TÜRKİYE İSTATİSTİK KURUMU

İSTATİSTİK ARAŞTIRMA DERGİSİ 2001

İSTATİSTİK ARAŞTIRMA DERGİSİ

JOURNAL OF STATISTICAL RESEARCH

JSRTR

E-ISSN 2791-7614

Cilt Volume 13 Sayı Issue 01 Yıl Year 2023

İSTATİSTİK ARAŞTIRMA DERGİSİ
ULUSLARARASI HAKEMLİ DERGİ

JOURNAL OF STATISTICAL RESEARCH
INTERNATIONAL PEER-REVIEWED JOURNAL

ISSN: 2791-7614
CİLT/VOL. 13 • SAYI / ISSUE 1 • 2023

Türkiye İstatistik Kurumu Adına Sahibi / Owner on Behalf of Turkish Statistical Institute
Erhan ÇETİNKAYA

Baş Editör / Editor - in - Chief
Prof. Dr. Selahattin GÜRİŞ

Editör Yardımcısı / Assistant Editor
Prof. Dr. Nurcan METİN
TÜİK Uzmanı Orçun AYDIN

Sorumlu Müdür / Responsible Manager
Stratejik Planlama, İzleme ve Değerlendirme Grup Başkanı Gönül KALENDER

Yabancı Dil Editörü / Foreign Language Editor
TÜİK Uzmanı Nilgün DORSAN

İletişim / Communication
Türkiye İstatistik Kurumu
Devlet Mah.
Necatibey Cad. No: 114
06420 Çankaya/Ankara/TÜRKİYE

Tel. / Phone: +90 312 454 73 50
Belgeç / Fax: +90 312 454 82 96
E-posta / E-mail: journal@tuik.gov.tr
İnternet Sayfası / Web Site:
https://journal.tuik.gov.tr

Türkiye İstatistik Kurumu yayını olan İstatistik Araştırma Dergisi, yılda iki kez yayımlanan uluslararası hakemli bir dergidir. Makalelerin bilimsel ve etik kurallara uygunluğu yazarların sorumluluğundadır; Türkiye İstatistik Kurumu sorumlu tutulamaz.
Journal of Statistical Research, the publication of the Turkish Statistical Institute, is an international peer-reviewed journal published twice a year. Compliance of the articles with scientific and ethical rules is the responsibility of the authors; Turkish Statistical Institute can not be held responsible.

Baş Editör'den

Değerli okuyucular,

İstatistik Araştırma Dergisi, Türkiye İstatistik Kurumu'nun 2002 yılında yayınlanmaya başladığı uluslararası akademik bir dergisidir.

Yayın politikası yenilenen ve kısa bir aradan sonra Temmuz 2021'de yeniden yayınlanmaya başlayan İstatistik Araştırma Dergisi'nde istatistik yanında ekonometri, yöneylem araştırması ve istatistiğin araç olarak kullanıldığı diğer bilim dallarında teorik çalışmalara yer verdiğimiz gibi, uygulamalı orijinal çalışmalara yer verilmektedir.

İstatistik Araştırma Dergisi'nin hedefi söz ettiğimiz konularda yayınlayacağı makaleler ile literatüre katkı sağlamaktır. Bu doğrultuda, derginin farklı endekslerde taranmasına yönelik çalışmalara da yer verilmektedir.

Bu sayıda makaleleri ile destek veren yazarlarımıza, dergimizin bilimsel çizgisine yön veren hakemlerimize, dergimizin Editörler Kuruluna üye olmayı kabul ederek bizi onurlandıran değerli bilim insanlarına, desteklerinden ötürü TÜİK Başkanı Sayın Erhan ÇETİNKAYA'ya ve emeği geçen herkese teşekkürlerimi sunarım.

Dergimizin Temmuz 2023 (Cilt 13, Sayı 1) sayısının bilim dünyasına katkı sağlamasını diler, bundan sonraki sayılarımıza da ilgi ve desteğinizin devamını diler, saygılarımı sunarım.

Prof. Dr. Selahattin GÜRİŞ
Baş Editör

Dergimize gönderilen makalelerin tümü intihal tespit aracıyla incelenmektedir.

All articles submitted to our journal are analyzed by plagiarism detection tools.

İÇİNDEKİLER / CONTENTS

ARAŞTIRMA MAKALELERİ / ORIGINAL RESEARCH ARTICLES

1-33 Cemil İSKENDER
Türkiye Buğday Üretimi ve Buğday Piyasası Üzerine Ekonometrik Çalışmalar (Verhulst Büyüme Fonksiyonu ve Örümcek Ağı Teoremlerinin İleri Uygulaması) / Econometric Studies on Wheat Production and Wheat Market of Türkiye (Advanced Application of Verhulst Growth Function and Cobweb Theorems)

34-47 Saygın DİLER, Yıldırım DEMİR
Sağdan Sansürlü Veriler için Veri Madenciliği Algoritmaları Performanslarının Karşılaştırılması / Comparison of Data Mining Algorithms Performances for Right-Censored Data

48-60 Abhishek AGARWAL, Himanshu PANDEY
Göç ve Mortalite Verilerine İlişkin Yaşamsal Olaylar için Himanshu Dağılımı Temelli Tek Boyutlu Önyargılı Olasılık Modeli /A One-Dimensional Biased Probability Model Based on Himanshu Distribution for Vital Events Related to Migration and Mortality Data

61-72 Esra N. KILCI
Para Arzı ve Kamu Borcundaki Artışın Euro Bölgesi Ekonomik Performansına Etkisinin Analizi /A Study on the Impact of Money Supply Growth and Government Debt Increase on the Economic Performance in the Euro-Area

73-84 Ceyda TUNÇ YILANCI , Mahmut Ünsal ŞAŞMAZ
Yolsuzluğun Vergi Gelirleri Üzerindeki Etkisi: AB Geçiş Ülkeleri Örneği / The Effect of Corruption on the Tax Revenues: Case of EU Transition Economies

İstatistik Araştırma Dergisi
Journal of Statistical Research
ISSN: 2791-7614

(2023), 13 (1)
Başvuru/Received: 24/08/2022
Kabul/Accepted: 01/11/2022

https://journal.tuik.gov.tr/

ARAŞTIRMA MAKALESİ RESEARCH ARTICLE

Econometric Studies on Wheat Production and Wheat Market of Türkiye (Advanced Application of Verhulst Growth Function and Cobweb Theorems)

Cemil İSKENDER
Researcher Economist
iscemil@outlook.com
Orcid No: 0000-0003-2841-5964

Abstract

In our study, we make economic analysis of wheat production and of the amounts traded together with the prices in Türkiye in an advanced econometrics and statistical framework. For this purpose, we create a data set of more than one hundred and ten years period using TurkStat data sources and explain the time trend of wheat production with the growth function. In the market research section, the relationships between supply, demand and prices are discussed and the time equilibrium paths of prices and quantities are determined with forty-year data and cobweb theorems. We then investigate the central points of price and quantity. To this end, we use the simple cobweb theorem with a single delay and Goodwin's theory that includes two delays and expectations. After determination of trend functions, cyclical fluctuations around the trend are determined with complementary functions. All functions have been solved simultaneously in computer enviroment. The results obtained with wheat data confirm the economic patterns of production and the market. Then, according to the equations obtained in the Goodwin model, we create the cobweb curve by simulating quantities and prices. We used econometrics and statistics procedures of SAS library for study.

Keywords: Cobweb Theorem, Wheat Production of Türkiye, Goodwin, Econometry

Corresponding Author / Sorumlu Yazar: 1-Cemil İSKENDER, Researcher Economist

Citation / Atıf: İSKENDER C. (2023). Econometric Studies on Wheat Production and Wheat Market of Türkiye (Advanced Application of Verhulst Growth Function and Cobweb Theorems). İstatistik Araştırma Dergisi, 13 (1), 1-33.

Citation / Atıf: İSKENDER C. (2023). Econometric Studies on Wheat Production and Wheat Market of Türkiye (Advanced Application of Verhulst Growth Function and Cobweb Theorems). İstatistik Araştırma Dergisi, 13 (1), 1-33.

Türkiye Buğday Üretimi ve Buğday Piyasası Üzerine Ekonometrik Çalışmalar (Verhulst Büyüme Fonksiyonu ve Örümcek Ağı Teoremlerinin İleri Uygulaması)

Özet

Çalışmamızda ileri ekonometri ve istatistik yöntemlerle, Türkiye'deki buğday üretiminin ve fiyatlarla birlikte işlem gören miktarların ekonomik analizini yapıyoruz. Bu amaçla TurkStat veri kaynaklarını kullanarak yüz on yılı aşkın bir veri seti oluşturup buğday üretiminin zaman trendini büyüme fonksiyonuyla açıklıyoruz. Piyasa araştırması bölümünde, kırk yıllık veriler ve örümcek ağı teoremleriyle arz, talep ve fiyatlar arasındaki ilişkiler ele alınmakta fiyatlar ve miktarların zaman denge patikaları tespit edilmektedir. Daha sonra fiyat ve miktarın merkez noktalarını araştırmaktayız. Bu amaçla tek gecikmeli basit örümcek ağı teoremini ve Goodwin'in iki gecikmeli ve bekleyişleri de içeren örümcek ağı teorisini kullanıyoruz. Trend tespitinden sonra, trend etrafındaki konjonktürel dalgalanmalar tamamlayıcı fonksiyonlarla tespit edilmektedir. Bütün fonksiyonlar bilgisayar ortamında eşanlı olarak çözülmüştür. Buğday verileriyle elde edilen sonuçlar üretim ve piyasaya ait ekonomik modelleri teyit etmektedir. Daha sonra Goodwin modelinde elde edilen denklemlere göre fiyat ve miktarların simülasyonuyla örümcek ağı eğrilerini oluşturmaktayız. Çalışmamızda SAS kütüphanesinin ekonometri ve istatistik prosedürleri kullanılmıştır.

Anahtar sözcükler: Örümcek Ağı Teoremi, Türkiye Buğday Üretimi, Goodwin, Ekonometri

1. Introduction and Scope of Research

The aim of this study is to carry out the *production and market analysis of wheat agriculture* which is a long-standing economic activity in the geography of Türkiye, with TurkStat data and according to the principles of economic theory, mathematics, econometric analysis and advanced statistical applications. With the mathematical modeling we made regarding wheat production and the market, the abstractly mathematical representation of Turkish wheat data was aimed and state of affairs was tested with advanced statistical study. After the successful results obtained in the mathematical representation and statistical proof phases, we did not make predictions for the future. We were content to give only brief opinions. With data and modeling, we aimed to stay within the determination of the current situation.

Our econometric study, which consists of two parts as production and market, covers the following details:

a) In the wheat production time series study, the nonlinear univariate Verhulst growth function applied for statistical tests.
b) In market section, first the time-based trends of quantities and prices then cyclical fluctuations around these trends are investigated.
c) Then, simultaneous solutions of quantity and price functions ascertained on the basis of simple cobweb and Goodwin's advanced cobweb theory based on expectations and two-phase delay.
d) Lastly, comments and results are presented.

In accordance with our aims, we have been determining trend and oscillations of the wheat production covering a period of more than a hundred years (1909-2020), ascertaining cobweb theory of price and quantity movements going back to a forty-year history of 1980-2020 and also including the foreign trade statistics with its twenty-year history (2000-2020).

2. Source and Explanation of Data Used

One of the oldest and most healthy statistics which goes back to Ottoman period in Türkiye is the amount of land sown and production of cereals and grains. TurkStat, which is at the same age as with Turkish Republic, has kept and published statistics for this group in a regular way ever since its establishment. Statistical publications series that we used in this research are "Statistical Yearbook of Turkey", "Statistical Indicators" and "The Summary of Agricultural Statistics". On the other hand, TurkStat also added agricultural statistics of Ottoman Period to its publications in 1997 with a separate book (Güran 1997). In this way, we have more than a century of data (1909-2020) of Turkish wheat agriculture.

3. Statistics and Econometry Application Procedures of SAS/STAT® and SAS/ETS®

With the advances made in statistical science over the last half century, the tests applied in data analysis have increased in addition to traditionel t-test, F-test, Durbin-Watson, correlation, determination coefficients etc. To mention a few, Godfrey's serial correlation, Shapiro-Wilk normality, and the application of generalized Durbin-Watson tests have gained space and become standardized. Theoretical statistical test studies on the structure of the mathematical functions used have also increased. On the other hand, in order for nonlinear mathematical functions to have sound results in statistical application, the conformity tests of these functions to the linearity have been developed. These include skewness, bias and global nonlinearity applications[1]. If the function does not appear to be linear when these tests are applied, then the necessary precautions should be taken to ensure linearity by mathematical transformations, simply by logarithmic and exponential applications, otherwise, the parameters found will be far from full representability. In our wheat study, in addition to the known tests, the volume of our article is expanding because we added the tests listed here to our outputs. However, we wanted a highly respectable and referrable econometric study and included tables in article as much as possible. Although we worked most of the estimation methods mentioned in the explanation of the SAS MODEL procedures below, but only included the "Full Information Maximum Likelihood" (FIML) estimation results here to be short. On the other hand, also collinearity tables, which take up a lot of space according to the volume of the article and the number of parameters, are not included in the article.

Econometric and statistic procedures of the SAS STUDIO library used are:

a) SAS PROC AUTOREG PROCEDURE:

   The AUTOREG procedure estimates and forecasts linear regression models for time series data when the errors are autocorrelated or heteroscedastic. The autoregressive error model is used to correct for autocorrelation, and the generalized autoregressive conditional heteroscedasticity (GARCH) model and its variants are used to model and correct for heteroscedasticity. (SAS Institute Inc. 2014,p. 302)

b) SAS PROC MODEL PROCEDURE:

   The MODEL procedure analyzes models in which the relationships among the variables form a system of one or more nonlinear equations. (SAS Institute Inc. 2018a,pp. 1423–1424)

PROC MODEL uses mainly following statistical calculation methods:

Ordinary least squares (OLS),

Seemingly unrelated regression (SUR) and iterative SUR (ITSUR),

Full information maximum likelihood (FIML),

Two-stage least squares (2SLS),

Three-stage least squares (3SLS) and iterative 3SLS (IT3SLS),

Generalized method of moments (GMM),

Simulated method of moments (SMM),

We extensively used test results of below given headings of SAS Model procedure:

Collinearity Diagnostics,

Nonlinear Summary of Parameter Estimates and Residual Errors,

Heteroscedasticity Test,

Godfrey's Serial Correlation Test,

Durbin-Watson Statistics,

Normality Test,

Structural Change Test: Chow test and

Dynamic Equation Simulation -Theil Statistics.

c) SAS PROC NLIN PROCEDURE: Nonlinear functions

[1]See (Gebremariam 2014; İskender 2018, pp. 103–105; SAS Institute Inc. 2020,pp. 6956–6964)

> The NLIN procedure fits nonlinear regression models and estimates the parameters by nonlinear least squaresor weighted nonlinear least squares…Nonlinear least-squares estimation involves finding those values in the parameter space that minimize the (weighted) residual sum of squares. (SAS Institute Inc. p. 6956, 2020)

NLIN procedure used in Verhulst groth function.

During our study, following statistical procedures also used when necessary.

d) SAS PROC GLM PROCEDURE
e) SAS PROC CAPABILITY PROCEDURE
f) SAS PROC UNIVARIATE PROCEDURE
g) SAS PROC ROBUSTREG PROCEDURE

4. Analysis of Turkish Wheat Production

With the increase in the amount of land sown since the establishment of the Republic, wheat production which is the main grain product, has had a continuous linear increase until 1975, after which the amount of land planted remained constant or decreased until 1995, while production amounts continued to increase with the increase in productivity, and since 1988 to the present day it has been fluctuating around nineteen million metric tons[2] per year. Depending on various factors, it is normal for agricultural products to fluctuate: especially weather and rain are the two most effective factors. The use of up to ten million hectares of arable land (1998) has decreased to 6.8 million hectares since then. The share of wheat in total arable grain land has been in the range of 60-70% in the last three decades. On the other hand, productivity per hectare starting from one tonne has reached to 2.96 tons today.

Since it is the primary source of nutritional needs as a traditional product and production is parallel to domestic consumption, it is worth explaining wheat production together with Turkey's population development. Turkey is not a major wheat exporter country. According to the annual consumption of production, imports were marginally realized in the years when supply was inadequate, and exports were marginal in the case of surplus supply. In other words, it is right to take wheat production as a function of the growing population. On the other hand, with the increase in per capita income, we also see that wheat production and consumption per population decrease with the increase in consumption of alternative and expensive nutrition such as; meat, milk and derivatives, vegetables and fruits, etc. It is possible to see the situation from Figure 2. Per capita production which increased up to 400 kg in the period of1965-1985, decreased steadily after this date to 200 kg.

Direct wheat export figures of Turkey are low: the figures for 2019 and 2020 are 135 and 125 thousand tons respectively. However, Turkey has imported wheat for input purposes in recent years and exported flour, pasta, bulgur, semolina biscuits etc. in significant quantities. Wheat imports increased from 3.8 million tons in 2013 to 9.6 million tons in 2020. As in export figures, also the minor parts of annual imports go to domestic consumption, but we know that most of them are used as inputs and exported as products. I aim to keep import and export information at this level, which is not our main topic.

[2] Throughout the article, the measure of ton means metric ton.

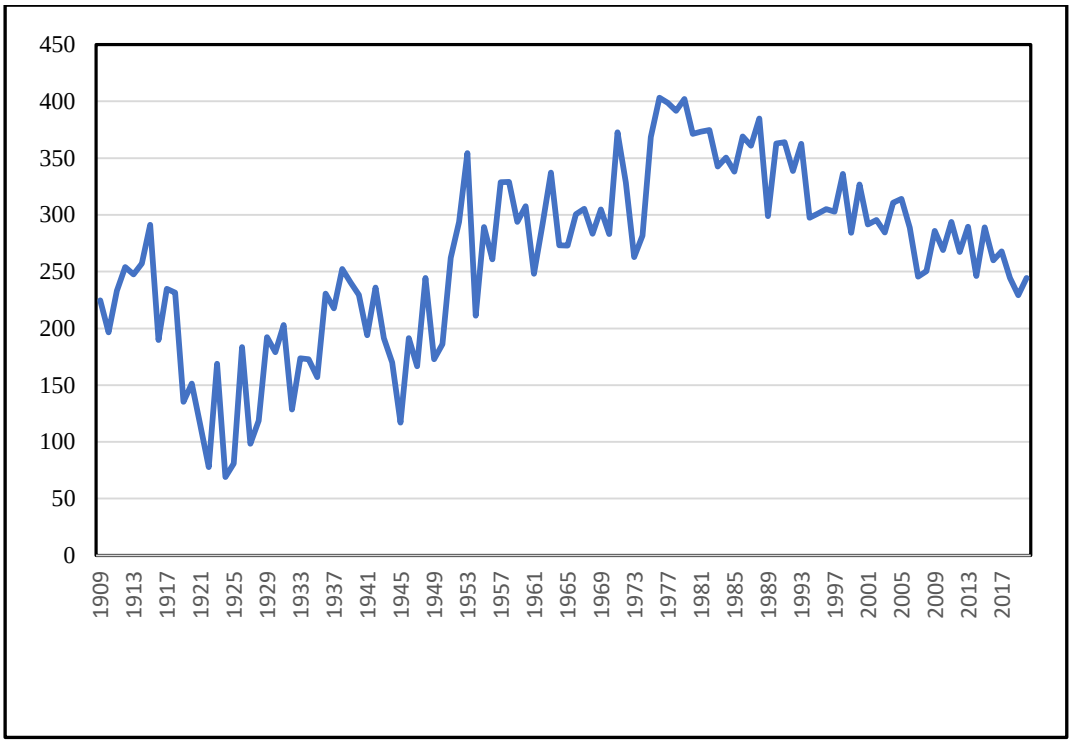


Figure 1: Per capita wheat production of Türkiye (Kg)

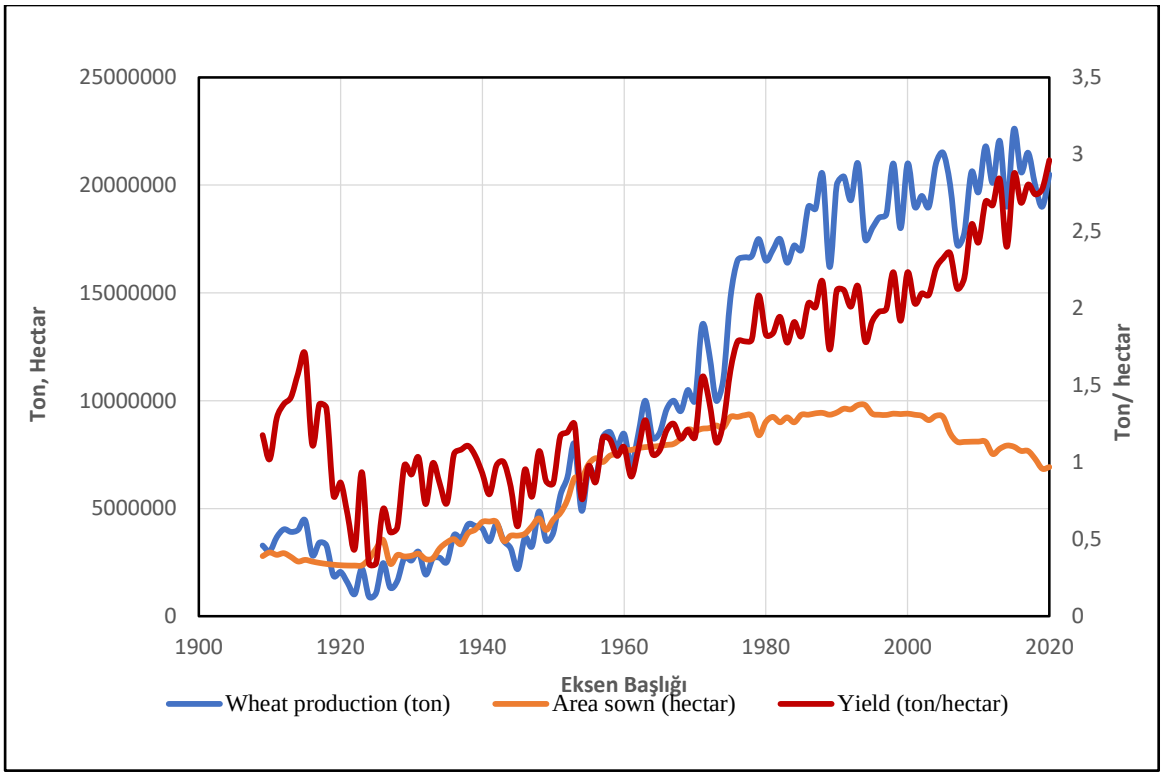


Figure 2: Wheat production and area sown of Türkiye (Production and area sown are shown at the lefthand side scale and yield on the right).

In mathematical modeling we applied Verhulst growth function and explained the development of wheat production in the geography of Türkiye. We have a data set of hundred and twelve years for the period 1909-2020 to use in modeling. TurkStat sources have a complete production data for the period 1924-2020. In a publication of TurkStat, we obtain statistics of the Ottoman Period of the years 1909, 1913 and 1914 (Güran

1997). Then we made estimates for the periods 1910-12 and 1915-1923 which cover extraordinary circumstances and have no data: For this purpose, we re-estimated the annual data for two periods applying Microsoft Excel (randbetween function) random numbers function in the range of plus minus seven hundred and fifty thousand tons to the function that is obtained from the fifth-degree polynomial curve fitting of the data of 1909-1950 period.

One of the reasons we included pre-republic period production figures in research is to obtain the data set of the lower asymptote region of the mathematical curve we use and obtain positive results from statistical application. In this way, the statistical application of growth curve has been successful. If we had subjected the wheat production figures for the period 1925-1988 to statistical analysis, there would have been a linear or close to linear production increase which will not reflect reality of the trend. However, adding statistics of Ottoman period and considering that the long-term production conjuncture has reached the upper asymptote region, it would be more accurate to think that wheat production is on a course following the Verhulst growth curve.

During our studies, we saw that the four-parameter Verhulst growth function would be suitable for ascertaining growth in time series analysis.

Symbols used:

Y: Wheat Production (dependent variable),

t: Time, (explanatory variable),

Z: Base year, 1940,

K: Upper asymptote,

L: Lower asymptote,

r: Intrinsic rate of growth,

Q: Coefficient of exponential growth base (*e*).

Verhulst function is a symmetric, non-linear structured mathematical growth function with two-variable, four-parameters (*K, L, Q, r*):

Function[3]:

$$Y_{(t-T)} = L + \frac{K - L}{1 + Qe^{-r(t-\mathrm{Z})}} \qquad \text{Eq (1)}$$

According to the statistical study, with 10% growth, the lower level of 2.6 million tons has been reached to the point where it is the upper level of 20.5 million tons. Production is not expected to grow any further. In more than a hundred years of data we used, significant weather-related fluctuations have been observed for every year around the main trend. Since we did not have series on the weather, especially about precipitation and drought, we could not include it as an independent variable in the function. Over the last half century weather-related fluctuations have continued around the trend in the band of plus minus 1.5 million tons. Time series explains 97.11% of the production. Trend growth, which was in a continuous increase until 1969, started to decrease from 1969 and reached zero growth at the level of 20.5 million tons. The price effect, which is important in balancing wheat production at the average trend level of 19.1 million tons, will be discussed in the market section. Wheat production rarely exceeded 4 million tons from 1909 to 1945, depending on weather conditions. The production capacity before the period tractors started to use in agriculture is around four million tons. From 1945 onwards, production has accelerated with the use of tractors and other agricultural equipment, and from 1960 onwards also with the increase in fertilizer and spraying usage. As we will see in the market section below, we expect production to continue at the trend of 19.1 million tons with the current domestic market wheat price, input prices and world wheat prices.

Wheat production function[4]:

[3] See (İskender, 2018, 2021a) for mathematical properties of function.

[4] Logarithmic and exponential equivalents of some figures used to achieve linearity.

$$Y_{(t-T)} = 2645954 + \frac{e^{16.8349} - 2645954}{1 + e^{e^{1.1284}}\, e^{(e^{-2.2463}(t-1940))}} \qquad \text{Eq (2)}$$

T=1940

Upper asymptote: $e^{16.8349} = 20478784$ tons

Starting level: $e^{e^{1.1284}} = 21.99$ tons

Intrinsic rate of growth: $e^{-2.4463} = 0.1058$

Lower asymptote: 2654954 tons.

