ij}^m, $i = 1,\ldots,n\ \ j = 1,\ldots,p$. In both matrices the indices i and j have the same range respectively, but observed values denoted by x, missing values denoted by x^m, while their positions in are defined by i and j. Theoretically imputation of missing values can either be done on column by column or row by row basis. This may lead to some differences in the imputed values. However, the process under study may require either of these approaches. In the case of the barley yield example it is wise to start imputation based on column wise approach, as the support variables, especially rain and temperature are susceptible to annual climatic changes, and row by row approach will result in larger fluctuations in the data values, reflecting on the estimates as well.

Column by column imputation can be performed using regression as follows.

1. Determine the l^{th} column ($1 \le l \le p$) in which the minimum number of missing values (k_l) is.
2. Missing values in column l are $x_i^m;\ i = 1,\ldots,k_l$ 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*. Do not include the values of the l^{th} column. Obtained averages forms the values of the first independent random variable X_{1a}. Assume the l^{th} column is the dependent variable X_l. Then $n-k_l$ tuples can be formed 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, the variables X_{2a}, X_{3a}, X_{4a}, X_5, X_6 can be defined 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.

Inclusion of the variables X_5 and X_6 in the regression process will provide useful information in the imputation process as they represent available data for the year imputation is undertaken.

5. Undertake multivariate regression for the column *l*

   $$x_l = b_0 + b_1 x_{1a} + b_2 x_{2a} + b_3 x_{3a} + b_4 x_{4a} + b_5 x_5 + b_6 x_6 .$$

6. Estimate the first missing value and impute in the l^{th} column. Then $k_l - 1$ missing values remains in the l^{th} column.

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

8. Then repeat steps 1 to 7, for each column till all missing values in all columns are imputed.

Following this algorithm imputation of missing values were undertaken and obtained results for the 40% missing data are given in Appendices J, K and N.

For the comparison of error levels for different imputation methods, expressing the Mean Square Error (MSE) as a percentage of the row data average is considered as a suitable method.

Mean square error for one row or column of the data matrix (MSER) computed from *k* estimated and corresponding observed values is given by

$$MSER = \sum_{i=1}^{k_l} \frac{(x_i - \hat{x}_i)^2}{k_l}$$

Expressing MSER as a percentage of the row or column average $RCA = \frac{\sum_{i=1}^{n} x_i^2}{n}$ (n: Number of observations in a row or column), is given as

$$MSE\% = \left(\frac{MSER}{RCA} \right) 100 .$$

For the whole data that is made up of *n* rows and *p* colums, the overall MSE% can be written as

$$MSE\% = n^{-1} \sum_{i=1}^{n} \left(\frac{MSER}{RCA} \right)_i 100$$

This enables the fair comparison of MSE values of different variables.

For the 10% missing data imputed values are given in Appendix L. In the case of 10% missing data, estimation errors are found to be very low (see Appendices M and O). The overall average MSE% obtained from 5 splits in 10% missing case using MRSV turned out to be 1.24% indicating the superiority of MRSV. Due to very low error levels subsequent work concentrated on the 40% missing case. This is done to assess the performance of the proposed imputation method MRSV when the missing value are a high percentage of overall data.

4.3.5 Using kernel regression for imputation

Kernel regression is a nonparametric method where local weights are used in the process of estimation. An estimate of a certain point is done by taking a linear combination of neighbouring observations. Estimation in kernel regression is undertaken using the Nadaraya – Watson estimator $\hat{g}(x) = \Sigma_{i=1}^{n} w_i k(u) / \Sigma_{i=1}^{n} k(u)$ where w_i represents local weights and *k* the kernel function. *u* which is defined as $u = (x_i - X)/h$, *h* being the band width. The weights *w* is estimated such that $SSE = \Sigma(x_i - \hat{x}_i)^2$ is minimum. On the other hand, the level of smoothing is an important point and as explained in Chapter 2, Section 2.6, it depends on the size of bandwidth used. Kernel values at data point *X* and within its neighbourhood, such that

on the left and right of *X* at locations x_i, at distances equivalent to some multiples of a small increment *dx* are taken and kernel values at these points are computed. Over smoothing occurs if the bandwidth *h* is too big obscuring the structure under study. If *h* is too small, then not appreciable smoothing will take place, not giving any idea about the variable representing the process. Then the optimum bandwidth has to be determined such that over or under smoothing does not occur. There are various formulae proposed by different researchers, amongst which $h = \left(\frac{4s^5}{3n} \right)^{1/5}$ is given in [44] p. 45 for the Gaussian kernel has some application areas, but never satisfactory. In various references including [9], [44] p.53 - 54, [38] pp. 96-98, and [22] p. 254, it is suggested that the cross validation method be used for fine tuning of the theoretically computed band width. This is based on optimizing the *h* value such that the average error is minimum.

In this study the Epanechniko $k(u) = 0.75(1-u^2)$; $|u| \le 1$, and Gaussian $k(u) = (2\pi)^{-1/2} \exp(-u^2/2)$ kernel functions are used.

Epanechnikov and Gaussian kernel estimation techniques were applied to the 10% and 40% missing cases with band width values 2, 4, and 6. Then MSE% values computed and given in Table 4.1 below for comparison. It is clearly evident that Epanechnikov kernel produced estimates with lower error levels than the Gaussian kernel method for the %10 missing data matrix. On the %40 missing data matrix Epanechnikov kernel produced estimates with lower error levels than the Gaussian kernel method for bandwidth values 2 and 6. For bandwidth value 4 Gaussian kernel estimates have lower

error levels than the Epanechnikov kernel. This may happen when working with data, but in general Epanechnikov performed better than the Gausian method.

Table 4.2: 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

4.3.6 Application of Markov Chain Monte Carlo Technique to imputation of missing values

Markov Chain Monte Carlo concept assumes that the conditional distribution of X_{n+1} given X_n does not depend on *n,* and has stationery transition probabilities. In application this leads to an iterative simulation concept for imputation. Conditioning information is used for the estimation of unknown parameters.

Assume X represents the data set with missing values, where missing values are denoted by X^m, observed values by X^o, and θ is the set of parameters to be estimated.

Then the posterior distribution of the parameter values conditioned on the observed data is given by [43] as

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

Here $f(\theta|x^o, x^m)$ 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.

An iterative approach is proposed by Tanner [48] which considers the imputed values together with the observed values for updating the distribution parameter θ. At the beginning the posterior values to be imputed are determined using some suitable algorithm considering the observed data.

The below given steps are followed in imputing missing values using MCMC.

i. Determining a suitable *dx* value to be used as an increment or decrement starting from an observed value say x^o_{i-1} neighboring a location with missing value x^m_i until the next observed value x^o_{i+1} is reached. This generates dummy data values between the two existing data values, in between which one or more locations with missing values exists.

ii. In a row or column for all locations with missing values x^m_i are subjected to this process. On the other hand the difference between two cells

neighboring a x_i^m cell may not be the same for all x_i^m cells, leading to different number of iterations for different x_i^m locations.

iii. Following the generation of dummy data values as explained in steps i and ii then the average of this row or column is computed and compared with the average of the corresponding row or column from the complete data. The row or column with imputed values that has an average closest to the average of the complete data of the same row or column is then selected as the imputed row or column.

The process given in steps i to iii is repeated for each row or column until the imputation process is complete.

Imputation results obtained from MCMC method are given in Appendix R. MSE% error levels obtained from imputation carried out using 5 different methods are summarized in Table 4.3. Clearly the proposed MRSV outperformed all other methods, indicating the importance of support variables in the estimation or imputation process.

Table 4.3: MSE% values obtained in different methods.

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

4.3.7 The Relative Aitchison Distance (RDA)

RDA is methodology that determines the robustness of estimates as explained under section 2.1.4. Hence can be used as a measure to compare of the accuracy of the estimates obtained from different estimation methods. RDA can be computed using the equation proposed by Templ M. et.al. [49].

$$RDA = \frac{1}{n_M} \sum_{i \in M} d_A(x_i, \hat{x}_i)$$

where $M \subset \{1, \ldots, n\}$, n_M is the number of locations with missing values in a variable, $d_A(x_i, \hat{x}_i)$ is the Aitchison distance.

