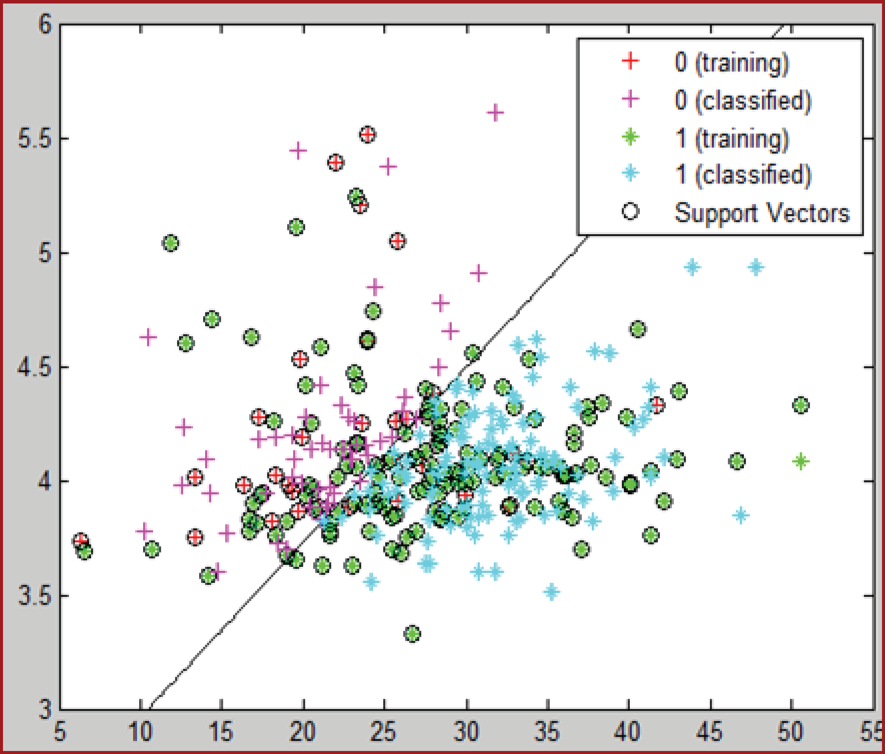
DIFFERENT STATISTICAL APPLICATIONS IN AGRICULTURE

EDITORS
Prof. Dr. İsmail KESKİN
Assoc. Prof. Dr. Nazire MİKAİL
Dr. Yasin ALTAY



İKSAD
Publishing House

DIFFERENT STATISTICAL APPLICATIONS IN AGRICULTURE

EDITORS

Prof. Dr. İsmail KESKİN
Assoc. Prof. Dr. Nazire MİKAİL
Dr. Yasin ALTAY

AUTHORS

Prof. Dr. Alaaddin YÜKSEL
Prof. Dr. Alaaddin YÜKSEL
Prof. Dr. Ali Rıza DEMİRKIRAN
Prof. Dr. İsmail KESKİN
Prof. Dr. Mehmet Bozkurt ATAMAN
Assoc. Prof. Dr. İbrahim AYTEKİN
Assoc. Prof. Dr. Nazire MİKAİL
Assoc. Prof. Dr. Suna AKKOL
Assoc. Prof. Dr. Şenol ÇELİK
Assist. Prof. Dr. Aslı AKILLI
Assist. Prof. Dr. Bülent BÜLBÜL
Assist. Prof. Dr. Derviş TOPUZ
Assist. Prof. Dr. Fatma İLHAN
Dr. Cem TIRINK
Dr. Gafur GÖZÜKARA
Dr. Mehmet ULUPINAR
Dr. Yasin ALTAY
Jabar Jalal FAQE
Kadir Miraç DAŞBİLEK
Mesut KIRBAŞ
Şükrü DOĞAN
Uğur DEMİRCİ

İKSAD
Publishing House

Institution of Economic
Development and Social
Researches Publications®
(The Licence Number of Publicator: 2014/31220)
TURKEY TR: +90 342 606 06 75
USA: +1 631 685 0 853
E mail: iksadyayinevi@gmail.com
www.iksadyayinevi.com

It is responsibility of the author to abide by the publishing ethics rules.

ISBN: 978-625-8061-07-9
Cover Design: İbrahim KAYA
December / 2021
Ankara / Turkey
Size = 16x24 cm

CONTENTS

CHAPTER I

POULTRY RATIONS PREPARATION POSSIBILITIES WITH CENTRAL COMPOSITE AND BOX-BEHNKEN DESIGN

Dr. Yasin ALTAY[*]

Assoc. Prof. Dr. Nazire MİKAİL[†]

Prof. Dr. İsmail KESKİN[‡]

[*] Eskisehir Osmangazi University, Faculty of Agriculture, Department of Animal Science, Eskisehir, Turkey. yaltay@ogu.edu.tr

[†] Siirt University, Faculty of Agriculture, Department of Animal Science, Siirt, Turkey. naziremikail@siirt.edu.tr

[‡] Selcuk University, Faculty of Agriculture, Department of Animal Science, Konya, Turkey. ikeskin@selcuk.edu.tr

INTRODUCTION

Optimization is a method that enables reaching certain goals by using the resources of a system in the most efficient way possible (Banga et al., 2003). In other words, it is a technique that is used to choose the best one among the possible alternatives under certain conditions. In optimization, the dependent variable and the relationships of the independent variables with each other are explained via linear or nonlinear regression models. These created models tend to convey the dependent variable to a minimum or maximum target by optimizing the independent variables. The most important element of the optimization technique is to determine the variables that will respond to the model in the best way and include them in the model. In fact, the independent variables are a function of the dependent variable.

Various explanatory variables Response Surface Methodology (RSM) is a method that explains the relationships between one or more response variables. It was developed by Box and Wilson in 1951 (Box and Wilson, 1992). In 1957, the concept of rotatability had revealed by Box and Hunter in the Central Composite Design. RSM methods that can be used for different purposes and developed by Box and Draper in 1959, were revealed (Mead, 1975). The main idea of the method is to construct a linear model or a quadratic polynomial model by using a set of experiments designed to obtain an optimal result. The contribution of the independent variables to the established model in the optimization are statistically significant. They are determined using one of the backward, stepwise and forward regression variable elimination methods. The basis of RSM is multivariate statistical methods, multiple regression and analytical space geometry. The optimum region is found by using the function of reaching the desired target by superimposing-desirability after drawing the isohips curves of the responses (Koç and Ertekin, 2010).

RSM is one of the multivariate analysis methods, and it can be seen in applied sciences, engineering, electricity-electronics, chemistry, food, genetics and agriculture applications in recent years. The method is proper to obtain important and effective results, especially in studies where time and budget are limited. Optimization has become one of the most important issues that have gained momentum with the development of computer software techniques in recent years. Optimization is not only important for economic reasons; it is also one of the effective methods in creating new products. In this way, preventing the waste of inputs as time, money, labor, equipment, etc. provides the opportunity to obtain the highest quality product with maximum profit (Yavuz and Keskin, 2020).

In cases where factor levels are very sensitive, it allows examining multiple dependent variables together and frequently used in areas that require optimization, especially in the agricultural sector. In the field of animal science, studies on this subject are limited. It is obvious that the savings to be achieved through the optimization of product costs, labor, proximity to the market and time will provide extra value to the manufacturers. Optimization practices in animal husbandry are at an insufficient level due to some difficulties in practice, although it is a suitable field of study. Optimization is inevitable because animal livestock activities can be carried out at the economic efficiency level by using scarce resources. In animal livestock, approximately 75-80 % of the total costs consist of feed costs and it is very important to use these scarce resources more carefully.

Thanks to the scientific and technological developments in the poultry farming sector in recent years, poultry meat has become a strategic and industrial product that needs to be produced more economically and with higher efficiency. It has become imperative for producers to consider economic production planning and costs by the industrialization of poultry meat production. Feed expenses constitute a large part of the cost of poultry meat production. For this reason, it is aimed to reach maximum live weight by preparing an optimum ration that meets all nutritional requirements of poultries (Faridi et al., 2013). Although there are some studies in the literature amounts that meet the metabolic and physiological optimum nutritional needs of poultries are available, the relevant literature is still limited.

By optimizing the protein and energy content of the diets of Japanese quails in terms of live weight, they reported that the RSM design is a suitable method for determining the nutritional needs of quails (Roush et al., 1979). They reported that diets of white Leghorn male chicks were optimized for protein, carbohydrates, and fat using a triangular chart in terms of body weight and multiple regression analysis and could be successfully applied in many nutritional fields to investigate the biological response to three factors from any nutrient present in the ration (Toyomizu et al., 1982). Roush (1983) emphasized that ration change time may be an important for improving carcass quality by optimizing protein levels in the start and finish rations in terms of body weight, carcass weight, feed efficiency, ration change time and net profit of male broilers. Sanders et al. (1992) studied the 10-days-old turkeys and determined with the RSM method that there is an optimum requirement of 12.5 g / kg of Ca and 10 g / kg of P in the ration. Faria Filho et al. (2008) applied RSM by using the results of eight other studies including weight gain, feed consumption, and feed efficiency performances in broiler chickens between 1995 and 2005 and found the preliminary study of the investigated characteristics successful. Ahmadi and Golian (2011) determined the optimization ration concentrations and broiler performance values using the RSM method in a study they conducted with 5 levels and 4 factors (digestible protein, lysine, total sulfur amino acid and threonine) in 420 broiler chickens 11–17 days old.

The purpose of this study is to estimate the optimum crude protein (CP), metabolic energy (ME) and Ca levels in the starting, growing and finishing rations in order to reach maximum body weight in broilers via Central Composite and Box-Behnken experimental designs.

1. MATERIAL

Since it is very difficult and troublesome to find real data suitable for the RSM experimental design in the study, a new data set was created via simulating by making use of different kinds of literature, expert opinions, and breeder handbooks. For rations in different periods, values between 17.5-24 % for CP and 2950-3300 kcal for ME were taken for Central Composite Design (CCD), while in Box-Behnken design (BBD) values are between 17.5-24% for CP, values between 2950-3300 kcal and Ca 0.7-1.1% were used.

2. METHOD

RSM was used to generate live weight (LW) response surfaces of ME, CP, and Ca levels. CDD and BBD were chosen as the experimental design.

The data obtained in terms of the features were analyzed using the Minitab statistical package program trail version 16. The adequacy of the model was evaluated by considering the coefficient of determination (R^2) and the goodness of fit (Lack of fit). Regression models were determined according to the quadratic equation with two and three variables, and the stepwise variable elimination method was used.

2.1. Responce Surface Methodology

RSM is a form of optimization based on the creation of an empirical model to evaluate the relationship between controllable experimental factors and the results obtained. While response variables are dependent variables, factors affecting the process are defined as independent variables (Myers and Montgomery, 2002). One of the most important features of RSM is that it effectively optimizes the independent variables that are thought to affect the response variable by using appropriate mathematical and statistical methods (Montgomery, 2001).

A model can be created using RSM multivariate regression models, at the same time a large number of dependent variables can be examined together and optimum results can be obtained with the least number of trials (Box and Draper, 1987). The fact that it allows the determination of the optimum point by considering a large number of responses makes RSM stand out among other optimization methods.

In factorial experimental desing, all treatment combinations must be included in the experiment. However, not all treatment combinations are used, as the RSM method is based on optimum points. RSM estimates the corresponding response values by constructing an appropriate mathematical function for the treatment combinations not included in the experiment, thus allowing to find the independent variables that will make the dependent variable optimal (Yılmaz, 2002). In this way, there is no harm in using the combinations that are not studied by researchers since they do not cause a loss of information (Yılmaz, 2002). For example, when in a factorial experiment with 3 factors and 3 levels, 3^3=27 treatment combinations are required, It is sufficient to use 2^n+(2n+1)= 2^3+(2*3+1)=15 treatment combinations in RSM, and this value can be reduced to 11 (Walker, 1984).

After the optimum points are found in the RSM, the selection of the treatment combinations to be used is extremely important. The most important criterion in determining the optimum points to be included in the model is to determine whether the relationship between them is linear (linear) or quadratic (Koç and Kaymak-Ertekin, 2010). One of the most important purposes of RSM is to determine the trial conditions in order to obtain the optimum desired responses. A model should be chosen in which all predicted responses have a high coefficient of determination (R^2) in the regression. Because the biggest disadvantage of RSM is the application of a linear regression model to independent variables that are not suitable for a linear regression model (Koç and Kaymak-Ertekin, 2010). The optimum exprimental conditions offered by the model can keep all of the responses in the desired range for the highest efficiency or at least one of them at the desired value (Myers and Montgomery, 2002). It is not recommended to be used only in cases where the number of factors is more than four, since it becomes difficult to interpret and establish a model (Fenwick et al., 2014).

If the relationship between dependent and independent variables is linear, the equation in (1) is used, and if the relationship is not linear, the equation in (2) is used (Myers and Montgomery, 2002). The RSM model is created by estimating the equations (1 and 2) and the partial regression coefficients (β) by utilizing the dependent variables. The least squares method is used to find these coefficients.

$$\hat{Y} = \beta_0 + \sum_{i=1}^{n} \beta_i X_i + \varepsilon_{ij}$$

$$\hat{Y} = \beta_0 + \sum_{i=1}^{n} \beta_i X_i + \sum_{i=1}^{n} \beta_{ii} X_i^2 + \sum_{i=1}^{n} \sum_{j=i+1}^{n} \beta_{ij} X_i X_j + \varepsilon_{ij}$$

In the equations, Y is the response variable, X is the independent variable, β_0 is the regression constant, β_i is the first-order (linear) equation coefficient, β_{ii} is the quadratic equation coefficient, β_{ij} is the interaction coefficient of the factors, and ε_{ij} is the error term.

The response surface plot represents the expected η responses corresponding to the x_1 and x_2 levels (Figure 1). Response Obtain the contours of the surface to visualize its shape. Contour shapes are obtained by taking fixed responses in the x_1, x_2 plane (Figure 2). Here, each contour corresponds to a certain height of the response surface.

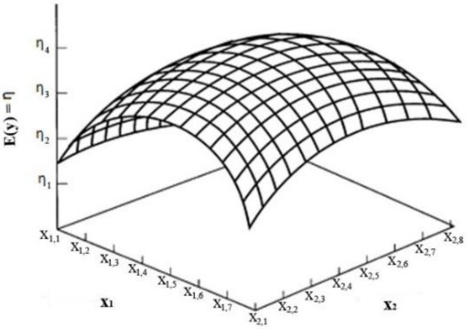


Figure 1. η response surface representation corresponding to the x_1 and x_2 planes

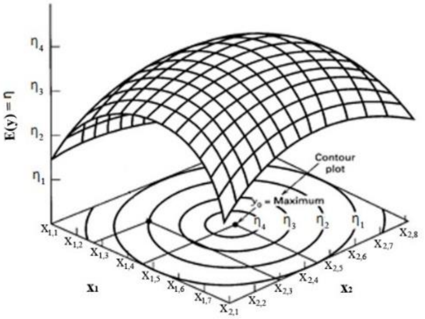


Figure 2. η response surface and contour representation corresponding to the x_1 and x_2 planes

2.1.1 Advantages of the RSM

- Allows easy optimization,
- Provides maximum information with a small number of experimental units,
- Possibility to use and change parameters simultaneously,
- Determing the interaction between the factors and enabling the removal of insignificant levels,
- Transformation can be done when the residual are not normally distributed,
- Being sensitive to outliers,
- Allows blocking,
- Giving advance information to the researchers in determining the level in a factorial experimental design,
- Ability to do without repetition.

2.1.2 Disadvantages of the RSM

- Data are not normally distributed because the number of treatment combinations is small,
- Factor levels being equidistant from each other,
- Which regression equation should be chosen,
- Performing analyzes without carrying the prerequisites of regression analysis of the data,
- Researchers can choose the wrong experimental design,
- As the number of factors increases, modeling and interpretation becomes more difficult.

Commonly used designs in RSM are the Central Composite and the Box-Behnken design. The difference between these two designs is due to the difference of the number of factors and treatment levels.

2.1.3 Central Composite Design (CDD)

CDD, one of the basic experimental designs of RSM, is used when two or more factors have two levels. After determining an α value that is equidistant from two levels of the independent variable in the CCD and making it five-level (–α, -1, 0, 1, α), the experimental is set up. The α value, called the closest orthogonal distances to these two levels, changes according to the desire of the researcher. In general, if the researcher has no influence on the experimental design, the hypotenuse of a right triangle with sides of 1 unit will be $\alpha = \sqrt{2}$ units (Figure 3). The representation of optimum points on a cube in CCD is shown in Figure 4.

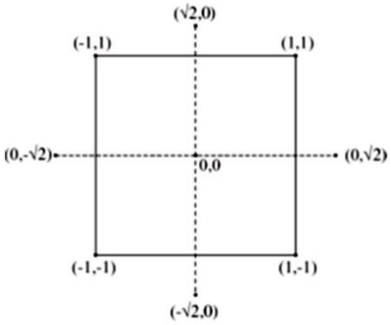


Figure 3. Representation of 2-Factor 2-level CDD in the coordinate system ($\alpha=\sqrt{2}$)

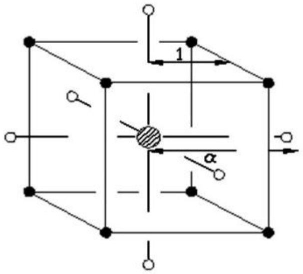


Figure 4. View of the cube-shaped optimum points of the CCD

In an experimental design with more than two independent variables (factors), many pattern points emerge. This situation may cause some problems both in the execution of the trial and in the interpretation at the end of the experiment. In such cases, the pattern point should be reduced. Box and Wilson (1951) revealed that the pattern point can be reduced by using a Central Composite Design. The orthogonal fragmentation of factor levels for CCD is given in Table 1.

Table 1. Orthogonal of factor levels for CDD*

Order	A	B
1	0	0
2	0	0
3	-1	1
4	1	1
5	0	1
6	0	0
7	0	-1
8	1	0
9	-1	0
10	-1	-1
11	0	0
12	1	-1
13	0	0

[*] Independent variables were randomly assigned

CCD ceases to be a reliable (robust) test when it contains more than two factors. It is recommended to use different experimental designs instead of CCD in studies where the number of factors is more than two.

In the factorial experimental design, if a study with two factors, five levels, and five replications is desired, there should be 2x5x5=50 experimental units. However, in CCD, the number of experimental units can be reduced to 13. In this way, it provides optimum results with the minimum experimental units and maximum information.

2.1.4 Box-Behnken Design (BBD)

The main difference between the Box-Behnken Design (BBD) and the Central Composite Design is that the factor levels are different. While five levels are used in CCD, three levels are used in BBD. In general, if the researcher has no influence on the experimental design, the hypotenuse of a right triangle with sides of one unit will be α = 1.68179 units. The representation of optimum points on a cube in BBD is shown in Figure 5.

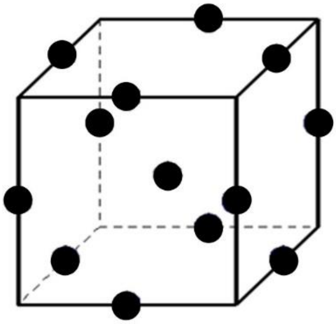


Figure 5. View of the cube-shaped optimum points of the BBD

In BBD, one of the factor levels is kept constant as the center value, while all the levels of the other factors take shape according to this center. As seen in Figure 6, factor C was fixed first, then all combinations of A and B factors were placed. Later, the same process was applied for the A and B factors (Montgomery, 2017).

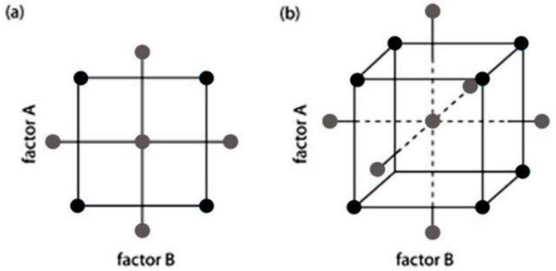


Figure 6. View of the square-shaped optimum points of the CCD

The orthogonal fragmentation of factor levels for BBD is given in Table 2.

Table 2. Orthogonal of factor levels for BBD*

Order	A	B	C
1	-1	-1	0
2	1	-1	0
3	-1	1	0
4	1	1	0
5	-1	0	-1
6	1	0	-1
7	-1	0	1
8	1	0	1
9	0	-1	-1
10	0	1	-1
11	0	-1	1
12	0	1	1
13	0	0	0
14	0	0	0
15	0	0	0

* Independent variables were randomly assigned.

In the factorial experimental design, if a trial with three factors, three levels, and five replications is desired, 3x3x5=45 trial units are needed, while the number of trial units can be reduced to 15 with BBD. BBD can be used as an alternative to factorial experimental designs. Thus, the experiments will become more efficient and the researchers will be able to save money in terms of cost and labor.

3. RESULTS

3.1. Central Composite Design Applications

Since the nutrition requirements of the animals are different in the starter, grower, and finisher periods in broilers, three different ration samples were created. Optimum amounts of ME and CP to maximize live weight (LW) at the end of each period were determined. LW values at the 10th day of the starter period, the 25th day of the grower period, and the 50th day of the finisher period were used as response variables. For all ration samples, a total of 13 points were determined as cube points (4), center points in a cube (5), and axial points (4), and the analysis was performed by taking the α value of 1.

3.1.1. Example of Starter Period (0-10 days) Ration with CDD

Maximum LW was aimed by taking the ration that meets the nutrient needs of animals in the 0-10 days, in the range of ME 2950-3050 kcal, and CP 22-24%. The LW values corresponding to the ME and CP combinations in the experimental units are presented in Table 3.

Tablo 3. LW results for ME and CP combinations

StdOrder	RunOrder	PtType	Blocks	ME	CP	LW
13	1	0	1	3000	23	325
4	2	1	1	3050	24	315
12	3	0	1	3000	23	322
5	4	-1	1	2950	23	305
6	5	-1	1	3050	23	310
8	6	-1	1	3000	24	316
2	7	1	1	3050	22	312
1	8	1	1	2950	22	300
9	9	0	1	3000	23	321
7	10	-1	1	3000	22	308
11	11	0	1	3000	23	324
3	12	1	1	2950	24	312

In the study, ME, and CP were modeled with the help of nonlinear regression in terms of LW values (Table 4). While ME and CP, which are the main effects of the established model, were statistically insignificant ($P>0.05$), the square effects of ME^2 and CP^2 were found to be significant ($P<0.05$). It was determined that there is a relationship between LW, ME and CP as "$LW=-35175 + 21.91ME + 216CP – 0.00364ME^2– 4.60CP^2$". When evaluated in terms of model performance, it was determined that 79.47 % of the change in LW was due to the change in the amount of ME and CP ($R^2 = 0.7947$).

Tablo 4. The coefficients and performance parameters of the prediction model of the LW corresponding to the ME and CP levels

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	321.17	1.82	176.58	0.000	
ME	3.33	1.79	1.86	0.099	1.00
CP	3.83	1.79	2.14	0.064	1.00
ME^2	-9.10	2.64	-3.45	0.009	1.17
CP^2	-4.60	2.64	-1.75	0.119	1.17

LW changes according to the amount of ME and CP of the starter ration are shown in Figure 7 and Figure 8. It is seen that the effect of the amount of ME on the LW values is in the form of a curve. When the model parameters were examined, it was determined that the second-order term ME was also significant (Table 4). LW reached its maximum value when the amount of ME reaches about 3000 kcal (Figure 8). With the increase in the CP ratio, there was an increase in the amount of LW. However, this increase also lost its speed over time. Although this is statistically insignificant, it shows the effect of the quadratic term of the CP ratio.

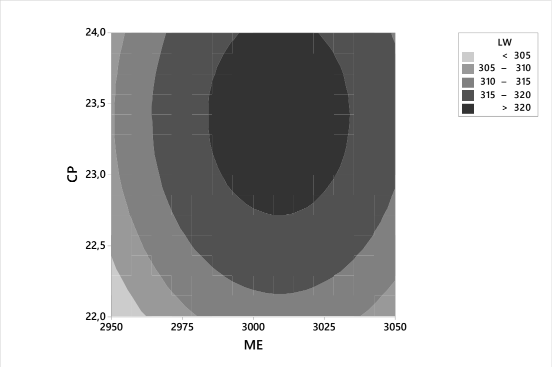


Figure 7. Contour plot representation of the effect of ME and CP changes on LW

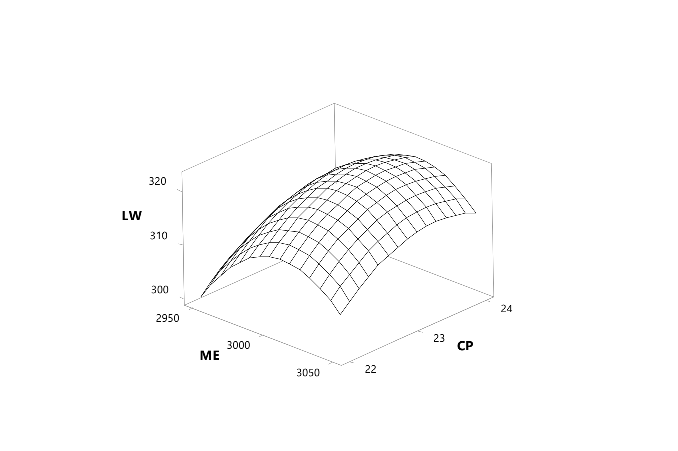
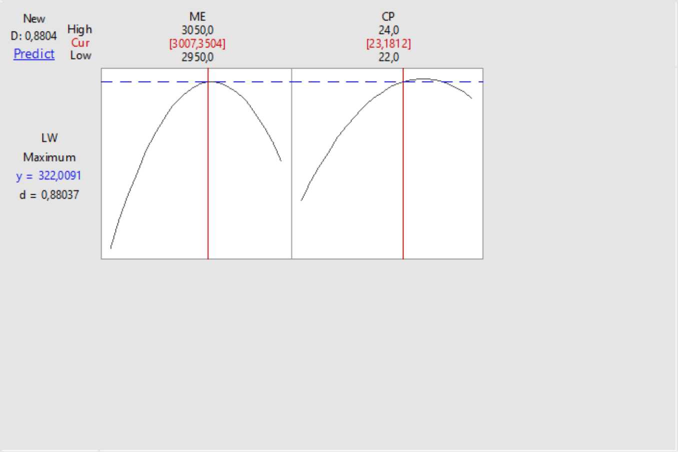


Figure 8. Surface plot representation of the effect of ME and CP changes on LW

In the study, LW was maximized using CDD and optimum ME and CP values were determined (Figure 9). As a result, it was estimated that broilers consuming the ration prepared at 3009.60 kcal ME and 23.4171% CP ratios would be a maximum of 322.27 g.



New
D: 0,8804
Predict

High
Cur
Low

ME
3050,0
[3007,3504]
2950,0

CP
24,0
[23,1812]
22,0

LW
Maximum
y = 322,0091
d = 0,88037

Figure 9. Optimization result of the ration in the starter period

3.1.2 Example of Grower Period (10-25 day) Ration with CDD

Maximum LW was aimed by taking the ration that meets the nutrient needs of broilers in the 10-25 days, in the range of ME 3000-3200 kcal, and CP 21-22 %. The response variable (LW) values corresponding to the ME and CP combinations are given in Table 5.

Tablo 5. LW results for ME and CP combinations

StdOrder	RunOrder	PtType	Blocks	ME	CP	LW
1	1	1	1	3000	21.0	1250
4	2	1	1	3200	22.0	1280
8	3	-1	1	3100	22.0	1285
7	4	-1	1	3100	21.0	1270
6	5	-1	1	3200	21.5	1300
12	6	0	1	3100	21.5	1315
2	7	1	1	3200	21.0	1275
13	8	0	1	3100	21.5	1313
10	9	0	1	3100	21.5	1314
3	10	1	1	3000	22.0	1260
11	11	0	1	3100	21.5	1316
5	12	-1	1	3000	21.5	1265

LW values and independent variables ME and CP were modeled with nonlinear regression (Table 6). In the model, regression constant, ME, ME^2, and CP^2 were statistically significant ($P<0.05$) but CP was not significant ($P>0.05$). There was a relationship between LW, ME, and CP as “LW=-71060 + 13.92ME + 4696CP – 0.002224ME^2 – 109.0CP^2”. When the model performance was evaluated, it was determined that the coefficient of determination of the prediction equation was $R^2 = 0.9439$.

Tablo 6. The coefficients and performance parameters of the prediction model of the LW corresponding to the ME and CP levels

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	1312.07	2.93	447.71	0.000	
ME	13.33	2.88	4.63	0.002	1.00
CP	5.00	2.88	1.74	0.121	1.00
ME^2	-22.24	4.25	-5.24	0.001	1.17
CP^2	-27.24	4.25	-6.41	0.000	1.17

LW amounts varying according to the amount of ME and CP of the grower ration created with CDD are given in Figure 10 and Figure 11. It was found that the ME value was in the range of 3000-3100, and the LWs of the animals were at the highest level. However, the increase in ME after this interval decreased the LW amount. Similarly, LW against the change in CP amount decreased when the CP ratio is greater than 21.5%. It showed the effect of the quadratic term on the LW of both independent variables.

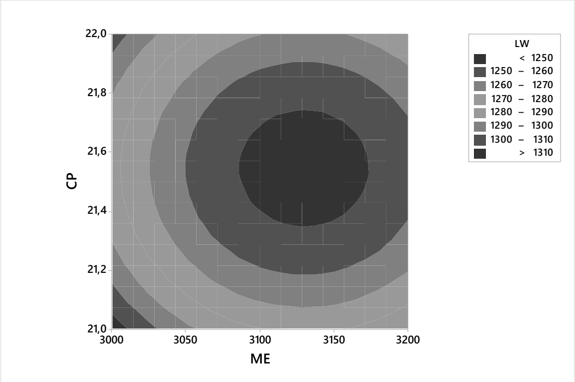


Figure 10. Contour plot representation of the effect of ME and CP changes on LW

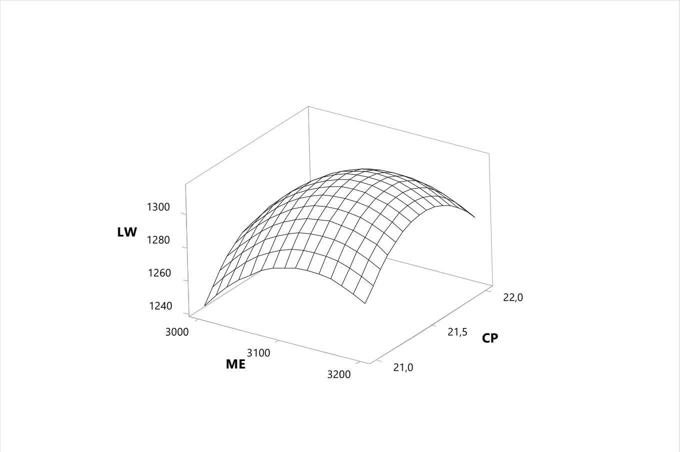


Figure 11. Surface plot representation of the effect of ME and CP changes on LW

In the grower ration, the optimum values of CP and ME were determined by maximizing LW (Figure 12). When broilers in the development period consume the ration prepared at the rate of 3129.29 kcal ME and 21.5455 % CP, it is estimated that it will be in the range of 1296.68-1331.91 g with a probability of 95 %.

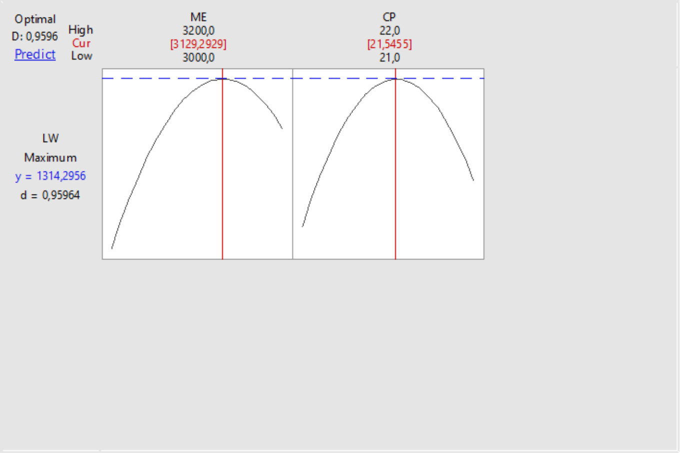


Figure 12. Optimization result of the ration in the grower period

3.1.3 Example of Finisher Period (25-50 day) Ration with CDD

The finisher ration meets the nutrient needs of the animals in the range of 25-50 days and it was determined in the range of 3150-3300 kcal ME and 17.5-19.5 % CP, and the maximum LW was tried to be reached. The dependent variable (LW) values corresponding to the ME and CP combinations were presented in Table 7.

Tablo 7. LW results for ME and CP combinations

StdOrder	RunOrder	PtType	Blocks	ME	CP	LW
4	1	1	1	3300	19.5	3600
7	2	-1	1	3225	17.5	3550
11	3	0	1	3225	18.5	3675
10	4	0	1	3225	18.5	3680
9	5	0	1	3225	18.5	3670
8	6	-1	1	3225	19.5	3570
3	7	1	1	3150	19.5	3500
12	8	0	1	3225	18.5	3685
5	9	-1	1	3150	18.5	3520
13	10	0	1	3225	18.5	3675
2	11	1	1	3300	17.5	3615
6	12	-1	1	3300	18.5	3650

In the finisher ration created with CDD; ME and CP were modeled via nonlinear regression in terms of LW values (Table 8). In the created model, the independent variables except for the main effect of CP were found to be statistically significant (P<0.05). It was determined that there was a relationship between LW and ME and CP as “LW=-112770

+ 56.0ME + 2706CP – 0.00857ME^2- 73.2 CP^{2}”. When the model performance was examined, the coefficient of determination (R^2) was found to be 0.8851.

Tablo 8. The coefficients and performance parameters of the prediction model of the LW corresponding to the ME and CP levels

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	3664.5	12.0	306.40	0.000	
ME	54.2	11.8	4.61	0.002	1.00
CP	-2.5	11.8	-0.21	0.837	1.00
ME^2	-48.2	17.3	-2.78	0.024	1.17
CP^2	-73.2	17.3	-4.22	0.003	1.17

The LW changes according to the ME and CP amount of the finisher ration are shown in Figure 13 and Figure 14. It is seen that the effect of the amount of ME and CP on the LW values is in the form of a curve. It is seen that the quadratic terms ME and CP were statistically significant (Table 8). While the LW reaches its maximum value when the ME amount reaches about 3250 kcal, there was a decrease in the LW amount after this value (Figure 14). While the LW was at its maximum when CP reached approximately 18.5 %, this increase decelerated over time. This shows that there is no effect of the first-order term of the CP ratio, but the effect of the second-order term.

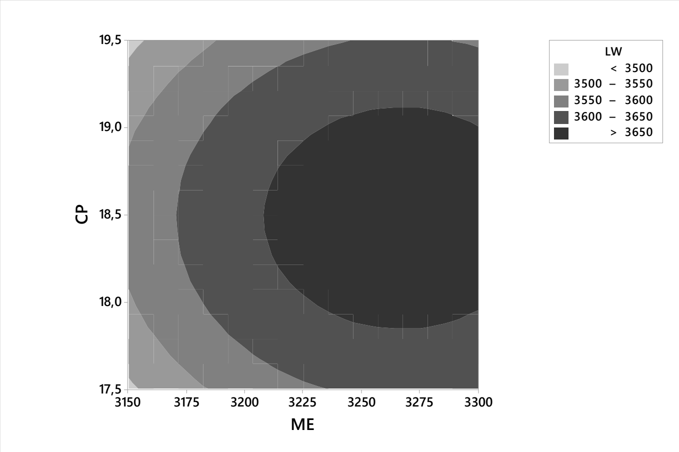


Figure 13. Contour plot representation of the effect of ME and CP changes on LW

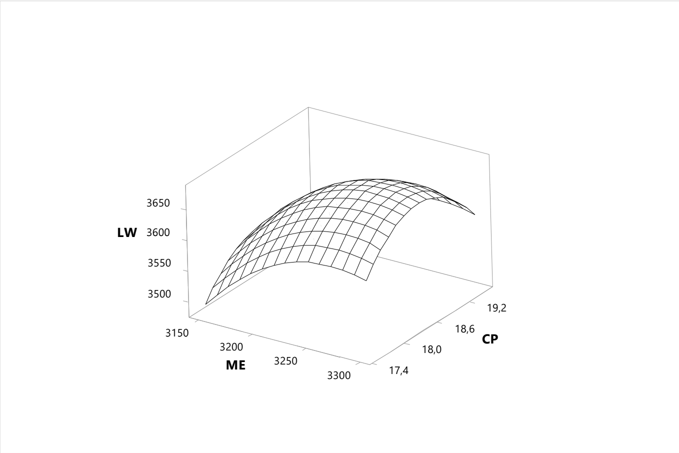


Figure 14. Surface plot representation of the effect of ME and CP changes on LW

In the finisher ration, the optimum amounts of ME and CP were revealed by maximizing LW (Figure 15). As a result, when broilers at the end of the grower period consumed the ration prepared at 3266.6667 ME and 18.4835 CP, it was estimated that it would be 3679.7233 g with a 5 % error.

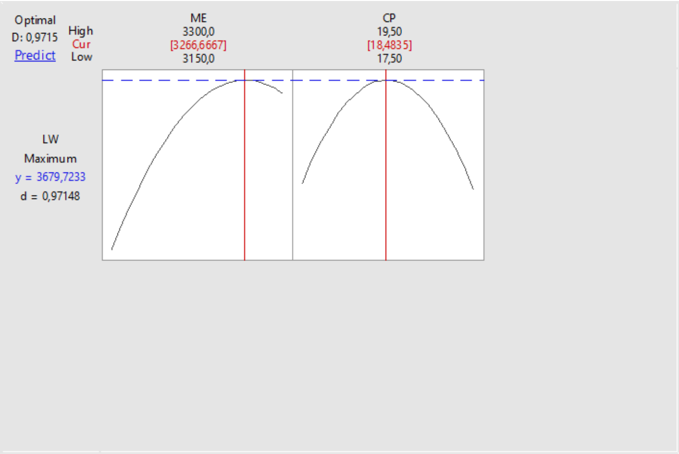


Figure 15. Optimization result of the ration in the finisher period

3.2. Box-Behnken Design Applications

As in CDD, three different ration samples were created in BBD for starter, grower, and finisher periods. Optimum amounts of ME, CP, and Ca were determined to maximize live weight (LW) at the end of each period. A total of 15 points were determined in the BBD for all ration samples and the analysis was performed by taking the α value of 1.

3.2.1 Example of Starter Period (0-10 days) Ration with BBD

In the experiment created in BBD, the maximum LW was tried to be reached by taking the ration that meets the nutrient needs of the animals in the range of 0-10 days, in the range of ME 2950-3050 kcal, CP 22-24 %, and Ca 0.9-1.1 %. Response variables (LW) values corresponding to experimental ME, CP, and Ca combinations are given in Table 9.

Tablo 9. LW results for ME, CP, and Ca combinations

StdOrder	RunOrder	PtType	Blocks	ME	CP	Ca	LW
12	1	2	1	3000	24	1.1	308
1	2	2	1	2950	22	1.0	305
4	3	2	1	3050	24	1.0	317
2	4	2	1	3050	22	1.0	308
14	5	0	1	3000	23	1.0	323
8	6	2	1	3050	23	1.1	310
5	7	2	1	2950	23	0.9	308
3	8	2	1	2950	24	1.0	315
13	9	0	1	3000	23	1.0	326
10	10	2	1	3000	24	0.9	312
7	11	2	1	2950	23	1.1	307
9	12	2	1	3000	22	0.9	304
6	13	2	1	3050	23	0.9	316
11	14	2	1	3000	22	1.1	307

LW values were modeled by the independent variables ME, CP and Ca via nonlinear regression (Table 10). While the main effects of the model, ME and CP, were statistically significant ($P<0.05$), the main effects of Ca were statistically nonsignificant ($P>0.05$). Besides, while ME2 and CP2 quadratic effects were statistically significant ($P<0.05$), Ca2 quadratic effects and CPxCa interaction were found to be nonsignificant ($P>0.05$). A relationship was found between LW and ME, CP, and Ca as “LW=-25024 + 13.14ME + 387.1CP + 2184Ca – 0.002183ME^2– 7.96CP^2 - 896Ca^2– 17.5CPxCa”. When the model performance was evaluated, it was determined that 95.60 % of the change in LW was due to the change in the amount of ME, CP, and Ca (R^2 =0.9560).

Tablo 10. The coefficients and performance parameters of the prediction model of the LW corresponding to the ME, CP and Ca levels

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	324.67	1.25	260.31	0.000	
ME	2.000	0.764	2.62	0.034	1.00
CP	3.500	0.764	4.58	0.003	1.00
Ca	-1.000	0.764	-1.31	0.232	1.00
ME^2	-5.46	1.12	-4.86	0.002	1.01
CP^2	-7.96	1.12	-7.08	0.000	1.01
Ca^2	-8.96	1.12	-7.97	0.000	1.01
CPxCa	-1.75	1.08	-1.62	0.149	1.00

The LW changes at one level of ME, CP and Ca of the starter ration created with BDD are given between Figure 16 and Figure 21. At the Ca=1 % level, when ME rised by about 3000 kcal, the amount of LW increases, while there is a decrease after this critical value. At the Ca=1 % level, it has seen that the effect of the CP amount on the LW values was curved. At the CP=23 level, it has revealed that the effect of ME and Ca on LW values is curved and the quadratic term is statistically significant (Table 10). At the ME=3000 kcal level, the amount of CP peaked about 23 % and there was a decrease in LW after this critical value. At ME=3000 kcal, LW peaked when Ca was around 1%, and with the increase of this value, it caused weight loss of the animals.

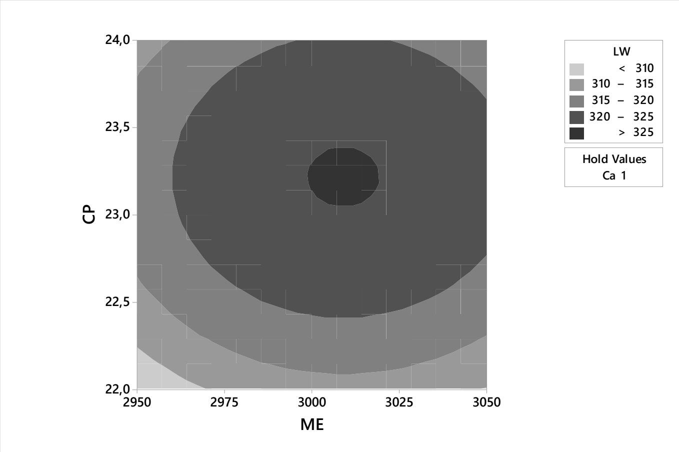


Figure 16. Contour plot representation of the effect of ME and CP changes on LW (Hold on Ca 1g)

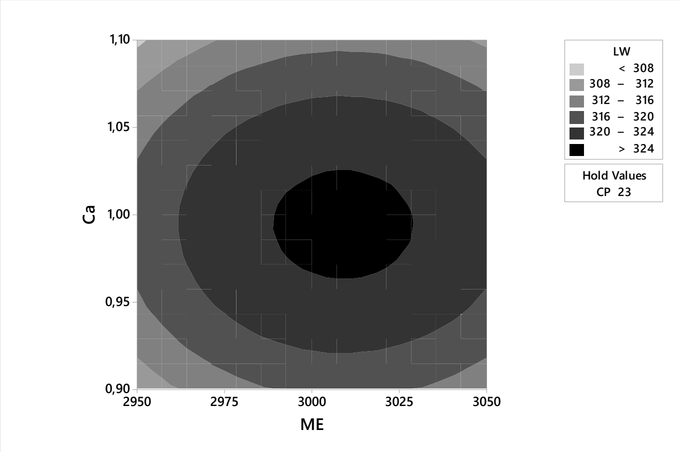


Figure 17. Contour plot representation of the effect of ME and Ca changes on LW (Hold on CP 23%)

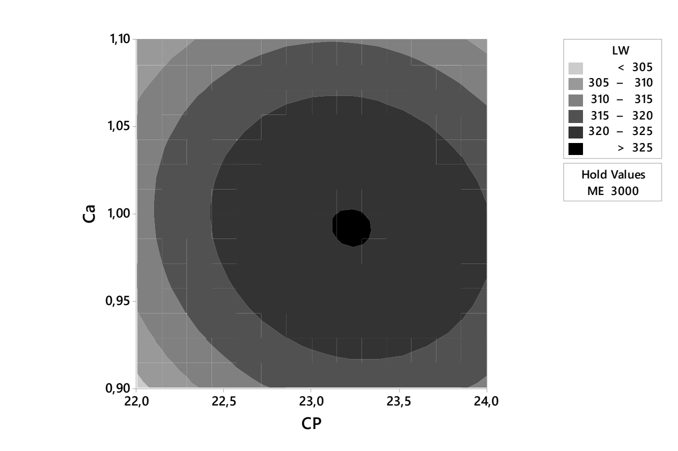


Figure 18. Contour plot representation of the effect of CP and Ca changes on LW (Hold on ME 3000 kcal)

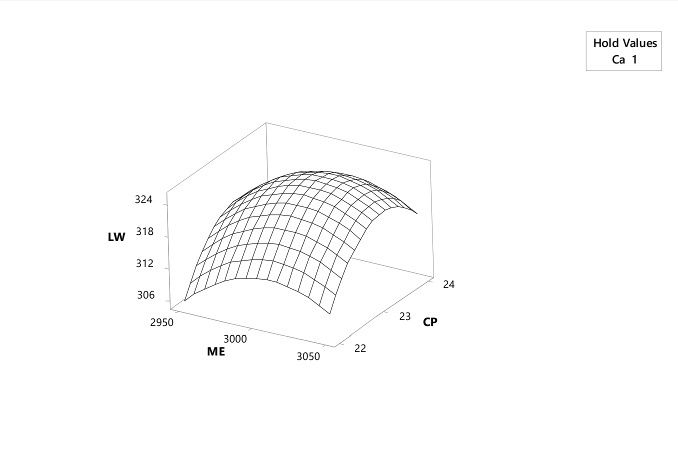


Figure 19. Surface plot representation of the effect of ME and CP changes on LW (Hold on Ca 1g)

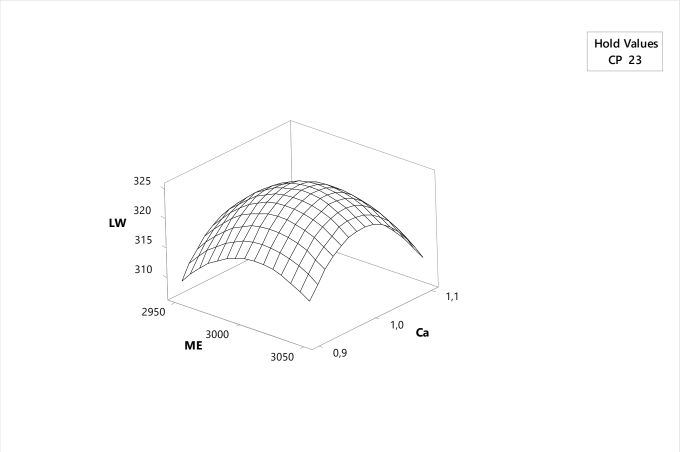


Figure 20. Surface plot representation of the effect of ME and Ca changes on LW (Hold on CP 23%)

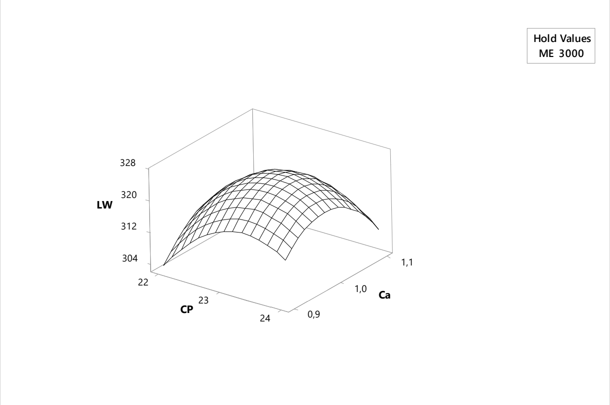


Figure 21. Surface plot representation of the effect of CP and Ca changes on LW (Hold on ME 3000 kcal)

In the study, the optimum amounts of ME, CP and Ca were determined by maximizing LW in the starter ration prepared using BBD (Figure 22). It was estimated that broilers consuming the ration prepared with 3005.9199 kcal ME, 23.1776 % CP and 0.9868 % Ca would have a maximum weight of 322.42-328.15 g with 95 % probability.

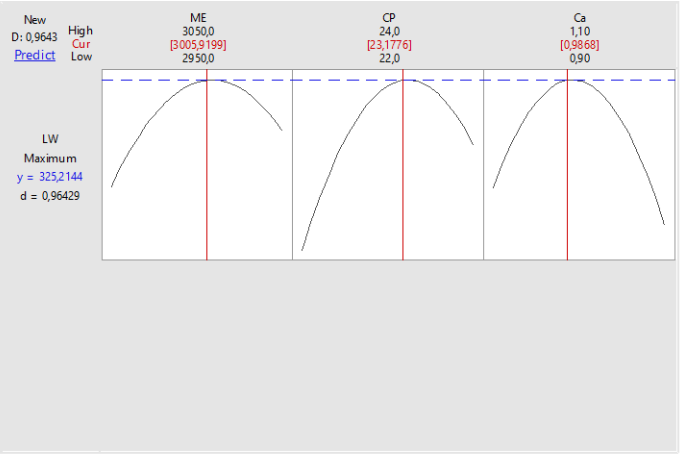


Figure 22. Optimization result of the ration in the starter period

3.2.2 Example of Grower Period (10-25 day) Ration with BBD

The maximum LW was tended to be reached by taking the ration that meets the nutrient needs of the animals in the range of 10-25th days, by taking the ration in the range of ME 3000-3200 kcal, CP 21-22 %, and Ca 0.80-0.95 %. The dependent variables (LW) values corresponding to the trial ME, CP, and Ca combinations were presented in Table 11.

Tablo 11. LW results for ME, CP, and Ca combinations

StdOrder	RunOrder	PtType	Blocks	ME	CP	Ca	LW
12	1	2	1	3100	22.0	0.950	1295
8	2	2	1	3200	21.5	0.950	1305
4	3	2	1	3200	22.0	0.875	1310
14	4	0	1	3100	21.5	0.875	1315
9	5	2	1	3100	21.0	0.800	1290
1	6	2	1	3000	21.0	0.875	1300
13	7	0	1	3100	21.5	0.875	1316
11	8	2	1	3100	21.0	0.950	1298

5	9	2	1	3000	21.5	0.800	1293
2	10	2	1	3200	21.0	0.875	1295
10	11	2	1	3100	22.0	0.800	1302
7	12	2	1	3000	21.5	0.950	1300
6	13	2	1	3200	21.5	0.800	1305
3	14	2	1	3000	22.0	0.875	1296

LW values were modeled by the independent variables ME, CP, and Ca using nonlinear regression (Table 12). In the model, the main effects of ME and CP were statistically significant (P<0.05), while the main effects of Ca were statistically nonsignificant (P>0.05). In the model, the quadratic effects of all independent variables (ME^{2}, CP^{2} and Ca^{2}) and CPxCa interaction were statistically significant (P<0.05). There was a relationship between LW and ME, CP and Ca as "LW= - 20523 + 1,682ME + 1540CP + 5881Ca - 0,000563 ME^{2}- 40,50 CP^{2}- 1711 Ca^{2} + 0,0950MExCP - 0,233MExCa - 100,0CPxCa". When the model performance was examined, it was determined that 98.57% of the change in LW was due to the change in the amount of ME, CP, and Ca (R^{2} = 0.9857).

Tablo 12. The coefficients and performance parameters of the prediction model of the LW corresponding to the ME, CP and Ca levels

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	1316.00	1.00	1316.00	0.000	
ME	3.250	0.612	5.31	0.003	1.00
CP	2.500	0.612	4.08	0.010	1.00
Ca	1.000	0.612	1.63	0.163	1.00
ME^{2}	-5.625	0.901	-6.24	0.002	1.01
CP^{2}	-10.125	0.901	-11.23	0.000	1.01

Ca^2	-9.625	0.901	-10.68	0.000	1.01
CPxCa	4.750	0.866	5.48	0.003	1.00

The LW changes at one level of ME, CP and Ca of the grower ration are given between Figure 23 and Figure 28. At the level of Ca=0.875 %, while ME was about 3150 kcal, the amount of LW increased and there was a decrease after this critical value. At the Ca=0.875 % level, the LW value peaked at approximately 21.5 % CP, and there was a decrease after this value. At the CP=21.5 % level, it is seen that the effect of Ca on the LW values was curved and the quadratic term is statistically significant (Table 12). At the ME=3100 kcal level, the amount of CP peaked at about 21.5 % LW, and there was a decrease in LW after this value. At ME=3100 kcal, Ca peaked at about 0.90 % LW, and when Ca was more than 0.90%, there was a decrease in LW in animals.

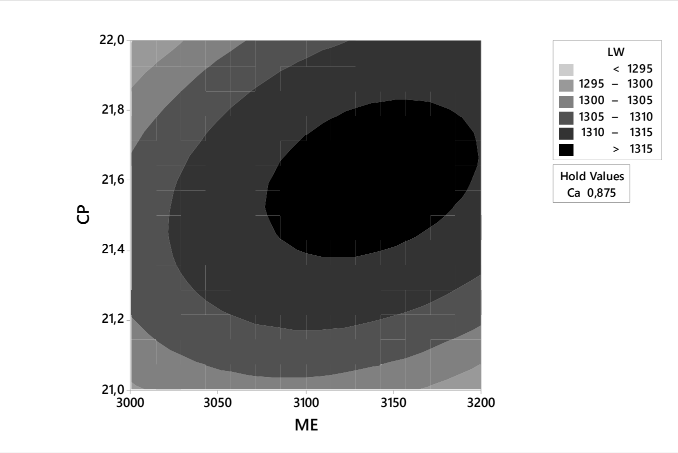


Figure 23. Contour plot representation of the effect of ME and CP changes on LW (Hold on Ca 0.875)

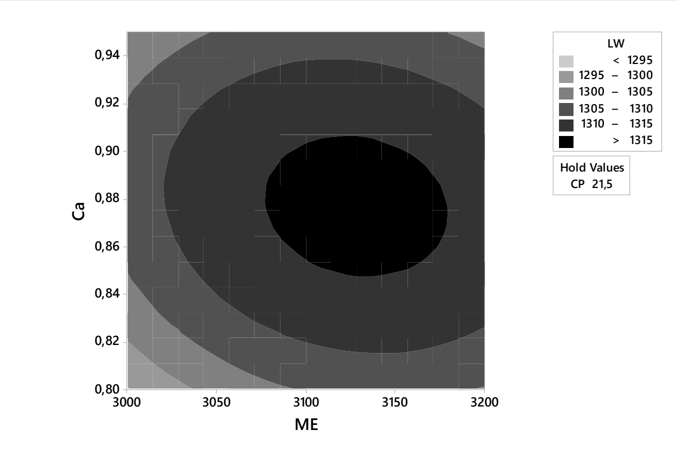


Figure 24. Contour plot representation of the effect of ME and Ca changes on LW (Hold on CP 21.5%)

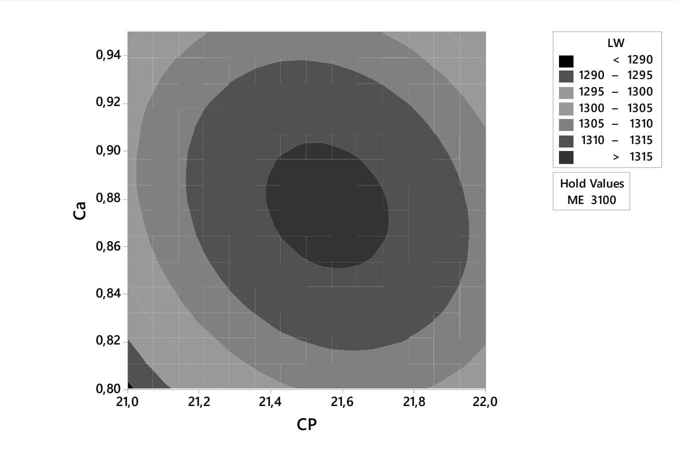


Figure 25. Contour plot representation of the effect of CP and Ca changes on LW (Hold on ME 3100 kcal)

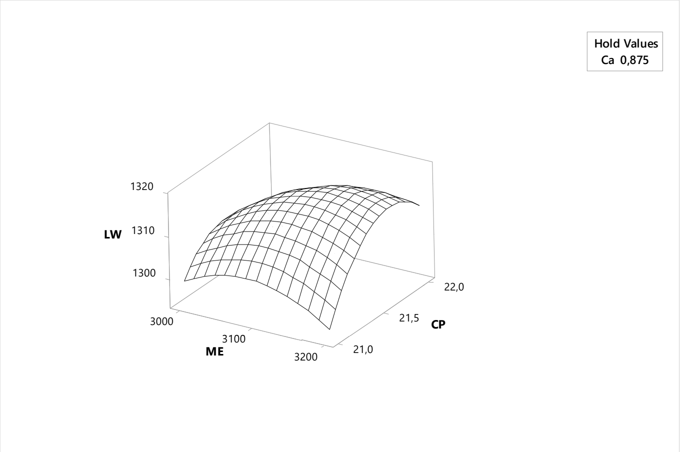


Figure 26. Surface plot representation of the effect of ME and CP changes on LW (Hold on Ca 0.875)

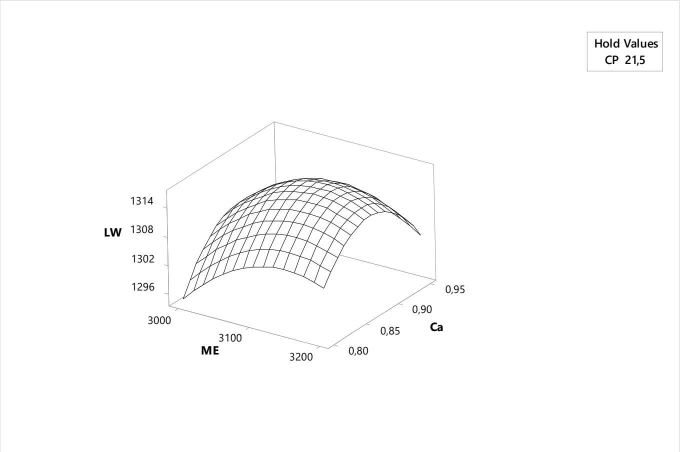


Figure 27. Surface plot representation of the effect of ME and Ca changes on LW (Hold on CP 21.5%)

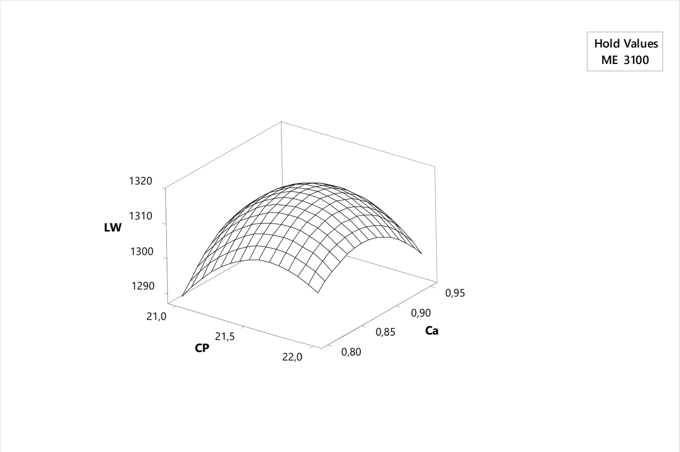


Figure 28. Surface plot representation of the effect of CP and Ca changes on LW (Hold on ME 3100 kcal)

In the grower ration, the LW maximum value was reached and the optimum values of ME, CP and Ca were found (Figure 7). As a result, it was estimated that broilers consuming the ration prepared with 3137.3737 kcal ME, 21.6061% CP and 0.8727% Ca amount could reach a maximum of 13166.885 g with 5% error.

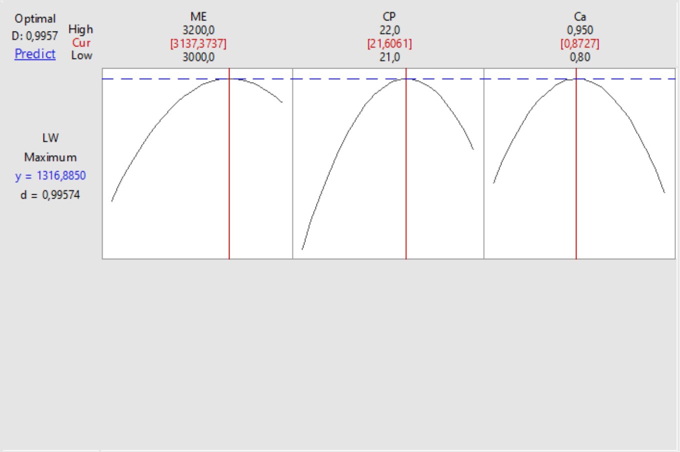


Figure 29. Optimization result of the ration in the grower period

3.2.3 Example of Finisher Period (25-50 day) Ration with BBD

The maximum LW was tended to be reached by taking the ration that meets the nutrient needs of the animals in the finisher period in the 25-50th days, by taking the ration in the range of ME 3150-3300 kcal, CP 17.5-19.5 %, and Ca 0.70-0.85 %. In the study, LW values corresponding to ME, CP and Ca combinations are presented in Table 13.

Tablo 13. LW results for ME, CP, and Ca combinations

StdOrder	RunOrder	PtType	Blocks	ME	CP	Ca	LW
1	1	2	1	3150	17.5	0.775	3600
5	2	2	1	3150	18.5	0.700	3580
10	3	2	1	3225	19.5	0.700	3590
9	4	2	1	3225	17.5	0.700	3595
14	5	0	1	3225	18.5	0.775	3680
11	6	2	1	3225	17.5	0.850	3625
13	7	0	1	3225	18.5	0.775	3685
6	8	2	1	3300	18.5	0.700	3590
3	9	2	1	3150	19.5	0.775	3595
12	10	2	1	3225	19.5	0.850	3620
7	11	2	1	3150	18.5	0.850	3610
8	12	2	1	3300	18.5	0.850	3650
2	13	2	1	3300	17.5	0.775	3625
15	14	0	1	3225	18.5	0.775	3683

LW values were modeled by the independent variables ME, CP, and Ca with the help of nonlinear regression (Table 14). As the main effects of the nonlinear model, ME and Ca were statistically significant ($P<0.05$), CP was found to be statistically insignificant ($P>0.05$). In this model, all quadratic effects and CPxCa interaction were found to be statistically significant ($P<0.05$). There was a relationship between LW and ME, CP and Ca as “LW= -77432 + 40.80ME + 1341.8CP + 6651Ca – 0.006459ME^2 – 36.33CP^2 - 6904Ca^2 + 1.333MExCa”. When the model performance was evaluated, it was determined that 99.29% of the change in LW was due to the change in the amount of ME, CP and Ca (R^2 =0.9929).

Tablo 14. The coefficients and performance parameters of the prediction model of the LW corresponding to the ME, CP and Ca levels

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	3682.67	2.44	1508.44	0.000	
ME	12.50	1.50	8.36	0.000	1.00
CP	-2.50	1.50	-1.67	0.138	1.00
Ca	18.75	1.50	12.54	0.000	1.00
ME^2	-36.33	2.20	-16.51	0.000	1.01
CP^2	-36.33	2.20	-16.51	0.000	1.01
Ca^2	-38.83	2.20	-17.65	0.000	1.01
CPxCa	7.50	2.11	3.55	0.009	1.00

LW changes at one level of ME, CP and Ca of the finisher ration developed with BDD were presented between Figure 30 and Figure 35. At the Ca=0.775 % level, the effects of the amount of ME and CP on the LW values were determined to be curved. At CP=18.5 %, while ME increased by approximately 3225 kcal of LW, there was a decrease in

LW after this value. At CP=18.5 %, the amount of LW increases when Ca reached approximately 0.8 %, while there was a decrease in the amount of LW at values greater than 0.8 %. At ME=3225 kcal, the amount of CP and Ca peaked around 18.50 % and 0.80 %, respectively, and after these values, LW caused weight loss.