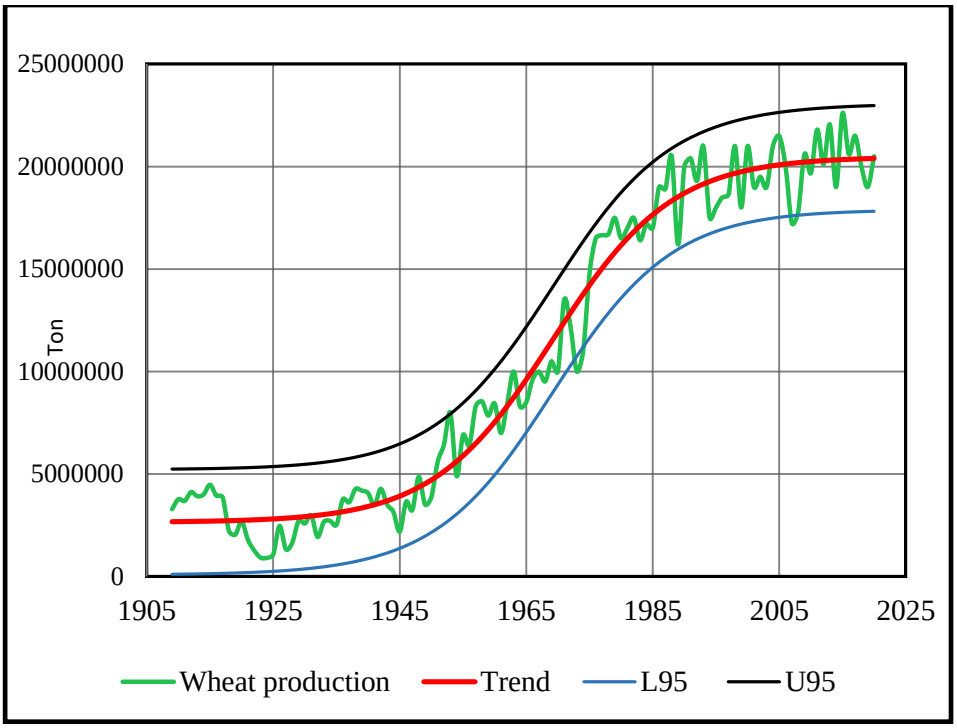


Figure 3: More Than a Century Wheat Production Trend of Türkiye

Linearity test results -skewness, bias and global nonlinearity measures- of Verhulst function given in Tables 1-3 below.

The guideline, according to Ratkowsky (1983), is that percentage bias greater than 1% is considered to be significantly nonlinear. Similarly, a value of the standardized Hougaard skewness measure greater than 0.25 in absolute value indicates nonlinear behavior. (Gebremariam 2014, p. 5)

According to Table 2, the skewness numbers of the parameters are below 0.25, bias values are below 1%, and global nonlinearity measures (table 3) are below curvature critical value which is 0.6381. Verhulst growth curve is very close to linear at the evaluation of Turkish wheat productin..

Table 1: Statistics of SAS NLIN Procedure

Source	DF	Sum of Squares Error	Mean Square	F Value	Approx Pr > F
Model	3	5.853E15	1.951E15	1208.65	<.0001
Error	108	1.743E14	1.614E12		
Corrected Total	111	6.028E15			

Table 2: Parameter and Linearity Estimates of NLIN Procedure

Parameter	Estimate	Approx Std Error	t value	Skewness	Bias	Percent Bias
L	2645954	256006	10.33552	-0.0984	-8021.4	-0.30
Log(r)	-2.2463	0.0732	-30.6872	0.0126	0.00160	-0.07
Log(log(Q))	1.1284	0.0770	14.65455	0.00281	0.00145	0.13
Log(K)	16.8349	0.0147	1145.231	0.1061	0.000451	0.003

Table 3: Global Nonlinearity Measures

Max Intrinsic Curvature	0.1307
RMS Intrinsic Curvature	0.0582
Max Parameter-Effects Curvature	0.2854
RMS Parameter-Effects Curvature	0.1241
Curvature Critical Value	0.6381
Raw Residual Variance	1614E9
Projected Residual Variance	1567E9

At the second stage, advanced statistical tests of both trend function and also complementary functions are added in order to explain sinusoidal variations around the trend line. Mathematical model of complementary function used is a two-component sinusoidal function and symbols as follows:

F_t *:* Conjunctural oscillations, Dependent variable,

t-Z: Time, independent variable adjusted with base Z,

$\pm A$: Initial amplitude of oscillation or peak and trough coefficient,

r: Damping or anti-damping multipliers, growth rates,

θ_1 , θ_2 are angular frequencies,

ε_1, ε_2 are phase lags

And numerical form.

$$F_t = A_1 e^{(r_1(t-Z))} cos(\theta_1(t-Z)+\varepsilon_1) + A_2 e^{(r_2(t-Z))} cos(\theta_2(t-Z)+\varepsilon_2) \quad \text{Eq (3)}$$

With the addition of complementary function, the sum of square error decreased from 1.743E14 to 1.2E14. The function accounts for 98% of production. Although we tried various mathematical models, it has not been possible to further reduce the sum of errors in production which followed a very fluctuating course. Thorughout the statistical study we added four more variables to eliminate the effects of autogression.

Trend function (particular integral, equilibrium path)[5]:

$$\bar{Y}_t = 2755058 + \frac{e^{17.02301} - 2755058}{1 + e^{e^{1.087176}\, e^{(e^{-2.30403}(t-1940))}}} \quad \text{Eq (4)}$$

Two-phase complementary function:

$$F_t = -400000e^{(0.0104(t-1940))}cos(0.50795(t-1940) + 2.635506) + 1150000e^{(-0.0187(t-1940))}cos(0.288015(t-1940) + 0.917351) \quad \text{Eq (5)}$$

$A_1 = -400000$, $A_2 = 1150000$, $r_1 = 0.0104$ and $r_2 = -0.0187$ taken as given.

,Autoregression corrections:

$$AR_t = 0.599928\text{W\_l1} - 0.79948\text{W\_l2} + 0.525013\text{W\_m1} - 0.71312\text{W\_m2} \quad \text{Eq (6)}$$

General solution (Complete primitive) consists of addition of three functions:

$$Y_t = \bar{Y}_t + F_t + AR_t \quad \text{Eq (7)}$$

SAS MODEL procedure FIML application Tables 4-8:

Table 4: Nonlinear FIML Summary Table of Residual Errors

Equation	DF Model	DF Error	SSE	MSE	Root MSE	$R^2$and R	Adj R^2	Durbin Watson
Wheat	12	100	1.2E14	1.2E12	1095593	0.9801 0.9900	0.9779	2.1597

[5] Since SAS MODEL procedure applies autoregression corrections, it is clear that the trend function parameters of eq. (4) differ slightly from those in NLIN application eq (2).

Table 5: Parameter Estimates

Parameters	Estimates	Approx Std Err	t Value	Approx Pr > |t|	Label
L	2755058	626481	4.40	<.0001	Lowest production level of wheat,
Log(log(Q))	1.087176	0.0790	13.77	<.0001	Exp of exp of e base coefficient
Log(r)	-2.30403	0.0688	-33.47	<.0001	Log of intrinsic rate of growth
Log(K)	17.02301	0.0970	175.52	<.0001	Highest level of production,
θ_1	0.50795	0.0107	47.64	<.0001	First angular frequency
ε_1	2.635506	0.5931	4.44	<.0001	First phase lag
θ_2	0.288015	0.00779	37.00	<.0001	Second angular frequency
ε_2	0.917351	0.2592	3.54	0.0006	Second phase
W_l1	0.599928	0.1573	3.81	0.0002	Lag1 parameter of dependent variable wheat
W_l2	-0.79948	0.1491	-5.36	<.0001	Lag2 parameter of dependent variable wheat
W_m1	0.525013	0.1906	2.76	0.0070	MA lag1 parameter of residuals (error term)
W_m2	-0.71312	0.1733	-4.11	<.0001	MA lag2 parameter of residuals (error term)

Notes of Table 5: a) Since all the p-values of parameters are lower than confidence level of α=0.05, the null hypothesis that parameters meaningless is rejected. b) Last four parameter applied for autoregressive correction of series. W_l1 and W_l2 are applied to dependant variable wheat instead of structurel residuals of equation. c) For the definitons and details of AR: Autoregressive Errors and MR: Moving-Average Models see (SAS Institute Inc. 2018a, pp. 1565–1580).

Table 6: Heteroscedasticity Test

Equation	**Test**	**Statistic**	**DF**	**Pr > ChiSq**	**Variables**
Wheat	White's Test	105.3	89	0.1143	Cross of all variables
	Breusch-Pagan	0.36	1	0.5484	1, E1

Note to Table 6:

$E = 1 + e^{e^{1.087172}}\, e^{e^{-2.30403(t-1940)}}$

$E1 = E + F$

The White test tests the null hypothesis. H_0: $\sigma_i^2 = \sigma^2$ for all i. The p-values>α=0.05 means that there is no heteroskedasticity.

Table 7: Serial Correlation and Autocorrelation Tests

Godfrey's Serial Correlation Test				Generalized Durbin-Watson Statistics				
Equation	**Alternative**	**LM**	**Pr > LM**	**Equation**	**Order**	**DW**	**Pr < DW**	**Pr > DW**
Wheat	1	1.81	0.1780	**Wheat**	1	2.16	0.4964	0.5036
	2	3.30	0.1920		2	1.76	0.0654	0.9346
	3	3.34	0.3424		3	2.18	0.8133	0.1867

Note of Table 7:a) There is no serial or auto correlation for p values greater than α=0.05 in Table 7. b) Pr < DW is the p-value for testing positive autocorrelation and Pr >DW is the p-value for negative autocorelation, also at Tables 16 and 23.

Table 8: Normality Tets

Equation	Test Statistic	Value	Prob
Wheat	Shapiro-Wilk	0.98	0.1400
System	Mardia Skewness	1.92	0.1657
	Mardia Kurtosis	0.04	0.9691
	Henze-Zirkler T	0.59	0.1361

Note of Table 8: The null-hypothesis of this test is that the population is normally distributed. Thus, if the p value is less than the chosen alpha level (which is 0.05 in our case), then the null hypothesis is rejected and there is evidence that the data tested are not normally distributed. https://en.wikipedia.org/wiki/Shapiro%E2%80%93Wilk_test

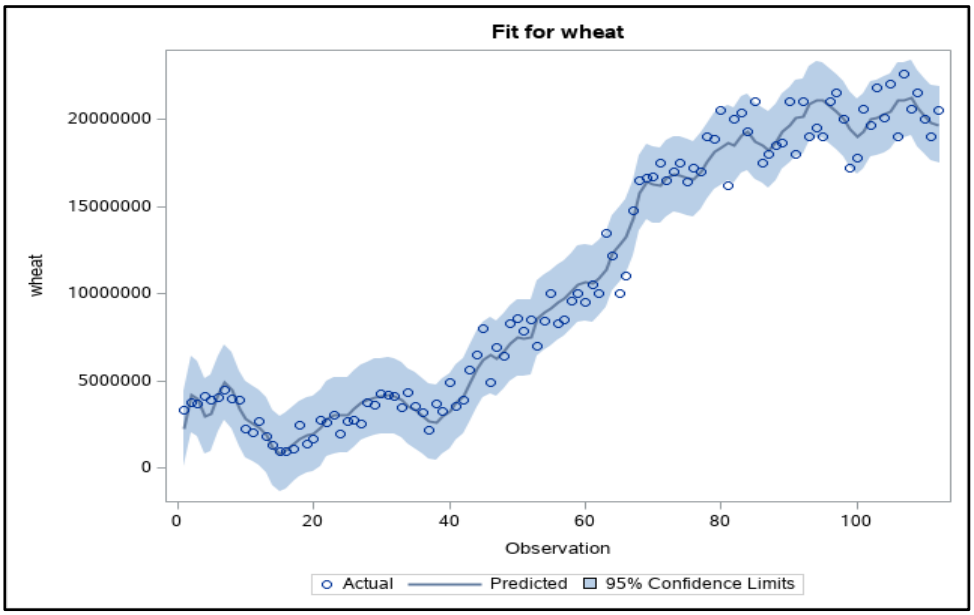


Figure 4: Curve Fitting of Wheat Production (SAS Model Procedure output)

5. Quantities and Prices: General Wheat Market Equilibrium of Türkiye

5.1. Methods

In this part of our study, we investigated the relationship between real wheat prices and traded quantities, with linear and nonlinear functions, simple dynamic cobweb theory, and also advanced cobweb model of Goodwin which includes producer expectations and two period lag of prices. For this purpose, we have compiled wheat producer prices, Producers Price Index (PPI), wholesale price index, production quantities, direct and indirect import and export statistics from the sources of TurkStat and made them ready for our analysis. Our statistical study cover the 1982-2020 period. We also included data from 1980 and 1981 where delayed variables were required. In this way, we have determined the price and supply-demand (traded quantities) trend of the wheat economy with reliable statistical applications for a period of forty years.

Since wheat production in Türkiye is essentially a primary product intended to meet domestic household food consumption and approximately meets the need in quantity, imports and exports figures of Türkiye have remained at marginal levels over decades. Therefore, ignoring the export and import figures and also stocks of which we do not have figures, we have assumed *production = domestic consumption* for the period 1982-1999.

Since 2000, for Türkiye's wheat imports have shown a serious upward trend as a result of growing export-oriented industries -flour, pasta, bulgur, semolina, biscuits and wafers producers- using wheat and flour as intermediate inputs, we have included direct wheat imports and exports figures, and wheat equivalents of exports and imports of the products listed above in the calculations in our work for the period 2000-2020. This approach have aslo been followed by the authors in sector (Polat 2020;Tarım ve Orman Bakanlığı, 2020, 2021). By defining *production +direct import-direct export-indirect export*, we obtained the 2000-2020 series and together with the data of 1982-1999 period, acquired thirty-eight-year quantity series as basis for statistical study.

We studied wheat prices for the period 1982-2020 that compiled from the web pages of Soil Produces Office of Türkiye (TMO), Polatlı Commodity Exchange, Konya Commodity Exchange and TurkStat, and observed that the series were very close to each other and decided to use TurkStat wheat prices together with the PPI (producer price index) index for statistical application. Starting from 1982, we obtained annual real wheat prices as *TurkStat's current price data received by wheat producer /Producer price index (PPI)* and matched this series with quantity figures. Our statistical calculations have been based on these two series.

In the market section of our article, which is already large enough, the linearity studies of the functions are not included. The statistical issues that will extend to the theoretical studies that we have faced in sinusoidal functions have also been effective in this decision.

5.2. Determination of Price and Quantity Trends of Wheat Market

The quantities traded in the market and the corresponding prices for these amounts for the period under review are given in Figure 5. The conclusions we obtained from Figure 5 and statistical study results are as follows:

The average of the real wheat price for the period 1982-2020 is 3.1 (*standart deviation: 0.29*) real lira. Over the period of nearly four decades, prices have hovered around this average. The relative share of wheat producers in the economy has not changed from price point of view. There have been years when the actual price has fallen as far below the average as it has been the years when it has risen above average. The real price, which was 3.1 in 1982, is still 3.1 real lira in 2020. Wheat prices, which are in the category of prices monitored by public institutions, have maintained their relative share in the economy. Although it retains its share in the economy, the first half of the 1982-2000 is the period of increase in the average price, while the second half 2001-2020 is the period of decline of the average price: According to our calculations, the actual price is 3.02 real lira in the second half, which was 3.18 real lira in the first period. The price, which fell below 3 real liras in the 2017-19 period, has recovered in 2020 and reached the average. The best year is 1996. This actual price has not been reached in subsequent years. The direction of the trend is still downwards. It is necessary to make a trend detection again with the data of the coming years. The outlook for manufacturers is not good. The state of affairs that is in accordance with economic theory is that prices fluctuate around the average over a long period of nearly forty years. This is the presumption of economic theory. It will be discussed in detail below. It is our personal opinion that the real price will be below average or at most as average in the coming periods.

On the other hand, it is also clear that the price of wheat in Türkiye has been under the influence of the world wheat price of $ 250 / tons. Imports will become attractive when the domestic price rises. The producer's reaction to the lack of real prices was to reduce the amount of land cultivated, in other words, to exclude the land from production that did not cover the cost of 3.1 liras. The amount of land cultivated in 2000 decreased from 9.4 million hectares to 6.9 million hectares in 2020, and it has been possibile to continue production at the level

of 19.1 million tons in the same period with the increase in productivity. The hectare yield which was 2.23 tons in 2000 increased to 2.96 tons in 2020 (see Figure 2).

Goodwin wrote:

> It is always assumed that decisions to produce lead to the same output at a later date, but this is not true, especially in agriculture, where the output of most crops is heavily affected by the weather. Thus acreage may consistently be controlled by prices, but yield per acre will have a nonsystematic or random character. (Goodwin 1947,pp. 186–187)

The39-year average of the supply and demand quantities is 19,125 thousand tons and this figure is the functional value of wheat supply and demand trends. Although this is the average, the amount increased from 18.66 million tons in the first half of the period to 19.56 million tons in the second half. There is an upward trend. During the period in question, traded amounts followed a sinusoidal course around the average amount but did not depart from the trend. Although there is an increase in the second half of the period, it is too early to see this as a trend change.

When Figure 5 is examined, it will be seen even with the naked eye that both price and quantity trends fluctuate sinusoidally around a fixed and horizontal axis of their own, each time they turn to trend lines – the invisible hand carefully manages state of affairs – and that cyclical periods are almost equal. This course of the charts has highly forced our econometric studies and has only allowed the *determination coefficient* in price and quantity functions to be obtained as 0.71 and 0.66, respectively in simple cobweb study. It's almost as if there is seemingly no relation between the two variables! Or everyone agree with administrative prices.

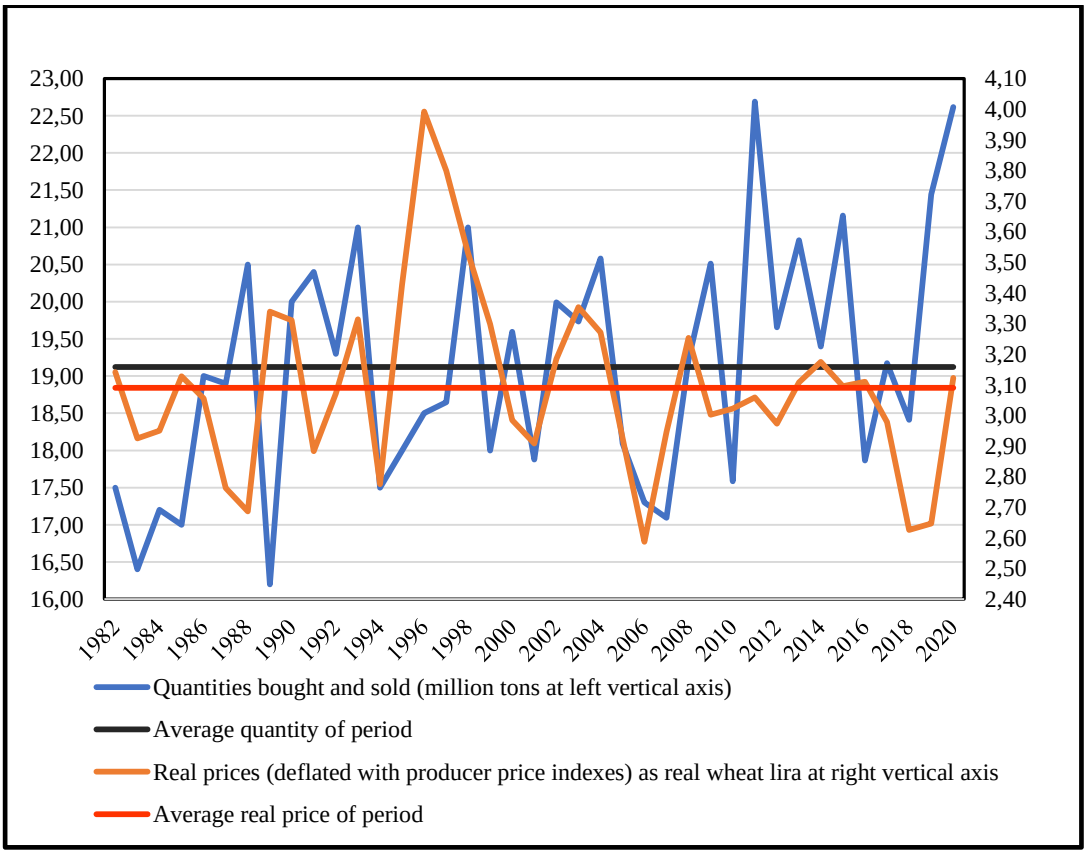


Figure 5: Wheat Quantities Bought and Sold and Price Trends of Türkiye

Table 9: Price Statistics of Wheat

N	39	**Sum Weights**	39
Mean	3.09539667	**Sum Observations**	120.72047
Std Deviation	0.29176979	**Variance**	0.08512961
Skewness	0.84673567	**Kurtosis**	1.71768165
Uncorrected SS	376.912666	**Corrected SS**	3.23492514
Coeff Variation	9.42592562	**Std Error Mean**	0.04672056

Table 10: Normality Test of Wheat Price

Test		**Statistic**	**p Value**	
Shapiro-Wilk	**W**	0.947530	**Pr < W**	0.0679

Table 11: Quantity Statistics of Wheat

N	39	**Sum Weights**	39
Mean	19.1247692	**Sum Observations**	745.866
Std Deviation	1.63533208	**Variance**	2.67431102
Skewness	0.26503539	**Kurtosis**	-0.5607078
Uncorrected SS	14366.1389	**Corrected SS**	101.623819
Coeff Variation	8.55085917	**Std Error Mean**	0.26186271

Table 12: Normality Test of Wheat Quantities

Test		**Statistic**	**p Value**	
Shapiro-Wilk	**W**	0.976315	**Pr < W**	0.5712

A point that we think will be of interest to the reader about the prices is the following Figure 6 comparing the prices received by the selected countries'producers': The wheat producer prices of Türkiye, United States, Australia, Canada and Argentina included in chart and also producer prices deflated with PPI are given in Figure 6. While Türkiye's producer price deflated by the dollar exchange rate of Turkish lira followed these countries' producer prices, the producer prices deflated with PII remained constant at the level of 3.1. In last fifteen years

they caught each other. This means Turkish producers have lost their real lira relative advantage when compared with dollar based prices.

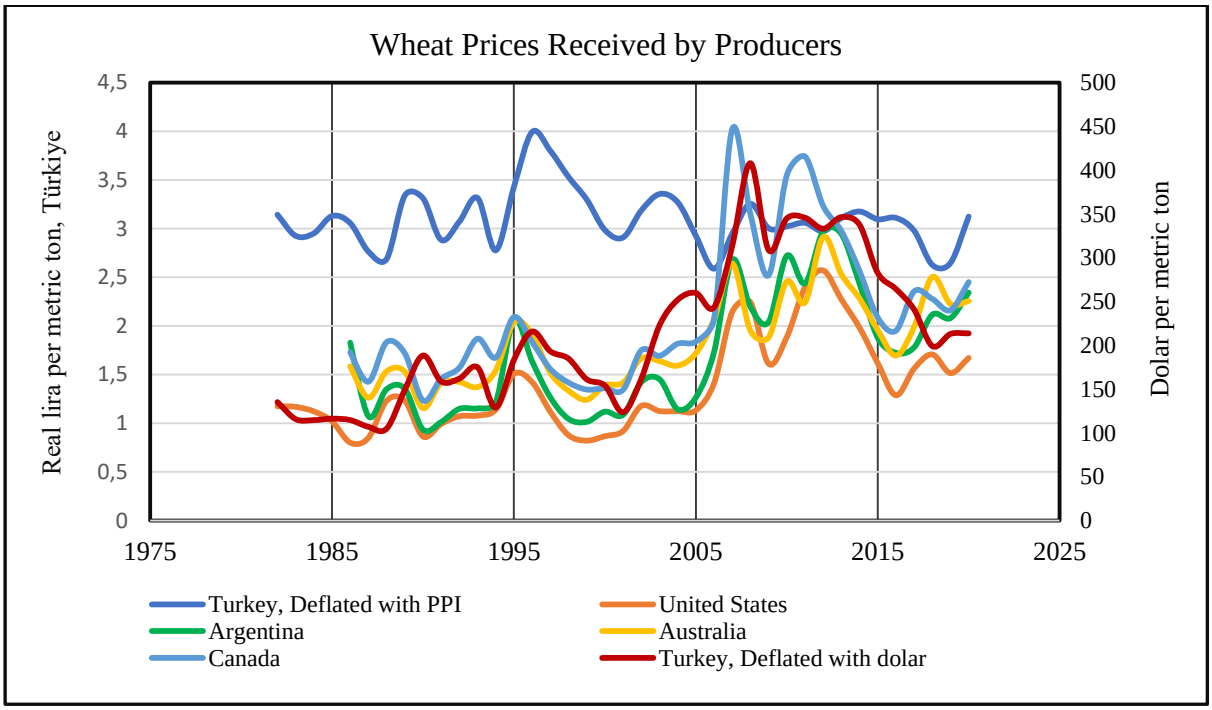


Figure 6: Wheat Prices of Selected Countries (Source: https://www.ers.usda.gov and TurkStat websites).

After determining the main trend of the prices and quantities of the wheat, we worked on the fluctuations around the trend with mathematical models that is explained in two sub-headings we have given below; i) Simple, dynamic cobweb theorem application that explains the price with last year's price, and, ii) Analysis with the advanced cobweb theorem (Goodwin 1947) which includes producer expectations in addition to two period delayed supply prices.

5.3. Simple Cobweb Theorem

Econometric analysis is the support of the hypotheses put forward by economic theory (wheat prices and quantities which is our subject) with equations established mathematically (differential equations and difference equations) and making them solvable and testing the equations with statistical methods and interpreting the results.

Adhering to approach, we developed mathematical polinomial functions with a lagged explanatory variable in simple cobweb theorem and used these functions with advanced statistical methods.

The annuall quantities of economic variables such as national income, price, production, etc. are never exactly on a linear line. In other words, annual growth rates are not at the same pace. Even if the equation we obtain as a result of our statistical study is linear, there will be variable figures that are not on this line. In this case, in addition to the function that determines the main trend, it is necessary to monitor the course of the data not included in the equation and to investigate the determination of the secondary equation (sinusoidal or cyclical complementary analysis).

Throgouhout the statistical tests, we ascertained that the parabolic fourth-degree function which has previous year's price as independent variable together with the sinusoidal complementary function had very high ability to represent the price trend. For the quantities function, second-degree price multiplied by time series as independent variable gave the best results in polinomial framework.

Symbols:

Price function :

P_t : Curent price, dependent variable,

P_{t-1}: Last year's price, independent variable,

t: Time, independent variable,

Z: Time base,

$a_1, a_2,$: Parameters,

a_3 , a_4 Constants of amplitude,

r_p: Growth rate,

a_5 : Angular frequncy.

Quantities function:

Q_t: Current quantity, dependant variable,

P_t : Curent price, independent variable,

t: Time, independent variable,

b_1, b_2 : Parameters,

b_3 : Amplitude,

r_q : Damping or anti-damping multipliers, growth rate,

b_4, b_5 angular frequency and phase lag of cos.

b_6, b_7 angular frequency and phase lag of sin.

If *r>1* anti-damped or explosive oscillations occur, when α <0 amplitude damped and when *r=1* amplitude has regular oscillations.

Trend and complementary functions used for price and quantities respectively are[6]:

$$P_t = a_1 + a_2 P_{t-1}^5 + a_3 \left[r_p^{(t-Z)} [cos(a_4(t-Z) + cos(a_5(t-Z) + a_6))] \right] \quad \text{Eq (8)}$$

$$Q_t = b_1 + b_2 P^3(t-Z) + b_3 \left[r_q^{(t-Z)} [\cos(b_4(t-Z) + b_5 + \sin(b_6(t-Z) + b_7))] \right] \quad \text{Eq (9)}$$

Last terms of both functions are complementary functions. Considering the characteristics of the data set used, we saw that complementary functions were more in weight than the trend functions. Although trends can be represented by mean values, it became necessary to make the constants of complementary functions exponential-based both in price and in quantity. In this way, determination coefficients obtained reached to very desirable levels (table 13).