The lower the RDA value the better or more robust is the estimator. RDA values computed for each of the methods used in this study are given in Table 4.3, Appendix P and also shown in Figure 4.7 Again

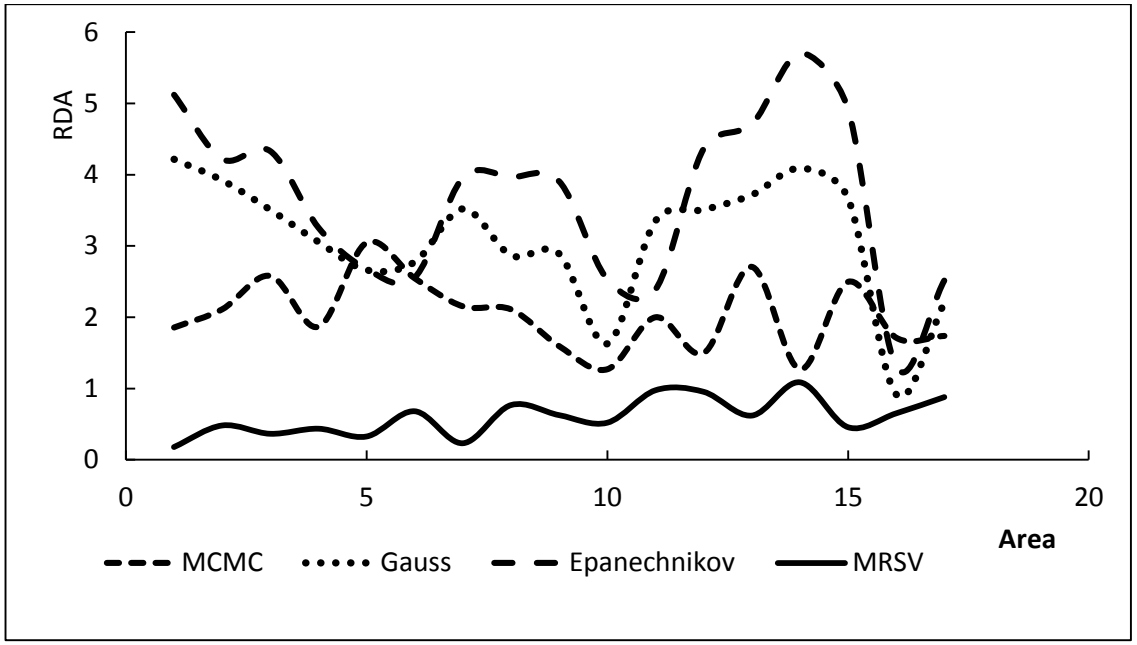

Figure 4.7 RDA performance of the methods used in imputation.

From Figure 4.7 it is clearly evident that the MRSV method performs better than the other methods as its RDA graph is much below the others.

The significance of the difference between the RDA values obtained from different methods can be checked using ANOVA and the F test. Mean and standard deviation values for RDA obtained from different methods are given in Table 4.4.

Table 4.4: Average and standard deviation of the RDA values obtained from each estimation method.

	MCMC Method 1	Gauss-Krnl h2 Method 2	Epan Krnl Method 3	MRSV Method 4
Mean	2.0343	3.0882	3.6726	0.6026
Std. Dev.	0.5111	0.8772	1.1873	0.2658

Based on these RDA sample statistics the null and alternative hypothesis set as below are to be tested at 0.05 significance level.

$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$
$H_1 : 2$ or more of the means not equal

As within group variance $s_1^2 = 30.867$ is much greater than between groups variance $s^2 = 0.6278$ an upper or right-hand tailed F test is appropriate.

Critical f value with $\alpha = 0.05,\ v_1 = k-1 = 3,\ v_2 = k(n-1) = 64 \rightarrow f \cong 2.75$.

Test statistics f value $s_1^2 / s^2 = 30.867 / 0.6278 = 49.17$ is much greater that the critical f value leading to the rejection of H_0. It means the mean RDA values cannot be accepted as being the same for all estimation methods.

To show that the proposed MRSV method's RDA values are better than other methods, a glance at the box and whisker diagrams given in Figure 4.8 is enough.

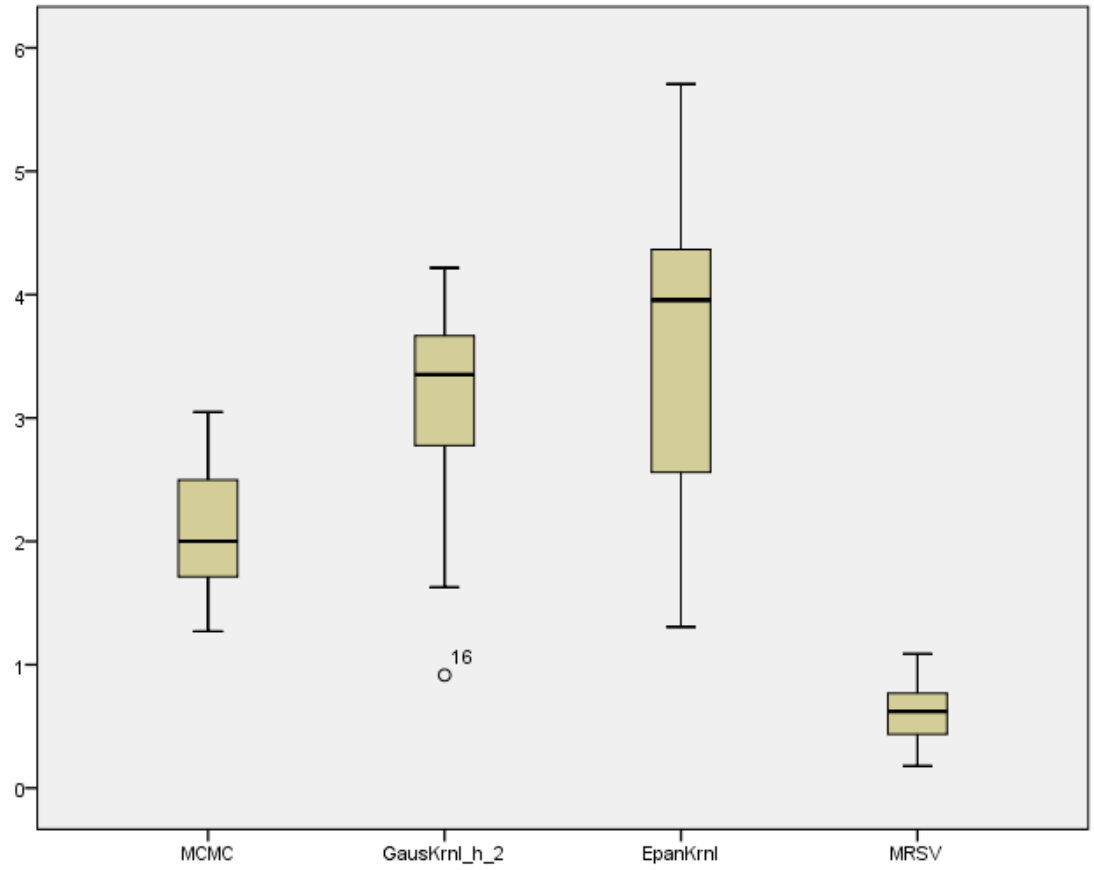

Figure 4.8: Box and whisker diagrams of RDA values for the estimation methods used in this study.

From Figure 4.8 it is evident that the proposed MRSV method upper limit for outlier values is below the lower limit for outliers of other methods. This is a good indication of the robustness of the MRSV method.

Chapter 5

CONCLUSION

Imputation of missing values is handled from a different perspective, such that the estimation of missing values in a certain variable is supported by other variables that are closely related with this variable. The concept developed is thus called imputation using support variables (SVs). In a data matrix of size $n \times p$ with missing values, imputation is carried out starting with the column with minimum number of missing values. In addition to the SVs other variables as explained under section 4.3.4 are included into the multivariate regression for the estimation of missing values (MRSV). This approach proved to be a better method for the imputation of missing values, compared with other methods implemented (bivariate regression, kernel regression, MCMC) in terms of errors (MSE%). The robustness of estimates measured using the relative Aitchison distance where the proposed method proved to be superior to other methods. When the RDA values for each method are used in an ANOVA test, the proposed MRSV method was the best performer.

It must be stressed that the units of the support variables to be used in the estimation of the missing values, must be converted to the same unit as the variable with missing values. This is a formidable task to be overcome and require the clear understanding of the process under study. Any ill designed algorithm for the conversion of the units will certainly result in high error margins in the estimation process.

Strengths of the proposed MRSV method compared with other methods used in this study is worth mentioning.

i. The MCMC method has two main disadvantages, namely time consuming, and obtaining optimum solution is not guaranteed.
ii. In kernel estimation uncertainty on the optimum value of bandwidth necessitates some kind of iterative simulation.
iii. In the MRSV method the only difficulty relates to the conversion of the units of SVs to that of the dependent variable.

The MRSV method handles imputation problems very efficiently, mainly due to the support variables introduced into the system. This is evident when the error levels are compared.

Considering the broadness in content and detail of implementing the support variables concept in imputation of missing values, there remains a vast field of research to find a generalized approach on how to handle the support variables. The methodology used in this study from the agricultural sector proved to be a success, providing encouragement for the application of the concept in different fields.

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] Aitchison, J. (1982). *The Statistical Analysis of Compositional Data*. Journal of the Royal Statistical Society. Series B (Methodological), Vol. 44, No. 2: 139-177.

[4] Aitchison, J., Barcel´o-Vidal, C., Mart´ın-Fern´andez, J. A. and Pawlowsky-Glahn, V. (2000). *Logratio Analysis and Compositional Distance*. Mathematical Geology, Vol. 32, No. 3.

[5] Asana, R. D., Williams, R. F. (1965). *The effect of temperature stress on grain development in wheat.* Aust. J. Agric. Res. **16**:1-3.