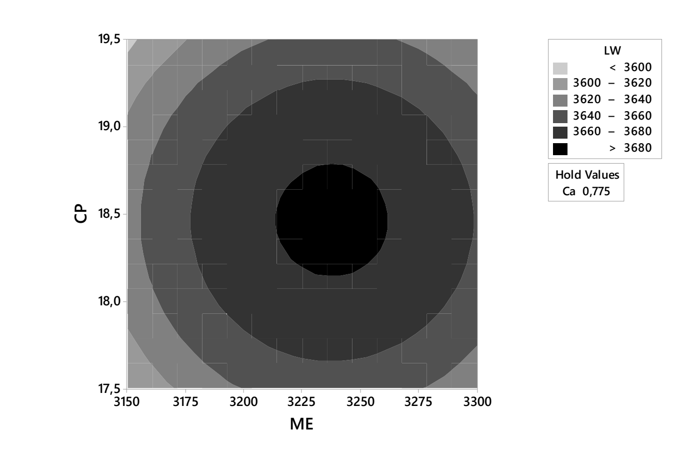


Figure 30. Contour plot representation of the effect of ME and CP changes on LW (Hold on Ca 0.775)

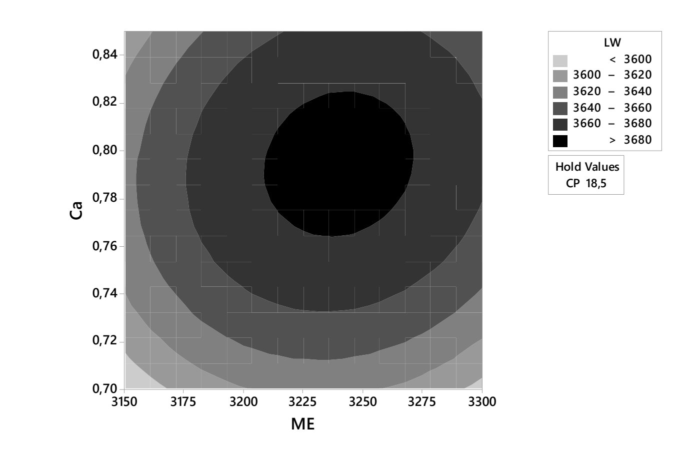


Figure 31. Contour plot representation of the effect of ME and Ca changes on LW (Hold on CP 18.5%)

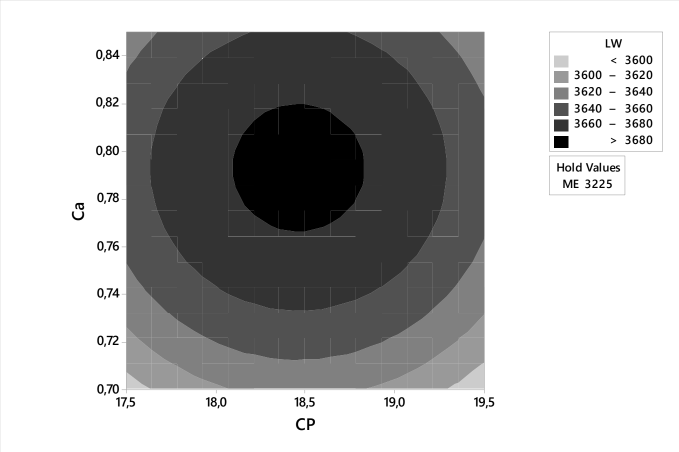


Figure 32. Contour plot representation of the effect of CP and Ca changes on LW (Hold on ME 3225 kcal)

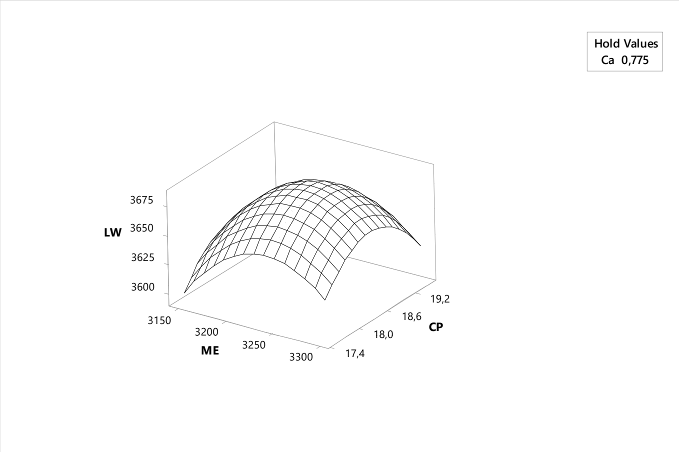


Figure 33. Surface plot representation of the effect of ME and CP changes on LW (Hold on Ca 0.775)

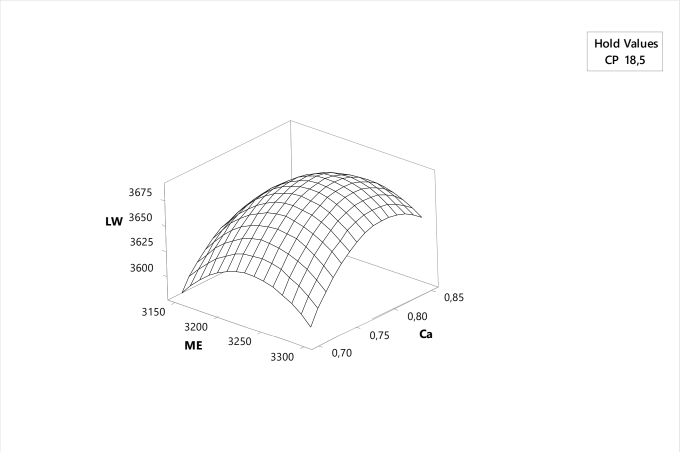


Figure 34. Surface plot representation of the effect of ME and Ca changes on LW (Hold on CP 18.5%)

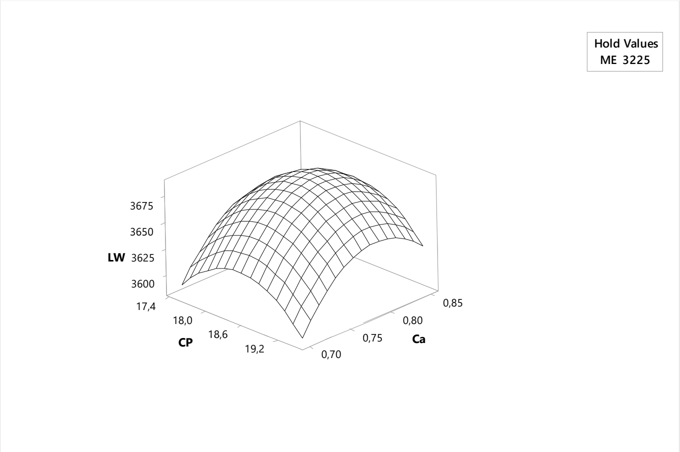


Figure 35. Surface plot representation of the effect of CP and Ca changes on LW (Hold on ME 3225 kcal)

In the finisher ration obtained using BBD, the optimum values of ME, CP and Ca were found by maximizing LW (Figure 7). As a result, it was estimated that the animals consuming the ration prepared with 3239.3939 kcal ME, 18.4697 % CP and 0.7939 % Ca amount would have an average weight of 3686.39 g, with a probability of 95%, in the range of maximum 3680.79-3692.0 g.

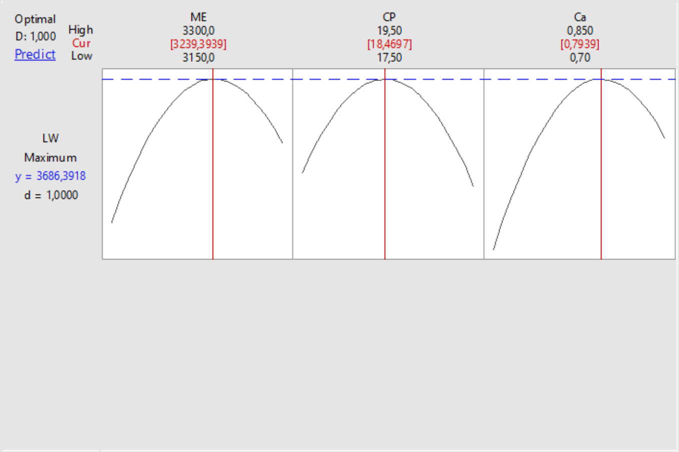


Figure 36. Optimization result of the ration in the finisher period

CONCLUSION

The RSM optimization method can be used as a preliminary step before planning a factorial experiment, as factor levels determine optimum points. In this way, the distances between the factor levels are shortened by providing robust results. In other words, the optimization is used to increase the significance and sensitivity of factor levels. In addition, appropriate transformations can be applied to non-normally distributed data sets. In this way, it becomes a robust test and enables more reliable results to be obtained.

In the preparation of poultry rations that require precise calculation in the field of livestock, it is possible to obtain the highest efficiency from the animals by preparing an optimum ration with fewer materials by using RSM. As a result, while the 10th-day maximum LW of broilers fed with 3009.60 kcal ME and 23.4171% CP in the starter ration prepared with CCD was 322.27 g, the ration prepared with BBD was estimated at 325.2144 g LW using 3005.9199 kcal ME, 23.1776% CP, and 0.9868% Ca. While the maximum LW of the animals is 1314.2956 g in the ration prepared with CDD at 3129.29 kcal ME and 21.5455% CP for the grower period, it has been estimated that the broilers consuming the ration prepared with BBD with the amount of 3137.3737 kcal ME, 21.6061% CP, and 0.8727% Ca reached a maximum of LW 1316.885 g. In the Finisher period, the optimum ME of the ration prepared with CCD was 3266.6667 kcal, and its CP was 18.4835% and its maximum LW was 3679.7233 g, while In BBD, broilers are estimated to have ME 3239.3939 kcal, CP 18.4697%, and Ca 0.7939% and maximum LW 3686.3918 g. According to these results, it is recommended to use CCD and BBD in the production of poultry rations economically.

REFERENCES

Ahmadi, H. & Golian, A. (2011). Response surface and neural network models for performance of broiler chicks fed diets varying in digestible protein and critical amino acids from 11 to 17 days of age. Poultry science, 90(9), 2085-2096.

Banga, J. R., Balsa-Canto, E., Moles, C. G. & Alonso, A. A., (2003), Improving food processing using modern optimization methods, Trends in Food Science & Technology, 14 (4), 131-144.

Box G. E., & Wilson K. B., (1951) On the experimental attainment of optimum conditions. J Roy Statist Soc Ser B 13(1):1–45 https://www.jstor.org/stable /2983966.

Box, G. E. & Hunter, J. S., (1957), Multi-factor experimental designs for exploring response surfaces, The Annals of Mathematical Statistics, 28 (1), 195-241.

Box, G. E., & Draper, N. R. (1959). A basis for the selection of a response surface design. Journal of the American Statistical Association, 54(287), 622-654.

Box, G. E. & Draper, N. R., (1987), Empirical model-building and response surfaces, Wiley New York, p.

Box, G. E. & Wilson, K. B., (1992), On the experimental attainment of optimum conditions, In: Breakthroughs in statistics, Eds: Springer, p. 270-310.

Faria Filho, D. E., Rosa, P. S., Torres, K. A. A., Macari, M., & Furlan, R. L. (2008). Response surface models to predict broiler performance and applications for economic analysis. Brazilian Journal of Poultry Science, 10(2), 139-141.

Faridi, A., Golian, A., France, J., & Heravi Mousavi, A. (2013). Study of broiler chicken responses to dietary protein and lysine using neural network and response surface models. British poultry science, 54(4), 524-530.

Fenwick, D., Scheidt, C., & Caers, J. (2014). Quantifying asymmetric parameter interactions in sensitivity analysis: application to reservoir modeling. Mathematical Geosciences, 46(4), 493-511.

Koç, B. & Kaymak-Ertekin, F., (2010), Yanıt Yüzey Yöntemi veGıda İşleme Uygulamaları, Gıda, 35 (1), 1-8.

Mead, R., & Pike, D. J. (1975). A biometrics invited paper. A review of response surface methodology from a biometric viewpoint. Biometrics, 31(4), 803-851.

Montgomery, D., C., (2001), Response surface methodology: process and product optimization using designed experiments, John Wiley & Sons.

Montgomery, D., C., (2017). Design and analysis of experiments. 10th ed., John Wiley & Sons, USA.

Myers, R.H., Montgomery, D.C., (2002). Response surface methodology: Process and product opti-mization using designed experiments, 2nd ed., John Wiley & Sons, USA.

Sanders, A. M., Edwards Jr, H. M., & Rowland III, G. N. (1992). Calcium and phosphorus requirements of the very young turkey as determined by response surface analysis. British journal of nutrition, 67(3), 421-435.

Roush, W., Petersen, R. G. & Arscott, G. H., (1979), An application of response surface methodology to research in poultry nutrition, Poultry Science, 58 (6), 1504-1513.

Roush, W. B. (1983). An investigation of protein levels for broiler starter and finisher rations and the time of ration change by response surface methodology. Poultry Science, 62(1), 110-116.

Toyomizu, M., Akiba, Y., Horiguchi, M., & Matsumoto, T. (1982). Multiple regression and response surface analyses of the effects of dietary protein, fat and carbohydrate on the body protein and fat gains in growing chicks. The Journal of nutrition, 112(5), 886-896.

Walker, C. E. (1984). Response surface analysis of baked-lab data with a personal computer. Cereal foods world, 29, 662-666.

Yavuz, C., & Keskin, İ. (2020). Using The Response Surface Method to Determine Optimum Temperature and Gam Usage in Egg Storage. Selcuk Journal of Agriculture and Food Sciences, 34(3), 178-182.

Yılmaz, M. T. (2002). Nitrit, glukono delta lakton ve askorbik asidin sucuğun bazı özellikleri üzerindeki etkisinin yanıt yüzeyi yöntemi ile modellenmesi. Yüzüncüyıl Üniversitesi.

CHAPTER II

TIME SERIES ANALYSIS AND AN APPLICATION IN AGRICULTURE

Assoc. Prof. Dr. Şenol Çelik[*]

[*] Bingöl University Faculty of Agriculture Animal Science, Biometry and Genetic Department Bingöl-Turkey, E-mail: senolcelik@bingol.edu.tr, ORCID ID: 0000-0001-5894-8986

INTRODUCTION

Time series is a significant subject that can be applied in econometrics, geophysics, meteorology, business, and economy, particularly in statistics. A time series is a sequence of measurements observed over time. Data such as monthly inflation rate in a country, annual export and import amounts of a country, annual investment and GNP revenues, annual unemployment rate, monthly precipitation in a place are examples of time series. It would be more appropriate to make production and sales plans by determining the production and sales amounts of the products grown region seasonally.

In time series study, the appropriate determination of the model and the suitability of the determined model to the data is important. A model which is incorrectly determined does not yield good results. After the model determination phase, the suitability of the determined model to the data should be tested. A better prediction can be made by creating the appropriate model (Akdi, 2010).

The concept of time series can be defined as a collection of random variables. Time series can be expressed as T , $\{X_t: t \in T\}$ being a set of indices. (Ω,U,P) being a probability space and T a set of indices, a time series (Ω*T) is a function going from a product space to real numbers. In short, time series is defined as

$$X(.,.): \Omega * T \to R$$

$$(w, f) \to X(w, t)$$

Time series is shown by $X(w, t)$ or X_t (Akdi, 2010).

The series consists of trend, seasonal variations and irregular components. The components that make up a series such as x_t are expressed by three equations.

i) Trend: $T_t = \beta_1 + \beta_2 t$ e.g: $(T_t = 1 + 0.1t)$

ii) Seasonal: $S_t = \beta_1 \sin\left(\frac{t\pi}{2}\right)$ e.g: $S_t = 1.6 \sin\left(\frac{t\pi}{6}\right)$

iii) Irregular: $I_t = \beta_1 I_{t-1} + \varepsilon_t$ e.g: $(I_t = 0.7 I_{t-1} + \varepsilon_t)$

T_t: value of trend component in t period

S_t: seasonal component in t period

I_t: irregular component in t period

ε_t: error in t period

(Enders, 2010).

SIGNIFICANT CONCEPTS ABOUT TIME SERIES

Trend, means the movement (tendency) of a series up and down, that is, in the direction of increase and decrease, over a certain period.

Stationary

$X_t: t \epsilon T$ with X_t time series is stationary if the following conditions are met (Günay et al., 2007).

- $E(X) = \mu$ (Expected value does not change according to time)
- $V(X_t) = \sigma^2$
- $Cov(X_t, X_{t+h})$ Its covariance is dependent on h and its time variable is independent of t.

Autocorrelation function and properties

The autocorrelation coefficient expresses the relationships between time series and the lagged series of that series. In short, there is a relationship between the (X_1, X_{1+h}), (X_2, X_{2+h}), …, (X_{t-h}, X_t) it is the correlation of the h^th lag. The autocorrelation function of a stationary time series of X_t:t∈T is defined as

$$\rho(h)=\frac{\gamma(h)}{\gamma(0)}=\frac{Cov(X_t,X_{t+h})}{\sqrt{Var(X_t)Var(X_{t+h})}}=\frac{\sum_{i=h+1}^{n}(X_t-\bar{X}_t)(X_{t+h}-\bar{X}_t)}{\sum_{i=1}^{n}(X_t-\bar{X}_t)^2}$$

(Wei, 2006; Box and Jenkins, 1976). The autocorrelation function has the following characteristics.

- $\rho(h)$. In other words, $\rho(-h)$= $\rho(h)$
- $|\rho(h)|$ for ∀ h≤1
- $\rho(h)=1$ (Wei, 2006).

Partial autocorrelation function

When X_t of $X_{t-1,}$ X_{t-2},…,X_{th} regression equation with respect to X is found, the coefficient of X_{th} is defined as the h^th partial autocorrelation, and is indicated as $\phi(h)$. The h^th partial autocorrelation of a time series is as follows:

$$\phi(h) = \frac{\begin{vmatrix} 1 & \rho_1 & \rho_2 & \dots & \rho_{h-2} & \rho_1 \\ \rho_1 & 1 & \rho_1 & \dots & \rho_{h-3} & \rho_2 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ \rho_{h-1} & \rho_{h-2} & \rho_{h-3} & \dots & \rho_1 & \rho_h \end{vmatrix}}{\begin{vmatrix} 1 & \rho_1 & \rho_2 & \dots & \rho_{h-2} & \rho_{h-1} \\ \rho_1 & 1 & \rho_1 & \dots & \rho_{h-3} & \rho_{h-2} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ \rho_{h-1} & \rho_{h-2} & \rho_{h-3} & \dots & \rho_1 & 1 \end{vmatrix}}$$

Here the first partial autocorrelation is $\phi_{11} = \rho_1$,

Second partial autocorrelation is

$$\phi_{22} = \frac{\begin{vmatrix} 1 & \rho_1 \\ \rho_1 & \rho_2 \end{vmatrix}}{\begin{vmatrix} 1 & \rho_1 \\ \rho_1 & 1 \end{vmatrix}}$$

Third partial autocorrelation is

$$\phi_{33} = \frac{\begin{vmatrix} 1 & \rho_1 & \rho_1 \\ \rho_1 & 1 & \rho_2 \\ \rho_2 & \rho_1 & \rho_3 \end{vmatrix}}{\begin{vmatrix} 1 & \rho_1 & \rho_2 \\ \rho_1 & 1 & \rho_1 \\ \rho_2 & \rho_1 & 1 \end{vmatrix}}$$

(Wei, 2006).

Stationary Time Series

Stationary time series consist of moving average (MA), autoregressive (AR), and autoregressive moving average (ARMA) models.

Moving average (MA(q)) series

If moving average (MA(q)) series is a stationary process with an autocovariance function such as X$_t$, mean zero, $|h|$ for >q and $\gamma(q) \neq 0$ then,

$$X_t = \mu - \theta_1\varepsilon_{t-1} - \theta_2\varepsilon_{t-2} - \cdots - \theta_q\varepsilon_{t-q} + \varepsilon_t$$

is in the form of MA(q), that is, the qth order moving average series. Here what is indicated as $\varepsilon_t{\sim}WN(0,\sigma^2)$ is the white noise series (Montgomery et al., 1990). In the qth order moving average series, that is, in MA(q), if q=1, first order moving average series, and if q=2, quadratic moving average series occurs. These series are denoted by MA(1) and MA(2), respectively.

First order moving average series, that is MA(1)

$$X_t = \mu - \theta_1\varepsilon_{t-1} + \varepsilon_t$$

is as follows. The autocovariance and autocorrelation function of this series are as follows, respectively.

$$\gamma(h) = \begin{cases} \sigma^2(1+\theta^2), & h = 0 \\ -\theta\sigma^2, & h = \mp 1 \\ 0, & |h| \geq 2 \end{cases}$$

$$\gamma(0) = V(X_t) = Var(\varepsilon_t - \theta\varepsilon_{t-1}) = \sigma^2(1+\theta^2)$$

Since autocorrelation function is $\rho(h) = \gamma(h)/\gamma(0)$

$$\rho(h) = \begin{cases} \dfrac{-\theta}{1+\theta^2}, & h = 1 \\ 0, & h > 1 \end{cases}$$

The autocorrelation function is cut off after the first lag, that is, it approaches zero (Montgomery et al., 1990). The 1st, 2nd and 3rd partial autocorrelation functions of the MA(1) series are expressed

$$\phi_{11} = \rho_1 = \frac{-\theta}{1+\theta^2} = \frac{-\theta(1-\theta^2)}{1-\theta^4}$$

$$\phi_{22} = \frac{\begin{vmatrix} 1 & \rho_1 \\ \rho_1 & 0 \end{vmatrix}}{\begin{vmatrix} 1 & \rho_1 \\ \rho_1 & 1 \end{vmatrix}} = \frac{-\rho_1^2}{1-\rho_1^2} = \frac{-\theta^2}{1+\theta^2+\theta^4} = \frac{-\theta^2(1-\theta^2)}{(1-\theta^6)}$$

$$\phi_{33} = \frac{\begin{vmatrix} 1 & \rho_1 & \rho_1 \\ \rho_1 & 1 & 0 \\ 0 & \rho_1 & 0 \end{vmatrix}}{\begin{vmatrix} 1 & \rho_1 & 0 \\ \rho_1 & 1 & \rho_1 \\ 0 & \rho_1 & 1 \end{vmatrix}} = \frac{-\rho_1^3}{1-2\rho_1^2} = \frac{-\theta^3}{1-2\rho_1^2} = \frac{-\theta^3(1-\theta^2)}{(1-\theta^8)}$$

as respectively. After the first lag, the autocorrelation function approaches zero, and the partial autocorrelation function decreases exponentially (Cooray 2008).

The autocorrelation function graph of the first-order moving average series MA(1) is as in Figure 1 and the partial autocorrelation function graph is as in Figure 2 (Çelik, 2013).

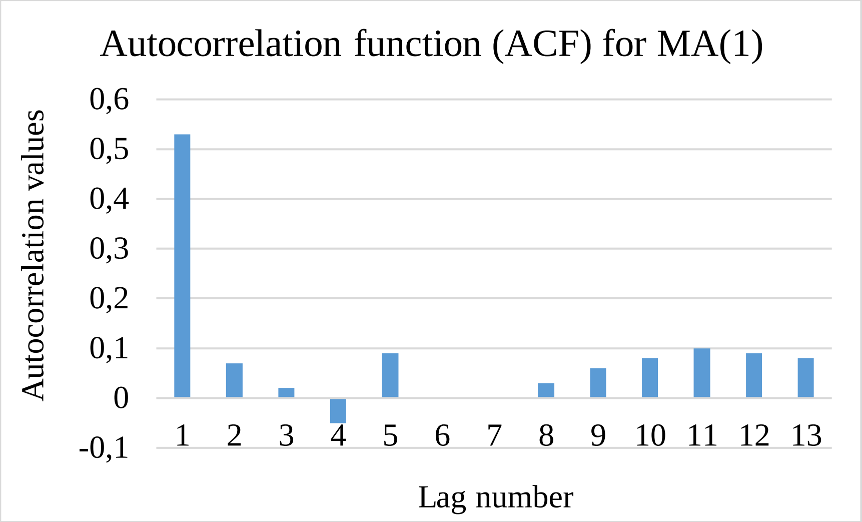


Figure 1. Graph of the autocorrelation function of the first-order moving average series MA(1)

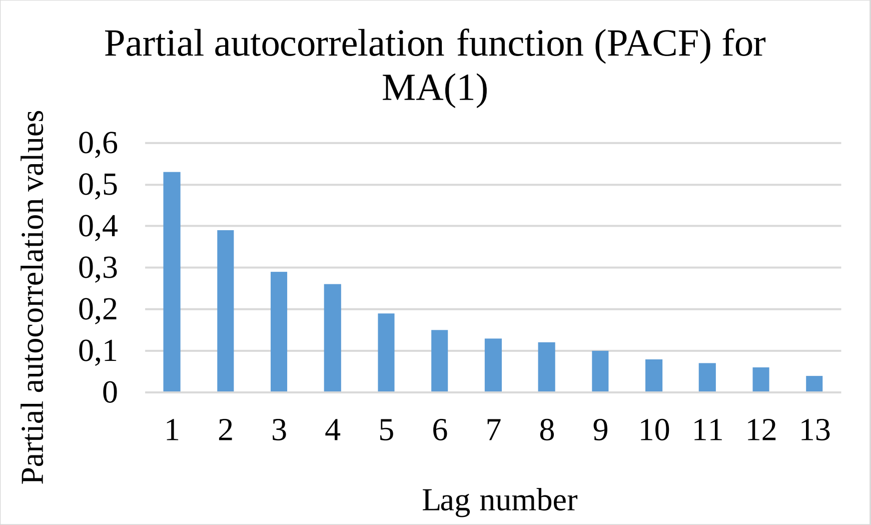


Figure 2. Graph of partial autocorrelation function of the first-order moving average series MA(1)

Quadratic moving average series of the second order, that is MA(2)

$$X_t = -\theta_1\varepsilon_{t-1} - \theta_2\varepsilon_{t-2} + \varepsilon_t$$

is expressed as and indicated by MA(2). The autocorrelation function of this series is as follows.

First autocorrelation coefficient

$$\rho_1 = \frac{-\theta_1(1-\theta_2)}{1+\theta_1^2+\theta_2^2}$$

Second autocorrelation coefficient

$$\rho_2 = \frac{-\theta_2}{1+\theta_1^2+\theta_2^2}$$

$$\rho_h = 0, \quad h > 2$$

is as (Montgomery et al., 1990). And the partial autocorrelation coefficients are as

$$\phi_{11} = \rho_1 = \frac{-\theta_1(1-\theta_2)}{1+\theta_1^2+\theta_2^2}$$

$$\phi_{22} = \frac{\begin{vmatrix} 1 & \rho_1 \\ \rho_1 & \rho_2 \end{vmatrix}}{\begin{vmatrix} 1 & \rho_1 \\ \rho_1 & 1 \end{vmatrix}} = \frac{\rho_2-\rho_1^2}{1-\rho_1^2}$$

$$\phi_{33} = \frac{\begin{vmatrix} 1 & \rho_1 & \rho_1 \\ \rho_1 & 1 & 0 \\ \rho_2 & \rho_1 & \rho_3 \end{vmatrix}}{\begin{vmatrix} 1 & \rho_1 & \rho_2 \\ \rho_1 & 1 & \rho_1 \\ \rho_2 & \rho_1 & 1 \end{vmatrix}} = \frac{-\rho_1^3-\rho_1\rho_2(2-\rho_2)}{1-\rho_1^2-2\rho_1^2(1-\rho_2)}$$

(Cooray, 2008). After the second lag, the autocorrelation function approaches zero, whereas the partial autocorrelation function decreases exponentially.

For the MA(2) model to be stationary,

a) $\theta_1 + \theta_2 < 1$

b) $\theta_2 - \theta_1 < 1$

c) $-1 < \theta_2 < 1$

conditions must be fulfilled (Wei, 2006). Otherwise, the series is made stationary by applying the operation of difference.

In general, autocorrelation (ACF) and partial autocorrelation (PACF) graphs for MA(2) are as in Figures 3 and 4 (Çelik, 2013).

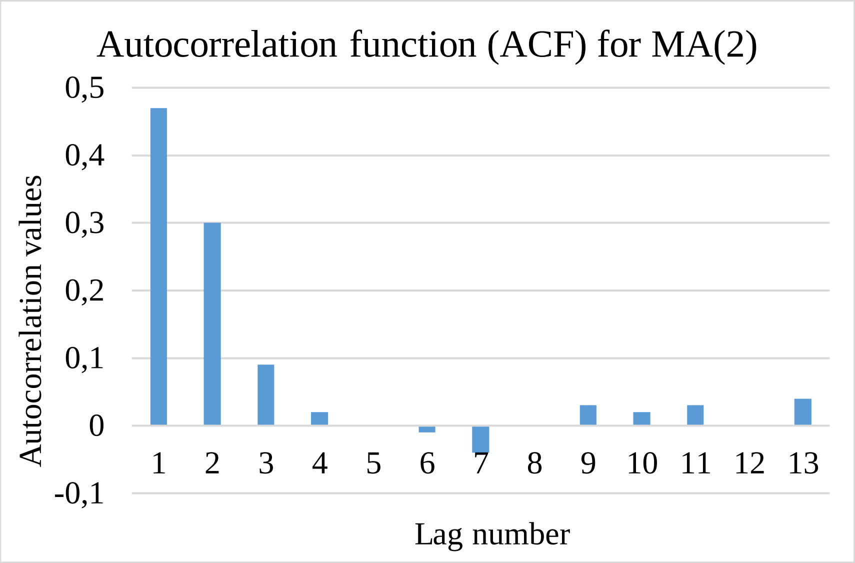


Figure 3. Graph of autocorrelation function for quadratic moving average series MA(2)

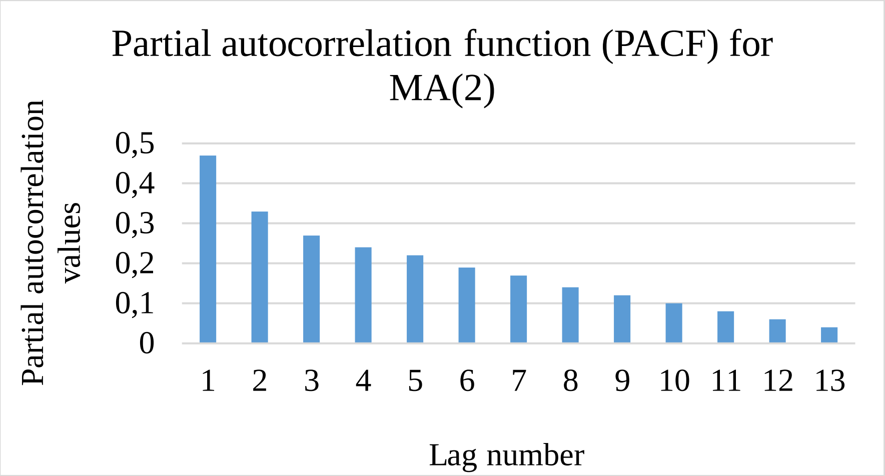


Figure 4. Graph of autocorrelation function for quadratic moving average series MA(2)

Autoregressive (AR) time series

A p-order autoregressive series, namely AR(p), is as follows (Hamilton, 1994):

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \cdots + \phi_p X_{t-p} + \varepsilon_t$$

In this series, when p=1, AR(1), that is, first-order autoregressive series occurs, and when p=2, AR(2), that is, second-order autoregressive time series occurs. While first order autoregressive series,

$$X_t = \phi_1 X_{t-1} + \varepsilon_t$$

is expressed as (Chatfield, 2000), quadratic autoregressive series

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \varepsilon_t$$

is written as (Cooray, 2008).

If autocorrelation function in AR(1) series

$$\rho_h = \phi \rho_{h-1} = \phi^h$$

and the partial autocorrelation function

$$\phi_{hh} = \begin{cases} \rho_1 = 0, & h = 1 \\ 0, & h \geq 2 \end{cases}$$

is as follows (Wei, 2006).

When the autocorrelation function in AR(1) series

$$\rho_h = \phi_1 \rho_{h-1} + \phi_2 \rho_{h-2}, \quad h \geq 1$$

becomes h=1 and h=2 here

$$\rho_1 = \phi_1 + \phi_2 \rho_1$$

$$\rho_2 = \phi_1 \rho_1 + \phi_2$$

and from here

$$\rho_1 = \frac{\phi_1}{1 - \phi_2}$$

$$\rho_2 = \frac{{\phi_1}^2}{1 - \phi_2} + \phi_2 = \frac{{\phi_1}^2 + \phi_2 - {\phi_2}^2}{1 - \phi_2}$$

is obtained (Cooray, 2008).

And the partial autocorrelation function

$$\phi_{11} = \rho_1 = \frac{\phi_1}{1 - \phi_2}$$

$$\phi_{22} = \frac{\begin{vmatrix} 1 & \rho_1 \\ \rho_1 & \rho_2 \end{vmatrix}}{\begin{vmatrix} 1 & \rho_1 \\ \rho_1 & 1 \end{vmatrix}} = \frac{\rho_2 - \rho_1^2}{1 - \rho_1^2} = \frac{\left(\frac{{\phi_1}^2 + \phi_2 - {\phi_2}^2}{1 - \phi_2}\right) - \left(\frac{\phi_1}{1 - \phi_2}\right)^2}{1 - \left(\frac{\phi_1}{1 - \phi_2}^2\right)}$$

$$= \frac{\phi_2\left[(1 - \phi_2)^2 - {\phi_1}^2\right]}{(1 - \phi_2)^2 - {\phi_1}^2} = \phi_2$$

$$\phi_{33} = \frac{\begin{vmatrix} 1 & \rho_1 & \rho_1 \\ \rho_1 & 1 & 0 \\ \rho_2 & \rho_1 & \rho_3 \end{vmatrix}}{\begin{vmatrix} 1 & \rho_1 & \rho_2 \\ \rho_1 & 1 & \rho_1 \\ \rho_2 & \rho_1 & 1 \end{vmatrix}} = \frac{\begin{vmatrix} 1 & \rho_1 & \phi_1 + \phi_2\rho_1 \\ \rho_1 & 1 & \phi_1\rho_1 + \phi_2 \\ \rho_2 & \rho_1 & \phi_1\rho_2 + \phi_2\rho_1 \end{vmatrix}}{\begin{vmatrix} 1 & \rho_1 & \rho_2 \\ \rho_1 & 1 & \rho_1 \\ \rho_2 & \rho_1 & 1 \end{vmatrix}} = 0$$

is as (Shumway and Stoffer, 2006).

Autoregressive moving average series (ARMA)

Series that can be expressed together with both AR(p) and MA(q) processes are shown as

$$\begin{aligned} X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \cdots + \phi_p X_{t-p} - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \cdots \\ - \theta_q \varepsilon_{t-q} + \varepsilon_t \end{aligned}$$

autoregressive moving average ARMA(p,q) model (Cryer 1986). Here $X_t \sim WN(0, \sigma^2)$is.

When q=0 in ARMA(p,q) model it is called pth order autoregressive model, and when p=0, it is called qth order moving average model.

Non-Stationary Time Series

In practice, it is not always possible for time series to be stationary. Such series are usually made stationary by taking the first or second difference. In short, when ARMA models become stationary by taking the difference, they turn into ARIMA models and are indicated by

ARIMA(p, d, q). Here p: expresses autoregressive degree, d: difference (1st or 2nd difference, ie d=1 or d=2), q: the degree of moving average series.

General ARIMA(p, d, q) model

Generally, to obtain ARMA(p, q) time series which is defined as $\phi_p(B)X_t = \theta_q(B)\varepsilon_t$, $(1-B)^d X_t$ difference series is applied. Thus,

$$\phi_p(B)(1-B)^d X_t = \theta_0 + \theta_q(B)\varepsilon_t$$

model is obtained. Here the stationary AR operator

$$\phi_p(B) = (1 - \phi_1 B - \phi_2 B^2 - \cdots \phi_p B^p)$$

and the MA operator is

$$\theta_q(B) = (1 - \theta_1 B - \theta_2 B^2 - \cdots \theta_q B^q)$$

.

θ_0 parameter is in a different state for d=0 and d>0. When d=0, the process is stationary. When it is $d \geq 1$, θ_0 is called a deterministic trend. An integrated autoregressive moving average model with non-stationary ARIMA(p, d, q) degrees is formed. When p=0, the ARIMA(p, d, q) model is called the (d, q) integrated moving average model and is defined as ARIMA(0, d, q) or IMA(d, q) (Wei, 2006).

Again, when q=0, the ARIMA(p, d, q) model is called (p, d)'s integrated autoregressive model and is defined as ARIMA(p, d, 0) or ARI(p, d). In other words, the ARIMA(p, d, q) model

$$\left(1-\phi_1 B-\phi_2 B^2-\cdots\phi_p B^p\right)(1-B)^d X_t=\theta_q(B)$$
$$=(1-\theta_1 B-\theta_2 B^2-\cdots\theta_q B^q)\varepsilon_t$$

is as (Kadılar, 2009). Here $(1-B)^d$ term is the d-order difference operation. Let's consider the ARIMA (p, d, q) process with $W_t=X_t-X_{t-1}$ more clearly.

$X_t-X_{t-1}=\phi_1(X_{t-1}-X_{t-2})+\phi_2(X_{t-2}-X_{t-3})+\cdots+\phi_p\left(X_{t-p}-X_{t-p-1}\right)-\theta_1\varepsilon_{t-1}-\theta_2\varepsilon_{t-2}-\cdots-\theta_q\varepsilon_{t-q}+\varepsilon_t$

It can be rewritten as

$X_t=(1+\phi_1)X_{t-1}+(\phi_2-\phi_1)X_{t-2}+(\phi_3-\phi_2)X_{t-3}+\cdots+\left(\phi_p-\phi_{p-1}\right)X_{t-p}-\phi_p X_{t-p-1}-\theta_1\varepsilon_{t-1}-\theta_2\varepsilon_{t-2}-\cdots-\theta_q\varepsilon_{t-q}+\varepsilon_t$

(Cryer 1986).

ARIMA(0, d, q) model

$$\phi_p(B)(1-B)^d X_t=\theta_0+\theta_q(B)\varepsilon_t$$

in the model, when p=0, d=1 and q=1

$$(1-B)^d X_t=(1-\theta B)\varepsilon_t$$

or

$$X_t=X_{t-1}-\theta\varepsilon_{t-1}+\varepsilon_t$$

becomes as. Here -1< $\theta$is 1. This ARIMA(0, 1, 1) model turns into a stationary MA(1) process after taking the first difference. The ARIMA(0, 1, 1) model, that is, the first-order integrated moving average model can be characterized by the ACF (autocorrelation function) graph of the first difference series (Wei, 2006).

Again, the model will be ARIMA(0, 2, 2) in an MA(2) series that is made stationary after taking the second difference. When p=0, d=2, and q=2 in this the general ARIMA(p, d, q) model and the second difference of the series is included in this equation,

$$(1-B)^2 X_t = (1-\theta_1 B - \theta_2 B^2)\varepsilon_t$$

a model such as is formed and the open form of such a model is

$$X_t = 2X_{t-1} - X_{t-2} - \theta_1\varepsilon_{t-1} - \theta_2\varepsilon_{t-2} + \varepsilon_t$$

ARIMA(p, d, 0) model is expressed as

ARIMA(1, 1, 0) process,

$$X_t - X_{t-1} = \phi(X_{t-1} - X_{t-2}) + \varepsilon_t$$

or

$$X_t = (1+\phi)X_{t-1} - \phi X_{t-2} + \varepsilon_t$$

Here $|\phi| < 1$ is (Cryer, 1986). Such series are called first-order integrated autoregressive series.

Similarly, different ARIMA(p, d, q) models can be written for various values of p and q (such as 0,1,2,3) that are taken first or second difference.

Unit Root Tests

Most of the time series are non-stationary series. If a series is not stationary, it should be tested and stationarity should be ensured. Working with stationary series provides reliable results. The autocorrelation function roughly shows whether there is a trend in a time series. A slowly decreasing ACF (Autocorrelation function) is indicative of a large characteristic root, right unit root process, or trend stationary process. Formal tests assist determine whether the system consists of a trend and whether this trend is deterministic or stochastic. However, these tests are not very effective in separating the approximate unit root or unit root process (Enders, 2010). There are different methods for testing stationarity in time series. Here Dickey-Fuller test, one of the most commonly used methods, will be explained.

Dickey-Fuller unit root test

The Dickey-Fuller method is based on the distribution of the least-squares estimator of the parameters under the assumption of a unit root. However, Dickey-Fuller tests are applied if the process has a unit root and can be eliminated using the difference taking method.

When a first-order autoregressive time series model such as $e_t \sim WN(0, \sigma^2)$ equation's

$$X_t = \phi X_{t-1} + \varepsilon_t$$

first order difference is taken

$$X_t - X_{t-1} = \phi X_{t-1} - X_{t-1} + \varepsilon_t$$

or

$$\Delta X_t = (\phi - 1)X_{t-1} + \varepsilon_t$$

it becomes as. When it is taken as $\gamma = \phi - 1$

$$\Delta X_t = \gamma X_{t-1} + \varepsilon_t$$

.

$X_t = \phi X_{t-1} + \varepsilon_t$ time series $H_0 : \phi = 1$ is not stationary under the null hypothesis. In non-stationary time series, t statistic is used to test in $H_0 : \phi = 1$ hypothesis. When the distribution is not a standard t distribution, the limit distribution is for the following three separate models,

$$\Delta X_t = \gamma X_{t-1} + \varepsilon_t$$
$$\Delta X_t = \alpha + \gamma X_{t-1} + \varepsilon_t$$
$$\Delta X_t = \alpha + \beta t + \gamma X_{t-1} + \varepsilon_t$$

is τ , τ_μ τ_τ respectively (Enders, 2010).

$H_0 : \phi = 1$ or when the hypothesis $H_0 : \gamma = 0$ is tested and the variance $n(\hat{\phi}_\tau - 1)$ is known, if the value is less than the critical value, the null hypothesis is rejected. This shows that the series is stationary.

Augmented Dickey-Fuller unit root test

In the extended Dickey-Fuller unit root test, it is aimed that the errors become non-autocorrelated. Therefore, the lagged values of the dependent variable can be added to the model. Here X_t series

$$\Delta X_t = \alpha_1 X_{t-1} + \sum_{j=1}^{p} \beta_j \Delta X_{t-j} + e_t$$

is written as and when the regression X_t on X_{t-1}, $\Delta X_{t-1}, \ldots, \Delta X_{t-p}$ is made, the coefficient of X_{t-1} is estimated and compared to the Dickey-Fuller critical values (Wooldridge, 2002). That is

$$\hat{\tau} = \frac{\hat{\alpha}_1}{S_{\alpha_1}}$$

statistics are used.

Model Determination Criteria

For time series models, the models are determined according to the behaviour of the autocorrelation and partial autocorrelation functions as follows (Table 1) (Sevüktekin and Nargeleçekenler, 2010).

Table 1. Model Determination Criteria

Model	Autocorrelation function	Partial autocorrelation function
AR(p)	The autocorrelation function decreases exponentially.	The partial autocorrelation function is cut off after the pth lag and approaches zero rapidly.
MA(q)	The autocorrelation function is cut off after the qth lag and approaches zero rapidly.	The autocorrelation function decreases exponentially.
ARMA(p,q)	The autocorrelation function decreases exponentially and is cut off after the pth lag.	The partial autocorrelation function decreases exponentially and is cut off after the qth lag.

In addition, there may be more than one model suitable for the series in practice. In this case, some criteria have been developed for the selection of the most suitable model for the series. One of these criteria is the Bayesian information criterion, namely the Schwartz Bayesian criterion (BIC).

$$BIC = nln\hat{\sigma}_{\varepsilon}^{2} + Mln\, n$$

It is calculated with the formula of (Cooray, 2008).

In some sources

$$BIC = ln\hat{\sigma}_{\varepsilon}^{2} + M\frac{ln\, n}{n}$$

is given as (Shumway and Stoffer, 2006).

APPLICATION

Time series analysis of lentil production amount in Turkey between the years 1960-2020 in Turkey has been examined. Data is taken from the Turkish Statistical Institute (TUIK) www.tuik.gov.tr in the Plant Production Statistics section of the website https://biruni.tuik.gov.tr /medas/?kn=92&locale=tr . Time series analysis of the data was carried out with SPSS 25 package program. First, the graph of the data was drawn (Figure 5).

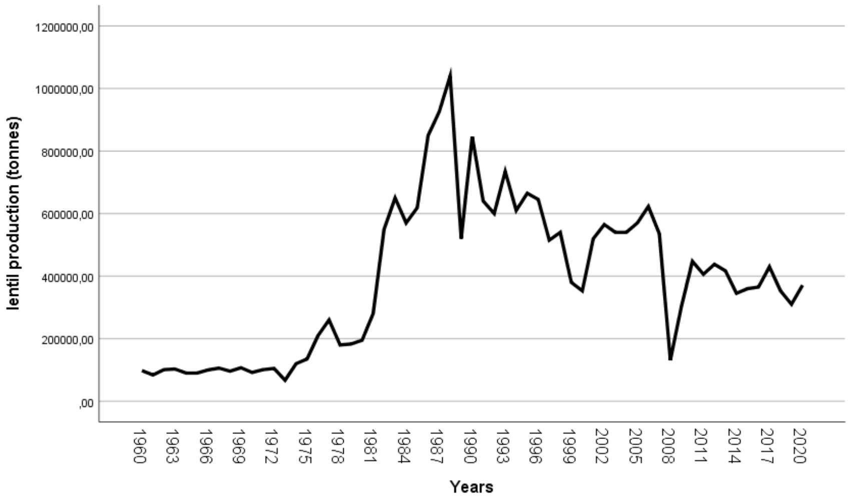


Figure 5. Time series graph

When Figure 5 is examined, it is seen that there is a trend in the series. In other words, it is seen that the series is not stationary. In order to see this more clearly, the autocorrelation (ACF) and partial autocorrelation (PACF) graph of the series should be examined. ACF and PACF graphs of the series belonging to lentil production are presented in Figure 6 and Figure 7, respectively.

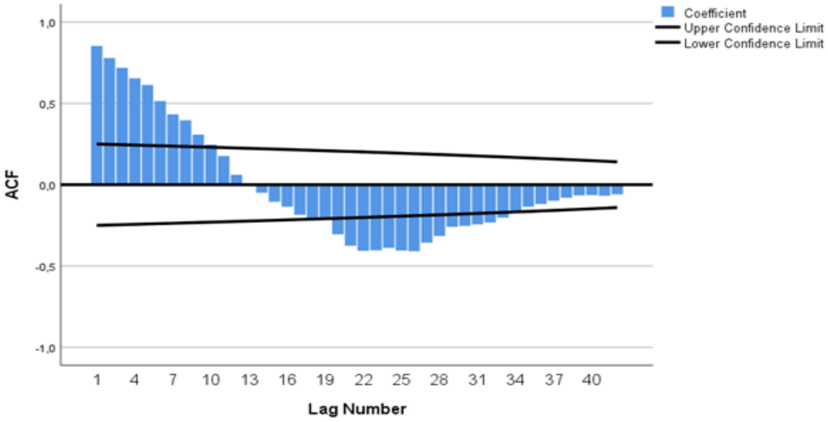


Figure 6. ACF graph

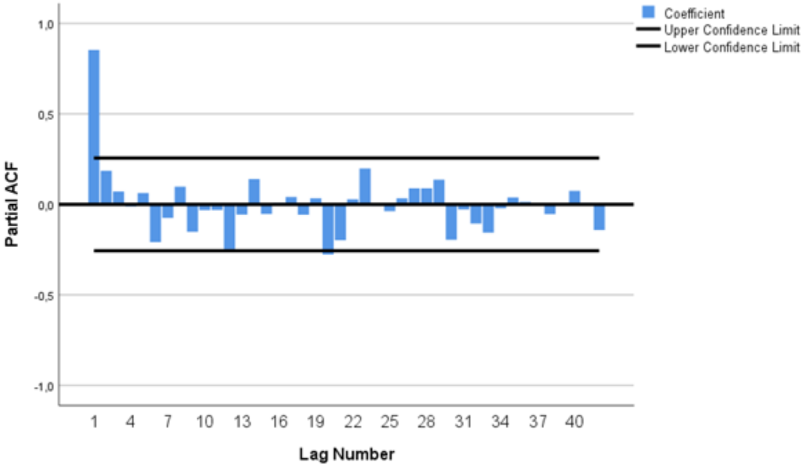


Figure 7. PACF graph

When Figure 6 was examined, it was seen that many relationships in the series exceeded the confidence limits. This shows that there is a trend in the series, in other words, that the series is not stationary. To make the series stationary, the first difference of the series will be taken.

The ACF graph of the first differenced series is presented in Figure 8, and the PACF graph is presented in Figure 9.

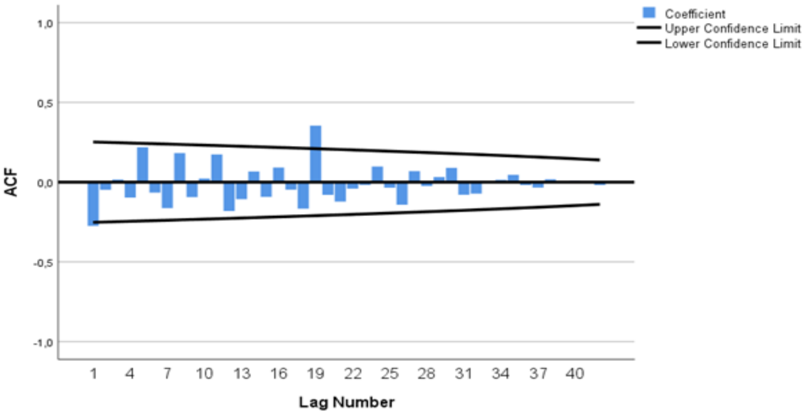


Figure 8. ACF graph of the first difference series

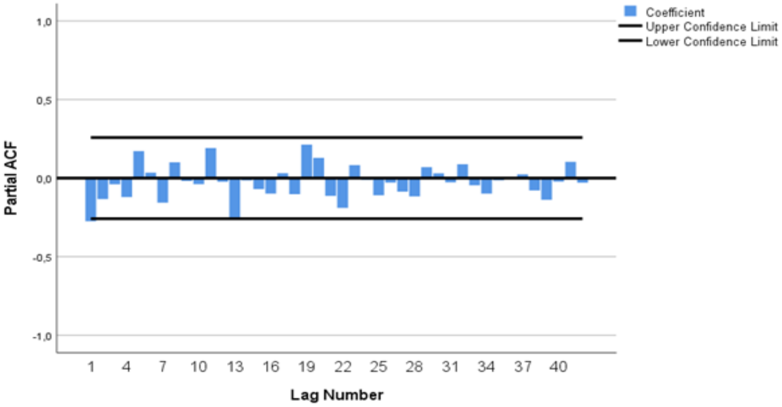


Figure 9. PACF graph of the first difference series

According to the ACF graph of the first difference of the series shown in Figure 8, it is seen that the values of the series after the first lag approach zero rapidly. All values are within confidence limits except for only the 19th lag value. Since this situation does not affect stationarity significantly, it can be said that the series is stationary. When the PACF graph of the first difference series is considered, all lag values after the first lag were within the confidence limits. When the ACF and PACF graphs are examined together, the lags after the first lag in the series in the ACF graph approached zero rapidly. In the PACF graph, the delays after the first lag also rapidly approached zero. When it is considered again to determine the model fit for the series and the order of the model, the near-zero after the first lag in the ACF graph is closer to the zero after the first lag on the PACF graph. Considering the explanations in Table 1, it is understood that the series has a moving average series. Since it gets very close to zero after the first lag, the model degree of the series is q=1 and p=0. Since the first difference of the series is taken, d=1. In this case, the series is written as ARIMA(0,1,1). This series is named as "***first order integrated moving average series”.*** Whether the series is stationary or not can also be determined by the Augmented Dickey-Fuller (ADF) unit root test. ADF unit root test was applied with EViews 9.0 package program. The unit root test was applied according to the situation in which it is the trend and intercept. The results of the original version of the series, namely the level ADF unit root test, are given in Table 2.

Tablo 2. Unit root results at the level of the series

Null Hypothesis: MERC has a unit root				
Exogenous: Constant, Linear Trend				
Lag Length: 0 (Automatic - based on SIC, maxlag=10)				
			t-Statistic	Prob.*
Augmented Dickey-Fuller test statistic			-2.232808	0.4631
Test critical values:	1% level		-4.118444	
	5% level		-3.486509	
	10% level		-3.171541	
*MacKinnon (1996) one-sided p-values.				
Augmented Dickey-Fuller Test Equation				
Dependent Variable: D(MERC)				
Method: Least Squares				
Date: 05/14/21 Time: 13:31				
Sample (adjusted): 1961 2020				
Included observations: 60 after adjustments				
Variable	Coefficient	Std. Error	t-Statistic	Prob.
MERC(-1)	-0.167334	0.074943	-2.232808	0.0295
C	50969.17	35395.60	1.439986	0.1553
@TREND("1960")	613.2982	1061.409	0.577815	0.5657
R-squared	0.085097	Mean dependent var		4546.917
Adjusted R-squared	0.052995	S.D. dependent var		128373.8
S.E. of regression	124925.9	Akaike info criterion		26.35754
Sum squared resid	8.90E+11	Schwarz criterion		26.46225
Log likelihood	-787.7261	Hannan-Quinn criter.		26.39850
F-statistic	2.650851	Durbin-Watson stat		2.346793
Prob(F-statistic)	0.079284			

MacKinnon (1996) are one-way p-values

According to the results given in Table 2, the t-test statistic of the ADF unit root test is -2.233. Critical values at 1%, 5% and 10% significance levels were found as -4.118, -3.487 and -3.172, respectively. Since the test statistic is smaller than the critical values in absolute value, H_0 the hypothesis is accepted and the series is not stationary (p=0.4631>0.05).

After taking the first difference of the series, the ADF unit root test was applied again and the results are given in Table 3.

Table 3. ADF unit root test results of the first difference of the series

Null Hypothesis: D(MERC) has a unit root				
Exogenous: Constant, Linear Trend				
Lag Length: 0 (Automatic - based on SIC, maxlag=10)				
			t-Statistic	Prob.*
Augmented Dickey-Fuller test statistic			-9.994096	0.0000
Test critical values:	1% level		-4.121303	
	5% level		-3.487845	
	10% level		-3.172314	
*MacKinnon (1996) one-sided p-values.				
Augmented Dickey-Fuller Test Equation				
Dependent Variable: D(MERC,2)				
Method: Least Squares				
Date: 05/14/21 Time: 13:28				
Sample (adjusted): 1962 2020				
Included observations: 59 after adjustments				
Variable	Coefficient	Std. Error	t-Statistic	Prob.
D(MERC(-1))	-1.283721	0.128448	-9.994096	0.0000
C	29676.54	34199.59	0.867746	0.3892
@TREND("1960")	-767.6064	966.6832	-0.794062	0.4305
R-squared	0.640765	Mean dependent var		1268.051
Adjusted R-squared	0.627935	S.D. dependent var		206526.8
S.E. of regression	125975.5	Akaike info criterion		26.37507
Sum squared resid	8.89E+11	Schwarz criterion		26.48071
Log likelihood	-775.0646	Hannan-Quinn criter.		26.41631
F-statistic	49.94334	Durbin-Watson stat		2.078919
Prob(F-statistic)	0.000000			

*MacKinnon (1996) are one-way p-values

According to the results given in Table 3, the t-test statistic of the ADF unit root test is -9,994. Critical values at 1%, 5% and 10% significance levels were found -4.121, -3.488 and -3.172, respectively. Since the test statistic is bigger than the critical values in absolute value, H_0 hypothesis is rejected. In addition, p=0.0000<0.05, the series became stationary.

According to the first difference series, the results of the time series analysis will be evaluated. The ACF and PACF values of the first difference of the series were calculated with the SPSS package program and results were given in Table 4 and Table 5, respectively.

Table 4. Autocorrelation (ACF) function values

Autocorrelations					
Series: lentil					
Lag	Autocorrelation	Std. Error[a]	Box-Ljung Statistic		
			Value	df	Sig.[b]
1	-0.274	0.126	4.735	1	0.03
2	-0.048	0.125	4.884	2	0.087
3	0.016	0.124	4.901	3	0.179
4	-0.097	0.123	5.524	4	0.238
5	0.219	0.122	8.767	5	0.119
6	-0.066	0.12	9.064	6	0.17
7	-0.163	0.119	10.92	7	0.142
8	0.182	0.118	13.29	8	0.102
9	-0.094	0.117	13.935	9	0.125
10	0.023	0.116	13.974	10	0.174
11	0.174	0.115	16.264	11	0.132

12	-0.181	0.114	18.807	12	0.093
13	-0.106	0.112	19.701	13	0.103
14	0.067	0.111	20.066	14	0.128
15	-0.092	0.11	20.77	15	0.144
16	0.091	0.109	21.477	16	0.161
17	-0.047	0.108	21.672	17	0.198
18	-0.166	0.106	24.12	18	0.151
19	0.354	0.105	35.502	19	0.012
20	-0.08	0.104	36.09	20	0.015
21	-0.122	0.102	37.509	21	0.015
22	-0.04	0.101	37.669	22	0.02
23	-0.017	0.1	37.699	23	0.027
24	0.098	0.098	38.698	24	0.029
25	-0.035	0.097	38.826	25	0.038

a. The underlying process assumed is independence (white noise).

b. Based on the asymptotic chi-square approximation.

Table 5. Partial autocorrelation (PACF) function values

Partial Autocorrelations

Series: lentil

Lag	Partial Autocorrelation	Std. Error
1	-0.274	0.129
2	-0.133	0.129
3	-0.039	0.129
4	-0.121	0.129
5	0.172	0.129
6	0.035	0.129
7	-0.157	0.129
8	0.101	0.129
9	-0.018	0.129

10	-0.038	0.129
11	0.192	0.129
12	-0.023	0.129
13	-0.252	0.129
14	-0.014	0.129
15	-0.07	0.129
16	-0.099	0.129
17	0.031	0.129
18	-0.102	0.129
19	0.213	0.129
20	0.129	0.129
21	-0.113	0.129
22	-0.189	0.129
23	0.083	0.129
24	-0.01	0.129
25	-0.109	0.129

The parameter estimates of the time series model obtained as ARIMA(0,1,1) were given in Table 6.

Table 6. ARIMA model parameters

		Estimate	SE	t	Sig.
Constant		4443.452	10765.802	0.413	0.681
Difference		1			
MA	Lag 1	0.330	0.124	2.652	0.010

In Table 6, the parameter estimation coefficient was found to be 0.330. The t test value of this coefficient was found to be t=2.652 and since it is statistically significant since the significance value is p=0.010<0.05. Hence, the model is considered fit. The model fit criteria calculated to see the fit of the model and the results of Ljung-Box statistics were given in Table 7.

Table 7. Model fit statistics

Model Fit statistics					Ljung-Box Q(18)		
R^2	RMSE	MAPE	MAE	BIC	Statistics	DF	Sig,
0.750	123353.299	22.24	75063.873	23.582	13.464	17	0.705

DF: Degree of Freedom

The coefficient of determination (R^2), square root of the mean square error (RMSE), mean absolute percentage error (MAPE), mean absolute error (MAE), and bayesian information criterion (BIC) values are given in Table 7. The efficiency of the model was determined according to the Ljung-Box Q test, the Ljung-Box Q statistic was found to be 13.464, and since p=0.705>0.05, it can be said that the model is fit.

Since the coefficient of the appropriate ARIMA(0,1,1) model is q=0.330, the following Back-Shift (regression operator) is used for the open software of the model.

$$(1-B)X_t = (1-\theta B)e_t$$

ARIMA(0,1,1) model which is

$$X_t - X_{t-1} = -\theta e_{t-1} + e_t$$

written as and

$$X_t = X_{t-1} - \theta e_{t-1} + e_t$$

expressed regularly as. Accordingly, the model in this application is

$$X_t = X_{t-1} - 0.330 e_{t-1} + e_t$$

written as. According to this model obtained, predictions can be made for future years. Before making a prediction, the graph of the error terms ACF and PACF is determined whether the series is white noise or not. From the ACF and PACF graphs of the error terms given in Figure 10, it was seen that all the lag values of the error terms of the series did not exceed the confidence limits and remained within the confidence limits. Therefore, the series is white noise. Since it is a white noise series, a healthier prediction can be made.

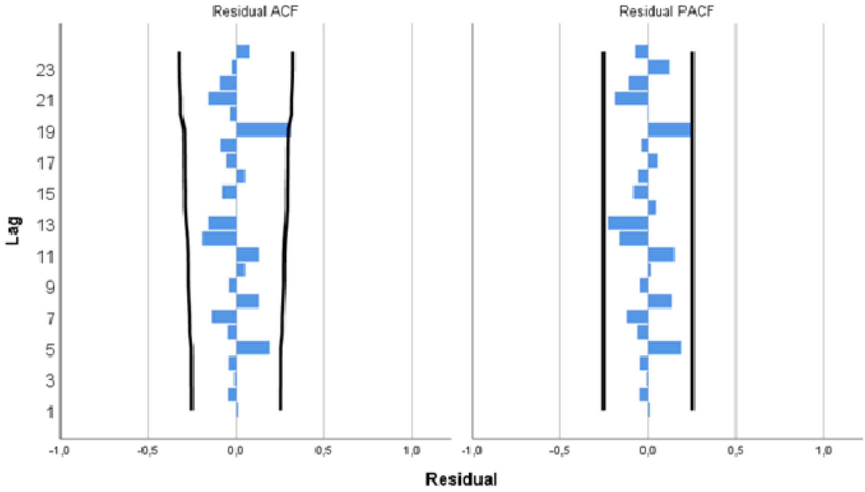


Figure 10. ACF and PACF graphs of the error terms in the first difference series

The prediction was made after obtaining the series with white noise. The prediction results covering the 2021-2025 period are given in Table 8.

Table 8. Lentil production prediction (2021-2025)

Years	Forecasting
2021	364053
2022	368497
2023	372940
2024	377384
2025	381827

As seen in Table 8, lentil production is expected to increase between 2021-2025. The graph, which includes the observed and predicted values as well as the predicted values, is shown in Figure 11.

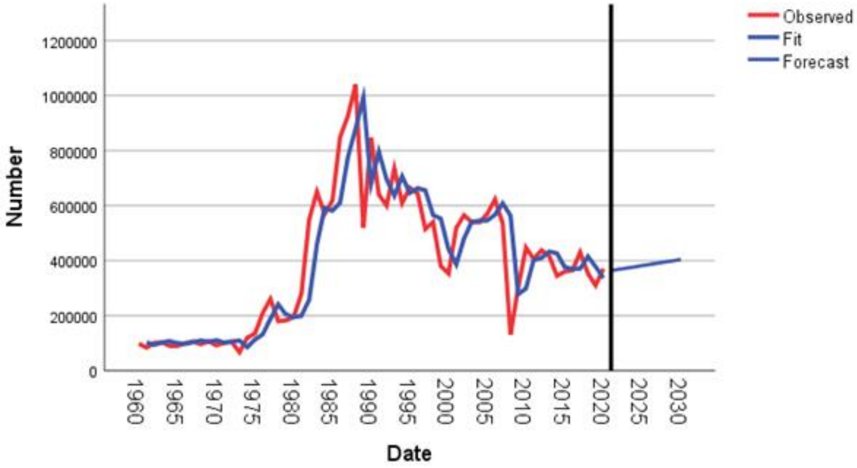


Figure 11. Estimated and predicted values of lentil production

CONCLUSION

In this study, information about time series is given and theoretical information is given in general terms. An application has been made on plant production in the field of agriculture. Time series analysis of lentil production in Turkey between 1960 and 2020 was made and a model such as ARIMA(0,1,1) was obtained. Since the series is not stationary, its first difference is taken and it becomes stationary after the first difference. The resulting model is

$$X_t = X_{t-1} - 0.330e_{t-1} + e_t$$

and named as “first order integrated moving average series." According to this model, when the prediction is made, it is expected that the lentil production amount will be between 364053-381827 tons. Lentil production in Turkey in 2020 is 370815 tons. It is estimated that the amount of lentil production in 2025 will increase by 2.97% compared to the production in 2020. This increase is very significant for Turkey. Increasing the amount of plant production in our country where the population is constantly growing is extremely important in terms of meeting the needs. However, plans and projects should be developed to ensure that this increase gets bigger. Time series analysis can be widely used in the development of plant production models in the field of agriculture.

REFERENCES

Akdi, Y. (2010). Zaman Serileri Analizi (Birim Kökler ve Kointegrasyon). Gazi Kitabevi.

Box, G. E. P. and Jenkins, G. M. (1976). Time Series Analysis Forecasting and Control Revised edition. Holden Day, pp. 25-36, San Francisco.

Chatfield, C. 2000. Time Series Forecasting. Chapman and Hall/CRC.

Cooray, T. M. J. A. (2008). Applied Time Series. Analysis and Forecasting. Narosa Publishing House Pvt. Ltd.

Cryer, J. D. (1986). Time Series Analysis.

Çelik, Ş. (2013). Zaman Serileri Analizi ve Trafik Kazası Verilerine Uygulanması. Ankara Üniversitesi Fen Bilimleri Enstitüsü Doktora Tezi, Ankara.

Enders, W. (2010). Applied Econometric Time Series Analysis. John Wiley and Sons, Inc, USA.

Günay, S., Eğrioğlu, E. ve Aladağ, Ç., H., 2007. Tek Değişkenli Zaman Serileri Analizine Giriş.

Hamilton, J. D. (1994). Time Series Analysis. Princeton University Press Princeton, New Jersey.

Kadılar, C. (2009). SPSS Uygulamalı Zaman Serileri Analizine Giriş.

MacKinnon, J. G. (1996). Numerical distribution functions for unit root and cointegration tests. Journal Applied Econometrics, 11: 601-618.

Montgomery, D. C., Johnson, L. A. and Gardiner, J. S. (1990). Forecasting and Time Series Analysis.

Sevüktekin, M., Nargeleçekenler, M. (2010). Ekonometrik Zaman Serileri Analizi Eviews Uygulamalı. Nobel Yayın Dağıtım Tic. Ltd. Şti., Ankara. ISBN: 978-975-591-755-9

Shumway, R. H., Stoffer, D. S. (2006). Time Series Analysis and its Applications with R Examples. Springer, New York.

Wei, W. W. S. (2006). Time Series Analysis, Addison Wesley Publishing Company.

Wooldridge, J. M. (2002). Introductory to Econometrics. A Modern Approach.

CHAPTER III

COMPARISON OF PENALIZED REGRESSION TECHNIQUES FOR LIVE WEIGHT IN KARAKAŞ SHEEP

Assoc. Prof. Dr. Suna AKKOL[*]

Assist. Prof. Dr. Aslı AKILLI[†]

[*] Van Yuzuncu Yil University, Faculty of Agriculture, Department of Animal Science, Van, Turkey. Email: sgakkol@yyu.edu.tr

[†] Ahi Evran University, Faculty of Agriculture, Department of Agricultural Economics, Kırşehir, Turkey. Email: asliakilli@ahievran.edu.tr

INTRODUCTION

Estimation of live weight in farm animals using body measurements is one of the important research topics for animal scientists and producers. The correct interpretation of the maturation process and related characters of animals affects the economic profitability of the producers and provides convenience in decision-making processes. At the same time, regular evaluation of body measurements and live weight information provides very useful information to producers on important issues such as monitoring animal health issues, adjusting drug dosage, adjusting ration performance and feed amount, breeding selection and weeding (Sabbioni et al., 2020). Regression analysis comes first among the methods used to evaluate the live weight and growth performance of animals and to obtain predictive values.

Regression analysis is one of the common statistical analysis methods used to describe and interpret the relationship between dependent and independent variables. Various model structures in linear and non-linear forms are used in the estimation of the predicted values of the dependent variable. Correlation analysis is another statistical analysis method used to numerically express and interpret the degree of relationship between variables. In some cases where the relationship between the variables is investigated, there may be strong relationships between the independent variables and the correlation values can be obtained at very high values. In such cases, researchers are faced with the multicollinearity problem and statistical assumptions required for regression analysis cannot be provided (Weisberg, 2005; Alpar, 2003;

Yakubu, 2009; Yakubu, 2010). The source of the multicollinearity problem may be due to small sample size, environmental effects, or data structure.

In the presence of multicollinearity problems, biased results can be obtained in hypothesis control and parameter estimation in regression models, and even results that lead to incorrectly obtaining the signs of the regression coefficients (Weisberg, 2005; Alpar, 2003; Yakubu, 2009; Yakubu, 2010).

In the modeling processes related to body weight estimation, multicollinearity problem is encountered in the data set (Yakubu, 2009; Yakubu, 2010; Dorman et al., 2013; Eyduran et al., 2013; Jahan et al., 2013; Akkol, 2018; Onk et al., 2018; Iqbal et al., 2019; Çankaya et al., 2019). Observing the data structure and eliminating unnecessary variables by using various statistical tools are among the basic methods of solving the multicollinearity problem. PCA and factor analysis, which allow resizing the number of variables, are common methods used by researchers to solve the multicollinearity problem (Tariq et al., 2012; Eyduran et al., 2013; Jahan et al., 2013, Çankaya et al., 2019).

Another widely used method for the multicollinearity problem is Ridge regression. Ridge regression is one of the penalized regression methods. It is shrinkage the regression coefficients by using L2-norm. This method makes predictions by overcoming multicollinearity (Hoerl & Kennard, 1970; Marquardt & Snee 1975; Hoerl & Kennard, 1988; Dorman et al., 2013). In the literature, there are studies on Ridge regression in which live weight estimation is performed in farm animals through various instruments in case of multicollinearity (Malau-Aduli et al., 2004; Yakubu, 2009; Yakubu, 2010; Tirink et al., 2020). However, in some cases, Ridge regression is insufficient for studies to reduce model complexity. Another feature of the method is that it makes predictions for all of the independent variables in the model, but cannot select the variable (Zou & Hastie, 2005). Tibshirani (1996) proposed the method that executed both automatic variable selection and continuous shrinkage simultaneously (Zou &d Hastie, 2005; Wang et al., 2011). This method was called the Least Absolute Shrinkage and Selection Operator (LASSO). Although this method is popular, it has been reported in the literature that it has some limitations (Fan & Li, 2001; Zou and Hastie, 2005; Wang et al., 2011). Zou (2006) introduced the penalized method called adaptive LASSO estimators to solve the limitations. For this end, a data defined weight was added to the original LASSO (Zou, 2006). According to Efron et al. (2004) revealed that LASSO has two restrictions under the conditions the number of variables is too large for the number of observations   the pairwise correlations of a group of variables are high. The Elastic Net (EN) method was proposed by Zou & Hastie (2005) to remove these restrictions emerging in the LASSO method.