Before work out the parameters of these functions, looking at the actual price and quantity diagram at Figure 7 (blue line) will be more instructive. The presumption of naming diagram as socalled cobweb is obvious. Two variables of diagram have been keeping going around in circles the actual price of 3.1 lira and the amount of 19.1 million tons for forty years. The graph shows the years from 1 to 39 in numbers for ease of tracking (1982=1… 2020=39). Our job as econometrician is to represent the cobweb route on the graph with mathematical functions and statistical applications.

[6] For details of theory and our recent application of complementary functions see (Allen 1956, p. 187-191; İskender 2021b) respectively.

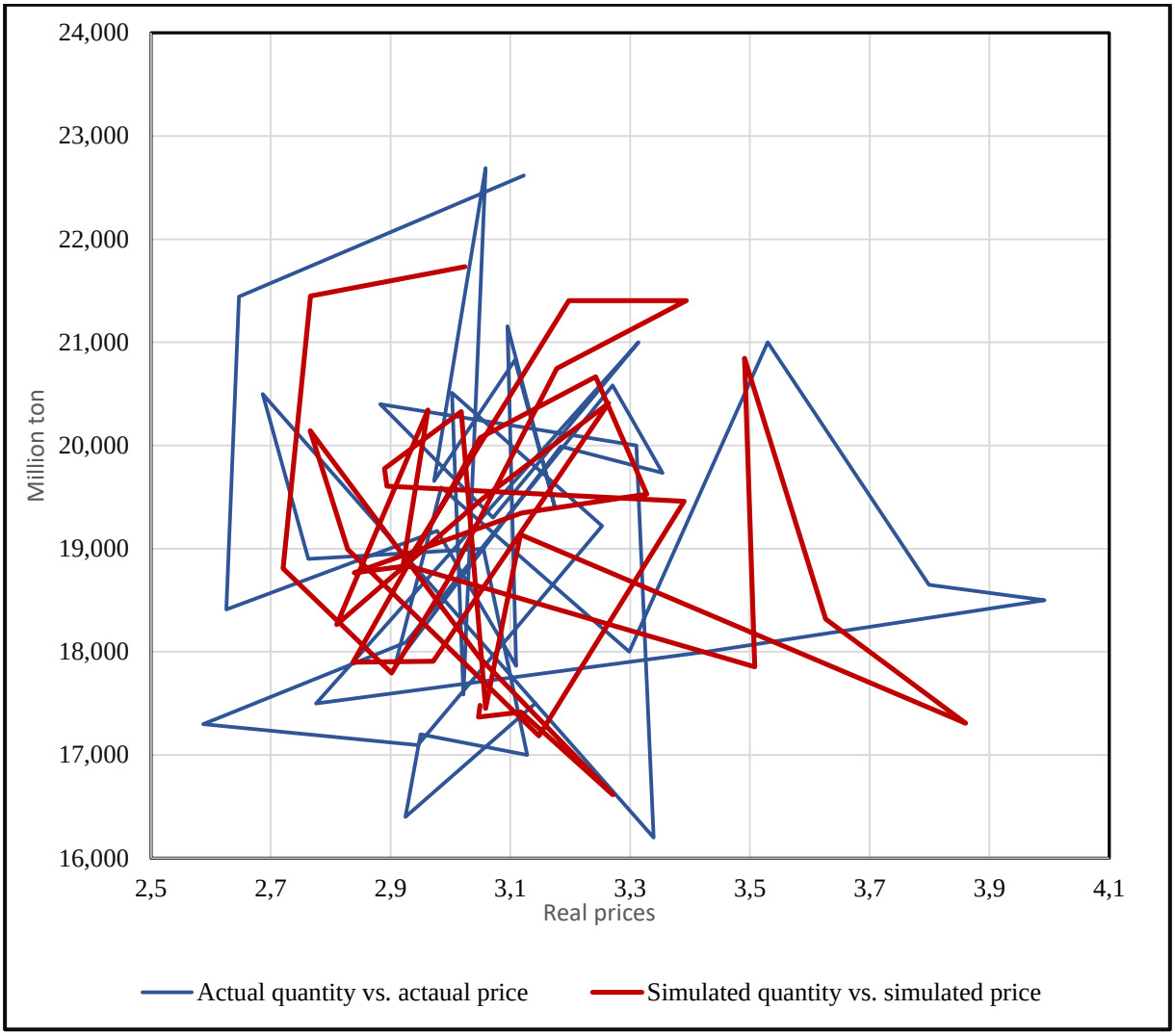


Figure 7: Harmony of Actual and Simulated Data : Our trend functions with the complementaries almost catch the actual data. Gravity center of state of affairs is 3.1 real lira on horizontal and 19.1 million tons wheat on vertical axis.

Based on the cobweb theorem, we took the current price as a function of the price of a year ago. In statistical studies, we obtained the best results from the fifth-degree polynomial application. In determining the quantity equation, the current price to the power of three multiplied by time is taken as independant variable, for in the simultaneous solution of only price-based supply equation did not yield the desired results at statistical tables. It seems that there has been no relationship simply between price and supply. This situation can be explained when we consider that the price of wheat is subject to very strict public monitoring, that there is no full competition, that its producers use the powers of institutional organization as well as the market to determine prices, that consumers also strictly control the prices of wheat made final products, etc. In summary, we preferred to work with the function based on price multiplied by time instead of price alone. Both functions are solved simultaneously.

As t: (1982…2020) is time and *Z*=1999 is base year, trend functions of prices and quantities:

$$P_t = \qquad \text{Eq (10)}$$
$$1.72127 + 0.00116P_{t-1}^{5} +$$
$$0.812015^{[0.998694^{(t-1999)}cos(1.02619(t-1999)+\cos(0.329373(t-1999)+2.152085))]} \_$$
$$0.43113P\_l2 + 0.454813P\_m3$$

$$Q_t = (17.81252 + 0.001902P^3(t-1999) + 0.375931^{\left[1.002928^{(t-1999)}[\cos(1.119806(t-1999)-1.64039+\sin(0.681446(t-1999)+1.106512))]\right]} + 0.704867Q\_l2 - 0.55422Q\_l4 \quad \text{Eq (11)}$$

We know from Figure 4 that the price and quantity curves run parallel to the horizontal (time) axis, and both functions almost explain the complete trends. Price and quantities are constant averages throughout the period. In Table 14, we may see those constant terms, α_1 and b_1. The average price trend of 3.1 (table 9) real lira versus the function constant value is 1.7 real lira. For quantities, it is 17.8 versus 19.1 (table 11). But on the other hand, although the parameters (a_2 and b_2) of independent variables are close to zero, their standard errors are very low and their significances are high. Price increases made every year on the basis of the previous year's price are effective in determining the current prices. And time is also an effective factor in explaining the course of quantity transections. This leads us to Goodwin's theory of expectations (next chapter). Although the coefficient of independant variable of quantity equation is small, the relationship between quantity and multiplication of time and square of current price as independent variable is high. Hence, the current price follows a course which is based on price of a year ago and the quantity in accordance with price and time variables.

We have obtained positive results from the statistical solutions made with OLS[7], SUR, FIML and ITSUR methods for these two functions, but preferred to give only SAS MODEL procedure[8] FIML (Full Information Maximum Likelihood) simultaneous solution results here with complementary functions of price and quantity functions for the sake of brevity.

Both the equilibrium path and the complementary functions represent the best values reached in statistical studies. We presume that the determination coefficients of the price and quantity equations obtained in our model (0.76 and 0.67) are satisfactory when the particular structure of the Turkish wheat market is taken into consideration.

As we mentioned earlier, it is difficult to measure supply and demand movements with annual data of wheat market which is under serious control by public authorities. If we had daily, weekly or monthly supply, demand and price data of Polatlı, Konya and Ankara Commodity Exchange in addition to the annual data, we could determine the supply and demand functions of these variables based on coefficients as well as time and price trend. Since it is not, what we do in this study consists of determining the path followed by the intersection points of supply and demand functions for a period of forty years.

With the determination of equations, also determined the conjuncture periods of the prices and quantities:

M: Oscillation time,

Price curve: $M_P = 2\pi/a_5$=2π /0.998694= 6.3 years,

Quantity curve: $M_q = 2\pi/b_4$=2π/1.002928= 6.3 years

The fact that cyclical durations of price and quantity functions are very close to each other indicates that the Turkish wheat market has a very stable appearance. Prices and quantities complete their the conjuncturel oscillations in average every six year. The presence of growth factors (*r*) which are close to one both in price and quantity complementary functions also indicate a stable structure, a regular oscillation (see *r* values at Table 14).

Tables 13-17 shows the results of FIML application. After simultaneous solution, the simulation values and graphs of the intersection points of these two equations are given in tables 18 and 19 and graphs 8 and 9, respectively.

In price and quantity equations with nineteen parameters, probabilities are almost at zero levels and the market representation of parameters is very high. At the following tables, the probabilities obtained are well above the 5% alpha level. The obtained test results confirm the validity of the simple cobweb theory of Turkish wheat market.

[7] For abbreviations see section 3b.

[8] For a very perfect and useful comparison of least squares and likelihood methods see (SAS Institute Inc. 2018b,pp. 37–43).

Table 13: Summary of Residuals

Equation	DF Model	DF Error	SSE	MSE	Root MSE	R^2 and R	Adj R^2	Durbin Watson
P[9]	9	30	0.7914	0.0264	0.1624	0.7553 0.8691	0.6901	1.9599
Q	10	29	33.9440	1.1705	1.0819	0.6660 0.8161	0.5623	2.4714

We think that the determination coefficients 0.7464 and 0.6680 provide a good representation considering seemingly simple nature of the data. It is obvious that the contribution of the application of variable-based amplitude coefficients, which we applied for the first time, has primary role.

[9] P: Price and Q: Quantity throughout the manuscript.

Table 14: Nonlinear FIML Parameter Estimates

Parameter	Estimate	Approx Std Err	t Value	Approx Pr > |t|	Label
P					
a_1	1.72127	0.0482	35.73	<.0001	Parameter
a_2	0.00116	0.000137	8.47	<.0001	Parameter
a_3	0.812015	0.0216	37.66	<.0001	First constant of amplitude
a_4	1.02619	0.0122	83.93	<.0001	Second constant of amplitude
r_p	0.998694	0.0120	83.50	<.0001	Growth factor
a_5	0.329373	0.0194	16.99	<.0001	Angular frequency
a_6	2.152085	0.2335	9.22	<.0001	Phase lag
P_l2	-0.43113	0.1312	-3.29	0.0026	AR(P) P lag2 parameter
P_l3	0.454813	0.1297	3.51	0.0015	AR(P) P lag3 parameter
Q					
b_1	17.81252	0.1869	95.32	<.0001	Parameter
b_2	0.001902	0.000605	3.15	0.0038	Parameter
b_3	0.375931	0.0360	10.45	<.0001	Initial amplitude
r_q	1.002928	0.00774	129.58	<.0001	Growth factor
b_4	1.119806	0.0128	87.45	<.0001	Angular frequency of cos
b_5	-1.64039	0.1282	-12.80	<.0001	Phase lag of cos
b_6	0.681446	0.0295	23.10	<.0001	Angular frequency of sin
b_7	1.106512	0.2143	5.16	<.0001	Phase lag of sin
Q_l2	0.704867	0.1372	5.14	<.0001	AR(Q) Q lag2 parameter
Q_l4	-0.55422	0.1393	-3.98	0.0004	AR(Q) Q lag4 parameter

Note of Table 14: P_l2, P_l3, Q_l2 and Q_l4 are coefficients of correction for serial autocorrelation See Table 5c note.

Table 15: Heteroscedasticity Test

Equation	Test	Statistic	DF	Pr>ChiSq	Variables
P	White's Test	39.00	38	0.4246	Cross of all vars
	Breusch-Pagan	5.25	5	0.3865	*S1*=(t-1999)(t-2015), *Q, P, 1*
Q	White's Test	39.00	38	0.4246	Cross of all vars
	Breusch-Pagan	2.71	5	0.7447	*S1*=(t-1999)(t-2015), *Q, P, 1*

Table 16: Autoregression Test Results of Prices and Quantities

Godfrey's Serial Correlation Test				**Generalized Durbin-Watson Statistics**				
Equation	**Alternative**	**LM**	**Pr> LM**	**Equation**	**Order**	**DW**	**Pr<DW**	**Pr> DW**
P	1	0.20	0.6586	**P**	1	1.96	0.2545	0.7455
	2	0.63	0.7280		2	2.09	0.8195	0.1805
	3	0.71	0.8701		3	1.95	0.7331	0.2669
Q	1	3.13	0.0767	**Q**	1	2.47	0.8653	0.1347
	2	4.44	0.1089		2	1.68	0.3257	0.6743
	3	4.44	0.2177		3	2.13	0.7483	0.2517

Table 17: Normality test

Equation	Test Statistic	Value	Prob
P	Shapiro-Wilk W	0.99	0.9225
Q	Shapiro-Wilk W	0.99	0.9021
System	Mardia Skewness	2.01	0.7348
	Mardia Kurtosis	-0.86	0.3871
	Henze-Zirkler T	0.37	0.6875

After determining the parameters of the price and quantity functions, we carried out the simulation work required for the comparison and graphing of the function estimates with the real data. Tables 18 and 19 and figures 8, 9 and 10 of this work give us the opportunity of comparison.

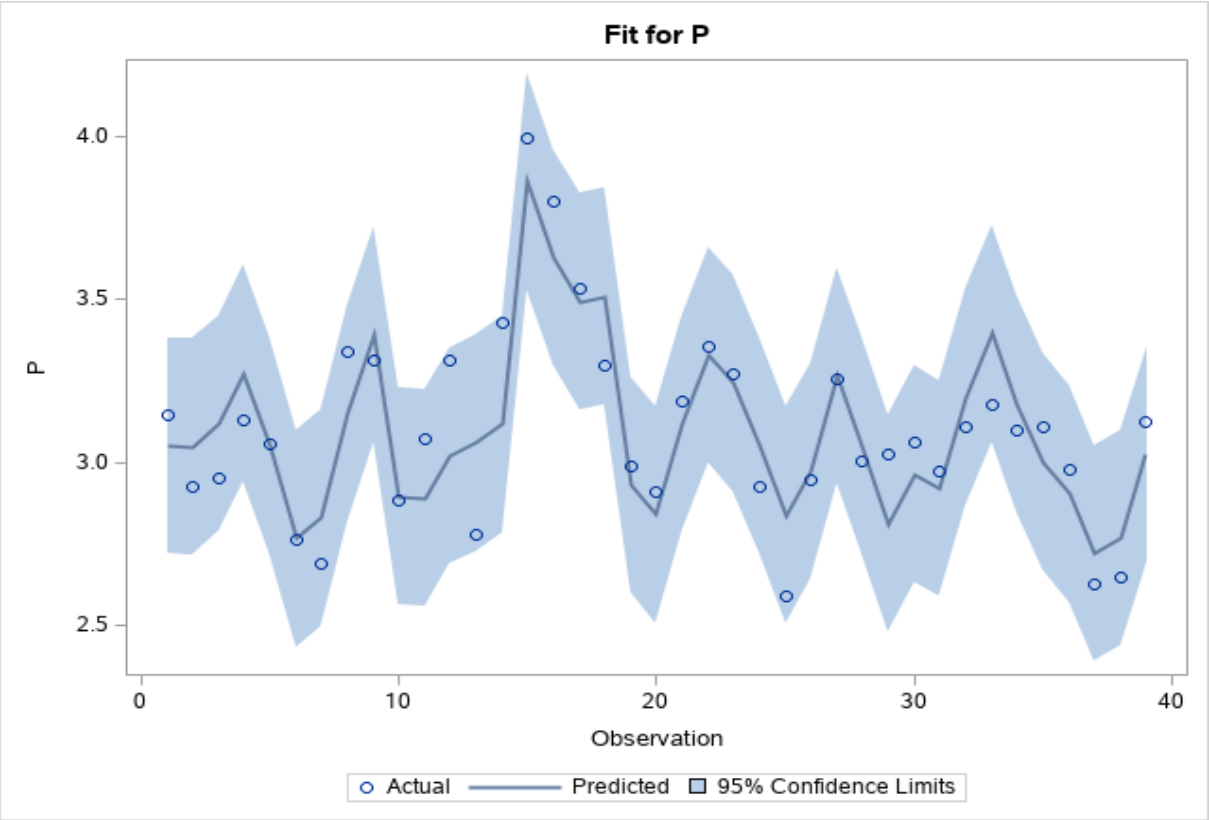


Figure 8: Fit for Price

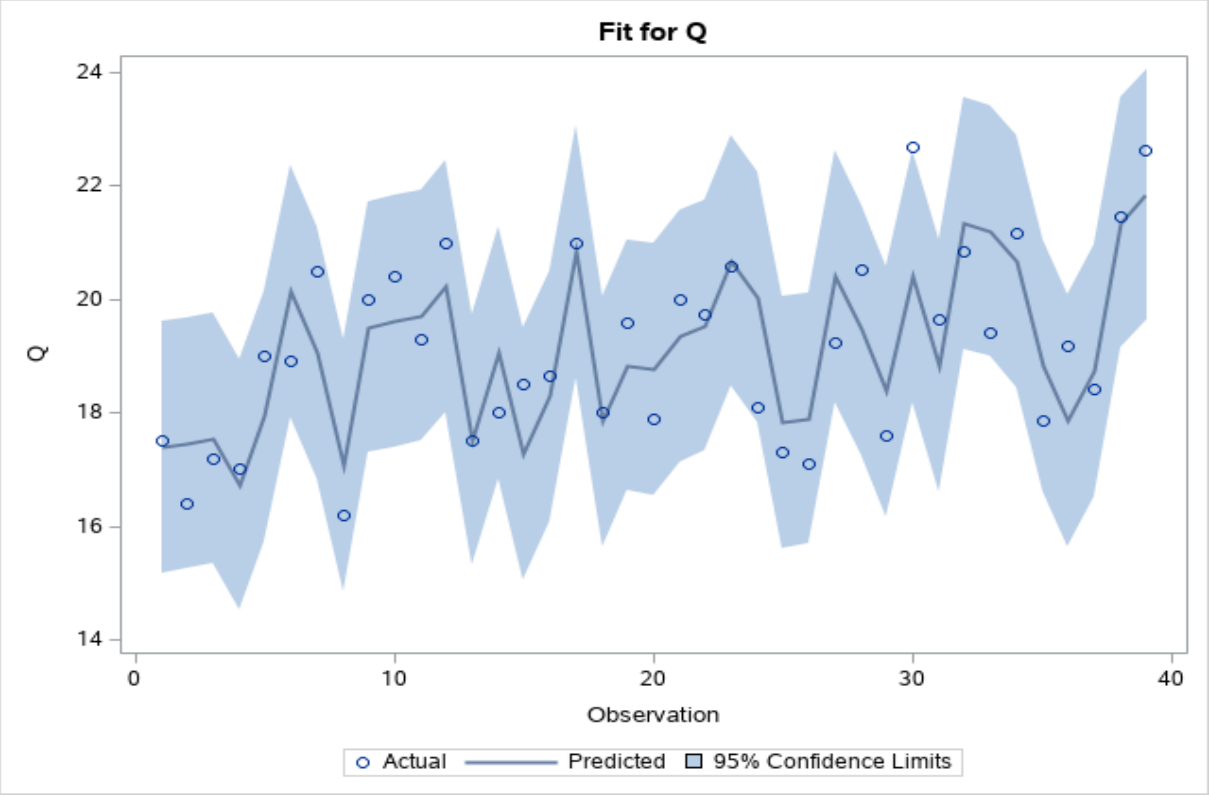


Figure 9: Fit for Q

Table 18: Simultaneous Simulation Resluts

			Actual		Predicted		
Variable	**N Obs**	**N**	**Mean**	**Std Dev**	**Mean**	**Std Dev**	**Label**
P	39	39	3.0954	0.2918	3.0933	0.2520	Price
Q	39	39	19.1248	1.6353	19.1183	1.3842	Quantity

Table 19: Simultaneous Simulation Statistics of Fit

Variable	**N**	**Mean Error**	**Mean % Error**	**Mean Abs Error**	**Mean Abs % Error**	**RMS Error**	**RMS % Error**	R^2
P	39	-0.00213	0.1465	0.1165	3.7820	0.1425	4.6365	0.7553
Q	39	-0.00650	0.1871	0.7813	4.0856	0.9549	4.9477	0.6501

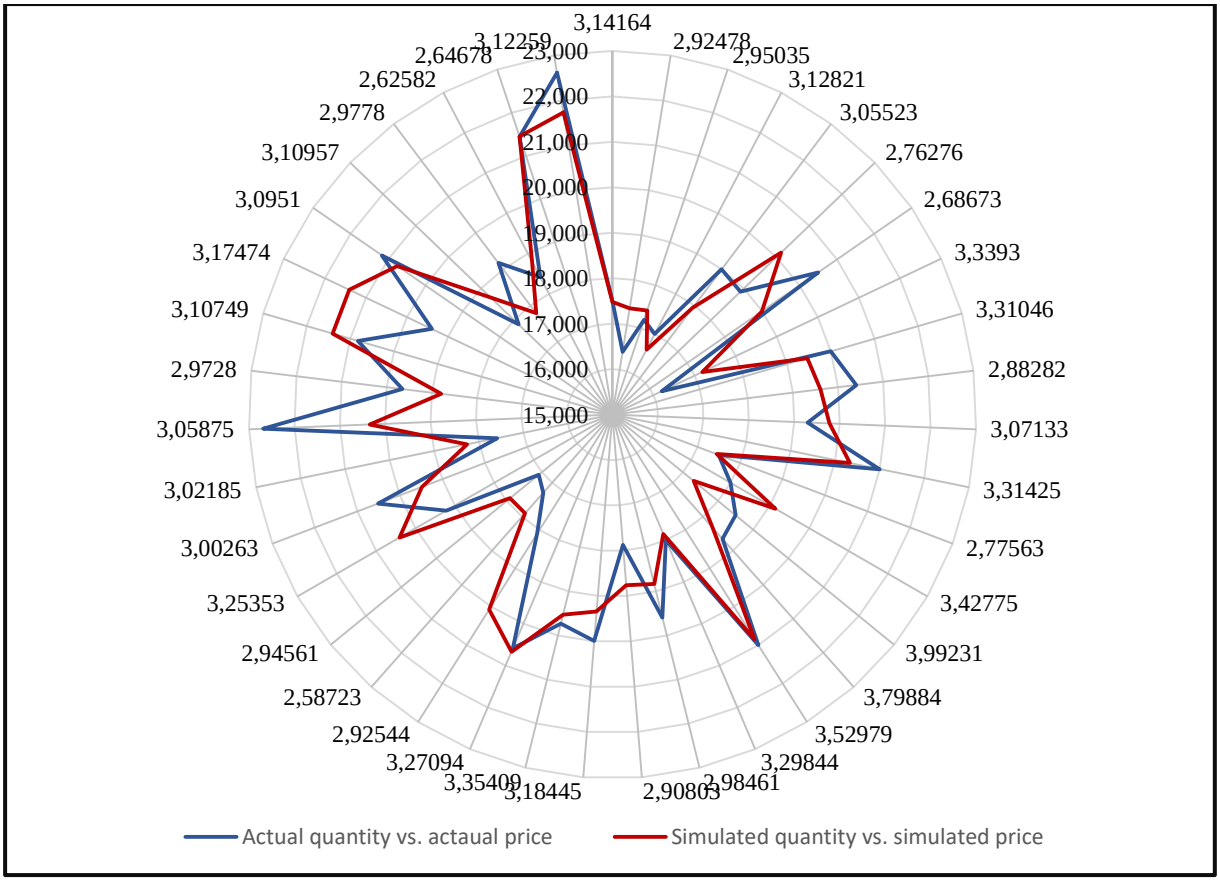


Figure 10: Satellite View of Turkish Wheat Market:Actual vs. Simulated Data (Prices are around the most outside circle, quantities (mil.tons) center to upwards, starting point is north pole 1982=3.14164), clockwise)

5.4. Producers' Expectations and Goodwin's Cobweb Model

In macro and microeconomic theories, expectations have an important place. While short-term expectations for producers cover price and sales amounts, long-term expectations include investments, production arrangements and employment according to price and quantity expectations (Keynes 1936, pp. 46–47). When the situation is looked at from the point of view of agricultural production and when there are effective public institutions in determining the price, it becomes even more important that price expectations direct production and investments. As we explained above, while the real prices remained stable, the area of cultivation in wheat in the period 1993-2020 decreased from 9.8 million hectares to 6.9 million hectares, the number of tractors increased from 746000 to 1370000. The sacrifice made from the land was closed by capital-intensive investment. Real prices, which have remained constant for forty years, have led to the abandonment of marginal land for wheat cultivation (which may have been used for other crops or purposes) and to increased investment in mechanization.

As the price expectations for annual production in agriculture gained importance, economists developed supply and demand models which include previous period price and price increases (Goodwin 1947) for commodity markets. Altough simple dynamic cobweb model sees supply as a function of the price of previous period P_{t-1}, Goodwin's theory explains the cobweb theorem with the price of the past two years and the expectations of price increase attributed and based on last year's price. Being Q_t:Quantity and P_t : Price,

Supply equation:

$$Q_t = \psi_{t-1} + \beta P_t \qquad \text{Eq (12)}$$

If we add a constant number (ρ) for the expected price of the next production period by the producers, then expected price equation may be written as:

$$Ex\ P_t = (1+\rho)P_{t-1} - \rho P_{t-2} \qquad \text{Eq (13)}$$

Substituting in eq (15):

$$Q_t = \psi_{t-1} + \beta((1+\rho)P_{t-1} - \rho P_{t-2}) \qquad \text{Eq (14)}$$

If $\rho = 0$, the simple cobweb theorem is obtained. If$\rho > 0$, it is expected that prices will increase in the same direction, and if $\rho < 0$, they will move in the opposite direction. Economically reasonable range is -1< ρ <1. This is a summary of Goodwin model.

In statistical application phase of our work, we took Goodwin's expectetions coefficient, added $(P_{t-1}^2 + (P_{t-1} - P_{t-2})^2)$, $\bar{P}$ and also complementary equation which was most suitable with wheat data for price function .