[6] Anderson, T. W. (1957). *Maximum likelihood estimates for a multivariate normal distribution when some observations are missing.* J. Amer. Stat. Assoc., **52**: 200-203

[7] Broks, S., Gelman, A. Jones, G. L. and Meng, X. L. edt. (2011). *Handbook of Markov Chain Monte Carlo*. Chapman & Hall.

[8] Cantero-Martinez, C., Villar, J. M., Romagosa, I., Fereres, E. (1995). *Growth and yield responses of two contrasting barley cultivars in a Mediterranean environment.* Eur. J. of Agronomy. **4**(3): 317-326.

[9] Chiu, S. T. (1991). *Bandwidth selection for kernel density estimation.* The Annals of Statistics. **19**(4): 1883 – 1905.

[10] 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'économétrie Faculté des sciences économiques et sociales Université de Genève.

[11] Derici, M. R., Kapur, S. A., Kaya, Z., Gök, M., Ortaş, İ. (Ed.) (2000,2002). *Kuzey Kıbrıs Türk Cumhriyeti detaylı toprak etüd ve haritalama projesi* (Cilt 1-2). Lefkoşa: KKTC Başbakanlık Devlet Basımevi.

[12] Donald, C. M. (1962*). In search of yield*. J. Aust. Inst. Agric. Sci. **238**:171-178.

[13] Ebrahimian, H., Playan, E. (2014). *Optimum management of furrow fertigation to maximize water and fertilizer application efficiency and uniformity*. J. Agr. Sci. Tech., **16**: 591 – 607.

[14] Edgett, G. L. (1956). Multiple regression with missing observations among the independent variables. J. Amer. Statist. Ass **51**: 122-131.

[15] Erbilen, U. S., Sahin, G. (2011). *Development of Greenhouse cultivations and its problems in T.R.N.C.*. ZfWT Vol. 3 , No. **3**: 197-219.

[16] Fasli, M., Riza, M., Erbilen, M. (2016). *The Assessment and Impact of Shopping Centres: Case study Lemar, Northern Cyprus*, Open House International, Vol. 41 No: 4.

[17] Fenton, M., Albers, C., Ketterings, Q. (2008). *Department of Crop and Soil Sciences*. College of Agriculture and Life Sciences, Cornell University, Fact Sheet 41.

[18] Fischer, R. A., Maurer, R. (1978). *Drought resistance in spring wheat cultivars. I. Grain yield responses.* Aust. J. Agric. Res. 29:897-912.

[19] Gönengil, B., Çavuş, E. (2006). *Kuzey Kıbrıs Türk Cuhuriyeti'nin iklimi*. İstanbul: Elçi Basimevi.

[20] Hansen, B. E. (2014). *Econometrics*. University of Wisconsin, Department of Economics, This Revision: January 3, 2014.

[21] Hasan (2002). *Physiological changes in wheat under late planting heat stress*. M.S. thesis. Dept. of Crop Botany. Bangabandhu Sheikh Mujibur Rahman Agricultural University. Salna.

[22] Haerdle, W. (2004). *Applied Nonparametric Regression. Economic Society Monographs.* Humboldt-Universitat zu Berlin, Wirtschaftswissenschaftliche Fakultat, Institut für Statistik und Ökonometrie, Spandauer Str. 1, D-10178 Berlin.

[23] Hossain, A., Teixeira da Silvia, J. A., Lozovskaya, M. V., Zvolinsky, V. P., 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.* Journal of Plant Breeding and Crop Science **4**(11): 184 -196.

[24] Hellevang, K. J. (1995). *Grain moisture content effects and management* . Department of Agricultural and Biosystems Engineering, North Dakota State University. Available online at:http://www.ag.ndsu.edu/pubs/plantsci/crops/ae905w.htm.

[25] Jinubala, V. and Lawrance, R. (2016). *Analysis of missing data and imputation on Agriculture data using predictive mean matching method*. Int. J. of Sci. and App. Inf. Tech., **5**(1): 1-4.

[26] Johnson, R. A., Wichern, D. W. (2007). *Applied multivariate statiscical analysis.* Pearson, PP 360 - 370.

[27] 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 Nutrition Inst. **95**(4): 9 -11.

[28] Karahan, T. Sabancı, C.O. (2010). *Güneydoğu anadolu ekolojik koşullarında bazı arpa (hordeum vulgare L.) çeşitlerinin verim ve verim öğelerinin belirlenmesi*. Batı Akdeniz Tarımsal Araştırma Enstitüsü Derim Dergisi, 27(1):1-11.

[29] Kipnis, C. and Varadhan, S. R. S (1986). *Central limit theorem for additive functionals of reversible Markov processes and applications to simple exclusions*. Communications in Mathematical Physics, V 104, No 1, 1-19 .

[30] Little, J. A. (1992). *Regression with missing X's: A review*. J. of the American Stat. Assoc., **87** No. 420, 1227 – 1237.

[31] Lopez, M., Arrue, J. (1997). *Growth, yield and water use efficiency of winter barley in response to conservation tillage in a semi-arid region of Spain*. Soil and Tillage Research, 44, 35–54.

[32] Lopez, M. V, 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).

[33] Meteoroloji Dairesi (2014). *KKTC Meteorolojık bilgi dairesi*. Lefkoşa: MTD.0.00.M9-14/038.

[34] Nahar, K., Ahamed, K. U., 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.

[35] Qin, J., Zhang B., and Leung H. Y. (2009*). Empirical likelihood in missing data problems.* J. of the American Stat. Assoc., **104** No. 488: 1492 – 1502.

[36] Quiroga, A., Funaro, D., Noellemeyer, E., Peinemann, N. (2005). *Barley yield response to soli organic matter and texture in the Pampas of Argentina.* Soil and Tillage Research 90: 63 – 68.

[37] Rao, C. R., Toutenberg, H. (1999). *Linear Models*: *Least square and alternatives.* 2nd. Edition . Spiringer. PP 241 - 248.

[38] Ramsey, J. O. .and Silverman, B. W. (2006), 2nd Edition. *Functional Data Analysis.* Springer, PP 96 – 98.

[39] Reeves, D. (1997). *The role of soil organic matter in maintaining soil quality in continuous cropping systems*. Soil Tillage Res. 43, 131– 167.

[40] Rubbin D. B. (1976). *Inference and missing data*. Biometrica, 63, 581 – 592.

[41] 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. of the American Stat. Assoc., 108 No. 501: 81 – 95.

[42] Samarah, N. H. (2005). *Effects of drought stress on growth and yield of barley*. Agro. for Sust. Dev. 25: 145-149.

[43] Schunk, D. (2008).*A Markov chain Monte Carlo algorithm for multiple imputation in large surveys*. American Statistical Association. Advances in Statistical Analysis, 92: 101 – 114.

[44] Silverman, B. W. (1998). *Density estimation for statistics and data analysis*. Chapman & Hall, Monographs on Statistics and Applied Probability.

[45] Stine, M. A. and R.R. Weil (2002). *The relationship between soil quality and crop productivity across three tillage systems in south central Honduras*. American J. of Alter. Agri., 17 **(**1): 2 – 8.

[46] Tandoğdu, Y., Camgöz, T. O. (1999). *An experimental approach for estimating evapotranspiration*. CIM Bulletin 92: 55-60.

[47] Tandoğdu, Y., Erbilen, M. (2018). *Imputing Missing Values using Support Variables with Application to Barley Grain Yield*. J. Agr. Sci. Tech. Vol. 20: 829-839.

[48] Tanner, M. A. and Wong, W.H. (1987). *The calculation of posterior distributions by data augmentation*. J. of the American Stat. Assoc. **82** No. 398: 528 – 540.

[49] 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

[50] Tiessen, H., Cuevas, E. Chacon, P. (1994). *The role of soil organic matter in sustaining soil fertility*. Nature 371, 783–785.

[51] 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*. Annals of Econ. and Fin. **6:** 289-301.

[52] Trawinski, I. M. and Bargmann, R. E. (1964). *Maximum likelihood estimation with incomplete multivariate data*. Annals of Math. Stat., **35:** 647-657.

[53] Walpole, E. R., Myers, H. R., Myers, L. S. and Ye K. (2012). 9th Edition. *Probability & Statistics for Engineers & Scientists*. Prentice Hall, pp: 507- 518.

[54] Wand, M. P. and Jones, M. C. (1995). *Kenel Smooting*. Springer, pp 58 – 86.

[55] Yorgancıoğlu, G. (1998). *Kıbrıs* Coğrafyası (Fiziki). İstanbul: Seçil Basimevi.

[56] Yozgatlıgil, C., Aslan, S., Iyigun, C., Batmaz, I. (2013). *Comparison of missing value imputation methods in time series: the case of Turkish meteorological data*. Theor. and App. Clim., **112**: 143–167.

[57] Zamar ,R. H. and Maronna ,R. A. (2002). *Robust Multivariate Estimates for High-Dimensional Datasets*. Technometrics **44**, 307–317.

[58] Zhang X., Song X., Wang H. and Zhang H. (2008). *Sequential local least squares imputation estimating missing value of microarray data*. Comp. in Biol. and Med. **38**: 1112 – 1120.