LASSO, Adaptive LASSO and Elastic Net methods are the subject of the study. These methods are examined within the scope of penalized/shrinkage regression techniques and are among the methods applied in cases where there is a multicollinearity problem. In recent years, there have been successful studies in the literature on body weight estimation using penalized/shrinkage regression techniques (Akkol, 2018; Iqbal et al., 2019; Zhang et al., 2021). In addition, studies in the literature using regularized regression techniques in case of multicollinearity have also been made in the literature (Das et al., 2018; Idrees & Kamal, 2021).

The aims of this study were to predict the live weight from morphological characteristics of Karakaş seep by using Ridge, LASSO, Adaptive LASSO, Naive Elastic Net, and Elastic Net and to make the variable selection for the purpose of to reduce the model complexity.

MATERIALS AND METHODS

MATERIALS

Animal material of the study which carried out at a Research and Application Farm of Van Yuzuncu Yil University consisted of 29 heads Karakaş lambs. In this study it was used the live weight and the body measurements of lambs recorded at six-month-old. Independent variables were the birth type of lambs (BT), the age of ewe (AE), the live weight at the birth of lambs (LWB), and some body measurements which were withers height (WH), body length (BL), chest width (CW), chest depth (CD), chest girth (CG), haunch girth (HG). In this work the live weight of lambs recorded at six-month-old (LW) was treated as dependent variable.

The statistical analyses were performed by using MEANS, CORR, GLM and GLMSELECT procedure in SAS (SAS, 2014).

METHODS

In this section; multiple linear regression, ordinary least square (OLS), Ridge, LASSO, Adaptive LASSO, and Elastic Net regularized estimation methods are given.

Multiple Linear Regression

The linear model remains one of most important tools for biometricians almost for many decades. In the multiple linear regression model, it is assumed the regression function, $E(X|Y)$ is a function of the input or independent variables. It is given an input vector $X^T = (X_1, X_2, ..., X_p)$ multiple linear regression model is written in Equation 1.

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ...\beta_p x_{ip} + e_i , \quad n = 1,2,...p \qquad 1$$

It is wanted to predict the output or dependent variable through the model in Equation 2.

$$\hat{Y} = \hat{\beta}_0 + \sum_{j}^{p} X_j \hat{\beta}_j \qquad 2$$

The linear regression model having an input vector is also written the form given in Equation 3.

$$f(X) = \hat{\beta}_0 + \sum_{j}^{p} X_j \hat{\beta}_j$$

The equation is written as

$$f(X) = \hat{\beta}_0 + \sum_{j}^{p} X_j \hat{\beta}_j = X^T \beta \qquad 3$$

The multiple linear regression model used to predict the live weight with OLS, LASSO Adaptive LASSO, naïve Elastic Net and Elastic Net model is given in Equation 4.

$$Y = \mu 1_n + X\beta + e \qquad 4$$

Where $Y = (y_1, y_2, ... y_n)^T$ is a vector of observed dependent variable, 1_n is a column vector of n ones $(i = 1,2,3..., n)$ and μ is intercept, X is an *nxp* matrix of explanatory variable, β is the vector of regression coefficients, $\beta = \left(\beta_0, \beta_1, ..., \beta_p\right)^T$, e is the vector of the residual with a mean zero and variance $I\sigma_e^2$. It was assumed that observed dependent variables have been mean-centered. The regression coefficients are, and the design matrix is *X*.

$$X = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & & & & \vdots \\ 1 & x_{i1} & x_{i2} & \cdots & x_{ip} \\ \vdots & & & & \vdots \\ 1 & x_{n1} & x_{n2} & \ldots & x_{np} \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_i \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}$$

Ordinary Least Square

The most familiar and frequently used estimation method to predict unknown parameters is the ordinary least squares (OLS). This method is based on the basis to choose the coefficients ($\hat{\beta}_j$)'s minimizing the residual sum of square (RSS) and shown via the Equation 5.

$$RSS = \sum_{i=1}^{n} (y_i - f(x_i))^2 \qquad 5$$

Mathematical representation of RSS is given in Equation 6.

$$RSS = \sum_{i=1}^{n} (y_i - \beta_0 + \sum_{j=1}^{p} x_{ij} \hat{\beta}_j)^2 \qquad 6$$

As an optimization problem, Equation 5 is written as Equation 7 for OLS.

$$\hat{\beta} = \arg\min_{\beta} \| Y - X\beta \|^2 \qquad 7$$

The mathematical form of Equation 7 is written as in Equation 8.

$$\hat{\beta} = \arg\min_{\beta} RSS(\hat{\beta}) = \arg\min_{\beta} \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 \qquad 8$$

The least square estimator is given in Equation 9.

$$\hat{\beta}_{OLS} = (X^T X)^{-1} X^T Y \qquad 9$$

When $\boldsymbol{X^T X}$ is nonsingular, then it has the unique solution. Hence fitted y-values ($\hat{y}$) are given by Equation 10 (Filzmoser, 2008; Hastie et all, 2009).

$$\hat{y} = X\hat{\beta} = X(X^T X)^{-1} X^T Y \qquad 10$$

The OLS estimator has the property of unbiasedness which desired in an estimator. This does not mean that the variance will always be low. The variance may be huge or approaches infinity due to two problems. First problem is highly correlation between the input variables and called collinearity or multicollinearity. Second problem called as the “large p and small n problem” ($p > n$), the number of input variables in the model are much bigger the number of observation (n) (Acharjee et al, 2013).

Ridge

Ridge regression is one of the most frequently used penalized methods. This method is based on the principle of minimizing the sum of error squares (RSS) in order to obtain the β coefficients (Hoerl & Kennard, 1970, Hoerl & Kennard, 1988). Ridge coefficients can be obtained by using Equation 11.

$$\hat{\beta}_{ridge} = \arg\min_{\beta} RSS(\beta) = \arg\min_{\beta} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \lambda_2 \sum_{j=1}^{p} \beta_j^2 \right\} \qquad 11$$

Here, λ≥0 is called the complexity constant controlling the amount of shrinkage (Marquardt & Snee 1975) and $\ell_2 = \lambda_2 \sum_{j=1}^{k} \beta_j^2$ is the Ridge penalty function (Hastie et al., 2009). The mathematical form of the expression given in Equation 11 is given in Equation 12.

$$\hat{\beta}_{ridge} = \arg\min_{\beta} \left\| Y - X\beta \right\|^2 + \lambda_2 \left\| \beta \right\|^2 \qquad 12$$

where $\left\| Y - X\beta \right\|^2 = \sum_{i=1}^{n} (y_i - x_i^T \beta)^2$ is the loss function, $\left\| \beta \right\|^2 = \sum_{j=1}^{p} \beta_j^2$ is the ℓ_2-norm penalty on β, $\lambda \geq 0$ is a tuning (penalty or shrinkage) parameter which regulates strength of penalty.

LASSO

LASSO regression is a penalized/shrinkage method proposed with the aim of improving the OLS method. This method is a successful procedure that makes prediction and simultaneously variable selection

(Tibshirani, 1996). Equation 13 in the Lagrangian form is used to calculate the regression coefficients with LASSO.

$$\hat{\beta}_{lasso} = \arg\min_{\beta} \|Y - X\beta\|^2 + \lambda_1 \|\beta\|_1 \qquad 13$$

where $\|\beta\|_1 = \sum_{j=1}^{p} |\beta_j|$ is the ℓ_1-norm penalty on β, $\lambda_1 \geq 0$ is a tuning (penalty or shrinkage) parameter which regulates strength of penalty and is important for the LASSO success. For the LASSO estimate Equation 13 is rewritten without an intercept (Hastie et al., 2007; Hastie et al., 2009)

$$\hat{\beta}_{lasso} = \arg\min_{\beta} \left\{ \frac{1}{2} \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \lambda_1 \sum_{j=1}^{p} |\beta_j| \right\} \qquad 14$$

The penalty function called ℓ_1 is important for the success of LASSO.

Adaptive LASSO

The adaptive LASSO proposed by Zou (2006) remedies the lack of the oracle properties of the original LASSO. The adaptive LASSO modifies the original LASSO penalty by adding weights for each parameter to the penalty term. These weights are data-defined weights, $\hat{\omega}_j$, and they control the shrinking of the zero coefficients more than the non-zero coefficients. The adaptive LASSO estimates $\hat{\beta}_{alasso}$ are given in Equation 15.

$$\hat{\beta}_{alasso} = \arg\min_{\beta} \left\| y_i - \sum_{j=1}^{p} x_{ij}\beta_j \right\|^2 + \lambda \sum_{j=1}^{p} \hat{\omega}_j \left|\beta_j\right| \qquad 15$$

Where a known weights vector, $\hat{\omega}_j = 1 \big/ \left|\hat{\beta}_{\mathrm{j}}^{\mathrm{ini}}\right|^{\gamma}$, γ is a positive constant and $\hat{\beta}_{\mathrm{j}}^{\mathrm{ini}}$ is initial consistent estimator of β obtained from ordinary least square or Ridge regression if there is a multicollinearity problem (Zou, 2006; Ogutu et al., 2012; Gunes, 2015). If the parameter estimates produced by adaptive LASSO are defined by $\hat{\beta}(\lambda_n)$. $\hat{\beta}(\lambda_n)$ is given in Equation 16.

$$\hat{\beta}(\lambda_n) = \arg\min \left\{ \left\| Y - \sum_{j=1}^{p} X_j \beta_j \right\|^2 + \sum_{j=1}^{p} \lambda_n \hat{\omega}_{jn} \left|\beta_j\right| \right\} \qquad 16$$

Fan and Li (2001), and Zou (2006) proved that the adaptive LASSO has oracle property when $\lambda_n \to \infty$ and $\lambda_n \big/ \sqrt{n} \to 0$.

Elastic Net

Both naïve elastic net and elastic net methods use a mixture of the LASSO (ℓ_1) and Ridge (ℓ_2) penalties (Friedman et al., 2010). The elastic net can be formulated by Equation 17 (Hastie et al., 2009).

$$\hat{\beta}_{naive} = \arg\min_{\beta} \left\| Y - X\beta \right\|^2 + \lambda_2 \left\|\beta\right\|^2 + \lambda_1 \left\|\beta\right\|_1 \qquad 17$$

$$\hat{\beta}_{naive} = \arg\min_{\beta}\left(\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{k} x_{ij}\beta_j\right)^2 + \lambda_2\sum_{j=1}^{k}\beta_j^2 + \lambda_1\sum_{j=1}^{k}\left|\beta_j\right|\right) \qquad 18$$

In the case of an orthogonal design naïve elastic net solution is, (Zou & Hastie, 2005)

$$\hat{\beta}_{j(naive)} = \frac{\left(\left|\hat{\beta}_{j(OLS)}\right| - \lambda_1/2\right)}{1+\lambda_2} sign\{\hat{\beta}_{j(OLS)}\}$$

$$\hat{\beta}_{naive} = \{1/\sqrt{(1+\lambda_2)}\}\hat{\beta}^* \qquad 19$$

Where,

$$\hat{\beta}^* = \arg\min\left|y^* - X^*\beta^*\right|^2 + \frac{\lambda_1}{\sqrt{(1+\lambda_2)}}\left|\beta^*\right|_1$$

Naïve elastic net has double shrinkage procedure. Firstly, it shrinkage to detect the Ridge coefficient then for lasso coefficient. This causes both extra bias and no reduction in variance (Zou & Hastie, 2005).

The elastic net, which is corrected estimates are,

$$\hat{\beta}_{elasticnet} = \sqrt{(1+\lambda_2)}\hat{\beta}^*$$

Elastic net is written as,

$$\hat{\beta}_{elasticnet} = (1+\lambda_2)\hat{\beta}_{naive} \tag{20}$$

Model Selection: The adjusted coefficient of determination (R^2_{adj}), Akaiki Information Criteria (AIC), corrected AIC (AICc) and Average Square Error (ASE) are cohesion criteria used to compare the Ridge, the LASSO, the adaptive LASSO, the naïve elastic net and the elastic net results. They are called the model-fit criteria (Maydeu-Olivares & Garci´a-Forero, 2010).

$$R^2_{adj} = 1-(1-R^2)\frac{n-1}{n-p-1} \tag{21}$$

Adjusted R^2 is given in Equation 21. Where, *p* is the total number of explanatory variables in the model not including the constant, and *n* shows the sample size. AIC (Akaike, 1974) and AICc (Hurvich and Tsai, 1989) are given in Equation 22.

$$\begin{aligned} AIC &= -2ll + 2p \\ AICc &= -2ll + 2kn/(n-p-1) \end{aligned} \tag{22}$$

Where *ll* shows the log-likelihood. The average square error (ASE) is another cohesion criteria (Equation 23).

$$ASE = \frac{\sum_{i=1}^{n}\left(Y_{new} - \left(\hat{\beta}_0 + \sum_{i=1}^{p-1}\hat{\beta}_j X_{new,j}\right)\right)^2}{n} \tag{23}$$

Where (Y_{new} and X_{new}) express new data. Performing the best model selection, the model being minimum AIC, AICc and ASE values is chosen.

RESULTS

The descriptive statistics of the 6th month live weight of Karakaş lambs used in the study and the biometric measurements (BT, AE, LWB, WH, BL, CW, CD, CG, and HG) of this period are given in Table 1. In the data set, the rate of singleton lambs was 51.72% and the rate of twins was 48.28%; The proportions of ewes aged 2, 3, 4 and 5 were found to be 14.14%, 20.69%, 48.28% and 6.90%, respectively. In Table 1 it was seen that LW, LWB, WH, BL, CW, CD, CG, and HG have the average 30.24 kg, 4.20 kg, 58.69 cm, 56.28 cm, 16.94 cm, 24.70 cm, 74.60 cm, 64.50 cm, respectively. The 6th month body weights of the lambs were consistent with the results of Gökdal et al.'s study (Gökdal et al., 2004).

Table 1. Descriptive statistics

Variable	Mean	SD	SE	CV	Min	Max
Live Weight (LW)	30.24	4.94	0.95	16.34	20.69	41.95
Birth Type (BT)	1.48	0.51	0.09	34.29	1	2
Age of Ewe (AE)	3.38	0.94	0.18	27.86	2	5
Live Weight at the Birth (LWB)	4.20	0.69	0.13	16.40	3.00	5.70
Withers Height (WH)	58.69	2.90	0.56	4.94	52.00	64.39
Body Length (BL)	56.28	3.51	0.68	6.25	48.14	64.56
Chest Width (CW)	16.94	1.42	0.27	8.36	13.39	20.02
Chest Depth (CD)	24.70	1.74	0.34	7.06	20.64	28.18
Chest Girth (CG)	74.60	5.11	0.98	6.85	63.84	86.66
Haunch Girth (HG)	64.50	4.19	0.81	6.50	56.84	74.13

SD: Standard Deviation; SE: Standard Error; CV: Coefficient of Variation; Min: Minimum; Max: Maximum

Table 2 shows the correlations between body weight and body measurements at 6 months. Accordingly, the phenotypic correlation between body weight and LWB has a weak and significant relationship (p<0.05). Phenotypic correlations between body weight and other body measurements (WH, BL; CW, CD, CG, and HG) were strong and significant (p<0.01).

Table 2. Correlation coefficients between the independent variables

	LW	BT	AE	LWB	WH	BL	CW	CD	CG
BT	-0.123	1							
AE	-0.087	0.424*	1						
LWB	0.383*	-0.232	-0.345*	1					
WH	0.714**	0.056	-0.069	0.464*	1				
BL	0.822**	-0.108	-0.135	0.305	0.672**	1			
CW	0.772**	-0.242	0.002	0.460*	0.597**	0.637**	1		
CD	0.767**	-0.068	-0.141	0.468*	**0.811****	0.685**	0.622**	1	
CG	0.861**	-0.211	-0.068	0.416*	0.716**	0.732**	0.733**	**0.886****	1
HG	0.821**	-0.092	-0.194	0.435*	0.610**	0.729**	0.635**	0.781**	**0.879****

LW: Live Weight in the third month, AE: Age of Ewe, LWB: Live Weight at the Birth, WH: Withers Height, BL: Body Length, CW: Chest Width, CD: Chest Depth, CG: Chest Girth, HG: Haunch Girth. *: p<0.05, **: p<0.01

Table 2 shows weak but significant correlations between body measurements LWB and AE, BT, WH, BL, CW, CD, CG, and HG (p<0.05). The phenotypic correlation value between WH, BL, CW, CD, CG, and HG was moderate and strong, and these were also significant (p<0.01) (Table 2). The results obtained from this study for the phenotypic correlations between body weight and body measurements were consistent with the results of various previous studies in the literature (Şeker & Kul 2000; Afolayan et al., 2006; Çankaya & Kayaalp, 2007; Yılmaz et al., 2013; Onk et al., 2018).

The F value in the model was obtained as 11.30 and was found to be statistically significant (p<0.01). Despite these results, it can be seen in Table 3 that none of the variables in the model (except for the intercept) are significant. In addition to the aforementioned results, the data set had to be evaluated in terms of multicollinearity due to the strong correlations between the independent variables (body measurements) (Table 2 and Table 3).

TV (Tolerans Value) and VIF (Variance Inflation Factor) values from multicollinearity diagnostics are given in Table 3, and eigenvalues and CI (Condition Index) values are given in Table 4. In Table 3, it can be seen that the TV values ranged from 11.30 to 0.0709 and the VIF values ranged between 1.7834- 14.1139. According to the criteria reported in the literature, if TV<0.1 and VIF>10, the relevant dataset has a multicollinearity problem. In the studied dataset, TV 0.0709 and VIF value of 14.13 for CG reveal that the dataset has a multicollinearity problem. This result was consistent with previous studies in the literature (Montgomery et al., 2001; Yakubu, 2010; Dormann et al., 2013). As can be seen in Table 4, the eigenvalues range from 9.805 to 0.00021254, while the CI ranges from 1 to 214.78. According to the reported information by Marquardt and Snee (1975), Belsley (1991) and Albayrak (2005), if the eigenvalues were very close to zero and the

CIs were greater than 30 then there was an multicollinearity problem in the dataset. Therefore, it is explained by the previous literature study that the data set has multicollinearity problem in terms of these diagnostics (Marquardt & Snee, 1975; Belsley, 1991; Albayrak, 2005).

Table 3. The estimation of coefficients obtained by using the OLS in the multiple linear regression

Independent Variable	Parameter Estimate	Standard Error	t	p	TV	VIF
Intercept	-48.54768	10.58202	-4.59	0.0003		
LWB	-0.35561	0.91440	-0.39	0.7022	0.4787	2.0890
BT	0.56022	1.29445	0.43	0.6706	0.5140	1.9455
AE	-0.21904	0.65436	-0.33	0.7419	0.5607	1.7834
WH	0.19549	0.30944	0.63	0.5360	0.2551	3.9209
BL	0.40438	0.22023	1.84	0.0839	0.3433	2.9133
CW	0.87823	0.53411	1.64	0.1185	0.3591	2.7845
CD	-0.34300	0.71886	-0.48	0.6393	0.1310	7.6312
CG	0.36569	0.33346	1.10	0.2881	0.0709	14.1139
HG	0.19062	0.28076	0.68	0.5063	0.1485	6.7364

Goodness of Fit Criteria	
MSE	4.9796
RMSE	2.2315
R^2	0.8568
R^2_{adj}	0.7810

LW: Live Weight in the third month, LWB: Live Weight at the Birth, WH: Withers Height, BL: Body Length, CW: Chest Width, CD: Chest Depth, CG: Chest Girth, HG: Haunch Girth

Table 4. Eigenvalues and CI values

	1	2	3	4	5	6	7	8	9	10
E	9.81	1.25x10^{-1}	4.96x10^{-2}	1.27x10^{-2}	3.06x10^{-3}	2.17x10^{-3}	1.26x10^{-3}	1.15x10^{-3}	3.72x10^{-4}	21.25x10^{-5}
CI	1	8.86	14.06	27.78	56.64	67.23	88.29	92.39	162.31	214.78

E: Eigenvalue, CI: Conditional Index

It has been revealed in previous studies that the multicollinearity implies that standard errors of regression coefficients are higher than expected and thus, it is difficult to find out the accuracy and robustness of the prediction models (Alpar, 2003; Weisberg, 2005; Yakubu, 2009, 2010). In order to overcome the multicollinearity problem, the data were analyzed using Ridge regression method. Numerical results which are regarding coefficient and standard errors are given in Table 5. It was determined that the estimation equation obtained using the Ridge regression method was statistically significant ($p<0.01$).

Table 5. In the linear regression model where LW is the dependent variable, the coefficients and standard errors for Ridge ($\lambda=0.2$), LASSO, Adaptive LASSO, Naïve Elastic Net and Elastic Net ($\lambda=0.05$)

Variables	Ridge	LASSO	Adaptive LASSO	Naïve Elastic Net	Elastic Net
LWB	-0.2047±-0.0274**	-	-	-	-
BT	0.0833±0.0085**	-	-	-	-
AE	-0.0085±-0.0017**	-	-	-	-
WH	0.1778±0.1044^{ns}	0.0735±0.0352	-	-	0.0998±0.0347
BL	0.3482±0.2477^{ns}	0.3991±0.0944	0.4231±0.1056	0.3860±0.0592	0.3704±0.1042
CW	0.7700±0.2208*	0.6711±0.3302	0.6771±0.4965	-	0.6622±0.1779
CD	0.1323±0.0467*	-	-	-	-
CG	0.1961±0.2027^{ns}	0.2651±0.0966	0.4559±0.0587	0.5328±0.0927	0.2376±0.0957

HG	0.2306±0.1954^{ns}	0.1704±0.0370		0.0214±0.0148	0.1908±0.1654
Model Sig	**	**	**	**	**

In the Ridge regression method, estimates were made for all coefficients. The coefficients except HG, and standard errors of all independent variables in the prediction equation for LW were smaller than those in OLS prediction (Table 3). Especially, the sign of the coefficients of CD changed. While no independent variable was found to be significant in the estimation equation obtained with OLS (Table 3), the results obtained with Ridge regression show that LWB, BT, AE, CW, CD variables are statistically significant (Table 5). All of these results were consistent with the literature (Montgomery et al., 2001; Yakubu, 2009; Yakubu, 2010; Dorman et al., 2013). In the Ridge regression method, the visual of the variation of VIF values and standardized regression coefficients with the Ridge parameter is given in Figure 1 (a) and (b), respectively. As can be seen in Figure 1(a), the VIF value decreased with the change of the Ridge parameter. It is seen in Figure 1 (a) that the Ridge parameter is λ=0.2 when all VIF values are observed to be around 1. Accordingly, the Ridge parameter was determined as λ=0.2 in the Ridge regression method. Figure 1(b) shows the variation of the coefficients standardized according to the Ridge parameter. It is seen that there is no significant change in the standardized coefficients after the selected λ=0.2, and the process becomes stagnant.

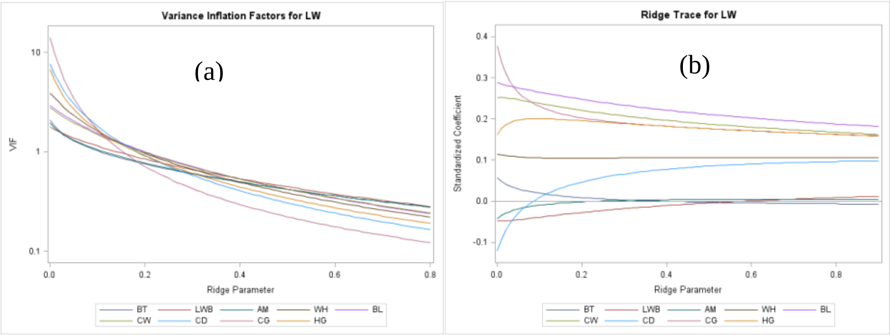


Figure 1. Variation in VIF values and coefficients according to the Ridge parameter

The Ridge regression method is an estimation method used to overcome the multicollinearity problem (Hoerl and Kennard, 1970; Marquardt and Snee 1975; Hoerl and Kennard, 1988; Dormann et al., 2013). The information that variable selection cannot be made using this method, only estimation is available in various studies in the literature (Tibshirani, 1996; Topal et al., 2010; Çiftsüren & Akkol, 2018). The aim of this study is to obtain the predicted values and to select the variable in order to reduce the model complexity. Therefore, the data set used in the study was analyzed by using LASSO, Adaptive LASSO, Naïve Elastic Net and Elastic Net methods, which are among the methods that make estimation and variable selection together, respectively. The coefficients and standard errors obtained as a result of the analysis are given in Table 5. For all the methods used in the study, the prediction equations were found significant (p<0.01). Table 5 shows the results related to the coefficient values obtained. Accordingly, the results are summarized as follows: WH, BL CW, CG, and HG for LASSO regression; CW and CG for Adaptive LASSO; BL, CG and HG for Naïve Elastic Net; WH, BL, CW, CG, and HG for Elastic Net. The selection of the variables in the estimation equation was carried out by removing the other variables from the model by forcing them to zero, as required by the working principle of these methods (Tibshirani, 1996; Efron et al., 2004; Zou & Hastie, 2005; Wang et al., 2011).

It can be seen in Table 5 that the results obtained from LASSO and Elastic Net methods are similar to each other in terms of the number of variables that should remain in the model. The results of the two methods show that it was decided to keep the same variables in the model. These are WH, BL CW, CG, and HG. At the same time, the coefficients and standard errors were generally quite close to each other. These estimates were 0.0735±0.0352, 0.0998±0.0347 for WH; 0.3991±0.0944, 0.3704±0.1042 for BL; 0.6711±0.3302, 0.6622±0.1779 for CW; It was 0.2651±0.0966, 0.2376±0.0957 for CG and 0.1704±0.0370, 0.1908±0.1654 for HG in LASSO and Elastic Net methods, respectively.

Since the number of independent variables in the data set used in the study was less than the number of observations and a group of variables did not have binary correlations, LASSO and Elastic Net methods produced close results and these results were found to be compatible with the literature (Efron et al., 2004; Zou & Hastie, 2005). However, it can be seen that the similarity of the results obtained using the LASSO and Elastic net method is compatible with the results of the study by

Akkol and Çiftdüren'in (2018) and the results of the study by Iqbal et al, (2019), especially for male Harnai lambs.

Since Adaptive LASSO and Naïve Elastic Net produced the same results in terms of the number of variables decided to remain in the model, the Adaptive LASSO and Naïve Elastic Net results given in Table 5 were evaluated together. Because in both methods, 3 variables were chosen to remain in the model. The variables that were decided to remain in the model with Adaptive LASSO were determined as BL, CW, and CG. In addition, the coefficient and standard errors were calculated as 0.4231±0.1056, 0.6771±0.4965, and 0.4559±0.0587, respectively. The variables selected using Naïve Elastic Net were determined as BL, CG, and HG, and their coefficients and standard errors were calculated as 0.3860±0.0592, 0.5328±0.0927 and 0.0214±0.0148, respectively. In both methods, BL and CG are included in the model, while the third variable is CW for Adaptive LASSO; It is designated as HG for Naïve Elastic Net.

The plots for LASSO, Adaptive LASSO, Naïve Elastic Net and Elastic Net were presented in Figure 2, Figure 3, Figure 4, and Figure 5 respectively. In all Figures, coefficient progression and AIC with cut off step were visualized in (a) and progression of Average Squared Errors with selected step was in (b). It was shown in these figures that how the selection process was made.

In Figure 2, it is seen that the LASSO selection is completed at the point where the AIC value is the smallest. Accordingly, 5 variables CG, BL, CW, HG and WH were included in the model. ASE was minimal in the 5 selected variables. After this stage, the reduction in ASE was not found to be significant.

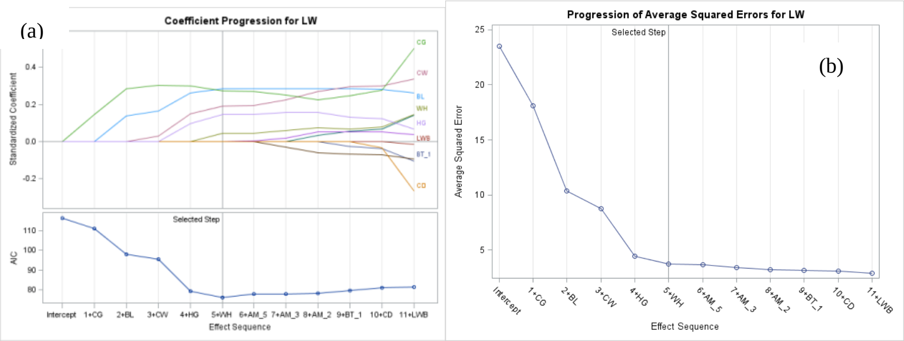


Figure 2. The coefficient progression (a) and progression of Average Squared Errors (b) for LASSO selection method

In Figure 3, it is clearly seen that the Adaptive LASSO selection reaches the best model with 3 variables. These variables are CG, CW, and BL. In Figure 3 (a), it can be observed that the selection phase is completed when the smallest AIC value is reached. After the selection step, it is seen that the coefficient progression is completed in a stabilized manner. At the point where the selection for the Adaptive LASSO method is complete, the ASE value very close to the ASE value of LASSO (Figure 3 (b)).

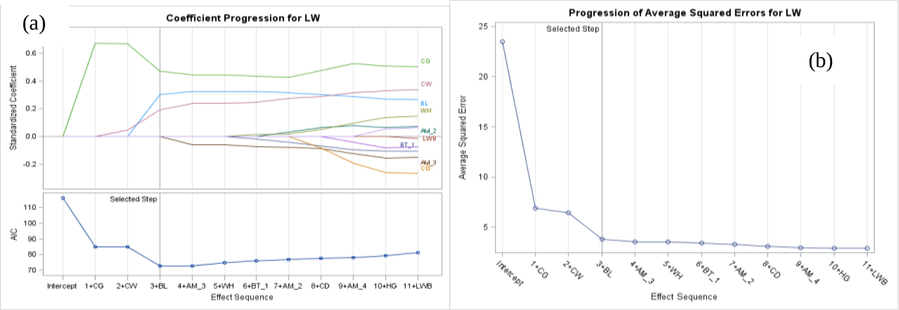


Figure 3. The coefficient progression (a) and progression of Average Squared Errors (b) Adaptive LASSO selection method

In Figure 4 (a), the coefficient and AIC progressions are monitored together in the Naïve Elastic Net method (Günes, 2015). Accordingly, it is seen that the fit of the model is made with the presence of 3 variables in the model. In Figure 4 (a), it is observed that the ASE value is minimum for the 3-variable model and the ASE value increases dramatically after this point.

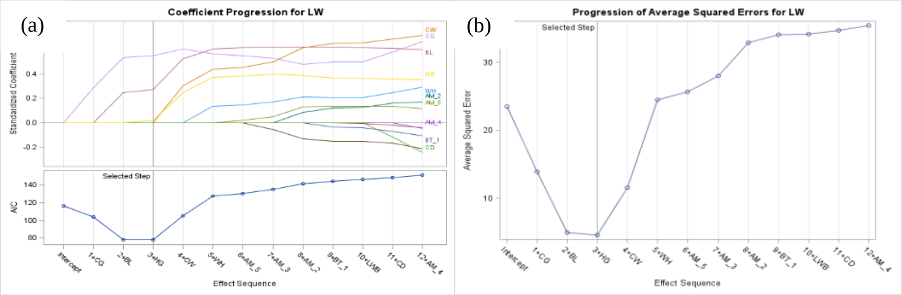


Figure 4. The coefficient progression (a) and progression of Average Squared Errors (b) for Naïve Elastic Net selection method

In Figure 5 (a), the variables that should be in the model and the change in the coefficients of the variables are observed at the point where the AIC is minimum according to the Elastic Net method. Accordingly, the model with four variables in the model, except for the intercept, was determined as the best model. In Figure 5 (b), the change in the ASE is observed by adding another variable to the model after the five variables that were decided to be included in the model. This change shows that the decrease in ASE changes very small when each variable is included in the model (Gunes, 2015).

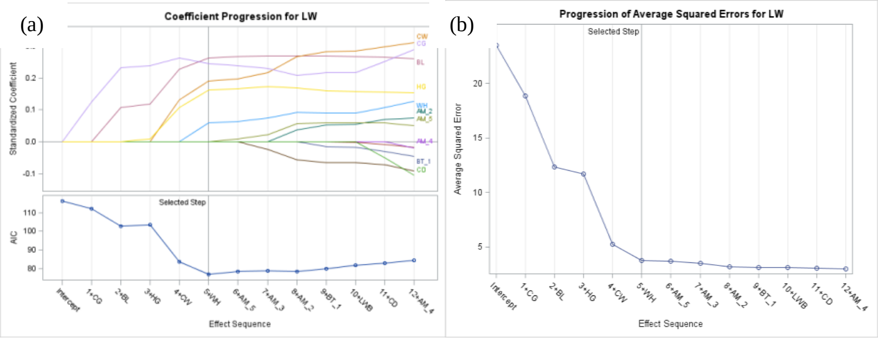


Figure 5. The coefficient progression (a) and progression of Average Squared Errors (b) for Elastic Net selection method

To compare Ridge, LASSO, Adaptive LASSO, naïve Elastic Net, and Elastic Net methods and to select the reduced model the model fir criteria were given in Table 6. Results regarding MSE, RMSE, number of variables for each method are given in the Table 6. In addition, AIC, AICc, and ASE values are given for non-Ridge methods.

Table 6. The model-fit criteria and standardized coefficient for Ridge, LASSO, Adaptive LASSO, Naïve Elastic Net, and Elastic Net regularization methods in multiple linear regression analysis

The Model-Fit Criteria	Ridge	LASSO	Adaptive LASSO	Naïve EN	Elastic Net
MSE	5.6626	4.7495	4.4238	5.3947	4.8579
RMSE	2.3761	2.1793	2.1033	2.32264	2.20407
R^2	0.8488	0.8429	0.8397	0.8046	0.8393
R^2_{adj}	0.7688	0.8055	0.8188	0.7791	0.8011
AIC		76.28148	72.81951	78.17677	76.89094
AICc		82.17622	75.67666	81.03391	82.78568

ASE		3.69404	3.76839	4.59544	3.77838
Variables	Standardized Coefficients				
LWB	-0.0274	-	-	-	-
BT	0.0085	-	-	-	-
AM	-0.0017	-	-	-	-
WH	0.1044	0.0432	-	-	0.0586
BL	0.2477	0.2839	0.3009	0.2746	0.2634
CW	0.2208	0.1924	0.1942	-	0.1899
CD	0.0467	-	-	-	-
CG	0.2027	0.2741	0.4713	0.5508	0.2457
HG	0.1956	0.1446	-	0.0182	0.1619
No of Variable	10	5	3	3	5

MSE: Mean Square Error; RMSE: Root MSE; R^2: Coefficient of determination; R^2_{adj} : adjusted coefficient of determination; AIC: Akaiki Information Criteria; AICc: Corrected AIC and ASE: Average Square Error.

In Table 6 it was seen that while fitting the model in Ridge regression, all variables were kept in the model and predictions were made for each of them. When Ridge regression was used, it was determined that the result was consistent with the knowledge that only estimation was made and no variable selection was made (Hoerl & Kennard, 1970; Hoerl & Kennard, 1988; Zou & Hastie, 2005). In other analysis methods used in the study, variable selection and estimation were made together. When the results of the analysis are examined, it is seen that the fit of the model is made with 3 variables in Adaptive LASSO and Naïve Elastic Net, and with 5 variables in the LASSO and Elastic Net methods. When Table 6 is examined considering the number of variables decided to remain in the model, it is seen that the largest MSE value was obtained from Ridge (5.6626), followed by the values obtained from Naïve Elastic Net, Elastic Net, LASSO and Adaptive LASSO, respectively. Similar ranking applies in RMSE. The RMSE values from largest to smallest are in Ridge, Naïve Elastic Net, Elastic Net, LASSO and Adaptive LASSO, respectively. Sorted from largest to smallest for the adjusted coefficient of determination, Adaptive LASSO (81.88%), LASSO (80.55%), Elastic Net (80.11%), Naïve Elastic Net 77.91%, and Ridge (76.88%). Since the best model is the model with the smallest MSE (and thus RMSE) and the largest adjusted coefficient of determination (Akaike, 1974; Hurvich & Tsai, 1989; Maydeu-Olivares & Garci´a-Forero, 2010), the fit of the data is best done, the ranking for the methods was Adaptive LASSO, LASSO, Elastic Net, Naïve Elastic Net and Ridge. Therefore, the most unsuccessful estimation equation for the data set used in this study was obtained by the Ridge method. The results where Ridge was more unsuccessful than other methods was consistent with previous studies (Ogutu et al., 2012, Çiftsüren & Akkol, 2018; Iqbal et al., 2019). When the other compliance criteria in Table 6 were evaluated in terms of AIC, AICc and ASE for methods other than Ridge, the ranking did not change. Accordingly, methods other than Ridge are listed as Adaptive LASSO, LASSO, Elastic Net, Naïve Elastic Net in terms of the best explanation of the model. Results on Naive Elastic Net and Elastic Net estimation methods are ranked after Ridge and were the other worst method for fitting the data set. These methods were developed when the study data had a group of highly correlated variables, and the number of variables is too large for the number of observations. The fact that our study data does not have these features explains the result in accordance with the literature (Zou & Hastie, 2005; Hastie et al., 2007; Friedman et al., 2010). In order of model selection, Adaptive LASSO was the best method for the study data. Accordingly, the best estimation equation for the data set used in the study was obtained with the Adaptive LASSO method. The result of the Adaptive LASSO method being the preferred method over LASSO was compatible with the literature (Fan & Li, 2001; Zou, 2006; Huang et al., 2008; Ogutu et al., 2012; Iqbal et al., 2018; Akkol, 2018).

The estimation equation obtained using Adaptive LASSO was created with standardized coefficients. According to this result, the biggest effect on the explanation of LW belonged to CG, followed by BL and CW, respectively. It was consistent with various studies in which the effect of CG, BL, and CW for LW was significant (Yılmaz et al., 2011; Yılmaz et al., 2013; Sabbioni et al., 2020; Tirink et al., 2020; İbrahim et al., 2021). The coefficient of determination and the adjusted coefficient of determination for this estimation equation were 83.97% and 81.88%, respectively. This rate was higher than that reported by İbrahim et al. (2021) (78.2% and 77.3%, respectively) and lower than that reported by Sabbioni et al. (2020) (the smallest adjusted coefficient of determination 86.6%).

CONCLUSION

The estimation and selection of variables, which are the aims of this study, which tried to obtain the estimation equation by using some morphological features of the live weight of Karakaş sheep, were made with LASSO, Adaptive LASSO, Naive Elastic Net, Elastic Net methods, apart from Ridge regression. As a result of the analyses made with the aforementioned methods, it was determined that there were very small differences when the predicted values were compared. The best prediction equation was obtained with Adaptive LASSO. It has been concluded that all of the estimation equations obtained by using penalized/shrinkage regression methods (LASSO, Adaptive LASSO, Naive Elastic Net, Elastic Net, except Ridge), which make estimation and variable selection together, are more reliable and accurate than OLS in cases of multicollinearity problems.

REFERENCES

Acharjee, A., Finkers, R., Visser, R. G., & Maliepaard, C. (2013). Comparison of regularized regression methods for~ omics data. Metabolomics, 3(3), 1-9.

Afolayan, R. A., Adeyinka, I. A., & Lakpini, C. A. M. (2006). The estimation of live weight from body measurements in Yankasa sheep. Czech Journal of Animal Science, 51(8), 343.

Akaike, H. (1974). A new look at the statistical model identification. IEEE transactions on automatic control, 19(6), 716-723.

Akkol, S. (2018). The prediction of live weight of hair goats through penalized regression methods: LASSO and adaptive LASSO. Archives animal breeding, 61(4), 451-458.

Albayrak, A. S. (2005). Çoklu bağlantı halinde en küçük kareler teknikleri ve bir uygulama. Uluslararası Yönetim İktisat ve İşletme Dergisi, 1(1), 105-126.

Alpar, R. (2003). Uygulamali çok değişkenli istatistiksel yöntemler. Detay Yayıncılık, Ankara, 853 pp.

Belsley, D. A. (1991). A guide to using the collinearity diagnostics. Computer Science in Economics and Management, 4(1), 33-50.

Çankaya, S., & Kayaalp, G. T. (2007). Estimation of relationship between live weights and some body measurements in German farm x hair crossbred by canonical correlation analysis. Hayvansal Üretim, 48(2).

Çankaya, S., Eker, S., & Abacı, S. H. (2019). Comparison of Least Squares, Ridge Regression and Principal Component Approaches in the Presence of Multicollinearity in Regression Analysis. Turkish Journal of Agriculture-Food Science and Technology, 7(8), 1166-1172.

Çiftsüren, M. N., & Akkol, S. (2018). Prediction of internal egg quality characteristics and variable selection using regularization methods: ridge, LASSO and elastic net. Archives Animal Breeding, 61(3), 279-284.

Das, B., Nair, B., Reddy, V. K. & Venkatesh, P. (2018). Evaluation of multiple linear, neural network and penalised regression models for prediction of rice yield based on weather parameters for west coast of India. International journal of biometeorology, 62 (10), 1809-1822.

Dormann, C. F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carré, G., Marquéz, J. R. G., Gruber, B., Lafourcade, B., Leitão, P. J., Münkemüller, T., McClean, C., Osborne, P. E., Reineking, B., Schröder, B., Skidmore, A. K., Zurell, D., & Lautenbach, S. (2013). Collinearity: a review of methods to deal with it and a simulation study evaluating their performance, Ecography, 36 (1), 027-046, doi: 10.1111/j.1600-0587.2012.07348.x

Efron, B., Hastie, T., Johnstone, I., & Tibshirani, R. (2004). Least angle regression. The Annals of statistics, 32(2), 407-499.

Eyduran, E., Waheed, A., Tariq, M. M., Iqbal, F., & Ahmad, S. (2013). Prediction of live weight from morphological characteristics of commercial goat in Pakistan using factor and principal component scores in multiple linear regression. The Journal of Animal & Plant Sciences, 23(6)-1532-1540.

Fan, J., & Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. Journal of the American Statistical Association, 96(456), 1348-1360.

Filzmoser, P. (2008). Linear and nonlinear methods for regression and classification and applications in R. Institut fr Statistik und Wahrscheinlichkeitstheorie Working Paper, 1(1), 1-52.

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. Journal of statistical software, 33(1), 1-20.

Gökdal, Ö., Ülker, H., Karakus, F., Cengiz, F., Temur, C., & Handil, H. (2004). Growth, feedlot performance and carcass characteristics of Karakas and crossbred lambs (F1)(Ile de France x Akkaraman (G1) x Karakas) under rural farm conditions in Turkey. South African Journal of Animal Science, 34(4), 223-232.

Gunes, F. (2015). Penalized regression methods for linear models in SAS/STAT®. In Proceedings of the SAS Global Forum 2015 Conference. Cary, NC: SAS

Institute Inc. http://support.sas.com/rnd/app/stat/papers/2015/Penalized Regression_LinearModels.

Hastie, T., Taylor, J., Tibshirani, R., & Walther, G. (2007). Forward stagewise regression and the monotone lasso. Electronic Journal of Statistics, 1, 1-29.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Prediction.Inference and Data Mining, 2nd edn. Springer Verlagi California.

Hoerl, A. E., & Kennard, R.,W. (1970). Ridge regression: biased estimation for non-orthogonal problems, Technometrrics, 12, 69-82.

Hoerl, A. E. & Kennard, R.W. (1988). Ridge regression, In Encyclopedia of Statistical Sciences, 8, 129–136.

Huang, J., Ma, S. & Zhang, C. H. (2008). Adaptive Lasso for sparse high-dimensional regression models. Statistica Sinica, 1603-1618.

Hurvich, C. M. & Tsai, C. L. (1989). Regression and time series model selection in small samples. Biometrika, 76(2), 297–307. https://doi.org/10.1093 /biomet/76.2.297.

Ibrahim, A., Artama, W. T., Budisatria, I. G. S., Yuniawan, R., Atmoko, B. A., & Widayanti, R. (2021). Regression model analysis for prediction of body weight from body measurements in female Batur sheep of Banjarnegara District, Indonesia. Biodiversitas Journal of Biological Diversity, 22(7).

Idrees, N., & Kamal, S. (2021). Modelling of wheat production in Punjab through the regularized regression approach while addressing multicollinearity. Pakistan Journal of Agricultural Sciences, 58(1).

Iqbal, F., Ali, M., Huma, Z. E., & Raziq, A. (2019). Predicting live body weight of harnai sheep through penalized regression models. J. Anim. Plant Sci, 29(6), 1541-1548.

Jahan, M., Tariq, M. M., Kakar, M. A., Eyduran, E., & Waheed, A. (2013). Predicting body weight from body and testicular characteristics of Balochi male sheep in Pakistan using different statistical analyses. JAPS, Journal of Animal and Plant Sciences, 23(1), 14-19.

Malau-Aduli, A. E. O., Aziz, M. A., Kojima, T., Niibayashi, T., Oshima, K., & Komatsu, M. (2004). Fixing collinearity instability using principal component and ridge regression analyses in the relationship between body measurements and body weight in Japanese Black cattle. Journal of Animal and Veterinary Advances, 3(12), 856-863.

Marquardt, D. W., & Snee, R. D. (1975). Ridge regression in practice. The American Statistician, 29(1), 3-20.

Maydeu-Olivares, A. & Garcia-Forero, C. (2010). Goodness-of-fit testing. International encyclopedia of education, 7(1), 190-196.

Montgomery, D. C., Peck, E. A., & Vining, G. G. (2001). Introduction to linear regression analysis. John Wiley & Sons.

Ogutu, J. O., Schulz-Streeck, T., & Piepho, H. P. (2012). Genomic selection using regularized linear regression models: ridge regression, lasso, elastic net and their extensions. In BMC proceedings. 6(2), 1-6.

Onk, K., Sari, M., & Gurcan, I. S. (2018). Estimation of live weights at the beginning and the end of grazing season in Tuj lambs via scores of factor analysis. Ankara Üniv Vet Fak Derg, 65, 261-266.

Sabbioni, A., Beretti, V., Superchi, P., & Ablondi, M. (2020). Body weight estimation from body measures in Cornigliese sheep breed. Italian Journal of Animal Science, 19(1), 25-30.

SAS: SAS/STAT User’s Guide: Version 9.4. SAS Institute Inc., Cary, NC, USA, 64 2014.

Şeker, İ. & Kul, S. (2000). İvesi ve Ost-Friz x İvesi (F1) koyunlarda beden ağırlığı, beden ölçüleri ve bunlar ile süt verimi arasındaki ilişkiler. Yüzüncü Yıl Üniversitesi Veteriner Fakültesi Dergisi, 11(2), 123-126.

Tariq, M. M., Eyduran, E., Bajwa, M. A., Waheed, A., Iqbal, F. & Javed, Y. (2012). Prediction of body weight from testicular and morphological characteristics in indigenous mengali sheep of Pakistan using factor analysis scores in multiple linear regression analysis. International Journal of agriculture and Biology, 14(4).

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso, Journal of the Royal Statistical Society: Series B (Statistical Methodology), 58, 267–288.

Tirink, C., Abacı, S. H., & Onder, H. (2020). Comparison of ridge regression and least squares methods in the presence of multicollinearity for body measurements in saanen kids. Journal of the Institute of Science and Technology, 10(2), 1429-1437.

Topal, M., Eyduran, E., Yağanoğlu, A. M., Sönmez, A., & Keskin, S. (2010). Çoklu doğrusal bağlantı durumunda ridge ve temel bileşenler regresyon analiz yöntemlerinin kullanımı. Atatürk Üniversitesi Ziraat Fakültesi Dergisi, 41(1), 53-57.

Wang, S., Nan, B., Rosset, S., & Zhu, J. (2011). Random lasso. The annals of applied statistics, 5(1), 468.

Weisberg, S. (2005). Applied linear regression. Third Edition, John John Wiley & Sons., New York, USA, 330 pp.

Yakubu, A. (2009). Fixing collinearity instability in the estimation of body weight from morpho-biometrical traits of west African dwarf goats. Trakia J.Sci. 7:61-66.

Yakubu, A. (2010). Fixing multicollinearity instability in the prediction of body weight from morphometric traits of White Fulani cows, Journal of Central European Agriculture, 11, 487-492.

Yilmaz, A., Tepeli, C., Tekin, M. E., Akmaz, A., Garip, M., Polat, E. S., Coşkun, B., & Çağlayan, T. (2011). Determination of live weights and body measurements of Kangal Type Akkaraman sheep in producers conditions. Journal of Food, Agriculture & Environment, 9(2), 366-370.

Yilmaz, O., Cemal, I., & Karaca, O. (2013). Estimation of mature live weight using some body measurements in Karya sheep. Tropical Animal Health and Production, 45(2), 397-403 doi: 10.1007/s11250-012-0229-7, 2013.

Zhang, L., Tedde, A., Ho, P., Grelet, C., Dehareng, F., Froidmont, E., Gengler, N., Brostaaux, Y., Hailemariam, D., & Soyeurt, H. (2021). Mining data from milk mid-infrared spectroscopy and animal characteristics to improve the prediction of dairy cow's liveweight using feature selection algorithms based on partial least squares and Elastic Net regressions. Computers and Electronics in Agriculture, 184, 106106.

Zou, H. (2006). The adaptive lasso and its oracle properties. Journal of the American statistical association, 101(476), 1418-1429.

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net, Statistical Society: Series B, 67, 301–320.

CHAPTER IV

CHI-SQUARE ANALYSIS IN GENETIC DATA

Assist. Prof. Dr. Fatma İLHAN[*]

[*] Selcuk University, Faculty of Agriculture, Department of Animal Science, Konya, Turkey. fatmailhan@selcuk.edu.tr

INTRODUCTION

Humans have been engaged in agriculture for thousands of years. At the end of the 19th century, a new perspective in agriculture was formed with the studies of Mendel. After Mendel's discovery, the laws of inheritance were expanded with new information and formed an important and large branch of science, genetics. Genetics is the cornerstone of breeding science and makes the most contribution compared to other sciences. Especially in recent years, where the world population has increased rapidly, the importance of agricultural improvement has increased even more. Since the aim of breeding studies is to increase the yield of agricultural products, we must know how to make the best use of genetic variations that affect these yields. For example, differences among animals in ability to tolerate disease are clearly inherited and, it is possible to make a more efficient agriculture with a selection made by taking advantage of this feature.

Statistics has an important place in genetics as in many branches of science. Mendel also shows that statistical methods should be used in heredity studies in his work that forms the basis of genetics. Statistics also forms the basis of bioinformatics, which has become very popular in the field of genetics in recent years. In this study, the chi-square method, which is a statistical method in genetic studies in the field of agriculture, is introduced, its usage areas and its calculation are shown.

1. Chi-Square

The chi-square statistic is used when the variable is measured at the nominal (also called categorical) level. In addition, the continuous variables specified by measurement can be qualified as more or less than a certain degree and the chi-square test can be applied. "Pearson's chi-square test" is the best known "chi-square test" and often the word "chi-square-test" is used for Pearson's chi-square test. Expected values in the Pearson chi-square test should not be too small. If the expected values are less than a certain number, this test should not be used and another suitable test should be chosen (Kılıç, 2016).

The chi-square test is characterized by degrees of freedom (df). The chi-square distribution can take various forms depending on the degree of freedom. Although the distribution for small degrees of freedom is skewed to the right, it approaches symmetry (normal distribution) with increasing degrees of freedom. If the number of n in the chi-square distribution is above 30, the distribution is similar to the normal distribution. The mean of the distribution is equal to df and the variance is equal to the twice df. Chi-square values take values between zero and plus infinity.

The graph below provides a comparative picture of the chi-square distribution with 1, 5 and 10 degrees of freedom.

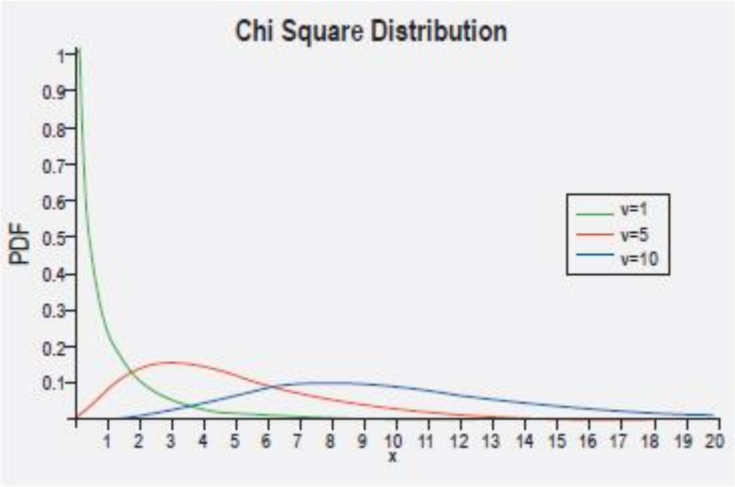


Figure 1. Shapes of chi-square distributions with different degrees of freedom

As it is seen from figure 1, one degree of freedom and chi-square distributions with higher degrees of freedom differ in shape and therefore need to be treated differently in statistical analysis (Singh, 2014). Yates correction is recommended for samples with a chi-square degree of freedom of 1 (Kocabaş, et. al., 2007).

Chi-square is generally used to test whether there is a difference between two or more groups, whether there is a relationship between two variables, whether there is homogeneity between groups, and whether the distribution obtained from the sample fits a desired theoretical distribution. Chi-square test is a test based on whether the difference between observed frequencies and expected frequencies is statistically significant.

$$\chi^2 = \sum_{i=1}^{k} \frac{(observed\ value - expected\ value)^2}{expected\ value}$$

If Yates correction is required, Chi-square is calculated as follows.

$$\chi^2 = \sum_{i=1}^{k} \frac{(|observed\ value - expected\ value| - 0.5)^2}{expected\ value}$$

The chi-square test is generally used in genetic data to check whether the population is in Hardy Weinberg equilibrium. The basis of the Hardy-Weinberg law for genetic inferences in the field of population genetics in terms of constituting has an important place. According to Hardy Weinberg's law, gene and genotype frequencies do not change from generation to generation in populations without selection, mutation and migration (Guo and Thomson, 1992). In cases where the sample size is sufficient and the class frequencies are not less than five, whether the population is in equilibrium or not can be determined by the chi-square test. If the data do not meet the prerequisites for the chi-square test, then the G statistics and exact tests are used instead of the chi-square (Çıtak and Kesici, 1999; Eyduran et al., 2005).

The chi-square method can be used to check whether the data obtained as a result of genetic studies are in Hardy Weinberg equilibrium and to check whether the results obtained in crossing studies comply with the Mendelian ratio.

2. Hardy Weinberg Equilibrium

Populations in which allele frequencies do not change across generations are in the Hardy Weinberg equilibrium. This phenomenon was first described in the early twentieth century (Wigginton et al., 2005). The allele frequency of a population remains constant over generations if all the following conditions are met (Klug et al, 2009):

-Random mating

-There should be no beneficial or harmful mutations that would alter the structure of genes.

-There should be no out-of or in-population migration and no mixing or hybridization with another neighboring population.

-The population should be sufficiently large or consist of an infinite number of individuals.

-There should be no selection for any trait or against any trait.

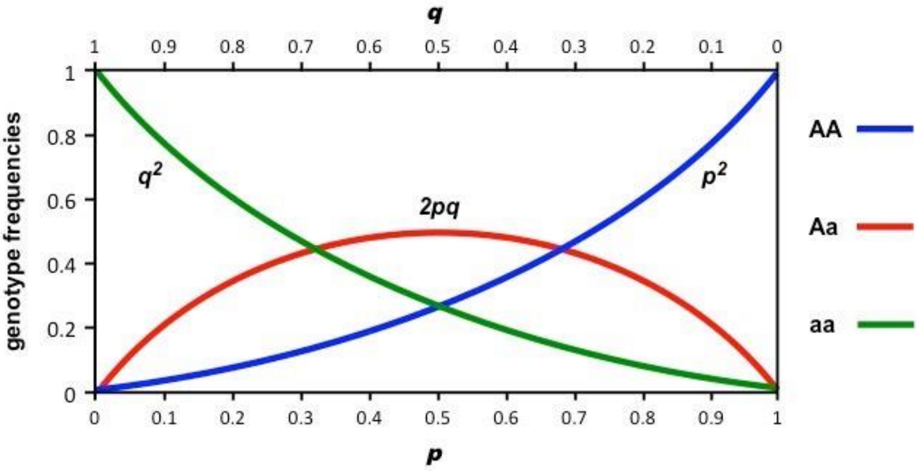


Figure 2. The relationship between the genotype frequencies originating from the Hardy-Weinberg equation and the allele frequencies

In general, the frequencies of all three genotypes can be estimated when the frequency of one of the alleles is known and Hardy-Weinberg conditions are assumed to exist (Klug et al, 2009). The relationship between genotype frequencies and allele frequencies is given in Figure 2. As it is seen from the figure, the ratio of heterozygotes increases as the p and q values move away from zero or one.

Hardy Weinberg balance is controlled in almost all molecular studies in agriculture. The most common test used to check that a population is in Hardy Weinberg equilibrium is the goodness-of-fit chi-square test (Wigginton et al., 2005). The null hypothesis is formed that randomly selected alleles provide the expected odds in the Hardy Weinberg equilibrium (ie p^2, 2pq and q^2). If the null hypothesis is rejected as a result of the chi-square test, it means that the population under consideration does not fulfill one or more of the above-mentioned conditions.

There are different opinions on how to calculate the degrees of freedom to be used when performing the chi-square test to check the Hardy-Weinberg equilibrium. Some researchers stated that it should be calculated with the formula k(k-1)/2. In this formula, k is the number of alleles at the studied locus (Weir, 1996; Wittke-Thompson, 2005; Zhou et al, 2009). Others have used the k-1-m formula. Here k is the number of genotypes and m is the number of independent allele frequencies estimated from the data (Freeman and Herron, 2004; Klug et al, 2009).

2.1. Gene and Genotip Frequensis

Genotype frequency is the proportion of a particular genotype formed by any gene pair to the number of all genotypes that gene pair can make. Gene frequency (allele frequency) is the proportion of the number of a particular allele in the population to the total number of alleles in the allele pair. For example, if there are N individuals in a population, there are 2N genes in the gene pool for trait or for a particular locus. Alleles for this feature are shown by the letters A and a. Allele frequencies are calculated as follows:

$$A\ allele\ frequensis\ P_A = \frac{n_A}{2N}$$

$$a\ allele\ frequensis\ P_a = \frac{n_a}{2N} = 1 - P_A$$

Where there are two alleles at one locus, A represents the dominant and a recessive allele. In this case, the frequency of A allele is p and the frequency of a allele is indicated by q. If the population is at equilibrium, (p + q) = 1 and the frequency of individuals with genotypes AA, Aa and aa will be p^2, 2pq and q^2 respectively. These are the terms of the binomial $(p+q)^2$ and their sum is equal to 1 (Düzgüneş and Ekingen, 1983).

For example, if a population has 350 individuals with A phenotype and 150 individuals with a phenotype and the population is in the Hardy Weinberg balance, allele and genotype frequencies are calculated as follows.

Since individuals with a phenotype have aa genotype, the frequency of aa genotype is q^2.

$aa=q^2= 150/500= 0.3$

$q=\sqrt{0.3}=0.55$

(p+q)=1

p+0.55=1

p=0.45

AA=p^2=0.45^2=0.2025

Aa=2pq=2x0.45x0.55=0.495

So;

Frequensis of AA=0.2025

Frequensis of Aa=0.495

Frequensis of aa=0.3

Sum =1.00

2.2. Two Alleles

In a population with two alleles for a trait, it is calculated as in the example below whether the population is in Hardy Weinberg equilibrium.

The genes that determine color in Shorthorn cattle show codominant characteristics. In a herd of 200 heads, there are 30 white, 80 red and 90 gray cattle. To check whether this cattle population is in Hardy Weinberg equilibrium, we first need to calculate the observation genotype frequencies.

Genotypes of red-colored cattle: AA

Genotypes of white-colored cattle: aa

Genotypes of gray -colored cattle: Aa

$$Frequensis\ of\ AA\ genotype\ = \frac{80}{200} = 0.4$$

$$Frequensis\ of\ Aa\ genotype\ = \frac{90}{200} = 0.45$$

$$Frequensis\ of\ aa\ genotype\ = \frac{30}{200} = 0.15$$

Allele frequencies are calculated with the following formula.

$$P(A) = \frac{(2xNumber\ of\ animals\ with\ AA\ genotype) + Number\ of\ animals\ with\ Aa\ genotype}{2xTotal\ number\ of\ animals}$$

Frequensis of A allele $P(A) = \frac{(2x80)+90}{2x200} = 0.625$

Frequensis of a allele $P(a) = \frac{(2x30)+90}{2x200} = 0.375$

The expected frequencies are calculated as follows.

AA=p^2xN=0.625^2x200=78.125

Aa=2xpxqxN=2x0.625x0.375x200=93.75

Aa= q^2xN=0.375^2x200=28.125

$$X^2 = \frac{(|80-78.125|-0.5)^2}{78.125} + \frac{(|90-93.75|-0.5)^2}{93.75} + \frac{(|30-28.125|-0.5)^2}{28.125} = 0.0439 + 0.156 + 0.117 = 0.712$$

The chi-square degrees of freedom are calculated using the formula k(k-1)/2. Since the number of alleles is 2 in this example, the degree of freedom is: 2(2-1)/2 = 1.

Chi-square table value with 1 degree of freedom at 5% significance level is 3.84. Since the chi-square value found above is less than the table value, it is decided that this cattle population is in Hardy Weinberg equilibrium in terms of color.

2.3. Multiple alleles

In a population with more than two alleles for a trait, whether the population is in Hardy Weinberg equilibrium is calculated as in the example below.

If the genotypes were found as follows as a result of any molecular study;

Genotypes	Observed	Observed frequensis	Expected frequensis
AA	28	15.05	0.36^2 x 186=24.11
AB	56	30.11	2 x 0.36 x 0.37 x 186=49.55
BB	35	18.82	0.37^2 x 186 =25.46
AC	23	12.37	2 x 0.36 x 0.27 x 186=36.16
BC	12	6.45	2 x 0.37 x 0.27 x 186=37.16
CC	32	17.20	0.27^2 x 186 =13.56
Sum	186	100	

To calculate the expected frequencies, the allele frequencies need to be calculated.

Frequensis of A allele $P(A) = \frac{(2x28)+56+23}{2x186} = 0.36$

Frequensis of B allele $P(B) = \frac{(2x35)+56+12}{2x186} = 0.37$

Frequensis of C allele $P(C) = \frac{(2x32)+23+12}{2x186} = 0.27$

$$X^2 = \frac{(28-24.11)^2}{24.11} + \frac{(56-49.55)^2}{49.55} + \frac{(35-25.46)^2}{25.46} + \frac{(23-36.16)^2}{36.16} + \frac{(12-37.16)^2}{37.16} + \frac{(32-13.56)^2}{13.56}$$
$$= 0.628 + 0.840 + 3.575 + 4.789 + 17.035 + 25.076 = 51.943$$

Since the number of alleles is 3 in this example, the degree of freedom is: 3(3-1)/2 = 3. Chi-square table value with 3 degree of freedom at 5% significance level is 7.815. Since the chi-square value found above is greater than the table value, it is decided that this population is not in Hardy Weinberg equilibrium. This means that this population does not fulfill one or more of the Hardy Weinberg assumptions mentioned earlier.

3. Mendelian Ratio

As a result of his studies, Mendel showed that the phenotypic ratio in the F_2 generation was 3:1 as a result of monohybrid crossing, and

9:3:3:1 as a result of dihybrid crossing. These ratios occur when the following conditions are met (Klug et al, 2009):

- Each allele is dominant or recessive
- Segregasyon occurs normally
- Independent expansion takes place
- There is random mating.

In the chi-square analysis to be done to check whether the F_2 generation fits the Mendelian ratio, the null hypothesis is formed as phenotypic ratio fits the Mendelian ratio.

An example of this is given below.

In cattle, hornless is dominant to horned (H>h) and black color is dominant to red (B>b) color. The numbers of animals with these characteristics in a herd are given below.

Fenotypes	Genotypes	Observed
Horned-Black	H-B-	571
Horned-Red	H-bb	402
Horniness-Black	hhB-	375
Horniness-Red	hhbb	252
	Sum	1600

Since dihybrid crossis results are given in this example, the expected frequencies should be calculated according to the 9:3:3:1 expansion of the Mendel.

$$Horned - Black\ (expected) = \frac{9x1600}{16} = 900$$

$$Horned - Red\ (expected) = \frac{3x1600}{16} = 300$$

$$Horniness - Black\ (expected) = \frac{3x1600}{16} = 300$$

$$Horniness - Red\ (expected) = \frac{1x1600}{16} = 100$$

Fenotypes	Genotypes	Observed	Expected
Horned-Black	H-B-	571	900
Horned-Red	H-bb	402	300
Horniness-Black	hhB-	375	300
Horniness-Red	hhbb	252	100
	Sum	1600	1600

$$X^2 = \frac{(571-900)^2}{571} + \frac{(402-300)^2}{402} + \frac{(375-300)^2}{375} + \frac{(252-100)^2}{252} = 189.56 + 25.88 + 13.99 + 91.68 = 321.11$$

Since there are 4 phenotpic groups in this example, the degree of freedom is calculated as 4-1=3.

Chi-square table value with 3 degree of freedom at 5% significance level is 7.815. Since the chi-square value found above is higher than the table value, it is decided to reject the null hypothesis. That is, the phenotypic ratios of the considered characteristics of the cattle population in the sample do not match the Mendelian ratios. This may be because this population does not meet the previously mentioned conditions. For example, the genes that determine these two traits may be on the same chromosome, that is, they may be linked.

CONCLUSION

Chi-square analysis is a method used in genetics as in many other fields. In genetics, it is mostly used to check whether the population is in balance and to check whether the F_2 generation complies with the Mendelian ratios. Many population genetics researchers want to check whether the population they are working on is in equilibrium. Although the researchers suggested that different methods can be used to control the Hardy-Weinberg balance, the Chi-square test is a widely used test in terms of being both convenient and practical for large samples (Hernandez and Weir, 1989; Guo and Thomson, 1992; Hosking et al., 2004) .

REFERENCES

Çıtak, B. & Kesici, T. (1999). Exact Test for Investigation of Hardy-Weinberg Equilibrium. Turkish Journal of Veterinary and Animal Sciences, 23(EK2), 435-440.

Düzgüneş, O. & Ekingen, H.R. (1983). Genetik. Ankara Üniversitesi Ziraat Fakültesi yayınları, Ankara.