Price expectation's function:

$$Ex\ P_t = (1+\rho)P_{t-1} + \rho(P_{t-1}^2 + (P_{t-1} - P_{t-2})^2)) \qquad \text{Eq (15)}$$

Price function:

$$P_t = a_1\big((1+\rho)P_{t-1} + \rho(P_{t-1}^2 + (P_{t-1} - P_{t-2})^2)\big) + \bar{P} + [a_2 P_{t-1}]^{r_p^{(t-Z)}[\cos(a_3(t-Z)+cos(a_4*(t-Z)+a_5))]} \qquad \text{Eq (16)}$$

Quantities function:

$$Q_t = b_1 + b_2\ P^4(t-Z) + b_3^{r_q^{(t-Z)}[cos(b_4(t-Z)+\boldsymbol{b_5})+\sin(b_6 P)]} \qquad \text{Eq (17)}$$

Since the supply and demand series data are not separately available in our hand, we have to work with price and quantity series. Although Goodwin established the equations according to the differences from the averages in order to explain the price and quantity fluctuations and to find the roots of auxiliery equation in his theoretical

explanations as so-called homogenous form, on the contrary we performed statistical applications by leaving the averages in the equation for best statistical test results.

$$P_t = -1.59938\big((1-0.19108)P_{t-1}\big) - 0.19108(P_{t-1}^2 + (P_{t-1}-P_{t-2})^2) + 3.1 + 0.267075^{1.002198^{(t-1999)}[cos(1.024524(t-1999)+\cos(0.327539(t-1999)+1.826656))]} \_ 0.56031P_{l2} + 0.307509P\_m3 \quad \text{Eq (18)}$$

Average price of period 3.1 lira also added to price equation as given.

$$Q_t = 16.91016 + 0.001138P_t^4(\mathrm{t}-1999) + 0.559911^{((0.99479)^{(t-1999)}[cos(1.128678(\mathrm{t}-1999)-1.4864)+\sin(-0.70981P)]} \_ 0.86449Q_{l2} - 2.47242Q_{m2} - 1.7301Q\_m4 \quad \text{Eq (19)}$$

In the statistical application of the Goodwin model, we preferred to analyze trend and complementary functions together and solved simultaneously. We found it appropriate to use exponential forms both in price and quantities complementary functions. Also, variables of auotcorrelation correction added both for price and quantitities equations. Detailed statistical tables are given below in Table 20-24.

Goodwin developed a two-dimensional chart, $\boldsymbol{\rho}$ (expectations coefficient) on the horizontal axis and $\boldsymbol{a}$ (the ratio of slopes of supply and demand curves) on the vertical axis to determine the effects of producer expectations on price and quantity conjunctures. The economically significant values of ρ are -1< ρ<+1 Value ($\rho = -0.19108$) founded in this statistical study meet this condition and near to zero point. Most stable place is zero point on horizontal axis. ρ -values moving away from zero will increase instability.

Although Goodwin did not give an explanation for the negative value of a_1, if we take it as an absolute value, the value we find $a_1 = 1.6$ (the growth of the slope of thesupply and demand curves) will increase the likelihood of instability. The stable pitch around the vertical axis is like a conical tent that sits on the horizontal axis, and large a_1 values will push the system to outside stabel area. The fact that the expectations coefficient is very close to zero prevents this situation. As we surely expected, the wheat market is in a stable place near the zero point on the horizontal axis of Goodwin[10]. We confirm a stable market outlook for Turkish wheat market with the Goodwin model. The determination coefficients we obtained indicate a stable market. Especially in the quantities equation, the appearance that the price has no effectiveness is obtained. In this case, the independent variable time is effective and explains the market of the product whose production period is one year. We presume that the determination coefficients obtained as 0.78 and 0.81 respectively for the price and quantities equations have an adequate explanation for the wheat data set we have. Notice that we have been evaluating a market which has public control, uncontrollable prices and supply quantities by producers, import factors etc.

When the statistical tables are examined, with sufficiently higher than expected t-values of seventeen parameters of Table 21 is also an indicator of the market representation capability of study. Autocorrelation arrangements are sufficient and contributive, heteroskedasticty, normality and the multicollinearity[11] results are very satisfactory, and the probabilities in every table exceed the 5% alpha level in all tests.

Cojuncturel cycles of prices and quantities which are very close to the case of simple cobweb theorem are given below. Price expectations and quantities bought and sold are completing their cycyle in every 6.3 and 6.3 years.

M: Oscillation time:

Price: $M_p = 2\pi/a_3$=2π /1.002198=6.3 years

years.

Quantities: $M_q = 2\pi/b_4$=2π/0.99479= 6.3 years.

[10] See (Allen 1956,pp. 13-14, 196-200) for a very detailed and advanced explanation of Goodwin model.

[11] Not included in manuscript because of very vast volume which has a matrix of 19 rows and 21 columns. Maximum "condition number" is at an acceptable level of 27.

Table 20: Nonlinear FIML Summary of Residual Errors

Equation	DF Model	DF Error	SSE	MSE	Root MSE	R^2 and R	Adj R^2	Durbin Watson	Label
P	9	30	0.7241	0.0241	0.1554	0.7762 0.8810	0.7165	1.7257	P
Q	10	29	18.8375	0.6496	0.8060	0.8146 0.9025	0.7571	2.0756	Q

Table 21: Nonlinear FIML Parameter Estimates

Parameter	Estimate	Approx Std	t Value	Approx	Label
P					
a_1	-1.59938	0.0920	-17.38	<.0001	Coefficient of price equation
$\boldsymbol{\rho}$	-0.19108	0.00282	-67.87	<.0001	Coefficient of expectations
a_2	0.267075	0.00750	35.60	<.0001	Amplitude of cosine
$\boldsymbol{r_p}$	1.002198	0.0118	85.18	<.0001	Growth factor
a_3	1.024524	0.0119	86.19	<.0001	Angular frequency of
a_4	0.327539	0.0164	19.93	<.0001	Angular frequency
a_5	1.826656	0.2283	8.00	<.0001	Phase lag
P_l2	-0.56031	0.1220	-4.59	<.0001	AR(P) P lag2 parameter
P_l3	0.307509	0.1247	2.47	0.0196	AR(P) P lag3 parameter
Q					
$\boldsymbol{b_1}$	16.91016	0.3940	42.92	<.0001	First constant of quantities
$\boldsymbol{b_2}$	0.001138	0.000256	4.45	0.0001	Second constant of quantities
$\boldsymbol{b_3}$	0.559911	0.0581	9.63	<.0001	Amplitude of cosine
$\boldsymbol{r_q}$	0.99479	0.00792	125.65	<.0001	Growth factor
$\boldsymbol{b_4}$	1.128678	0.0128	88.00	<.0001	Angular frequency of cos
$\boldsymbol{b_5}$	-1.4864	0.1664	-8.93	<.0001	Phase lag
$\boldsymbol{b_6}$	-0.70981	0.2055	-3.45	0.0017	Angular frequency of sinus
Q_l2	-0.86449	0.1060	-8.16	<.0001	AR(Q) Q lag2 parameter
Q_m2	-2.47242	0.1140	-21.69	<.0001	MA(Q) Q lag2 parameter
Q_m4	-1.7301	0.0511	-33.84	<.0001	MA(Q) Q lag4 parameter

Note of Table 21: P_l2, P_l3, Q_l2, Q_m2 and Q_m4 are coefficients of correction for serial autocorrelation. See Table 5c note.

Table 22: Heteroscedasticity Test

Equation	Test	Statistic	DF	Pr > ChiSq	Variables
P	White's Test	39.00	38	0.4246	Cross of all vars
	Breusch-Pagan	7.20	5	0.2061	s1=(t-2015)(t-1999),Q, P, P_{t-1}, P_{t-2}, 1
Q	White's Test	39.00	38	0.4246	Cross of all vars
	Breusch-Pagan	1.65	5	0.8956	S1=(t-2015)(t-1999),Q, P, P_{t-1}, P_{t-2}, 1

Table 23: Goodwin Model: Autoregression Test Results of Prices and Quantities

Godfrey's Serial Correlation Test				Generalized Durbin-Watson Statistics				
Equation	**Alternative**	**LM**	**Pr > LM**	**Equation**	**Order**	**DW**	**Pr < DW**	**Pr > DW**
P	1	1.03	0.3098	**P**	1	1.73	0.0858	0.9142
	2	3.84	0.1469		2	2.31	0.9172	0.0828
	3	3.97	0.2644		3	1.88	0.6108	0.3892
Q	1	0.39	0.5329	**Q**	1	2.08	0.4726	0.5274
	2	0.48	0.7861		2	1.96	0.8759	0.1241
	3	2.40	0.4940		3	1.66	0.2769	0.7231

Table 24: Normality Tests

Equation	Test Statistic	Value	Prob
P	Shapiro-Wilk	0.96	0.2296
Q	Shapiro-Wilk	0.98	0.8366
System	Mardia Skewness	4.22	0.3771
	Mardia Kurtosis	-1.01	0.3138
	Henze-Zirkler	0.52	0.3441

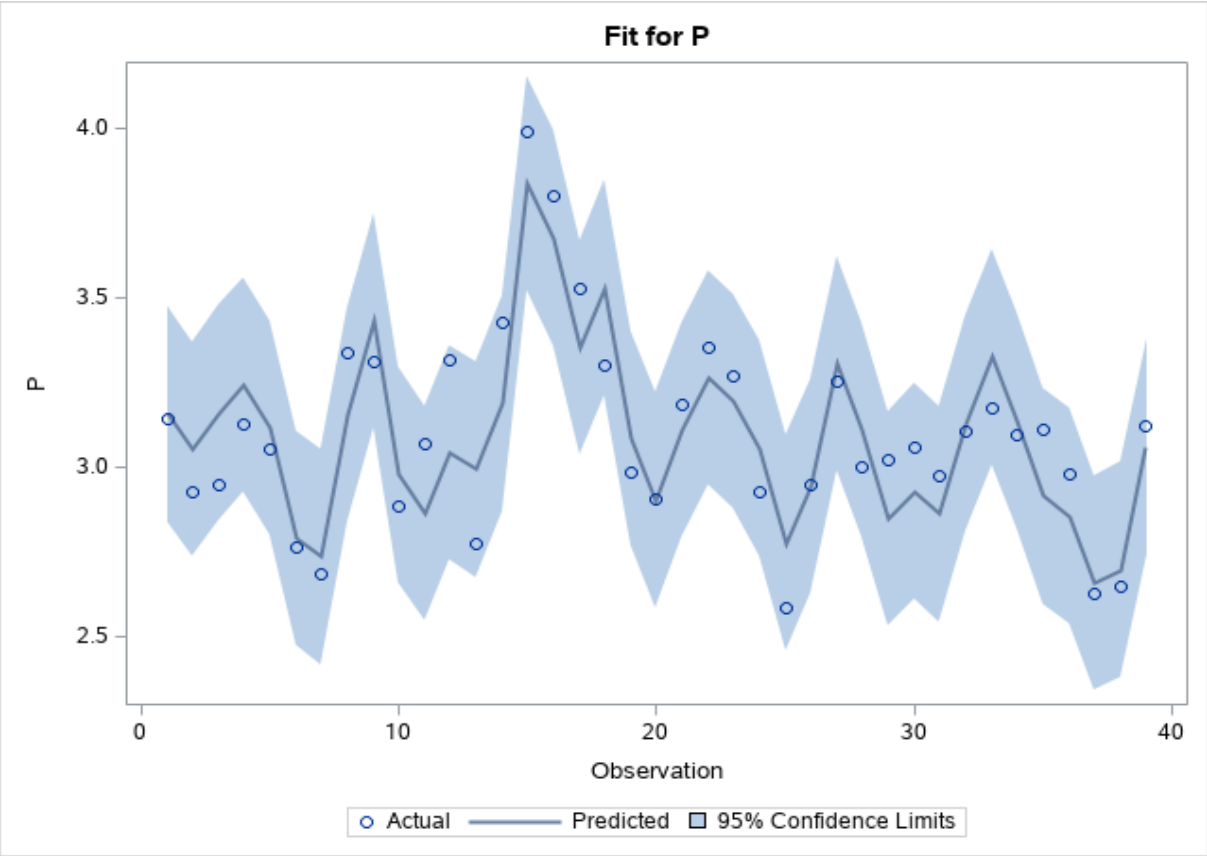


Figure 11: Wheat Prices

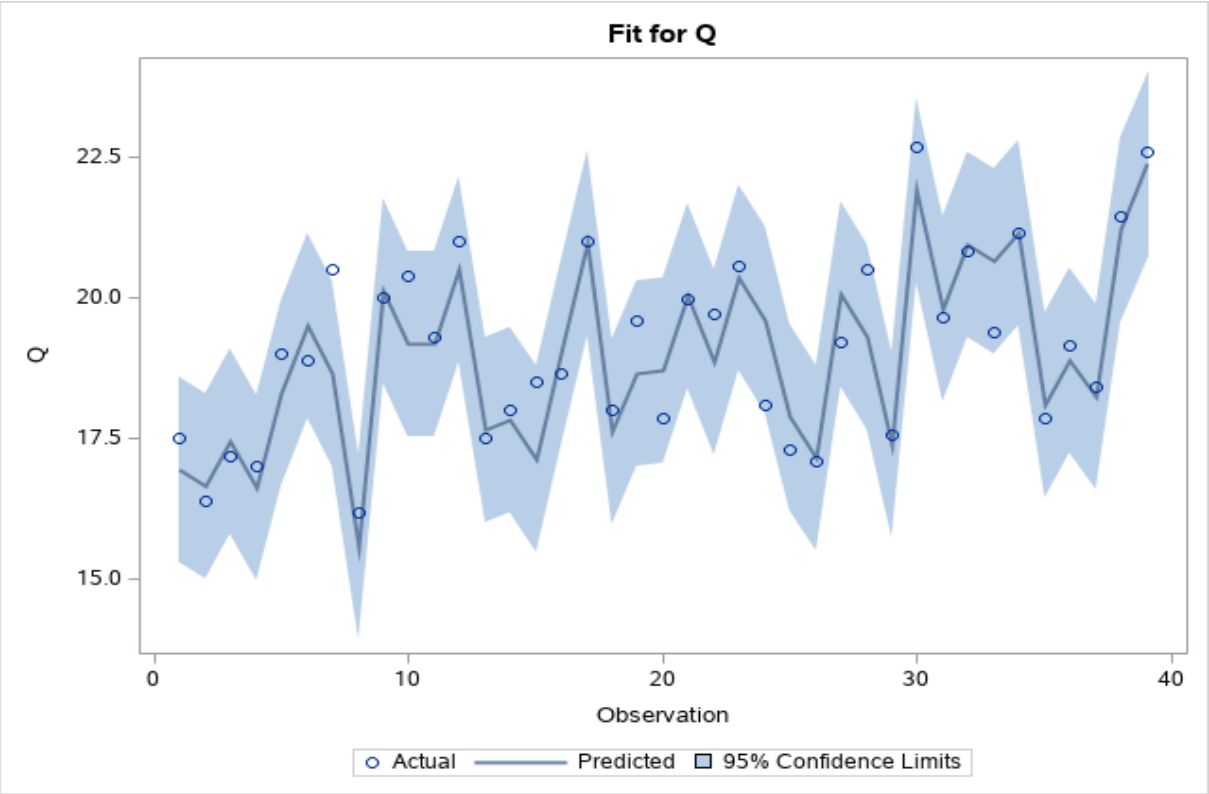


Figure 12: Fit for Quantities

Just as we find the estimated values of the dependent variable after the parameters are ascertained in case of a single function, we can also find the estimated values of the independent and dependent variables through these functions in cases where we use multiple functions. It is possible to solve equations simultaneously in SAS MODEL Procedure which takes lots of time to do manually. In the two-variable (price and quantities equations) model in our example, the statistics of simultanouous solution and graph of the estimated results are given Tables 25-26 and Figures 11-12. According to the Goodwin model, the center of the cobweb is the price of 3.1 real lira and the amount of 19.1 mil. tons of wheat. The entire cobweb is located around these coordinates. Results are almost same as in simple model. State of affairs confirms that producers' expectations theory is in conformity with cobweb model.

Below given statistical tables of price and quantity estimates based on equations and a graph of actual figures and forecast figures. We leave it to the readers to interpret the representation capability of tables and graphs in accordance with their purposes.

Table 25: Descriptive Statistics

			Actual		Predicted		
Variable	**N Obs**	**N**	**Mean**	**Std Dev**	**Mean**	**Std Dev**	**Label**
P	39	39	3.0954	0.2918	3.0892	0.2558	P
Q	39	39	19.1248	1.6353	18.9703	1.6365	Q

Table 26: Statistics of Fit

Variable	N	Mean Error	Mean % Error	Mean Abs Error	Mean Abs % Error	RMS Error	RMS % Error	R^2	Label
P	39	-0.00619	-0.0102	0.1166	3.7406	0.1363	4.3612	0.7762	P
Q	39	-0.1545	-0.7497	0.5325	2.7767	0.7250	3.7523	0.7983	Q

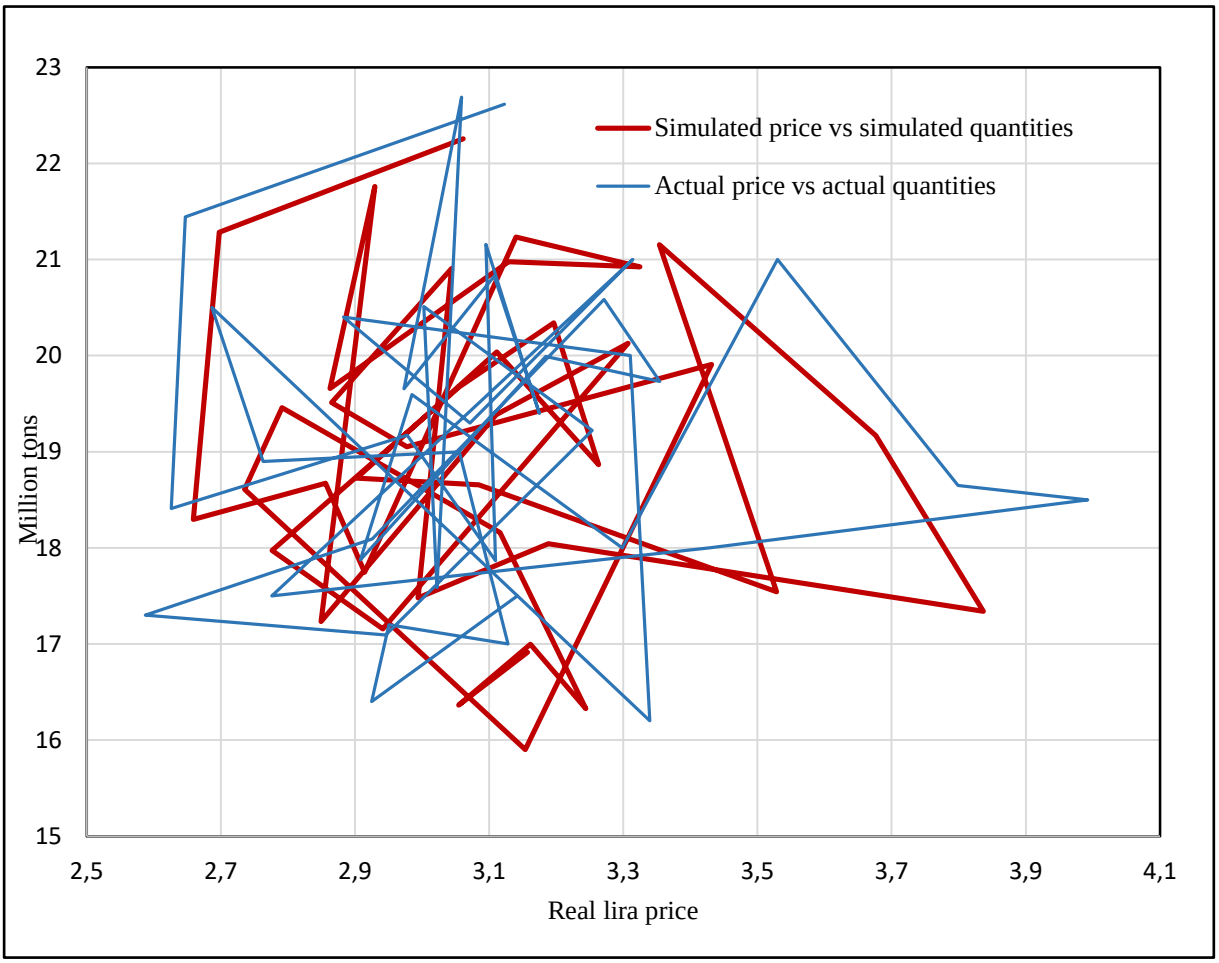


Figure 13: Actual and Simulated Figures of Goodwin Cobweb Functions

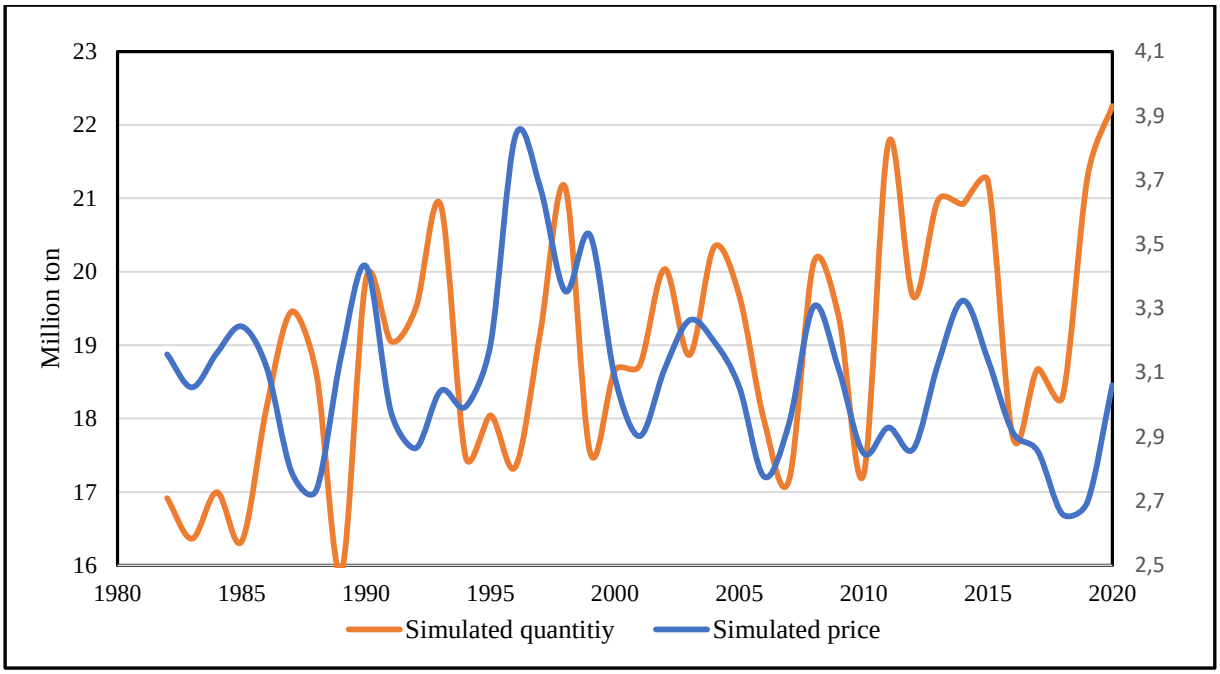


Figure 14: Simulated Quantity and Price Trends of Goodwin Model (Left axis: Quantities as million tons, right axis: Prices as real lira)

According to the Figure 14 simulation, production is on the rising trend, while prices are in stable and/or downward trend.

6. Comments and Results

Our wheat production and market study which we have chosen as a subject matter of referable and respectable econometric analysis of a product, has achieved its purpose in terms of the results obtained from economic theory, modeling and application. Successful and detailed results have benn only possible with advanced econometrics, statistical softwares. Although the developing and increasing number of theoretical econometrics and statistical studies increase the work volume of the research, the expectations are met more and the degree of health increases as a result of the expansion of the results in terms of the scope obtained.

Data we collected and the methods we applied confirm the validity of simple and advanced cobweb theorems. It is clear that prices and quantities do not form rectangular cobwebs as described in ecomomics textbooks. Supply and demand curves appear to form angular and complex cobwebs according to their slopes. In our study, supply and demand curves were not determined, but trend (equilibrium path) that the intersection points of supply and demand curves followed over the years ascertained. The axis of the price path and central point are 3.1 real lira. For quantity, the time axis and the center point of the cobweb is 19.1 million tons of wheat.

According to the production function, the wheat production capacity is 20.5 million tons. Changes below and above the trend value of 19.1 million tons occur as a result of weather conditions. With current input, output and world prices of wheat, trend value of production is expected to remain at 19.1 million tons and the price at the level of 3.1 real lira. Despite the fact that the amount of cultivated land decreased from 10 million hectares to 6.9 million hectares, it has been possible with mechanization (capital-intensive production) that production remained at the level of 19.1 million tons. No change is expected in terms of both supply and demand namely in production and consumption levels, which are under the supervision of the relevant public institutions.

Our simulation study reveals that the price trend is downward, while the quantity trend has yet set a new direction upwards for forty years of period. Interim periods might be quite different than that of long period.

Our study does not carry comprehensive objectives for the sectoral problems and solutions of wheat production, market and distribution. On the contrary, econometric study will contribute to the economy of the sector to the extent that it reveals valuable points related to the sectorial policies in accordance with its purpose and framework.

References

Allen, R. G. D. (1956). *Mathematical Economics* (2nd ed. 1959, reprinted 1960). Macmillan & Co Ltd.

Gebremariam, B. (2014). *Is Nonlinear Regression Throwing You a Curve? New Diagnostic and Inference Tools in the NLIN Procedure,* (Paper SAS384-2014). https://support.sas.com/resources/papers/proceedings14/SAS384-2014.pdf

Goodwin, R. M. (1947). Dynamical Coupling with Especial Reference to Markets Having Production Lags. *Econometrica, 15*(3), 181–204.

Güran, T. (1997), *Osmanlı Dönemi Tarım İstatistikleri 1909, 1913 ve 1914*, (Ş. Pamuk, ed.), Ankara: Devlet İstatistik Enstitüsü.

İskender, C. (2018), "Türkiye Nüfus Büyümesi ve Tahminleri: Matematiksel Büyüme Modelleri ve İstatistiksel Analiz ile Kuramsal ve Uygulamalı bir Yaklaşım," *Istanbul University Econometrics and Statistics e-Journal*, 14, 75–141. https://doi.org/10.26650/ekoist.2018.14.28.0004.

İskender, C. (2021a), "Mathematical Study of the Verhulst and Gompertz Growth Functions and Their Contemporary Applications," *Ekoist : Journal of Econometrics and Statistics*.

İskender, C. (2021b), "Econometric Analysis of Population Increase and Population Projections in Turkey," *Journal of Statistical Research*, 11, 30–55.

Keynes, J. M. (1936). *The General Theory of Employment Interest and Money* (1973 ed.). The Royal Economic Society.

Polat, K. (2020), *Durum ve Tahmin Buğday 2020*, Ankara: T. C. Tarım ve Orman Bakanlığı TEPGE.

SAS Institute Inc. (2014), *SAS/ETS® 13.2 User's Guide The AUTOREG Procedure,* North Carolina: SAS Institute Inc., Cary, NC, USA.

SAS Institute Inc. (2018a), *SAS/ETS® 15.1 User's Guide The MODEL Procedure*, Cary, NC, USA: SAS Institute Inc.

SAS Institute Inc. (2018b), *SAS/STAT 15.1 User ' s Guide Introduction to Statistical Modeling with SAS / STAT Software*, Cary, NC, USA: SAS Institute Inc., Cary, NC, USA.

SAS Institute Inc. (2020), *SAS/STAT® 15.2 User's Guide The NLIN Procedure,* Cary, NC, USA: SAS Insitute Inc.

Tarım ve Orman Bakanlığı (2020, 2021), *Ürün Masaları Buğday Bülteni,* Ankara.