APPENDICES

Appendix A: Barley Yield t/ha

Area	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	AreaAver
1	1.121	0.448	0.374	1.868	0.149	1.308	1.831	1.868	1.868	1.682	1.390	1.345	0.112	0.867	1.413	1.555	1.592	**1.223**
2	1.001	0.448	0.448	1.868	0.149	1.898	1.824	2.025	2.138	1.756	1.345	1.345	0.142	0.755	1.898	1.868	1.106	**1.295**
3	1.166	0.553	0.874	1.868	0.149	2.182	1.719	1.525	2.227	1.682	1.345	0.897	0.120	0.755	1.129	1.143	1.143	**1.205**
4	1.248	0.471	0.336	0.792	0.538	0.972	2.220	2.093	1.861	1.390	1.121	0.747	0.179	0.583	1.495	1.868	1.868	**1.164**
5	1.248	0.456	0.650	0.879	0.411	0.972	2.434	1.876	1.562	0.904	1.001	1.016	0.247	0.247	1.472	1.495	1.883	**1.103**
6	0.807	0.478	0.336	0.673	0.318	1.001	1.867	2.280	1.846	1.510	1.181	1.353	0.284	0.426	1.712	1.854	1.981	**1.171**
7	0.755	0.516	0.762	0.785	0.164	1.592	2.960	2.294	2.317	1.121	1.031	1.016	0.105	0.695	0.889	1.375	1.861	**1.191**
8	0.770	0.433	0.747	1.226	0.433	1.353	2.982	2.287	2.309	1.121	0.112	0.598	0.082	0.695	1.487	1.495	1.943	**1.181**
9	1.308	1.114	1.353	1.854	1.570	1.644	1.704	1.413	1.465	1.405	1.196	1.495	0.450	0.859	1.667	1.495	1.315	**1.371**
10	1.495	1.129	1.495	1.988	1.868	1.300	1.868	1.495	1.868	1.868	1.166	1.312	0.392	1.734	2.078	1.495	1.495	**1.532**
11	0.800	0.037	0.792	1.868	0.673	1.891	1.487	2.242	2.317	2.242	2.145	2.317	0.508	1.801	2.332	2.227	1.532	**1.601**
12	1.061	0.404	1.330	1.868	1.084	1.517	1.868	1.868	1.278	1.308	1.308	1.129	0.359	0.957	1.061	1.129	1.218	**1.220**
13	0.845	0.120	0.568	1.868	0.419	1.854	1.712	2.175	1.487	1.188	1.166	0.859	0.426	1.360	1.338	0.777	1.637	**1.165**
14	1.248	0.366	1.129	2.377	1.704	1.196	1.547	2.235	2.033	1.659	1.076	0.904	0.433	1.428	1.450	1.001	2.182	**1.410**
15	1.495	0.396	0.852	1.061	0.531	1.278	2.123	1.405	1.188	1.106	1.286	0.486	0.202	0.389	0.889	1.868	1.928	**1.087**
16	1.779	0.277	1.136	1.353	0.972	1.532	2.623	1.764	1.495	1.360	1.129	0.897	0.538	1.263	0.949	1.868	1.129	**1.298**
17	2.250	0.194	1.786	1.203	1.674	1.368	1.323	1.495	1.383	1.405	1.495	1.024	0.605	1.241	1.166	1.121	1.121	**1.286**
YrlyAver	**1.200**	**0.461**	**0.881**	**1.494**	**0.753**	**1.462**	**2.005**	**1.902**	**1.803**	**1.453**	**1.205**	**1.102**	**0.305**	**0.944**	**1.437**	**1.508**	**1.584**	1.265

Appendix B: Raw Rain Data in mm / m^2

Area	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	AreaAver
1	275.6	248.9	199.3	258.8	220.3	316.4	324.5	470.6	359.0	296.0	233.2	226.2	97.5	205.3	408.2	200.6	244.2	**269.7**
2	284.9	259.9	224.8	359.6	294.7	379.2	424.9	356.7	381.2	339.9	363.4	255.5	73.3	266.4	510.6	238.5	265.6	**310.5**
3	251.8	270.5	218.3	232.1	221.2	352.4	393.3	475.3	363.8	257.8	243.0	300.0	71.6	230.3	377.5	238.6	277.9	**280.9**
4	256.4	174.7	230.1	252.5	208.9	338.2	400.3	370.6	458.9	338.2	307.9	307.3	157.0	287.0	409.3	337.8	426.2	**309.5**
5	276.0	181.8	196.8	250.8	225.9	377.3	345.4	388.8	379.2	280.9	267.3	333.8	138.5	327.7	453.1	325.9	369.3	**301.1**
6	261.6	145.0	184.4	208.4	189.4	368.1	337.3	420.3	298.2	241.4	223.5	353.9	133.5	419.4	536.5	264.9	306.7	**287.8**
7	310.0	225.6	175.9	291.6	279.6	425.7	298.7	375.6	380.6	263.2	270.5	340.3	125.1	276.9	413.7	375.2	375.2	**306.1**
8	336.0	209.5	229.1	259.7	269.2	372.3	348.9	340.8	381.2	263.5	212.2	274.7	100.9	218.8	378.1	231.4	309.2	**278.6**
9	434.8	290.5	290.4	442.9	301.5	590.7	509.6	764.8	485.1	488.6	315.1	453.0	206.1	343.3	748.4	367.2	490.2	**442.5**
10	391.7	279.2	273.5	416.3	305.6	525.9	493.2	612.0	426.1	415.5	278.9	373.0	196.9	330.3	640.1	399.6	612.1	**410.0**
11	348.6	267.9	256.6	389.7	309.6	461.1	476.7	459.2	367.1	342.4	242.6	293.0	187.6	317.2	531.7	432.0	733.7	**377.5**
12	367.7	302.1	427.9	493.8	366.8	400.8	563.3	680.4	593.7	375.0	398.9	464.5	299.7	480.4	807.2	422.7	567.3	**471.3**
13	279.0	179.9	255.5	324.1	215.3	290.0	300.1	482.0	299.7	265.0	232.7	208.9	128.6	278.9	406.2	241.2	328.8	**277.4**
14	255.7	228.8	188.2	300.8	198.2	269.3	402.9	418.4	300.8	300.6	236.6	471.7	226.7	294.1	386.8	238.6	406.4	**301.4**
15	340.3	231.3	198.3	271.9	269.9	319.0	400.7	386.8	457.1	418.4	268.9	322.7	143.7	194.9	368.9	229.0	430.2	**308.9**
16	434.8	307.0	255.3	300.3	359.8	408.8	488.6	467.3	664.2	397.5	647.3	302.3	195.6	456.6	507.6	321.9	633.6	**420.5**
17	434.3	216.0	363.2	334.0	282.3	376.7	547.5	539.6	665.9	390.2	472.3	353.0	264.1	391.8	598.5	315.5	482.8	**413.4**
YrlyAver	**325.8**	**236.4**	**245.2**	**316.9**	**265.8**	**386.6**	**415.1**	**471.1**	**427.2**	**333.8**	**306.7**	**331.4**	**161.6**	**312.9**	**499.0**	**304.7**	**427.0**	**339.2**

Appendix C: Raw Temperature Data ^{0}C

Area	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	AreaAver
1	14.210	13.890	15.270	15.010	13.610	14.590	14.790	14.160	14.580	15.410	14.900	15.360	14.570	14.430	15.270	13.510	13.740	**14.547**
2	13.985	13.515	14.905	14.825	13.655	14.615	14.450	13.880	14.335	14.640	14.270	14.735	14.740	14.625	15.465	13.625	13.965	**14.366**
3	13.760	13.140	14.540	14.640	13.700	14.640	14.110	13.600	14.090	13.870	13.640	14.110	14.910	14.820	15.660	13.740	14.190	**14.186**
4	15.240	14.990	16.020	15.900	14.640	15.390	15.010	14.930	15.270	15.290	15.290	15.300	16.380	16.470	17.660	15.750	16.060	**15.623**
5	14.820	14.195	15.610	15.550	14.270	14.745	14.455	15.215	15.135	15.345	15.395	14.850	15.215	15.685	16.910	15.025	15.480	**15.171**
6	14.290	13.668	15.075	15.095	13.985	14.693	14.283	14.408	14.613	14.608	14.518	14.480	15.063	15.253	16.285	14.383	14.835	**14.678**
7	14.400	13.400	15.200	15.200	13.900	14.100	13.900	15.500	15.000	15.400	15.500	14.400	14.050	14.900	16.160	14.300	14.900	**14.718**
8	14.400	13.400	15.200	15.200	13.900	14.100	13.900	15.500	15.000	15.400	15.500	14.400	14.050	14.900	16.160	14.300	14.900	**14.718**
9	17.200	16.210	17.010	17.660	16.570	17.600	17.100	16.330	16.430	16.130	16.510	16.920	17.310	16.990	17.880	16.190	16.480	**16.854**
10	17.200	16.210	17.010	17.660	16.570	17.600	17.100	16.330	16.430	16.130	16.510	16.920	17.310	16.990	17.880	16.190	16.480	**16.854**
11	14.630	13.000	15.540	14.830	14.310	15.830	15.470	14.630	14.400	13.860	14.160	14.760	14.910	14.930	15.520	13.380	13.730	**14.582**
12	15.395	14.890	15.675	15.910	15.080	16.095	15.615	15.095	15.270	15.045	15.185	15.605	15.980	15.745	16.590	14.850	15.150	**15.481**
13	13.590	13.570	14.340	14.160	13.590	14.590	14.130	13.860	14.110	13.960	13.860	14.290	14.650	14.500	15.300	13.510	13.820	**14.108**
14	13.590	13.570	14.340	14.160	13.590	14.590	14.130	13.860	14.110	13.960	13.860	14.290	14.650	14.500	15.300	13.510	13.820	**14.108**
15	15.240	14.990	16.020	15.900	14.640	14.860	13.830	13.990	15.400	14.600	14.670	15.510	14.750	14.700	15.200	14.200	14.500	**14.882**
16	14.900	14.780	15.595	15.785	14.745	15.785	14.485	14.695	15.795	15.285	15.835	16.610	15.680	15.270	16.050	14.635	14.835	**15.339**
17	14.560	14.570	15.170	15.670	14.850	16.710	15.140	15.400	16.190	15.970	17.000	17.710	16.610	15.840	16.900	15.070	15.170	**15.796**
YrlyAver	**14.789**	**14.235**	**15.442**	**15.480**	**14.447**	**15.325**	**14.818**	**14.787**	**15.068**	**14.994**	**15.094**	**15.309**	**15.343**	**15.326**	**16.246**	**14.480**	**14.827**	**15.059**