Eyduran, E., Özdemir, T. & Küçük, M. (2005). Hayvancılıkta kategorik verilerde Ki-kare ve G testi. Yüzüncü Yıl Üniversitesi Veteriner Fakültesi Dergisi, 16(1), 1-3.

Freeman, S. & Herron, J.C. (2004). Evolutionary Analysis. third ed. Pearson Education, Inc., Upper Saddle River, NJ. 802p.

Guo, S. W. & Thompson, E. A. (1992). Performing the exact test of Hardy-Weinberg proportion for multiple alleles. Biometrics, 361-372.

Hernandez, J. L. & Weir, B. S. (1989). A disequilibrium coefficient approach to Hardy-Weinberg testing. Biometrics, 53-70.

Hosking, L., Lumsden, S., Lewis, K., Yeo, A., McCarthy, L., Bansal, A. & Xu, C. F. (2004). Detection of genotyping errors by Hardy–Weinberg equilibrium testing. European Journal of Human Genetics, 12(5), 395-399.

Kılıç, S. (2016). Ki-Kare Testi. Journal of Mood Disorders (JMOOD);6(3):180-2.

Klug, S., Cummings, M.R. 2000. Concepts of Genetic (Türkçe çeviri: Öner, C., Sümer, S., Öner, R., Öğüş, A. & Açık, L. (2002). Genetik kavramlar. Palme Yayıncılık, Ankara).

Kocabaş, Z., Özkan, M. & Başpınar, E. (2007). Temel Biyometri. Ankara Üniversitesi Ziraat Fakültesi yayınları, Ankara.

Singh, P. (2014). 2 x 2 Contingency Table: Fishers Exact Test. J Clin Prev Caridiol. 3(2):66-70.

Weir, B. S. (1996). Disequilibrium. In: Genetic data analysis II: methods for discrete population genetic data. Sinaur Associates, Sunderland, MA, pp 91–139.

Wigginton, J. E., Cutler, D. J. & Abecasis, G. R. (2005). A note on exact tests of Hardy-Weinberg equilibrium. The American Journal of Human Genetics, 76(5), 887-893.

Wittke-Thompson, J. K., Pluzhnikov, A. & Cox, N. J. (2005). Rational inferences about departures from Hardy-Weinberg equilibrium. The American Journal of Human Genetics, 76(6), 967-986.

Zhou, J. J., Lange, K., Papp, J. C. & Sinsheimer, J. S. (2009). A heterozygote–homozygote test of Hardy–Weinberg equilibrium. European journal of human genetics, 17(11), 1495-1500.

CHAPTER V

POTENTIAL USE OF SOME BODY MEASUREMENTS AS INDIRECT SELECTION CRITERIA AT WEANING PERIOD IN CENTRAL ANATOLIAN MERINO LAMBS: CANONICAL CORRELATION EXAMPLE

Şükrü DOĞAN[*]
Assist. Prof. Dr. Bülent BÜLBÜL[**]
Uğur DEMİRCİ[*]
Mesut KIRBAŞ[*]
Assoc. Prof. Dr. İbrahim AYTEKİN[***]
Prof. Dr. İsmail KESKİN[***]
Prof. Dr. Mehmet Bozkurt ATAMAN[****]

[*] Bahri Dağdaş International Agricultural Research Institute, Konya, Turkey
[**] Dokuz Eylul University, Faculty of Veterinary Medicine, İzmir, Turkey
[***] Selcuk University, Faculty of Agriculture, Department of Animal Science, Konya, Turkey
[****] Selçuk University, Faculty of Veterinary Medicine, Konya, Turkey

INTRODUCTION

In the breeding of any yield trait by selection, firstly knowing the effects of environmental factors and secondly standardization with correcting the yields according to these known factors will positively affect the success of the breeding (Boztepe, 1994). Identifying factors contributing to yield is important to increase breeding efficiency. For this reason, it is important to have easily measurable characters and useful relationship with yield to practice indirect selection for the high yield (Gashaw et al., 2007). In addition, knowing the effects of environmental factors can give an idea about which ones should be intervened in increasing efficiency. If the criteria determining the characteristics studied are obtained with difficult and costly methods, it has always been a referred way to focus on indirect criteria for selection in the early period. Such attempts in animal breeding are called indirect selection and are widely used (Düzgüneş, 1976). Morphometric traits can be used as indirect selection for body weight because they are of easy and low-cost measurement and are genetically correlated with body weight (Oliveira et al. 2021).

For this reason, determining the characters considered in animal breeding as early as possible allows the selection results to be obtained earlier and to prevent economic loss. For indirect selection, the relationships between early-detected traits and late-detected and economically important traits must be correctly determined. The linear relationship between traits can be determined by considering these traits by twos and calculating the correlation coefficient between them

(Keskin et al., 2005).

Correlation coefficient is an important statistical procedure used to evaluate breeding programs for high yield, as well as canonical correlation analysis is for determining the strength of relationships among physiological, morphological or chemical traits of plants or animals (Keskin and Yasar, 2007). Canonical correlation is a multivariate analysis technique focuses on the correlation between a linear combination of the variables in one set and a linear combination of the variables in another set (Johnson and Wichern, 2007; Altay and Yiğit, 2021). Canonical correlation analysis has been used in various fields including psychology, social science, political science, ecology, education, sociology, physical sciences, tourism, and marketing. In addition to this, there are also so many research results applied by this method in poultry science and animal science (Jaiswal et al. 1995; Akbas and Takma 2005; Yang et al. 2006; Kim et al. 2017; Karabacak et al. 2019a; Karabacak et al. 2019b; Altay et al. 2020).

Some studies have been conducted to determine the canonical correlations between the economically important traits in sheep (Tatar and Eliçin, 2002; Esmen and Davis, 2004; Keskin and Özsoy, 2004; Yaprak et al., 2008; Çankaya et al., 2009; Keskin and Dağ, 2009; Karabacak et al., 2009; Sahin et al., 2011; Tahtalı et al., 2012; Figueroa et al., 2020).

This study was carried out to evaluate the relationship between some body measurements during weaning period and body weights at different periods of Central Anatolian Merino lambs by using canonical correlation analysis method and to investigate the possibility of using them as indirect selection criteria.

MATERIAL AND METHOD

Animal Material

The animal material of the study consisted of all Central Anatolian Merino (Merino) lambs reared in the sheep breeding unit of Bahri Dağdaş International Agricultural Research Institute and born within ± 5 days of the average birth date of the populations in 2015 and 2016. Records of 355 merino lambs were used in the study. Merino was obtained by crossing G2 and G3 level German Mutton Merino x Akkaraman cross-breed sheep and rams in Konya Farm. Development was started in 1952 at the Konya State Farm. Merino genotype in this breed is over 85% (Kaymakçı and Taşkın, 2008; Yılmaz et al., 2013).

Breeding, Feeding and Data Collection

Since inseminations were recorded in the research institute, the average expected date of birth was calculated according to the insemination records and all lambs born within ± 5 days of this birth date were included in the study. Some other factors such as birth weight, sex, birth type related to lambs and ewe age, ram age and gestation period related to parents were recorded at birth. The breeding, feeding and data collection of the lambs within the scope of the study were performed according to the animal welfare rules stated in Article 9 in government law in Turkey (No. 5996) in the institute conditions and no different application was made. Lambs born in every 10-day period in the institute are divided into different groups. The lambs were offered ad libitum dry alfalfa from birth. At the age of one week, ad libitum standard lamb starter feed was started to be given by crep feeding method. Every other day suckling started at an average age of 85 days and weaned at the age of 90-95 days. Weaned lambs were acclimated to the pasture in addition to the defined feeding system. Up to 6 months of age, male and female lambs were fed together with a ration of pasture + concentrate feed + alfalfa (Medicago sativa). From the age of 6 months, they were separated into sex groups and fed in separate herds with a ration of pasture + concentrate feed + alfalfa (or a mixture of Vicia sativa (vetch) + Horduem vulgare (barley) grass) until adult age.

Body Measurements

Birth weight (BW0) was recorded with the birth of lambs. Live weight of lambs (up to 18 months) in weaning period (90th day) and every three months following this period were determined with 100 g precision weighing. Body weight of lambs were weighed in the morning before they were fed or sent to pasture.

Body length (BL), Withers height (WH), Rump height (RH), Chest depth (CD), Chest width (CW), Chest girth (CG), Shin girth (SG), Head length (HL), Brow length (BL), Head width (HW), Neck length (NL), Tail length (TL), Tail junction width (TW), Tail girth (TG) have been determined. Data sets were obtained for body weight (birth, 3rd, 6th, 12th, 15th and 18th month) and other body measurements (3rd month) to be used in statistical analysis. A measuring bench that can hold the animal steady is designed for measurements. Measurements according to the method reported by Ertuğrul (1996); it was carried out by the same person using a measuring stick, measuring tape and measuring caliper.

Statistical Analysis

The data in the sets were standardized according to the least squares method. For this, the application reported by Doğan et al. (2019) was used. For canonical correlation analysis, body measurements of lambs in the weaning period (3rd month) constituted the first variable set (13 traits), and body weight values determined in 7 different periods constituted the second set of variables (BW0, BW3, BW6, BW9, BW12, BW15, BW18). The relationship between the two data sets was analyzed using one of the multivariate statistical methods "Canonical Correlation Analysis". CANCORR procedure of SPSS (Version 21.0) package program was used in the analysis. The SPSS canonical correlation macro (canonical correlation.sps) is part of the SPSS Base system and can be found in the installation directory (Nimon et al. 2010).

Breeding values and genetic parameters of lambs according to live weight in the 3rd month were estimated based on the Single Trait Animal Model (Model 1) with the Wombat program (Version 04.04.2018) (Meyer, 2007). Genetic and phenotypic relationships between live weight and body measurements of this period were determined by the Harvey (1987) program. For the genetic and phenotypic correlation results calculated between live weight and some body measurements, records of 46 rams and 382 lambs were used. Records of 61 rams, 297 ewes and 410 lambs were used to estimate the parameters of body measurements. Data for the years 2014-2018 were used to calculate the heritability of live weight. For this, 162 rams, 1455 ewes and 3970 lambs records were used. The pedigree sets used in the calculations included 10189 animals, 440 rams and 2618 ewes records.

RESULTS

The results of the multi-link test performed in case the variables in the Y variable set are defined as dependent and the variables in the X variable set as independent variables are given in Table 1. Brauner and Shacham (1998) reported that there is no multicollinearity when VIF (Variance Inflation Factor) values are between 1 and 10. The sample size should be at least 50 if the canonical correlation coefficient between variables is estimated to be 0.90 or greater (Pituch and Stevens, 2016). When Table 1 was examined, it was seen that the model created for Merino did not have multiple connections. In addition, when Figure 1 is examined, it is seen that the relationship between the created variable sets is linear.

Table 1. Multiple connection test results

Model For BW0, 3, 6, 9, 12, 15 and 18	Tolerance	VIF
RH	0.304	3.288
BL	0.255	3.925
CW	0.331	3.020
CD	0.321	3.119
CG	0.211	4.742
SG	0.433	2.311
HW	0.500	1.999
HL	0.270	3.702
BL	0.481	2.079
NL	0.794	1.259
TW	0.645	1.551
TG	0.396	2.525
TL	0.679	1.472

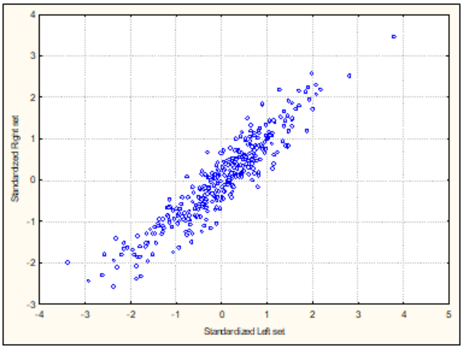


Figure 1. Relationship Between Merino Variables

Descriptive statistics of the traits examined in the study are given in Table 2. Correlation coefficients between traits are given in Table 3. The highest correlation was found between BW12 and BW15 (0.920).

Table 2. Descriptive statistics of the traits examined

Traits	N	$\bar{X} \pm S_{\bar{X}}$	CV (%)	Min.	Max.
BW0	355	4.7 ± 0.04	14.94	2.19	7.10
BW3	355	22.7 ± 0.17	14.26	14.30	33.51
BW6	355	33.6 ± 0.24	13.23	18.99	46.38
BW9	355	40.2 ± 0.25	11.67	22.74	54.14
BW12	355	45.7 ± 0.30	12.31	28.76	60.69
BW15	355	50.7 ± 0.36	13.23	29.37	70.27
BW18	355	53.7 ± 0.34	11.77	33.08	71.52
RH	355	56.9 ± 0.13	4.33	48.55	65.80
BL	355	57.6 ± 0.15	4.77	49.07	67.57
CW	355	15.3 ± 0.07	8.59	11.50	20.10
CD	355	21.7 ± 0.06	5.61	15.30	25.14
CG	355	66.2 ± 0.19	5.54	53.64	76.69
SG	355	8.0 ± 0.03	6.24	6.72	9.49
HW	355	10.9 ± 0.02	3.38	9.59	11.88
HL	355	17.6 ± 0.04	4.70	15.17	20.67
BL	355	9.3 ± 0.04	7.14	7.17	11.56
NL	355	20.3 ± 0.08	7.05	15.93	24.09
TW	355	3.7 ± 0.03	13.56	2.01	5.14
TG	355	11.8 ± 0.07	10.54	8.24	15.03
TL	355	29.0 ± 0.18	11.62	20.21	41.24

Table 3. The correlation matrix between variables in the Y and X variable sets (n= 355)

	BW0	BW3	BW6	BW9	BW12	BW15	BW18	RH	BL	CW	CD	CG	SG	HW	HL	BL	NL	TW	TG
BW 3	0.406**	1.000																	
BW 6	0.392**	0.821**	1.000																
BW 9	0.386**	0.706**	0.864**	1.000															
BW 12	0.395**	0.674**	0.768**	0.838**	1.000														
BW 15	0.345**	0.626**	0.710**	0.765**	**0.920****	1.000													
BW 18	0.354**	0.592**	0.717**	0.772**	0.847**	0.870**	1.000												
RH	0.420**	0.754**	0.674**	0.586**	0.584**	0.549**	0.545**	1.000											
BL	0.405**	0.836**	0.744**	0.648**	0.634**	0.596**	0.546**	0.737**	1.000										
CW	0.291**	0.757**	0.645**	0.574**	0.528**	0.500**	0.476**	0.590**	0.639**	1.000									
CD	0.349**	0.765**	0.660**	0.573**	0.576**	0.548**	0.529**	0.707**	0.734**	0.641**	1.000								
CG	0.369**	0.830**	0.713**	0.622**	0.613**	0.577**	0.548**	0.700**	0.721**	0.804**	0.761**	1.000							
SG	0.401**	0.697**	0.606**	0.520**	0.548**	0.523**	0.479**	0.558**	0.705**	0.582**	0.624**	0.653**	1.000						
HW	0.358**	0.678**	0.589**	0.544**	0.550**	0.505**	0.472**	0.606**	0.641**	0.517**	0.566**	0.580**	0.549**	1.000					
HL	0.376**	0.692**	0.597**	0.516**	0.488**	0.409**	0.400**	0.688**	0.651**	0.558**	0.546**	0.597**	0.544**	0.600**	1.000				
BL	0.237**	0.414**	0.364**	0.340**	0.332**	0.260**	0.220**	0.389**	0.404**	0.372**	0.352**	0.371**	0.403**	0.408**	0.698**	1.000			
NL	0.207**	0.419**	0.317**	0.281**	0.292**	0.280**	0.261**	0.404**	0.392**	0.220**	0.316**	0.299**	0.259**	0.304**	0.339**	0.145**	1.000		
TW	0.285**	0.470**	0.498**	0.443**	0.400**	0.360**	0.312**	0.333**	0.444**	0.364**	0.350**	0.446**	0.399**	0.355**	0.366**	0.194**	0.227**	1.000	
TG	0.273**	0.668**	0.587**	0.528**	0.478**	0.442**	0.379**	0.587**	0.653**	0.579**	0.533**	0.647**	0.549**	0.481**	0.573**	0.360**	0.304**	0.560**	1.000
TL	0.286**	0.528**	0.394**	0.344**	0.351**	0.348**	0.280**	0.423**	0.490**	0.360**	0.406**	0.457**	0.408**	0.328**	0.418**	0.280**	0.238**	0.258**	0.504**

**: Correlation is significant at the 0.01 level

For Merino lambs, data sets were created using live weights obtained at various periods and body measurements taken at 3 months of age, and the relationship between sets was examined by canonical correlation analysis. Canonical correlation between the sets was found to be statistically significant (P <0.01). 7 different canonical coefficients were calculated which in the study since there were 7 variables in the body weight set and 13 variables in the body measurement set (Table 4). The first three of the canonical correlation coefficients estimated as a result of canonical correlation analysis were found to be significant (P<0.05). It was observed that the canonical correlation coefficient calculated among the first canonical variable pair was 93.19%, and the coefficient values between the other pairs of variables took decreasing values.

Table 4. Canonical correlation coefficients and related test results

Pair of canonical variables	Canonical R	Canonical R^2	Chi-Square	DF	P Value	Wilks Lambda Value
U_1V_1	0.9319	0.868	806.117	91	0.000	0.096
U_2V_2	0.3114	0.097	109.634	72	0.003	0.727
U_3V_3	0.3073	0.094	74.609	55	0.041	0.805
U_4V_4	0.2152	0.046	40.528	40	0.447	0.889
U_5V_5	0.2068	0.043	24.238	27	0.617	0.932
U_6V_6	0.1371	0.019	9.217	16	0.904	0.974
U_7V_7	0.0884	0.008	2.696	7	0.912	0.992

The eigenvalues of the correlation matrix, which is the solution set of canonical correlations of the investigated traits, are given in Table 5.

86.80% of the total variation belonging to investigated traits can be explained by the first canonical variable pair calculated.

Table 5. Eigenvalues of the correlation matrix

Eigenvalues	λ_1	λ_2	λ_3	λ_4	λ_5	λ_6	λ_7
Value	0.868	0.097	0.094	0.046	0.043	0.019	0.008

Canonical coefficients to be used to reveal the relationship structure between data sets are given in Table 6 and canonical loadings in Table 7. The results are interpreted according to these coefficients.

Using the canonical coefficients in Table 6, the equation of the first canonical variable pair (U1 and V1) in which the highest relationship is predicted can be written as follows,

U_1 = 0.086 (RH) + 0.264 (BL) + 0.162 (CW) + 0.086 (CD) + 0.210 (CG) + 0.075 (SG) + 0.118 (HW) + 0.098 (HL) – 0.045 (BL) + 0.073 (NL) + 0.072 (TW) - 0.029 (TG) + 0.093 (TL)

V_1 = 0.083(BW0) + 0.840(BW3) + 0.086(BW6) - 0.022(BW9) + 0.039(BW12) + 0.081(BW15) - 0.023(BW18)

When the coefficients of the U_1 and V_1 canonical variable pair are examined; the highest contribution to the formation of the U_1 canonical variable was given by BL (0.264), later by CG (0.210), and by BW3 (0.840) in the formation of the V_1 canonical variable. Accordingly, it was seen that the increase in RH, BL, CW, CD, CG, SG, HW, HL, NL, TW and TL values and the decrease in BL and TG may cause an increase in the BW0, BW3, BW6, BW12 and BW15 values of the lambs.

Table 6. Standardized canonical coefficients for canonical variables

X Variable Set													
	RH	BL	CW	CD	CG	SG	HW	HL	BL	NL	TW	TG	TL
U_1	0.086	0.264	0.162	0.086	0.210	0.075	0.118	0.098	-0.045	0.073	0.072	-0.029	0.093
U_2	0.194	-0.588	-0.011	0.234	0.264	0.478	0.100	-0.023	-0.474	0.334	-0.666	-0.373	0.371
U_3	0.977	-0.143	-0.701	-0.210	0.027	0.643	-0.011	-0.071	0.051	-0.219	0.576	-0.800	-0.074
U_4	0.333	0.040	0.345	0.333	0.097	-0.364	-0.394	0.698	-0.645	-0.225	-0.015	-0.142	-0.707
U_5	0.414	0.441	0.291	0.017	-0.176	-0.001	0.151	-1.538	0.121	-0.102	0.198	0.203	-0.097
U_6	0.121	0.289	0.358	-0.289	-0.548	0.253	-0.801	0.537	-0.681	-0.112	0.282	-0.044	0.558
U_7	-0.868	0.831	-0.964	0.339	0.970	0.292	-0.614	0.139	0.052	-0.101	-0.268	0.067	-0.291

Y Variable Set							
	BW0	BW3	BW6	BW9	BW12	BW15	BW18
V_1	0.083	0.840	0.086	-0.022	0.039	0.081	-0.023
V_2	0.086	0.940	-1.091	-0.811	-0.606	0.341	1.210
V_3	0.766	-1.089	1.009	-0.987	0.645	-0.246	0.428
V_4	-0.396	-0.240	1.008	-0.307	-0.457	-1.290	1.580
V_5	-0.356	-0.640	0.349	0.295	-1.432	2.108	-0.186
V_6	0.425	0.049	0.989	-0.218	-2.133	1.405	-0.518
V_7	-0.425	-0.090	1.364	-1.981	1.342	-0.047	-0.533

When the canonical loads in Table 7 were examined, it was seen that the biggest contribution to the U_1 canonical variable was provided by the third month live weight of the lambs (0.992), and followed by the sixth month live weight (0.860). It was determined that the greatest contribution to the V_1 canonical variable was provided by BL (0.907), and followed by CG (0.894). In addition, CG (0.826), RH (0.823) and CW (0.807) also provided high contributions. It was determined that there was a positive correlation between body measurements and body weight changes of lambs.

Table 7. Canonical loadings of canonical variable pairs

X Variable Set													
	RH	BL	CW	CD	CG	SG	HW	HL	BL	NL	TW	TG	TL
V_1	0.823	0.907	0.807	0.826	0.894	0.764	0.740	0.745	0.450	0.449	0.524	0.717	0.567
V_2	0.084	-0.121	-0.020	0.134	0.058	0.104	-0.016	-0.193	-0.353	0.235	-0.577	-0.311	0.181
V_3	0.254	0.025	-0.309	-0.003	-0.087	0.245	0.083	0.044	0.025	-0.076	0.241	-0.276	-0.119
V_4	0.179	-0.028	0.200	0.158	0.112	-0.204	-0.160	0.043	-0.322	-0.149	-0.068	-0.135	-0.538
V_5	-0.018	0.094	0.084	0.079	0.037	0.026	-0.027	-0.611	-0.434	-0.120	0.144	0.063	-0.104
V_6	-0.023	0.066	0.022	-0.154	-0.100	0.061	-0.427	-0.022	-0.360	-0.037	0.211	0.136	0.410
V_7	-0.214	0.189	-0.185	0.176	0.141	0.221	-0.276	-0.080	0.010	-0.107	-0.091	-0.014	-0.087

Y Variable Set							
	BW0	BW3	BW6	BW9	BW12	BW15	BW18
U_1	0.484	0.992	0.860	0.754	0.740	0.695	0.651
U_2	0.032	0.027	-0.342	-0.371	-0.117	0.058	0.171
U_3	0.660	-0.111	0.189	0.095	0.297	0.266	0.349
U_4	-0.283	0.030	0.255	0.092	-0.107	-0.142	0.274
U_5	-0.267	-0.045	0.204	0.276	0.294	0.579	0.407
U_6	0.209	0.014	-0.004	-0.278	-0.502	-0.296	-0.384
U_7	-0.366	0.021	0.027	-0.351	0.007	-0.023	-0.191

When Table 8 was examined, the first canonical variable in the Y variable set explained the average 56.84% of the variation in this set. When the Redundancy Indeks values, which indicate the part of the variation in the Y variable set that can be explained by the X variable set, were examined, the value obtained for the first canonical correlation was found to be 49.36%. For the first canonical correlation, this situation showed that 56.84% of the explained variation in the Y variable set can be explained by the variables in the X variable set. In summary, for the first canonical correlation, 56.84% of the variation in the Y variable set resulted from the variation in the X variable set. When the total redundancy index was calculated, it was found as $TABI_{Y|X}$= 0.5167 (0.4936+0.0042+0.0104+0.0017+0.0048+0.0016 0.0003). This value showed that 51.67% of the variation in the Y variable set could be explained by the X variable set. It was seen that 95.92 % of this value belongs to the first canonical variable pair.

Table 8. The explained total variation ratio by canonical variables for the variable sets

X Variable Set					
	Variance Extracted	Additive Variance Extracted		Redundancy Index	Additive Redundancy
U_1	0.5258	0.5258	V_1	0.4566	0.4566
U_2	0.0565	0.5823	V_2	0.0055	0.4621
U_3	0.0302	0.6125	V_3	0.0029	0.4649
U_4	0.0475	0.6600	V_4	0.0022	0.4671
U_5	0.0490	0.7090	V_5	0.0021	0.4692
U_6	0.0452	0.7542	V_6	0.0009	0.4701
U_7	0.0250	0.7792	V_7	0.0002	0.4703

Y Variable Set					
	Variance Extracted	Additive Variance Extracted		Redundancy Index	Additive Redundancy
V_1	0.5684	0.5684	U_1	0.4936	0.4936
V_2	0.0433	0.6117	U_2	0.0042	0.4978
V_3	0.1105	0.7222	U_3	0.0104	0.5082
V_4	0.0374	0.7596	U_4	0.0017	0.5099
V_5	0.1113	0.8709	U_5	0.0048	0.5147
V_6	0.0868	0.9577	U_6	0.0016	0.5163
V_7	0.0423	1.0000	U_7	0.0003	0.5167

When the results given in Table 9 are examined, it is seen that the genetic correlation between the body weight of Merino lambs and the examined body measurements is higher than the phenotypic correlation values. It was determined that the highest genetic relationship with body weight was with VU (1.000), GG (0.999) and GC (0.945), respectively. The highest phenotypic relationship with body weight was with VU (0.863) and GC (0.860). All of the examined body measurement values were found to have a high phenotypic relationship with body weight.

Table 9. Genetic and phenotypic correlations between body weights and body measurements in the weaning (3rd month) period of Merino lambs

Trait	**BW**	**WH**	**RH**	**BL**	**CW**	**CD**	**CG**	**SG**
BW	-	0.718**	0.753**	0.863**	0.778**	0.786**	0.860**	0.724**
WH	0.936±0.079	-	0.949**	0.734**	0.554**	0.701**	0.682**	0.568**
RH	0.822±0.096	0.974±0.019	-	0.754**	0.610**	0.730**	0.715**	0.596**
BL	1.000±0.032	0.935±0.078	0.872±0.086	-	0.662**	0.760**	0.733**	0.727**
CW	0.999±0.068	0.761±0.171	0.709±0.167	1.000±0.101	-	0.661**	0.815**	0.620**
CD	0.872±0.080	0.820±0.112	0.803±0.107	0.859±0.091	0.888±0.120	-	0.783**	0.651**
CG	0.945±0.054	0.874±0.116	0.790±0.126	0.930±0.091	0.994±0.063	0.958±0.072	-	0.673**
SG	0.768±0.107	0.653±0.155	0.617±0.154	0.770±0.110	0.834±0.129	0.629±0.152	0.667±0.156	-

**: Correlation is significant at the 0.01 level. Values above the diagonal represent phenotypic correlations and the values below the diagonal represent genotypic correlations

When Table 10 was examined, it was seen that the heritability values of the body measurement characteristics of the lambs varied between 0.084 and 0.655. TL had the highest heritability (0.655). The heritability of WH, RH, BL, CW, CD, CG and SG traits with high genetic association with BW were estimated as 0.541, 0.561, 0.348, 0.218, 0.344, 0.248 and 0.484, respectively. It was determined that the heritability of related body measurements was higher than the heritability of body weight.

Table 10. Components of variance for some body measurements in the weaning period of Merino lambs

Traits	σ^2_a	σ^2_e	σ^2_p	h_a^2	$S_{\overline{h^2}}$
BW	3.95580	12.48680	16.44260	0.241	0.036
WH	3.62832	3.07857	6.70689	0.541	0.159
RH	3.85313	3.00999	6.86312	0.561	0.157
BL	2.87023	5.37037	8.24060	0.348	0.129
CW	0.40138	1.44347	1.84484	0.218	0.122
CD	0.55611	1.06081	1.61692	0.344	0.131
CG	3.65193	11.08440	14.73630	0.248	0.131
SG	0.13273	0.14167	0.27439	0.484	0.140
HW	0.06759	0.08158	0.14917	0.453	0.141
HL	0.33214	0.44370	0.77583	0.428	0.137
BL	0.06659	0.38369	0.45028	0.148	0.115
NL	0.19043	2.08768	2.27811	0.084	0.099
TW	0.07810	0.19387	0.27197	0.287	0.131
TG	0.90426	0.83983	1.74409	0.518	0.161
TL	7.65339	4.02812	11.68150	0.655	0.165

σ^2_a: Additive genetic variance, **σ^2_e:** Error variance, **σ^2_p:** Fenotipic variance, **h_a^2:** Heritability, $S_{\overline{h^2}}$ **:** Standard error

DISCUSSION and CONCLUSION

The first canonical correlation coefficient calculated between the first canonical variable pair was found to be 0.9319 (93.19%) (P<0.01). Although the variable sets under consideration were not similar, the predicted canonical correlation coefficient was found to be similar or higher than the results of other researchers. It was seen that 86.8% of the total variation could be explained with the first canonical variable pair of the properties examined. In some studies conducted by other researchers, they reported the explanation rate of the first canonical variable pairs as 70.8% (Çankaya and Kayaalp, 2007), 52.0% (Sahin et al., 2011), and 48.72% (Keskin and Yasar, 2007). It was found that the values obtained from this study were higher than the other researchers reported.

Standardized canonical correlation coefficients show the amount of change in the canonical variable in terms of standard deviation, corresponding to an increase of 1 standard deviation in the original variable. In other words, these coefficients are the coefficients that show the effects (contributions) of the original variables in that set in the formation of the canonical variable in a set (Keskin and Yasar, 2007). When the U1 and V1 canonical variable pair coefficients calculated for Merino lambs are examined; while BL (0.264), CG (0.210) and CW (0.162) provided the highest contribution in the U1 canonical variable set, BW3 (0.840) provided the highest contribution in V1 canonical variable set.

The striking point in this study is that the HW (0.118) and TL (0.093) traits in the U_1 canonical variable set contribute positively. Although HW makes a positive contribution, increasing this value can increase the risk of difficult or stillbirths. In addition, as it is seen that HL (0.098) contributes positively to live weight in Merino lambs, HL is thought to be an important criterion for Merino lambs. The BL trait contributed negatively in the canonical correlation analysis. Decreasing the BL trait and increasing the HL trait will increase the length of the nose.

When Table 9 is examined, it is seen that the genetic correlation between body weight and body measurements is higher than the phenotypic correlation values. In similar studies, genetic correlation results were reported to be higher than phenotypic correlation results in Kermani sheep (Bahreini Behzadi et al., 2007), Mehraban sheep (Gamasaee et al., 2010) and Merinolandschaf sheep (Petrović et al., 2012). It was determined that the highest genetic relationship with body weight was with BL (1.000), CW (0.999) and CG (0.945), respectively. This showed that the traits were genetically relationship. The highest phenotypic relationship with body weight was with BL (0.863) and CG (0.860). It was determined that all of the examined body measurement were in a high phenotypic relationship with bodyweight.

In similar studies, Abbasi and Ghafouri (2011) reported that live weight and chest circumference had the highest genetic correlation (0.74±0.15) in 1-year-old Makooei sheep. Jafari et al. (2014) reported that the highest genetic correlation between 1-year-old body weight in Makuie sheep was with WH (0.91±0.05), followed by BL with a value of 0.71±0.10. Ravimurugan et al. (2013) reported that body weight and body measurements were significantly related to each other in Kilakarsal sheep, and body weight had a higher correlation with chest girth. Oliviera et al. (2021) reported that CG is an indirect selection criterion because of its high genetic correlation (0.98±0.02) with body weight. In this study, the highest genetic relationship with BW was found with BL (1.000), CW (0.999) and CG (0.945). BL, CW and CG traits can contribute significantly to the improvement of body weight trait in Merino sheep. Considering this study and other research; although they are of different breeds and ages, it can be said that there is a high genetic correlation between body weight and body length and chest traits in sheep.

It is seen that the heritability values of body measurements traits in Table 10 change between 0.084 and 0.655. Janssens and Vandepitte (2004) reported middle and high heritability for three breeds of adult Belgian sheep. Researchers reported that heritability for BW, BL, CG and WH was 0.54, 0.30, 0.45 and 0.43 for Blue du Main sheep; 0.49, 0.35, 0.39 and 0.57 for Suffolk; 0.50, 0.28, 0.40 and 0.40 for Texel, respectively. Abbasi and Ghafouri (2011) reported heritability for BW, BL, CG in Makuei sheep as 0.22, 0.11 and 0.21, respectively. Supakorn et al. (2012) reported the heritability of the BW, CG, and BL in Thai sheep population as 0.32, 0.52, and 0.54, respectively. It is seen that the heritability calculated from body measurements is compatible with the literature reports, and in general, body measurements have a higher heritability than body weight. The use of body measurement as an indirect selection criterion in the breeding study for body weight may increase the success of the breeding study.

In addition to the presence of enough genotypic variation in the selection to be made in terms of more than one trait, the easy determination of one or more of these traits, which are considered together with the high relationship between the considered traits, and their early acquisition allow them to be used as indirect selection in terms of other traits. According to the results of canonical correlation analysis, it is thought that some body measurements in the 3-month period could be used as an indirect selection criterion, as it contributed positively to the body weight changes of Merino lambs up to adult age. Considering the first canonical variables (U_1V_1), it has been concluded that BL, CG, CW and HL can be used as early selection criteria, together with BW3, which provides the highest contribution from the body weight set.

As a result, the body weight trait of Merino lambs in the weaning period has low and middle heritability (0.241±0.036) and, due to some body measurements that have a high genetic correlation with body weight have medium and high heritability, these traits can be used as indirect selection criteria. In order to indirectly improve the genetics of body weight in Merino sheep, it is recommended to carry out breeding studies that take into account BL, CW and CG traits.

Acknowledgements

This study was supported by the General Directorate of Agricultural Research and Policy in Turkey, through the project called "Investigation of the Use of Some Body Measurements as Early Selection Criteria in the Weaning of Akkaraman and Central Anatolian Merino Lambs" (TAGEM/HAYSUD/2015/A07/P-01/01).

REFERENCES

Abbasi, M.A., Ghafouri-Kesbi, F., (2011). Genetic (Co)variance Components for Body Weight and Body Measurements in Makooei Sheep, Asian-Aust. J. Anim. Sci. 24 (6) : 739 – 743. DOI: 10.5713/ajas.2011.10277.

Akbas, Y., Takma, C., (2005). Canonical correlation analysis for studying the relationship between egg production traits and body weight, egg weight and age at sexual maturity in layers. Czech Journal of Animal Science 50, 163–168. DOI:10.17221/4010-cjas.

Altay, Y., Karabacak, A., Aytekin, İ. (2020). Determination of the relationship between ultrasonic measurements at different fattening times and carcass properties. Türkiye Tarımsal Araştırmalar Dergisi, 7(2), 183-191.

Altay, Y., Yiğit, S. (2021). Which Test is More Reliable for The Testing Statistical Significance of Canonical Correlation Coefficients?. Ziraat Mühendisliği, (372), 92-100.

Bahreini Behzadi, M.R., Shahroudi, F.E., Van Vleck, L.D., (2007). Estimates of genetic parameters for growth traits in Kermani sheep Journal of Animal Breeding and Genetics, 124: 296–301.

Boztepe, S., (1994). Some reproductive traits of Karacabey Merino: effect of enviromental factors,. Hayvancılık Araştırma Dergisi, 4 (2): 73-77. ISSN:1300-2031.

Brauner, N., Shacham, M., (1998). Role of Range and Precision of the Independent Variable in Regression of Data. AIChE Journal, 44 (3): 603-611. https://doi.org/10.1002/aic.690440311

Çankaya, S., Altop, A., Olfaz, M., Erener, G., (2009). Canonical correlation analysis for estimation of relationships between some traits measured at pre- and- post slaughtering periods in Karayaka hoggets. Anadolu J. Agric. Sci., 24 (1): 61-66.

Çankaya, S., Kayaalp, G.T., (2007). Estimation of relationship between live weights and some body measurements in German farm x hair crossbred by canonical correlation analysis. Hayvansal Üretim, 48 (2): 27-32.

Doğan, Ş., Pekgör, A., Soysal, M.İ., (2019). Standardization of Environmental Factor Effecting Production Traits by Least Squares Methods an Examples of Application By Excel (LSM-ex), 12th World Buffalo Congress, 18-20 September, İstanbul-Turkey.

Düzgüneş, O., (1976). Hayvan Islahı, Çukurova Üniv. Ziraat Fak. Yayınları No: 98, Adana.

Ertuğrul, M., (1996). Küçükbaş Hayvan Yetiştirme Uygulamaları (II. Baskı). Ders Kitabı. Ank. Üniv. Ziraat Fak., Yayın No: 1446. Ankara.

Emsen, E., Davis, M.E., (2004). Canonical correlation analyses of testicular and body measurements of awassi ram lambs. J. Anim. Vet. Adv., 3 (12), 842-845.

Figueroa, A., Caballero-Villalobos, J., Angón, E., Arias, R., Garzón, A., Perea, J. M., (2020). Using multivariate analysis to explore the relationships between color, composition, hygienic quality, and coagulation of milk from Manchega sheep. Journal of Dairy Science. DOI:10.3168/jds.2019-17201

Gamasaee, V.A., Hafezian, S.H., Ahmadi, A., Baneh, H., Farhadi, A., Mohamadi, A., (2010). Estimation of genetic parameters for body weight at different ages in Mehraban sheep. African Journal of Biotechnology. 9, 5218-5223.

Gashaw, A., Mohammed, H., Singh, H., (2007). Selection criterion for improved grain yields in Ethiopian durum wheat genotypes. Afri. Crop. Sci. J., 15, 25-31. DOI:10.4314/acsj.v15i1.54407

Harvey, W.R., (1987). Users guide for LSMLMW PC-1 Version mixed model least squares and maximumlikelihood computer program, Ohio State Uni. Colombus, Mimco, USA.

Jafari, S., Hashemi, A., Darvishzadeh, R., Manafiazar, G., (2014). Genetic parameters of live body weight, body measurements, greasy fleece weight, and reproduction traits in Makuie sheep breed, Spanish Journal of Agricultural Research, 12 (3) : 653-663. http://dx.doi.org/10.5424/sjar/2014123-4564

Jaiswal, U.C., Poonia, J.S., Kumar, J. (1995). Canonical correlation analysis for studying relationship among several traits: an example of calculation and interpretation. The Indian Journal of Animal Sciences 65, 765–769.

Janssens, S., Vandepitte, W., (2004). Genetic parameters for body measurements and linear type traits in Belgian Blue du Maine, Suffolk and Texel sheep. Small Ruminant Research. 54, 13-24. DOI: 10.1016/j.smallrumres.2003.10.008

Johnson, R.A., Wichern, D.W., (2007). Applied multivariate statistical analysis. 6.ed. Pearson Edication Inc., Upper Saddle River, NJ 07458: p: 539. ISBN:0-13-187715-1.

Karabacak, A., Aytekin, İ., Keskin, İ., Zülkadir, U., Boztepe, S., (2009). Investigation of Relationship Between Initial Fattening Weight and Some Body Measurements and, Carcass Traits of Akkaraman Lambs By Using Canonical Correlation Analysis. 1st International 5th National Vocational Schools Symposium'09, May 27-27, Konya, Turkey.

Karabacak, A., Altay, Y., Aytekin, İ. (2019a). Relationship between some body measurements and ultrasound measurements at the end of fattening of Akkaraman and Ivesi lambs. Bahri Dağdaş Hayvancılık Araştırma Dergisi, 8(2), 56-64.

Karabacak, A., Altay, Y., Aytekin, İ. (2019b). Akkaraman ve İvesi Kuzularının Besi Sonundaki Bazı Vücut Ölçüleri ile Ultrason Ölçüleri Arasındaki İlişkiler, 2. Internatıonal Turkish World Engineering and Science Congress, 7 - 10 Kasım, Antalya, Turkey.

Kaymakçı, M., Taşkın, T., (2008). Sheep Crossbreeding Studies in Turkey (Review). Hayvansal Üretim Dergisi, 49 (2) : 43-51.

Keskin, İ., Dağ, B., (2009). Investigation of Relationship Amongst Milk and Wool Yield Traits of Awassi Sheep by Using Canonical Corelation Analysis. J.Anim.Vet.Adv., 8 (3): 464-468.

Keskin, S., Kor, A., Başpınar, E., (2005). The Investigation of Relationships Between Some Traits Measured Pre-Slaughtering and Post-Slaughtering by Using of Canonical Correlation Analysis in Akkeçi Kids. Journal of Agricultural Sciences, 11 (2): 154-159.

Keskin, S., Özsoy, A.N., (2004). Canonical Correlation Analysis and Its an Application. Journal of Agricultural Sciences, 10 (1): 57-71.

Keskin, S., Yasar, F., (2007). Use of canonical correlatıon analysis for determination of relationships among several traits in egg plant (Solanum melongena L.) under salt stres. Pak. J. Bot., 39, 1547-1552.

Kim, T.W., Kim, I.S., Kwon, S.G., Hwang, J.H., Park, D.H., Kang, D.G., Ha, J., Kim, S.W., Kim, C.W. (2017). Identification of relationship between pork colour and physicochemical traits in American Berkshire by canonical correlation analyses. Animal Production Science, 57(6), 1179-1185.

Meyer, K. (2007). WOMBAT - A tool for mixed model analyses in quantitative genetics by restricted maximum likelihood (REML). Journal of Zhejiang University Science B. 8(11), 815–821. doi:10.1631/jzus.2007.B0815

Nimon, K., Henson, R.K., Gates, M.S., (2010). Revisiting Interpretation of Canonical Correlation Analysis: A Tutorial and Demonstration of Canonical Commonality Analysis. Multivariate Behav.Res., 45 (4) : 702–724.

Oliveira, E.J., Savegnago, R.P., de Paula Freitas, A., de Freitas, L.A., de Paz, A.C.A.R., El Faro, L., Simili, F.F., Filho, A.E.V., da CostaR.L.D., de Paz, C. C. P. (2021). Genetic parameters for body weight and morphometric traits in Santa Ines sheep using Bayesian inference. Small Ruminant Research, 106446.

Ravimurugan, T., Thiruvenkadan, A.K., Sudhakar, K., Panneersel-Vam, S., Elango, A., (2013). The estimation of body weight from body measurements in Kilakarsal sheep of Tamil Nadu, India. Iranian J. Appl. Anim. Sci. 3, 357-360.

Petrović, M.P., Caro Petrović, V., Ružić Muslić, D., Ilić, Z., Spasić, Z., Stojković, J., Milenković M., (2012). Genetic and Phenotypic Aspects of The Body Measured Traits in Merinolandschaf Breed of Sheep, Biotechnology in Animal Husbandry 28 (4) : 733-741. https://doi.org/10.2298/BAH1204733P.

Pituch, K.A. Stevens, J.P., (2016). Applied Multivariate Statistics for The Social Sciences. 6th Edition. Routledge, New York. p:621.

Sahin, M., Cankaya, S., Ceyhan, A., (2011). Canonical Correlation Analysis for Estimation of Relationships Between Some Traits Measured at Weaning Time and Six-Month Age in Merino Lambs, Bulg. J. Agric. Sci, 17 (5): 680-686.

Supakorn, C., Pralomkarn, W., Anothaisinthawee, S., (2012). Estimation of genetic parameters and genetic trends for weight and body measurements at birth in sheep populations in Thai-land. Songklanakarin J. Sci. Thechnol. 35, 1-10.

Tatar, M.A., Eliçin, A., (2002). Ile de The Research on the Relations Between Body Weight and Measurements in Sucking and Fattening Periods by the Method of Canonical Correlation in Ile de France x Akkaraman Bi Male Lamb. Journal of Agricultural Sciences, 8 (1): 67-72.

Tahtali, Y., Cankaya, S., Ulutas, Z., (2012). Canonical correlation analysis for estimation of relationships between some traits measured at birth and weaning time in Karayaka lambs. Kafkas Univ Vet Fak Derg, 18(5), 839-844. doi:10.9775/kvfd.2012.6578

Yang, Y., Mekki, D.M., Lv S.J., Yu, J.H., Wang, L.Y., Wang, J.Y., Xie, K.Z., Dai, G.J. (2006). Canonical correlation analysis of body weight, body measurement and carcass characteristics of Jinghai Yellow chicken. J.Anim.Vet.Adv., 5 : 980–984. https://medwelljournals.com/abstract/?doi=javaa.2006.980.984

Yaprak, M., Koycegiz, F., Kutluca, M., Emsen, E., Ockerman, H.W., (2008). Canonical Correlation Analysis of Body Measurements, Growth Performance and Carcass Traits of Red Karaman Lambs. J.Anim.Vet.Adv., 7 (2): 130-136.

Yilmaz, O., Cengiz, F., Ertugrul, M., Wilson, R.T., (2013). The domestic livestock resources of Turkey: sheep breeds and cross-breeds and their conservation status, Animal Genetic Resources, 52: 147–163.

CHAPTER VI

FUZZY LINEAR REGRESSION AND FUZZY PEARSON CORRELATION ANALYSIS

Assist. Prof. Dr. Derviş TOPUZ[*]

[*]Niğde Ömer Halisdemir University, Niğde Zübeyde Hanım Vocational School of Health Services, Department of Health Services Science, Niğde, Turkey. Corresponding author email: topuz@ohu.edu.tr.

INTRODUCTION

Although there are striking changes in every field in the developing world, the necessity of living in a complex environment also emerges. It is important that researchers make some predictions for the future and determine their strategies according to these estimation results in order for them to survive and adapt to the competitive environment in the face of this obligation. Researchers working in different fields of science wanted to make the necessary strategic plans by using the information they obtained from the events that took place, predicting the consequences of future situations (Topuz, 2018). Forecasting methods have also been developed and diversified to meet these needs and have become the basis for the development of new methods. Human beings have had an idea about the possible consequences of the events that they expect to occur in the future since ancient times and tried to make some predictions (Topuz, 2020). The understanding of making predictions and making strategic decisions according to the results of the estimates has reached the present day, and the importance of making estimates and studies related to forecasting has increased day by day.

Regression method constitutes one of the most important subjects of statistics science that provides calculation of predictions about the future. The regression method is widely used in science fields such as biology, mathematics, economics, engineering, medicine and agriculture. When applying the regression method, the observation values and the affected events must be displayed with a mathematical notation, that is, with the help of a function. The regression method is a statistical analysis method that allows us to find the cause-effect relationship between the dependent variable and the independent variables, and is used to examine the relationships and determine the coefficients for the appropriate values of a function (Arnold,1990).

Since regression analysis is a method used to predict unknown situations by using known values, considering the fact that even a single value in the data set can have a great effect on the parameter estimates in the regression model, the importance of creating a valid and reliable regression model is better understood. In the case of small data groups, it was concluded that the creation of such a model would be possible with the development of systems that would minimize the errors that occur in the collection of each value belonging to the variable to which the data belongs (which may be caused by measurement devices that are not suitable for their features, researcher or lack of information). However, it is often not possible to provide these conditions in real life. For example, when it is desired to determine the relationships among the various characteristics of vertebrate animals belonging to the species Spemrophylus xantphorus (according to the length of hind leg, tail length, total height, etc.), this situation may cause many uncertainties among researchers. For these reasons, researchers who prepare reports on the subject for analyzing and modeling problems explain their ideas with verbal (linguistic) expressions rather than numerical values. Therefore, the fact that the assumptions of classical statistical models cannot be realized has revealed the necessity of changing these traditional model structures and perspectives. Thus, the need for fuzzy sets and logic applications to gain effectiveness in many fields has arisen.

The uncertainties of fuzzy models are assumed to arise from imprecise data sets and people. Fuzzy approximations perform operations using fuzzy data sets and classical crisp data sets. Many fuzzy approaches have been developed using the basic features of fuzzy sets and logic theory.

In this study, Tanaka's fuzzy regression analysis approach based on the application of fuzzy membership functions and fuzzy Pearson correlation coefficient will be is to explain.

1. CLASSIC AND FUZZY CLUSTER

It is a set of clusters that is developed on the logic of any set of elements belonging to the properties that are examined on classic clusters or which is not the element of the cluster. This type of cluster requires characteristic functions to reveal the shape of the distribution of random variables. So classic clusters, $f_B(x)$ are defined with characteristic functions (Tanaka & Guo, 1999).

Using the characteristic function, each element examined is assigned only one of the values 0 and 1. The characteristic function of the B;

$$f_B(x) = \left\{\begin{matrix} 1, & x \in B & \text{if} \\ & & \\ 0, & x \notin B & \text{if} \end{matrix}\right\} \qquad (1)$$

It is possible to define (Ross, 2004; Tansu, 2012; Bede, 2013; Trillas & Eciolaza, 2015). Here it is $f_B(x): E \rightarrow \{0,1\} \rightarrow R$ (set of real numbers) the function, it is the characteristic (membership) function of cluster B, which is a subset of the universal cluster.

Each element in the cluster either belongs to a cluster or not. Partial membership is never allowed, hence the classic set theory, identifies the boundaries of the clusters and the properties of the elements of the cluster. However, the limits of any set to be formed in practice and the general characteristics of the elements that will form this cluster cannot always be determined precisely. In such cases, it is clear that the basic knowledge of classical set theory is insufficient to classify some uncertainties in our daily lives (Topuz & Keskin, 2021).

To solve such situations, there is a need for different scientific methods with basic knowledge that can measure the linguistic uncertainty of words or groups of words used in a very complex or unambiguously defined language of life. Membership functions that use fuzzy numbers to objectively are created to examine the uncertainty of scientific methods needed. With these membership functions, many ambiguities can be expressed mathematically and the values of the variables can be evaluated numerically (Zadeh, 1965; Vrusias, 2009).

Fuzzy sets membership functions with expressed $\mu_{\tilde{B}}(x_i)$ is the cluster type and they form the basis of fuzzy set theory. Quantitative variables are made more meaningful by giving membership degrees with the help of membership functions. Membership function $\mu_{\tilde{B}}(x_i)$, the degree of belonging of the cluster belonging to any fuzzy set $\mu_{\widetilde{B}}(x_i): E \rightarrow [0.0,1.0]$ the mathematical function that corresponds to a number with various degrees of membership is called "*membership function*" (Zimmermann, 1996; Tanaka, 1997; Nguyen & Wakler, 2000; Bede, 2013; Topuz & Keskin, 2021). Values calculated with the help of membership functions are also called membership degrees (Zadeh, 1978). The degree of membership refers to the degree to which any object is a member of the universal cluster.

For example; $E = \{x_i|i = 1,2,.,n\}$ Get a fuzzy universal cluster, in this universal cluster, fuzzy one $\widetilde{B}$ subset, a list of the properties of a property that is examined on the corresponding clusters by the list method,

$$\widetilde{B} = \{(x_1,\ \mu_{\widetilde{B}}(x_1)), (x_2,\ \mu_{\widetilde{B}}(x_2)), \ldots, (x_n,\ \mu_{\widetilde{B}}(x_n)) | x_n, \mu_{\widetilde{B}}(x_n)\ \ x \in E\} \quad (2)$$

in the form of. Here $\mu_{\widetilde{B}}(x_i)$: Blurred clusters are members $\widetilde{B}$ is the value that indicates how much each x_i element belongs to the $\widetilde{B}$ fuzzy set. Equality at 3;

$$\mu_{\widetilde{B}}(x_i) = \left\{\begin{array}{ll} \mu_{\widetilde{B}}(x_i) = 1 \text{ if,} & x_i \text{ completely } \widetilde{B} \text{ is a member of the cluster } x_i \in \widetilde{B} \\ 0 < \mu_{\widetilde{B}}(x_i) < 1 \text{ if,} & x_i \text{ partially } \widetilde{B} \text{ is a member of} \\ \mu_{\widetilde{B}}(x_i) = 0 \text{ if,} & x_i\ \widetilde{B} \text{ is not a member of } x_i \notin \widetilde{B} \end{array}\right\} \quad (3)$$

identified (Zadeh, 1965; Abdalla, 2012; Atanassov, 2012). Membership degree $\mu_{\widetilde{B}}(x_i)$' when the element with a value close to 1 is a high-order element of the cluster, those with a membership degree of

0 are the lower element of this cluster (Sakawa, 1993; Topuz & Keskin, 2021).

In order to solve the many uncertainties encountered in determining the degree of relations between the clusters, it is tried to show that more reliable and consistent results can be obtained by using the alternative fuzzy pearson correlation coefficient calculation method. In what follows, elements of fuzzy set theory, proposed by Tanaka et al. (1982) and the fuzzy linear regression are presented in method section.

2. FUZZY LINEAR REGRESSION MODELS

Fuzzy linear regression models are the most frequently used technique in statistical models to obtain the quantitative relationship between a dependent variable and a number of independent variables in a fuzzy environment (Akbari et al., 2012). A fuzzy linear regression model was first introduced by Tanaka et al. (1982). Since then, numerous studies for the fuzzy (linear) regression theory have been conducted, and its applications to the fuzzy (linear) regression analysis have received increasing popularity lately (Diamond, 1988; Kim & Bishu, 1998). Their applications were improved by Tanaka & Watada (1988) and Tanaka (1997). They formulated a linear regression model with fuzzy response data, crisp predictor data and fuzzy parameters as a mathematical programming problem. In real world applications, due to parameters uncertainty, instrumental, methodological, environmental and human errors, data are usually not exact but fuzzy numbers. However, due to the vagueness of data in many practical situations, the traditional least squares method cannot be applied. Therefore, the classical regression models experience some limitations, and they need to be adjusted accordingly to deal with fuzzy data (Akbari et al., 2012).

The least squares method has become widely popular in fuzzy linear regression analysis. Its popularity derives from the variety of real-world applications, and it has been applied to determine the correlation between various factors, and provide more accurate predictive guidance to practitioners and decision-makers. For instance, the fuzzy linear regression analysis has been used to analyze the correlation of temperature, relative humidity, atmospheric pressure, rainfall amount, radiation, wind speed, and carbon monoxide (CO) concentration in the atmosphere (Zhuang et al., 2009). The fuzzy regression analysis introduced by Tanaka et al. (1982) considers crisp inputs and fuzzy outputs.

2.1. Fuzzy Least Squares Regression (FLSR) Analysis

Fuzzy regression analysis is a fuzzy (or possibility) type of ordinary regression analysis. The approach is one of the most widely used statistical techniques for evaluating the functional relationship between dependent and independent variables in uncertainty situations. In fuzzy regression analysis, the relationship between dependent variables and independent variables is not as precise as in ordinary regression analysis (Wang & Tsaur, 2000). In these uncertain cases, fuzzy techniques can explain the effects of independent variables in a more realistic way. The fuzzy least squares (FLS) technique, which is an extension of the least squares technique to fuzzy set theory, was used by to estimate fuzzy parameters (Chang & Ayyup, 2001; Diamond, 1987; Wang & Tsaur, 2000). The approach is based on blurring the coefficients. Blurring can be done in two ways. It is possible by 1) blurring the model coefficients estimated by the ordinary least squares technique at a specified "h level", or 2) estimating the coefficients as fuzzy numbers (Hong et al., 2001; Nasrabadi & Nasrabadi, 2004; Ming et al., 1997).

Most of these developed fuzzy regression models are evaluated with fuzzy outputs and fuzzy parameters, but non- fuzzy (net) inputs. Fuzzy least squares regression (FLSR) analysis technique, which is generally based on linear programming (LP), is proposed in order to minimize the fuzziness of the analyzed data and the total spread of the output (see, for example (Diamond,1987; Hong et al., 2001).

According to the FLSR approach, it is assumed that the deviations between the observed values and the predicted values are caused by the uncertainty of the system structure or the blurring of the regression coefficients, not from measurement and observation errors, contrary to the OLSR analysis method (Chang & Ayyup, 2001). That is, it assumes that the coefficients of the regression analysis model are related to its blur. For this purpose, the formula below is employed to estimate parameters of FLSR;

$$\mathrm{f} = \mathrm{X} \times \tilde{\beta} \to \widetilde{\mathrm{Y}}_{\mathrm{i}}, \widetilde{\mathrm{Y}}_{\mathrm{i}} = \mathrm{f}\left(\tilde{\beta}, \mathrm{X}\right) \tag{4}$$

It is given by the function 1. Here; $\widetilde{\mathrm{Y}}_{\mathrm{i}}$, denotes the fuzzy dependent variable in the symmetric triangular property structure estimated and is shown as $\widetilde{\mathrm{Y}}_{\mathrm{i}} = (\tilde{\mathrm{y}}_{\mathrm{c}}, \tilde{\mathrm{e}}_{\mathrm{s}})$. $\tilde{\mathrm{y}}_{\mathrm{c}}$, denotes the mean value (center) and $\tilde{\mathrm{e}}_{\mathrm{s}}$, denotes the spread value.

In the case of fuzzy observations, consider a fuzzy linear regression for crisp explanatory and fuzzy response observations as follows:

$$\widetilde{\mathrm{Y}}_{\mathrm{i}} = \mathrm{f}(\tilde{\beta}, \mathrm{X}) = \tilde{\beta}_0 + \tilde{\beta}_1 \mathrm{X}_{\mathrm{i}1} + \cdots + \tilde{\beta}_{\mathrm{p}-1} \mathrm{X}_{\mathrm{i}(\mathrm{p}-1)} = \tilde{\beta}_0 + \sum_{\mathrm{i}=1}^{\mathrm{n}} \tilde{\beta}_{\mathrm{i}} \mathrm{X}_{\mathrm{i}} \quad (4.\mathrm{a})$$

$$\widetilde{\mathrm{Y}}_{\mathrm{i}} = \{\mathrm{c}_0, \mathrm{s}_0\} + \{\mathrm{c}_1, \mathrm{s}_1\} \mathrm{X}_{\mathrm{i}1} + \cdots + \{\mathrm{c}_{\mathrm{p}-1}, \mathrm{s}_{\mathrm{p}-1}\} \mathrm{X}_{\mathrm{i}(\mathrm{p}-1)} \qquad (4.\mathrm{b})$$

in which $\tilde{\beta}_{\mathrm{j}} = [\tilde{\beta}_0 \text{ ve } \tilde{\beta}_1, \tilde{\beta}_2, \tilde{\beta}_3, \ldots \tilde{\beta}_{\mathrm{j}} \ldots \ldots, \tilde{\beta}_{\mathrm{p}-1}]^{\mathrm{t}}$ are the coefficient values of the independent variables in the function and it is a set of dependent and independent variables formed in the form of $\{\mathrm{Y}_{\mathrm{i}}, \mathrm{X}_{\mathrm{i}1}, \mathrm{X}_{\mathrm{i}2}, \mathrm{X}_{\mathrm{i}3}, \ldots, \mathrm{X}_{(\mathrm{p}-1)\mathrm{n}}\} = \{\mathrm{Y}_{\mathrm{i}}, \mathrm{X}_{\mathrm{ij}}\}$, and each dependent variable observation is expressed as $\mathrm{x} \in \mathrm{X}$ $(\mathrm{i} = 1, \ldots, \mathrm{n},\ \mathrm{j} = 1,2, \ldots, \mathrm{p} - 1)$. That is, they are crisp values of the explanatory variables. It is defined by $(\mathrm{i} = 1,2,3, \ldots, \mathrm{n})$. In the fuzzy LS regression model, the data of the dependent $\widetilde{\mathrm{Y}}_{\mathrm{i}}$ variable can be real numbers or fuzzy numbers. It is generally assumed that the data for the dependent Y_{i} variable are symmetrical fuzzy numbers of interval type (Kim & Bishu, 1998).

$\tilde{\beta}_{\mathrm{j}} = [\tilde{\beta}_0 \text{ ve } \tilde{\beta}_1, \tilde{\beta}_2, \tilde{\beta}_3, \ldots \tilde{\beta}_{\mathrm{j}} \ldots \ldots, \tilde{\beta}_{\mathrm{p}-1}]^{\mathrm{t}}$ are fuzzy regression coefficients vectors with a symmetric triangular fuzzy number structure and they are fuzzy numbers in the form of $\tilde{\beta}_j = (c_j, s_j)$ $\tilde{\beta}_j$, (j: 0,1,2,3,….,p-1). c_j, is the $\mu_{\tilde{\beta}_i}(c_j) = 1$ value representing the midpoint of the coefficients, that is, the center value, and has the form $c_j = [c_1, c_2 c_3, \dots, c_n]^t$. s_j, shows the spread of the coefficients belonging to the fuzzy regression analysis model and is $s_j = [s_1, s_2, s_3, \dots, s_n\ ]^t$shaped (Wang & Tsaur, 2000; Ghoshray, 1997). Each coefficient value $\tilde{\beta}_i = \{c_j, s_j\} = \{\tilde{\beta}_i : c_j - s_j \leq \tilde{\beta}_i \leq c_j + s_j\}$ has a symmetric triangular property structure and is $\tilde{\beta}_i$ (j: 0,1,2,3,…,p-1) (Topuz, 2020).

The $\tilde{\beta}_i = \{c_j, s_j\}$ value of the fuzzy coefficients was estimated by the minimum blur method proposed by Tanaka. The method is given in the following equation (Arnold, 1990).

In Possibilistic Regression Analysis proposed by Tanaka and Watada (1988), the linear programming (LP) formulation considers triangular membership functions (not necessarily symmetric). The spreads of the calculated fuzzy coefficients are calculated with the help of equation. The LP formulation is as follow (Tanaka & Watada, 1988):

$$\min\ Z(x)\ \ s^t|X_i| = \min_{c,s}\left[s_0 + \sum_{j=0}^{n} s_j\ |X_{ij}|\right] \tag{5}$$

$\min_{c,s} J = c_1, c_2, \dots, c_n\ ,\ c_j \geq 0, \forall i;\ i = 1,2, \dots, m$ and

$\min_{c,s} J = s_1, s_2, \dots, s_n, s_j \geq 0\ \ \forall i;\ j = 0, 1,2, \dots, n$

$$\sum_{j=0}^{n} c_j X_{ij} + (1-h)\left[\sum_{j=0}^{n} s_j \left|X_{ij}\right|\right] \geq \tilde{y}_c + (1-h)\tilde{e}_s \ \forall i;\ i = 1,2,..,n$$

$$(6)$$

$$\sum_{j=0}^{n} c_j X_{ij} - (1-h)\left[\sum_{j=0}^{n} s_j \left|X_{ij}\right|\right] \leq \tilde{y}_c - (1-h)\tilde{e}_s \quad \forall i;\ i = 1,2,\dots,n$$

$$s_j \geq 0,\quad c_j\ \in R,\qquad X_{i0} = 1(\ 0 \leq h \leq 1, \forall_i = 0,1,2,\dots,n)$$

Here, Z(x): function shows the total blur in the model. m: is the number of observations regarding the dependent variable. j: the number of independent variable x_{ij}: is the ith observation value of the jth independent variable. For each predicted $\widetilde{Y}_i$ observation value, the constraint number must be 2x*n* (Moskowitz & Kim, 1993). In order to minimize the total spread, the level h, $\widetilde{Y}_i$, the predictor of each observation value Y_i, is assumed to have a turbidity tolerance $\mu_{\widetilde{Y}_i}(Y_i) \geq$ h i=1,2,..,m. (Hojati et al., 2005). In Equation 3, the objective function is weighted with the absolute values of the measurements of the distributions of the independent variables. The application of bootstrap resampling technique in fuzzy least squares regression analysis is given below.

3. Fuzzy Pearson Correlation Coefficient

It is very important to determine the relationship direction and the degree and the statistical significance of this relationship in the interaction of two variables. Classical statistical methods are widely used to calculate the relationship between two or more variables, such as x and y, and to define conventional data sets. However, classical statistical methods cannot manage uncertainties in the natural structure of data very well. Because the source of uncertainties in the natural structure of the data affects many factors. We can define the uncertainties in the natural structure of the data with fuzzy measurements and mathematically express the measurements in each step. For example, if the correlation coefficient is calculated for a fuzzy data set, then the assumption that the independent variable (X) is the same for all values in defining the inherent ambiguities between the two variables can be assumed (Şentürk & Aşan, 2007).

The fuzzy logic theory calculates the degree of linear relationship between fuzzy sets with a certain degree of membership, resulting in more reliable and consistent results (Xie & Wu, 2012; Yang, 2016). Chiang and Lin (1999) developed a crisp correlation coefficient between two fuzzy sets. The crisp correlation coefficient lies in the interval [-1, 1]. Their method takes a random sample from a crisp set, with corresponding pairs of membership functions of the two fuzzy sets to compute the correlation between those two fuzzy sets. This method developed demonstrates not only the strength of the relationship between fuzzy sets, but also whether the relationship between fuzzy sets are positive (increasing) or negative (decreasing) (Lin et al., 2007). A graphical representation of a fuzzy triangle number type data set is as shown in Figure 1 (Yongshen, 2005).

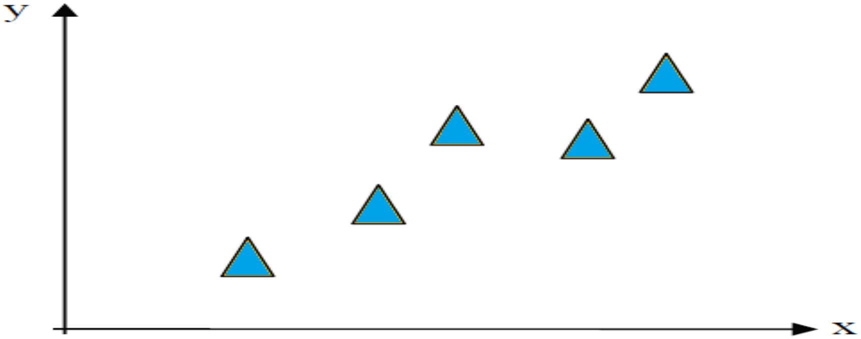


Figure 1. A blurred Data Set With a Triangle Membership Function

The relationship between two fuzzy clusters such as $\widetilde{\mathrm{A}}$ and $\widetilde{\mathrm{B}}$ is linear as in classical clusters. The value that determines the degree of this linear relationship is also called the fuzzy correlation coefficient and can be shown in the form of $\widetilde{\mathrm{A}} = \{\mathrm{x_i},\ \mu_{\widetilde{\mathrm{A}}}(\mathrm{x_i}) | \mathrm{x} \in \mathrm{X}\}$ and $\widetilde{\mathrm{B}} = \widetilde{\mathrm{A}} = \{\mathrm{x_i},\ \mu_{\widetilde{\mathrm{B}}}(\mathrm{x_i}) | \mathrm{x} \in \mathrm{X}\}$. Fuzzy correlation coefficient $\tilde{\mathrm{r}}_{\mathrm{A,B}}$ is calculated with the help of Equation (10) (Chiang & Lin, 2000; Topuz & Keskin, 2021). There must be at least two different fuzzy sets for the implementation of Equation 10. If fuzzy correlation coefficient $\tilde{\mathrm{r}}_{\mathrm{A,B}} > 0$; $\widetilde{\mathrm{A}}$ and $\widetilde{\mathrm{B}}$ fuzzy sets are positively related, and the fuzzy correlation coefficient value for membership values is $\tilde{\mathrm{r}}_{\mathrm{A,B}} = 1$ (Yu, 1993; Chiang & Lin, 2000).