İstatistik Araştırma Dergisi
Journal of Statistical Research

ISSN: 2791-7614

(2023), 13 (1)
Başvuru/Received: 27/10/2022
Kabul/Accepted: 17 /01/2023

https://journal.tuik.gov.tr/

ARAŞTIRMA MAKALESİ **RESEARCH ARTICLE**

Sağdan Sansürlü Veriler için Veri Madenciliği Algoritmaları Performanslarının Karşılaştırılması

Saygın DİLER
Van Yüzüncü Yıl Üniversitesi / Doktora Öğrencisi
saygin.diler@tuik.gov.tr
Orcid No: 0000-0002-9056-412X

Yıldırım DEMİR
Van Yüzüncü Yıl Üniversitesi / Dr. Öğr. Üyesi
ydemir@yyu.edu.tr
Orcid No: 0000-0002-6350-8122

Özet

Veri madenciliği algoritmaları ile gerçekleştirilen modelleme çalışmaları bilgisayar teknolojisinin gelişmesiyle birlikte artış göstermiştir. Ancak bu algoritmalar ile yapılan çalışmalarda veri kalitesinin bozulması elde edilecek sınıflandırma performanslarında önemli rol oynamaktadır. Bu çalışmada veri madenciliği sınıflandırma algoritmalarının performanslarının veri kalitesini bozan etmenlerden biri olan sansürlü verinin veri setinde yer alması durumunda nasıl etkilendiği incelenmiştir. Sansürlü verilerinin etkisini veri setinde gösterilebilmesi amacı ile K en yakın komşu algoritması (kNN) imputasyon yöntemi kullanılmıştır. Daha sonra sınıflandırma algoritmalarından olan Naive Bayes (NB), Lojistik Regresyon (LR) ve K en yakın komşu algoritması (kNN) ile uygulamalar gerçekleştirilmiştir. Yöntemlerin performanslarının incelenmesi için simülasyon çalışması ve gerçek veri seti çalışmaları yapılmış, sonuçlar sunulmuştur. Analiz sonuçlarına göre, yüksek sansür seviyesinde ve düşük sansür seviyesinde Lojistik Regresyon algoritmasının sansür ile baş etmede dikkate değer performans gösterdiği belirlenmiştir. Ayrıca örneklem büyüklüğü arttıkça genel olarak algoritmaların doğru sınıflama performanslarının arttığı gözlenmiştir. Özetle büyük örneklemeli veri setlerinde Lojistik Regresyon algoritmasının doğru sınıflandırma oranı ile başarılı sınıflandırma performansı gösterdiği söylenebilir.

Anahtar Sözcükler: Sağdan Sansürlü Veri, Sınıflandırma, Veri Madenciliği

Sorumlu Yazar / Corresponding Author: 1-Saygın DİLER, Van Yüzüncü Yıl Üniversitesi, Fen Bilimleri Enstitüsü İstatistik

2-Yıldırım DEMİR, Van Yüzüncü Yıl Üniversitesi, İktisadi ve İdari Bilimler Fakültesi, Ekonometri Bölümü

Atıf / Citation: DİLER S., DEMİR Y. (2023). Sağdan Sansürlü Veriler için Veri Madenciliği Algoritmaları Performanslarının Karşılaştırılması. İstatistik Araştırma Dergisi, 13 (1), 34-47.

Bu makale Saygın DİLER tarafından yazılmakta olan "Veri Kalitesinin Bozulduğu Durumlarda Veri Madenciliği Sınıflandırma Algoritmalarının Performanslarının Karşılaştırılması" isimli doktora tezinden üretilmiştir.

Atıf / Citation: DİLER S., DEMİR Y. (2023). Sağdan Sansürlü Veriler için Veri Madenciliği Algoritmaları Performanslarının Karşılaştırılması. İstatistik Araştırma Dergisi, 13 (1), 34-47.

Comparison of Data Mining Algorithms Performances for Right-Censored Data

Abstract

Modeling studies performed with data mining algorithms have increased with the development of computer technology. However, the deterioration of data quality in studies with these algorithms plays an important role in the classification performances to be obtained. In this study, it has been examined how the performance of data mining classification algorithms is affected when censored data, which is one of the factors that deteriorates data quality, is included in the data set. In order to show the effect of the censored data in the data set, the K nearest neighbor algorithm (KNN) imputation method was used. Then, applications were carried out with Naive Bayes (NB), Logistic Regression (LR) and K nearest neighbor algorithm (KNN), which are among the classification algorithms. To inspect the performance of the mentioned methods, simulation study and real data example are carried out. According to the results of the analysis, it was determined that Logistic Regression algorithm at high and low censorship level showed remarkable performance in dealing with censorship. In addition, it was observed that the correct classification performance of the algorithms increased as the sample size increased. In summary, it can be said that the correct classification success of Logistic Regression algorithm in data sets with large samples show successful classification performance with values.

Keywords: Classification, Data Mining, Right-Censored Data

1. Giriş

Günümüzde teknolojik gelişmeler çok hızlı ilerleme göstermektedir. Bu hızlı gelişmelerle birlikte teknolojinin aktif olarak kullanılması sonucunda veriler üretilmektedir. Üretilen veri çeşitliliği, miktarı ve hacmi her geçen gün daha da artmaktadır. Bu artış ile veri tabanlarında analiz edilmeyi bekleyen sayısızca veri bulunmaktadır. Bu verilerin analiz edilmesi için çeşitli yöntemler geliştirilmiş ve geliştirilmeye de devam edilmektedir. Yaygın olarak kullanılan bu yöntemlerden birisi de veri madenciliği algoritmaları yaklaşımıdır. Veri madenciliği algoritmaları sağlık, güvenlik, finans, lojistik, ekonomi gibi birçok farklı bilim alanında sıklıkla kullanılmaktadır (Han ve ark., 2012). Ancak veri madenciliği için geliştirilen algoritmaların sınıflandırma performanslarını etkileyen önemli etkenlerden biri de veri setinin kalitesidir. Zira algoritmalar veri setini kullandığı ve çıkarımlar veri setinden yapıldığı için algoritma performansları veri kalitesinden etkilenmektedir (Batista & Monard, 2002). Veri kalitesini bozan durumlara; eksik veri, sansürlü veri, aykırı değer, çoklu doğrusal bağlantı sorunu örnek gösterilebilir.

Başarılı bir şekilde veri madenciliği uygulamasının yapılabilmesi, büyük ölçüde veri kalitesi ile doğru orantılıdır. Veri bütünlüğü ve verilerin analize uygun olması veri madenciliği çalışmalarının ön koşulları arasında yer almaktadır. Birçok bilim alanında toplanan verilerde, veri yapısı ve kalitesini etkileyen sorunlarla karşılaşmak mümkündür. Söz konusu sorunların varlığında veri madenciliği uygulanmalarından veya istatistik analizlerden güvenilir tahminler yapmak zorlaşabilir. Bu nedenle, veri kalitesinin bozuk olduğu durumlarda klasik veri madenciliği yöntemleri hantal hale gelmektedir. Özellikle yüksek boyutlu ve çeşitli süreç verilerinde varsayım ve hipotezlerin derinlemesine incelenmesi araştırmacıları zorlamaktadır (Zhu ve ark., 2018). Veri madenciliği çalışmalarında, veri yapısı ve kalitesini etkileyen önemli faktörlerden birisi de sansürlü verilerdir. Sansürlü veriler üzerine yapılan çalışmalarda, genellikle sağdan sansürlü veriler kullanılmaktadır (Eröz & Tutkun, 2020). Bu çalışmada da sağdan sansürlü verilere odaklanılmıştır.

Veri madenciliğinde sınıflandırma amacı ile kullanılan algoritmalar olay bilgilerinin bütün veriler için bilindiği varsayımı ile çalışmaktalar. Sansürlü veri içeren veri setlerinde bilgileri takip edilemeyen denekler için belirsiz bir durum bulunmaktadır (Vock ve ark., 2016). Sınıflandırma algoritmaları gerçek veri setleri üzerinde başarılı performanslar göstermesine rağmen, literatürde sansürlenmiş verilere bu algoritmaların uyguladığı çok az çalışma bulunmaktadır (Goldberg ve Kosorok, 2012).

Literatürde, sansürlü verilerde makine öğrenmesi yöntemlerini kullanan çalışmalardan bazıları şu şekildedir. Shivaswamy ve ark. (2007) ile Khan ve Zubek (2008), sansürlü veriyi hesaba katmak için kayıp fonksiyonunu değiştirerek destek vektör makinelerini uyarlamayı önermişlerdir. Her iki çalışmada da sansürlü veri setlerine uyarlanmış destek vektör makinelerinin geleneksel yöntemlere göre daha başarılı sonuçlar verdiği gözlenmiştir. Ishwaran ve ark. (2008), sağdan sansürlenmiş hayatta kalma verilerinin analizi için rasgele orman algoritmasının bir versiyonu olan rastgele hayatta kalma orman algoritmasını tanıtmışlar. Hayatta kalma dağılımını sınıflandırma ağaçları ile göstermişler. Štajduhar ve ark. (2009), sansürlü verilerde bayes ağları ile oluşturulan modellerin yaşam sürelerinin hesaplanmasındaki etkisini incelemişler. Hafif, orta ve ağır sansür altında sentetik veriler üzerinde bir simülasyon çalışması kullanarak sansürlemenin etkisini ve sansürün bayes ağlarının öğrenme olasılığını nasıl etkilediğini belirlemeye çalışmışlar. Bandyopadhyay ve ark. (2015), sansürlü elektronik sağlık verileriyle kardiyovasküler riskleri

tahmin etmek amacıyla bayes ağı modeli önermişlerdir. Sağdan sansürlü verilerde bayes ağı modelinin Cox orantılı risk analizine göre daha başarılı tahmin etme performansı gösterdiğini belirtmişlerdir. Vock ve ark. (2016), sağdan sansürlü elektronik sağlık verilerine; sansürleme ağırlığının ters olasılığını hesaba katarak makine öğrenmesi/veri madenciliği algoritmalarını uygulamışlar. İleri sürdükleri yaklaşımın daha başarılı sonuçlar verdiğini göstermişlerdir.

Sağdan sansürlü veriler olması durumunda, literatürde yer alan çalışmalarda izlenen yollardan birisi sağdan sansürlü veriler hariç tutularak analizler gerçekleştirilmektedir. Bir diğer yol ise veri ön işleme adımlarında sansürlü verilerin tahminleme yöntemi ile tamamlanmasıdır. Bu çalışmalar genel olarak değerlendirildiğinde şu sonuçlar elde edilmektedir; sansürlü veriler ile ilgili yapılan çalışmalar tek bir sınıflandırma algoritmasına özgüdür ve genel olarak uygulanabilirler ancak analizler sonucunda elde edilen tahminler hata payı içermektedir (Vock ve ark., 2016).

Bu çalışmada, sağdan-sansürlü verilerde veri madenciliği sınıflandırma algoritmalarının sınıflandırma performansları üzerinde durulmuştur. Sansürlenmiş gözlemleri tahmin etmeye yarayan yerine koyma yöntemlerinden biri olan KNN imputasyon yöntemi ile sansürlü veriler tahmin edilerek analize uygun hale getirilmiş ve daha sonra uygulamalar, sansürlü veri setleri üzerinde gerçekleştirilmiş ve çeşitli sansür seviyelerine göre algoritma başarıları karşılaştırılmıştır.

Çalışmanın temel amacı, sağdan sansürlü verilerde makine öğrenmesi algoritmalarının performanslarını ve bu tür verilerde mevcut kullanılan algoritmalardan hangisinin daha iyi olduğunu belirleyerek literatüre katkı sağlamaktır.

2. Materyal ve Yöntem

2.1 Sınıflandırma Algoritmaları

Sınıflandırma yöntemlerinde amaç bir bağımlı ve birden fazla bağımsız değişkenden oluşan veri setlerinde bağımsız değişkenlerden bağımlı değişkene eşleyen bir model oluşturmaktır. Bağımlı değişkenler kategorik verilerden oluşmaktadır. Veri madenciliği alanında sınıflandırma amacı için geliştirilen algoritmalar istatistik, yapay zeka, makine öğrenmesi gibi bilim alanlarından faydalanmaktadır. Her algoritma için geçerli varsayımlar ve sınıflandırma görevleri, kullanılan yönteme göre birbirinden farklılık göstermektedir (Davidson ve Tayi, 2009).

Sınıflandırma yöntemlerinde genellikle kullanılan başlıca teknikler arasında; K-En Yakın Komşu, Naïve Bayes, Lojistik Regresyon, Destek Vektör Makineleri, Karar Ağaçları, Yapay Sinir Ağları ve Genetik Algoritmalar yer almaktadır (Silahtaroğlu, 2013). Bu çalışmada literatürde fazlaca uygulama yapılan algoritmalar arasında yer alan KNN, Naïve Bayes ve Lojistik Regresyon algoritmaları ile uygulama gerçekleştirilmiştir ve performansları karşılaştırılmıştır.

2.1.1 K-En Yakın Komşu Algoritması

K-en yakın komşu (KNN) algoritması veri madenciliğinde en çok kullanılan algoritmalar arasında yer almaktadır. Parametrik olmayan yöntemler arasında yer alan algoritma (Bishop, 2006), sınıfları belli olan veri setinden sınıfı bilinmeyen yeni bir veriyi en yakın komşusuna atama mantığına dayanmaktadır (Mucherino ve ark., 2009).

KNN algoritmasının doğru sınıflandırma başarısı büyük ölçüde k değerinin (komşu sayısı) optimum seçimine bağlıdır. k değerini seçmenin bir çok yolu bulunmakta ve en basit seçim, farklı k değerleri kullanılarak en iyi performans sergileyen k değerini belirleme yöntemidir (Guo ve ark., 2003). Ayrıca bu değer, örnek sayısı ve öznitelikler göz önünde bulundurularak da seçilebilir. Örnek sayısı n ise k değeri; $k = \sqrt{n}$ şeklinde seçilebilir (Balaban ve Kartal, 2015). k en yakın komşu algoritmasının doğru sınıflama başarısında k değerinin seçimi kadar en yakın komşuya olan uzaklık (benzerlik) ölçüsü de büyük önem arz etmektedir. Öznitelikler arasındaki mesafeyi ölçmenin farklı yöntemleri bulunmakta ve bu çalışmada uzaklık ölçüsü olarak Öklid kullanılmıştır. Öklid ölçüsü;

$$d(x,y) = \sqrt{\sum_{j=1}^{N}\left(x_j - y_j\right)^2} \qquad (1)$$

olarak gösterilebilir (Bramer, 2007).

K-en yakın komşu algoritması ile veriler basit ve etkili bir şekilde sınıflandırılabilir. Ancak algoritma örnek tabanlı olduğundan uygulama yapabilmek için eğitim verilerinin mevcut olması gerekmekte ve bu nedenle büyük veri kümeleri için büyük miktarda depolama alanına ihtiyaç duyulmaktadır. Verilerin yapısı hakkında bilgi vermemesi (model oluşturmaz) KNN'nin bir diğer dezavantajıdır (Harrington, 2012).

2.1.2 Naive Bayes Algoritması

Naive bayes algoritması gözlemelere dayalı olasılıkları ve olasılık dağılımındaki parametreleri hesaplayan bir yöntemdir (McNamara ve ark., 2006). İstatistik tabanlı algoritmalar arasında yer alan Bayesci sınıflama tekniği (Bramer, 2007), veri seti içerisinde sınıfı bilinen verileri kullanarak yeni bir verinin mevcut sınıflardan herhangi birine girme olasılığını belirlemektedir (Silahtaroğlu, 2013).

Naive Bayes, makine öğrenmesi ve veri madenciliği için en verimli ve etkili tümevarımsal öğrenme algoritmalarından biridir. Bu algoritma basit olmasına rağmen, çeşitli öğrenme problemlerinde iyi performans sergilemektedir (Frank ve ark., 2003). Naive Bayes sınıflandırıcı, test örneklerinin sınıfını doğru tahmin eden ve sınıf bilgisi taşıyan eğitim örnekleri gibi danışmanlı tümevarım işlemlerinde kullanılmak üzere tasarlanmıştır (Balaban & Kartal, 2015).

Naive Bayes sınıflandırıcı ile başarılı bir sınıflandırma modeli oluşturmak için eğitim veri setinin büyük olmasına gerek yoktur. Konu ile ilgili olmayan niteliklere karşı güçlü olup bu sınıflandırıcı için örneklem boyutu arttıkça ilgisiz veriler önemsiz hale gelmekte ve gerçek zamanlı çevrimiçi sistemlere kolayca entegre edilebilmektedir (Lewis, 2017).

Algoritma, sınıfları belirlemek için koşullu olasılıkları kullanmaktadır. Burada $X = \{x_1, x_2, \dots, x_n\}$ nitelik değerlerinden oluşan ve sınıf üyeliği bilinmeyen veri örneğini, $C_1, C_1, \dots, C_m$ m sınıfın sınıf değerlerini göstermektedir. Sınıfı belirleyecek olan örneğe ilişkin olasılıklar,

$$P(C_i|X) = \frac{P(X|C_i)P(C_i)}{P(X)} \tag{2}$$

olarak hesaplanır. $P(X|C_i)$ olasılığı basitleştirilerek hesaplamadaki işlem yükü azaltılabilir. Bunun için, x_k değerlerinin birbirinden bağımsız olduğu kabul edilmekte ve Eşitlik (3) kullanılmaktadır.

$$P(X|C_i) = \prod_{k=1}^{n} P(X_k|C_i) \tag{3}$$

Bilinmeyen X örneğini sınıflandırmak için eşitlik (2) deki en büyük değer belirlenmekte ve bilinmeyen örneğin bu sınıfa ait olduğuna karar verilmektedir.

$$arg\max_{Ci}\{P(X|C_i)P(C_i)\} \tag{4}$$

Sonlu olasılıkları kullanan Eşitlik (4), en büyük sonlu sınıflandırma yöntemi (Maximum A Posteriori classification=MAP) olarak da bilinir ve Bayes sınıflayıcısı Eşitlik (5)'i kullanmaktadır (Özkan, 2008):

$$C_{MAP} = arg\max_{Ci} \prod_{k=1}^{n} P(X_k|C_i) \tag{5}$$

2.1.3 Lojistik Regresyon Analizi

Lojistik regresyon modeli doğrusal regresyon modelinin özel bir hali olup bu modelde bağımlı değişken iki sınıflı kategorik bir değişkendir (Lewis, 2017). Lineer regresyonda aranan varsayımların aranmaması, değişken tipi ve dağılımı ile ilgili varsayımların az olması nedeniyle lojistik regresyon araştırmacıların ilgisini çekmektedir (Balaban & Kartal, 2015). Lojistik regresyon analizinde amaç en az değişken ile regresyon katsayılarını optimize ederek model oluşturmaktır (Akpınar, 2014). Veri madenciliği çalışmalarında ise sınıflandırma amacı ile kullanılmaktadır. Lojistik regresyon analizinde sigmoid fonksiyon olarak adlandırılan lojistik fonksiyon aşağıdaki gibi yazılabilir (Harrington, 2012):

$$P(Y=1)=\frac{e^{\beta_0+\beta_1X_1}}{1+e^{\beta_0+\beta_1X_1}}=\frac{1}{1+e^{-(\beta_0+\beta_1X_1)}} \quad (6)$$

Burada β'lar model parametrelerini, x'ler açıklayıcı değişkenleri, Y ise genellikle 0 ve 1 değerlerinden oluşan kategorik bağımlı değişkenini ifade etmektedir. Logaritmik dönüşüm uygulanarak bu ilişki doğrusal bir şekilde incelenebilir. Bu dönüşüm lojit dönüşüm olduğundan yöntem lojistik regresyon olarak adlandırılır (Gamgam & Altunkaynak, 2017). Bağımsız değişken sayısı p olduğu zaman regresyon modeli:

$$P(Y=1|\,x)=\frac{e^{\beta_0+\beta_1X_1+\cdots+\beta_pX_p}}{1+e^{\beta_0+\beta_1X_1+\cdots+\beta_pX_p}}=\frac{1}{1+e^{-(\beta_0+\beta_1X_1+\cdots+\beta_pX_p)}} \quad (7)$$

olur. Burada $P(Y|\,x)$, bağımsız değişkenlerin değeri bilindiğinde bağımlı değişkenin olasılığını göstermektedir. Bu model Eşitlik (8)'deki gibi düzenlenebilir.

$$ln\left(\frac{P(Y=1|\,x)}{1-P(Y=1|\,x)}\right)=\beta_0+\beta_1X_1+\cdots+\beta_pX_p \quad (8)$$

Bir olayın gerçekleşme olasılığının gerçekleşmeme olasılığına oranı olan $P(Y=1|\,x)$/(1- $P(Y=1|\,x)$) ifadesi odds oranını göstermektedir (Hosmer ve ark., 2013).

Bağımlı değişkenin iki kategoriden oluşması ikili (Binary), kategorilerin ikiden fazla ve sıralı olması sıralı (Ordinal), ikiden fazla ve sıralı olmaması durumunda ise nominal lojistik regresyon analizi geçerlidir (Özdamar, 2019).

2.2. Sağdan-sansürlü veri

Sansürlü veriler; sağdan sansürlü (right censored), soldan sansürlü (left censored) ve aralıklı sansürlü (interval censored) veriler olmak üzeri üç ana başlık altında incelenmektedir (Bardakçı & Kartal, 2018). Bu çalışmada sağdan sansürlü veriler olması durumunda sınıflandırma algoritmalarının performansları karşılaştırıldığı için sağdan sansürlü veri tanıtıldıktan sonra, sansürün çözümlenmesi için kullanılan K en yakın komşu algoritması (kNN) imputasyon yöntemi belirtilmiştir.

Bir araştırma için klinik bir deney yapıldığı ve çalışmaya katılan n adet hastanın yaşam sürelerinin takip edildiği düşünüldüğünde, gözlem süresince maalesef çalışmadaki tüm (n) deneklerin hayatta kalma süreleri gözlemlenemez (sansürlenir). Bu tür durumlar aşağıdaki sebeplerden dolayı yaşanmaktadır;

- **Çalışmanın bitmesi**: Bazı bireyler için çalışma süresince ilgilenilen olay henüz gerçekleşmemiştir (ölüm, hastalık belirtisi vb.)
- **Bırakma**: Bireyin artık çalışmaya katılmak istememesi, bırakıp gitmesi.
- **Takipte kayıp**: Çalışmadaki kişi artık çalışmada görünmez, kişinin izi kaybedilmiştir.

Böyle durumlarda çalışmada bulunan bazı bireyler için kısmi bilgi elde edilir. "c" çalışmanın gerçek bitiş süresini gösterdiğinde bazı bireyler için çalışma sonuna kadar ilgilenilen durum gerçekleşmediğinde en kötü ilgili bireyin hayatta kalma süresi c_i olacaktır. Böylece i'inci birey için kısmi bilgiye sahip olunur ve y_i hayatta kalma gözleminin c_i tarafından sansürlendiği söylenebilir. Dolayısıyla her bir özne veya nesne için c_i ve y_i değerleri arasından en küçük olanının seçilmesiyle yaşam sürelerini sansür durumlarıyla birlikte sansür bilgisini içeren yeni bir yanıt değişkeni elde edilir. Gözlemler;

$$Z_i=\min(y_i,c_i)\ ve\ \delta_i=I(y_i\le c_i) \quad (9)$$

şeklinde olur. Böylece sansürü taşıyan gözlem çiftleri elde edilmekte ve burada δ_i'ler sansür bilgisini taşıyan gösterge fonksiyonunu vermektedir. Sansür varsa "0" yoksa "1" değerini göstermektedir (Gijbels, 2010; Yılmaz & Aydın, 2019).

Sağdan sansürlü verilerin tahmin edilmesi ile ilgili çalışmalarda önemli bir varsayım olan bağımsızlık varsayımı: y_i hayatta kalma süresinin tüm c_i sansürleme süresinden bağımsız olduğudur (Gijbels, 2010). Bu sansür mekanizması, sağ rastgele sansür olarak anılmaktadır. İkinci varsayım ise $P(y_i\le c_i|y_i,x_i)=P(y_i\le c_i|y_i)$'dir. Yani olay (ölüm, belirti vb.) zamanı için açıklayıcı değişkenler veri sansürlü olsa da olmasa da elde edilenden daha fazla bilgi sağlamamaktadır (Yılmaz & Aydın, 2019).

2.2.1. k-NN yerine koyma yöntemi ile sansürün çözümlenmesi

kNN yerine koyma yöntemi, sağdan sansürlü veri noktalarını tahmin etmek için kullanılabilir. Bu yöntem dağılımdan tamamen bağımsız bir yöntem olup hiçbir varsayım kullanmadığı için parametrik olmayan bir yöntemdir. Kesikli ve sürekli veri yapısındaki değişkenler için kullanılabilir. Yerine koymak için tahmin edilen değerler gerçek verilerden elde edildiği için açıklayıcı değişkenler hakkında daha fazla bilgi içermektedir. Bunlar yöntemin önemli avantajları arasında yer almaktadır. Bu, veri noktaları arasındaki mesafelere bağlı olarak benzerliğe dayalı çalışan bir yöntemdir. Genellikle bu mesafe, Eşitlik (1)'de verilmiş olan Öklid uzaklıkları ile ölçülmektedir. k-NN yerine koyma algoritmasının uygulama adımları Tablo 1'de verilmiştir (Ahmed ve ark., 2020);

Tablo 1. k-NN için algoritma adımları

Algoritma: $Sağdan-sansürlü\ veriler\ için\ kNN\ yerin\ koyma\ yöntemi$
Girdi: Sağdan sansürlü veriseti z_i
Sansür işaretçisi δ_i
En yakın komşu sayısı k
Açıklayıcı değişken değerleri x_i
Çıktı: Sansürlü verilerin tahminlerini içeren yeni değişken $\mathbf{y}^{knn} = \left(y_1^{knn}, \dots, y_n^{knn}\right)^T$
1 **başla**
2 **for** $(i = 1$ to $n)$
3 **Eğer** $(\delta_i = 0)$ **yap** (eğer veri noktası sansürlü ise)
4 **for**$(j = 1$ to $n)$
5Her bir sansürlü gözlem için x_j ve x_i için eşitlik $(1)'$te verilen öklit uzaklıkları bulunur
6 Mesafeler küçükten büyüğe sıralanır
7 **for** $(j = 1$ to $k)$
8 Sıralanan mesafelerle ilişkili ilk k adet sansürlü olmayan gözlem alınır.
9 i′nci sansürlü veri tahmini $\left(y_i^{knn}\right)$ en yakın k adet y_i değerinin ortalaması alınarak hesaplanır.
10 Tahmin edilen gözlemler $\left(y_i^{knn}\right)$ sansürlü gözlemler $(z_i, \delta_i = 0)$ile yer değiştirilir.
11 $\mathbf{y}^{knn} = \left(y_1^{knn}, \dots, y_n^{knn}\right)^T$ oluşturulur.
12 **Son**

Böylece veri setinde sansürün etkisi modele dahil edilerek yaşam sürelerini temsil eden veriler yerine $\mathbf{y}^{knn}$ kullanılabilir.

2.3. Sınıflandırma Performanslarının Değerlendirilmesi

Sınıflandırma algoritmaları ile kurulan modellerin başarısını değerlendirmek için Doğruluk, Duyarlılık, Seçicilik, Kesinlik ve F-Ölçütü gibi metrikler bulunmaktadır. Hata matrisi tablosunda yer alan gerçek ve tahmin değerlerinden elde edilen bazı metriklerle sınıflandırma model performansları değerlendirilmektedir (Balaban ve Kartal, 2015).