Appendix D: Raw Soil Organic Matter Ratio Data

Area														AreaAver
1	0.7	1.96	1.5	1.89	1.53									1.516
2	1.92	1.96	1	1.35	2.54									1.754
3	2.6	1.67	1.73	1.44	1.18									1.724
4	1.53	2.25	1.74	1.99										1.878
5	1.66	1.94	1.99	1.26										1.713
6	1.44	2.05	2.06	1.74	1.16	1.18								1.605
7	1.28	1.95	3.18	1.66	1.22	1.16	2.5	1.15						1.763
8	1.76	1.55	1.96	2.18	1.86	1.78								1.848
9	3.3	2.8	3.14	4.15										3.348
10	2.02	1.44	2.19	2.02	2.21	2.69	1.79	5	1.93					2.366
11	1.58	1.94	3.31	6.13	2.43									3.078
12	1.51	1.5	5	2.27	2.5	1.13	1.16							2.153
13	1.12	1	1.41	2	2.03	2.5	4	1.07	2.91	1.41	1.15	1.06	2.43	1.853
14	1.6	1.5	2.43	2.16	3.53	2.45	1.57	2.69	1.44					2.152
15	2.76	1.97	2.67	0.51	1.67	0.71	1.71							1.714
16	1.74	3.11	2.8	1.75										2.350
17	1.32	1.73	2.16	2.91	1.79	2.03	2.86	2.56						2.170

Appendix E: Rain Equivalent Barley Grain Yield in t/ha

Area	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	AreaAver
1	1.213	1.095	0.877	1.139	0.969	1.392	1.428	2.071	1.580	1.302	1.026	0.991	0.429	0.903	1.796	0.883	1.074	**1.186**
2	1.254	1.144	0.989	1.582	1.297	1.668	1.870	1.569	1.677	1.496	1.599	1.124	0.323	1.172	2.247	1.049	1.169	**1.366**
3	1.108	1.190	0.961	1.021	0.973	1.551	1.731	2.091	1.605	1.134	1.069	1.320	0.315	1.013	1.661	1.050	1.223	**1.236**
4	1.128	0.769	1.012	1.111	0.919	1.048	1.761	1.631	2.019	1.488	1.355	1.352	0.691	1.263	1.801	1.486	1.875	**1.336**
5	1.140	0.703	0.912	1.014	0.876	1.334	1.623	1.740	1.666	1.275	1.169	1.455	0.639	1.554	2.081	1.326	1.612	**1.301**
6	1.151	0.638	0.811	0.917	0.833	1.620	1.484	1.849	1.312	1.062	0.983	1.557	0.587	1.845	2.361	1.166	1.349	**1.266**
7	1.364	0.993	0.774	1.283	1.230	1.873	1.754	1.653	1.675	1.158	1.190	1.497	0.550	1.218	1.820	1.651	1.651	**1.373**
8	1.478	0.922	1.008	1.143	1.184	1.638	1.535	1.500	1.677	1.159	0.934	1.209	0.444	0.963	1.664	1.018	1.360	**1.226**
9	1.913	1.278	1.278	1.949	1.327	2.599	2.242	3.365	2.134	2.150	1.386	1.993	0.907	1.511	3.293	1.616	2.157	**1.947**
10	1.583	1.329	1.883	2.173	1.614	1.764	2.479	2.994	2.612	1.650	1.755	2.044	1.319	2.114	3.552	1.860	2.496	**2.072**
11	1.534	1.179	1.129	1.715	1.362	2.029	2.097	2.020	1.615	1.507	1.067	1.289	0.825	1.396	2.339	1.901	3.228	**1.661**
12	1.583	1.329	1.883	2.173	1.614	1.764	2.479	2.994	2.612	1.650	1.755	2.044	1.319	2.114	3.552	1.860	2.496	**2.072**
13	1.228	0.787	1.124	1.426	0.947	1.276	1.320	2.121	1.319	1.166	1.024	0.919	0.566	1.227	1.787	1.061	1.447	**1.220**
14	1.125	1.007	0.828	1.324	0.872	1.185	1.773	1.842	1.324	1.323	1.041	2.075	0.997	1.294	1.702	1.050	1.788	**1.326**
15	1.497	1.018	0.873	1.196	1.188	1.404	1.763	1.702	2.011	1.841	1.183	1.420	0.632	0.858	1.623	1.008	1.893	**1.359**
16	1.931	1.351	1.123	1.321	1.583	1.799	2.150	2.056	2.922	1.749	2.848	1.330	0.861	2.009	2.233	1.416	2.788	**1.851**
17	1.911	0.950	1.607	1.470	1.242	1.657	2.409	2.374	2.930	1.717	2.078	1.553	1.162	1.724	2.633	1.388	2.124	**1.819**
YrlyAver	**1.420**	**1.040**	**1.122**	**1.409**	**1.178**	**1.624**	**1.876**	**2.092**	**1.923**	**1.460**	**1.380**	**1.481**	**0.739**	**1.422**	**2.244**	**1.340**	**1.867**	**1.507**

Appendix F: Temperature Equivalent Barley Grain Yield In t/ha

Area	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	AreaAver
1	1.206	0.515	0.361	1.851	0.155	1.216	1.768	2.054	1.781	1.481	1.239	1.274	0.099	0.900	1.338	1.747	1.653	**1.214**
2	1.099	0.534	0.442	1.860	0.155	1.746	1.804	2.302	2.087	1.623	1.272	1.343	0.124	0.781	1.789	2.030	1.154	**1.303**
3	1.305	0.683	0.881	1.870	0.155	1.986	1.743	1.794	2.228	1.634	1.355	0.946	0.103	0.777	1.058	1.202	1.199	**1.231**
4	1.333	0.510	0.327	0.780	0.544	0.892	2.199	2.267	1.820	1.325	1.062	0.745	0.160	0.574	1.286	1.804	1.863	**1.146**
5	1.333	0.528	0.648	0.863	0.420	0.920	2.476	2.032	1.503	0.843	0.934	1.024	0.222	0.246	1.320	1.486	1.903	**1.100**
6	0.904	0.590	0.339	0.673	0.330	0.911	1.893	2.682	1.847	1.467	1.189	1.428	0.246	0.439	1.605	1.949	2.077	**1.210**
7	0.806	0.642	0.781	0.767	0.170	1.557	3.095	2.486	2.194	1.023	0.947	1.034	0.095	0.702	0.832	1.409	1.906	**1.203**
8	0.822	0.539	0.766	1.199	0.447	1.323	3.118	2.478	2.187	1.023	0.103	0.608	0.075	0.702	1.391	1.531	1.991	**1.194**
9	1.239	1.141	1.265	1.652	1.459	1.368	1.492	1.355	1.345	1.290	1.066	1.348	0.383	0.784	1.442	1.381	1.246	**1.250**
10	1.416	1.156	1.398	1.771	1.737	1.082	1.636	1.434	1.716	1.716	1.039	1.183	0.334	1.582	1.798	1.381	1.416	**1.400**
11	0.836	0.049	0.790	2.084	0.708	1.665	1.357	2.412	2.406	2.266	2.168	2.311	0.456	1.750	2.195	2.430	1.672	**1.621**
12	1.098	0.439	1.309	1.839	1.077	1.340	1.770	1.954	1.238	1.243	1.231	1.096	0.315	0.951	0.990	1.114	1.252	**1.192**
13	0.963	0.139	0.590	2.054	0.447	1.744	1.765	2.500	1.524	1.170	1.163	0.905	0.386	1.484	1.352	0.824	1.839	**1.226**
14	1.424	0.424	1.172	2.612	1.820	1.125	1.596	2.569	2.082	1.633	1.073	0.952	0.393	1.558	1.465	1.062	2.452	**1.495**
15	1.596	0.429	0.828	1.045	0.536	1.195	2.189	1.706	1.140	1.082	1.265	0.512	0.183	0.412	0.911	1.954	2.089	**1.122**
16	1.970	0.302	1.118	1.336	0.982	1.349	2.590	1.960	1.368	1.272	1.027	1.015	0.472	1.289	0.913	1.888	1.184	**1.296**
17	2.586	0.214	1.781	1.192	1.692	1.138	1.253	1.531	1.211	1.259	1.265	1.254	0.515	1.221	1.058	1.096	1.141	**1.259**
YrlyAver	**1.290**	**0.520**	**0.870**	**1.497**	**0.755**	**1.327**	**1.985**	**2.089**	**1.746**	**1.373**	**1.141**	**1.116**	**0.268**	**0.950**	**1.338**	**1.546**	**1.649**	**1.262**