If the value of the fuzzy correlation coefficient $\tilde{\mathrm{r}}_{\mathrm{A,B}} < 0$, $\widetilde{\mathrm{A}}$ and $\widetilde{\mathrm{B}}$ The relationship between fuzzy sets is negative. If value of fuzzy correlation coefficient $\tilde{\mathrm{r}}_{\mathrm{A,B}} = 0$, there is no correlation between $\widetilde{\mathrm{A}}$ and $\widetilde{\mathrm{B}}$ fuzzy sets and the value of the fuzzy correlation coefficient for the membership values of the non-cluster members is $\tilde{\mathrm{r}}_{\mathrm{A,B}} = 0$. Intermediate values that are not included in the set are the values of the fuzzy correlation coefficient, which is calculated by using the equation of uncertainty (8) and the equation (11) $0 \leq \mu_{\tilde{A},\tilde{B}}(x_i) \leq 1$ receives very different membership levels (Yu, 1993; Chiang & Lin, 1999; Lin et al., 2007).

The formula used in this study is Pearson’s product sum correlation coefficient; a pair of membership function values replaces the original data values as follows.

$$\bar{\mu}_A = \frac{\sum_{i=1}^{n}(\mu_A(x_i)}{n} \quad \text{and} \quad \bar{\mu}_B = \frac{\sum_{i=1}^{n}(\mu_B(y_i)}{n} \tag{7}$$

$$\tilde{S}_A^2 = \frac{\sum_{i=1}^{n}((\mu_A(x_i)-\bar{\mu}_A)^2}{n-1} \quad \text{and} \quad \tilde{S}_B^2 = \frac{\sum_{i=1}^{n}((\mu_B(x_i)-\bar{\mu}_B)^2}{n-1} \tag{8}$$

$$\tilde{S}_A = \sqrt{\tilde{S}_A^2} \Rightarrow \tilde{S}_A = \sqrt{\frac{\sum_{i=1}^{n}((\mu_A(x_i)-\bar{\mu}_A)^2}{n-1}} \quad \text{and}$$

$$\tilde{S}_B = \sqrt{\tilde{S}_B^2} \Rightarrow \tilde{S}_B = \sqrt{\frac{\sum_{i=1}^{n}((\mu_B(x_i)-\bar{\mu}_B)^2}{n-1}} \tag{9}$$

$$\tilde{r}_{A,B} = \frac{\frac{\sum_{i=1}^{n}((\mu_A(x_i)-\bar{\mu}_A)x\sum_{i=1}^{n}((\mu_B(x_i)-\bar{\mu}_B)}{n-1}}{\sqrt{\frac{\sum_{i=1}^{n}((\mu_A(x_i)-\bar{\mu}_A)^2 x\sum_{i=1}^{n}((\mu_B(x_i)-\bar{\mu}_B)^2}{n-1}}} \tag{10}$$

$$\tilde{r}_{A,B} = \frac{\frac{\sum_{i=1}^{n}((\mu_A(x_i)-\bar{\mu}_A)x(\mu_B(x_i)-\bar{\mu}_B)}{n-1}}{\tilde{S}_A * \tilde{S}_B} \tag{11}$$

(Chiang & Lin, 1999; Chiang & Lin, 2000; Yongshen & Cheung, 2003; Lin et al., 2007).

Here, $\mu_{\tilde{A}}(x_i)$: the membership function, which expresses the equivalent of x exact numbers in a fuzzy set such as $\tilde{A}$, $\mu_{\tilde{B}}(x_i)$: the membership function, which expresses the equivalent of x exact numbers in a fuzzy set such as $\widetilde{\mathrm{B}}$, ~ : represents values for fuzzy sets.

$\bar{\mu}_{\tilde{A}}$ and $\bar{\mu}_{\tilde{B}}$: $\widetilde{\mathrm{A}}$ and $\widetilde{\mathrm{B}}$ mean values of membership functions of fuzzy sets,

$\tilde{S}_{A}$ and $\tilde{S}_{B}$: $\widetilde{\mathrm{A}}$ and $\widetilde{\mathrm{B}}$ shows the standard deviation values of the averages of membership functions of fuzzy sets (Chiang & Lin, 1999).

Graphical fuzzy Perason correlation coefficient values are as in Figure 2 (Yongshen, 2005).

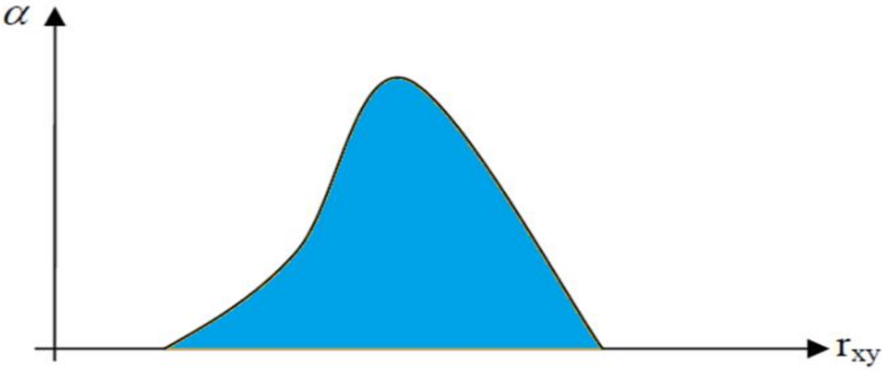


Figure 2. The fuzzy Pearson correlation coefficient values with the membership degrees

Fuzzy hypothesis testing for significance testing of fuzzy Pearson correlation coefficients can be established as follows (Buckley, 2006).

$\mathrm{H}_0: \bar{\tilde{\mu}} = 0$ and $\mathrm{H}_1: \bar{\tilde{\mu}} \neq 0$

Here; $\bar{\tilde{\mu}}$ is the average of fuzzy Pearson correlation coefficients.

The hypothesis control of fuzzy Pearson correlation coefficients and the test statistic to be used for estimation of confidence limits are $\tilde{t}$ test

(Buckley, 2006). Test statistics for hypothesis testing are obtained in Equation (12) (Chiang & Lin, 2000; Yongshen, 2005).

$$\tilde{t} = \frac{\tilde{r}_{A,B} - \bar{\tilde{\mu}}}{\tilde{S}_r} \quad t_{n-2,\frac{\alpha}{2}} \tag{12}$$

Here; $\tilde{S}_r$: is the fuzzy standard error value, supplied with:

$$\tilde{S}_r = \sqrt{\frac{1-\tilde{r}_{A,B}^2}{n-2}} \tag{13}$$

For estimation of confidence limits;

$$\bar{\tilde{\mu}} = \tilde{r}_{A,B} \pm \tilde{t}_{\frac{\alpha}{2}}; \tilde{S}_r$$

equality is used. Fuzzy Pearson correlation coefficient $\tilde{r}_{A,B}$ of %95 confidence interval is estimated as in Equality (14) (Chiang & Lin, 2000; Buckley, 2006; Topuz & Keskin, 2021),

$$\tilde{r}_{A,B} - \tilde{t}\frac{\alpha}{2}; \tilde{S}_r < \bar{\tilde{\mu}} < \tilde{r}_{A,B} + \tilde{t}\frac{\alpha}{2}; \tilde{S}_r = 1 - \alpha \tag{14}$$

The fuzzy Pearson correlation coefficient calculated with Equation 8 shows the power and the distribution of the relationship between random variables (Yongshen & Cheung, 2003; Lin et al., 2007). For this correlation method, the values of the correlation coefficient will be in the interval [-1, 1] (Chiang & Lin, 1999). As we have just described, the resultant correlation is a crisp value. A major contribution of this model is the development of partial correlation of fuzzy sets. If a random sample with multiple fuzzy attributes, Chiang and Lin's method can compute the correlation coefficient between the two fuzzy attributes.

4. NUMERICAL EXAMPLE

The material of the study was composed of data of 5 male and female species, namely, Spemrophylus xantphorus (Niğde University Project no 98.FEB-07; Çakır, 2004). The relationships between Total Length (X_1), Hind Leg Length (X_2), Tail Length (X_3), dependent variable Live Weight (gr) (Y), belonging to this rodent (Spemrophylus xantphorus), were calculated by using Pearson and fuzzy Pearson correlation coefficients calculation method and the mean fuzzy correlation coefficient values for each coefficient were compared by creating confidence intervals. For analyses EXCEL 2016, Matlab R2013a, LİNGO 16.0 and SPSS for WINDOWS Version 24.0 were used.

4.1. Estimation of Live Weights of Spemrophylus Xantphorus Species Using Fuzzy Linear Regression Analysis Method

The following systematic procedure has been followed for fuzzy regression analysis.

Step 1: Data on dependent and independent variables were obtained as in Table 1.

Table 1. 5 Sample Data Set of Various Characteristics of the Rodent of the Spemrophylus Xantphorus Species

Observation No	**LW(Y) (gr)**	**TL(X_1) (mm)**	**HLL (X_2) mm**	**TU(X_3) (mm)**
1	270	251,20	38,30	40,10
2	372	273,40	41,60	49,15
3	320	248,30	36,60	46,85
4	222	233,40	38,65	41,15
5	202	230,60	37,40	32,70
Total	1386	1236.9	192.55	209.95

LW: Live Weight (gr), TL: Total Length (mm), HLL: Hind Leg Length, TU: Tail Length (mm)

Step 2: The turbidity tolerance level was set at h = 0.5.

Step 3: Using the data of 5 Spemrophylus xantphorus vertebrates at h = 0.5 turbidity tolerance in Table 1, 10 (5 observations x 2) constraints were created as in Equation 15.

MIN = 5*s0+1236.9*s1+192.55*s2+209.95*s3;

$$\underset{a_{c1},a_{s1}}{\min} J = \begin{bmatrix} c0 + 251.20 * c1 + 38.30 * c2 + 40.10 * c3 * 0.5 * s0 + 251.20 * 0.5 * s1 + 38.30 * 0.5 * s2 + 40.10 * 0.5 * s3 \geq= 270; \\ c0 + 251.20 * c1 + 38.30 * c2 + 40.10 * c3 - 0.5 * s0 - 251.20 * 0.5 * s1 - 38.30 * 0.5 * s2 - 40.10 * 0.5 * s3 \leq 270; \end{bmatrix}$$

.

. **(15)**

$$\underset{a_{c10},a_{s10}}{\min} J = \begin{bmatrix} c0 + 230.60 * c1 + 37.40 * c2 + 32.70 * c3 * 0.5 * s0 + 230.60 * 0.5 * s1 + 37.40 * 0.5 * s2 + 32.70 * 0.5 * s3 \geq= 202; \\ c0 + 230.60 * c1 + 37.40 * c2 + 32.70 * c3 - 0.5 * s0 - 230.60 * 0.5 * s1 - 37.40 * 0.5 * s2 - 32.70 * 0.5 * s3 \leq 202; \end{bmatrix}$$

@FREE(c0); @FREE(c1); @FREE(c2); @FREE(c3); END

Step 4: Spemrophylus xantphorus 10 constraints created for the body weights of vertebrates were analyzed in the LINGO 16.0 package program to calculate the coefficient values $\widetilde{\mathrm{A}}_\mathrm{i}$, i = 0,1,..,3 and the values in Table 2 were calculated.

Table 2. Center and Spread Values of Coefficient Values Calculated at h = 0.5 Turbidity Tolerance Level

Variables	$\widetilde{\mathrm{A}}_\mathrm{i} = \{\mathrm{a}_\mathrm{i}^\mathrm{c}, \mathrm{a}_\mathrm{i}^\mathrm{s}\}$		
		h = 0.5	
	Coefficients	Center value ($\mathrm{a}_\mathrm{i}^\mathrm{c}$)	Spread ($\mathrm{a}_\mathrm{i}^\mathrm{s}$)
Constant	$\widetilde{\mathrm{A}}_0$	-1485806	0.00
TL (X_1)	$\widetilde{\mathrm{A}}_1$	1.76867	0.102
HLL (X_2)	$\widetilde{\mathrm{A}}_2$	37.879	0.00
TU (X_3)	$\widetilde{\mathrm{A}}_3$	-3.805	0.00

LW: Live Weight (gr), TL: Total Length (mm), HLL: Hind Leg Length, TU: Tail Length (mm)

Step 5: Equation 16 was created by applying the value of the objective function Z (x), which represents the blur level of the equation to be formed with the coefficients in Table 2, with the total values calculated in Table.1 and the calculated diffusion values in Table.2 in Equation 5;

$$\mathrm{Z(x)} = \left[5\ \mathrm{a}_0^\mathrm{S} + \mathrm{a}_1^\mathrm{S} \sum_{\mathrm{i}=1}^{5} \mathrm{x}_{1\mathrm{i}} + \mathrm{a}_2^\mathrm{S} \sum_{\mathrm{i}=1}^{5} \mathrm{x}_{2\mathrm{i}} + \ \mathrm{a}_3^\mathrm{S} \sum_{\mathrm{i}=1}^{5} \mathrm{x}_{3\mathrm{i}}\right] \tag{16}$$

$$\mathrm{Z} = \left[5\ \mathrm{x}\ \mathrm{a}_0^\mathrm{S} + 1236.9\mathrm{x}\ \mathrm{a}_1^\mathrm{S} + 192.55\mathrm{x}\ \mathrm{a}_2^\mathrm{S} + 209.96\mathrm{x}\ \mathrm{a}_3^\mathrm{S}\right]$$

$$\mathrm{Z} = [5\ \mathrm{x}\ 0.00 + 1236.9\ \mathrm{x}\ 0.102 + 192.55\ \mathrm{x}\ 0.00 + 209.96\ \mathrm{x}\ 0.00]$$

The $\mathrm{Z} = 126.16$ value has been calculated. This calculated value shows the level of uncertainty in the data.

Step 6: Fuzzy linear regression analysis model created using the coefficient values in Table 3.2, calculated at a very high turbidity level of Z (x) = 126.16;

$$\tilde{\mathrm{Y}}_{\mathrm{i}} = \{-1485.8; 0.0\} + \{1.768; 0.1\}\mathrm{X}_{\mathrm{i}(1)} + \{37.9; 0.0\}\mathrm{X}_{\mathrm{i}(2)} + \{-3.8; 0.0\}\mathrm{X}_{\mathrm{i}(3)} \tag{17}$$

in the form.

Step 7. 8. 9. and 10. : Using Equation 17, the mean canlı ($\tilde{\mathrm{Y}}_{c}$) live weight values estimated for 5 Spemrophylus xantphorus and the distribution, lower limit and upper limit values of these values were calculated with the equation 9,11,12 and the values in Table 3. were obtained.

Table 3. Statistics of The Estimated Average Live Weight Values of Spemrophylus Xantphorus Vertebrates Using Fuzzy Linear Regression Analysis Approach

Observation No	**Observed Body weight values (gr) ($\mathbf{Y_i}$)**	**Statistics on estimated average body weight ($\tilde{\mathbf{Y}}_{\mathbf{i}}$) values**			
		$\tilde{\mathrm{Y}}_{\mathrm{c}}$	$\tilde{\mathrm{Y}}_{\mathrm{s}}$	Lower limit values	Upper limit values
1	270	256.50	25.62	230.87	282.12
2	372	386.31	27.88	358.42	414.20
3	320	161.29	25.32	135.96	186.62
4	222	234.29	23.80	210.48	258.09
5	202	214.14	23.52	190.62	237.66
Mean	277	250.5			
S.Error	69.95	83.70	37.432		
Change	13.75				

Skewness	0.410 1.238			
Kurtosis	-1.454 2.322			
Mean of of the differences. t-statistics p-value				
13.75		0.799	0.469	
R^2: 0.540 F=1.234>0.348*				

There is no statistically significant difference between the observed live weight values (g) and the average live weight values (g) measurements of Spemrophylus xantphorus vertebrates in Table 3.

Step 7: The graphic in Figure 4 was obtained to show the consistency between the observed live weights and the estimated average weights of Spemrophylus xantphorus vertebrates in Table 3.

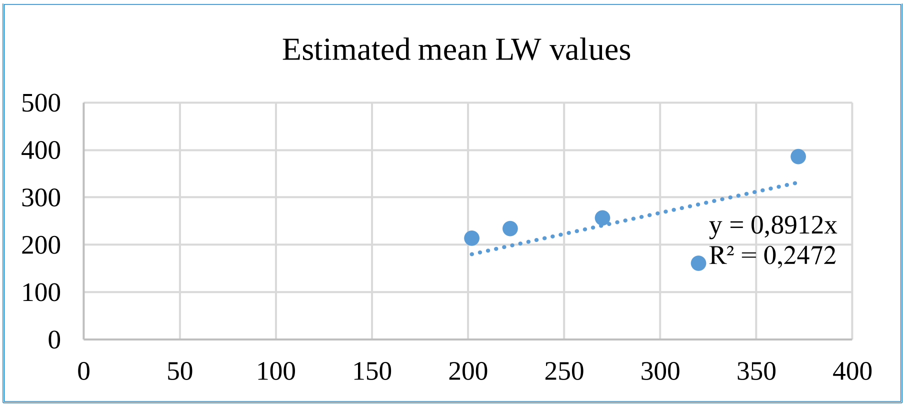


Figure 3. Graphical Representation of Observed and Predicted Average ($\tilde{Y}_c$) Live Weight Values of Spemrophylus Xantphorus Vertebrates Together

4.2. Calculation of the Relationship Between Body Weights and Total Height of Spemrophylus Xantphorus Species Using Fuzzy Pearson Correlation Analysis Method

In cases where the data does not show normal distribution, it is possible to obtain more reliable and consistent results by calculating the fuzzy Pearson correlation coefficient value which can be calculated instead of classical Pearson correlation coefficient value. In the calculations made for such studies, it has been shown that there is a strong relationship between the size of the correlation coefficients and similar studies, and it is a general fact that the researchers make wrong interpretations and make wrong decisions. Matlab codes were created to calculate these results. The fuzzy Pearson correlation coefficient calculation method mentioned in the method was used to measure the correlation coefficients and confidence level between the independent variables, Total Length (X_1), Hind Leg Length (X_2), Tail Length (X_3) and dependent variable Live Weight (gr) (Y_i) values. Relationship coefficients were interpreted by creating confidence intervals. For Spemrophylus xantphorus type, calculation of the relationship between Live weight (gr) (mm) and total length (mm) is as in Table 4.

Table 4. Live Weight (Y) (gr) and Total Length (X_1) (mm) Values of the Spemrophylus Xantphorus Species

No	LW (Y) (gr)	TL (X) (mm)	Y^2	X^2	Y.X
1	(270;0.0)	(251,20;0.0)	(72.900;0.0)	(63101.44;0.0)	(67.824;0.0)
2	(372;0.6)	(273,40;0.6)	(138384;0.36)	(7474756;0.36)	(101704.8;0.36
3	(320;1.0)	(248,30;1.0)	(102.400;1.0)	(61652.89;1.0)	(79.456;1.0)
4	(222;0.6)	(233,40;0.6)	(492.84;0.36)	(54475.56;0.36)	(51814.8;0.36)
5	(202:0.0)	(230,60;0.0)	(408.04;0.0)	(53176.36;0.0)	(46.581.2;0.0)
Total	1386;2.2	1236,9;1.72	403.772;1.72	307153.81;1.72	347.380.8;1,72

LW: Live Weight (gr), TL: Total Length (mm)

Assuming that the relationship between the Live weight (gr) of the Spemrophylus xantphorus and the Total Length (mm), is a linear relationship, the fuzzy Pero correlation coefficient value explaining this relationship was performed by using the equation of the hand calculation and significance test by fuzzy data (7, 8, 9, 10, 11, 12, 13.14) (Table 2).

$$\sum_{i=1}^{5} d_Y d_X = \sum XY - \frac{(\sum X)(\sum Y)}{N}$$

$$= 347380.8; 1.72 - \frac{(1386; 2.2) * (1236.9; 2.2)}{5}$$

$$= 347380.8; 1.72 - \frac{1714343.4; 4.84}{5}$$

$$\sum_{i=1}^{5} d_Y d_X = 347380.8; 1.72 - 342868.68; 0.97 = 14512.12; 0.75$$

$$\sum_{i=1}^{5} d_{\tilde{Y}}^2 = \sum Y^2 - \frac{(\sum Y)^2}{n} = 403772;1.72 - \frac{(1386;2.2)^2}{5}$$
$$= 403772;1.72 - 384199.2;0.97$$

$$\sum_{i=1}^{5} d_{\tilde{Y}}^2 = 19573;0.75$$

$$\sum_{i=1}^{5} d_{\tilde{X}}^2 = \sum Y^2 - \frac{(\sum X)^2}{n} = 307153.81;1.72 - \frac{(1236.9.15;2.2)^2}{5}$$
$$= 307153.81;1.72 - 305984.32;0.97$$
$$\sum_{i=1}^{5} d_{\tilde{X}}^2 = 1169.49\,;0.75$$

$$\tilde{r}_{X,Y} = \frac{\sum_{i=1}^{n} d_Y d_X}{\sqrt{\sum_{i=1}^{n} d_{\tilde{Y}}^2 d_{\tilde{X}}^2}} = \frac{14512.12;0.75}{\sqrt{(19573;0.75) * (1169.49;0.75)}}$$
$$= \frac{14512.12;0.75}{\sqrt{22890427.77;0.562}} = \frac{14512.12;0.75}{4784.39;0.75}$$

$\tilde{r}_{X,Y} = (0.943;1.0)$ the average coefficient value is calculated and the value of this coefficient value with the help of equality (12) $\tilde{t} = \frac{0.943-\bar{\tilde{\mu}}}{0.134} \rightarrow t_{5-2},\frac{0.05}{2}$ fuzzy standard error value with the help of equality (13), $\tilde{S}_r = \sqrt{\frac{1-(0.943;1.0)}{5-2}} = (0.134;0.577)$ calculated as.

The mean coefficient value is calculated and the fuzzy statistical values calculated by their representation with the degree of membership of the other calculated coefficients are shown in Figure 4.a, Figure 4.b. and Table 5. It is created as in.

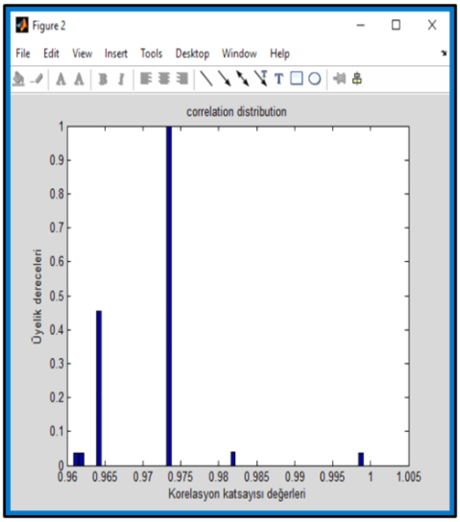


Figure 4.a) The distribution of calculated fuzzy correlation coefficients and membership degrees (Pearson = 0.9731)

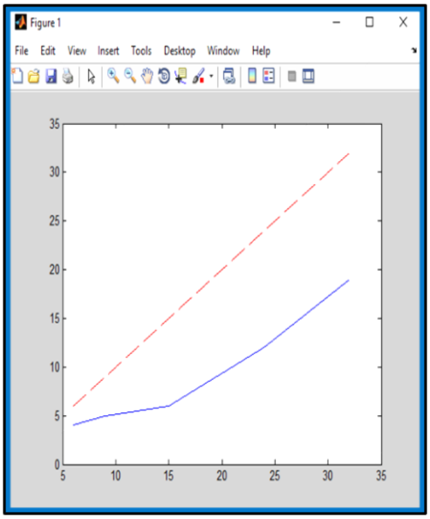
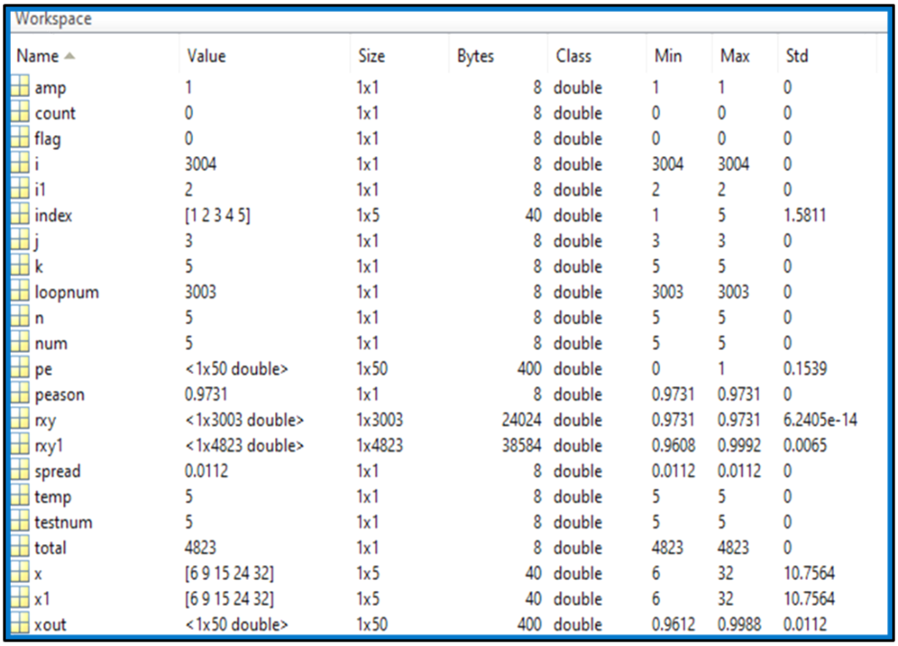


Figure 4.b) The spread value of the calculated fuzzy correlation coefficients (spread = 0.0112)

Table 5. Calculated Matter Output Showing the Mean Value of the Calculated Pearson Correlation Coefficients



Workspace

Name ▲	Value	Size	Bytes	Class	Min	Max	Std
amp	1	1x1	8	double	1	1	0
count	0	1x1	8	double	0	0	0
flag	0	1x1	8	double	0	0	0
i	3004	1x1	8	double	3004	3004	0
i1	2	1x1	8	double	2	2	0
index	[1 2 3 4 5]	1x5	40	double	1	5	1.5811
j	3	1x1	8	double	3	3	0
k	5	1x1	8	double	5	5	0
loopnum	3003	1x1	8	double	3003	3003	0
n	5	1x1	8	double	5	5	0
num	5	1x1	8	double	5	5	0
pe	<1x50 double>	1x50	400	double	0	1	0.1539
peason	0.9731	1x1	8	double	0.9731	0.9731	0
rxy	<1x3003 double>	1x3003	24024	double	0.9731	0.9731	6.2405e-14
rxy1	<1x4823 double>	1x4823	38584	double	0.9608	0.9992	0.0065
spread	0.0112	1x1	8	double	0.0112	0.0112	0
temp	5	1x1	8	double	5	5	0
testnum	5	1x1	8	double	5	5	0
total	4823	1x1	8	double	4823	4823	0
x	[6 9 15 24 32]	1x5	40	double	6	32	10.7564
x1	[6 9 15 24 32]	1x5	40	double	6	32	10.7564
xout	<1x50 double>	1x50	400	double	0.9612	0.9988	0.0112

The fuzzy Pearson correlation coefficient value (7, 8, 9, 10, 11, 12, 13, 14) is calculated with the help of equations and the membership degrees of the data in Table 1.

$$\bar{\mu}_A = \frac{\sum_{i=1}^{5}(\mu_A(x_i)}{n} = \frac{0.0 + 0.60 + 1.0 + 0.60 + 0.0}{5} = \frac{2.2}{5} = 0.44$$

$$\bar{\mu}_B = \frac{\sum_{i=1}^{5}(\mu_B(y_i)}{n} = \frac{0.0 + 0.60 + 1.0 + 0.60 + 0.0}{5} = \frac{2.2}{5} = 0.44$$

$$\tilde{S}_A^2 = \frac{\sum_{i=1}^{5}((\mu_A(x_i) - \bar{\mu}_A)^2}{n-1}$$

$$= \frac{(0.0 - 0.44)^2 + (0.60 - 0.44)^2 + (1.0 - 0.44)^2 + (0.0 - 0.44)^2}{5-1}$$

$$= \frac{0.19 + 0.16 + 0.56 + 0.19}{4}$$

$$\tilde{S}_A^2 = \frac{1.1}{4} = 0.275$$

$$\tilde{S}_B^2 = \frac{\sum_{i=1}^{n}((\mu_B(x_i) - \bar{\mu}_B)^2}{n-1}$$

$$= \frac{(0.0 - 0.44)^2 + (0.60 - 0.44)^2 + (1.0 - 0.44)^2 + (0.0 - 0.44)^2}{5-1}$$

$$= \frac{0.19 + 0.16 + 0.56 + 0.19}{4}$$

$$\tilde{S}_B^2 = \frac{1.1}{4} = 0.275$$

$$\tilde{S}_A = \sqrt{\tilde{S}_A^2} \Rightarrow \tilde{S}_A = \sqrt{\frac{\sum_{i=1}^{n}((\mu_A(x_i) - \bar{\mu}_A)^2}{n-1}} = \sqrt{0.275} = 0.524$$

$$\tilde{S}_B = \sqrt{\tilde{S}_B^2} \Rightarrow \tilde{S}_B = \sqrt{\frac{\sum_{i=1}^{n}((\mu_B(x_i) - \bar{\mu}_B)^2}{n-1}} = \sqrt{0.275} = 0.524$$

$$\tilde{r}_{A,B} = \frac{-0.44 * (-0.44) + 0.16 * 0.16 + 0.56 * 0.56 + 0.44 * (-0.44)}{0.524 * 0.524}$$

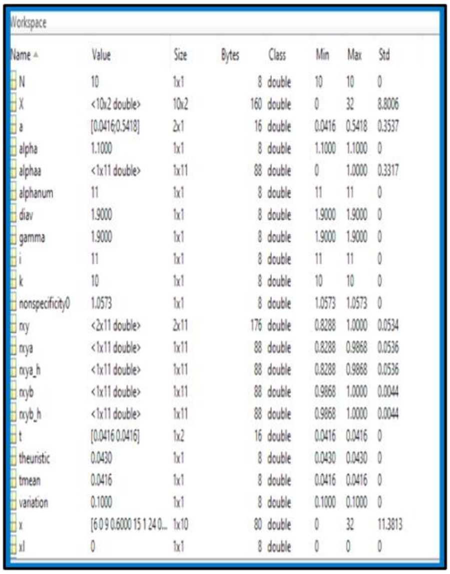
$$\tilde{r}_{A,B} = \frac{0.1936+0.0256+0.3136+0.1936}{0.274} = \frac{\frac{0.7264}{4}}{0.274} = \frac{0.1816}{0.274} = 0.66$$

$$\tilde{r}_{A,B} = \frac{\frac{\sum_{i=1}^{n}((\mu_A(x_i)-\bar{\mu}_A)X(\mu_B(x_i)-\bar{\mu}_B)}{n-1}}{\tilde{S}_A * \tilde{S}_B} = 0.66$$

$\tilde{r}_{A,B} = (0.66)$ The mean coefficient value was calculated and the value of this coefficient was calculated as $\tilde{t} = \frac{0.066-\bar{\bar{\mu}}}{0.134}\ \ t_{5-2}, \frac{0.05}{2}$ with the help of equality (12). With the help of Equation (13), the fuzzy standard error value was calculated as $\tilde{S}_r = \sqrt{\frac{1-(0.66)}{5-2}} = 0.433.$

In other words, while the coefficient value calculated by the classical method was 0.973, the coefficient value calculated using the fuzzy membership degrees was calculated to be 0.66. Fuzzy statistics and plots of the calculated values are shown in Table 6 and Figure 5a.

Table 6. Matlab Output Showing Calculated Fuzzy Correlation



Workspace

Name ▲	Value	Size	Bytes	Class	Min	Max	Std
N	10	1x1	8	double	10	10	0
X	<10x2 double>	10x2	160	double	0	32	8.8006
a	[0.0416;0.5418]	2x1	16	double	0.0416	0.5418	0.3537
alpha	1.1000	1x1	8	double	1.1000	1.1000	0
alphaa	<1x11 double>	1x11	88	double	0	1.0000	0.3317
alphanum	11	1x1	8	double	11	11	0
diav	1.9000	1x1	8	double	1.9000	1.9000	0
gamma	1.9000	1x1	8	double	1.9000	1.9000	0
i	11	1x1	8	double	11	11	0
k	10	1x1	8	double	10	10	0
nonspecificity0	1.0573	1x1	8	double	1.0573	1.0573	0
rxy	<2x11 double>	2x11	176	double	0.8288	1.0000	0.0534
rxya	<1x11 double>	1x11	88	double	0.8288	0.9868	0.0536
rxya_h	<1x11 double>	1x11	88	double	0.8288	0.9868	0.0536
rxyb	<1x11 double>	1x11	88	double	0.9868	1.0000	0.0044
rxyb_h	<1x11 double>	1x11	88	double	0.9868	1.0000	0.0044
t	[0.0416 0.0416]	1x2	16	double	0.0416	0.0416	0
theuristic	0.0430	1x1	8	double	0.0430	0.0430	0
tmean	0.0416	1x1	8	double	0.0416	0.0416	0
variation	0.1000	1x1	8	double	0.1000	0.1000	0
x	[6 0 9 0.6000 15 1 24 0...	1x10	80	double	0	32	11.3813
xl	0	1x1	8	double	0	0	0

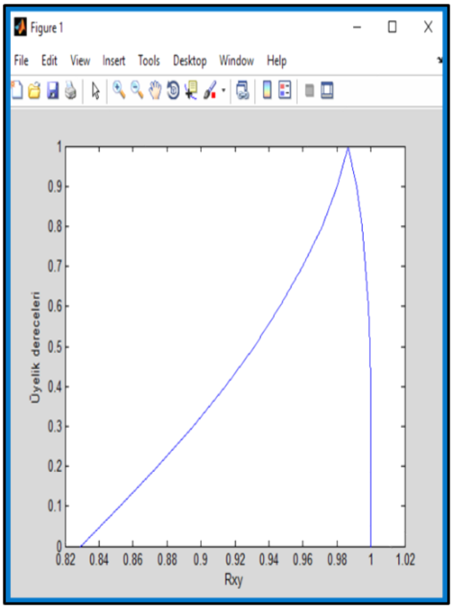


Figure 5a. The Distribution of Calculated Fuzzy Correlation Coefficients and Membership Degrees (Pearson=0.9868)

CONCLUSION

Pearson’s product-sum formula has been widely accepted to compute the correlation coefficient between two crisp random variables. In recent years, changing problems in daily life according to the conditions and continuous development of researches have led to more complex structure. The developments that have taken place during this time have revealed the necessity to change the standart analysis methods and perspectives depending on the development of science and technique. Fuzzy set theory is widely used in applications in different fields recently (Chiang et al., 2004). However, one of the fuzzy logic approaches in practice is the fuzzy Pearson correlation coefficient, which has been done to analyze the correlation of fuzzy data. For example, applications in many fields such as engineering, medicine, psychology, business, agriculture, pharmacy and veterinary medicine have been done. In this study, an easy method based on theoretical foundations of classical Pearson correlation coefficient was used. The range of the calculated fuzzy coefficient is a fuzzy number with interval [−1, 1], which consists with the classic range of Pearson correlation. In addition, it was found that the coefficient value calculated by the classical Pearson correlation method was the same as the coefficient values calculated with the fuzzy Pearson correlation method.

The aim of this study is to calculate the fuzzy linear regression and Fuzzy Pearson correlation coefficient value and membership degrees on the exact dataset using the written MATLAB codes. Frequently preferred techniques based on this fuzzy logic approach are fuzzy linear regression and Pearson correlation coefficient calculation techniques. It is observed that if the original data set is in symmetric shape, then the fuzzy correlation takes on roughly symmetric distribution; otherwise, the membership function of the fuzzy correlation is a skew distributed. In cases where the data used in the model contain uncertainty, the method works more efficiently. It is concluded that it is appropriate for researchers to use the fuzzy linear regression and fuzzy Pearson correlation coefficient method when they want to reach the approximate results with the information that does not make any certainty when deciding for any situation.

Acknowledgements

We thank Prof. Dr. Mehmet Çakır for all data used in this study (The project named Histo-anatomy of Spermophylus xantoprymnus living in Niğde Region. Niğde University Project no 98.FEB-07).

REFERENCES

Abdalla, H.A. (2012). Possibilistic logistic regression: In fuzzy environment. Lap Lambert Aca-demic Publishing. Saarbrücken- Germany

Akbari, M.G, Mohammadalizadeh, R. and Rezaei, M. (2012). International Journal of Fuzzy Systems, Vol. 14, No. 4, December.

Arnold, S.F. (1990). Mathematical Statistics. New Jersey. Prentice-Hall..

Atanassov, K. (2012). On intuitionistic fuzzy sets theory. Studies İn Fuzziness and Soft Computing, Berlin- Germany.

Bede, B. (2013). Mathematics of fuzzy sets and fuzzy logic. Springer, Heidelberg New York Dordrecht London.

Buckley, J.J. (2006). Fuzzy probability and statistics. Studies İn Fuzziness and Soft Computing, Publ., No: 2, Springer, Berlin- Germany.

Chang, Y.H.O. & Ayyub, B.M. (2001). Fuzyy regression methotds-a comparative assessment, Fuzzy Sets and Systems, Vol.119, pp 187-203.

Chiang, D.A. & Lin, N.P. (1999). Correlation of fuzzy sets. Fuzzy Sets and Systems, Vol. 2, No.102, pp 221-226.

Chiang, D.A. & Lin, N.P. (2000). Partial correlation of fuzzy sets. Fuzzy Sets and Systems, Vol.2, No.110, pp 209-215.

Chiang, J.H, Yue, S., Yin, Z. (2004). A new fuzzy cover approach to clustering. IEEE Transactions on Fuzzy Systems, Vol.2, No.12, pp 199- 208.

Çakır, M. (2004). Morphometric and Hysto-Anatomic Researches on Partes Genitales Femininae Internae of Female. *Spermophilus xanthoprymnus* (Rodentia: Sciuridae) of Turkey. Известия Вузов (İzvestiya VUZOV). No (Vıp): Vol.4, pp 74-81.

Diamond, P. (1987). Least squares fitting of several fuzzy variables, Proc. of the Second IFSA Congress, Tokyo; July 20–25. No.1, pp 339-332

Diamond, P.(1988). Fuzzy least squares, Information Science, No. **46**, pp 141-157.

Ghoshray, S. (1997). Fuzzy linear regression analysis by symmetric triangular. IEEE Int Conf Intell, pp 307 – 313.

Hojati, M., Bector, C.R., Smimou, K.A. (2005). Simple method for computation of fuzzy linear regression. European Journal of Operational Research, Vol. **166**, pp 172-184.

Hong, D.H., Song, J.K. and Young, H. (2001). Fuzzy least-squares linear regression analysis using shape preserving operations. Information Sciences, Vol, 138, pp, 185-193

Kim, B & Bishu, R.R.(1998). Evaluation of fuzzy linear regression models by comparison membership functions. Fuzzy Sets Syst. Vol.100, pp 343–352.

Lin, N.P, Chen, J.C, Chueh, H.E, Hao, W.H, Chang, C.I. (2007). A fuzzy statistics based method for mining fuzzy correlation rules. WSEAS Transactions on Mathematic, Vol.6, No.11, pp 852-858.

Ming, M., Friedman, M., Kandel, A. (1997). General fuzzy least squares, Fuzzy Sets and Systems, Vol, 88, pp 107- 118.

Moskowitz, H. & Kim, K. (1993). On assesing the H value in fuzzy linear regression, Fuzzy Sets and Systems, Vol.**58**, pp 303-27.

Nasrabadi, M.M. & Nasrabadi, E.A. (2004). Mathematical-programming approach to fuzzy linear regression analysis, Applied Mathematics and Computation, Vol.3, No.155, pp 873-881.

Nguyen, H.T. & Wakler, E.A. (2000). First course in fuzzy logic. 2nd edition. Chapman & Hall/CRC. Boca Raton, FL, 1 January, NewYork, pp.359.

Ross, T.J. (2004). Fuzzy logic with engineering appli-cations, John Willey and Sons Inc, Fuzyy Sets, Wiley- New York..

Sakawa, M. (1993). Fuzzy sets and ınteractive multi-objective optimization. Plenum Pres, New York, pp 305.

Şentürk, S. & Aşan, Z. (2007). Correlation coefficient in fuzzy logic; an application in meterological events. Eskisehir Osmangazi University Journal Of Engineering And Architecture,Vol.1, No.20, pp 149-158.

Tanaka, H. (1997). Fuzzy data analysis by possibilistic linear models. Fuzzy Sets and Systems, Vol.3, No.24, pp 363-375.

Tanaka, H. & Guo, P. (1999). Possibilistic data analysis for operations research. Physica-Verlag Heidelberg, New York.

Tanaka, H., Uejima, S., Asai, K. (1982). *Linear regression analysis with fuzzy model*, IEEE Transactions on Systems, Man, and Cybernetics, Vol.**12**, pp 903 – 907.

Tanaka, H. & Watada, J. (1988). Possibilistic linear systems and their application to the linear regression model, Fuzzy Sets and Systems. Vol.3 No.27, pp 275-289.

Tansu, A. (2012). Fuzzy linear regression: fuzzy regression. Lambert Academic Publishing, Springer, Berlin- Germany.

Topuz, D.(2018). Süt sığırcılığında Bulanık regresyon modellerinin kullanımı. Doktora Tezi, Selçuk Üniversitesi Fen Bilimleri Enstitüs, pp.196. Konya.

Topuz, D.(2020). Clinical Data Obtained For Prediction Of The Weight Of The Newborn Analysis With Classical And Fuzzy Linear Regression Models. Turkiye Klinikleri J Biostat, Vol.3, No.12, pp 320-34.

Topuz, D. & Keskin,İ.(2021). An Application of Fuzzy Pearson Correlation Methods in Animal Sciences. Selcuk Journal of Agriculture and Food Sciences, Vol.3, No.35, pp 265-271

Trillas, E. & Eciolaza, L. (2015). Fuzzy logic studies in fuzziness and soft computing. Springer. ISBN 978-3-319-14203-6.

Vrusias, L. B. (2009). Fuzzy Logic. Artificial Intelligence Lectures Notes, In, Eds, London, pp 140

Wang, H.F. & Tsaur, R.C. (2000). Insight of a fuzzy regression model, Fuzzy Sets and Systems, Vol.**112**, pp 355-69.

Xie, M.C. & Wu, B. (2012). The relationship between high schools students time management and academic performance: an application of fuzzy correlation. Educational Policy Forum, Vol. 1, No.15, pp 157–176.

Yang, C.C. (2016). Correlation coefficient evaluation for the fuzzy interval data. Journal of Business Research, Vol.6, No. 69, pp 2138–2144.

Yongshen, N. & Cheung, J.Y. (2003). Correlation coefficient estimate for fuzzy data. In Intelligent Systems Design And Applications, Publ. No:23,Berlin- Germany

Yongshen, N. (2005). Fuzzy correlation and regression analysis. University of Oklahoma, Graduate College; UMI number: 3163014.

Yu, C. (1993). Correlation of fuzzy numbers. Fuzzy sets and Systems, Vol.3, No.55, pp 303-307.

Zadeh, L.A. (1965). Fuzzy sets. Information and Con-trol, Vol.3, No.8, pp 338-353.

Zadeh, L.A. (1978). Fuzzy sets as a basis for a theory of possibility. Fuzzy Sets and Systems, Vol.1, No.1, pp 3–28.

Zhuang S, Shao H., Guo F. et al. (2009). Sns and kirre, the drosophila orthologs of nephrin and neph1, direct adhesion, fusion and formation of a slit diaphragm-like structure in insect nephrocytes. Development, Vol.14, No.136, pp 2335--2344.

Zimmermann, H. J. (1996). Fuzzy set theory and its applications, springer science+business media. 3rd Edition, Kluwer-Nijhoff, Boston, New York, pp. 203-240.

Annex 1

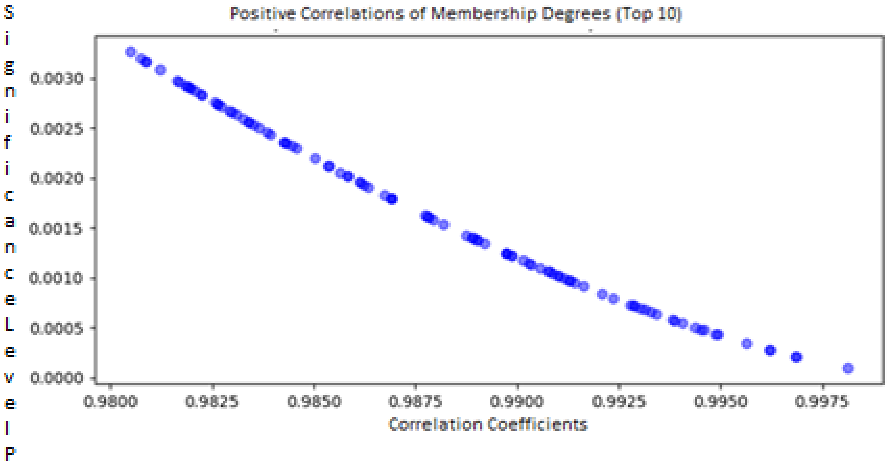
The Correlation of Membership Degrees (Total 20 in Negative and Positive Directions) and Python Codes Used in Calculating Significance Levels

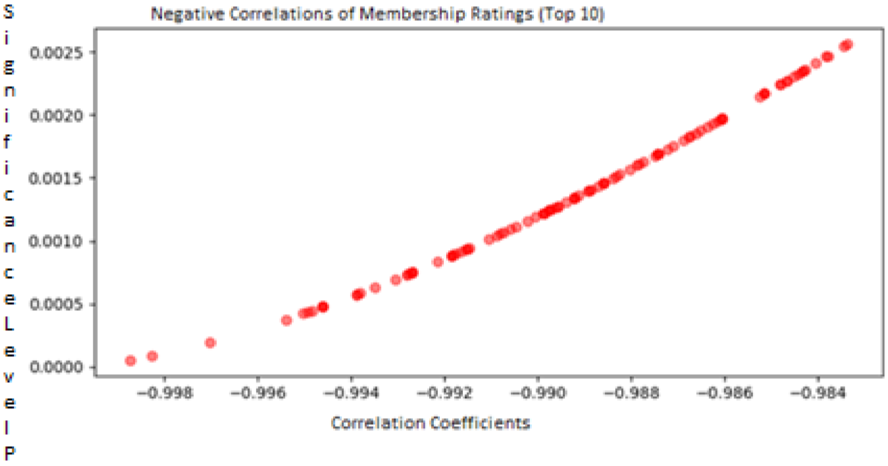
```
import pandas as pd
import numpy as np
import scipy.stats as stat
import matplotlib.pyplot as plt
import random
import itertools

w = [0.0,0.1,0.2,0.3,0.4,0.5,0.6,0.7,0.8,0.9,1]

def uyelikHesapla(w,n):
    sonuc = list(itertools.product(w,repeat=2))
    weights = [ random.sample(sonuc,n) for i in range(100000)]
    xr_ = []
    yp_ = []
    uyelik = []
    for i in range(100000):
        x = []
        y = []
        count = 0
        for j in weights[i]:
            x.append(j[0])
            y.append(j[1])
            count+=1
            if count == 5:
                cikti = stat.pearsonr(x,y)
                r = cikti[0]
                p = cikti[1]
                if (p<0.05):# and (r>0.7 or r<-0.7)):
                    xr_.append(r)
                    yp_.append(p)
                    uyelik.append(weights[i])
    data = pd.DataFrame({"r":xr_,
                         "p":yp_,
                         "uyelik":uyelik})
    return data

df = uyelikHesapla(w,5)
df = df.sort_values("r",ascending=False)
pozitif_korelasyon = df[:20]
negatif_korelasyon = df[-20:]



fig,ax = plt.subplots(nrows = 2, figsize =(8, 8))
```

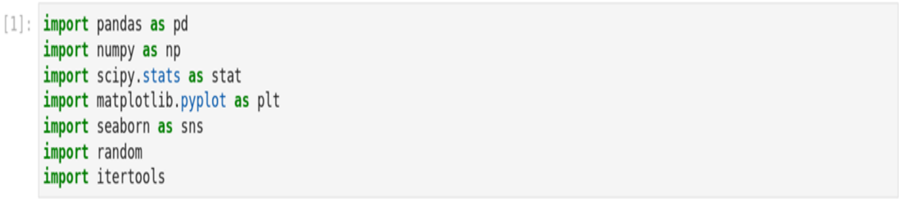
```
    ax[0].scatter(pozitif_korelasyon.r,pozitif_korelasyon.p,color="blue",alpha=0.5
    ax[0].set_title('Üyelik Derecelerinin Pozitif Korelasyonları(İlk 20)')
    ax[0].set_xlabel("Korelasyon")
    ax[0].set_ylabel("Anlamlılık(p)")

    ax[1].scatter(negatif_korelasyon.r,negatif_korelasyon.p,color="red",alpha=0.5
    ax[1].set_title('Üyelik Derecelerinin Negatif Korelasyonları(İlk 20)')
    ax[1].set_xlabel("Korelasyon")
    ax[1].set_ylabel("Anlamlılık(p)")

plt.tight_layout()
```

Correlation Coefficient(r)	p	Membership degree
0.998113214994492	9.83545669492308E-05	[(1, 0.6), (0.0, 0.0), (0.2, 0.1), (0.9, 0.5), (0.7, 0.4)]
0.996870607401959	0.000210048685155	[(1, 1), (0.6, 0.4), (0.4, 0.1), (0.7, 0.6), (0.5, 0.3)]
0.996831449970143	0.000214002176593	[(0.1, 0.1), (0.6, 0.5), (0.8, 0.6), (0.5, 0.4), (1, 0.8)]
0.996231324468294	0.000277569229603	[(0.4, 0.6), (0.1, 0.3), (0.9, 1), (0.6, 0.8), (0.0, 0.2)]
0.996182063332873	0.00028302712828	[(0.7, 0.5), (1, 0.7), (0.4, 0.2), (0.1, 0.0), (0.2, 0.1)]
0.995629298081616	0.000346637632588	[(0.8, 0.9), (0.8, 0.8), (0.2, 0.0), (0.9, 1), (0.3, 0.2)]
0.994936676326182	0.000432172643508	[(0.7, 0.7), (1, 0.8), (0.1, 0.4), (0.3, 0.5), (0.9, 0.8)]
0.994874389621496	0.000440167585905	[(0.1, 0.3), (0.8, 0.9), (0.4, 0.5), (0.9, 0.9), (0.0, 0.2)]
0.994591241480368	0.000477119957689	[(0.9, 1), (0.6, 0.7), (0.2, 0.3), (0.1, 0.1), (0.0, 0.1)]
0.994527278179062	0.00048560380465	[(0.6, 0.6), (0.8, 0.8), (1, 0.9), (0.0, 0.0), (0.7, 0.7)]
0.994376712684369	0.000505769463802	[(0.2, 0.3), (0.3, 0.4), (0.6, 0.7), (0.1, 0.1), (0.9, 1)]
0.994070141267795	0.000547663384115	[(0.0, 0.0), (0.2, 0.1), (1, 0.9), (0.1, 0.1), (1, 1)]
0.993883734673619	0.000573673109332	[(0.0, 0.1), (0.4, 0.8), (0.2, 0.4), (0.3, 0.6), (0.1, 0.2)]
0.99381202330337	0.000583785549527	[(0.3, 0.2), (0.2, 0.2), (1, 0.8), (0.6, 0.5), (0.9, 0.7)]
0.993418364102507	0.000640332311969	[(0.1, 0.2), (0.4, 0.4), (0.7, 0.7), (0.3, 0.3), (0.9, 0.9)]
0.993292582595468	0.000658763345145	[(0.8, 0.9), (0.7, 0.8), (0.3, 0.4), (0.3, 0.3), (0.9, 1)]
0.993134144692148	0.000682225662644	[(1, 0.9), (0.0, 0.1), (0.0, 0.2), (0.7, 0.7), (0.9, 0.9)]
0.993043784063013	0.000695728452038	[(0.5, 0.6), (1, 1), (0.7, 0.7), (0.4, 0.5), (0.9, 0.9)]
0.99289460944547	0.000718211529985	[(0.5, 0.9), (0.2, 0.4), (0.0, 0.0), (0.2, 0.3), (0.3, 0.5)]
0.992863372595271	0.000722949458425	[(0.1, 0.0), (0.0, 0.0), (0.8, 1), (0.9, 1), (0.1, 0.1)]
-0.991853721521983	0.000881538616383	[(0.3, 0.9), (0.6, 0.5), (0.3, 1), (1, 0.0), (0.5, 0.6)]
-0.992156741649222	0.000832850471825	[(0.0, 0.9), (0.0, 1), (0.4, 0.2), (0.2, 0.6), (0.1, 0.8)]
-0.992673894012538	0.000751909938	[(0.0, 0.9), (0.3, 0.6), (0.9, 0.0), (0.5, 0.3), (0.8, 0.1)]
-0.992690446228782	0.000749365004616	[(0.8, 0.4), (0.6, 0.6), (0.2, 0.9), (0.0, 1), (0.1, 1)]
-0.992715672337676	0.000745491964164	[(0.9, 0.2), (0.6, 0.5), (0.1, 0.9), (0.3, 0.7), (0.0, 0.9)]
-0.992773966542046	0.000736567438133	[(0.2, 0.7), (0.9, 0.1), (0.5, 0.5), (0.0, 0.9), (0.3, 0.7)
-0.992818197314672	0.000729819823281	[(0.7, 0.2), (0.7, 0.3), (0.1, 0.8), (0.0, 0.9), (0.3, 0.6)]
-0.993042049238722	0.000695988550355	[(0.4, 0.6), (0.7, 0.2), (0.8, 0.0), (0.3, 0.7), (0.3, 0.8)]
-0.993489737200892	0.000629951446504	[(1, 0.1), (0.1, 1), (0.5, 0.6), (0.2, 0.8), (0.7, 0.4)]
-0.99381202330337	0.000583785549527	[(0.4, 0.6), (0.7, 0.4), (0.8, 0.3), (0.1, 0.9), (1, 0.2)]
-0.993877743361828	0.00057451572807	[(0.8, 0.6), (0.9, 0.6), (0.3, 0.9), (0.1, 1), (0.5, 0.8)]
-0.993892435273715	0.000572450188453	[(0.8, 0.2), (0.9, 0.0), (0.4, 0.9), (0.7, 0.3), (0.4, 0.8)]

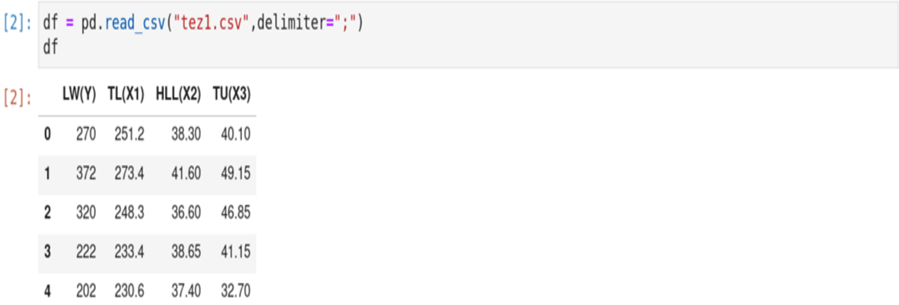
-0.994606403100465	**0.000475116280723**	**[(0.5, 0.0), (0.2, 0.5), (0.0, 0.8), (0.4, 0.1), (0.1, 0.7)]**
-0.994606403100465	**0.000475116280723**	**[(0.0, 1), (0.4, 0.6), (0.9, 0.0), (0.1, 1), (0.2, 0.8)]**
-0.994606799205504	**0.000475063971246**	**[(1, 0.0), (0.0, 0.8), (0.4, 0.5), (0.5, 0.4), (0.3, 0.5)]**
-0.994835057326806	**0.000445241219895**	**[(0.2, 0.8), (0.1, 0.9), (0.1, 1), (0.6, 0.0), (0.3, 0.6)]**
-0.994920484287555	**0.000434246315472**	**[(0.5, 0.7), (0.3, 1), (0.9, 0.1), (0.7, 0.5), (1, 0.0)]**
-0.99501623938769	**0.000422031326079**	**[(1, 0.0), (0.9, 0.1), (0.2, 0.7), (0.6, 0.4), (0.5, 0.5)]**
-0.995393803240412	**0.000375013836227**	**[(0.7, 0.2), (0.9, 0.0), (0.1, 1), (0.1, 0.9), (0.5, 0.5)]**
-0.997029348806553	**0.0001942753651**	**[(0.7, 0.3), (0.8, 0.2), (0.9, 0.1), (0.6, 0.5), (0.2, 1)]**
-0.998276073093128	**8.59013114063105E-05**	**[(0.9, 0.1), (0.1, 0.8), (1, 0.0), (0.6, 0.4), (0.8, 0.2)]**
-0.998735019100828	**5.39979641133576E-05**	**[(0.3, 0.9), (0.9, 0.1), (0.7, 0.4), (0.6, 0.5), (0.2, 1)]**





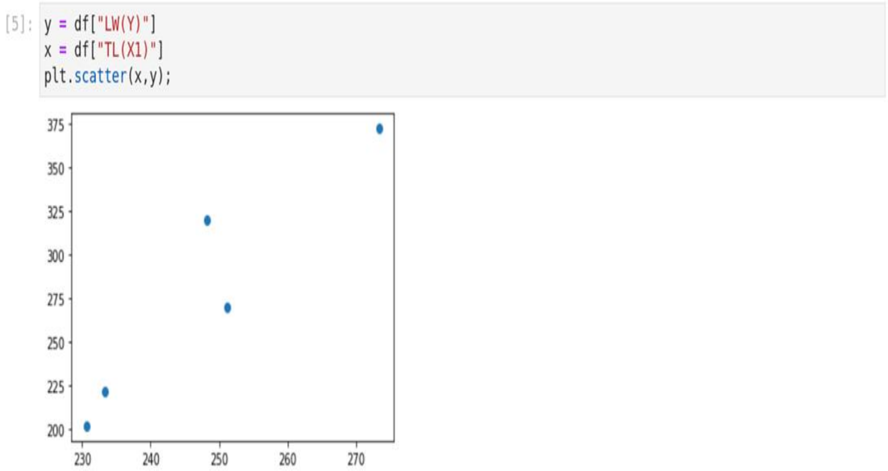


```
[1]: import pandas as pd
     import numpy as np
     import scipy.stats as stat
     import matplotlib.pyplot as plt
     import seaborn as sns
     import random
     import itertools
```

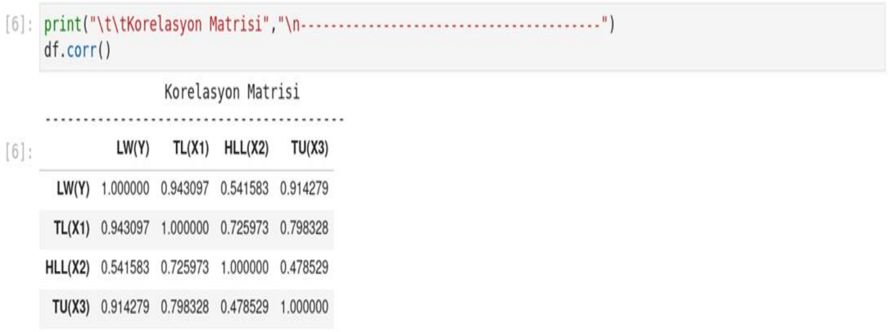


```
[2]: df = pd.read_csv("tez1.csv",delimiter=";")
     df
```

```
[2]:    LW(Y)  TL(X1)  HLL(X2)  TU(X3)
     0    270   251.2    38.30   40.10
     1    372   273.4    41.60   49.15
     2    320   248.3    36.60   46.85
     3    222   233.4    38.65   41.15
     4    202   230.6    37.40   32.70
```

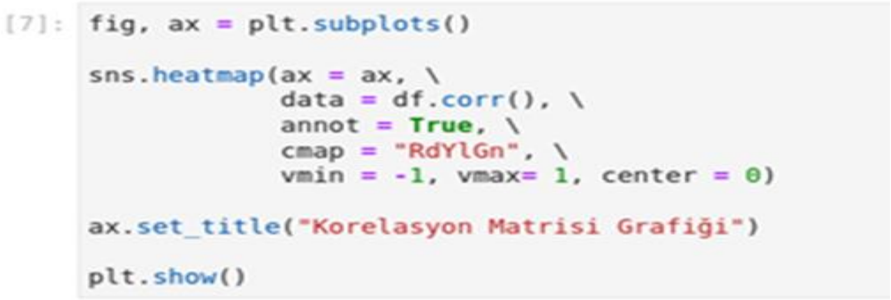


```
[5]: y = df["LW(Y)"]
     x = df["TL(X1)"]
     plt.scatter(x,y);
```

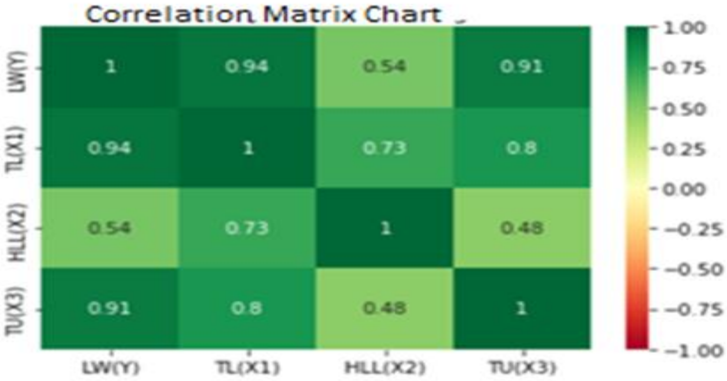


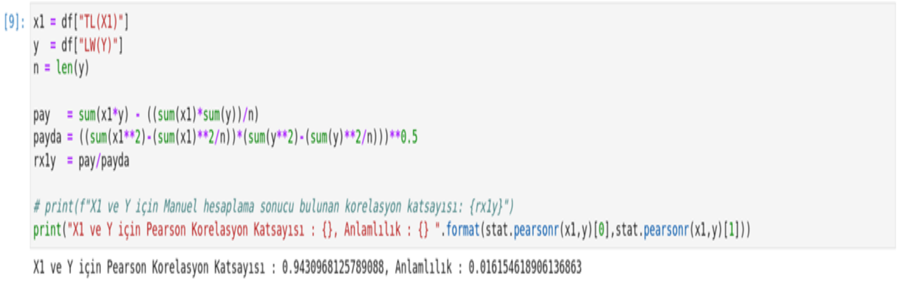
```
[6]: print("\t\tKorelasyon Matrisi","\n--------------------------------------")
     df.corr()
```

```
                   Korelasyon Matrisi
     ---------------------------------------
[6]:             LW(Y)    TL(X1)   HLL(X2)    TU(X3)
      LW(Y)   1.000000  0.943097  0.541583  0.914279
     TL(X1)   0.943097  1.000000  0.725973  0.798328
     HLL(X2)  0.541583  0.725973  1.000000  0.478529
      TU(X3)  0.914279  0.798328  0.478529  1.000000
```

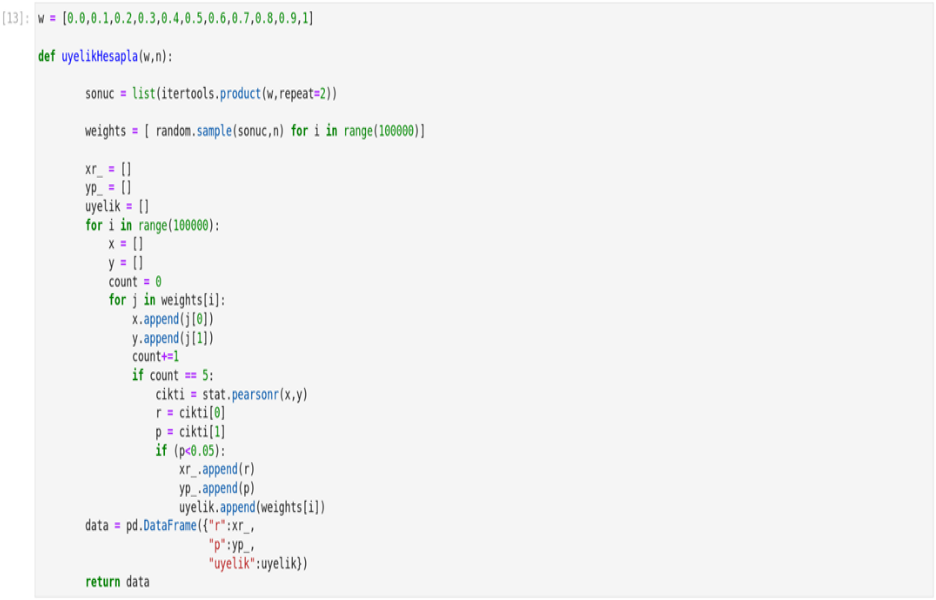


```
[7]: fig, ax = plt.subplots()

     sns.heatmap(ax = ax, \
                 data = df.corr(), \
                 annot = True, \
                 cmap = "RdYlGn", \
                 vmin = -1, vmax= 1, center = 0)

     ax.set_title("Korelasyon Matrisi Grafiği")

     plt.show()
```

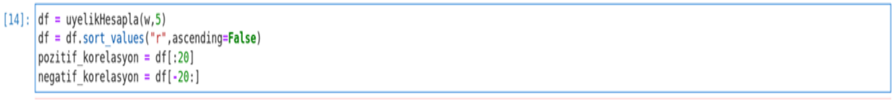




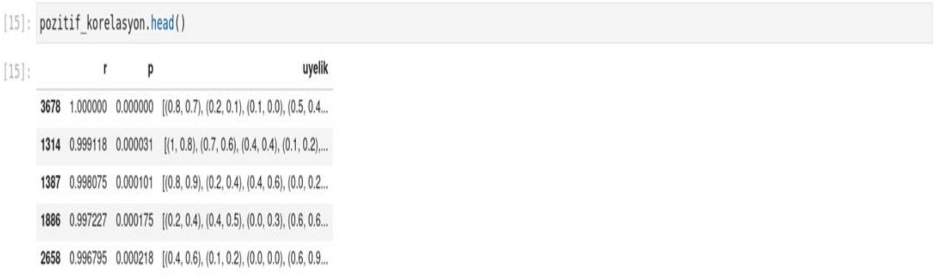
```
[9]: x1 = df["TL(X1)"]
     y  = df["LW(Y)"]
     n = len(y)

     pay   = sum(x1*y) - ((sum(x1)*sum(y))/n)
     payda = ((sum(x1**2)-(sum(x1)**2/n))*(sum(y**2)-(sum(y)**2/n)))**0.5
     rx1y  = pay/payda

     # print(f"X1 ve Y için Manuel hesaplama sonucu bulunan korelasyon katsayısı: {rx1y}")
     print("X1 ve Y için Pearson Korelasyon Katsayısı : {}, Anlamlılık : {} ".format(stat.pearsonr(x1,y)[0],stat.pearsonr(x1,y)[1]))

     X1 ve Y için Pearson Korelasyon Katsayısı : 0.9430968125789088, Anlamlılık : 0.016154618906136863
```



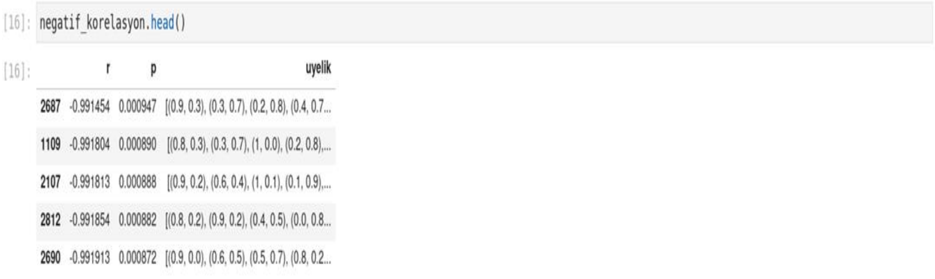
```
[13]: w = [0.0,0.1,0.2,0.3,0.4,0.5,0.6,0.7,0.8,0.9,1]

      def uyelikHesapla(w,n):

          sonuc = list(itertools.product(w,repeat=2))

          weights = [ random.sample(sonuc,n) for i in range(100000)]

          xr_ = []
          yp_ = []
          uyelik = []
          for i in range(100000):
              x = []
              y = []
              count = 0
              for j in weights[i]:
                  x.append(j[0])
                  y.append(j[1])
                  count+=1
                  if count == 5:
                      cikti = stat.pearsonr(x,y)
                      r = cikti[0]
                      p = cikti[1]
                      if (p<0.05):
                          xr_.append(r)
                          yp_.append(p)
                          uyelik.append(weights[i])
          data = pd.DataFrame({"r":xr_,
                               "p":yp_,
                               "uyelik":uyelik})
          return data
```



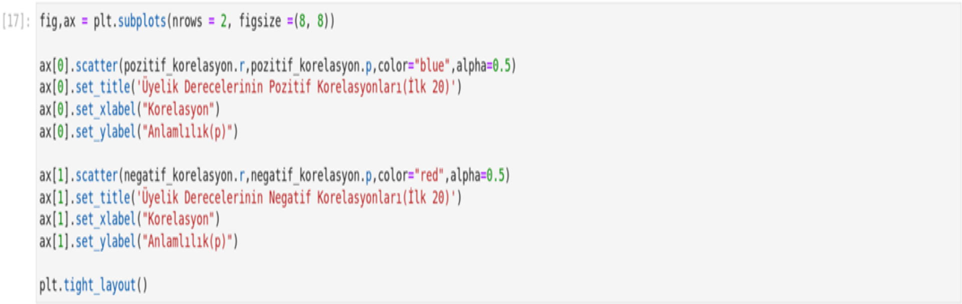
```
[14]: df = uyelikHesapla(w,5)
      df = df.sort_values("r",ascending=False)
      pozitif_korelasyon = df[:20]
      negatif_korelasyon = df[-20:]
```



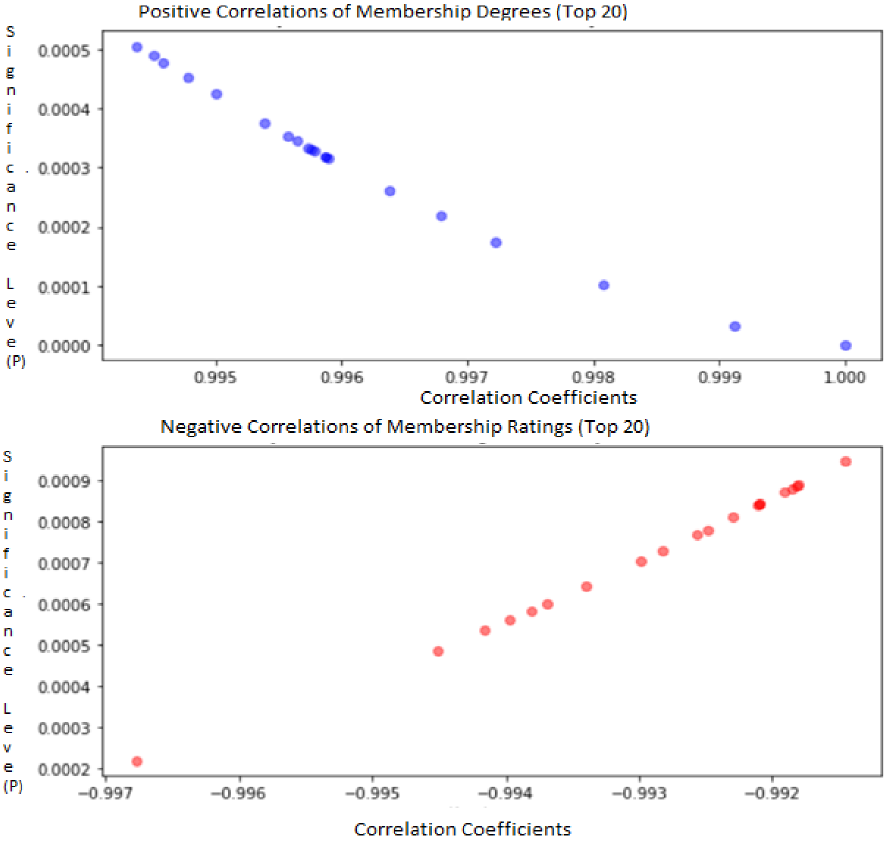
```
[15]: pozitif_korelasyon.head()

[15]:             r         p                                             uyelik
      3678  1.000000  0.000000  [(0.8, 0.7), (0.2, 0.1), (0.1, 0.0), (0.5, 0.4...
      1314  0.999118  0.000031   [(1, 0.8), (0.7, 0.6), (0.4, 0.4), (0.1, 0.2),...
      1387  0.998075  0.000101  [(0.8, 0.9), (0.2, 0.4), (0.4, 0.6), (0.0, 0.2...
      1886  0.997227  0.000175  [(0.2, 0.4), (0.4, 0.5), (0.0, 0.3), (0.6, 0.6...
      2658  0.996795  0.000218  [(0.4, 0.6), (0.1, 0.2), (0.0, 0.0), (0.6, 0.9...
```



```
[16]: negatif_korelasyon.head()

[16]:              r         p                                             uyelik
      2687 -0.991454  0.000947  [(0.9, 0.3), (0.3, 0.7), (0.2, 0.8), (0.4, 0.7...
      1109 -0.991804  0.000890   [(0.8, 0.3), (0.3, 0.7), (1, 0.0), (0.2, 0.8),...
      2107 -0.991813  0.000888   [(0.9, 0.2), (0.6, 0.4), (1, 0.1), (0.1, 0.9),...
      2812 -0.991854  0.000882  [(0.8, 0.2), (0.9, 0.2), (0.4, 0.5), (0.0, 0.8...
      2690 -0.991913  0.000872  [(0.9, 0.0), (0.6, 0.5), (0.5, 0.7), (0.8, 0.2...
```



```
[17]: fig,ax = plt.subplots(nrows = 2, figsize =(8, 8))

      ax[0].scatter(pozitif_korelasyon.r,pozitif_korelasyon.p,color="blue",alpha=0.5)
      ax[0].set_title('Üyelik Derecelerinin Pozitif Korelasyonları(İlk 20)')
      ax[0].set_xlabel("Korelasyon")
      ax[0].set_ylabel("Anlamlılık(p)")

      ax[1].scatter(negatif_korelasyon.r,negatif_korelasyon.p,color="red",alpha=0.5)
      ax[1].set_title('Üyelik Derecelerinin Negatif Korelasyonları(İlk 20)')
      ax[1].set_xlabel("Korelasyon")
      ax[1].set_ylabel("Anlamlılık(p)")

      plt.tight_layout()
```



CHAPTER VII

ANALYSIS OF CONTINUOUS PROPORTIONAL DATA: CASE STUDY OF DETERMINATION OF MILK PROTEIN

Dr. Cem TIRINK[*]

[*]Igdir University, Faculty of Agriculture, Department of Animal Science, Biometry and Genetic Unit, TR76000, Igdir, Turkey. Email: cem.tirink@igdir.edu.tr

INTRODUCTION

Milk and dairy products are one of the most important foods of animal origin in the diet of humans. According to FAO data, cattles constitute 90% of the total milk production in the world (FAO, 2019). The quality of milk produced for economic purposes also affects its price. Milk protein are very important variables that determine the qualify of the milk. In addition, milk protein is one of the most important factors affecting milk processing. When evaluated on the dry matter of milk, the share of milk protein in milk content was approximately 25% (Ozek, 2015). Coagulation properties of milk is important and essential factor for cheese production industries (Duchemin et al., 2020). Coagulation properties of milk, which has an important place in milk processing, are affected by many factors such as SCC, milk protein composition and lactation stage (Tyrisevä et al., 2004; Cassandro et al., 2008; Duchemin et al., 2020). From this point of view, there are many factors that affect the composition of milk which is depending on both genetic structure, and various factors such as feeding, season, number of days in lactation, and diseases (Ozek, 2015).