Tablo 2. Hata matrisi (Confusion matrix)

		Tahmin Edilen Sınıf	
		Pozitif	**Negatif**
Gerçek Sınıf	**Pozitif**	TP (True Positive)	FN (False Negative)
	Negatif	FP (False Positive)	TN (True Negative)

Doğruluk (Accuracy): Modelin ortalama performansını ve sınıflandırma başarısını gösteren en basit ve en sık kullanılan ölçüttür. Doğru sınıflandırılmış değerlerin toplam örnek sayısına bölünmesi ile hesaplanmaktadır.

$$Doğruluk = \frac{TP + TN}{TP + TN + FP + FN} \tag{10}$$

Duyarlılık (Sensitivity): Doğru sınıflandırılmış pozitif değerlerin toplam gerçek pozitif örnek sayısına bölünmesi ile hesaplanmaktadır.

$$Duyarlılık = \frac{TP}{TP + FN} \tag{11}$$

Seçicilik (Spesificity): Doğru sınıflandırılmış negatif örneklerin toplam negatif tahmin edilen örneklere bölünmesi ile hesaplanmaktadır.

$$Seçicilik = \frac{TN}{TN + FP} \tag{12}$$

Kesinlik (Precision): Doğru sınıflandırılmış pozitif örneklerin toplam pozitif tahmin edilen örneklere bölünmesi ile hesaplanmaktadır

$$Kesinlik = \frac{TP}{TP + FP} \tag{13}$$

F-Ölçütü (F-Measure): Kesinlik ve Duyarlılık metriklerinin harmonik ortalaması alınarak her iki ölçüyü beraber değerlendirme imkânı vermektedir (Mulla ve ark., 2021).

$$F\ Ölçütü = 2\ x\ \frac{Duyarlılık\ x\ Kesinlik}{Duyarlılık +\ Kesinlik} \tag{14}$$

3. Uygulama

Sağdan sansürlü veriler bulunması durumunda sınıflandırma algoritma performanslarının karşılaştırılması amacı ile iki gerçek veri seti ve daha sonra ise simülasyon çalışması ile uygulama gerçekleştirilmiştir. Uygulamalarda R programlama dili kullanılmıştır.

%75'i eğitim ve %25'i test olacak şekilde veri setleri iki gruba ayrılmıştır. Modeller eğitim veri seti ile oluşturulmuş daha sonra model performansları test veri setiyle ölçülmüştür. Model performanslarını karşılaştırmak üzere doğruluk, duyarlılık, seçicilik, kesinlik ve F-ölçütü değerleri kullanılmıştır.

3.1. Rektum Kanseri veri seti

Rektum veri setinde (Aydin ve Yilmaz, 2018) yer alan değişkenler Tablo 3'de gösterilmiştir. Veri seti 97 gözlemden oluşmakta ve bu gözlemlerin 32'si sağdan sansürlüdür. Böylece veri seti %33 orta düzey sansür içermektedir. Sınıflandırılan veri setinde bir bağımlı ve dört bağımsız değişken bulunmaktadır.

Tablo 3. Rektum veri setinde yer alan değişkenler.

Öznitelik Adı	Veri Türü	**Nümerik**=Min, Max **Kategorik**= Değerler	**Nümerik**=Ort.±s.s **Kategorik**= n (%)
Yaş	Nümerik	18- 85	56.6±14.3
Preop.Alb<3gr:1 >3gr:2	Nümerik	1, 2	1: n=15 (%15) 2: n=82 (%85)
Toplam kan tx	Nümerik	0- 17	3.4±3.5
Operasyon süresi (dk.)	Nümerik	60- 330	205.8±69.0
Sansür Durumu	Kategorik	0=sansürlü, 1=sansürsüz	0: n=32 (%33) 1: n=65 (%67)
Yaşam Süresi*	Nümerik	1- 94	25.9±23.3
İmputed Data (Yaşam Süresi)	Nümerik	1- 98.2	31.0±27.2
Tümörün evresi**	Kategorik	1, 2, 3, 4	1: n=3 (%3) 3: n=20 (%21) 2: n=28 (%29) 4: n=46 (%47)
Bağımlı Değişken (Tümörün evresi)	Kategorik	1: Başlangıç safhası 2: İleri safha	1: n=23 (%24) 2: n=74 (%76)

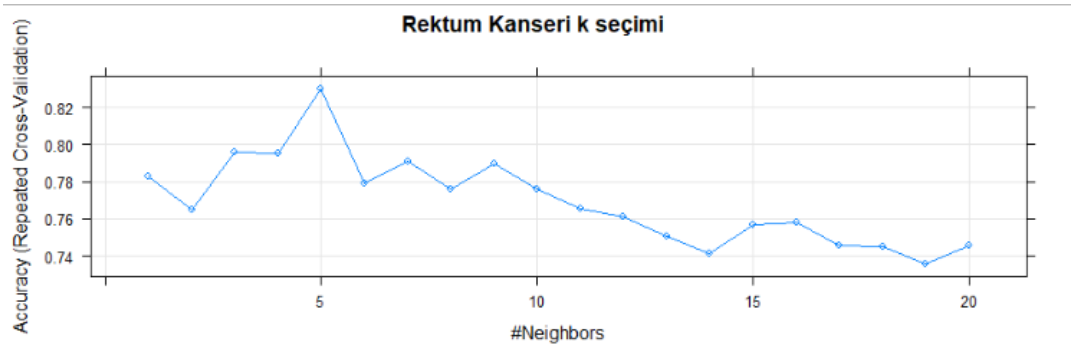
*yaşam süresi değişkeni sağdan sansürlü verileri temsil etmekte ve modelde bu değişken yerine İmputed data kullanılmıştır.
**tümörün evresi değişkeni yeniden sınıflandırılmış ve veri setinde bağımlı değişken olarak kullanılmıştır.

Rektum veri seti için beş sınıflandırma performans değerlendirme kriteri dikkate alınarak sınıflandırma algoritmalarının sınıflandırma performans sonuçları Tablo 4'de verilmiştir.

Tablo 4. Rektum veri setinde için sınıflandırma performansları.

Ölçütler / Algoritma	Doğruluk	Duyarlılık	Seçicilik	Kesinlik	F-Ölçütü
K-NN	0.78	0.60	0.83	0.50	0.54
Naive Bayes	0.70	**0.60**	0.72	0.38	0.46
Lojistik Regresyon	**0.83**	**0.60**	**0.89**	**0.60**	**0.60**

Sağdan sansürlü rektum veri setinde oluşturulan modellerden en iyi performans Lojistik Regresyon algoritması tarafından %83 doğrulukla elde edilmiştir. En düşük performans ise %70 doğrulukla Naïve Bayes algoritmasıyla elde edilmiştir. Duyarlılık, Seçicilik, Kesinlik ve F-Ölçütü metriklerine göre incelendiğinde en yüksek değerler Lojistik Regresyon algoritması ile elde edildiği gözlenmiştir.



Şekil 1. Rektum veri seti için k değerinin seçimi (k=5)

3.2. Hepatosellüler veri seti

Hepatosellüler veri seti 26 Mayıs 2022 tarihinde (https://rdrr.io/cran/asaur/man/hepatoCellular.html) adresinden alınmış ve veri setinde yer alan değişkenler Tablo 5'de verilmiştir. Veri seti 227 gözlemden oluşmakta ve bu gözlemlerin 84'ü sağdan sansürlüdür. Böylece veri seti %37 orta düzey sansür içermektedir. Sınıflandırma yapılan veri seti, 1 bağımlı ve

11 bağımsız değişkenden oluşmaktadır.

Tablo 5. Hepatosellüler veri setinde yer alan değişkenler.

Öznitelik Adı	**Veri Türü**	**Nümerik**=Min, Max **Kategorik**= Değerler	**Nümerik**=Ort.±s.s **Kategorik**= n(%)
Yaş	Nümerik	13- 79	49.8 - 12.4
RFS	Nümerik	1- 81	26.3 - 23.2
CXCL17T	Nümerik	0- 1184.4	99.3 - 181.4
CXCL17P	Nümerik	2,0- 1016.0	100.5 - 130
CXCL17N	Nümerik	0- 1171.1	100 - 181.4
Tümör boyutu	Kategorik	1,2	1: n=96 (%42.3) 2: n=131 (%57.7)
Cinsiyet	Kategorik	0,1	0: n=30 (%13.2) 1: n=197 (%86.8)
HBsAg	Kategorik	0,1	0: n=18 (%7.9) 1: n=209 (%92.1)
Sansür Durumu	Kategorik	0=sansürlü, 1=sansürsüz	0: n=84 (%37.0) 1: n=143 (%63.0)
Tumormultiplicity	Kategorik	0,1	0: n=170 (%74.9) 1: n=57 (%25.1)
OS*	Nümerik	2- 83	36.5- 22.2
İmputed Data (OS)	Nümerik	2- 94.9	42- 26.6
Bağımlı Değişken (Vascularinvasion)	Kategorik	0, 1	0: n=186 (%81.9) 1: n=41 (%18.1)

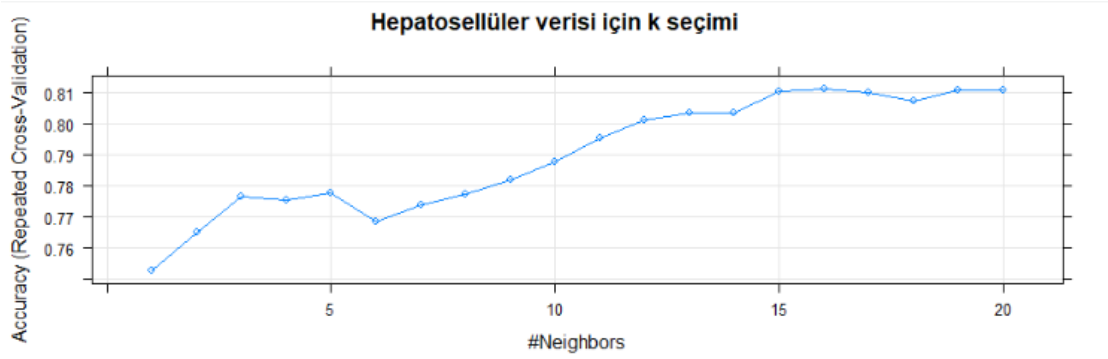
*OS değişkeni sağdan sansürlü verileri temsil etmektedir ve modelde bu değişken yerine impüstasyon ile elde edilen veri kullanılmıştır.

Beş sınıflandırma performans değerlendirme kriteri dikkate alınarak hepatosellüler veri seti için sınıflandırma algoritmalarının sınıflandırma performans sonuçları Tablo 6'da verilmiştir.

Tablo 6. Hepatosellüler veri setinde için sınıflandırma performansları.

Ölçütler / **Algoritma**	**Doğruluk**	**Duyarlılık**	**Seçicilik**	**Kesinlik**	**F-Ölçütü**
K-NN	**0.82**	1.00	0.00	0.82	**0.90**
Naive Bayes	0.77	**0.87**	**0.30**	0.85	0.86
Lojistik Regresyon	0.79	0.83	0.25	**0.93**	0.88

Sağdan sansürlü hepatosellüler veri setinde oluşturulan modellerden en iyi performans %82 doğrulukla K-NN algoritmasıyla ve en düşük performans ise %77 doğrulukla Naive Bayes algoritmasıyla elde edilmiştir. Duyarlılık ve Seçicilik için en yüksek değerler Naïve Bayes algoritmasıyla, kesinlik için ise en yüksek değer %93 ile Lojistik Regresyon algoritması ile elde edilmiştir.



Şekil 2. Hepatosellüler veri seti için k değerinin seçimi (k=16)

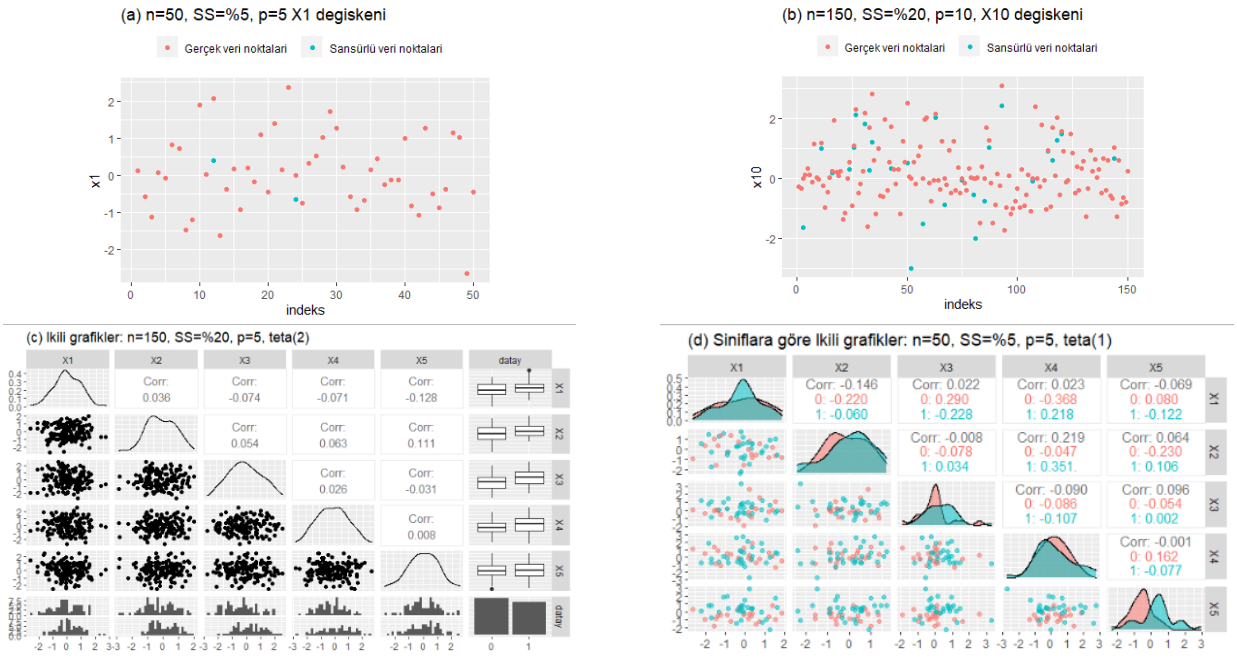
3.3. Simülasyon Çalışması

Simülasyon çalışması için veri üretimi ve simülasyon kurgusu Tablo 7'de verilmiştir. Her bir simülasyon 1000 tekrarlı olup sonuçlar tablolar halinde sunulmuştur. Simülasyon çalışmalarında bağımlı değişkenin içerdiği sınıf sayısı her zaman iki olarak alınmıştır. Simülasyon kurgusu göz önüne alındığında, sınıflandırma algoritmalarının performanslarına etkileri bakımından hem sansür seviyesinin hem örneklem büyüklüğünün hem de bağımsız değişken sayısının performansa nasıl etki yaptıkları gözlemlenebilmektedir. Eğitim ve test veri setleri gerçek veride olduğu gibi ayarlanmıştır. Ayrıca, elde edilen sonuçlar, gerçek veri çalışmaları ile de karşılaştırılmış ve uyum ya da uyumsuzluklar tartışılmıştır.

Tablo 7. Simülasyon kurgusu ve veri üretimi

Simülasyon kurgusu		
Örneklem büyüklüğü	Sansür seviyesi	Açıklayıcı değişken sayısı
$n = 50, 150$	$SS = (\%5, \%20)$	$p = (5, 10)$
Verilerin üretilmesi		
Bağımsız değişkenler	Olasılıklar üretilmesi (sınıflar için)	Kategorik bağımlı değişken
$X \sim U[0,1] \in \mathbb{R}^{n \times p}$	$z = 1 + \theta_p X,$ $P_y = \frac{1}{(1 + e^{-z})}$	$y = binom(n, P_y)$

Burada θ_p, $(p \times 1)$ bir vektörü temsil eder ve açıklayıcı değişkenin katsayılarını ifade eder. Veri üretiminde iki farklı katsayı vektörü ele alındı. Bunlar; $\theta_p^{(1)} = (0.5, 0.5, \ldots, 0.5)$ ve $\theta_p^{(2)} = (3, 3, \ldots, 3)$. Böylece, açıklayıcı değişkenlerin, sınıf değişkeni üzerindeki farklı seviyedeki etkilerinin incelenmesi planlanmıştır.



Şekil 3. Simülasyonda belirli konfigürasyonlar için üretilen verilere ait tanımlayıcı grafikler

Şekil 1, üretilen verileri, sansürlülük durumlarını ve değişkenlerin birbirleriyle ilişkilerini gözlemleyebilmek amacıyla elde edilmiştir. Dört panelden oluşmakta ve her bir panel, farklı simülasyon konfigürasyonlarını temsil etmektedir. Panel (a), düşük sansür ve panel (b) yüksek sansür seviyeleri için sırasıyla küçük ve büyük örneklem büyüklüklerine göre verilerin dağılımını göstermektedir. Burada sansürlenen değişkenler, açıklayıcı değişkenlerdir. Her bir değişken aynı dağılımdan üretildiğinden, temsilen bir tanesi için saçılım grafiği elde edilmiştir. Panel (c), verilerin tamamı için ikili ilişkilerin yoğunluklarını ve korelasyon değerlerini gösterirken, Panel (d) bağımlı değişkene göre sınıfları içeren ikili grafikleri göstermektedir. Grafiklere göre bağımsız değişkenler arasında güçlü bir ilişkinin olmadığı görülmüş ki bu, lojistik regresyon modelinin doğru çalışabilmesi için gerekli bir varsayımdır. Burada belirtmek gerekir ki

korelasyon değerleri olarak Spearman korelasyon katsayıları hesaplanmıştır. Aksi halde, tanımsızlıkların elde edilmesi mümkündür. Diğer makine öğrenmesi yöntemlerinin sıkı varsayımlar gerektirmemesi nedeniyle veri üretiminde başka bir varsayım ele alınmamıştır. Sonuçlardan önce belirtmelidir ki, kNN için "k" gerçek veri çalışmalarında olduğu gibi doğruluk kriterini maksimum yapacak şekilde çapraz geçerlilik (cross-validation) ile optimize edilerek seçilmiştir.

Sınıflandırma algoritmalarının genel performanslarına ait bütün muhtemel simülasyon senaryoları Tablo 8 ve Tablo 9'da verilmiş ve Tablo 8, bağımsız değişken sayısı $p = 5$ olduğu durum içindir. En iyi performans gösteren yönteme ait değerler tablolarda kalın şeklide vurgulanmıştır.

Tablo 8. $p = 5$ olduğunda örnek büyüklüğü ve sansür seviyesine göre sonuçlar

		%5						%20					
n	Algoritma	K-NN		Naive Bayes		Lojistik Reg.		K-NN		Naive Bayes		Lojistik Reg.	
		$\theta_p^{(1)}$	$\theta_p^{(2)}$	$\theta_p^{(1)}$	$\theta_p^{(2)}$	$\theta_p^{(1)}$	$\theta_p^{(2)}$	$\theta_p^{(1)}$	$\theta_p^{(2)}$	$\theta_p^{(1)}$	$\theta_p^{(2)}$	$\theta_p^{(1)}$	$\theta_p^{(2)}$
50	**Duyarlılık**	0.652	0.900	**0.840**	0.843	**0.840**	**0.942**	0.608	0.850	**0.804**	0.810	0.768	**0.864**
	Seçicilik	0.917	**0.886**	0.787	**0.886**	**0.920**	0.880	0.847	0.840	0.747	0.780	**0.876**	**0.850**
	Kesinlik	0.786	0.876	0.722	0.881	**0.840**	**0.843**	0.700	**0.888**	0.751	0.764	**0.864**	0.810
	F-Ölçütü	0.687	**0.880**	0.762	0.858	**0.832**	0.878	0.604	**0.842**	0.774	0.779	**0.809**	0.816
	Doğruluk	0.784	0.893	0.813	0.865	**0.880**	**0.911**	0.777	0.845	0.725	0.795	**0.822**	**0.857**
150	**Duyarlılık**	0.711	0.792	**0.937**	0.855	0.809	**0.931**	0.605	0.779	**0.871**	0.769	0.793	**0.857**
	Seçicilik	0.777	**0.920**	0.878	0.907	**0.928**	0.898	0.810	0.899	**0.909**	**0.907**	0.839	0.829
	Kesinlik	0.733	**0.906**	0.792	0.901	**0.885**	0.870	0.744	0.838	**0.844**	**0.898**	0.789	0.892
	F-Ölçütü	0.712	0.842	0.754	0.871	**0.833**	**0.898**	0.754	0.804	**0.854**	0.824	0.797	**0.872**
	Doğruluk	0.744	0.856	0.808	0.892	**0.861**	**0.915**	0.790	0.839	**0.890**	0.838	0.824	**0.893**

Tablo 8 incelendiğinde, genel çerçevede lojistik regresyon yönteminin diğer üç yönteme göre üstün olduğu görülmektedir. Ayrıca, ölçüt değerlerindeki düşüş ile sansürün performans üzerinde negatif etkiye sahip olduğu gözlemlenebilir. Fakat bu genel bir doğru olarak sunulmamalıdır. Zira hem sansürlülük çözümü için kullanılan k-NN imputasyon yönteminin, hem sınıflandırma algoritmalarının sıkı varsayımlara dayanamaması nedeniyle sansür seviyesinin kesinlikle negatif etki oluşturacağı söylenemez. Bu durum hem Tablo 8 hem de Tablo 9'da pek çok kez gözlenmiştir. Aynı şekilde, örneklem büyüklüğü arttığında performansların artması beklenir. Fakat aynı sebepten bu beklenti gerçekleşmeyebilir. Bu bağlamda, yöntemlerin performansı incelendiğinde, örneklem büyüklüğü artıkça performansın arttığı gözlenmiştir. Diğer yandan, sansür seviyesi arttıkça hem Tablo 8 hem de Tablo 9'da Naive-Bayes algoritmasının ağır sansür altında ve n=150 olduğunda lojistik regresyon ile birlikte performansının kNN modeline göre daha iyi sınıflandırma performansı gösterdiği açıktır. Ek olarak kNN yönteminin her iki örneklem büyüklüğü için de seçicilik ve kesinlik kriterleri özelinde fakat veri üretiminde $\theta_p^{(2)}$ kullanıldığında iyi sonuçlar gösterdiği saptanmıştır. Burada belirtmek gerekir ki katsayı vektörü θ_p tahmin performanslarını önemli ölçüde etkilemiştir. $\theta_p^{(1)}$ için elde edilen performanslar $\theta_p^{(2)}$'ye göre çok daha düşük elde edilmiştir ki bu beklenen bir sonuç olmasına rağmen, çalışmanın önemli sonuçlarından sayılabilir. Çünkü açıklayıcı değişkenlerin sınıf değişkeni üzerindeki etkisi arttıkça sınıflandırma performansının iyileşeceği öngörülebilir ve bu öngörü verilen Tablo 8 ve Tablo 9'daki değerlerle ispatlanmıştır.

Tablo 9. $p = 10$ olduğunda örnek büyüklüğü ve sansür seviyesine göre sonuçlar

		%5						%20					
n	Algoritma	K-NN		Naive Bayes		Lojistik Reg.		K-NN		Naive Bayes		Lojistik Reg.	
		$\theta_p^{(1)}$	$\theta_p^{(2)}$	$\theta_p^{(1)}$	$\theta_p^{(2)}$	$\theta_p^{(1)}$	$\theta_p^{(2)}$	$\theta_p^{(1)}$	$\theta_p^{(2)}$	$\theta_p^{(1)}$	$\theta_p^{(2)}$	$\theta_p^{(1)}$	$\theta_p^{(2)}$
50	**Duyarlılık**	0.600	**0.853**	**0.800**	0.767	0.753	0.785	0.590	0.826	0.640	**0.900**	**0.653**	0.800
	Seçicilik	0.826	0.876	**0.892**	0.590	0.740	**0.893**	**0.735**	0.666	0.726	0.790	0.718	**0.966**
	Kesinlik	0.733	0.820	**0.820**	0.680	0.580	**0.933**	0.614	0.761	**0.708**	0.736	0.616	**0.950**
	F-Ölçütü	0.545	0.820	**0.803**	0.719	0.613	**0.847**	0.522	0.778	0.647	0.801	**0.658**	**0.863**
	Doğruluk	0.658	**0.864**	**0.846**	0.678	0.747	0.839	0.667	0.746	**0.683**	0.845	0.676	**0.883**
150	**Duyarlılık**	0.734	0.788	**0.936**	0.844	0.840	**0.920**	0.774	0.753	0.766	**0.963**	**0.822**	0.910
	Seçicilik	0.735	0.831	0.833	0.834	**0.955**	**0.889**	0.841	0.841	0.774	0.832	**0.842**	**0.929**
	Kesinlik	0.733	0.814	0.803	0.824	**0.941**	**0.872**	0.782	0.796	0.746	0.809	**0.825**	**0.895**
	F-Ölçütü	0.721	0.792	0.864	0.829	**0.885**	**0.889**	0.775	0.767	0.753	0.876	**0.823**	**0.899**
	Doğruluk	0.729	0.810	0.885	0.839	**0.897**	**0.905**	0.807	0.797	0.770	0.898	**0.832**	**0.919**

Tablo 9, bağımsız değişken sayısı $p = 10$ olduğunda muhtemel tüm simülasyon senaryolarını içermektedir. Tablo incelendiğinde, $n = 50$ ve CL=5% olduğunda Tablo 8'deki sonuçlardan farklı olarak Naive-Bayes yönteminin $\theta_p^{(1)}$ vektörü için diğer iki yöntemden daha iyi performans gösterdiği görülmüştür. Buna göre Naive-Bayes yönteminin açıklayıcı değişken sayısı arttığında ve bu değişkenlerin katkılarının küçük olduğu durumda kullanılması önerilebilir ki bu sonuç, çalışmanın katkılarından biri olarak sunulabilir. Diğer yandan, genel tablo incelendiğinde, lojistik regresyon modelinin diğer iki yönteme göre çok daha iyi sonuçlar sunduğu ve Naive-Bayes'in lojistik regresyon yakın değerler verdiği görülmüştür. kNN sınıflandırma yöntemi her ne kadar $n = 150$ olduğunda ve $\theta_p^{(2)}$ katsayıları kullanıldığında tatmin edici sonuçlar verse de, diğer iki yöntem kadar iyi bir performans gösterememiştir. Bu bağlamda, $p = 10$ özelinde, yüksek sayıda açıklayıcı değişken içeren veri setlerinde kNN yerine lojistik regresyon veya uygun konfigürasyon altında ($n = 50, CL = 5\%$) Naive-Bayes kullanılması önerilebilir. Burada ayrıca belirtmek gerekir ki $n = 150$ olduğunda performanslar dikkate değer şekilde artmıştır. Ayrıca duyarlılık ölçütü bakımından en iyi performansı lojistik regresyon ve Naive Bayes yöntemleri göstermiştir. Sansür seviyesi $SS = \%20$ olduğunda Naive-Bayes yönteminin özellikle $\theta_p^{(1)}$ için ayırt edici şekilde sansürle baş etmede lojistik regresyon ile kNN algoritmasından daha iyi sonuçlar verdiği söylenebilir.

Gerçek verilerle simülasyon çalışmalarının uyumlu sonuçlar verdiği belirlenmiştir. Hem Rektum kanseri hem de Hepatosellüler veri setinde lojistik regresyon yönteminin performans ölçütleri bağlamında iyi sonuçlar göstermesi, simülasyon sonuçlarının gerçek veri ile uyumlu olduğu şeklinde yorumlanabilir.