Appendix G: Soil Organic Matter Equivalent Yield Data in t/ha

Area	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	AreaAver
1	1.704	0.682	0.568	2.840	0.227	1.988	2.783	2.840	2.840	2.556	2.113	2.045	0.170	1.318	2.147	2.363	2.420	**1.859**
2	1.753	0.785	0.785	3.270	0.262	3.322	3.191	3.544	3.741	3.074	2.354	2.354	0.249	1.321	3.322	3.270	1.936	**2.267**
3	2.005	0.951	1.504	3.214	0.257	3.754	2.957	2.622	3.831	2.892	2.314	1.543	0.206	1.298	1.941	1.967	1.967	**2.072**
4	2.346	0.885	0.632	1.489	1.012	1.827	4.173	3.934	3.499	2.613	2.108	1.405	0.337	1.096	2.810	3.513	3.513	**2.188**
5	2.134	0.780	1.112	1.504	0.703	1.661	4.162	3.208	2.671	1.546	1.713	1.738	0.422	0.422	2.518	2.556	3.221	**1.886**
6	1.300	0.770	0.541	1.083	0.511	1.612	3.005	3.670	2.972	2.431	1.901	2.178	0.457	0.686	2.756	2.984	3.189	**1.885**
7	1.329	0.908	1.342	1.381	0.289	2.802	5.209	4.038	4.078	1.973	1.815	1.789	0.184	1.223	1.565	2.420	3.275	**2.095**
8	1.424	0.802	1.383	2.268	0.802	2.503	5.517	4.231	4.272	2.074	0.207	1.106	0.152	1.286	2.751	2.765	3.595	**2.185**
9	4.382	3.731	4.532	6.209	5.258	5.508	5.709	4.732	4.907	4.707	4.006	5.007	1.509	2.879	5.583	5.007	4.407	**4.593**
10	3.543	2.675	3.543	4.712	4.428	3.082	4.428	3.543	4.428	4.428	2.763	3.109	0.930	4.109	4.924	3.543	3.543	**3.631**
11	2.463	0.115	2.440	5.755	2.072	5.824	4.581	6.906	7.136	6.906	6.607	7.136	1.565	5.548	7.182	6.860	4.719	**4.930**
12	2.282	0.868	2.860	4.017	2.330	3.262	4.017	4.017	2.748	2.812	2.812	2.426	0.771	2.057	2.282	2.426	2.619	**2.624**
13	1.562	0.221	1.051	3.457	0.774	3.429	3.166	4.024	2.751	2.198	2.157	1.590	0.788	2.516	2.475	1.438	3.028	**2.155**
14	2.683	0.787	2.426	5.110	3.664	2.571	3.326	4.805	4.371	3.567	2.314	1.944	0.932	3.069	3.117	2.153	4.692	**3.031**
15	2.556	0.677	1.457	1.815	0.907	2.185	3.630	2.403	2.032	1.891	2.198	0.831	0.345	0.665	1.521	3.195	3.297	**1.859**
16	4.180	0.650	2.670	3.179	2.283	3.601	6.165	4.145	3.513	3.197	2.652	2.108	1.265	2.968	2.231	4.391	2.652	**3.050**
17	4.882	0.422	3.876	2.611	3.633	2.968	2.871	3.244	3.000	3.049	3.244	2.222	1.314	2.692	2.530	2.433	2.433	**2.790**
YrlyAver	**2.502**	**0.983**	**1.925**	**3.171**	**1.730**	**3.053**	**4.052**	**3.877**	**3.694**	**3.054**	**2.546**	**2.384**	**0.682**	**2.068**	**3.039**	**3.134**	**3.206**	**2.653**

Appendix H: Barley Yield Data With 40% Missing Values

Area	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	AreaAver
1	1.121		0.374	1.868		1.308		1.868	1.868	1.682		1.345	0.112	0.867	1.413		1.592	**1.285**
2		0.448			0.149		1.824		2.138	1.756	1.345			0.755	1.898	1.868		**1.354**
3	1.166	0.553	0.874			2.182	1.719	1.525				0.897	0.120			1.143	1.143	**1.132**
4	1.248		0.336	0.792	0.538		2.220	2.093	1.861	1.390	1.121			0.583			1.868	**1.277**
5								1.876		0.904		1.016		0.247	1.472	1.495		**1.168**
6	0.807		0.336	0.673		1.001				1.510	1.181		0.284		1.712		1.981	**1.054**
7		0.516	0.762		0.164	1.592	2.960	2.294	2.317		1.031	1.016		0.695	0.889	1.375	1.861	**1.344**
8				1.226			2.982		2.309	1.121	0.112		0.082		1.487			**1.331**
9	1.308	1.114	1.353	1.854	1.570	1.644	1.704	1.413		1.405		1.495		0.859		1.495	1.315	**1.425**
10	1.495	1.129			1.868			1.495	1.868		1.166			1.734	2.078		1.495	**1.592**
11			0.792	1.868		1.891			2.317	2.242	2.145	2.317	0.508	1.801		2.227	1.532	**1.786**
12	1.061	0.404	1.330		1.084	1.517			1.278	1.308	1.308		0.359	0.957		1.129		**1.067**
13				1.868	0.419		1.712	2.175	1.487			0.859		1.360	1.338		1.637	**1.428**
14	1.248	0.366	1.129	2.377	1.704	1.196	1.547	2.235		1.659	1.076	0.904	0.433	1.428	1.450	1.001		**1.317**
15				1.061		1.278			1.188	1.106	1.286	0.486		0.389			1.928	**1.090**
16	1.779	0.277	1.136		0.972		2.623	1.764		1.360		0.897	0.538		0.949	1.868		**1.288**
17	2.250		1.786		1.674	1.368	1.323		1.383		1.495		0.605	1.241	1.166		1.121	**1.401**
YrlyAver	**1.348**	**0.601**	**0.928**	**1.510**	**1.014**	**1.498**	**2.061**	**1.874**	**1.820**	**1.454**	**1.206**	**1.123**	**0.338**	**0.993**	**1.441**	**1.511**	**1.589**	1.314

Appendix I: Barley Yield Data With 10% Missing Values

Area	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	AreaAver
1	1.121	0.448	0.374	1.868	0.149	1.308	1.831	1.868	1.868	1.682	1.390	1.345	0.112	0.867	1.413	1.555	1.592	**1.223**
2	1.001	0.448	0.448	1.868	0.149	1.898	1.824	2.025	2.138	1.756	1.345	1.345	0.142	0.755	1.898	1.868	1.106	**1.295**
3	1.166	0.553	0.874	1.868	0.149	2.182	1.719	1.525	2.227	1.682	1.345	0.897	0.120	0.755	1.129	1.143	1.143	**1.205**
4	1.248		0.336	0.792	0.538	0.972	2.220	2.093	1.861	1.390	1.121	0.747		0.583	1.495	1.868	1.868	**1.276**
5	1.248					0.972		1.876	1.562	0.904	1.001	1.016		0.247	1.472	1.495	1.883	**1.243**
6	0.807		0.336	0.673		1.001		2.280		1.510	1.181	1.353	0.284	0.426	1.712	1.854	1.981	**1.184**
7	0.755	0.516	0.762	0.785	0.164	1.592	2.960	2.294	2.317	1.121	1.031	1.016	0.105	0.695	0.889	1.375	1.861	**1.191**
8	0.770		0.747	1.226	0.433		2.982	2.287	2.309	1.121	0.112	0.598	0.082	0.695	1.487	1.495	1.943	**1.219**
9	1.308	1.114	1.353	1.854	1.570	1.644	1.704	1.413		1.405	1.196	1.495		0.859	1.667	1.495	1.315	**1.426**
10	1.495	1.129	1.495	1.988	1.868	1.300	1.868	1.495	1.868	1.868	1.166			1.734	2.078	1.495	1.495	**1.623**
11		0.037	0.792	1.868	0.673	1.891		2.242	2.317	2.242	2.145	2.317	0.508	1.801	2.332	2.227	1.532	**1.662**
12	1.061	0.404	1.330		1.084	1.517	1.868	1.868	1.278	1.308	1.308	1.129	0.359	0.957	1.061	1.129	1.218	**1.180**
13	0.845	0.120	0.568	1.868	0.419		1.712	2.175	1.487	1.188	1.166	0.859	0.426	1.360	1.338	0.777	1.637	**1.122**
14	1.248	0.366	1.129	2.377	1.704	1.196	1.547	2.235		1.659	1.076	0.904	0.433	1.428	1.450	1.001	2.182	**1.371**
15			0.852	1.061		1.278	2.123	1.405	1.188	1.106	1.286	0.486	0.202	0.389	0.889	1.868	1.928	**1.147**
16	1.779	0.277	1.136		0.972	1.532	2.623	1.764	1.495	1.360	1.129	0.897	0.538		0.949	1.868	1.129	**1.296**
17	2.250	0.194	1.786	1.203	1.674	1.368	1.323	1.495	1.383	1.405	1.495	1.024	0.605	1.241	1.166	1.121	1.121	**1.286**
YrlyAver	**1.207**	**0.467**	**0.895**	**1.521**	**0.825**	**1.443**	**2.022**	**1.902**	**1.807**	**1.453**	**1.205**	**1.089**	**0.301**	**0.924**	**1.437**	**1.508**	**1.584**	1.291