In this context, it is also very important to determine the factors affecting the protein content of milk, which are the two most important quality criteria of milk. In determining these factors, it is important to examine the relationships between the variables used in determining the milk quality criteria within the scope of cause-and-effect relationship.
Determining the relationships between variables is an important issue in statistics. Regression analysis is the most widely used method in

determining of these relationships between variables. Regression analysis, which is used in many fields, is a statistical analysis method that examines the relationship between two or more variables with a cause-and-effect relationship and is used to model this relationship (Ari and Onder, 2013). Incorrect and misleading results are obtained as a result of using inappropriate regression methods in determining these relationships. One another aim of the regression analysis is to find a suitable coefficient for the prediction model. The selection of best model within the scope of regression analysis is evaluated depending on the success of the model selection criteria. In this context, the results of the model selection vary according to the regression method applied and the information criterion based on the model selection (Dunder, 2017). The regression analysis method to be applied varies according to the structure and distribution for response variable of the data set i.e., percentage data.

Percentage data are the case of taking values in the standard unit range (0,1) such as ratios, percentages, proportions and fractions (Dunder et al., 2015). Analysis of percentage data emerges as a general problem for research to be conducted with quantitative data (Schmid et al., 2013). Many studies have been conducted in recent years on statistical modeling for continuous percentage data. Commonly used regression models such as linear or nonlinear regression models are not suitable for percentage continuous variables (Espinheira et al., 2008; Ospina and Ferrari, 2012). Different methods have been proposed for modeling continuous percentage variable that are thought to be related to other

variables. One of the proposed methods is based on the beta distribution by Ferrari and Cribari-Neto (2004), which is useful for modeling continuous variables that take values (0, 1) (Espinheira et al., 2008).
The main purpose of this study is to evaluate the best model for estimating milk protein, using somatic cell count (logarithmic), which is a milk quality trait, and days in milk, a non-nutritional factor. The main point here is to use beta regression, which is recommended to be applied in situations where limited data are available.

MATERIAL AND METHODS

In order to determine the factors affecting milk protein, which are the two most important quality criteria of milk, milk yield records from a private farm in Bafra district of Samsun were used. For this purpose, 150 milk yield records were used. For this purpose, somatic cell count (logarithmic values), number of days in milk were used to explain protein ratios.

Methods of analyzing continuous and proportional data are common in biological studies. Since the normal distribution allows values in the range of $-\infty$ to $+\infty$ and has a constant variance assumption, the predictions to be made as a result of modeling the ratios with models based on the normal distribution may not be reliable. Therefore, various transformations can be applied to the data to provide statistically reliable models (Douma and Weedon, 2019). Proportional data can be derived from a variety of different basic data types. The methods used in the analysis of this type of data are important. Therefore, the statistical methods that can be used in the analysis of proportional data are presented in Figure 1. According to Figure 1, the analysis of proportional data should focus on the case of discrete or continuous as well as the number of categories (Douma and Weedon, 2019). In modeling continuous proportional data, in addition to the use of various transformations, the least squares method without transformation can be used (Sokahl and Rohlf, 1995; Kieschnick and McCullough, 2003; Warton and Hui, 2011; Douma and Weedon, 2019). However, it is seen that there is swelling in the variances of the proportional continuous data analyzed by transformation, and therefore the bias increases; therefore, it is recommended to model the observations as original as possible in proportional continuous data (Douma and Weedon, 2019).

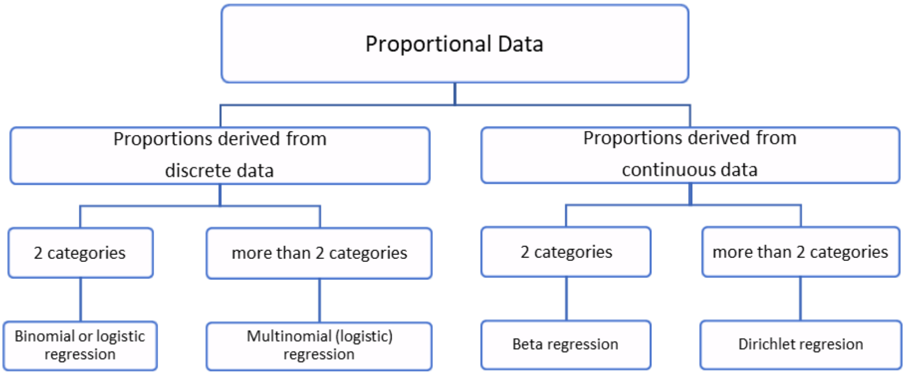

Proportional Data
Proportions derived from discrete data
Proportions derived from continuous data
2 categories
more than 2 categories
2 categories
more than 2 categories
Binomial or logistic regression
Multinomial (logistic) regression
Beta regression
Dirichlet regresion

Figure 1. Flowchart for statistical analysis that should be used in proportional data (Douma and Weedon, 2019)

Commonly used regression models such as linear or nonlinear regression models are not suitable for percentage continuous variables (Espinheira et al., 2008; Ospina and Ferrari, 2012). Different methods have been proposed for modeling continuous percentage variable that are thought to be related to other variables. One of the proposed methods is based on the beta distribution by Ferrari and Cribari-Neto (2004), which is useful for modeling continuous variables that take values (0, 1) (Espinheira et al., 2008).

The beta distribution with two shape parameters such as α_1 and α_2, has the density function

$$f(v;\alpha_1,\alpha_2)=\frac{\Gamma(\alpha_1+\alpha_2)}{\Gamma(\alpha_1)\Gamma(\alpha_2)}v^{\alpha_1-1}(1-v)^{\alpha_2-1}, \qquad v\in[0,1]$$

where $\Gamma(\cdot)$ is the gamma function. The parameterization utilized in the density function is not the usual one, but it is convenient for continuous proportional data modeling purposes (Ospina and Ferrari, 2012; Zhou et al., 2020). α_1 and α_2 can be calculated by using the distribution mean (μ) and precision parameter (ϕ) with as $\alpha_1=\mu\phi$ and $\alpha_2=(1-\mu)\phi$ equations. When α_1 and α_2 are greater than 1, there is a unique mode for beta distribution provided by θ that can be calculated by $\theta=(\alpha_1-1)/(\alpha_1+\alpha_2-2)$. Including the mode directly in the parameterization makes it more convenient to make inferences for the mode. Therefore, $\alpha_1=1+m\theta$ and $\alpha_2=1+m(1-\theta)$ for m>0 (Zhou et al., 2020).

The density function of generalized biparabolic distribution (GBP) is given as below.

$$f_{GBP}(v;\theta,m) = \frac{(2m+1)(m+1)}{(3m+1)} d^m(2-d^m)$$

In the density function, d can be calculated by $d = I(0 < v \leq \theta)v/\theta + I(\theta < v \leq 1)(1-v)/(1-\theta)$, $I(\cdot)$ is an indicator function for the d and m is a positive shape parameter (Garcia et al., 2009; Zhou et al., 2020).

In beta regression, the parameterization distribution is modeled as beta with the help of mean and precision parameters. Here, the mean linked with the response, as in generalized linear models (GLMs), by means of a link function and a linear predictor. Besides, the precision parameter can be linked to another set of regressors via a second link function, which then results in a variable dispersion model. Many link functions such as probit, log, cauchit, log-log can be used for this process. Estimation is performed with maximum likelihood via link one of these link functions.

Log-likelihood function for beta mode and GBP mode model are given below, respectively.

$$\begin{aligned}\ell_{beta}(\Omega,\mathcal{D}) = {} & n log\Gamma(2+m) \\ & -\sum_{i=1}^{n} log(\Gamma\{1+m\theta(X_i)\}\Gamma[1+m\{1-\theta(X_i)\}]) \\ & + m\sum_{i=1}^{n}[\theta(X_i)\log Y_i + \{1-\theta(X_i)\}\log(1-Y_i)]\end{aligned}$$

$$\ell_{GBP}(\Omega, \mathcal{D}) = nlog\left\{\frac{(2m+1)(m+1)}{3m+1}\right\} + m\sum_{i=1}^{n}\log d_i + \sum_{i=1}^{n}\log(2-d_i^m)$$

where $\ell_{beta}(\Omega, \mathcal{D})$ regarding $\Omega = (\beta^T, m)^T$ produces the maximum likelihood estimation for Ω under this model and β can be calculated by $(\beta_0, \beta_1^T)^T$. The probability function corresponding to the beta distribution is concave, which can easily provide the regularity conditions for beta regression required for the maximum likelihood estimation (MLE) to be consistent and asymptotically normal (Zhou et al., 2020). Unlike the beta distribution, GBP is not an exponential family, so it can also be shown that the log-probability of GBP is concave in a neighborhood of reality. Also, no additional conditions are needed to establish asymptotic normality for MLE in GBP (Cox and Hinkley, 1979; Zhou et al., 2020).

As a result of basing the statistical inference on a certain parametric model, misleading model estimation can be made. To overcome this situation, Zhou et al. (2020), correct interpretation can be made by using the diagnosis methods. For this aim, the first method is graphical interpretation methods were applied. Semi-normal residual plots are useful graphical tools for checking the goodness of fit of the model in situations where response distributions are complex (Zhou et al., 2020). A second method that provides a reliable interpretation of the model is the score test method. The score test evaluates the model that makes better predictions by using the score functions created to quantitatively evaluate the adequacy of a predicted regression model.

All statistical analysis were performed by using R software (R Core Team, 2020). Descriptive statistics of the quantitative characteristics were estimated by using “psych” package in R package (Revelle, 2020). For performing all statistical analysis for the present study, “betareg” and “maxLik” package accessible in R environment were used with the scope of predicting milk protein content (Cribari-Neto and Zeileis, 2010; Henningsen and Toomet, 2011).

RESULTS

Milk protein values formed by values between 0 and 1 and so, different methods such as beta mean, beta mode and GBP mode were used due to the compatibility of the data structure with beta distribution. For this purpose, somatic cell count (logarithmic), which is a milk quality feature, and number of days in milk, which is a non-nutritive factor, were used to estimate milk protein.

Descriptive statistics of the data used in the present study are given in Table 1. Histogram plot is given in Figure 1 to show the structure of the distribution of milk protein. In Figure 1, it is seen that milk protein yields show a slightly right skewed distribution.

Table 1. Descriptive statistics

Variables	Mean	Standard deviation	Median
Milk Protein	0.03	0.00	0.03
Days in Milk (DIM)	195.39	0.64	185.50
SCC (LG10)	5.07	104.69	4.92

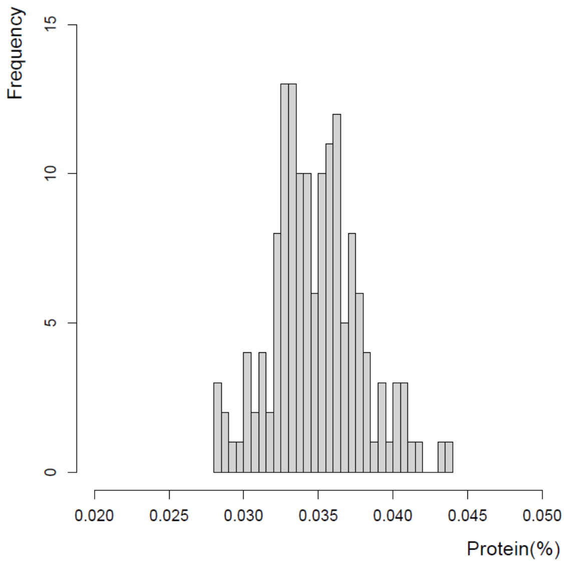


Figure 1. Histogram of milk protein (%)

In Table 2, besides the unknown parameter values for the beta mean regression model, the standard error values of the parameters and test statistics are presented. According to the results shown for the beta mean model in Table 2, the contribution of other explanatory variables to the model, except somatic cell count (LG10), was found to be significant.

Table 2. Coefficients of beta mean model

	Estimate	Std. Error	z value	Pr(>|z|)	
β_0 (Intercept)	-1.257e+00	1.343e-02	-93.607	<2e-16	***
β_2 (LG10)	3.940e-03	2.684e-03	1.468	0.142	
β_3 (DIM)	1.347e-04	1.638e-05	8.223	<2e-16	***
ϕ	5793.7	669.2	8.658	<2e-16	***

The estimation results made according to the beta mode model are given in Table 3. According to the results in Table 3, the explanatory variables that were important for the beta mean model were also important in the beta mode model.

Table 3. Coefficients of beta mode model

	Estimate	Std. Error	t value	Pr(>|z|)	
β_0 (Intercept)	-1.2586496	0.0125321	-100.434	<2e-16	***
β_2 (LG10)	0.0039513	0.0025225	1.566	0.117	
β_3 (DIM)	0.0001352	0.0000164	8.244	<2e-16	***
$\log m$	8.6655858	0.0347977	249.028	<2e-16	***

In the GBP mode model, in which LG10 and DIM explanatory variables were used to predict milk protein, all explanatory variables were found to be statistically significant (Table 4).

Table 4. Coefficiencts of GBP mode model

	Estimate	Std. Error	t value	Pr(>|z|)	
β_0 (Intercept)	-1.250e+00	6.304e-03	-198.216	<2e-16	***
β_2 (LG10)	-6.051e-03	1.206e-03	-5.017	5.24e-07	***
β_3 (DIM)	1.353e-04	9.024e-06	14.998	<2e-16	***
$\log m$	5.416e+00	7.337e-02	73.827	<2e-16	***

The residuals graphs were given as graphical diagnosis that was used primarily for the study, in which graphical diagnosis and score test diagnosis methods were used to compare models. For this aim, residuals were given in Figure 2 and Figure 3 which were showed half-normal residual plot for beta mode and GBP mode model.

According to these figures, both beta mode and GBP mode show that the model has a good fit. As a result of the score test diagnosis, which is another criterion used in comparing models other than graphical diagnosis, the p-value for the beta mode model was 0.73, and the p-value for the GBP mode model was determined as 0.000.

In the light of this information, it is possible to interpret the beta mode model as making a more reliable prediction than the GBP mode model.

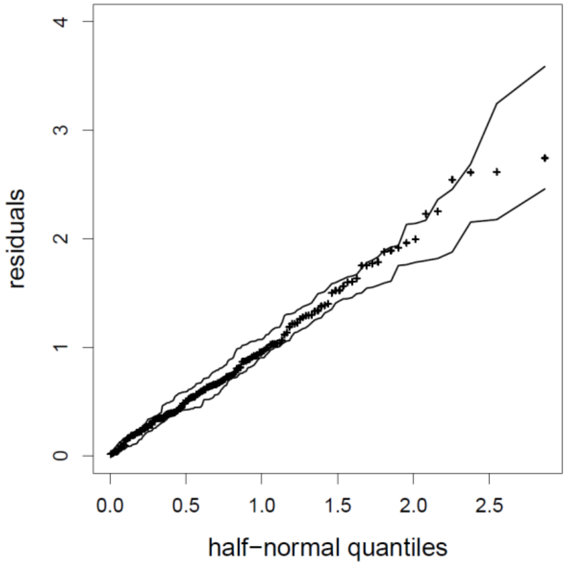


Figure 2. Half-normal residual plots for beta mode model

Figure 4 and Figure 5 show the histogram plot of residuals for both beta mean and beta mode regression. Structurally, it is seen that both histogram diagrams show similar results. Although the histograms of the errors are close to each other, the interpretation of the two methods is different.

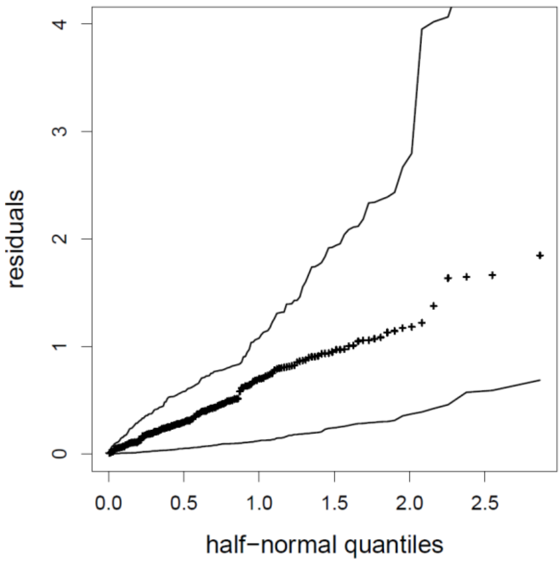


Figure 3. Half-normal residual plots for GBP model

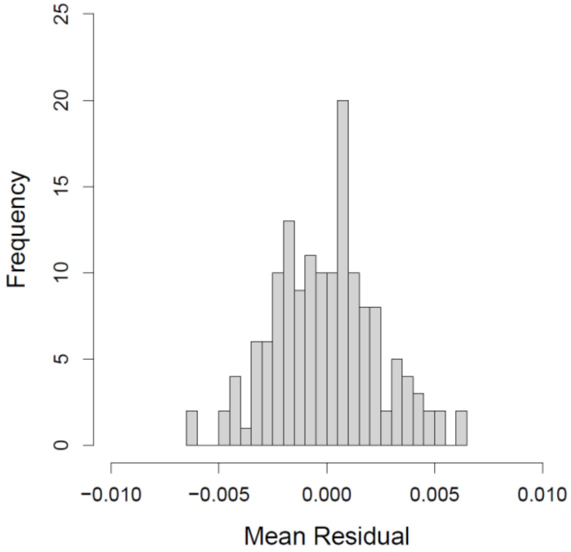


Figure 4. Histogram of residuals for beta mean regression

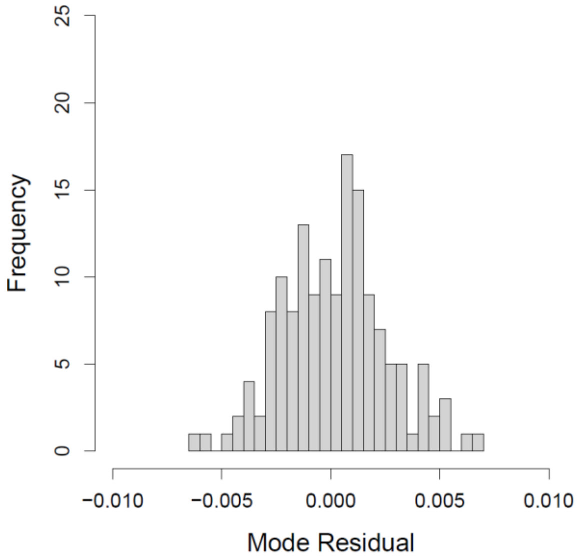


Figure 5. Histogram of residuals for beta mode regression

CONCLUSION

In the milk processing process, determination of milk protein content is an important and necessary task in the production of quality dairy products. From this point of view, it should not be forgotten that there are many factors affecting the composition of milk depending on both genetic structure and various factors such as nutrition, season, number of days in lactation and diseases. In this context, the number of studies in which the relationships between the conditions that affect the milk quality criteria and the milk quality criteria are interpreted is rare. In addition, the number of studies in which multivariate statistics were used to explain these relationships is almost non-existent. Due to being lack of published reports on the prediction of milk protein by many factors in the milk processing process, an attempt was made in the present investigation to find somatic cell count and DIM influential on the content of milk protein with the scope of bounded data process. In this context, we use the methods recommended for limited data in multivariate statistics.

In conclusion, examining the graphical and score test diagnostics, it is concluded that the beta mode model is potentially a better model than the GBP mode model.

REFERENCES

Ari, A., & Onder, H. (2013). Regression Models Used for Different Data Structures. Anadolu Journal of Agricultural Sciences. 28(3), 168-174.

Cassandro, M., Comin, A., Ojala, M., Dal Zotto, R., De Marchi, M., Gallo, L., Carnier, P., & Bittante, G., (2008). Genetic Parameters of Milk Coagulation Properties and Their Relationships with Milk Yield and Quality Traits in Italian Holstein Cows. Journal of Dairy Science 91:371–376.

Cox, D. R., & Hinkley, D. V. (1979). Theoretical statistics. Boca Raton, FL: Chapman and Hall/CRC.

Cribari-Neto, F. & Zeileis, A. (2010). Beta Regression in R. Journal of Statistical Software. Volume: 34 Issue: 2, Page: 1-24. URL: http://www.jstatsoft.org /v34/i02/.

Douma, J. C., & Weedon, J. T. (2019). Analysing continuous proportions in ecology and evolution: A practical introduction to beta and Dirichlet regression. Methods in Ecology and Evolution, 10(9), 1412-1430.

Duchemin, S.I., Nilsson, K., Fikse, W.F., Stålhammar, H., Johansen, L.B., Hansen, M.S., Lindmark-Månsson, H., de Koning, D.J., Paulsson, M., & Glantz, M. (2020). Genetic parameters for noncoagulating milk, milk coagulation properties, and detailed milk composition in Swedish Red Dairy Cattle. Journal of Dairy Science, 103(9), 8330-8342.

Dunder, E. (2017). Model Selection in Beta Regression Analysis Using Heurisic Optimization Algorithms. Doctoral Dissertation. Ondokuz Mayis University Institute of Science.

Dunder, E., Gumustekin, S., & Cengiz, M.A. (2015). Evaluation of Determinants of Employment Efficiency Using Stochastic Frontier Analysis and Beta Regression. Journal of Mathematical and Computational Science. No.6, 848-856.

Espinheira, P.L., Ferrari S.L.P., & Cribari-Neto, F. (2008). On beta regression residuals. Journal of Applied Statistics, 35:4, 407-419.

FAO, (2019). Crops and livestock products. (Last Accessed Time: 21/11/2021). URL: https://www.fao.org/faostat/en/#data/QCL.

Ferrari, S.L.P., & Cribari-Neto, F. (2004). Beta regression for modelling rates and proportions. Journal of Applied Statistics. 31, pp. 799–815.

Henningsen, A. & Toomet, O. (2011). maxLik: A package for maximum likelihood estimation in R. Computational Statistics 26(3), 443-458.

Kieschnick, R., & McCullough, B. D. (2003). Regression analysis of variates observed on (0, 1): percentages, proportions and fractions. Statistical Modelling, 3(3), 193–213.

Ospina, R., & Ferrari, S.L. (2012). A general class of zero-or-one inflated beta regression models. Computational Statistics & Data Analysis, 56(6), 1609-1623.

Ozek, K. (2015). Factors Affecting Composition of Milk in Dairy Cattle and Relation between Nutrition and Milk Composition. Journal of Bahri Dagdas Animal Research 4 (2):37-45.

R Core Team. (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL: https://www.R-project.org/.

Revelle, W. (2020). psych: Procedures for Personality and Psychological Research, Northwestern University, Evanston, Illinois, USA, https://CRAN.R-project.org/package=psych Version = 2.0.12.

Schmid, M., Wickler, F., Maloney, K.O., Mitchell, R., Fenske, N., & Mayr, A. (2013). Boosted Beta Regression. PLoS ONE 8(4): e61623.

Sokahl, R., & Rohlf, F. (1995). Biometry (3rd ed.). New York: W. H. Freeman.

Tyrisevä, A.M., Elo, K., Kuusipuro, A., Vilva, V., Jänönen, I., Karjalainen, H., Ikonen, T., & Ojala, M. (2008). Chromosomal Regions Underlying Noncoagulation of Milk in Finnish Ayrshire Cows. Genetics. 180:1211–1220.

Warton, D. I., & Hui, F. K. C. (2011). The arcsine is asinine: The analysis of proportions in ecology. Ecology, 92(1), 3–10.

Zhou, H., Huang, X., & Alzheimer's Disease Neuroimaging Initiative. (2020). Parametric mode regression for bounded responses. Biometrical Journal, 62(7), 1791-1809.

CHAPTER VIII

DETERMINATION OF SPATIAL DISTRIBUTION PATTERN OF SOME SOIL PROPERTIES USING DIFFERENT INTERPOLATION METHODS IN URBAN SAZOVA PARK IN ESKISEHIR

Dr. Gafur GÖZÜKARA[*]

[*] Eskisehir Osmangazi University, Faculty of Agriculture, Department of Soil Science and Plant Nutrition, Eskisehir, Turkey. ggozukara@ogu.edu.tr

INTRODUCTION

The sustainability of soils depends on determining the chemical and physical properties of soils with minimum cost and time and managing them according to their properties. Moreover, some limitations on soil properties such as soil salinity, high or low soil pH, erosion, low organic matter, limited soil formation affected agricultural practice and sustainable agriculture. With this perspective, interpolation techniques, which is a critical role to save time and cost, have been increasingly used to determine distribution soil physical and chemical properties.

Determining the spatial variation pattern of physical and chemical soil properties via different interpolation models allows predicting the value of the investigated soil properties at any point in the research area with the lowest error. Therefore, the distribution maps obtained as a result of the interpolation analysis of soil properties allow the most appropriate planning and management decisions related to land management for the research area (Öztaş, 1996; Özyazıcı et al., 2015; Gülser et al., 2016; Çelik and Dengiz, 2017; Alaboz et al., 2020). In soil science, many studies have been commonly utilized some interpolation methods such as ordinary, simple, universal kriging (Taşan and Demir, 2017; Arslan et al., 2018; Celilov and Dengiz, 2019; Şimşek et al., 2020; Altunbaş et al., 2020; Alaboz et al., 2020), inverse distance weighting (IDW) (with different power levels) (Özyazıcı et al., 2015; Çelik and Dengiz, 2017; Sancan and Karaca, 2017; Taşan and Demir, 2017; Arslan et al., 2018; Celilov and Dengiz, 2019; Alaboz et al., 2020; Şenol et al., 2020), and radial basis function (RBF) (Taşan and Demir, 2017; Arslan et al., 2018; Celilov and Dengiz, 2019; Özdemir et al., 2019; Alaboz et al., 2020) to determine distribution of soil physical and chemical properties. Moreover, they reported that these interpolation methods have been successfully used to determine distribution maps of soil properties. However, according to the soil physical and chemical properties they focused on, some researchers reported that the kriging interpolation methods were more successful than IDW and RBF interpolation methods for some soil properties, whereas some researchers reported that the IDW model was more successful than kriging and RBF interpolation methods for some soil properties (Taşan and Demir, 2017; Özdemir et al., 2019; Alaboz et al., 2020). These differences in prediction performances are affected by many factors such as parent material, land use, soil formation rate, depth, soil diversity. Thus, these interpolation methods should be tested at different soils which are developing under different parent materials, climate conditions, and land use to develop prediction performance. According to the studies, kriging, IDW, and RBF interpolations methods were applied on agricultural and pasture soil. However, these interpolation models were not used to predict soil properties in urban park areas. However, the sustainability of large urban park areas such as research areas depends on chemical and physical soil properties.

Therefore, the aims of this study were i-) determine some chemical and physical soil properties and ii-) compare ordinary, simple, universal kriging with different semivariogram (spherical, exponential, and Gaussian), IDW with different power levels (1, 2, and 3), and RBF with different types completely regularized spline interpolation, spline with tension, and thin plate spline.

MATERIAL AND METHODS

2.1. Study Area and Soil Sampling

The soil sampling points are located in Sazova Park, Odunpazarı, Eskisehir, Turkey (4404545.96 – 4404226.30 K latitudes and 283965.88 – 283303.21 D longitude (WGS-84, UTM-m, 36 Zone) (Fig. 1). The study area covers almost 28 ha. This area is categorized as cold semi-arid climate (BSk) according to the Köppen-Geiger classification system with mean annual precipitation of 522.2 mm and mean annual temperature of 13.6 ˚C. The soil temperature and moisture regimes were classified as mesic and xeric (dry xeric in the subclass), respectively. The soils were developed in a flat landscape (slopes: 0.5–1%) and elevation of 794-797m above sea level. The soil samples were collected based on a regular grid sampling strategy with 60 m x 60 m. A total of 60 soil samples were collected from 0-20 cm depth and transported to the laboratory for chemical and physical analysis.

2.2. Soil Physical and Chemical Analyses

The soil samples were air-dried, disaggregated, and sieved via a 2 mm diameter sieve for analyses. The air-dried and sieved soil samples were analyzed for physical and chemical properties. Sand, silt, and clay content were determined using the hydrometer method (Bouyoucos, 1953). Soil reaction (pH) and electrical conductivity (EC) were measured in a 1:1 soil water suspension using a digital Isolab pH-EC meter (Isolab Inc., Eschau, Germany) (Soil Survey Staff, 2014). $CaCO_3$ was measured using a Scheibler calcimeter (Allison and Moodie, 1965). Soil organic matter (OM) content was measured by the modified Walkey-Black method (Black, 1965).

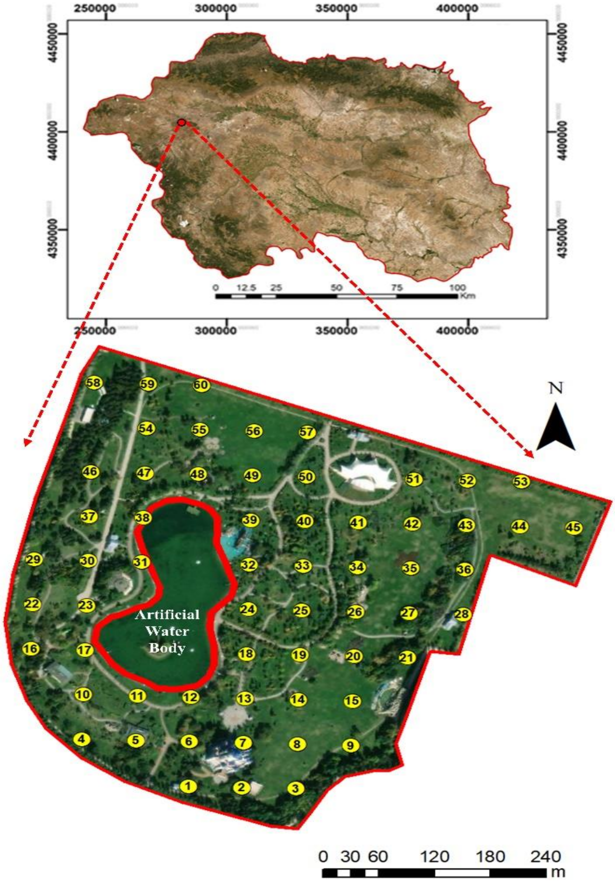


Figure 1. Location of Study Area (Odunpazari, Eskisehir, Turkey) and Sampling Points (yellow 60 sampling points) and Inside of Bold Red Polygon Represent Artificial Water Body

2.2. Interpolation methods

The ordinary, simple, universal kriging (scholastic approaches), inverse distance weighting (IDW), and radial basis function (RBF) with methods were applied to interpolate EC, pH, SOM, $CaCO_3$, sand, silt and clay. The spherical, exponential, and Gaussian semivariogram models for each kriging interpolation methods power level (1, 2, and 3) for IDW, and completely regularized spline interpolation, spline with tension, and thin plate spline for RBF were tested to select the best method for interpolating for each variable.

Kriging is based on the logic of estimating the value of a variable at an unknown point, using the values of known points to estimate the value of a variable, similar to other estimation models in general (Li and Heap, 2008). The general equation of the kriging interpolation method is given in equation 1.

$$\hat{Z}(x_0) - \mu = \sum_{i=1}^{n} \lambda_i [\, Z(Xi) - \mu(xo)] \qquad \text{(Eq.1)}$$

Where, μ = still average, λ_i= kriging weight, n = the number of points used to estimate, $\mu(x\_0)$ = research area samples average, $\hat{Z}(x_0)$ = true value of the predicted point.

One of the most widely used interpolation methods, IDW is based on estimating the unknown value from the known point. This assumption was based on the logic that as the distance from the known point to the target point increases, the similarities decrease (Shepard, 1968; Li and

Heap, 2008). The general equation of the IDW method is given in Equations 2 and 3.

$$Z\,(x) = \frac{\sum_{i=1}^{n} w_i z_i}{\sum_{i=1}^{n} w_i} \qquad \text{(Eq.2)}$$

$$W_i = d_i^{-u} \qquad \text{(Eq.3)}$$

Where, Z(x)= Estimated value at the interpolated point, Z_i=The value at the known point, n=the total number of points whose value is known, d_i=The distance between the point and the point whose value is to be estimated, W_i=weight assigned to the point, u=strength of the parameter

The RBF method is a method used in the interpolation of multidimensional data. It is generally used for estimating a limited number of data or points that are difficult to predict. The biggest advantage of this method is that it can be easily used in any size due to the low general restrictions (Wright, 2003). The general equation of the RBF method is given in Equation 1.

$$S(\underline{x}) = \sum_{\mathrm{j}=1}^{\mathrm{n}} \lambda_j \phi(\|\mathrm{x} - \mathrm{x_j}\|) \qquad \text{(Eq. 4)}$$

Where, x = free variable, λ_j = expansion coefficient, ϕ= stands for a single variable.

Normal distribution of soil properties was tested with the Kolmogorov-Smirnov test in ArcGIS software. The non-normally distributed variables are approximated to the normal distribution by applying the logarithmic transformation.

2.3. Comparison of Interpolation Methods and Evaluation

The root mean squared error (RMSE) (Eq.5) was calculated for each method to determine the best interpolation method and semivariogram model for mapping the spatial distribution of variables. Many studies were commonly used the RMSE value to select the best prediction model (Mihalikova et al., 2015; Tunçay et al., 2018; Alaboz et al., 2020). The distribution maps of EC, pH, SOM, $CaCO_3$, sand, silt and clay were created through ArcGIS 10.7.1 software.

$$RMSE \ = \sqrt{\frac{\sum(Z_i - Z)^2}{n}} \qquad \text{(Eq. 5)}$$

Where z_i, z and n are predicted values, observed values, and the number of observations, respectively.

RESULTS AND DISCUSSION

Descriptive statistics and box plots of soil samples collected from urban Sazova park are given in Table 1 and Figure 2. Soil properties changed from pH 8.40 to 9.11, EC 0.23 to 0.67 dS m^{-1}, OM 1.02 to 4.71%, $CaCO_3$ 5.67 to 13.41%, sand 15.64 to 69.48%, silt 15.64 to 69.48%, and clay 15.64 to 69.48% at study area. The soil texture of soil samples was described as clay (C) (1 soil sample), silty clay loam (SiCL) (1 soil sample), sandy loam (SL) (1 soil sample), and silty loam (SiL) (57 soil samples. In particular, according to the results of the physical and chemical analysis of the soils, the maximum coefficient of variation

(CV) value was determined in clay content with CV = 105.43%, whereas the minimum coefficient of variation value was determined in the pH with CV = 1.69%. Generally, pH had low CV value compared to other soil physical and chemical properties. Same results reported by Gulmezoğlu et al., (2017), Horuz and Dengiz (2018), Şenol et al., (2020), Gözükara et al., (2021), Davutoğlu et al., (2021). In general, the soil particle size distribution of the study area had a high CV value compared to the other soil properties. In particular, while the clay content varies between 2.54-9.32%, clay content clearly increased at some parts of the study area. The high CV values (105.43%) in the clay content may be affected by the differences in the soils at the stage of establishment of the Sazova park area.

Table 1. Descriptive Statistics of Soil Physical and Chemical Properties

Soil properties	Min	1st. Qu	Med.	Mean	3rd Qu.	Max	Skewness	Kurtosis	CV (%)
pH	8.40	8.60	8.70	8.70	8.80	9.11	0.38	0.52	1.69
EC (dS m^{-1})	0.23	0.41	0.44	0.45	0.49	0.67	0.31	1.74	16.05
OM (%)	1.02	2.17	2.64	2.68	3.21	4.71	0.11	0.21	28.47
$CaCO_3$ (%)	5.67	6.91	8.13	8.41	9.35	13.41	0.74	0.15	21.42
Sand (%)	15.64	23.66	28.38	29.28	32.60	69.48	2.09	9.18	27.70
Silt (%)	28.80	60.44	65.46	63.52	67.78	77.72	-1.91	5.77	13.39
Clay (%)	0.82	2.54	4.72	7.21	9.32	43.82	2.75	9.66	105.43

CV: coefficient of variation

The Pearson correlations coefficients (r) between soil chemical and physical properties were calculated and given in Table 2. As a result of the study, 9 out of 21 correlation pairs were found to be statistically significant ($p<0.05$; $p<0.01$) (Table 2). In general, a high considerable correlation was observed between soil particle size (sand, silt, and clay). The EC had negatively correlated (r = –0.47**) with sand content, whereas positively correlated with silt content (r = 0.27*). As expected, the silt content highly correlated with sand content (r = –0.58**) and clay content (r = –0.50**). Moreover, the pH was negatively correlated with OM (r = –0.32*) and silt (r = –0.28*), while positively correlated with $CaCO_3$ (r = –0.31*).

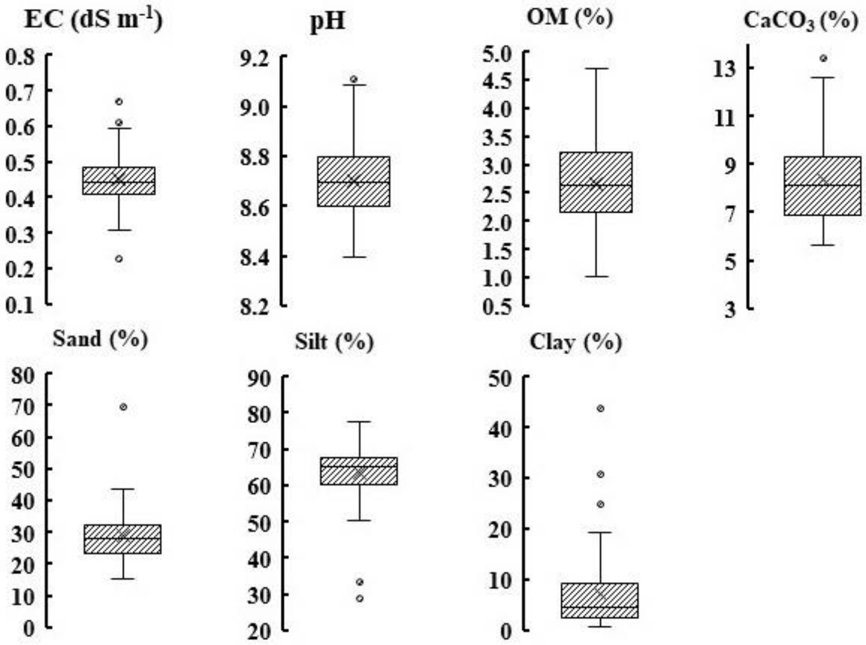


Figure 2. Distribution of Soil Physical and Chemical Properties

The Pearson correlation between soil physical and chemical properties is affected by many factors which soil parent material, soil formation, depth, climate condition, land use, different soil horizon, and soil types. Based on these differences, many studies reported different results on the Pearson correlation between soil physical and chemical properties (İmamoğlu et al., 2016; Celilov and Dengiz, 2019; Gözükara et. al., 2021).

Table 2. The Pearson Correlations Between Soil Physical and Chemical Properties

Soil Properties	EC	pH	OM	$CaCO_3$	Sand	Silt	Clay
EC	1	0.17	0.25	0.10	-0.47**	0.27*	0.20
pH		1	-0.32*	0.31*	0.25	-0.28*	0.05
OM			1	-0.30*	0.01	0.22	-.025
$CaCO_3$				1	-0.10	-0.18	0.31*
Sand					1	-0.58**	-0.42**
Silt						1	-0.50**
Clay							1

*： p<0, 05 **： p<0. 01

The EC, pH, OM, $CaCO_3$, sand, silt, and clay maps were created via kriging, IDW, and RBF interpolation methods based on a digital soil mapping approach. The result of the RMSE value for each soil property is given in Table 3. Each soil property was tested by Kolmogorov-Smirnov normality test for determination of normal or not normal distribution. Results of the normality test, EC, pH, and OM were normally distributed, whereas $CaCO_3$, sand, silt, and clay contents were not normally distributed. Therefore, logarithmic transformation was applied on $CaCO_3$, sand, silt, and clay contents to obtain normally distribution for these soil properties. According to Table 5, fifteen interpolation methods were applied to select the best model based on the lowest RMSE value for the distribution of EC, pH, OM, $CaCO_3$, sand, silt, and clay in Sazova park. Among fifteen interpolation methods, kriging had the lowest RMSE value for distribution maps of EC, pH, OM, $CaCO_3$, sand, silt, and clay. Moreover, among kriging interpolation methods, the simple kriging method had the lowest RMSE value compared to ordinary and universal interpolation methods. EC and pH had the lowest RMSE value with the Gaussian semivariogram model. The OM, $CaCO_3$, sand, and silt had the lowest RMSE value with the spherical semivariogram model, whereas clay had the lowest RMSE value with the exponential semivariogram model. When used the IDW interpolation method, the power levels considerably affected the RMSE value. In particular, the RMSE value of EC, pH, sand, silt, and clay content increased with increasing power level (1 to 3), while the RMSE value of OM and $CaCO_3$ content decreased with decreasing power level (3 to 1). RBF interpolation methods had the highest RMSE value compared to kriging and IDW interpolation methods. Moreover, ST had the lowest RMSE value among RBF interpolation methods. Distribution maps of soil properties were shown in Fig. 3.

Table 3. The Results of RMSE Value of Kriging and IDW Interpolations

IM	Model	Type	EC	pH	OM	$CaCO_3$	Sand	Silt	Clay
Kriging	Ordinary	Spherical	0.074	0.148	0.768	1.882	8.215	9.021	7.921
		Exponential	0.074	0.149	0.778	1.885	8.246	9.046	8.879
		Gaussian	0.073	**0.146**	0.778	1.871	8.194	9.015	7.887
	Simple	Spherical	0.073	0.147	**0.732**	**1.814**	**8.047**	**8.976**	7.590
		Exponential	0.073	0.149	0.744	1.817	**8.047**	8.983	**7.557**
		Gaussian	**0.072**	0.146	0.740	1.826	**8.047**	8.985	7.564
	Universal	Spherical	0.074	0.148	0.768	1.882	8.215	9.052	7.922
		Exponential	0.074	0.149	0.777	1.885	8.246	9.046	7.879
		Gaussian	0.073	0.146	0.775	1.871	8.194	9.015	7.887
IDW	Power Levels	1	0.074	0.149	0.790	1.827	8.263	8.942	7.698
		2	0.077	0.152	0.773	1.819	8.582	9.042	7.801
		3	0.081	0.156	0.767	1.822	9.059	9.340	7.957
RBF	Kernel Functions	CRS	0.080	0.158	0.759	1.882	8.851	9.229	8.113
		ST	0.078	0.155	0.759	1.867	8.621	9.064	7.955
		TPS	0.106	0.196	0.874	2.234	12.58	12.37	9.886

Abbreviations: IM; interpolation methods, IDW; Inverse Distance Weighting, RBF; Radial Basis Functions, CRS; Completely Regularized Spline, ST; Spline with Tension, TPS; Thin Plate Spline

The EC and pH had almost similar distribution. In particular, the north and northwest of the study area had the highest value for EC and pH. Northeast and southeast of the study area had highest value OM. Soil particle size had heterogeneous distribution, whereas $CaCO_3$ had homogeneous distribution compared to other soil properties.

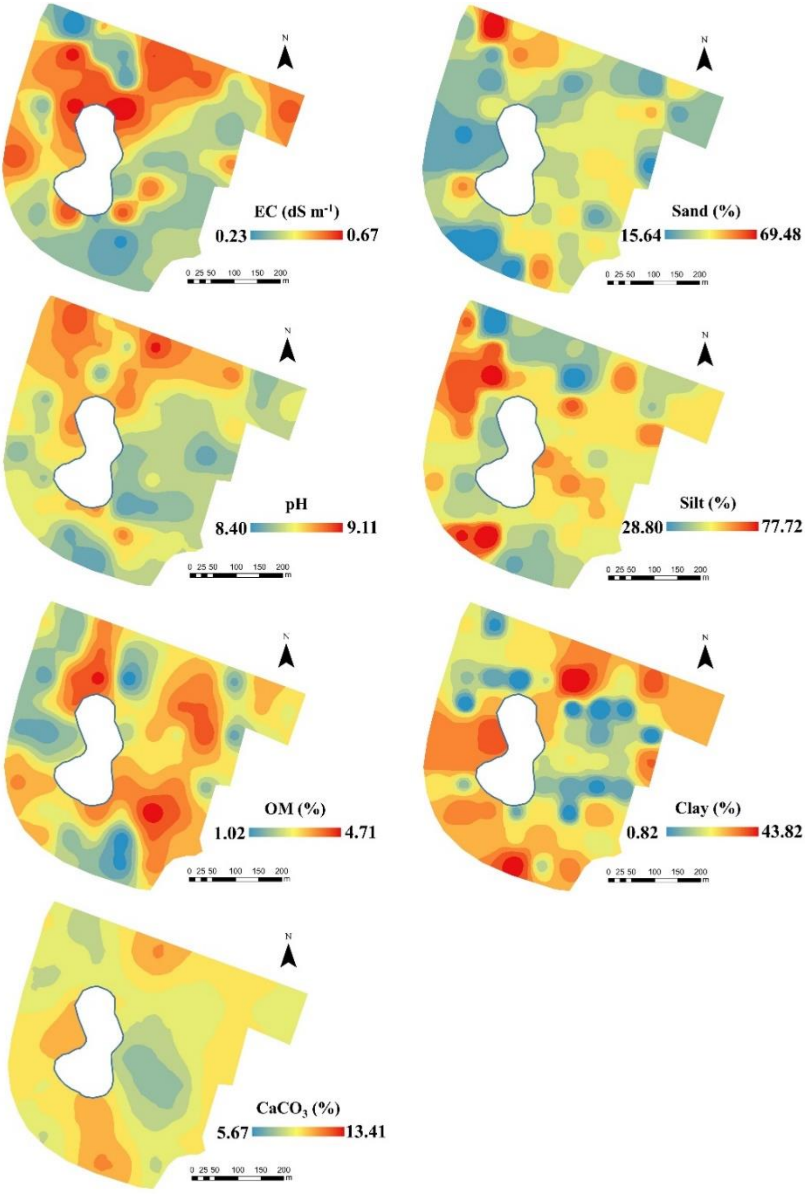


Figure 3. Distribution of Electrical Conductivity (EC), pH, $CaCO_3$, Soil Organic Matter (OM), Sand, Silt, and Clay Contents at Study Area.

CONCLUSION

Interpolation techniques, which is a critical role to save time and cost, have been increasingly used to determine distribution soil properties. In this context, fifteen interpolations methods, which is include kriging, IDW, and RBF, were applied to select the best prediction model for distribution maps of EC, pH, OM, $CaCO_3$, sand, silt, and clay. In general, soil samples were described as having no risk of salinity, alkaline, moderate level $CaCO_3$ and OM content, the silty loam of urban Sazova Park soils.

As a result of this study, it was concluded that kriging had the lowest RMSE value compared to IDW and RBF for distribution maps of EC, pH, OM, $CaCO_3$, sand, silt, and clay. Moreover, among kriging interpolation methods, the simple kriging method had the lowest RMSE value compared to ordinary and universal interpolation methods. EC and pH had the lowest RMSE value with the Gaussian semivariogram model. In addition, distribution maps of soil physical and chemical properties must be taken into account in the management of grass areas in the Sazova Park.

ACKNOWLEDGMENTS

I gratefully thank to Dr. Onur İLERİ, Meliha Feryal SARIKAYA, and Elifnur ELEREN from Eskisehir Osmangazi University, Faculty of Agriculture, Department of Field Crops for their helping in the soil sampling at urban Sazova Park.

REFERENCES

Alaboz, P., Demir, S. & Dengiz, O. (2020). Determination of spatial distribution of soil moisture constant using different interpolation model case study, Isparta Atabey Plain. Journal of Tekirdag Agricultural Faculty, 17(3): 432-444.

Allison, L. E. & Moodie, C. D. (1965). Carbonate. Agronomy monograph, methods of soil analysis. Part 2. In: Chemical and Microbiological Properties, Agronomy. 9.2. American Society of Agronomy, Wisconsin, pp. 1379–1396.

Altunbaş, S., Gözükara, G. & Demirel, B. Ç. (2020). Determination of soils properties and distributions developing on different fluvial deposits in Aksu Plain. Journal of Agricultural Faculty of Ege University, 57(3): 381-391.

Arslan, E., Çaycı, G., Dengiz, O., Yüksel, M. & Atikmen, N. Ç. (2018). Determination of spatial distribution of some macro nutrient contents of soils under different agricultural uses. Soil Water Journal, 7(2): 28-37.

Black, C. A. (1965). Methods of soil analysis Part 2, Amer. Society of Agronomy Inc., Publisher Madison, Wisconsin, U.S.A., 1372-1376.

Bouyoucos, G. J. (1953). An improved type of soil hydrometer. Soil Science, 76, 377-378.

Celilov, C. & Dengiz, O. (2019). Determination of the spatial distribution for erodibility parameters using different interpolation methods: Ilgaz National Park Soils, Turkey. Turkish Journal of Agricultural Research, 6(3): 242-256.

Çelik, P. & Dengiz, O. (2017). Determination of basic soil properties and nutrient element states of agricultural soils of akselendi plain and formation of distribution maps. Turkish Journal of Agricultural Research, 5(1): 9-18.

Davutoğlu, A., Gözükara, G. & Gülmezoğlu, N. (2021). Effects of soil properties on yield and quality of roasted pumpkin. Turkish Journal of Agriculture - Food Science and Technology, 9(5): 909-918.

Gözükara, G., Demirel, B. Ç. & Altunbaş, S. (2021). Effect of soil horizons on the relationship between digital color parameters and soil properties. Mediterranean Agricultural Sciences, 34(1): 125-133.

Gulmezoglu, N., Aytac, Z., Kutlu, I., Kulan, E. G. & Gozukara, G. (2017). Mapping boron and beneficial heavy metal ions for wheat-cultivating soils in turkey's boron-mining zone. Applied Ecology and Environmental Research, 15(3): 1119-1130.

Gülser, C., Ekberli, İ., Candemir, F. & Demir, Z. (2016). Spatial variability of soil physical properties in a cultivated field. Eurasian Journal of Soil Science, 5(3): 192-200.

Horuz, A. & Dengiz, O. (2018). The relationships between some physico-chemical properties and nutrient element content of paddy raised on alluvial land in Terme Region. Anadolu Journal of Agricultural Sciences, 33(1): 58-67.

İmamoğlu, A., Bahadır, M. & Dengiz, O. (2016). Determination of spatial distribution pattern of soil erodibility factor using different interpolation Models in Çorum-Alaca Watershed. Soil Water Journal, 5(1): 8-15.

Li, J. & Heap, A. D. (2008). A Review of spatial ınterpolation methods for environmental scientists. Geoscience Australia, Record 2008/23, 137 pp.

Mihalikova, M., Başkan, O. & Dengiz, O. (2015). Capability of different interpolation models and pedotransfer functions to estimate soil hydraulic properties in Büyükçay Watershed. Environmental Earth Sciences, 74: 2425-2437.

Özdemir, Ş., Günal, H., Acir, N., Arslan, H., Özaydın, K. A., Kahyaoğlu, S. E. & Ağar, A. M. (2019). Use of deterministic and stochastic interpolation methods for estimating soil salinity in çerikli irrigation area. Soil Water Journal, 8(1): 55-67.

Öztaş, T. (1996). Eğimli bir arazide erozyonla kaybolan toprak derinliğindeki değişimin Kriging analizi ile belirlenmesi. Tarım-Çevre İlişkileri Sempozyumu, "Doğal Kaynakların Sürdürülebilir Kullanımı", 13- 15 Mayıs, Mersin, s. 327-335.

Özyazıcı, M. A., Dengiz, O., Aydoğan, M., Bayraklı, B., Kesim, E., Urla, Ö., Yıldız, H. & Ünal, E. (2015). Concentrations of some macro and micro plant nutrient of cultivated soils in Central and Eastern Blacksea Region and their mapping by inverse distance weighted (IDW) method. Artvin Çoruh University Journal of Forestry Faculty, 16(2): 187-202.

Sancan, M. & Karaca, S. (2017). Determination of some soil properties and mapping by geographical information systems in vineyard areas of Bayramlı village, Erciş county Van. Journal of Soil Science and Plant Nutrition, 5(2): 55-62.

Soil Survey Staff. (2014). Keys to Soil Taxonomy. Twelfth Edition Edition, United States Department of Agriculture, Natural Resources Conservation Service ISBN 0-16-048848-6. Washington DC.

Shepard, D. (1968). A two-dimensional interpolation function for irregularly-spaced data. Proceedings of the 1968 ACM National Conference. pp. 517–524.

Şenol, H., Alaboz, P. & Dengiz, O. (2020). Evaluation of the pysico-chemical and nutrient elements status of soils formed on different parent materials using interpolation method. Anadolu Journal of Agricultural Sciences, 35(3): 505-516.

Şimşek, O., Altunbaş, S., Demirel, B. Ç. & Gözükara, G. (2020). Land evaluation studies on different soils developing on alluvial physiographies. Mediterranean Agricultural Sciences, 33(1): 129-135.

Taşan, M. & Demir, Y. (2017). Determination of spatial distribution of iron and manganese contents with different interpolation methods at rice cultivated areas. Anadolu Journal of Agricultural Sciences, 32(1): 64-73.

Tunçay, T., Başkan, O., Bayramin, İ., Dengiz, O. & Kılıç, Ş. (2018). Geostatistical approach as a tool for estimation of field capacity and permanent wilting point in semiarid terrestrial ecosystem. Archives of Agronomy and Soil Science, 64 (9): 1240-1253.

Wright, G. (2003). Radial Basis Function Interpolation: Numerical and Analytical Developments. Ph. D. thesis, University of Colorado, Boulder.

CHAPTER IX

CLASSIFICATION OF SOME PLANTS IN TURKEY ACCORDING TO GEOGRAPHICAL REGIONS IN TERMS OF YIELD USING THE QUEST ALGORITHM

Assoc. Prof. Dr. Şenol Çelik[*]

[*] Bingöl University Faculty of Agriculture Animal Science, Biometry and Genetic Department Bingöl-Turkey, E-mail: senolcelik@bingol.edu.tr

INTRODUCTION

The QUEST (Quick, Unbiased, Efficient, Statistical Tree) algorithm supports univariate and linear combination splits (Loh and Shih, 1997). For each split, the association between each input attribute and the target attribute is computed using the ANOVA F–test or Levene's test or Pearson's chi–square (for nominal attributes). If the goal attribute is multinomial, two means clustering is used to create two super classes. The attribute that obtains the highest association with the target attribute is selected for splitting. QUEST has negligible bias and it yields binary decision trees. Ten–fold cross–validation is used to prune the trees (Maimon and Rokach , 2010). As a hole the QUEST can be classified as a complex system of data analysis which gives ability to analyze various variants of predictors and use optimization procedures to choose them (Loh and Shih, 1997).

QUEST algorithm (Loh and Shih, 1997) employs modification of the recursive quadratic discriminant analysis and includes a number of innovative features for improving the reliability and efficiency of the classification tree that it computes. The QUEST algorithm is fast and unbiased. Its lack of bias in variable selection for splits is a distinct advantage when some predictor variables have few levels and other have many. Furthermore, QUEST does not sacrifice the predictive accuracy for speed (Lim et al., 1997).

Loh (2014) pointed out that QUEST has two steps, which are based on the "significance tests to split each node". During the first test, the association of each X with Y is tested. The variable selection is based on level of significance. The highest significant variable is selected. If each of the X is independent of Y, then each X has the same selection chance. As a result, selection bias is not presented in this approach. QUEST uses different tests based on the nature of the variables (Loh, 2014). For categorical variables, it utilizes chi-squared tests. For ordered variables, analysis of variance tests is utilized.

QUEST is used for data classification and mining in variety of combinations. These include linear or univariate combination splits. The unique aspect of QUEST that the bias in its attribute selection method is negligible. QUEST tree construction process consists of split predictor selection, split point selection for the split predictor, and stopping (Lutfi, 2020).

Classification performance can be appraised by computing the number of correctly identified class examples (true positives; tp), the number of correctly recognized examples that do not belong to the class (true negatives; tn), and the number of examples that were either incorrectly assigned to the class (false positives; fp) or that were unrecognized as class examples (false negatives; fn) (Sokolova and Lapalme, 2009).
In order to evaluate the model performance, classification performance measure can be obtained as follow (Ferri et al., 2009; Kim, 2010; Chou et al., 2014).

$$Accuracy = \frac{tp + tn}{tp + fn + fp + tn}$$

$$Precision = \frac{tp}{tp + fp}$$

$$Sensitivity = \frac{tp}{tp + fn}$$

$$Specificity = \frac{tn}{fp + tn}$$

$$AUC = \frac{1}{2}\left[\left(\frac{tp}{tp + fn}\right) + \left(\frac{tn}{fp + tn}\right)\right]$$

Based on the above measures, the following overall average performance score (S) is proposed

$$S = \frac{1}{m}\sum_{i=1}^{m} P_i$$

here m represents the number of distinct performance measures and P_i denotes the i^{th} performance measure. The S range is 0–1; the coefficient positively correlates with the effectiveness of the overall evaluation measures.

STUDIES ON THE QUEST ALGORITHM

The subject and content of the studies on the QUEST Algorithm are summarized in Table 1.