4. Tartışma ve Sonuç

Bu çalışmada sansürlü veri setleri üzerinde üç farklı sınıflandırma yöntemi performanslarının karşılaştırılması hedeflenmiştir. Bu bağlamda sağdan sansürlü veriler için k-NN imputasyon yöntemi ile sağdan-sansürlü veriler tamamlanmış ve sınıflandırma yöntemleri elde edilen yeni veriler kullanılarak bağımlı değişken için sınıflandırma modelleri kurulmuş ve performansları ölçülmüştür. Sonuçları gözlemlemek için Rektum kanseri veri seti ve Hepatosellüler veri seti ile yapılan uygulamalara ek olarak farklı senaryolardan yöntemlerin davranışlarını gözlemlemek amacıyla simülasyon çalışması gerçekleştirilmiş ve sonuçlar sunulmuştur. Elde edilen sonuç ve öneriler aşağıda sıralanmıştır.

- Rektum kanseri veri seti sonuçları incelendiğinde, en iyi modellerin lojistik regresyon yöntemi ile elde edildiği görülmüştür. Sansürlü veri bağlamında oldukça yüksek sayılabilecek bir sansür seviyesi için 0.83 doğruluk değeriyle Lojistik regresyon algoritması tatmin edici sonuçlar vermiştir. Böylece sansürle baş etmede bu yöntemin dikkate değer performans gösterdiği görülmüştür.
- Hepatosellüler veri seti sonuçları incelendiğinde, hemen hemen aynı sansür seviyesi için (SS=%37) her yöntem farklı performans kriteri için iyi sonuçlar vermiştir. Bu durum, $n = 227$ geniş örneklem büyüklüğü söz konusu olduğunda elde edilmiş ve $n = 150$ olduğunda simülasyon sonuçları bunu doğrular niteliktedir. Doğruluk kriterinde en iyi sonuçlar K-NN yöntemiyle sağlanmıştır.
- Simülasyon sonuçlarına göre ise, sansüre karşı en dayanıklı iki yöntemin lojistik regresyon ve Naive-Bayes olduğu görülmüştür. Ayrıca lojistik regresyon yönteminin büyük örneklemlerde iyi sonuçlar verdiği gözlemlenmiş ve bu durum Hepatosellüler veri seti sonuçlarıyla uyumluluk göstermektedir. K-NN simülasyon çalışmasında her ne kadar diğer yöntemlere yakın performanslar gösterse de dikkate değer bir farklılık ortaya koyamamıştır.

Bu bağlamda gerçek veri setleri ile simülasyon çalışmaları beraber değerlendirildiğinde doğruluk ölçütüne göre; yüksek ve düşük sansür seviyesinde Lojistik Regresyon algoritmasının sansür ile baş etmede dikkate değer performans gösterdiği söylenebilir. Ayrıca örneklem büyüklüğü arttıkça genel olarak algoritmaların doğru sınıflama performanslarının arttığı gözlenmiştir. Özetle, büyük örneklemli veri setlerinde Lojistik Regresyon algoritmasının doğru sınıflandırma oranı ile başarılı sınıflandırma performansı gösterdiği söylenebilir.

Kaynakça

Ahmed, S. E., Aydin, D., & Yılmaz, E. (2020). Nonparametric regression estimates based on imputation techniques for right-censored data. *Advances in Intelligent Systems and Computing, 1001*, 109–120. https://doi.org/10.1007/978-3-030-21248-3_8

Akpınar, H. (2014). *Data : Veri Madenciliği Veri Analizi* (Genişletil). Papatya Bilim Yayınevi.

Aydin, D., & Yilmaz, E. (2018). Modified spline regression based on randomly right-censored data: A comparative study. *Communications in Statistics: Simulation and Computation, 47*(9), 2587–2611. https://doi.org/10.1080/03610918.2017.1353615

Balaban, M. E., & Kartal, E. (2015). *Veri Madenciliği ve Makine Öğrenmesi Temel Algoritmaları ve R Dili İle Uygulamaları* (Birinci Ba). Çağlayan Kitapevi.

Bandyopadhyay, S., Wolfson, J., Vock, D. M., Vazquez-Benitez, G., Adomavicius, G., Elidrisi, M., Johnson, P. E., & O'Connor, P. J. (2015). Data mining for censored time-to-event data: a Bayesian network model for predicting cardiovascular risk from electronic health record data. In *Data Mining and Knowledge Discovery* (Vol. 29, Issue 4, pp. 1033–1069). https://doi.org/10.1007/s10618-014-0386-6

Batista, G. E. A. P. A., & Monard, M. C. (2002). A study of k-nearest neighbour as an imputation method. In *Frontiers in Artificial Intelligence and Applications* (Vol. 87, pp. 251–260).

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer.

Bramer, M. (2007). *Principles of Data Mining. Undergraduate Topics in Computer Science*. Springer Verlag.

Davidson, I., & Tayi, G. (2009). Data preparation using data quality matrices for classification mining. In *European Journal of Operational Research* (Vol. 197, Issue 2, pp. 764–772). https://doi.org/10.1016/j.ejor.2008.07.019

Eröz, İ., & Tutkun, N. A. (2020). Aralıklı Sansürlü Veriler için Sağkalım Modelleri. *Süleyman Demirel Üniversitesi Fen Bilimleri Enstitüsü Dergisi*, *24*(2), 267–280. https://doi.org/DOI: 10.19113/sdufenbed.652776

Frank, E., Hall, B., & Pfahringer, B. (2003). Locally weighted naive bayes. *Proceedings of Conference on Uncertainty in Artificial Intelligence*, 249–256.

Gamgam, H., & Altunkaynak, B. (2017). *SPSS Uygulamalı Regresyon Analizi* (2. Basım). Seçkin Kitapevi.

Gijbels, I. (2010). Censored data. *Wiley Interdisciplinary Reviews: Computational Statistics*, *2*(2), 178–188. https://doi.org/10.1002/wics.80

Goldberg, Y., & Kosorok, M. R. (2012). Q-learning with censored data. *Annals of Statistics*, *40*(1), 529–560.

Guo, G., Wang, H., Bell, D., Bi, Y., & Greer, K. (2003). KNN Model-Based Approach in Classification. In R. Meersman, Z. Tari, & D. C. Schmidt (Eds.), *On The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE* (pp. 986–996). Springer Berlin Heidelberg.

Han, J., Kamber, M., & Pei, J. (2012). *Data Mining Concepts and Techniques* (Third Edit). Morgan Kaufman Publishers.

Harrington, P. (2012). *Machine Learning In Action*. Manning Publications.

Hosmer, D. W., Lemeshov, S., & Sturdivant, R. X. (2013). *Applied Logistic Regression* (Third Edit). John Wiley & Sons, Inc.

Ishwaran, H., Kogalur, U. B., Blackstone, E. H., & Lauer, M. S. (2008). Random survival forests. *Ann. Appl. Stat.*, *2*(3), 841–860.

Khan, F. M., & Zubek, V. B. (2008). Support vector regression for censored data (SVRc): A novel tool for survival analysis. *Proceedings - IEEE International Conference on Data Mining, ICDM*, 863–868. https://doi.org/10.1109/ICDM.2008.50

Lewis, N. D. (2017). *Machine Learning Made Easy with R: An Intuitive Step by Step Blueprint for Beginners*. CreateSpace Independent Publishing Platform.

McNamara, J. M., Green, R. F., & Olsson, O. (2006). Bayes' Theorem and Its Applications in Animal Behaviour. *Oikos*, *112*(2), 243–251. http://www.jstor.org/stable/3548663

Mucherino, A., Papajorgji, P. J., & Paradalos, P. M. (2009). *Data Mining In Agriculture*. Springer.

Mulla, G. A. A., Demir, Y., & Hassan, M. (2021). Combination of PCA with SMOTE Oversampling for Classification of High-Dimensional Imbalanced Data. *Bitlis Eren Üniversitesi Fen Bilimleri Dergisi*, *10*(3), 858–869. https://doi.org/10.17798/bitlisfen.939733

Özdamar, K. (2019). *Paket Programları İle İstatistiksel Veri Analizi-1* (11. Baskı). Nisan Kitapevi.

Özkan, Y. (2008). *Veri Madenciliği Yöntemleri*. Papatya Yayınevi.

Shivaswamy, P. K., Chu, W., & Jansche, M. (2007). A support vector approach to censored targets. *Proceedings - IEEE International Conference on Data Mining, ICDM*, 655–660. https://doi.org/10.1109/ICDM.2007.93

Silahtaroğlu, G. (2013). *Veri Madenciliği Kavram ve Algoritmaları*. Papatya Yayınevi.

Štajduhar, I., Dalbelo-Bašić, B., & Bogunović, N. (2009). Impact of censoring on learning Bayesian networks in survival modelling. In *Artificial Intelligence in Medicine* (Vol. 47, Issue 3, pp. 199–217).

https://doi.org/10.1016/j.artmed.2009.08.001

Vock, D. M., Wolfson, J., Bandyopadhyay, S., Adomavicius, G., Johnson, P. E., Vazquez-Benitez, G., & O'Connor, P. J. (2016). Adapting machine learning techniques to censored time-to-event health record data: A general-purpose approach using inverse probability of censoring weighting. *Journal of Biomedical Informatics*, *61*, 119-131. https://doi.org/https://doi.org/10.1016/j.jbi.2016.03.009

Yılmaz, E., & Aydın, D. (2019). Regresyon Analizinde Sağdan Sansürlü Veriler İçin Önerilen Çözüm Yöntemleri Üzerine Bir İnceleme. *Turkiye Klinikleri Journal of Biostatistics*, *11*(3), 224–238. https://doi.org/10.5336/biostatic.2019-66838

Zhu, J., Ge, Z., Song, Z., & Gao, F. (2018). Review and big data perspectives on robust data mining approaches for industrial process modeling with outliers and missing data. *Annual Reviews in Control*, *46*(1), 107–133. https://doi.org/https://doi.org/10.1016/j.arcontrol.2018.09.003

İstatistik Araştırma Dergisi
Journal of Statistical Research
ISSN: 2791-7614

(2023), 13 (1)
Başvuru/Received: 12/11/2022
Kabul/Accepted: 04/07/2023

https://journal.tuik.gov.tr/

ARAŞTIRMA MAKALESİ RESEARCH ARTICLE

A One-Dimensional Biased Probability Model Based on Himanshu Distribution for Vital Events Related to Migration and Mortality Data

Abhishek AGARWAL
D.D.U Gorakhpur University, INDIA
abhishek.agarwal014@gmail.com
Orcid No: 0000-0002-2425-7471

Himanshu PANDEY
D.D.U Gorakhpur University, INDIA
himanshu_pandey62@yahoo.com
Orcid No: 0000-0002-3839-6560

Abstract

In this paper a one dimensional biased probability model based on Himanshu distribution has been obtained by size biasing the Himanshu distribution introduced by Agarwal and Pandey (2022). Its raw moments and central moments has been obtained. Hence expression of coefficient of variation, Index of dispersion, Skewness, Kurtosis have also been given. The parameter involved in the proposed Model has been obtained by the estimation techniques. The suitability of the one dimensional biased probability model tested through the real data sets related to human migration and mortality patterns of different regions.

Keywords: One Dimensional Biasing, Himanshu Distribution, Moments, Estimation of Parameter, Goodness of Fit

Corresponding Author / Sorumlu Yazar: 1-Abhishek AGARWAL, D.D.U Gorakhpur University, Department of Mathematics and Statistics

2-Himanshu PANDEY, D.D.U Gorakhpur University, Department of Mathematics and Statistics

Citation / Atıf: AGARWAL A., PANDEY H. (2023). A One-Dimensional Biased Probability Model Based on Himanshu Distribution for Vital Events Related to Migration and Mortality Data. İstatistik Araştırma Dergisi, 13 (1), 48-60.

Citation / Atıf: AGARWAL A., PANDEY H. (2023). A One-Dimensional Biased Probability Model Based on Himanshu Distribution for Vital Events Related to Migration and Mortality Data. İstatistik Araştırma Dergisi, 13 (1), 48-60.

Göç ve Mortalite Verilerine İlişkin Yaşamsal Olaylar için Himanshu Dağılımı Temelli Tek Boyutlu Önyargılı Olasılık Modeli

Özet

Bu çalışmada, Agarwal ve Pandey (2022) tarafından tanıtılan Himanshu dağılımına dayalı tek boyutlu önyargılı bir olasılık modeli elde edilmiştir. Ham ve merkezi momentleri çıkarılmıştır. Değişim katsayısı, dağılım indeksi, çarpıklık, basıklık değerleri de verilmiştir. Önerilen modelde yer alan parametre, uygun tahmin yöntemleri ile elde edilmiştir. Tek boyutlu yanlı olasılık modelinin uygunluğu, farklı bölgelerdeki göç ve ölüm örüntülerine ilişkin gerçek veri setleri üzerinden test edilmiştir.

Anahtar kelimeler: Uyum iyiliği, Tek boyutlu yanlılık, Himanshu Dağılımı, Momentler, Parametre.

1. Introduction

Probability model based on distributions play an important role in the various fields of Social Sciences, Medical Sciences, Environmental Sciences etc. Recently Pandey and Jai kishun (2009,2010), Dubey and Pandey (2022) has suggested probability model for vital events.

Himanshu distribution has probability mass function (PMF) as below. It is used by Agarwal and Pandey (2022).

$$P(X=x)=p^n(1-p^n)^x \quad ; \quad \begin{matrix} x=0,1,2,\dots \\ 0<p<1 \\ n\in I^+ \end{matrix} \tag{1.1}$$

for modelling count data from different field of real life problems.

$$\text{with Mean}=\frac{1-p^n}{p^n} \quad \text{and Variance}=\frac{1-p^n}{p^{2n}}$$

One dimensional biased distribution is a special type of weighted distribution. They can be seen while observation from a sample are selected with probabilities proportional to any measures of unit size. It was first given by Fisher (1934). One dimensional biased observation occurs in many research areas related to real life problems includes Environmental Science, Medical Science, Population Science, Psychology, Ecology etc. Van Deusen (1986) has been discussed the application of one dimensional biased distribution theory in fitting distributions of diameter at breast height (DBH) data arising from horizontal point sampling. Lappi and Bailey (1987) have applied one dimensional biased distribution to analyze horizontal point sampling diameter increment data.

Patil and Rao (1977, 1978) has studied detailed statistical application of one dimensional biased distribution to analysis of data relating to human population and ecology.

One dimensional biasing is defined in the following way.

Let a random variable X has original probability distribution(PD) $P_o(x;\theta)$; $\begin{matrix} x=0,1,2,\dots \\ \theta>0 \end{matrix}$

Suppose the sample units are weighted or selected with probability proportional to some measure. x^{α}. Here α is a positive integer. Then the corresponding one dimensional biased PD of order α can be defined by its PMF.

$$P_1(x;\theta) = \frac{x^{\alpha}.P_0(x;\theta)}{\mu'_{\alpha}} \tag{1.2}$$

$$\text{where } \mu'_{\alpha} = E(X^{\alpha}) = \sum_{x=0}^{\infty} x^{\alpha}\, P_0(x\,;\theta)$$

If $\alpha = 1$, then the distribution is known as one dimensional biased and is applicable for one dimensional biased sampling in sampling theory.

The PMF of the one dimensional biased Himanshu distribution with parameter *p* can thus be obtained as;

$$P_2(x;\theta) = \frac{x.P_0(x;\theta)}{\mu'_1} = x.p^{2n}(1-p^n)^{\,x-1}\,; \quad \begin{matrix} x = 1,2,3,\ldots \\ 0 < p < 1 \\ n \in I^+ \end{matrix} \tag{1.3}$$

$$\text{where } \mu'_1 = \frac{1-p^n}{p^n} \text{ is the Mean of Himanshu distribution with p. m. f (1.1)}$$

Mortality is the population's other important occurrence. In the actual world, mortality has emerged as the key concern for hospitals and the insurance sector. Living things have the ability and may even be required to die at some point. Many demographers and social scientists are able to quantify the event mortality, and for the smooth study of mortality pattern they adopted the form of modeling in a probabilistic environment. This is due to the increased interest in recent decades in understanding the mortality pattern and risk. The many models have been suggested in this manner by Pandey and Shukla (2014), Pandey etal. (2015), Agarwal and Pandey (2022) etc.

Migration is the third event of the population change, the other two being Mortality and Fertility. The nature of migration is very complex as a factor affecting population size different from that of mortality and fertility. The Migration affect the Social, Cultural, Economical, Political Characteristic of the society. Any region's population distribution and the expansion of its work force are significantly influenced by migration. Studying the migratory pattern and its effects on society at the micro level may roughly be done using the probability model provided by Aryal (2003, 2011), Singh etal.(2016), Dubey and Pandey (2021), Agarwal and Pandey (2022) etc.

2. Proposed Model

The PMF of the proposed model can be formed by using (1.1) and (1.2) in the following way.

$$P(X = x) = x\,p^{2n}(1-p^n)^{\,x-1} \quad ; \quad \begin{matrix} x = 1,2,3,\ldots \\ 0 < p < 1 \\ n \in I^+ \end{matrix} \tag{2.1}$$

The Moment Generating Function of (2.1) is given as

$$M_X(t) = \sum_{x=1}^{\infty} e^{tx}\, x p^{2n}(1-p^n)^{x-1}$$

$$M_X(t) = \frac{p^{2n}e^t}{\{1-e^t(1-p^n)\}^2} \tag{2.2}$$

Then the first four moments (about origin) are as follows.

$$\text{Mean} = \mu_1' = \frac{2}{p^n} - 1 \qquad \mu_2' = \frac{p^{2n} - 6p^n + 6}{p^{2n}}$$

$$\mu_3' = \frac{-(p^{3n} - 14p^{2n} + 36p^n - 24)}{p^{3n}}$$

$$\text{and} \quad \mu_4' = \frac{p^{4n} - 30p^{3n} + 150p^{2n} - 240p^n + 120}{p^{4n}} \tag{2.3}$$

and central moments as,

$$\left.\begin{aligned} \mu_2 = 2\left(\frac{1-p^n}{p^{2n}}\right) \text{ , } \mu_3 = \frac{2(1-p^n)(2-p^n)}{p^{3n}} \\ \text{and} \quad \mu_4 = \frac{2(1-p^n)(p^{2n} - 12p^n + 12)}{p^{4n}} \end{aligned}\right\} \tag{2.4}$$

$$\left.\begin{aligned} \gamma_1 = \frac{2-p^n}{\sqrt{2(1-p^n)}} \\ \text{and } \gamma_2 = \frac{p^{2n} - 6p^n + 6}{2(1-p^n)} \end{aligned}\right\} \tag{2.5}$$

$$\left.\begin{aligned} \text{Index of dispersion } = \frac{2}{p^n}\left(\frac{1-p^n}{2-p^n}\right) \\ \text{and C.V} = \frac{\sqrt{2(1-p^n)}}{2-p^n} \end{aligned}\right\} \tag{2.6}$$

3. Parameter Estimation

The parameter p of the Model (2.1) can be estimated using Maximum likelihood estimation in the following way.

$$L = \prod_{i=1}^{k} f(x_i\,; p)$$

$$L = (p^{2n})^k \prod_{i=1}^{k} x_i\, (1-p^n)^{\sum_{i=1}^{k} x_i - k}$$

Taking log and upon differentiating w.r.t *p* and equating to zero we get

$$\hat{p} = \left(\frac{2}{\bar{x}+1}\right)^{\frac{1}{n}}$$

Again the parameter *p* estimated by method of moments after using (2.3) we get

$$\bar{x} = \mu_1'$$
$$\Rightarrow \hat{p} = \left(\frac{2}{\bar{x}+1}\right)^{\frac{1}{n}}$$

4. Application

The proposed probability model at *n=2* is fitted using some real data sets collected from different sample surveys entitled "Demographic survey of Chandauli district (Rural Area- 2001-2002)"-(2015); Rural development and population growth survey 1978-PRC, BHU" (2015); and "the survey collected by the researcher in Varanasi district (2018) for the single adult male migrant of 15 years and above.

Mortality does not depend only on the biological and epidemiological factors. It also depends on some socioeconomic and cultural factors. Prevailing health conditions, medical facilities, environmental conditions also worth mentioning. In developing besides under developed countries the mortality among infants and children is much higher than youngsters. These reasons the high infant mortality has thrown a serious challenge to the doctors and medical personnel. So, it can be seen as one of the sensitive position of existing medical and health facilities in the population. Therefore, we have studied the infant mortality pattern using proposed model. In this respect the real data set of Sri Lanka taken from the survey by Meegama (1980) and the real data set of India is taken from Lal (1955).

Table 1: Observed and Expected number of households(NoH) with at least one male migrant according to the number of male migrants aged 15 years and above (2001 survey)

No of migrants	**Observed no. of households**	**Exp. no. of households**
1	97	91.60
2	35	44.56
3	19	16.25
4	6 } 9	7.58
5	3 }	
Total	**160**	**160**
Mean=1.643 Variance=0.8494 $\hat{p}$ =0.8698	$\chi^2 = 3.08$ (after pooling) p-value=0.2143 $\chi^2_{(2)} = 5.99$ at 5% level of significance	

Table 2: Same data with Table 1 according to the 1978 survey

Number of migrants	Observed no. of households	Expected no. of households
1	375	367.2
2	143	154.9
3	49	49.0
4	17	13.7
5	6	5.2
Total	**590**	**590**
Mean=1.535 Variance=0.6780 $\hat{p}$ =0.8882	$\chi^2 = 1.96$ (after pooling) p-value=0.5807 $\chi^2_{(3)} = 7.815$ at 5% level of significance	

Table 3: Same data with Table 1 according to 1978 survey in three types of households.

	Type of households					
	Semi Urban		**Remote**		**Growth Centre**	
	Observed	**Expected**	**Observed**	**Expected**	**Observed**	**Expected**
1	95	86.60	176	169.66	154	146.02
2	19	31.25	59	66.74	47	59.31
3	10	11.15	18	19.69	18	18.06
4	2 } 15		6 } 10		9 } 11	
5	3		4	6.90	2	6.60
Total	**129**	**129**	**263**	**263**	**230**	**230**
Mean=	1.44		1.49		1.51	
Variance=	0.5367		0.6098		0.6397	
χ^2(after pooling) =	6.92		2.65		5.91	
d.f=	1		2		2	
p-value=	0.0085		0.2658		0.0520	

Table 4: Observed and Expected NoHs having adult Male Migrants aged 15 years and above.

Number of migrants	Observed no. of households	Expected no. of households
1	97	93.00
2	42	47.19
3	16	17.95
4	7 } 12	8.86
5^+	5	
Total	**167**	**167**
Mean=1.68 Variance=0.9111 $\hat{p}$ =0.8638	$\chi^2 = 2.07$ (after pooling) p-value=0.3552 $\chi^2_{(2)} = 5.99$ at 5% level of significance	

Table 5: Observed NoHs having adult Male Migrants aged 15 and above in North Eastern Bihar.

Number of migrants	**Observe NoHs**	**Expected NoHs**
1	95	88.10
2	41	48.98
3	15	20.42
4	12	7.57
5	6	3.93
Total	**169**	**169**
Mean=1.77 Variance=1.066 $\hat{p}$ =0.8497	$\chi^2 = 6.96$ (after pooling) p-value=0.0731 $\chi^2_{(3)} = 7.815$ at 5% level of significance	

Table 6: The number of mothers(NoM) of the Rural Area having at least one live birth and one neonatal death(ND).

Number of NDs	**Observed no. of mothers**	**Expected no. of mothers**
1	409	402.67
2	88	97.92
3	19	17.86
4	5 } 6	3.55
5	1 } 6	
Total	**522**	**522**
Mean=1.27 Variance=0.3152 $\hat{p}$ =0.9372	$\chi^2 = 2.86$ (after pooling) p-value=0.2393 $\chi^2_{(2)} = 5.99$ at 5% level of significance	

Table 7: The NoMs of the Estate Area having at least one live birth and one ND.

Number of NDs	**Observed no. of mothers**	**Expected no. of mothers**
1	71	69.27
2	32	32.38
3	7	11.34
4	5 } 8	5.01
5	3 } 8	
Total	**118**	**118**
Mean=1.61 Variance=0.7961 $\hat{p}$ =0.8753	$\chi^2 = 3.48$ (after pooling) p-value=0.1755 $\chi^2_{(2)} = 5.99$ at 5% level of significance	

Table 8: The NoMs of the Urban Area with at least two live births by the number of infant and child deaths.

No. of Infant and child deaths	Observed no. of mothers	Expected no. of mothers
1	176	168.01
2	44	57.19
3	16	14.60
4	6 } 8	4.20
5	2 } 8	
Total	**244**	**244**
Mean=1.41 Variance=0.4943 $\hat{p}$ =0.9109	$\chi^2 = 6.97$ (after pooling) p-value=0.0306 $\chi^2_{(2)} = 9.21$ at 1% level of significance	

Table 9: The NoM of the completed fertility having experienced at least one child death.

No. of child deaths	Observed NoM	Expected NoM
1	89	79.88
2	25	36.85
3	11	12.75
4	6 } 10	5.52
5	3 } 10	
6	1 } 10	
Total	**135**	**135**
Mean=1.60 Variance=0.7797 $\hat{p}$ =0.8770	$\chi^2 = 8.72$ (after pooling) p-value=0.0127 $\chi^2_{(2)} = 9.21$ at 1% level of significance	

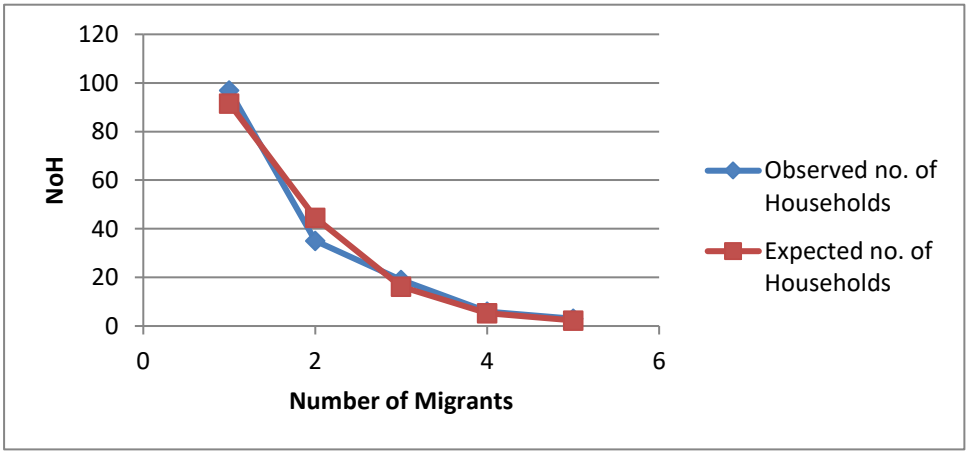


Figure 1: Graphical presentation showing observed and expected NoH aged 15 years and above (survey 2001 data).