Appendix J: Imputed Barley Yield Data With 40% Missing Values

Area	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	AreaAver
1	1.121	0.570	0.374	1.868	0.820	1.308	2.271	1.868	1.868	1.682	1.159	1.345	0.112	0.867	1.413	1.404	1.592	**1.273**
2	1.432	0.448	0.959	1.471	0.149	1.501	1.824	1.890	2.138	1.756	1.345	1.109	0.364	0.755	1.898	1.868	1.588	**1.323**
3	1.166	0.553	0.874	1.151	0.405	2.182	1.719	1.525	1.600	2.217	1.080	0.897	0.120	0.570	1.386	1.143	1.143	**1.161**
4	1.248	0.386	0.336	0.792	0.538	1.434	2.220	2.093	1.861	1.390	1.121	0.988	0.305	0.583	1.421	1.344	1.868	**1.172**
5	1.335	0.460	0.895	1.345	0.915	1.485	2.387	1.876	1.649	0.904	1.122	1.016	0.329	0.247	1.472	1.495	1.683	**1.213**
6	0.807	0.207	0.336	0.673	0.326	1.001	2.918	2.124	1.376	1.510	1.181	0.341	0.284	0.415	1.712	1.080	1.981	**1.075**
7	1.405	0.516	0.762	1.438	0.164	1.592	2.960	2.294	2.317	1.539	1.031	1.016	0.354	0.695	0.889	1.375	1.861	**1.306**
8	1.275	0.518	0.925	1.226	0.807	1.465	2.982	1.933	2.309	1.121	0.112	0.972	0.082	0.959	1.487	1.357	1.623	**1.244**
9	1.308	1.114	1.353	1.854	1.570	1.644	1.704	1.413	1.870	1.405	1.282	1.495	0.407	0.859	1.473	1.495	1.315	**1.386**
10	1.495	1.129	1.108	1.903	1.868	1.638	1.369	1.495	1.868	1.842	1.166	1.676	0.494	1.734	2.078	1.869	1.495	**1.543**
11	2.055	1.391	0.792	1.868	2.138	1.891	0.846	1.589	2.317	2.242	2.145	2.317	0.508	1.801	1.598	2.227	1.532	**1.721**
12	1.061	0.404	1.330	1.349	1.084	1.517	2.593	2.033	1.278	1.308	1.308	0.626	0.359	0.957	1.368	1.129	1.736	**1.261**
13	1.496	0.658	0.996	1.868	0.419	1.526	1.712	2.175	1.487	1.642	1.300	0.859	0.395	1.360	1.338	1.592	1.637	**1.321**
14	1.248	0.366	1.129	2.377	1.704	1.196	1.547	2.235	1.774	1.659	1.076	0.904	0.433	1.428	1.450	1.001	1.607	**1.361**
15	1.202	0.313	0.799	1.061	0.691	1.278	2.569	2.044	1.188	1.106	1.286	0.486	0.282	0.389	1.356	1.278	1.928	**1.133**
16	1.779	0.277	1.136	1.467	0.972	1.459	2.623	1.764	1.703	1.360	1.157	0.897	0.538	0.873	0.949	1.868	1.640	**1.321**
17	2.250	0.870	1.786	1.761	1.674	1.368	1.323	1.830	1.383	1.608	1.495	1.303	0.605	1.241	1.166	1.636	1.121	**1.436**
YrlyAver	**1.393**	**0.599**	**0.935**	**1.498**	**0.955**	**1.499**	**2.092**	**1.893**	**1.764**	**1.547**	**1.198**	**1.073**	**0.351**	**0.925**	**1.438**	**1.480**	**1.609**	1.309

Appendix K: Absolute Error for 40% Missing Values

Area	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	AreaAver
1		0.122			0.670		0.440				0.231					0.151		**0.323**
2	0.431		0.511	0.397		0.397		0.136				0.236	0.222				0.482	**0.351**
3				0.718	0.255				0.628	0.535	0.266			0.185	0.258			**0.406**
4		0.085				0.462						0.240	0.126		0.073	0.525		**0.252**
5	0.087	0.004	0.245	0.466	0.504	0.514	0.047		0.087		0.121		0.082				0.200	**0.214**
6		0.271			0.008		1.051	0.155	0.470			1.011		0.011		0.773		**0.469**
7	0.650			0.653						0.418			0.249					**0.492**
8	0.506	0.084	0.178		0.373	0.112		0.354				0.374		0.264		0.138	0.320	**0.270**
9									0.405		0.086		0.043		0.194			**0.182**
10			0.386	0.085		0.338	0.499			0.026		0.364	0.102			0.375		**0.272**
11	1.255	1.354			1.465		0.641	0.653							0.734			**1.017**
12				0.520			0.725	0.165				0.503			0.306		0.518	**0.456**
13	0.652	0.539	0.428			0.328				0.453	0.134		0.031			0.815		**0.422**
14									0.259								0.576	**0.418**
15	0.293	0.083	0.053		0.161		0.446	0.639					0.080		0.467	0.591		**0.313**
16				0.115		0.073			0.208		0.029			0.390			0.511	**0.221**
17		0.675		0.558				0.335		0.203		0.279				0.515		**0.428**
YrlyAver	**0.553**	**0.357**	**0.300**	**0.439**	**0.491**	**0.318**	**0.550**	**0.348**	**0.343**	**0.327**	**0.144**	**0.430**	**0.117**	**0.212**	**0.339**	**0.485**	**0.434**	**0.383**

Appendix L: Imputed Barley Yield Data With 10% Missing Values

Area	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	AreaAver
1	1.121	0.448	0.374	1.868	0.149	1.308	1.831	1.868	1.868	1.682	1.390	1.345	0.112	0.867	1.413	1.555	1.592	**1.223**
2	1.001	0.448	0.406	1.868	0.149	1.898	1.824	1.971	2.138	1.756	1.345	1.345	0.142	0.755	1.804	1.868	1.106	**1.284**
3	1.166	0.553	0.874	1.868	0.011	2.182	1.719	1.525	2.163	1.682	1.320	0.897	0.120	0.755	1.129	1.143	1.143	**1.191**
4	1.248	0.471	0.336	0.792	0.538	0.972	2.220	2.093	1.861	1.390	1.105	0.747	0.179	0.583	1.495	1.868	1.868	**1.163**
5	1.248	0.456	0.650	0.879	0.411	0.972	2.434	1.876	1.562	0.904	1.001	1.016	0.247	0.479	1.472	1.434	1.883	**1.113**
6	0.807	0.478	0.336	0.775	0.318	1.001	1.867	2.280	1.846	1.510	1.181	1.305	0.284	0.426	1.712	1.854	1.981	**1.174**
7	0.755	0.516	0.762	0.785	0.164	1.418	2.878	2.294	2.317	1.121	1.031	1.016	0.105	0.695	0.889	1.375	1.861	**1.175**
8	0.770	0.433	0.747	1.283	0.433	1.353	2.982	2.287	2.309	1.121	0.112	0.598	0.082	0.695	1.487	1.495	1.943	**1.184**
9	1.308	1.114	1.353	1.854	1.777	1.644	1.704	1.713	1.465	1.405	1.196	1.495	0.450	0.859	1.691	1.495	1.315	**1.402**
10	1.495	1.129	1.495	1.988	1.868	1.300	1.868	1.495	1.868	1.868	1.166	1.312	0.392	1.734	2.078	1.495	1.495	**1.532**
11	0.800	0.037	0.784	1.868	0.673	1.891	1.487	2.242	2.373	2.196	2.145	2.317	0.508	1.789	2.332	2.227	1.532	**1.600**
12	1.061	0.404	1.330	1.868	1.084	1.612	1.868	1.868	1.278	1.308	1.308	1.129	0.359	0.957	1.061	1.114	1.218	**1.225**
13	0.845	0.120	0.568	1.868	0.419	1.854	1.712	2.175	1.487	1.188	1.166	0.859	0.426	1.360	1.338	0.777	1.675	**1.167**
14	1.248	0.366	1.276	2.377	1.704	1.196	1.547	2.235	2.033	1.659	1.076	0.904	0.440	1.428	1.450	1.001	2.182	**1.419**
15	1.495	0.396	0.852	1.061	1.305	1.278	2.123	1.405	1.188	1.106	1.286	0.486	0.202	0.389	0.931	1.868	1.928	**1.135**
16	1.779	0.277	1.136	1.353	0.972	1.532	2.623	1.764	1.495	1.360	1.129	0.897	0.538	1.263	0.949	1.868	1.129	**1.298**
17	2.250	0.194	1.786	1.203	1.674	1.368	1.323	1.495	1.383	1.405	1.495	1.024	0.605	1.241	1.166	1.121	1.121	**1.286**
YrlyAver	**1.200**	**0.461**	**0.886**	**1.504**	**0.803**	**1.458**	**2.001**	**1.917**	**1.802**	**1.451**	**1.203**	**1.100**	**0.305**	**0.957**	**1.435**	**1.504**	**1.587**	1.269