Table 1. Some studies on the QUEST Algorithm

Topic	**Results**	**References**
A case study of construction defects in Taiwan	The CHAID algorithm generates more rules than the CART and QUEST algorithms do, and this can be considered an informational advantage for the CHAID algorithm. The prediction accuracy observed for the CHAID algorithm was 75.45% in the study.	Lin and Fan (2019)
Reviewing and comparing some of the algorithms such as QUEST, GUIDE, CRUISE, C4.5 and RPART with their strengths, weakness and capabilities.	The limitation of RPART is, it has fever child nodes comparing QUEST, GUIDE and CRUISE even though C4.5 trees often have the most by far. The accuracy of CRUISE and QUEST are high if it is using the linear combination splits.	Shamy and Dheeba (2016)
An application on computer and internet security	The accuracy rate in the applied QUEST algorithm was found to be 62.33%.	Çalış et al. (2014)
Predicting causes of traumatic brain injuries	If a person's age is less than or equal to 36.61 and place of incident is Zahedan, then the probability is 28 percent. If a person's age is less than or equal to 36.61, or the incident location is outside the city or unknown, then the likelihood of death is 62.22 percent.	Raeesi et al. (2014)
Dietary protein is the strong predictor of coronary artery disease	According to QUEST model, the accuracy of the tree was 84.36% for training dataset and 82.94% for testing dataset.	Soflaei et al. (2021).
Feature-based decision rules for control charts pattern recognition	CART-based decision trees result in better recognition performance show but lesser consistency, whereas, the QUEST-based decision trees give better consistency but lesser provide performance.	Bag et. al. (2012)
Survival prediction of patients with breast cancer	The C5.0 algorithm performed better results in predicting breast cancer survival than CHAID, QUEST, CART	Momenyan et al. (2018)

	algorithms and the logistic regression.	
Optimizing parameters of support vector machine	All classification models (GASVM, C5.0, CART, CHAID) achieved at least 80% accuracy except QUEST.	Chou et al. (2014)
Urban flood risk mapping	In the QUEST model applied to analyze the flood vulnerability, AUC-ROC=89.2%, Kappa=0.79.	Darabi et al. (2019)
Entrepreneurial intentions among the youth	QUEST classification tree algorithm has managed to provide additional insight into the influence of potential predictors on entrepreneurial intentions.	Djordjevic et al. (2021)

APPLICATION

The QUEST algorithm was used to classify field crops such as barley, rye, oats, beans, chickpeas, potatoes, sugar beets, alfalfa and sainfoin produced in Turkey according to 7 geographical regions in terms of yield per decare. In the QUEST algorithm, the parent node: child node ratio is 6:3. The number and rate of classification for each region are given in Table 2.

Table 2. Classification by geographic regions

Classification								
	Predicted							
Observed	R1	R2	R3	R4	R5	R6	R7	Percent Correct
R1	4	0	0	0	1	2	0	57.1%
R2	0	0	0	4	1	2	1	0.0%
R3	0	0	6	1	1	3	0	54.5%
R4	0	0	0	11	0	2	0	84.6%

R5	3	0	0	1	2	8	0	14.3%
R6	1	0	0	3	1	13	0	72.2%
R7	1	0	0	0	0	0	5	83.3%
Overall Percentage	11.7%	0.0%	7.8%	26.0%	7.8%	39.0%	7.8%	53.2%
Growing Method: QUEST, Dependent Variable: region								

R1: Mediterranean, R2: Aegean, R3: Marmara, R4: Central Anatolia, R5: Eastern Anatolia, R6: Black Sea, R7: Southeast Anatolia

When Table 2 is examined, the accuracy rates of classification according to regions are seen. The highest accuracy rate was 84.6% in R4 (Central Anatolian Region) and the lowest accuracy rate was 0% in R2 (Aegean Region). The overall accuracy rate was 53.2%. Decision tree based on the QUEST algorithm for classification by geographic regions was displayed in Figure 1.

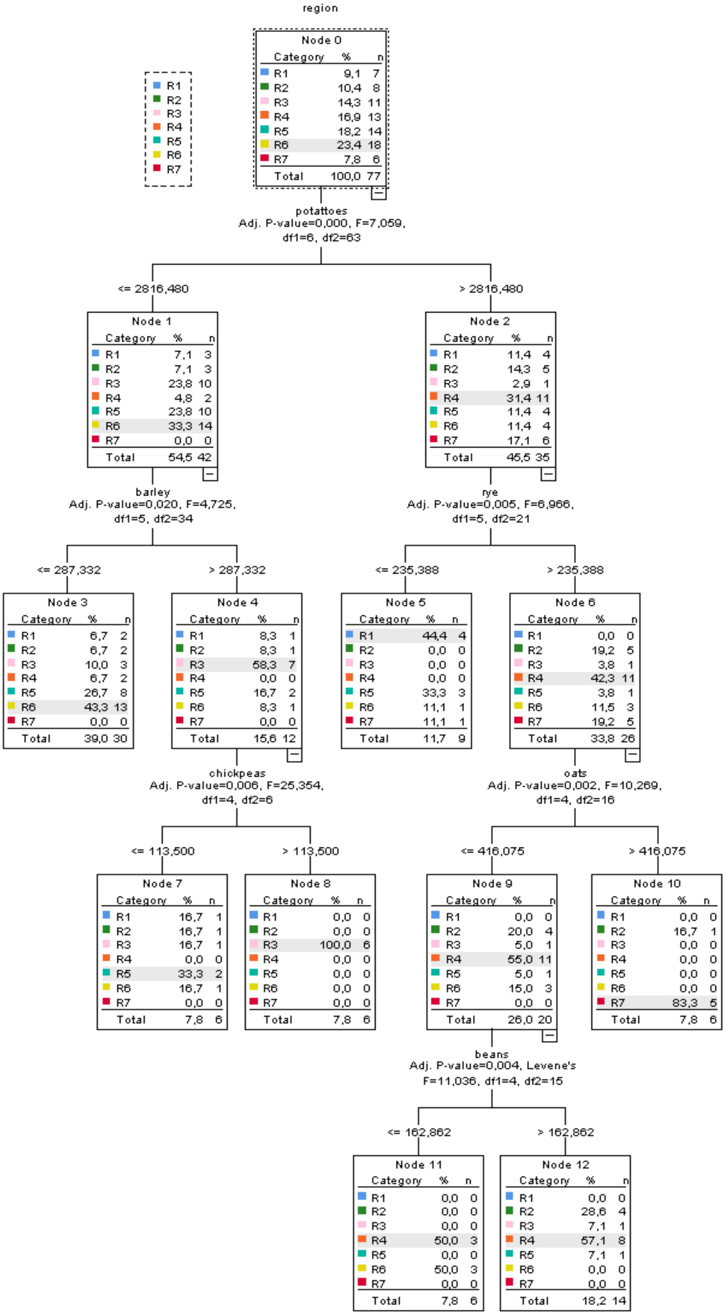


Figure 1. Decision tree based on the QUEST algorithm

The rules explaining the decision tree in Figure 1 are as follows:

- Node 1: if (potatoes ≤ 2816.48) then class R1 = 7.1%, R2=7.1%, R3=23.8%, R4=4.8%, R5= 23.8%, R6=33.3% and class R7 = 0%
- Node 2: if (potatoes > 2816.48) then class R1 = 11.4%, R2=14.3%, R3=2.9%, R4=31.4%, R5= 11.4%, R6=11.4% and class R7 = 17.1%
- Node 3: if (barley ≤ 287.332) then class R1 = 5.7%, R2=5.7%, R3=10%, R4=5.7%, R5= 26.7%, R6=43.3% and class R7 = 0%
- Node 4: if (barley > 287.332) then class R1 = 8.3%, R2=8.3%, R3=58.3%, R4=0%, R5= 16.7%, R6=8.3% and class R7 = 0%
- Node 5: if (rye ≤ 235.388) then class R1 = 44.4%, R2=0%, R3=0%, R4=0%, R5= 33.3%, R6=11.1% and class R7 = 11.1%
- Node 6: if (rye > 235.388) then class R1 = 0%, R2=19.2%, R3=3.8%, R4=42.3%, R5= 3.8%, R6=11.5% and class R7 = 19.2%
- Node 7: if (chickpeas ≤ 113.5) then class R1 = 16.7%, R2=16.7%, R3=16.7%, R4=0%, R5= 33.3%, R6=16.7% and class R7 = 0%
- Node 8: if (chickpeas > 113.5) then class R1 = 0%, R2=0%, R3=100%, R4=0%, R5= 33.3%, R6=16.7% and class R7 = 0%
- Node 9: if (oats ≤ 416.075) then class R1 = 0%, R2=20%, R3=5%, R4=55%, R5= 5%, R6=15% and class R7 = 0%
- Node 10: if (oats > 416.075) then class R1 = 0%, R2=16.7%, R3=0%, R4=0%, R5= 0%, R6=0% and class R7 = 83.3%

- Node 11: if (beans ≤ 162.862) then class R1 = 0%, R2=0%, R3=0%, R4=50%, R5= 0%, R6=50% and class R7 = 0%
- Node 12: if (beans > 162.862) then class R1 = 0%, R2=28.6%, R3=7.1%, R4=57.1%, R5= 7.1%, R6=0% and class R7 = 0%

CONCLUSION

In this study, some field crops produced in Turkey were classified according to 7 geographical regions by using the QUEST algorithm. In addition, the correct classification rate according to the regions was determined. The overall accuracy of this model is 53.2%. The most important effect in classification is the potato plant. While the number of cities with potatoes ≤ 2815.480 kg is 42, the number of cities with potatoes > 2815.480 kg is 35. Other important plants affecting the classification were barley, rye, chickpeas, oats and beans, respectively.

REFERENCES

Bag, M., Gauri, S. K., Chakraborty, S. (2012). Feature-based decision rules for control charts pattern recognition: A comparison between CART and QUEST algorithm. International Journal of Industrial Engineering Computations, 3:199-210

Chou, J. S., Cheng, M. Y., Wu, Y. W., Pham, A. D. (2014). Optimizing parameters of support vector machine using fast messy genetic algorithm for dispute classification. Expert Systems with Applications, 41:3955-3964

Çalış, A., Kayapınar, S., Çetinyokuş, T. (2014). Veri Madenciliğinde Karar Ağacı Algoritmaları ile bilgisayar ve internet güvenliği üzerine bir uygulama. Endüstri Mühendisliği Dergisi, 25(3-4):2-19

Darabi, H., Choubin, B., Rahmati, O., Haghighi, A. T., Pradhan, B., Kløve, B. (2019). Urban flood risk mapping using the GARP and QUEST models: A comparative study of machine learning techniques. Journal of Hydrology 569:142–154

Djordjevic, D., Cockalo, D., Bogetic, S., Bakator, M. (2021). Predicting Entrepreneurial Intentions among the Youth in Serbia with a Classification Decision Tree Model with the

QUEST Algorithm. Mathematics, 9:1487, https://doi.org/10.3390/math9131487

Ferri, C., Hernández-Orallo, J., Modroiu, R. (2009). An experimental comparison of performance measures for classification. Pattern Recognition Letters, 30, 27–38.

Kim, Y. S. (2010). Performance evaluation for classification methods: A comparative simulation study. Expert Systems with Applications, 37: 2292–2306.

Lim, T. S., Loh, W. Y., Shih, Y. S. (1997). An empirical comparison of decision trees and other classification methods. Madison: University of Wisconsin, Department of Statistics, Technical Report 979

Lin, C. L., Fan, C. L. (2019). Evaluation of CART, CHAID, and QUEST algorithms: a case study of construction defects in Taiwan. Journal of Asian Architecture and Buildıng Engineering, 18(6):539-553

Loh W. Y, Shih X. (1997). Split selection methods for classification trees. Statistica Sinica, 7: 815-840.

Loh, W. Y. (2014). Classification and regression trees. Institute for Mathematical Sciences, 10, 19.

Lutfi, H. M. A. (2020). Applying Decision Tree Algorithms to Develop Go/No Go Decision Model for Owners. Qatar University College of Engineering, Master thesis.

Maimon, O., Rokach, L. (2010). Data Mining and Knowledge Discovery Handbook. Classification Tree, pp. 149-174. Springer Science Business Media, LLC, New York, NY, USA, e-ISBN: 978-0-387-09823-4

Momenyan, S., Baghestani, A. R., Momenyan, N., Naseri, P., Akbari, M. E. (2018). Survival Prediction of Patients with Breast Cancer: Comparisons of Decision Tree and Logistic Regression Analysis. Int J Cancer Manag., 11(7):e9176.

Raeesi, A., Ebrahimi, S., Nia, L. I., Arji, G., Askani, M. (2014). An investigation of data mining techniques of the performance of a decision tree algorithm for predicting causes of traumatic brain injuries in Khatamolanbya Hospital in Zahdan city, 2012 to 2013. Journal of Health Management and Informatics, 1(2):28-30

Shamy, S., Dheeba, J. (2016). Review of QUEST, GUIDE, CRUISE, C4.5 and RPART Classification Algorithms. International Journal of Advanced Technology in Engineering and Science, 4(6):116-123

Soflaei, S. S., Shamsara, E., Sahranavard, T., Esmaily, H., Moohebati, M., Shabani, N., Asadi, Z., Tajfard, M., Ferns, G. A., Ghayour-Mobarhan, M. (2021). Dietary protein is the strong predictor of coronary artery disease; a data mining approach. Clinical Nutrition ESPEN, 43:442-447

Sokolova, M., Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. Information Processing and Management, 45:427-437

CHAPTER X

AN APPLICATION OF BOOTSTRAP RESAMPLING METHOD IN REGRESSION ANALYSIS ON CERTAIN RODENTS

Assist. Prof. Dr. Derviş TOPUZ[*]

[*] Niğde Ömer Halisdemir University, Niğde Zübeyde Hanım Vocational School of Health Services, Department of Health Services Science, Niğde, Turkey. Email: topuz@ohu.edu.tr

INTRODUCTION

In statistics, firstly observations are made on a population and data is collected in order to predict the parameter or parameters about that population. In cases where obtaining data by means of a complete count is possible, definite information on the feature may be achieved with the exception of certain compilation errors. However, carrying out a complete examination on entire population is often impossible. Even if a complete examination on population is possible, such examination often requires many operations, great amount of time and great numbers of qualified personnel. In addition, since assessment will take a long period of time, results acquired from the gathered data might become out of date and useless. Due to all these reasons given above, sample data is necessary with the exception of cases where complete count is absolutely necessary (Topuz, 2002). Validity of the data obtained from the sample depends on accurate selection of sample and statistical methods. Every sampling method has its pros and cons. Therefore, it is of great importance to select a sampling method to be the most appropriate for any kind of study.

In this research, if the value(s) of variable or variables vary depending on many factors that are beyond our control, coming to a general judgment with regards to such population becomes increasingly difficult. Especially in natural sciences, occurrence of many events is dependent on many factors that happen in a systematic scheme whose reasons are unknown to us and often beyond our control. Inability in fully understanding possible reasons for variation in data may result in inability in reducing sample errors to a desired level. The issue we wish to discuss here is on the efficiency of conventional or parametric sampling methods.

Sümbüllüoğlu and Sümbüllüoğlu (1987) listed some of the weak aspects of conventional sampling methods as follows:

- Applying an inaccurate sampling method,
- ignored (uninformed) selection of sample units,
- biased sample selection, knowingly or inadvertently,
- using some other subject as a substitute for a subject that could not been found,
- failure in desired level of effective data control,
- failure in generating a sample that is formed of desired number of individuals.

Surely, it is clear that teamwork under the control of expert individuals is essential in order to eliminate such defects. In parallel to the innovations in computer technologies, it is widely discussed in literature that more efficient and consistent forecasts can be made in some cases with use of bootstrap method that is often known as “resampling methods”(Efron,1982; Fox, 1997; Topuz, 2002).

Bootstrap sampling method, as a non-parametric approach to statistical methods, has the purpose of eliminating some weak aspects of conventional sampling methods (Efron,1982). Depending on the capability of the samples taken out of the representing population, bootstrap method is based on the principle of selecting sub-samples out of selected samples. Forecasting population parameters using such samples is then attempted. Such forecasted values are the accuracy criteria for the parameters of a model (Efron & Tibshirani, 1993).

In cases where a study had not been carried out on sample number of individuals, parametric predictions are erroneous and predictors fail in fulfilling certain theoretical features desired (such as accuracy, impartiality, efficiency etc). It must be kept in mind that some of these features are asymptotic features. Consequently, non-parametric sampling methods will minimize sampling errors of great number of random samples. It will generate powerful confidence intervals (Efron, 2003). Amongst the most widely applied methods is “bootstrap” resampling method.

Bootstrap resampling method is also being adapted for smallest square roots regression analysis in recent years. In smallest square roots regression analysis, certain hypothesis on observed values, variables and data are set forth. Calculations are only valid if and when such hypotheses are proven to be true. Otherwise, any process that has been correct in terms of algebra has no statistical value unless such hypotheses are proven to be true. Failure in proving the hypothesis true indicates failure in model’s capability of representation of the population and failure in a healthy generalization. Therefore, by using such models, the likelihood of failure of future predictions also become very high. Bootstrap sampling method will be useful in predicting parameters and calculating prediction errors without the need for hypothesis of smallest square roots analysis (Shao & Tu, 1995).

In this study, usage of bootstrap method in regression analysis will be shown. In sections two and three, introduction of bootstrap method and its use in regression analysis will be given, respectively. In section four, an application carried out with data on 39 *Spermophilus xanthoprymnus* species, both male and female is included. In addition, how the application was performed in S-Plus 2018 software suit is presented along with some screenshots.

2. MATERIAL AND METHOD

2.1. Material

In this study, data on 39 males and females mixed *Spermophilus xanthoprymnus* species, collected from Nigde region, was used as material (Niğde University Project no 98.FEB-07; Çakır, 2004). Data on Total Length (X_1), Hind Leg Length (X_2), Tail Length (X_3), dependent variable Live Weight (gr) (Y) of the species (*Spermophilus xanthoprymnus*) has been used and Table.1, Table.2, Table.3, Figure.1 and Figure.2 have been generated. Smallest square roots have been used in order to analyze the data generated by means of bootstrap resampling method (Wu, 1986; Liu, 1988; Shao, 1996; Fox, 1997). Correlations between the variables that take part in the relevant groups have been assessed by means of S-PUS 2018 for Windows statistical software suite.

Using the model $Y_i = \beta_0 + \beta_1 X_i + e_i$, effects of Total Length, Hind Leg Length and Tail Length of *Spermophilus xanthoprymnus* on Live Weight have been calculated by means of smallest square roots method and regression coefficient values obtained are (β_i) presented in Table.1. Bootstrap samples are analyzed by means of smallest square roots method. Only conventional sampling method (random) and bootstrap sampling method are discussed here. Assessments that were made in accordance with bootstrap technique in regression analysis are presented in Table.1. Results that had been generated by means of conventional sampling method in smallest square roots regression method were compared to results that were obtained by means of placing samples that had been generated by means of bootstrap resampling method in smallest square roots regression analysis method and are summarized in Table.1. It is attempted to demonstrate the applicability of bootstrap resampling method to examine the correlations between the several variables of *Spermophilus xanthoprymnus* species.

2.2. Method

To describe the resampling methods we start with an n sized sample $w_i = (Y_i, X_{ji})'$ and assume that w_i's are drawn independently and identically from a distribution of F, where $Y_i = (y_1, y_2, .., y_n)'$contains the responses, $X_{ji} = (x_{j1}, x_{j2}, \ldots, x_{jn})'$is a matrix of dimension nxk, where j = 1,2,...k, i = 1,2,3,...,n.

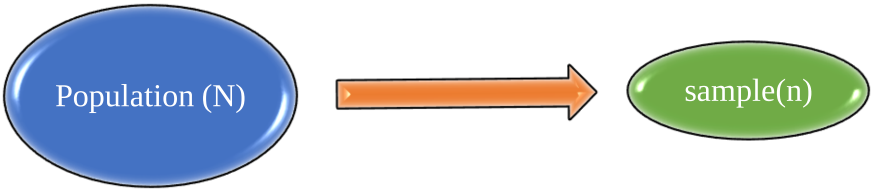
2.2.1. Bootstrap Resampling Technique

One of the most important purposes of statistical analysis is that the sample taken from the population must represent the population. The bootstrap resampling technique was developed by Efron (1979) as a general technique for assessing the statistical accuracy of an estimator. The main purpose of the technique is to calculate the predictive $\hat{\theta}$ value by choosing random samples with width n volumes, independent of a certain unknown distribution $f(x;\theta)$ and accept it as the predictor of the parameter θ. The bootstrap resampling technique is theoratically used to estimate values associated with the sampling distribution of estimators and test statistics.

Briefly, it is a technique of re-sampling in which n new samples, each having the same volume as the observed data, are obtained in return, when there is no information about the distribution of random samples drawn from the population (Friedl & Stampfer, 2002). A statistic is calculated for each of the new data sets, and the calculated statistics (mean, mode, median, proportion, odds ratio, correlation coefficient or regression coefficientetc) form a bootstrap distribution. The bootstrap technique are nonparametric and specific resampling technique whose purpose is calculate standard errors, confidence intervals for estimators of unknown parameters and p statistical values for hypothesis tests. The ordinary sampling techniques use some assumptions related to the form of the estimator distribution. These are the cases where standard assumptions are invalid, e.g. n volume is small, data contains uncertainity, data shows non-standard twist. In these situations, the use of these standard techniques may not give reliable and valid results. When these assumptions are doubtful or when the calculation of standard errors is necessary when parametric inference is impossible, the bootstrap resampling technique makes calculations without the need for these distributive assumptions because the sample population is considered (Efron, 1982). The results calculated by the estimators can be used as an experimental distribution for statistics (Efron,1990; Casella, 2003). The technique has been rarely used, although it is used to generate the estimation of the standard error of a statistic, confidence intervals and distributions by repeated use of the observed data (Efron, 1982; Efron & Tibshirani, 1998; Efron, 2003; Topuz, 2002).

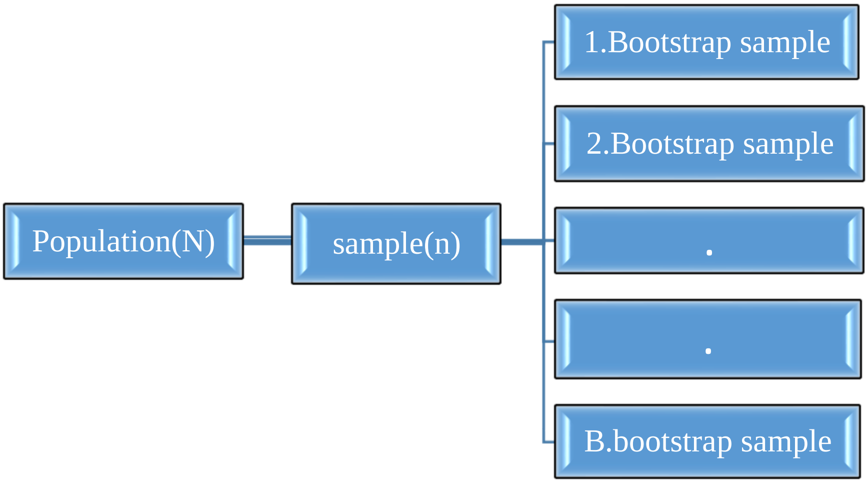
The main purpose of obtaining the sampling distribution of $\hat{\theta}$‘ which is the predictor of the population parameter, is to predict or test the population parameter. The obtaining the sampling distribution of the estimator is, if not impossible, extremely difficult and time consuming. Also, the bootstrap resampling technique proposed to obtain the experimental sampling distribution of the estimator removes this drawback. In line with this logic, the required bootstrap resampling algorithm can be defined as follows (Fox, 1997):

1) Obtaining a original sample of n volume from the population;



Calculation of estimators of population parameters using this example.

2) The main sample obtained is considered to be the only and best estimator of the population when there is no other information about the population. This main sample is considered as population. With a return selection each time, taking the probability of entering each observation into the sample as $\frac{1}{n}$, a sample of n volume is obtained and this process is repeated B times(Stine, 1990; Sahinler & Topuz, 2007).



Calculation of the related estimator for each bootstrap sample,

4) The obtaining the sampling distribution of these estimators based on B number of samples.

5) The obtaining parameter estimation values from this distribution with important estimators such as mean, standard bias and standard error,

6) As a result, the making comments about the population with using these estimates

As a result, as a good guide to the sampling distribution method, a bootstrap instance is created by recalculated resampling from the sample. The sample created is considered the population. This treated example is repeated many times to create an experimental distribution for the predictor. However, in the bootstrap technique, some data in the original sample can be selected more than once, some samples may not be selected at all.

The resampling number B is application dependent. In fact, although it is theoretically possible to create n^n bootstrap samples from a sample of n volume, this is both unnecessary and time consuming (Stine, 1990). The number of bootstrap repeats is generally suggested to be 50-200 to estimate the standard bias of $\hat{\theta}$ and to be at least 1000 within the estimate of confidence intervals of θ' (Efron, 1990; Leger et al., 1992). With the application of the bootstrap algorithm, the bias between population parameters and estimators will be reduced without increasing the sample size, and by obtaining the sampling distributions of the estimators, it will be provided to calculate the standard error of the estimators more accurately (Efron, 1979).

2.2.2. Bias Prediction of Bootstrap Distribution

If a predictor is biased and such bias is not known, definite predictions may not be made with regards to population parameter. Some predictors are unbiased. There are however, many cases where impartiality (lack of bias) hypothesis cannot be proven. It is known that predictors of most parameters are biased. Although working with such biased predictors brings about certain drawbacks, conventional sampling methods are not capable of providing clarification in such issue or of solving such issue. Bootstrap method is a non-parametric method that can be used both for generating sampling distribution of predictors and also for reduction of the bias of such predictors in order to overcome such drawbacks. Bootstrap bias prediction may be expressed as follows:

$$b^{*}(\hat{\theta})=\hat{\theta}-E(\hat{\theta}^{*})\ =\hat{\theta}-\bar{\hat{\theta}}_{B}^{*} \qquad (1)$$

In above, $\hat{\theta}$ represents the parameter predictor, $b^{*}(\hat{\theta})$ represents bootstrap bias prediction, $E(\hat{\theta}^{*})$ represents expected value of the bootstrap predictor, and $\bar{\hat{\theta}}_{B}^{*}$ represents the arithmetic average of the bootstrap predictors. In order to calculate the value of $\bar{\hat{\theta}}^{*}$, following formula will be made use of:

$$\bar{\hat{\theta}}_{B}^{*}=\frac{\sum_{b=1}^{B}\hat{\theta}_{b}^{*}}{B} \qquad (2)$$

In this formula bootstrap number b is indicated with the symbol $\hat{\theta}_{b}^{*}$. In addition, it must be remembered that number of repetitions (copies) is indicated with B (Birch, 1995; Topuz, 2002).

Efron stated that the value $b^{*}(\hat{\theta})$ was smaller than the difference between predictor and population value and that it was possible to reduce the bias all while predicting the bias.

2.2.3. Standard Error Prediction of Bootstrap Distribution

As the width of the samples drawn out of the distributions other than normal distribution widens, sampling distribution of the average approximates to normal. However, how much the predicted standard error approximates to its real value is not known definitely. Based on such opinion, bootstrap method argues that standard errors can be guessed more effectively originating from bootstrap distribution (Efron, 1981).

Bootstrap repetition number depends on previous experiences in data being processed. This number is recommended to be between 50– 200 for standard error predictions and at least 1000 for prediction of confidence interval of θ parameter. On the other hand, there are opinions of that the number of repetitions must be 100 for sample sizes between 10 and 80 (Efron, 1990; Topuz, 2002).

Bootstrap prediction of standard error is standard error of statistic’s bootstrap repetitions and calculated in the following equation:

$$s_B^* = \sqrt{\frac{\sum_{b=1}^{B} (\hat{\theta}_b^* - \overline{\hat{\theta}}_B^*)^2}{B-1}} \qquad (3)$$

2.2.4. Bootstrap Confidence Intervals

Sampling distribution of the predictor must be predicted in order to generate confidence intervals regarding population parameter. In conventional sampling methods, it is assumed that empirical distribution function $F(\hat{\theta})$ has normal distribution. In line with this assumption, confidence limits of population parameter are generated at $1\text{-}\alpha$ ambiguity level. While population variance is known, proposed confidence intervals for θ is (n > 30):

$$\bar{x} \pm z_{\alpha/2} \frac{\sigma}{n} \tag{4}$$

Such interval will contain population parameter at the level of 100% ambiguity (1-α) so far as the assumption on the form of sampling distribution. However, assumption made here may not be always true. In this case sampling methods will no longer be valid (Mooney & Duval, 1993).

Bootstrap confidence intervals is a non-parametric approach that can be used in order to eliminate the weak aspects of the sampling methods that were discussed above (Hall, 1986). It is an approach that bears great resemblance to normal method and the method that is used for determining parametric confidence intervals. Such approach is a method that is used in cases where normality hypothesis for a given statistics is true but it is difficult or impossible to find standard error analytically. Bootstrap confidence intervals is given as follows (Mooney & Duval, 1993):

$$\hat{\theta}^* \pm z_{\alpha/2} s^*_B \tag{5}$$

Under the hypothesis of normal distribution of sample values, a very small bootstrap repetition number, in other words B = 50 – 200, will be sufficient.

2.2.5. Bootstrap Resampling Method in Smallest Square Roots Regression Analysis

In this section, we introduce bootstrap resampling technique procedure. In general, regression technique for bootstrap is divided into two approaches: the first is based on the resampling observations' approach and the second is based on the resampling errors. Bootstrap technique based on resampling errors is known as more suitable for deterministic cases, whereas bootstrap resampling technique based on the drawing i.i.d. sample from the observation pairs is more appropriate for the case of random. But bootstrap resampling technique pairs can also be used for deterministic (Efron, 1982). The bootstrap is a "model-dependent" technique in terms of its implementation and performance although the bootstrap requires no theoretical formula for the quantity to be estimated and is less model-dependent than the traditional approach. In this paper, we show an application carried out the data on 39 *Spermophilus xanthoprymnus* using bootstrap technique based on the observation values. The bootstrap regression analysis procedure is as follows:

2.2.5.1. Bootstrapping Regression Algoritm

Here, two approachs for bootstrapping regression methods were given. The coise of either methods depends upon the regressors are fixed or random. If the regressors are fixed, the bootstrap uses resampling of the error term. If the regressors are random, the bootstrap uses resampling of observation sets $\mathrm{w_i}$. The bootstrap procedure based on the resampling of observation sets $\mathrm{w_i}$ follows (Sahinler & Topuz, 2007).

2.2.5.2. Bootstrap Based On The Resampling Observations

This approach is usually applied when the regression models built from data have regressors that are as random as the response. Let the (k+1)x1 vector $\mathrm{w_i = (Y_i, X_{ji})'}$ denote the values associated with ith observation. In this case, the set of observations are the vectors $\mathrm{(w_1, w_2, w_3, .., w_n)}$. The bootstrap procedure based on the resampling observations is as follows (Sahinler & Topuz, 2007).

1$^{(o)}$. Draw a n sized bootstrap sample ($\mathrm{w_1^{(b)}, w_2^{(b)}, w_3^{(b)}, \dots, w_n^{(b)}}$) with replacement from the observations giving $1/n$ probability each wi values and label the elements of each vector

$\mathrm{w_i^{(b)} = (y_i^{(b)}, x_{ji}^{(b)})'}$ where j = 1,2,..,k, i = 1,2,...n. From these form the vector

$\mathrm{Y_i^{(b)} = (y_1^{(b)}, y_2^{(b)}, \dots, y_n^{(b)})'}$ and the matrix $\mathrm{X_{ji}^{(b)} = (x_{j1}^{(b)}, x_{j2}^{(b)}, .., x_{jn}^{(b)})'}$

2$^{(o)}$. Calculate the OLS coefficients from the bootstrap sample:

$$\hat{\beta}^{(b1)} = (\mathrm{X^{(b)\prime}X^{(b)}})^{-1}\mathrm{X^{(b)\prime}Y_i^{(b)}} \qquad (6)$$

3$^{(o)}$. Repeat steps 1 and 2 for r = 1,2,.., B, where B is the number of repetition.

4$^{(o)}$. Obtain the probability distribution $\left(F\left(\hat{\beta}^{(b)}\right)\right)$ of bootstrap estimates $\hat{\beta}^{(b1)}, \hat{\beta}^{(b2)}, \dots, \hat{\beta}^{(bB)}$ and use the $\left(F\left(\hat{\beta}^{(b)}\right)\right)$ to estimate regression coefficients, variances and confidence intervals as follows. The bootstrap estimate of regression coefficient is the mean of the distribution $\left(F\left(\hat{\beta}^{(b)}\right)\right)$ (Fox,1997),

$$\hat{\beta}^{(b)} = \sum_{b=1}^{B} \hat{\beta}^{(br)} \Big/ B = \bar{\hat{\beta}}^{(br)} \qquad (7)$$

5$^{(o)}$. Thus, the bootstrap regression equation is

$$\hat{Y}_i = f\left(\hat{\beta}, x\right) = \hat{\beta}_0^{(b)} + \hat{\beta}_1^{(b)} x_{j1}^{(b)} + \hat{\beta}_2^{(b)} x_{j2}^{(b)} + \cdots + \hat{\beta}_s^{(b)} x_{jn}^{(b)} + \varepsilon \qquad (8)$$

where $\hat{\beta}^{(b)}$ is unbiased estimator of β (Shao, 1995; Sahinler & Topuz, 2007).

An illustrative example that presents how the regression parameters are estimated from the bootstrap based on the the resampling observations was given in Table 1.

2.2.5.3 The bootstrap bias, variance, confidence and percentile interval

The bootstrap bias equals,

$$\widehat{bias_b} = \hat{\beta}^{(b)} - \hat{\beta} \qquad (9)$$

(Further discussion are described in Efron and Tibshirani, 1993). The bootstrap variance from the distribution $\left(\mathrm{F}\left(\hat{\beta}^{(b)}\right)\right)$ are calculated by (Sahinler & Topuz, 2007).

$$\mathrm{var}\left(\hat{\beta}^{(b)}\right) = \sum_{i=1}^{B}\left[\left(\hat{\beta}^{(br)} - \hat{\beta}^{(b)}\right)\left(\hat{\beta}^{(br)} - \hat{\beta}^{(b)}\right)'\right]/(B-1)\,,r$$
$$= 1,2,..,B \ (10)$$

The bootstrap confidence interval by normal approach is obtained by

$$\left(\hat{\beta}^{(b)} - t_{n-p,\ \frac{\alpha}{2}} * S_e\left(\hat{\beta}^{(b)}\right) < \beta < \hat{\beta}^{(b)} + t_{n-p,\ \frac{\alpha}{2}} * S_e\left(\hat{\beta}^{(b)}\right)\right) = 1-\alpha \quad (11)$$

where $t_{n-p,\ \frac{\alpha}{2}}$ is the critical value of t with probability α/2 the right for n-p degrees of freedom; and $S_e\left(\hat{\beta}^{(b)}\right)$ is the standard error of the $\hat{\beta}^{(b)}$. If sample size is n ≥30, then Z distribution values are used instead of t in estimation of confidence intervals (Diciccio & Tibshirani, 1987).

A nonparametric confidence interval named percentile Interval can be constructed from the quantiles of the bootstrap sampling distribution of $\hat{\beta}^{(b)}$. The (α/2)% and (1-α/2)% percentile interval is

$$\hat{\beta}^{(br)}_{(lower)} < \beta < \hat{\beta}^{(br)}_{(upper)} \qquad (12)$$

where $\hat{\beta}^{(br)}$ is the ordered bootstrap estimates of regression coefficient from Equation 9 or 10, lower = (α/2)B, and upper = (1-α/2)B.

3. RESULTS

In this study, data on 39 males and females mixed *Spermophilus xanthoprymnus* species, collected from Nigde region, has been used as material (Table.3).

Current data were deemed to have been the data pertaining to the population. Then, a regression model was generated by means of smallest square roots method and parameters were calculated. Later, comparison is made in order to find out whether these parameters would better be predicted by using predictors obtained through classic sampling methods or by using predictors obtained by means of a resampling method, namely bootstrap sampling method. Samples of n=20 have been selected on random basis from the same data and statistics pertaining to parameters were calculated (Table.1). This data was then used as the sample, and samples have been generated by means of bootstrap sampling method that had been discussed in method section (Table.3). Samples pertaining to error terms that had been generated for bootstrap, regression line of smallest square roots pertaining to such data, and relevant parameters were predicted and results have been compared for each data group.

Arithmetic average and standard deviation of *Spermophilus xanthoprymnus* species that had been obtained by means of classic sampling method have been 230,63 ∓ 18,65 and distribution interval has been (2.1396)-(3.9442) mm. As for the values pertaining to Hind Leg Length have been 37.52 ± 1.79 mm. Tail Length has been 38,19 ∓ 5,04 mm. Histogram graphs that indicate the average values and distribution interval pertaining to Total Length, Hind Leg Length (X_2), Tail Length (X_3), and Live Weight (gr) of species that had been obtained by means of bootstrap sampling method have been shown in Figure.1.

3.1. Regression Analysis Results Pertaining to the Sample Generated by Means of Bootstrap Resampling Method

Bootstrap algorithm based on observation values has been applied to data pertaining to 39 male and female *Spermophilus xanthoprymnus* as follows:

Table 1. Data Set Pertaining to Population (N=39)

No	**Live Weight (Y)gr**	**Total Length (X_1)**	**Hind leg length (X_2) mm**	**Tail length (X_3) mm**
1	300	244,30	38,50	34,10
2	175	212,30	38,55	36,15
3	275	254,60	38,30	40,10
4	375	272,55	41,60	49,15
5	125	191,70	33,70	31,50
6	250	233,50	42,80	35,00
.	.	.	.	.
35	250	226,10	37,10	33,30
36	200	223,05	33,55	42,7
37	225	235,70	38,65	41,15
38	200	230,9	37,40	32,70
39	125	215,35	37,25	34,15

LW: Live Weight (gr), TL: Total Length (mm), HLL: Hind Leg Length, TU: Tail Length (mm)

$1^{(o)}$. Sample of unit n=20 has been selected out of the population.

$2^{(o)}$. Smallest square roots regression line that belongs to the sample selected is generated.

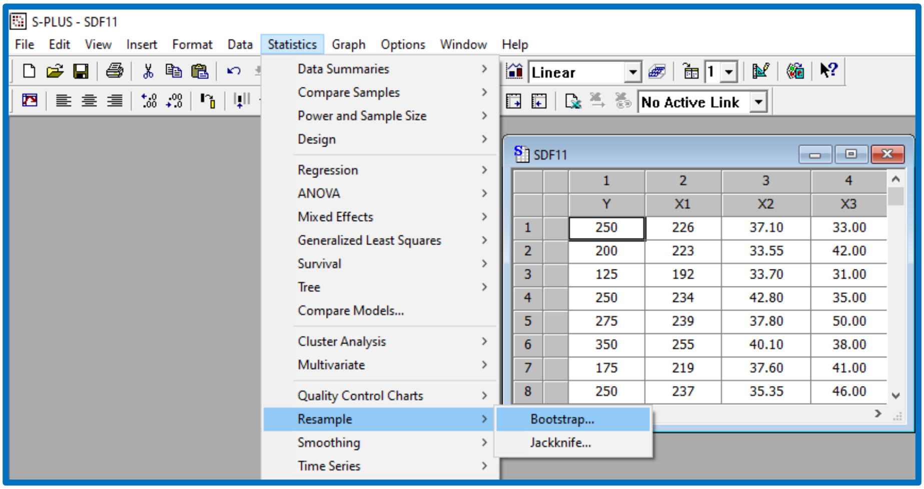
3$^{(o)}$. Repeat steps 1 and 2 for r = 1,2,...,100, where 100 is the number of repetition.

4$^{(o)}$. Giving 1/20 probability for each observation value, 100 instances of bootstrap observation samples made of 20 volumes have been generated with the help of “sampling” module of EXCEL software.

Table 2. The illustration of the bootstrap (B=100 bootstrap samples, each of size n=20) regression procedure from the data given in Figure 1, calculating the bootstrap estimates of the regression parameters for each sample for Spermophilus xanthoprymnus Live Weight model

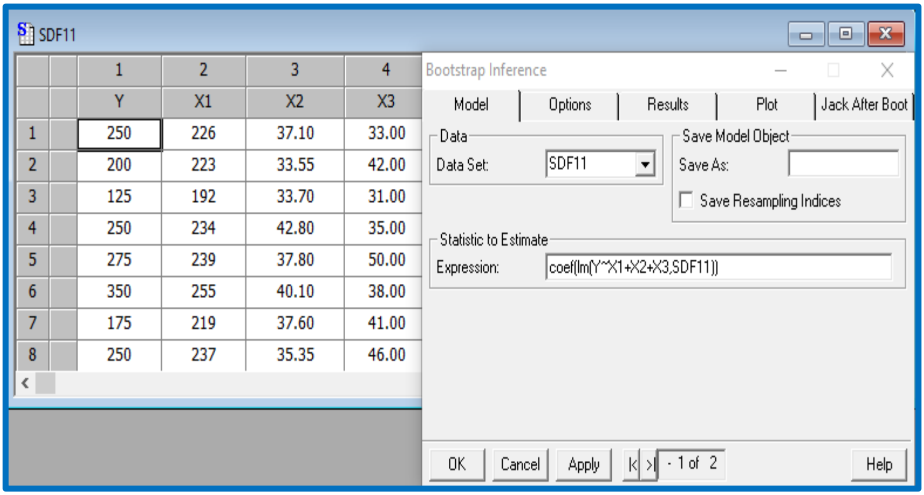
		Observation sets						**OLSBR**			
r	**Variables**	$w_1^{(b)}$	$w_2^{(b)}$	$w_3^{(b)}$	$w_4^{(b)}$	.	$w_{20}^{(b)}$	$\hat{\beta}_0^{(b)}$	$\hat{\beta}_1^{(b)}$	$\hat{\beta}_2^{(b)}$	$\hat{\beta}_3^{(b)}$
1	LW(Y)	300	175	275	375	.	250	−543.2	4.20	-2.06	-2.63
	TL(X_1)	244.3	212.3	254.6	272.55	.	236.8				
	HHL(X_2)	38.50	38.55	38.30	41.60	.	35.35				
	TU(X_3)	34.10	36.15	40.10	49.15		46.60				
2	LW(Y)	375	125	350	250	.	175	.	.	.	.
	TL(X_1)	272.5	191.7	254.6	236.8	.	212.3				
	HHL(X_2)	41.6	33.7	40.10	35.35	.	38.55				
	TU(X_3)	49.1	31.5	38.1	46.60		36.15				
3	LW(Y)	200	375	375	125	.	350	.	.	.	.
	TL(X_1)	233.0	272.5	272.55	254.6	.	254.6				
	HHL(X_2)	33.55	41.6	41.60	38.30	.	40.10				
	TU(X_3)	42.7	49.1	49.15	40.10		38.1				
	.	.	.	.	.	.	.	.	.	.	.
100	LW(Y)	175	200	175	250	.	375	.	.	.	.
	TL(X_1)	218.8	223.0	212.3	236.8	.	272.55				
	HHL(X_2)	37.6	33.5	38.55	35.35	.	41.60				
	TU(X_3)	41.6	42.7	36.15	46.60		49.15				
	$\hat{\beta}^{(b)} = \sum_{b=1}^{100} \hat{\beta}^{(br)} \Big/ 100\,,\quad \tilde{\beta}^{(b)} = \sum_{b=1}^{100} \tilde{\beta}^{(br)} \Big/ 100 = \bar{\bar{\beta}}^{(br)}$							-571,88	3,54	1,17	-1,41

Bootstrap error predictors of these samples that had been generated have been calculated by means of relevant equations.



		1	2	3	4
		Y	X1	X2	X3
1		250	226	37.10	33.00
2		200	223	33.55	42.00
3		125	192	33.70	31.00
4		250	234	42.80	35.00
5		275	239	37.80	50.00
6		350	255	40.10	38.00
7		175	219	37.60	41.00
8		250	237	35.35	46.00

Figure 1. Statistics Dialog Box



		1	2	3	4
		Y	X1	X2	X3
1		250	226	37.10	33.00
2		200	223	33.55	42.00
3		125	192	33.70	31.00
4		250	234	42.80	35.00
5		275	239	37.80	50.00
6		350	255	40.10	38.00
7		175	219	37.60	41.00
8		250	237	35.35	46.00

Data

Data Set: SDF11

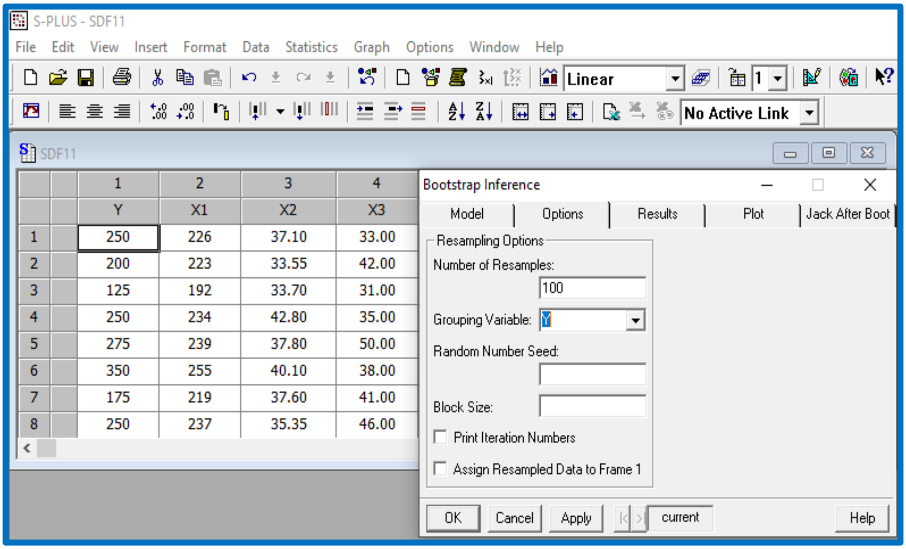
Save As:

☐ Save Resampling Indices

Expression: coef(lm(Y~X1+X2+X3,SDF11))

Figure 2. Data Dialog Box

Bootstrap observation predictors obtained have been placed in relevant equations and bootstrap Y values (Y_i*)) have been calculated.



		1	2	3	4
		Y	X1	X2	X3
1		250	226	37.10	33.00
2		200	223	33.55	42.00
3		125	192	33.70	31.00
4		250	234	42.80	35.00
5		275	239	37.80	50.00
6		350	255	40.10	38.00
7		175	219	37.60	41.00
8		250	237	35.35	46.00

Resampling Options

Number of Resamples: 100

Grouping Variable: Y

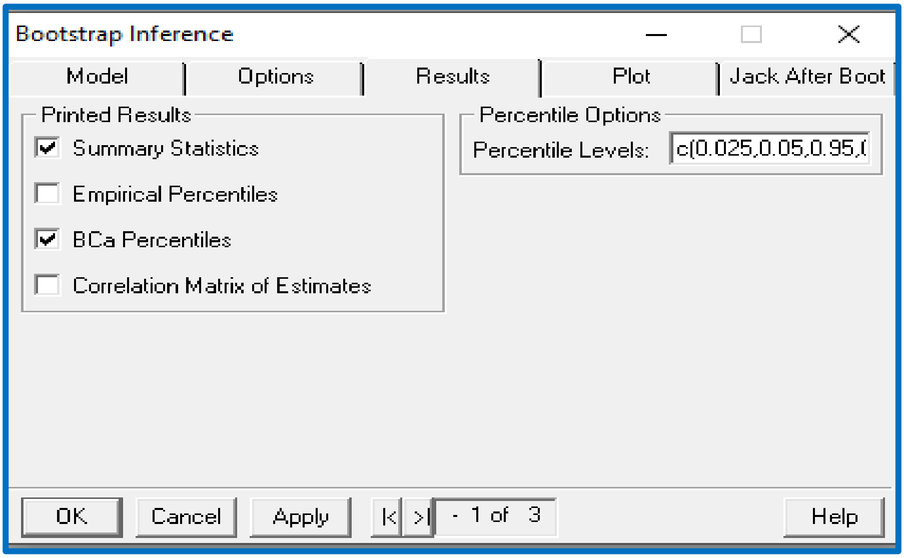
Random Number Seed:

Block Size:

□ Print Iteration Numbers

□ Assign Resampled Data to Frame 1

Figure 3. Resampling Options Dialog Box



Printed Results

☑ Summary Statistics

□ Empirical Percentiles

☑ BCa Percentiles

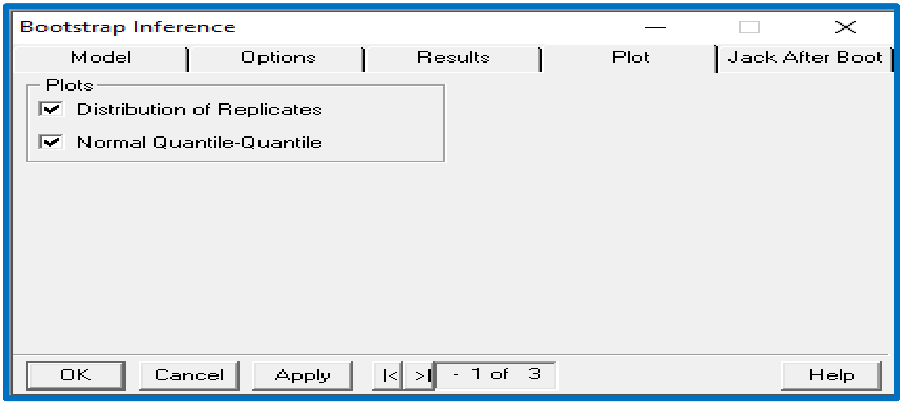
□ Correlation Matrix of Estimates

Percentile Options

Percentile Levels: c(0.025,0.05,0.95,(

- 1 of 3

Figure 4. Printed Results Dialog Box



Plots

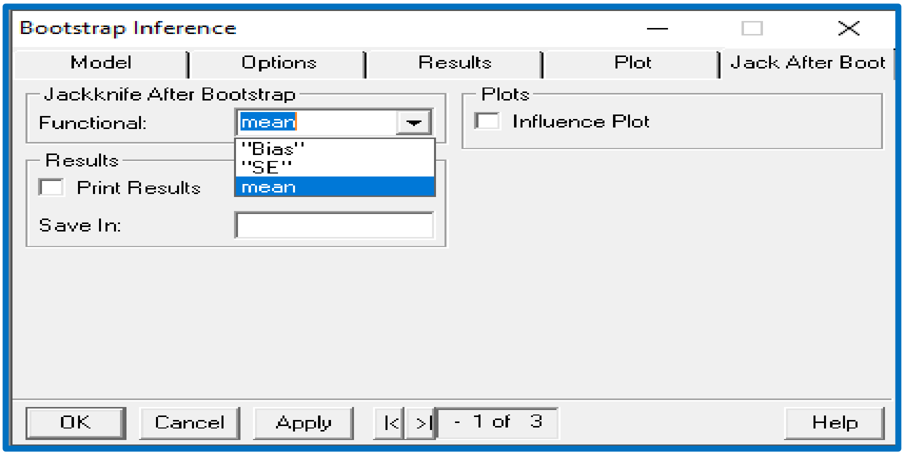
☑ Distribution of Replicates

☑ Normal Quantile-Quantile

Figure 5. Plots Dialog Box

5$^{(o)}$. By means of new observation points generated, parameters are predicted by smallest square roots method and predicted values on bootstrap regression coefficients have been calculated as follows:

$$\hat{Y}^{*} = -571{,}879 + 3{,}537(X_1) + 1{,}177(X_2) - 1{,}414(X_3) \tag{13}$$



Jackknife After Bootstrap

Functional: mean

Results

□ Print Results

Save In:

Plots

□ Influence Plot

Figure 6. Jackknife After Bootstrap Functional Dialog Box

```
*** Bootstrap Results ***
Call:bootstrap(data = SDF11, statistic = coef(lm(Y ~ X1 + X2 + X3, SDF11)), B =
          100, group = Y, trace = F, assign.frame1 = F, save.indices = F)
Number of Replications: 100 Summary Statistics: Values calculated have been
summarized in Table-3.
```

Table 3. Certain Definitive Values Pertaining to Sample Groups Examined

Methods			Variables			
			Live Weight (Y) (gr)	**Total Length (X_1) mm**	**Hind Leg Length (X_2) mm**	**Tail Length (X_3) mm**
Population (N=39)		β	-538,56	3,042	1,45	0,57
		Std. Error	112,33	0,46	2,69	1,38
		t	-4,79	6,60	0,54	0,42
		Pr(>|t|	0,00	0,00	0,59	0,67
		Interval	(-758.7)-(-318.4)	(2.13)-(3.94)	(-3.8)-(6.7)	(-2.12)-(3.28)
Conventional Sample	**n=20**	$\hat{\beta}$	-554,48	3,48	1,04	-1,44
		Std. Error	173,98	0,53	4,10	1,62
		t	-3,18	6,48	0,25	-0,88
		Pr(>|t|	0,005	0,00	0,80	0,38
		Interval	(-895.4)-(-197.5)	(2.44)-(4.58)	(-6.9)-(9.0)	(-4.79)-(1.73)
Bootstrap Values	**n=20**	$\hat{\beta}^{*}_{ort}$	-571,88	3,54	1,177	-1,41
		Deviation	-17,46	0,05	0,14	0,026
		Std. Error	98,52	0,24	1,56	1,17
		t	22.04	0.05	0.35	0.26
		Interval	(-928,32)-(-167,6)	(2,308)-(4,66)	(-6,1)-(10,5)	(-4,41)-(1,75)

CONCLUSION

In this study, different samples from two different (namely classic sampling and resampling (Bootstrap)) methods were gathered to be used in smallest square roots regression analysis in order to point out with which one of these more accurate parameter predictions can be made. As a result of the study on *Spermophilus xanthoprymnus* species, it has been pointed out that bootstrap resampling method has minimum deviation and smaller standard error value in comparison to classic sampling method and that bootstrap resampling method generates more reliable confidence intervals, enabling more efficient prediction of parameter values pertaining to the population. It is attempted to point out that greater advantages in terms of the cost by means of easier calculation of statistical values in shorter amount of time without any negative effects have been possible (**Figure 5**). It is recommended for further researches that researchers can use bootstrap sampling method instead of random sampling so the selected sample may represent the population better because it is pointed out that studies carried out in this manner will generate more efficient results.

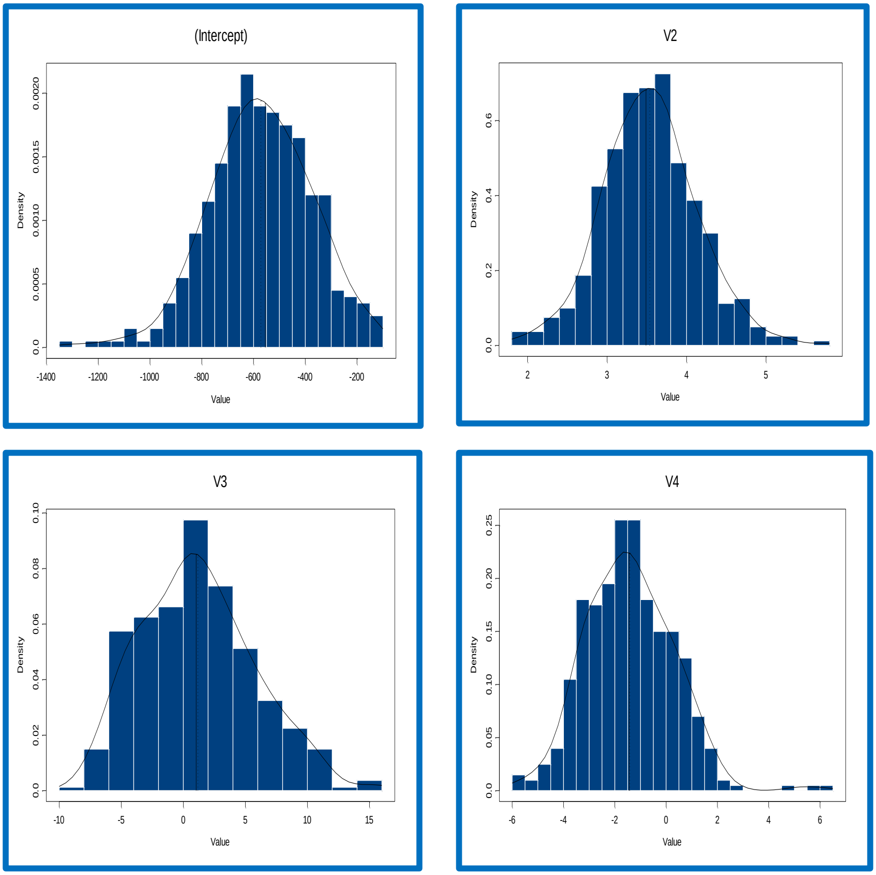


Figure 5. Graphic Demonstrating the Distribution of Averages of Live Weight (Y), Total Length (V2), Hind Leg Length (V3) and Tail Length (V4)

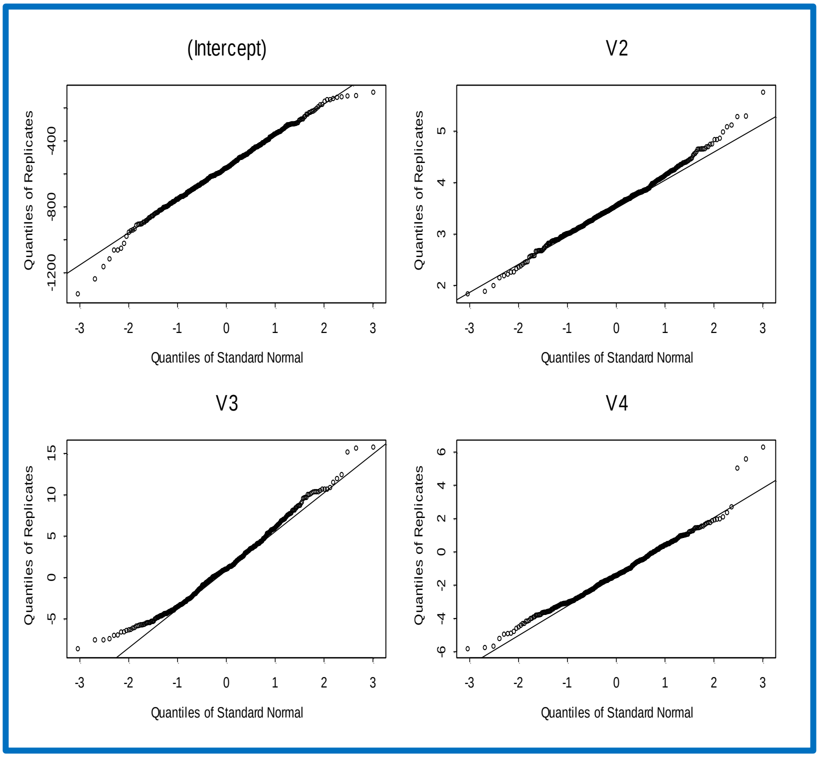


Figure 6. Graphic Demonstrating The Distribution of Error Terms of Live Weight (Y), Total Length (V2), Hind Leg Length (V3) and Tail Length (V4)

ACKNOWLEDGEMENTS

We thank Prof. Dr. Mehmet Çakır for all data used in this study (The project named Histo-anatomy of Spermophylus xantoprymnus living in Niğde Region. Niğde University Project no 98.FEB-07)

REFERENCES

Birch, J.B. (1995). Exploratory and Robust Data Analysis Using MINITAB, Virgini Tech, Blachsburg,VA.pp.3-13.

Casella, G. (2003). Introduction to the silver anniversary of the bootstrap, Statistical Science, Vol .2, No.18, pp 133-134.

Çakır, M. (2004). Morphometric and Hysto-Anatomic Researches on Partes Genitales Femininae Internae of Female. *Spermophilus xanthoprymnu* (Rodentia: Sciuridae) of Turkey. Известия Вузов (İzvestiya VUZOV). No (Vıp): Vol.4, pp 74-81.

Diciccio, T. & Tibshirani, R. (1987). Bootstrap confıdence intervals and bootstrap approximations. JASA. Vol. 82, pp 163-170.

Efron, B.(1979). Bootstrap methods: Anoher look at the Jackknife. Ann. of Stat. Vol.7, pp 1-26.

Efron, B.(1981). Censored Data and Bootstrap. JASA, Vol.76, No.374. pp.313-315.

Efron, B. (1982). The jackknife, the bootstrap and other resampling plans, Society for industrial and applied mathematics, Vol. 38, pp 29-35.

Efron, B.(1990). More efficient bootstrap computations. JASA. Vol. 55, pp 79-89.

Efron, B. & Tibshirani, R.(1993). An Introduction to the Bootstrap. Chapman and Hall, New-York.

Efron, B.&Tibshirani, R.J. (1998). An ıntroduction to the bootstrap, Library of Congress Cataloging in Publication Data, CRC Press, Boca Raton, Florida, pp 60-82.

Efron, B.(2003). Second thought on bootstrapping, Statistical Science, Vo.2, No.18, pp 135-140.

Friedl, H. & Stampfer, E.(2002). Resampling methods. In Encyclopedia of Environmetrics. Wiley: Chichester, Vol.3, pp, 1768-1770.

Fox, J.(1997). Applied regression analysis, linear models, and related methods, Sage Publications. London. pp 494-520.

Hall, P.(1986). On The Bootstrap and Confidence Intervals. The Ann. of Stat., Vol.14, No.4, pp.1431.

Liu, Y. R.(1988). Bootstrap Procedures Under Some Non-I:I.D. Models. Ann. of Stat., Vol.16, No. 4 ,pp 1706.

Leger, C., Politis, D.N & Romano, J.P. (1992). Bootstrap technology and applications. Technometrics. Vol. 34, pp 378-397.

Mooney, C.Z.& Duval, R.D. (1993). Bootstrapping: A nonparametric approach to statistical inference. Sage pub. Inc. London. pp. 80

Sümbülloğlu, K. & Sümbülloğlu, V. (1987). Biyoistatistik 7. Baskı, Hatiboğlu Yayınevi, Ankara.

Stine, R.(1990). Modern methods of data analysis. Sage Pub. Inc. Scotland. pp 325-373.

Sahinler, S.& Topuz, D. (2007). Bootstrap and Jackknife resampling algorithms for estimation of regression parameters, Journal of Applied Quantitative Methods, Vol.2, pp 188-199.

Shao, J.& Tu, D.(1995). Applications to time series and other dependent data, The Jackknife and Bootstrap. Springer Series in Statistics. Springer, New York, NY, PP 386-415.

Shao, J.(1996). Bootstrap model selection. JASA. Vol. 91, pp 655-665.

Topuz, D.(2002). Regresyonda Yeniden Örnekleme Yöntemlerinin Karşılaştırmalı olarak İncelenmesi. Yüksek lisans Tezi, Niğde Üniversitesi Fen Bilimleri Enstitüs, p.68. Niğde.

Wu, C.F.J.(1986). Jackknife, Bootstrap and Other Resampling Methods in Regression Analysis. Ann. of Stat., Vol.4, No.14, pp 1261-1265.

CHAPTER XI

THE NITRATE STATUS OF THE WATERS IN THE CENTER OF BİNGÖL

Prof. Dr. Ali Rıza DEMİRKIRAN[*]
Dr. Mehmet ULUPINAR[†]
Prof. Dr. Alaaddin YÜKSEL[*]
Kadir Miraç DAŞBİLEK[‡]

[*] Bingol University, Faculty of Agriculture, Department of Soil Science and Plant Nutrition, Bingol, Turkey, ademirkiran@bingol.edu.tr; ayuksel@bingol.edu.tr

[†] Bingol University, Faculty of Agriculture, Department of Biosystem, Bingol, Turkey, mulupinar@bingol.edu.tr

[‡] Ministry of Agirculture and Forestry, Bingol, Turkey

INTRODUCTION

It is known that the water quality is almost and always associated with all of living creators in the world. The water quality is also related by what people do and some climatic factors in the world.

Concern about water quality for using such as domestical, agricultural and industrial is old matter and standarts for its evaluation have existed since late 1800's. The various users of water have different quality parameters for the same water supplies (Rhoades 1972).

It was devepoled that the criteria of the water quality, for examle pH, temperature, phosphates, nitrates, dissolved oxygen (DO), solid materials (Chaturvedi and Bassin 2009). NH_4-N, NO_3-N, DO, and turbidity are the most used and determined paramaters. The minimum water quality index (WQI) was determined based on a stepwise lineer regression analysis, consisted of 5 parameters as NH_4-N, NO_3-N, permanganate index (COD_{Mn}), DO, and turbidity (Wu at al. 2018). Nutrients such as nitrogen (nitrite, ammonia, and nitrate) and phosphorus are most important parameters for water quality (Şener et al. 2013). Nitrate (NO_3; molecular weight, 62,008; valence 1) contain generally trace concentrations in surface waters.

Some illness (e.i. blue babies, methemoglobinemia, gastric carcinomas, abnormal pain, nervous system defects, diabetes) may occur by using of water with high nitrate concentration (Vasanthavigar et al. 2010; Varol and Davraz 2015).

Nitrogen in soils and waters takes place in the chemical, physical and biological reactions. Generally, sources of the contaminated waters can be attributed to having septic materials, animal and human wastes, and chemical fertilizers (Keeney and Olson 1986).

The nitrate levels varied from 1.10 to 1.59 for the 6 river systems as total 384 water samples in the Lake Taihu Basin, China (Wu at al. 2018). It was indicated that the high nitrate levels (up to 56 mg L^{-1}) in the central Wisconsin sand plains are significantly above background in the irrigated region (Saffigna and Keeney 1977). A study was evaluated changes of NO_3 in ground water samples of USA in predominantly agricultural areas during 1988–2004. The concentrations of NO_3 (in 495 wells; three of those eight wells) increased according to the USEPA maximum contaminant level of 10 mg L^{-1} (Rupert 2008).

In the Walnut Creek watershed, USA, nitrate levels of surface water were founded between 3.1 and 29.3 mg L^{-1} that often exceeded 10 mg L^{-1} during May, June, and July, while decreasing during September and October months. During runoff events, it was indicated that NO_3 concentrations decreased (Jaynes et al. 1999).

Nitrate concentrations increased in U.S. Rivers strongly during 1945-1980 at most of the stations (Goolsby and Battaglin 2001; Smith et al. 2003; Robinson et al. 2003; Stets et al. 2015; Broussard and Turner 2009). In the Midwest, Eastern and Western U.S, nitrate levels increased to a greater extent. In Maumee River, the highest concentration of the flow-weighted concentration NO_3-N (FWC NO_3-N)as 7.4 and 9.1 mg N L^{-1} was founded during 1975-1985 and 1998-2008, respectively (Stets et al. 2015).

The concenteration of NO_3 was founded that the variety between 1.35 and 6.42 mg L^{-1} in the Aksu River, located Isparta and Antalya (Şener et al. 2017). Saraydüzü Dam Lake (Sinop), the low nitrate concentration (BDL) showed that a declining from autumn to winter (from October to February) was determined while high level (3.62 mg L^{-1}) in early autumn (September). According to the Intra-Continental Surface Water Resources Regulation of Turkey (SWOR), the nitrate quality of Saraydüzü Dam Lake was found in the criteria ‘very good’ (Kükrer and Mutlu 2019). According to the Varol et al. (2012), the nitrate concentrations of Kralkızı, Dicle and Batman Dams on the Tigris River were higher in winter and spring with comparing to autumn and summer. This may be interpreted due to stream inputs and surface runoff, which carry more nutrients into the reservoirs during the rainy season (winter and spring).

In the study of Duran and Suicmez (2007), nitrate levels of drinking water in city of Bitlis, were investigated. The mean nitrate levels were found 5.68±0.31 mg L^{-1} ($p<0.05$) in accordance with recommended standards (Alemda et al. 2009). The nitrate levels of Cekerek stream, Tokat, were found in the autumn, winter, and spring, summer at the ranges of 2.5-4.5, 2.1-3.1, 3.3-4.3, and 5.0-12.3 mg L^{-1}, respectively. Korkanç et al (2017) observed the high nitrate values at the Karasu stream (21.3 mg L^{-1}) and Akkaya Dam, Niğde (64.8 mg L^{-1}). It was also observed that seasonal nitrate values increased in spring and decrease in autumn. Korkanç et al (2017) stated that nitrate levels increased during autumn to spring season with effecting of the fertilizing at the spring season in Karasu areas.

By the study of Yolcubal et al. (2016), nitrate contents of Dil stream, İzmit, were determined between 0.1 and 50.6 mg L^{-1}. The high value was impied the impact of uncontrolled industrialization in the Gulf of İzmit. By Tanrıverdi et al. (2010) were shown that nitrate levels of Ceyhan River, Kahramanmaraş, were range of 0.01 in Fırnız stream - 22.2 mg L^{-1} in Aksu stream. Nitrate levels of Murat River, part of the Euphrates River, were determined as "first-class water quality" according to the “Regulation on Surface Water Quality Management (SWOR)” criteria by Koyun et al. (2020). The evidence has interpreted that there is not significant nitrate problem in the Murat River. The lowest nitrate was 1.52 mg L^{-1} in autumn, and the highest nitrate was 3.41 mg L^{-1} in winter. Nitrate was seasonally reached the highest value in spring (2.83 mg L^{-1}).

The aim of this study is to know the criteria of nitrate levels in the surface water of Bingol, Turkey.

1. MATERIAL AND METHOD

The surface waters were collected from 13 stations in 2019 + 12 months in Bingol, Turkey, (Table 1).

Table 1. The surface water station in Bingol, Turkey

No	Station	Location
1	Gayt	Gayt water, Kaleönü, Industry road, Center-Bingol
2	Ekinyolu	Ekinyolu bridge, Bingol-Muş Road, Center-Bingol
3	Garip	Garip village, Garip bridge, Center-Bingol
4	Genç	Genç bridge, Genç-Bingol
5	Sarıçiçek	Sarıçiçek village, Gülbahar stream, Center-Bingol
6	Şeyh Ahmet	Şeyh Ahmet stream, Bingol-Genç Road, Genç-Bingol
7	Leymalun	Leymalun bridge, Arakonak, Solhan-Bingol
8	Masala	Dilektepe village, Masala bridge, Solhan-Bingol
9	Paymerg	Paymerg village, Ekinyolu, Center-Bingol
10	Beyaztoprak	Beyaztoprak, Center-Bingol
11	Çeltiksuyu	Çeltiksuyu village, Center-Bingol
12	Elmalı	Elmalı village, Elmalı bridge, Center-Bingol
13	Alatepe	Alatepe village, Elmalı bridge, Center-Bingol

Nitrate concentrations in water samples were determined with using spectrophotometer (MN-1911) and its reagents. The nitrates criteria in the water determined by SWOR (2016) was shown in Table 2.

Table 2. The nitrates criteria of surface water by the Turkish standards

Parameter	Water quality classes				
	I (very good)	II (good)	III (medium)	IV (poor)	Reference
Nitrate nitrogen (mg NO_3^{-}N L^{-1})	< 3	10	20	> 20	SWOR (2016)
Nitrate (mg NO_3^{-}N L^{-1})	< 5	10	20	> 20	TSE (1988)
Nitrate (mg NO_3^{-} L^{-1})	22	44	88	> 88	TSE (2005)

Random Blocks Trial Arrangement: If the trial material differs in a feature, the trial must be arranged in the Random Blocks trial layout. The plants to be used for a trial may differ in terms of variety, trees variety or age, animals’ race or age, field water holding capacity. In this case, the trial material is divided into homogeneous parts in terms of different properties. Each of the homogeneous parts created is called a block. Random blocks trial layout eliminates variation in trial material. For this reason, it reduces trial error and provides more reliable results. Since each treatment must be tried in each block, each treatment is repeated an equal number of times, ie the number of repetitions in each treatment is equal. For this reason, the analysis of the obtained data is easy. Despite the positive aspects of the random blocks trial scheme, there are also negative aspects. Missing observation causes difficulties in analyzing data, as every transaction must be tried in every block. If the trial material differs in two ways in terms of the effects of the investigated factors, this trial is not an appropriate and effective scheme. The point to be considered when choosing a randomized block trial setup as a trial setup is that as the number of treatments to be investigated increases, the size of the blocks will increase and their homogeneity will deteriorate. Random blocks are used in cases where it is necessary to reduce the variation in one direction in the trial layout (Düzgüneş, 1987). The mathematical model in the random blocks experiment is as follows:

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

Here, μ: population mean, α_i: group influence, β_j: block influence, ε_{ij}: margin of error.

If the difference between groups is observed as statistically important as a result of variance analysis, multiple comparison tests are applied to determine between which groups this difference is. One of these tests is the Tukey test. The Tukey test was proposed by Tukey (1953) to enable the use of a "studertized range of variation". The Tukey test is calculated as follows:

$$W = q\alpha * \sqrt{(MSE/r)}$$

or

$$HSD = q\alpha * \sqrt{(MSE/n)}$$

Here, *HSD or W*: Honestly Significant Difference, $q\alpha$: Scale value for Tukey W test (Error SD; Operation SD; probability value), r or n: Number of recurrences, MSE: Error Mean Squares.

If the critical value is too high, it may result that the mean differences, which are actually important in the Tukey test, are insignificant (Yıldız and Bircan, 1991).

2. RESULTS

The nitrates levels (mg L^{-1}) after analysis of surface water in Bingol are given in Table 3 and Figure 1 and 2. In this research, it is observed that nitrate concentration between 0.1 and 15 mg L^{-1} in surface water of Bingol. According to the months, the high levels of nitrate were in the December, February, March and April months, while the low levels of nitrate in the June, July, August, October and November of 2019 year. According to the stations, the high levels of nitrate were in the Leymalun, Masala and Beyaztoprak, while the low levels of nitrate in the Gayt, Ekinyolu and Genç stations.

Table 3. The average nitrate levels of surface water in Bingol

Location	Months (2019)											
	Jan.	Feb.	Mar.	Apr.	May.	Jun.	Jul.	Aug.	Sep.	Oct.	Nov.	Dec.
1.Gayt	4	1	1	4	4	2	2	0.1	2	2	0.1	1
2.Ekinyolu	3	1	1	9	6	1	0.1	0.1	2	3	0.1	2
3.Garip	4	3	2	6	4	2	5	3	5	6	2	7
4.Sarıçiçek	4	3	2	5	4	3	6	4	2	5	4	5
5.Genç	4	5	2	3	2	3	1	0.1	2	3	1	3
6.Şeyh Ahmet	5	5	4	3	1	0.1	1	6	1	0.1	6	7
7.Leymalun	7	11	10	11	7	1	3	0.1	1	1	0.1	15
8.Masala	1	2	5	5	2	1	0.1	2	3	0.1	1	14
9.Paymerg	3	4	2	5	3	4	6	4	2	4	3	5
10.Beyaztoprak	3	1	2	12	7	1	2	0.1	6	0.1	0.1	1
11.Çeltiksuyu	4	3	2	3	2	4	1	1	2	2	3	5
12.Elmalı	3	2	4	6	5	0.1	0.1	0.1	2	4	2	3
13.Alatepe	2	3	2	5	5	1	5	0.1	2	3	0.1	3

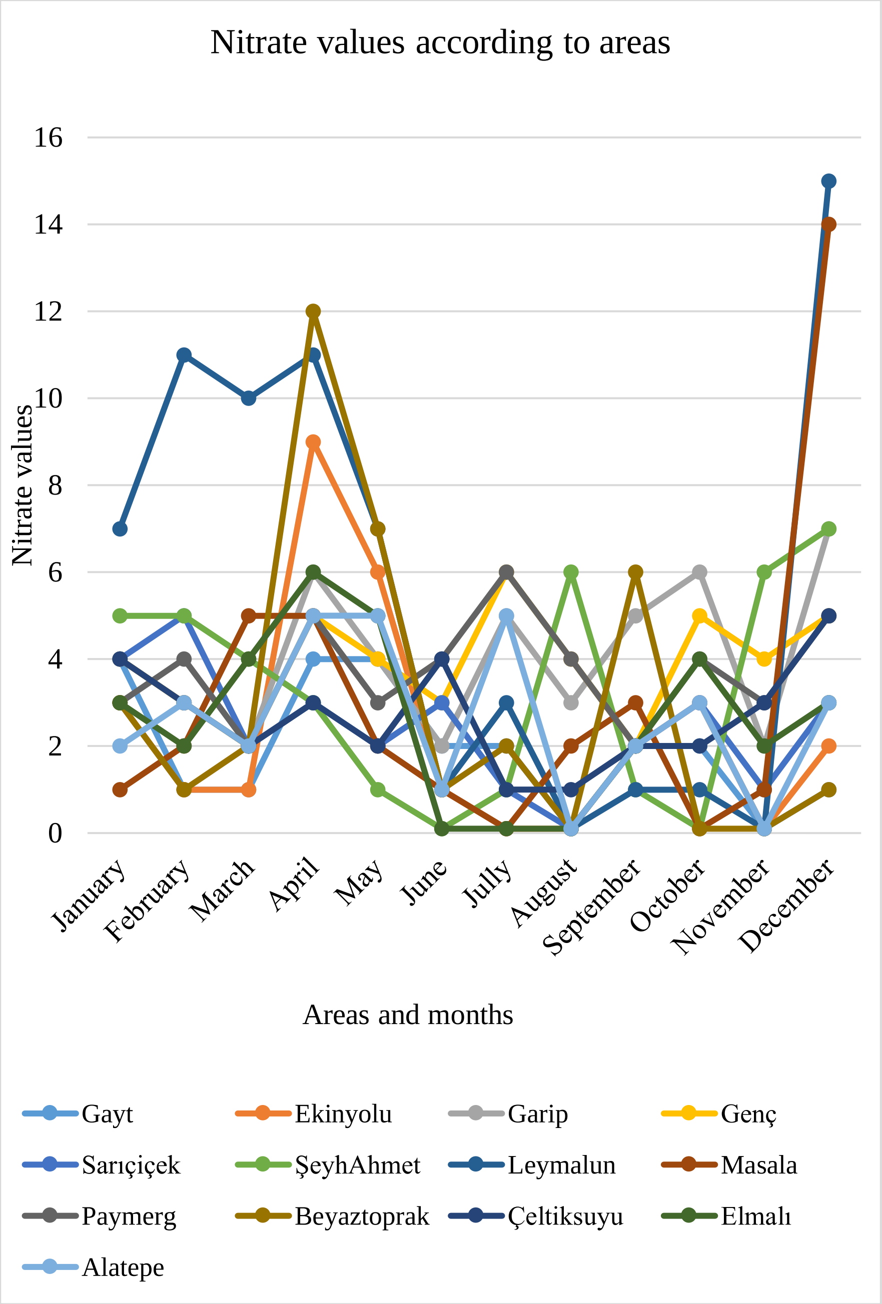


Figure 1. The nitrate levels (mg L^{-1}) of surface water in 2019 year with months in Bingol

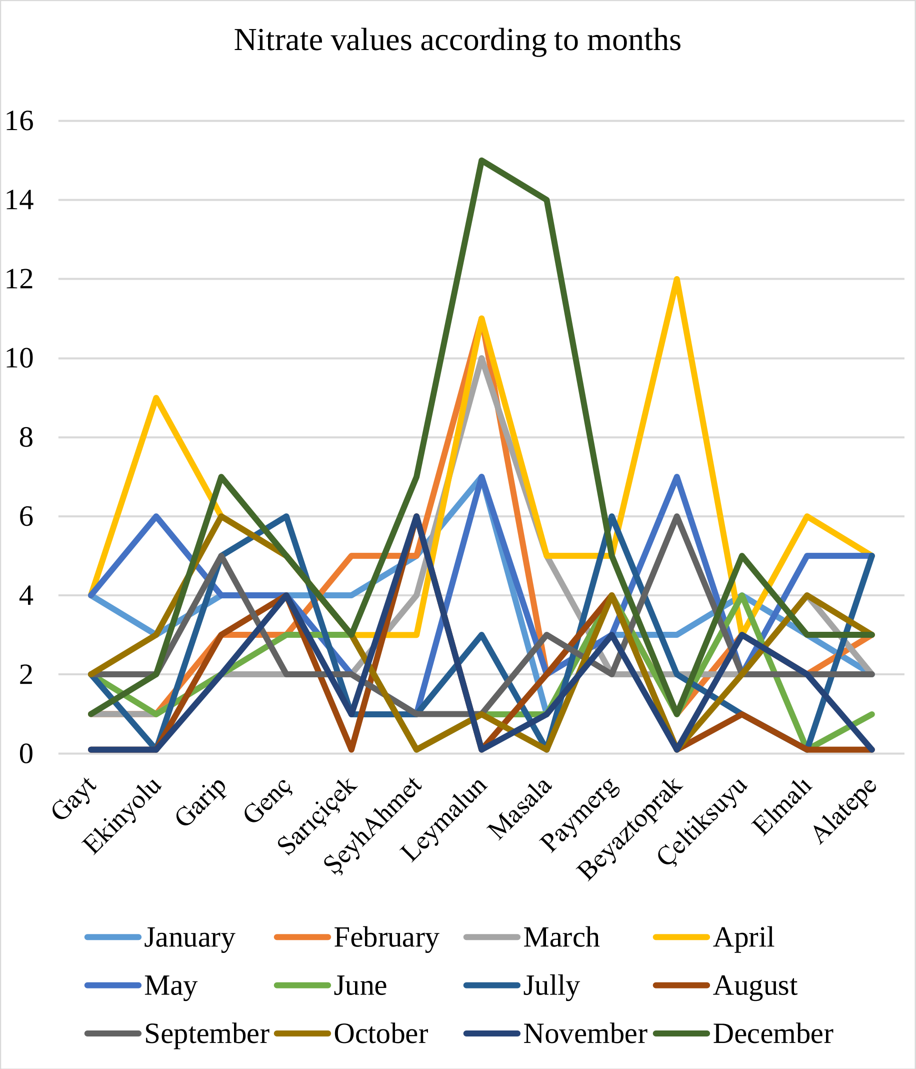


Figure 2. The nitrate levels (mg L^{-1}) of different surface water in Bingol

Analysis of variance was arranged in the randomized blocks experimental design and it was examined whether the difference between nitrates amounts in the surface water according to the different regions of Bingöl. The regions were handled as a group. Nitrate values in each region were measured in all months. Months were considered to be repeated. So it's a 12 replication experiment. Considering that the nitrate values will change according to the months due to the season, the repetitions were considered as blocks to ensure homogeneity. According to the analysis of variance, it was seen that the nitrate difference according to the regions was significant ($p<0.05$) (Table 4).

Table 4. Analysis of variance results

Source	DF	Seq SS	Adj SS	Adj MS	F	P
Region	12	136,257	136,257	11,355	2.22	0.014
Bloc	11	280,987	280,987	25,544	4.98	0.000
Error	132	676,623	676,623	5,126		
General	155	1093,867				

Tukey's test, one of the multiple comparison tests, was used to observe which regions were the difference. Tukey test data are given in Table 5. As seen in Table 5, as a result of Tukey test, the differences of the surfaces waters are follows:

- 7-Leymalun (Group A) > 5-Genç (Group B)
- 7-Leymalun (Group A) > 2-Ekinyolu (Group B)
- 7-Leymalun (Group A) > 1-Gayt (Group B)

The nitrate difference between the regions was significant ($p<0.05$). While the highest nitrate value was in 7-Leymalun (5.6 mg L^{-1}) on average, the lowest nitrate was found on 5-Genç, 2-Ekinyolu (2.4 mg L^{-1}) and 1-Gayt (1.9 mg L^{-1}).

Table 5. Tukey test results

Region	N	Mean (Nitrate, mg L^{-1})	Grouping
7-Leymalun	12	5,600	A
3-Garip	12	4,083	AB
4-Sarıçiçek	12	3,917	AB
9-Paymerg	12	3,750	AB
6-Şeyh Ahmet	12	3,267	AB
8-Masala	12	3,017	AB
10-Beyaztoprak	12	2,942	AB
11-Çeltiksuyu	12	2,667	AB
12-elmalı	12	2,608	AB
13-Alatepe	12	2,600	AB
5-Genç	12	2,425	B
2-Ekinyolu	12	2,368	B
1-Gayt	12	1,933	B

CONCLUSION

In the Turkey; In the Saraydüzü Dam Lake (Sinop), the low nitrates were observed in autumn and winter (from October to February) while high level were in early autumn (September) (Kükrer and Mutlu 2019). In the Kralkızı, Dicle and Batman Dams (on the Tigris River), high nitrate were determined in winter and spring with comparing to autumn and summer (Varol et al. 2012). The low nitrate levels of Cekerek stream (Tokat) were found in the autumn, winter, and spring, while high were in summer by Duran and Suicmez (2007). Korkanç et al (2017) observed the high nitrate values in the spring, while low values were autumn in Niğde. Koyun et al. (2020) were observed that the lowest nitrate was in autumn, while the highest was in winter. Nitrate was seasonally reached the highest value in spring.

In this study, surface water quality of the Bingöl, Turkey, determined as "first-class water quality" according to criteria of SWOR (2016), TSE (1988, 2005). The nitrate results in the surface waters of Bingol were found with accordance by results of Koyun et al. (2020), Şener at al. (2017), Kükrer and Mutlu (2019), Duran and suicmez (2007), and Tanrıverdi et al. (2010).

The nitrate levels of Leymalun-Solhan (in February, April and December), Masala-Solhan (in December) and Beyaztoprak-Bingöl (in April) water stations were found >10 mg L^{-1} in the some months. This situation has been interpreted as the fact that cattle breeding is intense in those regions and their wastes pass into these waters. Consequently, water quality can be negatively affected by unorganized and unplanned animal breeding. Therefore, necessary protection measurements and organized farming and breeding sould be taken as using and planning of the water sources.

REFERENCES

Alemda, S., Kahraman, T., Agaoglu, S., & Alisarli, M. (2009). Nitrate and Nitrite Levels of Drinking Water in Bitlis Province, Turkey, Journal of Animal and Veterinary Advances, 8(10): 1886-1892.

Broussard, W. & Turner, R. E. (2009). A Century of Changing Land-Use and Water-Quality Relationships in the Continental US, Frontiers in Ecology and the Environment, 7(6): 302-307.

Chaturvedi, M. K., & Bassin, J. K. (2009). Assessing the Water Quality Index of Water Treatment Plant and Bore Wells, in Delhi, India, Environ. Monit. Assess., 163(1/4): 449–453.

Duran, M., & Suicmez, M. (2007). Utilization of Both Benthic Macroinvertebrates and Physicochemical Parameters for Evaluating Water Quality of The Stream Cekerek (Tokat, Turkey), Journal of Environmental Biology, 28(2): 231-236.

Düzgüneş, O. (1987). Araştırma ve Deneme Metodları (İstatistik Metodları-II), Ankara Üniversitesi Ziraat Fakültesi Yayınları (in Turkish).

Goolsby, D. A. & Battaglin, W. A. (2001). Long-Term Changes in Concentrations and Flux of Nitrogen in the Mississippi River Basin, USA, Hydrological Processes 15: 1209-1226.

Jaynes, D. B., Hatfield, J. L., & Meek, D. W. (1999). Water Quality in Walnut Creek Watershed: Herbicides and Nitrate in Surface Waters, Journal of Environmental Quality, 28(1): 45-59.

Keeney, D., & Olson, R. A. (1986). Sources of Nitrate to Ground Water, Critical Reviews in Environmental Science and Technology, 16(3): 257-304.

Korkanç, S. Y., Kayıkçı, S., & Korkanç, M. (2017). Evaluation of Spatial and Temporal Water Quality in the Akkaya Dam Watershed (Niğde, Turkey) and Management Implications, Journal of African Earth Sciences, 129: 481-491.

Koyun, M., Ulıupınar, M., Şen Özdemir, N., Kırıcı, M., & Caf, F. (2020). Seasonal Changes in Water Quality of Murat River (Bingöl, Turkey) in Terms of Physico-Chemical and Biological Parameters, Acta Aquatica Turcica, 16(3): 305- 312.

Kükrer, S., & Mutlu, E. (2019). Assessment of Surface Water Quality Using Water Quality Index and Multivariate Statistical Analyses in Saraydüzü Dam Lake, Turkey, Environmental Monitoring and Assessment, 191(2): 71.

Rhodes, J. D. (1972). Quality of Water for Irrigation, Soil Science, 113(4): 277-285.

Robinson, K. W., Campbell, J. P. & Jaworski, N. A. (2003). Water Quality Trends in New England Rivers during the 20th Century, U.S. Geological Survey Water-Resources Investigations Report 03-4012.

Rupert, M. G. (2008). Decadal-Scale Changes of Nitrate in Ground Water of the United States, 1988–2004, Journal of Environmental Quality, 37(S5): S-240.

Saffigna, P. G., & Keeney, D. R. (1977). Nitrate and Chloride in Ground Water under Irrigated Agriculture in Central Wisconsin, Groundwater, 15(2): 170-177.

Smith, R. A., Alexander, R. B. & Schwarz, G. E. (2003). Natural Background Concentrations of Nutrients in Streams and Rivers of the Conterminous United States, Environmental Science & Technology 37(14): 3039-3047.

Stets, E. G., Kelly, V. J., & Crawford, C. G. (2015). Regional and Temporal Differences in Nitrate Trends Discerned from Long-term Water Quality Monitoring data, JAWRA Journal of the American Water Resources Association, 51(5): 1394-1407.

SWQR. (2016). Turkey's Ministry of Forestry and Water Affairs Surface Water Quality Regulations (in Turkish), http://www.resmigazete.gov.tr/eskiler/2016/08/20160810-9.htm.

Şener, Ş., Davraz, A., & Karagüzel, R. (2013). Evaluating the Anthropogenic and Geologic Impacts on Water Quality of the Eğirdir Lake, Turkey, Environ. Earth Sci., 70: 2527–2544.

Şener, Ş., Şener, E., & Davraz, A. (2017). Evaluation of Water Quality Using Water Quality Index (WQI) Method and GIS in Aksu River (SW-Turkey), Science of the Total Environment, 584: 131-144.

Tanrıverdi, Ç., Alp, A., Demirkıran, A. R., & Üçkardeş, F. (2010). Assessment of Surface Water Quality of the Ceyhan River Basin, Turkey, Environmental Monitoring and Assessment, 167(1): 175-184.

TSE. (1988). Turish Standards, Regulations of Water Pollution Control, The Official Gazette; Su Kirliliği Yönetmeliği, Resmi Gazete, Sayı: 19919 (in Turkish).

TSE. (2005). Regulation for Concerning Water Intended for Human Consumption (in Turkish), Offi. J. 25730 (TSE 266).

Tukey, J. (1953). Multiple Comparisons, Journal of the American Statistical Association, 48(263): 624-625.

Varol, S., & Davraz, A. (2015). Evaluation of the Groundwater Quality with WQI (Water Quality Index) and Multivariate Analysis: A Case Study of the Tefenni Plain (Burdur/ Turkey), Environ. Earth Sci., 73: 1725–1744.

Varol, M., Gökot, B., Bekleyen, A., & Şen, B. (2012). Spatial and Temporal Variations in Surface Water Quality of The Dam Reservoirs in The Tigris River Basin, Turkey, Catena, 92: 11-21.

Vasanthavigar, M., Srinivasamoorthy, K., Vijayaragavan, K., Rajiv Ganthi, R., Chidambaram, S., Anandhan, P., Manivannan, R., & Vasudevan, S. (2010). Application of Water Quality Index for Groundwater Quality Assessment: Thirumanimuttar Sub-Basin, Tamilnadu, India, Environ. Monit. Assess. 171: 595–609.

Wu, Z., Wang, X., Chen, Y., Cai, Y., & Deng, J. (2018). Assessing River Water Quality Using Water Quality Index in Lake Taihu Basin, China, Science of the Total Environment, 612: 914-922.

Yıldız N., & Bircan H. (1991). Araştırma ve Deneme Metodları. Atatürk Üniversitesi Yayınları No: 697, Ziraat Fakültesi No: 305, Ders Kitapları No: 57, Erzurum, s. 6–20 (in Turkish).

Yolcubal, I., Gündüz, Ö. C., & Sönmez, F. (2016). Assessment of Impact of Environmental Pollution on Groundwater and Surface Water Qualities in a Heavily Industrialized District of Kocaeli (Dilovası), Turkey, Environmental Earth Sciences, 75(2): 170.

CHAPTER XII

EFFECT OF CONSUMED-TEA RESIDUE APLICATION LEVELS ON AVAILABILITY OF SOME NUTRIENTS IN A CALCAREOUS SOIL

Jabar Jalal FAQE[*]
Prof. Dr. Ali Rıza DEMİRKIRAN[†]

[*] Sulaymany Polytechnic University, Northern Iraq, Iraq. jabar.jalal@gmail.com

[†] Bingol University, Faculty of Agriculture, Department of Soil Science and Plant Nutrition, Bingol, Turkey. ademirkiran@bingol.edu.tr

INTRODUCTION

It is known that the soil degradation and decreasing of the soil fertility is almost and always associated with loss of soil organic matter. The organic matter of soil is also related features of soils such as soil structure, related features of soils such as lower water infiltration, soil compaction, erodibility, and leaching. Such degradations of soils in fact lead to decrease in nutrient holding capacities and a poorer environment for biological activities (Joergensen and Potthoff 2005).

The organic matter decomposing and mineralizing in the soil can led to increasing the solubility of nutrients due to the organic anion compute with ions for making layers by binding on the soil colloids. Generally it is well known that using different kind of plant residue as organic matter with good tillage may improve the physio-chemical properties of soil.

There is a few studies about using consumed-tea residue as a direct organic fertilizer source to soil especially to calcareous soil without decomposition process:

Dilshad and Kocher (2010) stated that tea residues contains some micronutrients especially Fe, Mn and Zn hence after decomposing these nutrients can easily release to soil solution.

Jayasinghe et al. (2010) pointed that food and plant parts consumed the huge amounts partially residues can be used for soil amendment. Also, CTR may be regarded as an environmentally friendly way of increasing soil organic matter content. There are requirements for new technical information on waste materials as compost and have clearly increased over the last decade. This fact is important both developed and undeveloping countries as the practice and interest in commercial production of plant residues to soil sustanability and plant production.

In order to soils' organic matters generally are low, CTR as an organic matter is accepted a sustainable cost-effective and reasonable way to efficiently consume nutrients from pre and post-consumed food waste and vegetative wastes from modern agriculture. Because composed waste and CTR can be specifically prepared for use as a soil organic fertilizer and nutrient sources for plant (Ingham 2005).

Truly CTR or composting tea residues is a green method of disposal and fantastic for the health of all plants, providing organic matter to increase total soil quality and more parameters of soils (Lee et al. 2004).

Uncomposted CTR and its compost application of to agricultural land can result in changes in soil physical properties such as structure, water retention and infiltration rates, biological properties and crop yields. Moreover, organic materials such as compost especially can act as a valuable source of plant available nutrients such as nitrogen, phosphorus, potassium, sulphur and magnesium and thereby reduce the need for manufactured fertilizer inputs (Rollett et al. 2010).

According to laboratory analysed by Kumaraswamy (2014) CTR were showed positively effect to plant harvest residue has a good role to supply soil with N, P and C. Many studies showed plant residue has a positive role among accelerating the mineralization of nitrogen and microbial biomass activity because the residue contains enough amount of necessary nutrients and energy that plants and microrganisms needs for growth.

Sabrina and Hasmah (2008) impied that tea residue waste was also used as low cost adsorbent for removal of heavy metals and recycling wastewater, because tea waste has a strong capacity of binding of Pb and Cu from aqueous solutions. The experiments showed that 96.4% of ions of Pb removed by tea waste, the adsorption capacity is highest at solution of pH between 5 and 6.

Iraqi farmers have always concerned about the quality and fertility of their fields because of most of these soil has high contain of $CaCO_3$ and low availability of nutrients. Therefore should consider to locally supplied a cheap and safe fertilizer such as CTR or composted plant material as organic fertilizer source. The organic residues and materials are not only makes available essential nutrients to plants, it can also increases soil fertility (Arumugam 2012).

The aim of this study is to know the effect of 6 different levels of air dried consumed-tea residue on availability of some nutrients such as N, Fe, Zn and Mn concentrations in calcareous soil and amount of availability of those nutrients uptake by plant (*Zea mays* L.).

1. MATERIAL AND METHOD

1.1. Material

Preparation of consumed tea residue (CTR): Wet consumed tea-residues were collected from local Café and house keeper in North Region of Iraq. The collected wet CTR about 15 kg and placed under shadow at open area for few days with room temperature until became air dried.

In this study, surface soil samples (0-20 cm) were collected from the Agricultural Research Center of Sulaymani–Bakrajo North Region of Iraq, and air-dried then passed through a 4 mm screen. *Zea mays* L. Dkc 6724 Monsanto (American), F1 variant seeds were taken from Department of Field Crops, Faculty of Agriculture, Bingöl University, Turkey (Faqe, 2017).

1.2. Method

The pot experiment were arranged with totally 20 pots used and was set up in completely randomized design with three replicates. The doses of CTR were 0, 1, 1.5, 2, 2.5, and 3% on weight basis (Figure 1). To maintain these conditions 0 (L1), 180 (L2), 270 (L3), 360 (L4), 450 (L5), and 540 g (L6) of CTR were thoroughly mixed with 18 kg of air-dry calcareous 20 L plastic pots. These application and 6840 kg da^{-1} in the field by considering the bulk density of rates were equivalent to 0, 2280, 3420, 4560, 5700, soils. Then the mixture filled in experimental soil. Then 3 seeds of *Zea mays* L. Dkc 6724 Monsanto (American) was sowed in each pot on 17 June 2016 (Figure 2).

Figure 1. Drying of wet consumed tea residue, CTR (Faqe, 2017)

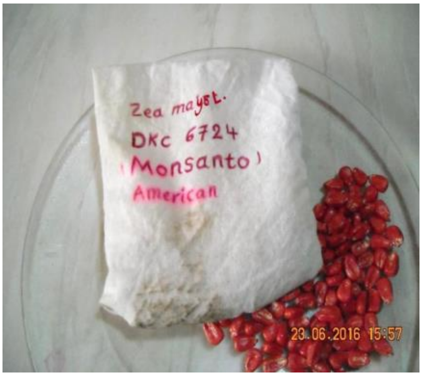


Figure 2. Used seeds of *Zea mays* L. (Faqe, 2017)

1.2.1. Soil Analysis

Texture of soil determined by Bouyoucos (1962) and saturation (%) by Richards (1954). The pH determined in with ratio 1:2.5 by Grewling and Peech (1960) and EC determined in saturated paste with EC meter by Richards (1954). Total salt determined by Jackson (1962), CEC was determined using the fixed-pH ammonium acetate method of by Dawson et al (1974), organic matter determined by wet oxidation method (Walkley and Black 1934) and $CaCO_3$ was determined in with ratio 1:3 HCl 37% by Balázs et al. (2005), Total-N determined by Bremner (1965), available-P determined by Olsen et al. (1954), extractable-K determined by Kalra and Maynard (1991), extractable-K, Ca, Mg determined by Kalra and Maynard (1991) and available-Fe, Mn, Cu, Zn and Cr determined by using DTPA extract with AAS (Lindsay and Norvell 1978).

1.2.2. Plant Analysis

Corn plant, *Zea mays* L. were harvested just above the soil surface after two months of growth period (from 17 June 2016 until 17 August 2016) and partitioned into leaf, stem and root. Fresh and oven dry (dried in an oven at 70^{o}C for 48 hours) weights of the plant parts were weighted. The plant samples were homogenized by reducing the particle size below 40 meshes. The samples were then wet-ashed and analyzed for nutrient content in the Soil Department and Plant Nutrition Laboratories and Central Laboratories of Bingol University, Turkey. One gram of plant samples were digested in a mixture of 4:1 nitric acid 65% and perchloric acids $HClO_4$ 72% (Isaac and Kerber 1971) and then total Fe, Mn, Cu and Zn concentrations of the digests were determined by a Carl Zeiss Atomic Absorption Spectrophotometer.

1.2.3. Consumed Tea Residue (CTR) Analysis

2 g of CTR was grinded and boiled with 150 ml distillated water on 80^{o}C temperature. After colding pH level determined pH meter (Grewling and Peech, 1960), EC determined by EC meter (Richards, 1954), total nitrogen determined by method of Chapman and Pratt (1961), some total micro nutrients as iron, manganese, and zinc determined by DTPA method using Carl Zeiss Atomic Absorption Spectrophotometer.

1.2.4. Statistical Analysis

The data were subjected to ANOVA by using JAMP 5.01 statistical package. The effect of CTR treatments on the measured nutrient and plant parameters were separated by Tukey-Kramer multiple comparison test. The optimum doses of CTR for different traits were determined by regression analysis.

The Tukey test was proposed by Tukey (1953) to enable the use of a "studertized range of variation". The Tukey test is calculated as follows:

$$W = q\alpha * \sqrt{MSE/r}$$

or

$$HSD = q\alpha * \sqrt{MSE/n}$$

Here, HSD: Honestly Significant Difference, $q\alpha$: Scale value for Tukey W test (Error SD; Operation SD; probability value), r or n: Number of recurrences, MSE: Error Mean Squares.

2. RESULTS

2.1. Soil Physico-Chemical Parameters

The soil which used for experimental pots was silty clay texture with a clay type and according to my primary study determinations show that the soil contains low organic matter (1.34 %), and alkaline in reaction with a pH value as 8.163. The salinity soil of research area was determined as 0.125% its means is not salinity problem. The calcium carbonate content of this soil was determined very high level (25.4%)

is classified as a calcareous soil. The CEC was determined 29.48 mg/100 g soil was low level. Extractable potassium, calcium and magnesium contents of the used soil were obtained high level as 671, 12155, and 1190 mg kg^{-1} respectively. DTPA-Fe, Zn, Mn and Cu contents were determined as 10.0, 0.8, 18.9, 2.0 mg kg^{-1} respectively.

2.2. Properties of Consumed Tea Residue (CTR)

Consumed tea-residue were also determined pH as 5.56 and EC as 526 µmhos/cm and some nutrients were determined also total N, Fe, Mn and Zn as 2.789%, 0.65, 2.512 and 0.323 mg kg^{-1} respectively.

2.3. Effect of CTR on the Plant Nutrients

ANOVA analysis revealed that there were significant differences between the CTR application rates at either $p \leq 0.01$ or $P \leq 0.05$ probability levels. In general the minimum values were obtained for the control treatment with no CTR addition whereas the maximum values were recorded for either 180 (L2) or 360 (L4) g CTR pot^{-1} treatments. In case of micronutrients, availability of Fe, Zn and Mn were significantly treatment-induced. The micro element contents were found reliable with results of Salem and El-Gizawy (2012). The highest availability values or the maximum nutrients uptakes were recorded for 360 (L4) g CTR application dose as observed for N (0.395 g pot^{-1}) by leaf, Fe (2.346 mg pot^{-1}) by root, Mn (2.465 mg pot^{-1}) by root and Zn (0.739 mg pot^{-1}) by stem respectively. While the lowest nutrient uptake from soil by corn were recorded for control treatment (L1, no CTR) as determined for N (0.106 g pot^{-1}) by root, Fe (0.519 mg pot^{-1}) by leaf, Mn (0.579 mg pot^{-1}) by leaf and Zn (0.250 mg pot^{-1}) by root of corn, respectively. This means the CTR were positively affected on the availability of total N, Fe, Zn and Mn to be uptake from soil by plant and in variant parameters (Table 1).

Table 1. The effects of applications of CTR on some nutrients concentration (N, Fe, Mn and Zn) in different parts of corn

Doses	Plant (*Zea mays* L.) parts		
	Leaf	Stem	Root
	Nitrogen content (g pot^{-1})		
L1 (0 g CTR)	0.31	0.16	0.11d
L2 (180 g CTR)	0.40	0.23	0.14cd
L3 (270 g CTR)	0.32	0.21	0.18bc
L4 (360 g CTR)	0.36	0.24	0.27a
L5 (450 g CTR)	0.38	0.23	0.23ab
L6 (540 g CTR)	0.35	0.16	0.13cd
	ns	ns	$P\leq0.01$
	Iron content (g pot^{-1})		
L1 (0 g CTR)	0.519b	0.46	1.25b
L2 (180 g CTR)	0.859a	0.72	1.51ab
L3 (270 g CTR)	0.927a	0.62	1.64ab
L4 (360 g CTR)	1.069a	0.73	2.35a
L5 (450 g CTR)	0.930a	0.65	1.85ab
L6 (540 g CTR)	0.994a	0.79	1.77ab
	$P\leq0.01$	ns	$P\leq0.05$
	Manganese content (g pot^{-1})		
L1 (0 g CTR)	0.58c	1.22	1.18b
L2 (180 g CTR)	0.87a	2.09	1.25b
L3 (270 g CTR)	0.54c	1.85	1.78ab
L4 (360 g CTR)	0.75bc	2.13	2.47a
L5 (450 g CTR)	0.69bc	1.83	1.76ab
L6 (540 g CTR)	0.81ab	1.91	1.87ab
	$P\leq0.01$	ns	$P\leq0.05$
	Zinc content (g pot^{-1})		
L1 (0 g CTR)	0.24c	0.24c	0.25c
L2 (180 g CTR)	0.33a	0.37bc	0.37bc
L3 (270 g CTR)	0.23bc	0.43b	0.42abc
L4 (360 g CTR)	0.27abc	0.74a	0.59a
L5 (450 g CTR)	0.22c	0.72a	0.44ab
L6 (540 g CTR)	0.28ab	0.39bc	0.44ab
	$P\leq0.01$	$P\leq0.01$	$P\leq0.01$

ns: Non-significant, $P\leq0.01$ and $P\leq0.05$; Means values in columns followed by the same alphabets indicate not significantly different between themselves according to Tukey test at $P\leq0.01$ and $P\leq0.05$.

The results of some nutrients in used soil indicated also the second good affected level is L2 (180 g CTR 18 kg^{-1} soil), after the first one L4 (360 g CTR 18 kg^{-1} soil). In another hand high addition of CTR as indicated in L4, L5 and L6 because of using high doses of CTR may even refer to releasing much more nutrients to the soil solution in a high concentration and this may cause to toxicity either for plant or to those beneficial microorganisms that related with mineralization process, this can clearly be seen with treatments as well.

The data showed that the second best application dose was 180 g CTR pot^{-1} treatment. On the other hand high addition of CTR above 360 g pot^{-1} was supposed to release higher amounts of nutrients. In fact during decomposition of CTR a considerable amount of N mineralized at 180 and 360 g CTR pot^{-1} treatments. However, the cease of increase in N uptake by plant upon increasing CTR application indicate that the mineralization of CTR is likely to be limited due to lack of nitrogen in the soil. Since the soil is poor in nitrogen and organic matter contents and no N fertilization was practiced in this study the N limited decomposition of CTR resulted in lesser amounts of N uptake by plant. On the other hand non synchronized release of nutrient elements from the CTR may result in smaller uptake of N.

Much nitrogen and phosphorus and some other nutrients like Fe uptake occurred before flowering compared to only one-half of P, S, are translocated from vegetative plant parts to the developing grain later in the season. Nutrient uptake with an estimated 55% and 80% of P and K that uptake occurring before flowering (Hanway 1962).

More doses and good environment of CTR as shown in treatments may increase the availability or amount of some nutrients in stem and root for a unknown and optimum time, means it is clearly seen in some replicate with Level-540 in Table 1 but in some others is not positive significant. For example bioavailability of Mn in soil which is strongly influenced by the amount and the quality of organic matter that can react with it, OM forming complexes and chelates of varying stability (Leita et al. 1999).

CTR as an organic fertilizer and natural organic materials as well, such as peat moss, compost, and wheat and clover straw and plant residue have increased the solution and exchangeable Mn (Tisdale et al. 1993).

The addition of CTR as an or similar to OM or organic fertilizer (OF) for soil might thus have increased the uptake of Zn either by increasing the potential mobility of the investigated Zn by formation of soluble organic metallic complexes or improving the growth conditions of microorganisms through the additional nutrients provided, my result may agree with those reported by Almas and Singh (2001).

These results are in agreement also with who reported that materials such as sewage sludge, animal manure, humates and compost may be rich in iron Fe^{+2}, Fe^{+3} and in metal binding biochemical that help keep Fe and other metals in solution through chelation which make Fe more available (O'Hallorans et al. 2005).

Because tea plant can uptake very high amounts of Mn during its growth under severe acid condition and low redox potential of soil. The application of higher levels of consumed tea residue to soil, and these higher levels may be decomposed or mineralized by microbial activities, this mineralization led to release of some micronutrients such as Fe, Mn, and Cu with present originally in CTR of soil solution by Dilshad and Kocher (2010). A similar result was also reported by Somani and Kanthaliya (2004) and Adiloglu and Adiloglu (2006).

Phosphorus availability was indicated by Dilshad and Kocher (2010) studied the effect of tea residues (0, 5, 10, 15, 20 kg ha^{-1}) and P levels (0, 50, 100 and 150 kg ha^{-1} of P_2O_5) with their combination on chickpea (*Cicer arietinum* L.) growth. The results were indicated that the application of tea residues in different levels had a significant effect at ($p \leq 0.01$) and ($p \leq 0.05$) on available P in soil and plant.

2.4. Effect of CTR on Calcareous Soil

The comparisons between primary study of the soil of research areas with experimental analysis of the same soil after experiment as showed in (Table 2) and according to Tukey test analysis of variances ($P \leq 0.01$ and $P \leq 0.05$) the CTR were significantly affected amount of nutrients positively by increasing the rate of CTR for additional levels. The amount of organic matter of soil increased with CTR at L2, L3, L4, L5 and L6 levels. With CTR levels, pH decreased from 8.15 to 8.00. Electrical condactivity also decreased from 0.57 mmhos m^{-1} to 0.52 or 0.53 mmhos m^{-1} with apllication of CTR. Nitrogen content of soil also increased positively with CTR like as organic matter content. From micro nutrients Mn and Zn contents of soil increased with CTR levels. The highest iron content of soil was found at L6 of CTR.

Table 2. The effects of applications of CTR on some nutrients concentration (N, Fe, Mn and Zn) in different parts of corn

Doses	pH	EC (mmhos m^{-1})	OM (%)	N (%)	Fe (mg kg^{-1})	Mn (mg kg^{-1})	Zn (mg kg^{-1})
L1	8.15	0.570	1.35b	0.046c	10.3abc	17.3c	1.11b
L2	8.00	0.523	2.15a	0.094abc	10.9bc	23.1bc	1.57ab
L3	8.02	0.530	2.21a	0.074bc	10.3abc	21.8bc	1.61ab
L4	8.03	0.523	2.27a	0.095abc	9.17bc	25.0b	1.65a
L5	7.98	0.520	2.42a	0.129a	8.73c	28.3ab	1.75a
L6	8.00	0.533	2.58a	0.116ab	12.0a	31.9a	1.86a
	ns	ns	$P\leq0.01$	$P\leq0.01$	$P\leq0.01$	$P\leq0.01$	$P\leq0.01$

ns: Non-significant, $P\leq0.01$ and $P\leq0.05$; Means values in columns followed by the same alphabets indicate not significantly different between themselves according to Tukey test at $P\leq0.01$ and $P\leq0.05$.

CONCLUSION

Daily waste of CTR as a wet garbage, rich contain of benefit nutrients and concerning about negative environmental consequences, soil degradation and yield intensification which leads to high inputs of nutrients in the form of chemical fertilizers and in another hand low productivity of calcareous soil because of leaching some elements, fixation and unavailability of nutrients especially micro nutrients such as Fe, Mn and Zn.

High contain of calcium carbonate in Iraqi-Kurdistan Region soil and Since there are little or no studies about the role of organic fertilizers especially local organic fertilizer source such as CTR and using it direct to soil may increasing micro nutrient availability in such soil, for those reasons may leading me to think and finding way for recycling and using CTR as alternative of organic fertilizer and can be used to improve physiochemical soil properties of calcareous soil. On the results obtained, it might be concluded that CTR application of micronutrients could be useful for improving the nutrient status in soil, physiological performance and may to decrease the pH value in cancerous soil which contain a high amount of carbonate and calcium carbonate ions. The Fe and Zn contents of both shoot and root were inversely proportional to rhizosphere pH. The Mn contents also increased with decreasing pH but a sharp increase was apparent below pH 5.5. The shoot Fe, Zn and Mn content were significantly correlated with the extractable levels determined in soil (Lutz et al. 1972).

The results showed that CTR with good management significantly affected of the availability of some nutrients in calcareous soil that up taken by Maize (*Zea mays* L.) and total dry matter of crop. The best CTR dose is L4 (360 g CTR) which superior over all doses and after that is L2 (180 g CTR).

REFERENCES

Adiloglu, A., & Adiloglu, S. (2006). An Investigation on Nutritional Status of Tea (*Camellia sinensis* L.) Grown in Eastern Blacksea Region of Turkey, Pakistan Journal of Biologica Science, 9(3): 365-370.

Almås, Å. R., & Singh, B. R. (2001). Plant Uptake of Cadmium-109 and Zinc-65 at Different Temperature and Organic Matter Levels, Journal of Environmental Quality, 30(3): 869-877.

Arumugam, R. (2012). Feasibility of Plant-based Composts and Compost Tea on Soil Health, Crop Protection and Production, Journal of Green Bioenergy, 1(1): 78-100.

Balázs, H., Opara-Nadib, O., & Beesea, F. (2005), A Simple Method for Measuring the Carbonate Content, SSSA, 69: 1066–1068.

Bouyoucos, G. S. (1962). Hydrometer Method Improved for Making Particle Size Analysis of Soils, Agronomy Journal, 54: 464–465.

Bremner, J. M. (1965). Organic Forms of Soil Nitrogen, In Methods of Soil Analysis, In: Black CA et al. (Eds.) ASA, USA 9, Part 2, pp 1238–1255

Chapman, H. D., & Pratt, P. F. (1961). Methods of Analysis for Soils, Plants, and Waters, Priced Publication 4034, Division of Agriculture Sciences, University of California, Berkeley.

Dawson, M., Foth, M. D., Page, A. L., & McLean, E. O. (1974). Cation Exchange Properties of Soils, A Slide Show, Div. S-2, Soil Chemistry, SSSA, pp 8-45

Dilshad, A. D., & Kocher, R. M. (2010). Influence of Consumed Tea Residues and Phosphorus Fertilizer on Phosphorus Availability and Growth of Chickpea (*Cicer arietinum* L), Journal of Environmental Studies, 4: 9-13.

Faqe, J. J. (2017). Effect of Consumed Tea Residue Application Levels on Availability of Some Nutrients in a Calcareous Soil, Master Thesis, Bingol University, Bingol, Turkey.

Grewling, T., & Peech, M. (1960). Chemical Soil Tests, Cornell University Agricultural Experiment Station Bulletin, p 960.

Hanway, J. J. (1962). Corn Growth and Composition in Relation to Soil Fertility: II. Uptake of N, P, and K and Their Distribution in Different Plant Parts during the Growing Season, Agronomy Journal, 54: 145-148.

Ingham, E. R. (2005). The Compost Tea Brewing, Manual Fifth Edition, Soil Food Web Incorporated, Corvallis, Oregon, pp 13-17.

Issac, R. A. & Kerber, J. D. (1971). Atomic Absorption and Flame Photometry: Techniques and Uses in Soil, Plant and Water Analysis, In Instrumental Methods for Analysis of Soil and Plant Tissues (L.M. Walsh, Ed.), Soil Science Society of America, pp. 17-37.

Jackson, M. C. (1962). Soil Chemical Analysis, Prentice Hall. Inc., Eng. Cliff. USA

Jayasinghe, G. Y., Tokashiki, Y., Arachchi, I. D., & Arakaki, M. (2010). Sewage Sludge Sugarcane Trash Based Compost and Synthetic Aggregates as Peat Substitutes in Containerized Media for Crop Production, Journal of Hazardous Materials, 174: 700–706.

Joergensen, R. G., & Potthoff, M. (2005). Microbial Reaction in Activity, Biomass, and Community Structure after Long-term Continuous Mixing of a Grassland Soil, Soil Biology and Biochemistry, 37: 1249-1258.

Kalra, Y. P., & Maynard, D. G. (1991). Methods Manual for Forest Soil and Plant Analysis, Forestry Canada, Northern Forestry Center, Information Report NORX-319, Vol. 15, pp 84–94.

Kumaraswamy, S., Mendham, D. S., Grove, T. S., O’Connell, A. M., Sankaran, K. V., & Rance, S. J. (2014). Harvest Residue Effects on Soil Organic Matter, Nutrients and Microbial Biomass in Eucalypt Plantations in Kerala, India, Forest Ecology and Management, 328: 140–149.

Lee, J. J., Park, R. D., Kim, Y. W., Shim, J. H., Chae, D. H., & Rim, Y. S. (2004). Effect of Food Waste Compost on Microbial Population, Soil Enzyme Activity and Lettuce Growth, Bioresource Technology, 93(1): 21-8.

Leita, L., De Nobili, M., Mondini, C., Muhlbachova, G., Marchiol, L., Bragato, G., & Contin, M. (1999). Influence of Inorganic and Organic Fertilization on Soil Microbial Biomass, Metabolic Quotient and Heavy Metal Bioavailability, Biology and Fertility of Soils, 28(4): 371-376.

Lindsay, W. L., & Norvell, W. A. (1978). Development of a DTPA Soil Test for Zinc, Iron, Manganese, and Copper, SSSA, 42: 421–428.

Lutz, J. A. J., Genter, C. F., & Hawkins, G. W. (1972). Effect of Soil pH on Element Concentration and Uptake by Maize, Agronomy Journal, 64: 583–585.

O'Hallorans, J. M., Lindemann, W. C., & Steiner, R. (2005). Iron Characterization in Manure Amended Soils, Communications in Soil Science and Plant Analysis, 35(15-16): 2345-2356.

Olsen, S. R., Cole, C. V., Watanabe, F. S., & Dean, L. A. (1954). Estimation of Available Phosphorus in Soils by Extraction with Sodium Bicarbonate, US Department of Agriculture Circular 939, Washington, DC.

Richards, L. A. (1954). Saline and Alkali Soil, United States Agriculture, Handbook No. 60, p 84.

Rollett, A. J., Bhogal, A., Taylor, M. J., & Chambers, B. J. (2010). Green/Food Compost: Crop Available Nitrogen Supply and Soil Fertility Benefits, Use of Manures and Organic Wastes to Improve Soil Quality and Nutrient Balances, pp 10-15.

Salem, H. M., & El-Gizawy, N. K. B. (2012). Importance of Micronutrients and its Application Methods for Improving Maize (*Zea mays* L.) Yield Grown in Clayey Soil, American Eurasian J. of Agr. and Env. Sci., 12: 954-959.

Sabrina, K., & Hasmah, S. I. (2008). Tea Waste as Low Cost Absorbent for Removal of Heavy Metals and Turbidity from Synthetic Wastewater, Proceedings Int. Conference on Environmental Research and Technology, pp 238-241.

Somani, L. L. & Kanthaliya, P. C. (2004). Soils and Fertilizers, ATA Glance, 1st edition, Agrotech Publishing Academy, pp 648- 690.

Tisdale, S. L., Nelson, W. L., Beaton, J. D., & Havlin, J. L. (1993). Soil Fertility and Fertilizers, 5th edition, Prentice- Hall., Incorpation, pp 25-36.

Tukey, J. W. (1953). The Problem of Multiple Comparisons, Unpublished Manuscript, In The Collected Works of John W. Tukey VIII. Multiple Comparisons: 1948–1983 1–300. Chapman and Hall, New York.

Walkley, A. & Black, I. A. (1934). An Examination of Degtjareff Method for Determining Soil Organic Matter and a Proposed Modification of the Chromic Acid Titration Method, Soil Science, 37: 29-38.

CHAPTER XIII

SUPPORT VECTOR MACHINES AND APPLICATIONS IN AGRICULTURE

Assoc. Prof. Dr. Nazire MİKAİL[*]
Dr. Yasin ALTAY[†]

[*] Siirt University, Faculty of Agriculture, Department of Animal Science, Siirt, Turkey. naziremikail@siirt.edu.tr

[†] Eskisehir Osmangazi University, Faculty of Agriculture, Department of Animal Science, Eskisehir, Turkey. yaltay@ogu.edu.tr

INTRODUCTION

In statistical classification such as linear and quadratic discriminant analyses, the classification rules were constructed using criteria such as Mahalanobis distance minimization and posterior probability maximization. A basic property of these classification methods is their reliance on estimators such as sample mean vectors or sample variance-covariance matrices that summarize information contained in the observed training data from each class. Support vector machines (SVMs) proposed by Vapnik (1996, 1998), in contrast, are characterized by an internal process of constructing classification rules quite different in basic concept from those of the statistical methods. SVMs are often effective in cases for which ordinary classification methods are not effective, such as problems based on high dimensional data and data with nonlinear structure, and they have been adapted and applied in many fields.

Many classification methods have been proposed by machine learning classifiers in the fields of expert systems, statistical methods, and cellular biology. The support vector machine method is an application of machine learning based on statistics.

Generally, there are 2 situations in statistical learning: 1. learning/computing from training examples, 2. estimating for test samples (examples to be encountered later). According to the definition made in the statistics section, training and test data should be in the same statistical distribution. Learning processes in statistics can be grouped into the following three groups;

- Classification or pattern recognition,
- Calculation of the continuous function from regression or noisy samples,
- Estimation of probability density using samples.

The SVM method also has the infrastructure to perform these operations.

The aim of SVM is to find the optimum separator plane that classifies the data points as well as possible and divides them into two class points as best as possible. In other words, it is aimed to find the situation where the distance between the two classes is maximum. The cornerstones of this classification logic are the support vectors at the endpoints of both classes and selected from among the training samples.

During the classification process, some of the available data is reserved for training and the rest for testing. Because the use of training data in the accuracy estimation of the classifier causes us to obtain optimistic (higher than real applications) results. The ratio of these data to each other directly affects the accuracy (as well as the error rate) of the classification process. Another factor affecting the accuracy rate is the distribution of the data. Mathematically, SVM is based on distribution-independent formula structure. In the SVM technique, the classifier model is created by first taking the training data and training the SVM. Then, the output values that the system will calculate for the test data that we know beforehand are determined. Then, the classification performance of the SVM is evaluated according to the difference ratio between these two values (Mammadova & Keskin, 2013).

In this chapter, we focus on the basic concept in SVMs for the solution of recognition and classification problems in the field of agriculture.

1. SUPPORT VECTOR MACHINE

In Figure 1 shows a support vector machine (SVM), which is a non-recurrent static two-layer ANN. The two classes are separated by a hyperplane, but unlike MLP, this hyperplane provides maximum separation.

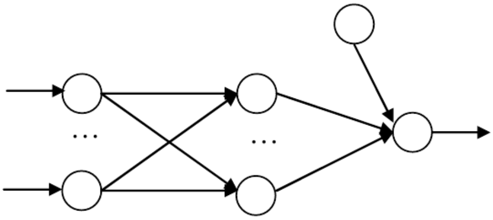


Figure 1. Support vector machine

The main idea behind creating SVM is to select a subset training data as support vectors. This subset represents the robust properties of the entire training set. The SVM supervised learning algorithm provides maximizing the Lagrange function. This algorithm is based on minimizing empirical risk (learning errors).

For SVM the following are specified:

1. A training set is given as

$$\left\{(x_\mu, d_\mu) \middle| x_\mu \in R^{N^{(0)}},\quad d_\mu \in \{-1{,}1\}\right\},\ \ \mu\epsilon\overline{1,P},$$

where, x_μ - μ^{th} training input vector, d_μ - μ^{th} training output, $N^{(0)}$ - the number of neurons in the input layer, P is the power of the training sets, and the parameter C>0.

2. Calculation of the output signal for the first layer in the form of a kernel K(x$_i$,x$_j$), $i, j \in \overline{1, P}$

- if we want to get a polynomial learning machine, then

$$K(x_i, x_j) = (x_i^T x_j)^p \quad or \quad K(x_i, x_j) = (x_i^T x_j + 1)^p;$$

- if we want to get a two-layer perceptron, then

$$K(x_i, x_j) = tanh(k_0 + k_1 x_i^T x_j),$$

where, the parameters k$_0$, k$_1$ satisfy the condition k$_0$>0 or k$_1$>0;

- if we want to get a radial basis function, then

$$K(x_i, x_j) = exp\left(-\frac{1}{2\sigma^2}\|x_i - x_j\|^2\right),$$

where parameter σ is the width of the function $K(x_i, x_j)$,

$$x_i^T x_j = \sum_{k=1}^{N(0)} x_{ik} x_{jk}, \quad \|x_i - x_j\| = \sqrt{\sum_{k=1}^{N^{(0)}} (x_{ik} - x_{jk})^2}.$$

3. The variables are defined (Lagrange multipliers) λ$_i$, maximizing the Lagrange function, i.e.

$$L(\lambda) = \sum_{i=1}^{P} \lambda_i - \frac{1}{2}\sum_{i=1}^{P}\sum_{j=1}^{P} \lambda_i \lambda_j d_i d_j K(x_i, x_j) \to \max_{\lambda} \quad ,$$

with restrictions

$$0 \le \lambda_i \le C,$$

$$\sum_{i \in I} \lambda_i y_i = 0, \quad I = \{i\colon\ 0 \le \lambda_i \le C\}$$

If 0<λ$_i$<C, then the pair (x$_i$, d$_i$) is a support vector, i.e. is located on the border of the dividing strip.

The number of support vectors will be denoted as $N^{(1)}$, since it corresponds to the number of neurons in the first layer. We enumerate the training set and the Lagrange multipliers so that we first go support vectors and associated Lagrange multipliers.

4. Determine the optimal value of the vector of weight coefficients w

$$w = \sum_{i=1}^{N^{(1)}} \lambda_i d_i x_i.$$

5. Determine the optimal bias (threshold) b using any support vector x_i

$$b = \frac{1}{d_i} - w^T x_i.$$

The function is:

$$y = f(x) = sgn(b + w^T x) = sgn\left(b + \sum_{i=1}^{N^{(1)}} \lambda_i d_i K(x, x_i)\right)$$

(Fedorov, 2016).

1.1. Advantages

1. Used to classify samples.
2. It is a universal approximator. Provides a global approximation for the nonlinear mapping of the input signal to the output signal.
3. Provides good generalization quality.
4. The number of hidden layers is automatically determined (equal to one).
5. The number of neurons in the hidden layer is automatically determined (equal to the number of support vectors).
6. In contrast to the gradient learning methods MLP, RBFNN, ME, HME, which randomly choose the position of the dividing hyperplane,

SVM learning is based on the principle of the optimal dividing hyperplane, which leads to maximization of the dividing strip width between classes, therefore, to more confident classification.

7. Unlike gradient learning methods MLP, RBFNN, ME, HME, which reduce to multi-extremal problems, SVM learning is reduced to a quadratic programming problem in a convex domain that has a unique solution.

1.2. Disadvantages

1. The number of neurons in the hidden layer is greater than that of MLP, RBFNN, especially in the case of linear inseparability and noisy data. This results in slower SVM performance compared to MLP, RBFNN.
2. There can be only two classes, so you have to use a binary tree, the nodes of which are SVMs.
3. There is no general approach to the automatic choice of the kernel (and the construction of the rectifying subspace as a whole) in the case of linear inseparability of classes.
4. In the general case, when linear separability is not guaranteed, it is necessary to select the control parameter C.
5. Unlike ART, it does not solve the problem of ductility and stability.
6. Represents the image as a vector (Fedorov, 2016).

2. SVM APPLICATIONS IN AGRICULTURE

Many studies have been carried out using the svm algorithm in agriculture and successful results have been found. A few of them are given below as examples:

Karimi et al. (2008) evaluated the ability of SVM regression models to extract continuous vegetation variables using aerial hyperspectral observations. The study showed that by using reflectance data collected at the tasseling stage, crop parameters can be estimated with reasonable accuracy. The coefficients of determination were found greater than 0.9 for biomass, yield and plant height. The results were also compared with those obtained with a stepwise approach, and the SVM results were found to be superior.

Mathur & Foody (2008) classified with support vector machine agricultural and non-agricultural areas in a particular land using remote sensing data. Agricultural classes were cotton, basmati rice and a local variety of rice while the non-agricultural classes were land and sand.

Ahmed et al. (2011) used SVM and Bayesian classifier as machine learning algorithm for efficient classification of crops and weeds in digital images. A total of 22 features characterizing the plants and weeds in the images were tested to find the optimal combination of features that yielded the highest classification rate for both methods. Analysis of the results reveals that SVM achieves over 98% accuracy, where Bayesian classifier achieves over 95% accuracy on the same set of images.

Kumar et al. (2017) presented a model for monitoring of sugarcane crop. The proposed model continuously monitor parameters (temperature, humidity and moisture) responsible for healthy growth of the crop in addition KNN clustering along with SVM classifier is utilized for infection identification if any through images obtained at regular intervals. The data has been transmitted wirelessly from the site to the control unit. Model achieves an accuracy of 96% on a sample of 200 images.

Pulido et al. (2017) presented a classification system for weeds and vegetables from outdoor crop images. The classifier is based on SVM with its extension to the nonlinear case, using the Radial Basis Function (RBF) and optimizing its scale parameter σ to smooth the boundary decision. The feature space is the result of Principal Component Analysis (PCA) for 10 texture measurements calculated from Gray Level Co-occurrence Matrices (GLCM). The results indicate that classifier performance is above 90%, validated with specificity, sensitivity and precision calculations.

Akbarzadeh et al. (2018) in their study developed a SVM algorithm for weed-crop separation. Experimental results showed that the developed Gaussian SVM algorithms can classify maize and silver beet with a separation accuracy of 97%, while the maximum accuracy obtained using the traditional NDVI-based method does not exceed 70%.

Xiao et al. (2018) developed a model for recognition and classification of vegetable pests by means of SVM. The average accuracy rate of recognition classification is higher than 90% in many cases even if the test image contains a complex environmental background.

Kumar et al. (2019) developed SVM based classification models for the prediction of rice yield in India. Experiments have been conducted involving one against-one multi classification method, k-fold cross validation and polynomial kernel function for SVM training. The best prediction accuracy for the 4-year relative average increase has been achieved as 75.06% using 4-fold cross validation method.

Zhao et al. (2020) explored the genomic-based prediction performance of SVM model. They selected the most suitable kernel function and hyperparameters for the SVM model in eight published genomic data sets on pigs and maize. Next, they compared the SVM model with RBF and the linear kernel functions to the two most commonly used genome-enabled prediction models (GBLUP and BayesR) in terms of prediction accuracy, time, and the memory used. The results showed that the SVM model had the best prediction performance in two of the eight data sets. According to the results, SVM is a competitive method in animal and plant breeding, and there is no universal prediction model.

Ou & Zhang (2021) in their study used fuzzy least square support vector machine (FLS-SVM) to investigate the recognition performance of harvesting robot using regions of interest histogram of oriented gradients feature. The detection accuracy for the learning samples, the isolated fruit, the overlapped fruit, and the background can achieve 99.50%, 96.0%, 89.9%, and 97.0%, respectively.

Peng et al. (2021) tried to identificate different grape varieties with support vector machine. The experimental results showed that the method proposed, can achieve fast and accurate identification of grape varieties. Based on the proposed algorithm, the smart machinery in agriculture can take more targeted measures based on the different characteristics of different grape varieties for further improvement of the yield and quality of grape production.

Suyoto et al. (2021) used first-order feature extraction with the SVM algorithm, which aims to recognize civet and robusta coffee beans' patterns using texture analysis on grayscale images and feature extraction. The data used in this study were 120 images, consisting of 110 training data and 10 test data. The accuracy in identifying the types of coffee beans by using this method is 87.27%.

CONCLUSION

Support Vector Machines are applied to many real world problems. In this section, only a few applications in agriculture are mentioned. The purpose of this section is to show how the method can be used successfully in different fields. The results obtained in the studies also show that SVMs give better results than other machine learning methods.

REFERENCES

Ahmed, F., Bari, A.S.M.H., Hossain, E., Al-Mamun, H. A., Kwan, P. H. (2011). Performance Analysis of Support Vector Machine and Bayesian Classifier for Crop and Weed Classification from Digital Images. World Applied Sciences Journal, 12(4), p. 432-440.

Akbarzadeh, S., Paap, A., Ahderom, S., Apopei, B., Alameh, K. (2018). Plant discrimination by Support Vector Machine classifier based on spectral reflectance. Computers and Electronics in Agriculture. Volume 148, pp: 250-258.

Karimi, Y., Prasher, S.O., Madani, A. and Kim, S. (2008). Application of support vector machine technology for the estimation of crop biophysical parameters using aerial hyperspectral observations. Canadian Biosystems Engineering/Le génie des biosystèmes au Canada 50: 7.13 - 7.20.

Kumar, S., Kumar, V., Sharma, R. K. (2019). Rice Yield Forecasting Using Support Vector Machine. International Journal of Recent Technology and Engineering, 8 (4): 2588-2593.

Kumar, S., Mishra, S., Khanna, P., Pragya. (2017). Precision Sugarcane Monitoring Using SVM Classifier. Information Technology and Quantitative Management (ITQM 2017).

Mammadova, N. & Keskin, İ. (2013). Application of the Support Vector Machine to Predict Subclinical Mastitis in Dairy Cattle. The Scientific World Journal, vol. 2013, Article ID 603897, 9 pages, 2013. https://doi.org/10.1155/2013/603897

Mathur, A. & Foody, G. M. (2008). Crop classification by support vector machine with intelligently selected training data for an operational application, International Journal of Remote Sensing, 29:8, 2227-2240.

Ou, J. & Zhang, J. (2021). Investigation on Recognition Performance of Harvesting Robot Using Regions of Interest Histogram of Oriented Gradients Feature Based on Improved Fuzzy Least Square Support Vector Machine. Hindawi Mathematical Problems in Engineering. Volume 2021, Article ID 6650367, 10 pages, https://doi.org/10.1155/2021/6650367.

Pulido, C., Solaque, L., Velasco, N. (2017). Weed recognition by SVM texture feature classification in outdoor vegetable crop images. Ingeniería e Investigación, vol. 37, núm. 1, April, 2017, pp. 68-74.

Peng, Y., Zhao, S., Liu, J. (2021). Fused Deep Features-Based Grape Varieties Identification Using Support Vector Machine. Agriculture. 11, 869. https://doi.org/10.3390/agriculture11090869

Suyoto, R. Z. H., Komarudin, M., Nama, G. F. & Yulianti, T. (2021). Classification of Civet and Canephora coffee using Support Vector Machines (SVM) algorithm based on order-1 feature extraction. IOP Conf. Ser.: Mater. Sci. Eng. 1173 012006

Vapnik, V. (1996). The Nature of Statistical Learning Theory. Springer, New York.

Vapnik, V. (1998). Statistical Learning Theory. Wiley, New York

Xiao, D. Q., Feng, J. Z., Lin, T. Y., Pang, C. H., Ye, Y. W. (2018). Classification and recognition scheme for vegetable pests based on the BOF-SVM model. Int J Agric & Biol Eng, 11(3): 190–196.

Zhao, W., Lai, X., Liu, D., Zhang, Z., Ma, P., Wang, Q., Zhang, Z. & Pan, Y. (2020). Applications of Support Vector Machine in Genomic Prediction in Pig and Maize Populations. Front. Genet. 11:598318. doi: 10.3389/fgene.2020.598318

İKSAD
Publishing House

ISBN: 978-625-8061-07-9