Appendix M: Absolute Error for 10% Missing Values

Area	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	AreaAver
1																		
2			0.042					0.055							0.094			**0.064**
3					0.139				0.064		0.025							**0.076**
4							0.001				0.016							**0.008**
5														0.232		0.061		**0.146**
6				0.102								0.048						**0.075**
7						0.174	0.082											**0.128**
8				0.057														**0.057**
9					0.207			0.300							0.024			**0.177**
10																		
11			0.008						0.057	0.047				0.012				**0.031**
12						0.095										0.015		**0.055**
13																	0.039	**0.039**
14			0.148										0.006					**0.077**
15					0.775										0.042			**0.408**
16																		
17																		
YrlyAver			**0.066**	**0.080**	**0.374**	**0.134**	**0.041**	**0.178**	**0.061**	**0.047**	**0.021**	**0.048**	**0.006**	**0.122**	**0.053**	**0.038**	**0.039**	**0.087**

Appendix N: Absolute Error Square for 40% Missing Values

Area	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	AreaAver
1		0.015			0.449		0.193				0.053					0.023		**0.147**
2	0.186		0.261	0.158		0.158		0.018				0.056	0.049				0.232	**0.140**
3				0.515	0.065				0.394	0.286	0.071			0.034	0.066			**0.204**
4		0.007				0.214						0.058	0.016		0.005	0.275		**0.096**
5	0.007	0.000	0.060	0.217	0.254	0.264	0.002		0.008		0.015		0.007				0.040	**0.079**
6		0.074			0.000		1.105	0.024	0.221			1.023		0.000		0.598		**0.381**
7	0.422			0.427						0.174			0.062					**0.271**
8	0.256	0.007	0.032		0.139	0.013		0.125				0.140		0.070		0.019	0.102	**0.090**
9									0.164		0.007		0.002		0.038			**0.053**
10			0.149	0.007		0.114	0.249			0.001		0.133	0.010			0.140		**0.100**
11	1.576	1.833			2.147		0.411	0.426							0.539			**1.155**
12				0.270			0.526	0.027				0.253			0.094		0.268	**0.240**
13	0.425	0.290	0.183			0.108				0.206	0.018		0.001			0.664		**0.237**
14									0.067								0.332	**0.199**
15	0.086	0.007	0.003		0.026		0.199	0.409					0.006		0.218	0.349		**0.145**
16				0.013		0.005			0.043		0.001			0.152			0.261	**0.079**
17		0.456		0.311				0.112		0.041		0.078				0.265		**0.211**
YrlyAver	**0.422**	**0.299**	**0.115**	**0.240**	**0.440**	**0.125**	**0.384**	**0.163**	**0.150**	**0.142**	**0.027**	**0.248**	**0.019**	**0.064**	**0.160**	**0.292**	**0.206**	**0.225**

Appendix O: Absolute Error Square for 10% Missing Values

Area	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	AreaAver
1																		
2			0.042					0.055							0.094			**0.064**
3					0.139				0.064		0.025							**0.076**
4							0.001				0.016							**0.008**
5														0.232		0.061		**0.146**
6				0.102								0.048						**0.075**
7						0.174	0.082											**0.128**
8				0.057														**0.057**
9					0.207			0.300							0.024			**0.177**
10																		
11			0.008						0.057	0.047				0.012				**0.031**
12						0.095										0.015		**0.055**
13																	0.039	**0.039**
14			0.148										0.006					**0.077**
15					0.775										0.042			**0.408**
16																		
17																		
YrlyAver			**0.066**	**0.080**	**0.374**	**0.134**	**0.041**	**0.178**	**0.061**	**0.047**	**0.021**	**0.048**	**0.006**	**0.122**	**0.053**	**0.038**	**0.039**	**0.087**

Appendix P: Relative Aitchison Distance (RDA) for The 40% Missing Data for Each Method of Imputation

	MCMC	GausKrnl_h=2	EpanKrnl	MRSV
Area	RDA=UsingImputedvals	RDA compFrmİmptdOnly	RDAimpt	RDA=UsingImputedOnly
1	1.858	4.217	5.121	0.179
2	2.114	3.921	4.221	0.482
3	2.582	3.507	4.330	0.365
4	1.868	3.056	3.256	0.436
5	3.049	2.662	2.676	0.327
6	2.548	2.777	2.561	0.681
7	2.154	3.521	3.956	0.233
8	2.112	2.868	3.968	0.768
9	1.595	2.889	3.905	0.627
10	1.273	1.628	2.532	0.521
11	2.001	3.353	2.391	0.976
12	1.499	3.508	4.365	0.954
13	2.708	3.717	4.707	0.621
14	1.271	4.091	5.706	1.087
15	2.499	3.666	4.914	0.457
16	1.714	0.916	1.306	0.652
17	1.740	2.202	2.518	0.880

Appendix Q: Imputed Barley Yield Data With 40% Missing Values (MCMC)

Area	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	AreaAver
1	1.121	0.570	0.374	1.868	0.820	1.308	2.271	1.868	1.868	1.682	1.159	1.345	0.112	0.867	1.413	1.404	1.592	**1.273**
2	1.432	0.448	0.959	1.471	0.149	1.501	1.824	1.890	2.138	1.756	1.345	1.109	0.364	0.755	1.898	1.868	1.588	**1.323**
3	1.166	0.553	0.874	1.151	0.405	2.182	1.719	1.525	1.600	2.217	1.080	0.897	0.120	0.570	1.386	1.143	1.143	**1.161**
4	1.248	0.386	0.336	0.792	0.538	1.434	2.220	2.093	1.861	1.390	1.121	0.988	0.305	0.583	1.421	1.344	1.868	**1.172**
5	1.335	0.460	0.895	1.345	0.915	1.485	2.387	1.876	1.649	0.904	1.122	1.016	0.329	0.247	1.472	1.495	1.683	**1.213**
6	0.807	0.207	0.336	0.673	0.326	1.001	2.918	2.124	1.376	1.510	1.181	0.341	0.284	0.415	1.712	1.080	1.981	**1.075**
7	1.405	0.516	0.762	1.438	0.164	1.592	2.960	2.294	2.317	1.539	1.031	1.016	0.354	0.695	0.889	1.375	1.861	**1.306**
8	1.275	0.518	0.925	1.226	0.807	1.465	2.982	1.933	2.309	1.121	0.112	0.972	0.082	0.959	1.487	1.357	1.623	**1.244**
9	1.308	1.114	1.353	1.854	1.570	1.644	1.704	1.413	1.870	1.405	1.282	1.495	0.407	0.859	1.473	1.495	1.315	**1.386**
10	1.495	1.129	1.108	1.903	1.868	1.638	1.369	1.495	1.868	1.842	1.166	1.676	0.494	1.734	2.078	1.869	1.495	**1.543**
11	2.055	1.391	0.792	1.868	2.138	1.891	0.846	1.589	2.317	2.242	2.145	2.317	0.508	1.801	1.598	2.227	1.532	**1.721**
12	1.061	0.404	1.330	1.349	1.084	1.517	2.593	2.033	1.278	1.308	1.308	0.626	0.359	0.957	1.368	1.129	1.736	**1.261**
13	1.496	0.658	0.996	1.868	0.419	1.526	1.712	2.175	1.487	1.642	1.300	0.859	0.395	1.360	1.338	1.592	1.637	**1.321**
14	1.248	0.366	1.129	2.377	1.704	1.196	1.547	2.235	1.774	1.659	1.076	0.904	0.433	1.428	1.450	1.001	1.607	**1.361**
15	1.202	0.313	0.799	1.061	0.691	1.278	2.569	2.044	1.188	1.106	1.286	0.486	0.282	0.389	1.356	1.278	1.928	**1.133**
16	1.779	0.277	1.136	1.467	0.972	1.459	2.623	1.764	1.703	1.360	1.157	0.897	0.538	0.873	0.949	1.868	1.640	**1.321**
17	2.250	0.870	1.786	1.761	1.674	1.368	1.323	1.830	1.383	1.608	1.495	1.303	0.605	1.241	1.166	1.636	1.121	**1.436**
YrlyAver	1.393	0.599	0.935	1.498	0.955	1.499	2.092	1.893	1.764	1.547	1.198	1.073	0.351	0.925	1.438	1.480	1.609	**1.309**