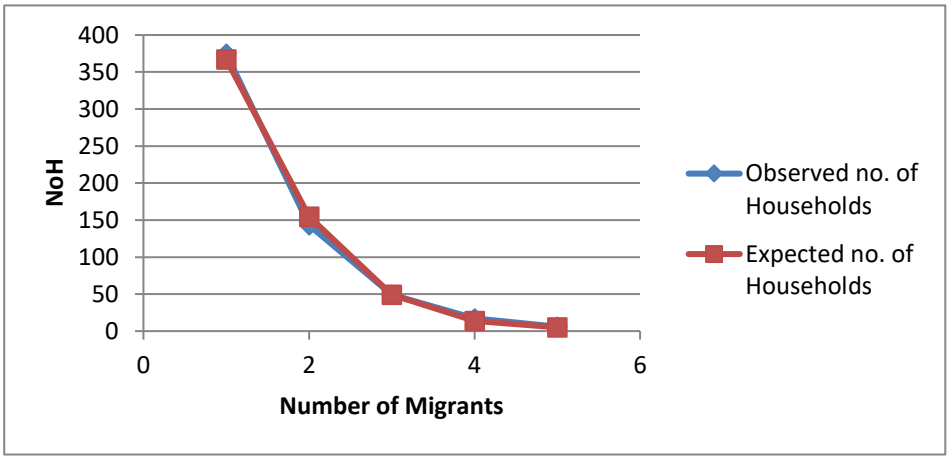


Figure 2: Graphical presentation showing observed and expected NoH with at least one male migrant according to the number of male migrants aged 15 years and above (survey 1978 data).

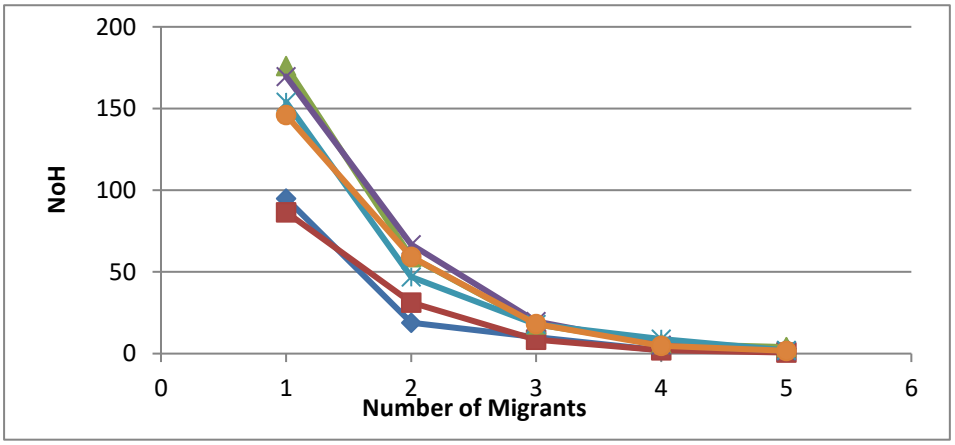


Figure 3: Graphical presentation showing observed and expected NoHs aged 15 years and above (survey 1978 data) in three types of households.

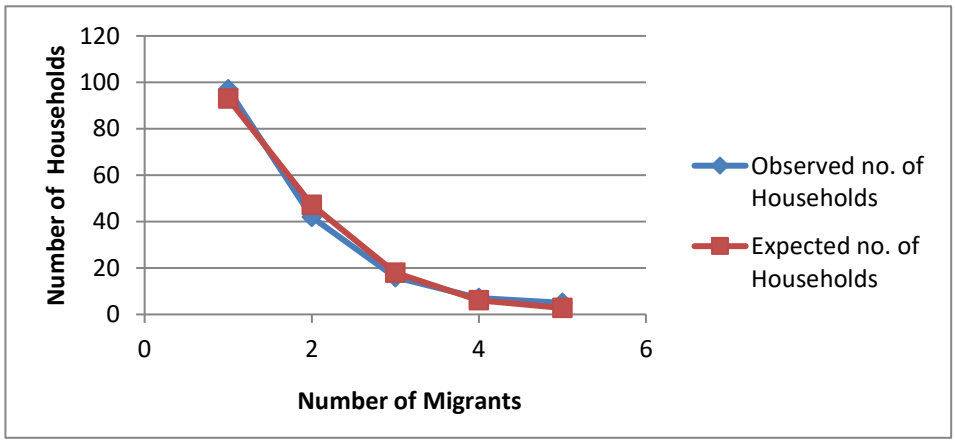


Figure 4: Graphical presentation showing observed and expected number of households having adult Male Migrants aged 15 years and above.

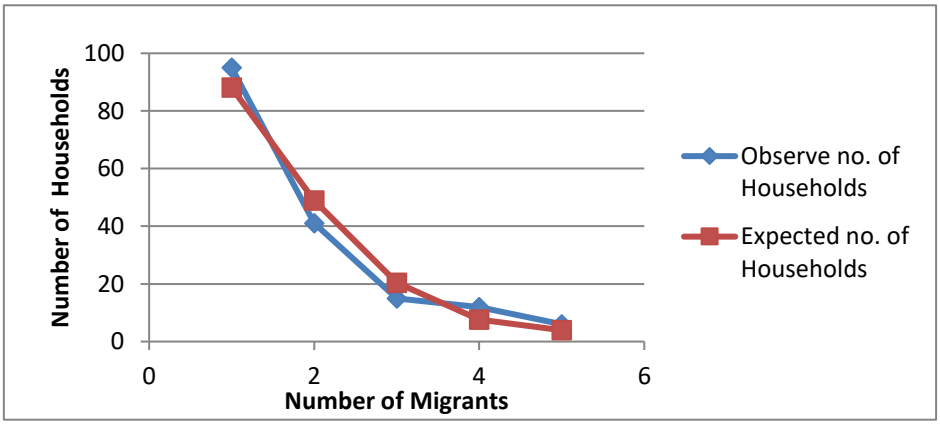


Figure 5: Graphical presentation showing observed and expected NoHs having adult Male Migrants aged 15 and above in North Eastern Bihar.

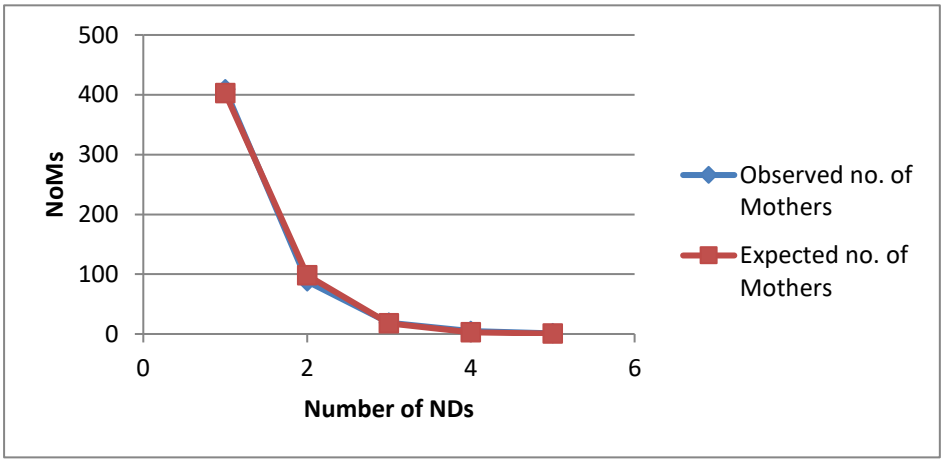


Figure 6: Graphical presentation showing observed and expected no. of mothers of the Rural Area having at least one live birth and one ND.

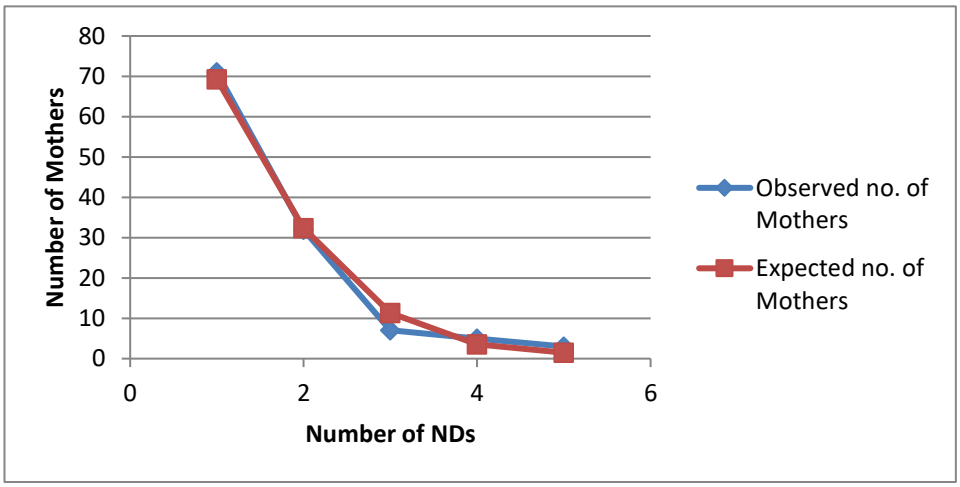


Figure 7: Graphical presentation showing observed and expected no. of mothers of the Estate Area having at least one live birth and one ND.

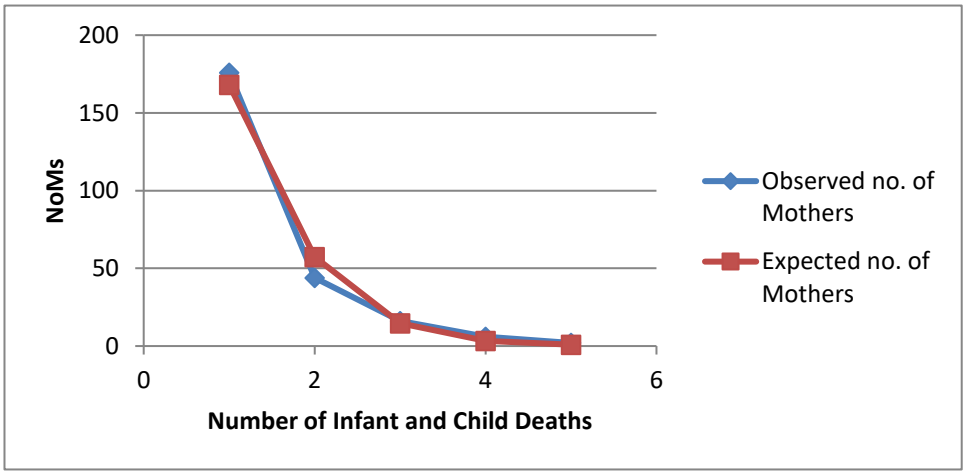


Figure 8: Graphical presentation showing observed and expected no. of mothers of the Urban Area with at least two live births by the number of infant and child deaths.

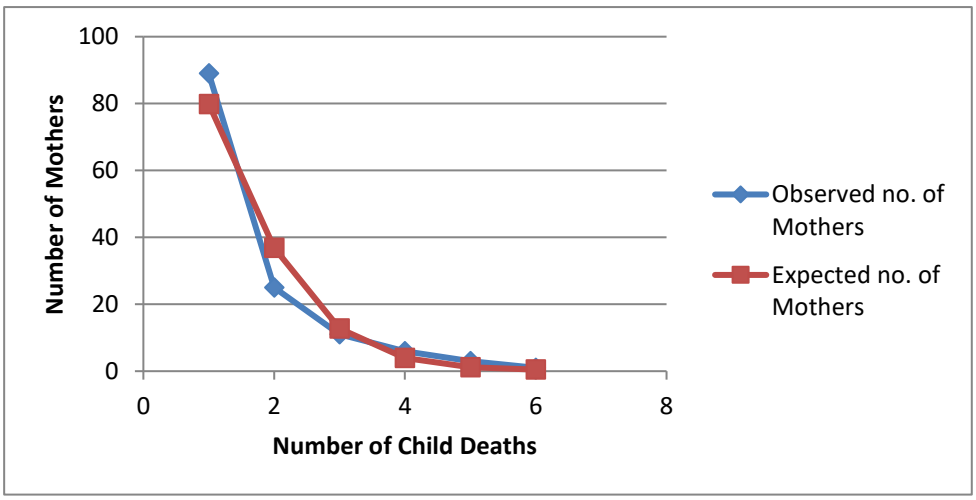


Figure 9: Graphical presentation showing observed and expected no. of mothers of the completed fertility having experienced at least one child death

5. Conclusion

A chi-square goodness of fit test determines if sample data matches a population. A **low value** for chi-square means there is a **high correlation** between your two sets of data. From tables(1 to 9), it clearly indicates calculated χ^2 is less than the critical χ^2 value at 1% and 5% level of significance, Hence we conclude there is no significant difference between the observed and expected value of the given data set. For a Chi-square test, a **p-value** that is less than or equal to your significance level indicates there is sufficient evidence to conclude that the observed distribution is not the same as the expected distribution. According to the value of χ^2 and p-value from (tables 1,2,3,4,5,6,7,8,9) and graphical representation between O_i and E_i, the nature and behavior of proposed one dimensional biased probability model found suitable for the migration pattern as well as infant mortality pattern of different regions. The overall studies shows that the proposed one dimensional biased probability model could also be helpful in policy making, Rural development, Fresh Environment, Medical Facilities for the betterment of the society.

Probability is used in Bayesian analysis for both data and hypotheses. It pertains to a subjective assessment of the veracity of an occurrence. A different approach to traditional statistics is provided by Bayesian statistics. It stands out for its capacity to characterize uncertain values using probability distributions, which leads to elegant solutions to several challenging statistical problems and is extensively useful in the fields of demography, medicine, and insurance. Now that these viewpoints and the work of Rao and Pandey (2020, 2021) have been taken into consideration, it is possible to employ the Bayesian Analysis of the suggested model by figuring out various loss functions.

Acknowledgment

The authors are very thankful to referees for their valuable suggestions.

References

Aryal T.R. (2011). Inflated Geometric distribution to study the distribution of rural out migrants, *Journal of Institute of Engineering*, 8(1&2), 266-268.

Agarwal A., Pandey H., Himanshu distribution and its applications, *Bulletin of Mathematics and Statistics Research*.2022;10(4):24-35.

Agarwal, A., & Pandey, H. (2022). Probability Model Based on Zero Truncation of Himanshu Distribution and Their Applications. *Asian Journal of Probability and Statistics*, 20(4), 57-67.

Agarwal, A., & Pandey, H. (2022). An Inflated Probability Model for the Adult Male Migrants. *International Journal of Statistics and applied Mathematics*, 7(6), 48-52.

Dubey A., Pandey H. (2021). An inflated probability models for Commutation pattern. *Bulletin of Mathematics and Statistics Research*, 8(4),101-108.

Dubey, A., & Pandey, H. (2022). An inflated probability models for adult out migration. *International Journal of Statistics and Applied Mathematics*, 7(5), 80-82.

Fisher, R. A. (1934). The effect of methods of ascertainment upon the estimation of frequencies. *Annals of eugenics*, 6(1), 13-25.

Jai Kishun, Pandey H. (2009). Inflated Probability Model for Risk of vulnerability to HIV/AIDS infection among female Migrants. *Journal of Reliability and Statistical study*, 2(2), 77-84.

Lal DN (1955). Patna in 1955, A Demographic Sample Survey, *Demographic Research Centre*, Department of Statistics, Patna University, India.

Meegama SA (1980). Socio-economic determinants of infant and child mortality in Sri Lanka, An analysis of post war experience, *International Statistical Institute* (World Fertility Survey), Netherland.

Lappi, J. and bailey, R.L. (1987): Estimation of diameter increment function or other tree relations using angle-count samples, *Forest science*, 33, 725 – 739.

Pandey A, Pandey H, Shukla V.K. (2015). An inflated Probability for rural out migration, *Journal of Computer and Mathematical Sciences*, 6(12), 702-711.

Pandey H., Jai kishun, (2010) A probability model for the child mortality in a family, *Revista Colombiana de Estadistica*, 33(1), 1-11.

Patil, G.P. and Rao, C.R. (1977): The Weighted distributions: A survey and their applications. In applications of Statistics, Ed P.R. Krishnaiah, 383 – 405, *North Holland Publications Co*., Amsterdam.

Patil, G.P. and Rao, C.R. (1978): Weighted distributions and size-biased sampling with applications to wild-life populations and human families, *Biometrics*, 34, 179 – 189.

Singh, B. P., & Singh, N. K. (2016). On the distribution of risk of migration and its estimation. *International Journal of Statistical Distributions and Applications*, 2(4), 67-71.

Sahu K.K., Pandey A., Pandey H. (2015). A displaced probability model for child mortality for fixed parity, *International Journal of Science and Research*, 4(12), 1681-1685.

Van Deusen, P.C. (1986): Fitting assumed distributions to horizontal point sample diameters, For. Sci., 32, 146 -148.

Rao A.K, Pandey H. (2020). Parameter estimation of Area biased Ailamujia distribution. *International Journal of Physics and Mathematical Sciences*, 10(1), 21-28.

Rao A.K, Pandey H. (2021). Bayesian estimation for the parameter of Gamma Lomax distribution under different loss function. *International Journal of Scientific Research in Mathematical and Statistical Sciences*, 8(1), 42-51.

İstatistik Araştırma Dergisi
Journal of Statistical Research
ISSN: 2791-7614

(2023), 13 (1)
Başvuru/Received: 14/06/2023
Kabul/Accepted: 27/06/2023

https://journal.tuik.gov.tr/

ARAŞTIRMA MAKALESİ RESEARCH ARTICLE

A Study on the Impact of Money Supply Growth and Government Debt Increase on the Economic Performance in the Euro-Area

Esra N. KILCI
Istanbul University / Assoc. Prof.
esra.kilci@iuc.edu.tr
Orcid No: 0000-0002-2239-4560

Abstract

The severe economic and financial challenges of the COVID-19 pandemic have become increasingly noticeable since March 2020. The European Central Bank has attempted to intervene swiftly and effectively in response to these challenges and has carried out some facilities, including providing additional bank liquidity measures, promoting collateral easing initiatives, and large supplementary acquisitions of government and private sector assets. As a result, we witness an expansion in the balance sheet of the European Central Bank. The objective of this paper is to analyze the impact of money supply (M3) growth and government debt to GDP ratio on the real economy between 2000:Q1 and 2020:Q3. In the analysis, we employ Autoregressive Distributed Lag (ARDL) bounds test is developed by Pesaran et al. (2001). The findings from the ARDL bounds test support the evidence of the impact of money supply (M3) growth and government debt increase on inflation.

Keywords: Economic Growth, COVID-19, Money Supply, Inflation

Corresponding Author / Sorumlu Yazar: 1-Esra N. KILCI, Istanbul University - Cerrahpaşa, Department of Health Economics

Citation / Atıf: KILCI N. E. (2023). A Study on the Impact of Money Supply Growth and Government Debt Increase on the Economic Performance in the Euro-Area. İstatistik Araştırma Dergisi, 13 (1), 61-72.

Citation / Atıf: KILCI N. E. (2023). A Study on the Impact of Money Supply Growth and Government Debt Increase on the Economic Performance in the Euro-Area. İstatistik Araştırma Dergisi, 13 (1), 61-72.

Para Arzı ve Kamu Borcundaki Artışın Euro Bölgesi Ekonomik Performansına Etkisinin Analizi

Özet

COVID-19 salgını Mart 2020'den bu yana tüm dünyada yıkıcı ekonomik ve finansal sonuçları beraberinde getirmiştir. Diğer merkez bankaları gibi, Avrupa Merkez Bankası da, bu sonuçların etkisini hafifletebilmek adına, likiditenin arttırılması, teminatların gevşetilmesi, kamu ve özel sektör tahvillerinin alınması gibi önlemleri içeren bir dizi aksiyon almıştır. Sonuç olarak, Avrupa Merkez Bankası'nın bilançosunda belirgin bir genişleme yaşanmıştır. Bu çalışmanın amacı, 2000:Ç1 ve 2020:Ç3 döneminde, para arzındaki (M3) ve kamu borcundaki artışın reel ekonomi üzerindeki etkisini araştırmaktır. Analizde, Pesaran ve diğ. (2001) tarafından geliştirilen ARDL Sınır Testi Yaklaşımı kullanılmıştır. Analiz sonuçları, para arzındaki (M3) ve kamu borcundaki artışın enflasyon üzerinde etkiye sahip olduğunu göstermiştir.

Anahtar kelimeler: Ekonomik Büyüme, COVID-19, Para Arzı, Enflasyon

Introduction

The European Central Bank (ECB) has responded to the COVID-19 recession by rapidly supporting banks and injecting the Euro-Area with liquidity contrary to its poor reaction to the 2008-09 global financial crisis. However, the recession has made the fiscal imbalances worse, particularly in the countries like Italy and Spain. In the second quarter of 2020, the GDP declined by 12.1 percent. The recovery is expected to rely on effective management of the COVID-19 crisis, while the economic policy, in particular, the degree of monetary and fiscal stimulus, is of critical importance (International Institute for Strategic Studies, 2020). We see that the outbreak of the COVID-19 pandemic and subsequent steps during the first half of 2020 not only led to the closure of several activities in economies but also adversely impacted the banking sector in these countries. Even though banks entered the crisis at higher capital and liquidity levels compared to the 2008-09 global financial crisis period, their resilience has been tested by the sharp tightening observed in financial markets, the increasing funding burdens, and substantial re-pricing of risky assets. Such developments and the growing risk for more detrimental scenarios contributed to extraordinary policy actions. A variety of new monetary and prudential facilities have been established with the policy reaction to the COVID-19 crisis in several countries. In the Euro-Area, the EBC's response involved the readjustment of the targeted longer-term refinancing operations (TLTROs) and the easing of capital requirements by the national macroprudential authorities and centralized micro prudential authority (Baldwin and Weder di Mauro 2020; Acharya and Steffen 2020; Altavilla et al., 2020).

In this study, we attempt to give brief information about the monetary policy response of the European Central Bank since the beginning of the COVID-19 pandemic. We also try to test the impact of the policy response of ECB on the real economy by using the variables as money supply (M3) growth, government debt to GDP ratio, industrial production index, economic growth, inflation, and unemployment level. Our research contributes to academic literature in several ways. First, there are very few studies focusing on the impact of the ECB's response against adverse impacts of the COVID-19 pandemic on the real economy. Secondly, in the empirical analysis, we employ the bootstrap ARDL bounds test to examine the existence of level-relationship using the quarterly data from 2000:Q1 through 2020:Q3. Since the bootstrap ARDL bounds test is powerful even in small samples and allows to investigate the relationship between the variables when the regressors mixed of I(0) and I(1), we prefer to use this test in this research.

The study has the following structure: After mentioning the central bank policy instruments such as quantitative easing since the start of the COVID-19 in Section 1, we will proceed with reviewing the policy response of the ECB in Section 2. Section 3 briefly addresses the academic literature, and Section 4 outlines the empirical methodology. Finally, the last section concludes by giving some policy implications.

The Central Bank Instrument Combination Against COVID-19

Although the COVID-19 shock is mainly a real shock, its impact on financial markets also was significant. When the COVID-19 extends from a Chinese economic epidemic to a global pandemic, equities have plummeted, and financial uncertainty has increased worldwide. At the beginning of February, the major stock markets worldwide have reacted, as it is clear that the outbreak has been spilled over. With cases reported particularly in Italy, Iran, South Korea, and Latin America, the local benchmark indices responded more strongly. The shifts of the stock market represent the most badly hit industries such as travel and leisure or food and catering (CEIC, 2020). In the United States, for instance, recent levels of volatility rival or exceed the levels seen in October 1987 and December 2008 and, before, at the end of 1929 and the early 1930s (Baker et al., 2020). In the U.S, the stock market index has fallen by 30 percent in a few weeks. Figure 1 shows that the CBOE volatility index (VIX) has risen to levels comparable with those during the 2008-2009 Global Financial Crisis. Other indicators signaling financial volatility, such as high yield spreads and investment grades, have shown similar adverse trends. The Federal Reserve had to allocate nearly 20 percent of U.S GDP in financing a wide variety of lending and market supporting facilities in order to avoid the free fall (Federal Reserve, 2020).

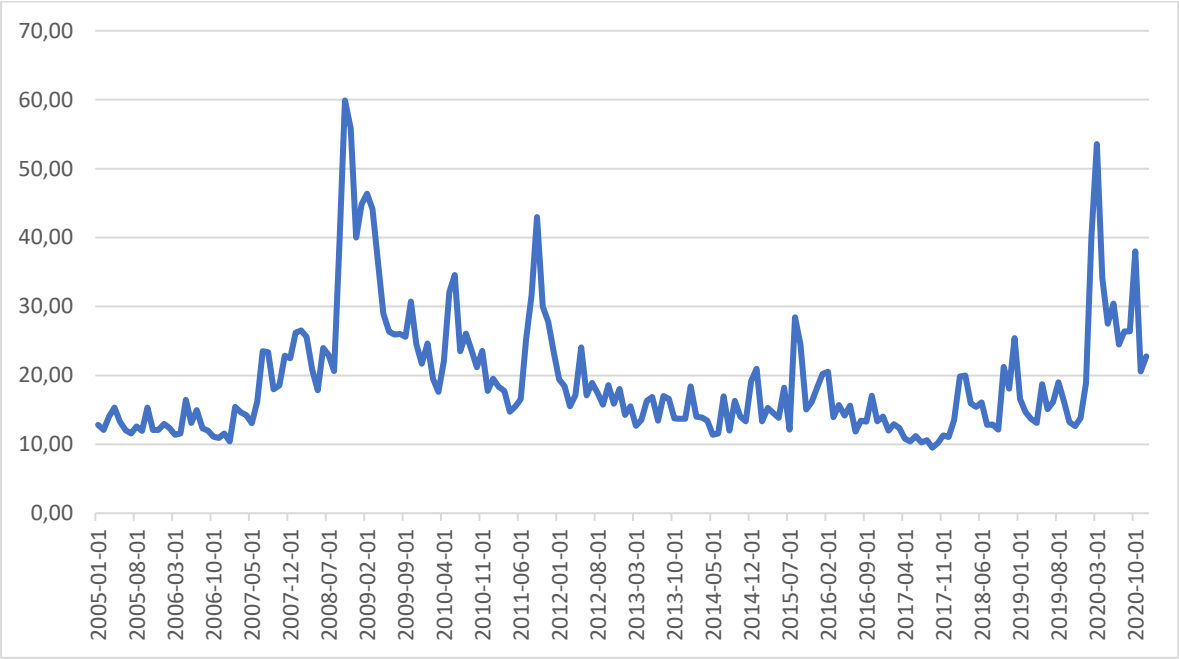


Fig.1: CBOE Volatility Index (VIX)
Source: Federal Reserve Bank of St. Louis (2020).

Until now, many central banks implemented expansionary monetary policies to boost the slowing economies as a result of the pandemic. Policy rate cuts, in other words, interest rate reductions, have been a commonly used tool by central banks. The idea behind interest rate cuts is quite simple; that is, if a central bank pushes downward on short-term interest rates, which reduces the overall borrowing costs, it stimulates corporate spending and consumer expenditure preferably. When short-term rates are almost close to zero, further reductions might have little to no effect. This is why central banks have relied on asset purchase schemes, which is called quantitative easing, to pressure downward on long-term interest rates. That strategy has been a pillar of the Federal Reserve's response against COVID-19, in which newly generated currency purchases hundreds of billions of dollars of assets such as government bonds (Lu, 2020).

Caballero and Simsek (2020) emphasize that conventional monetary policy might reduce the downward pressure if the interest rate is unconstrained. If so, massive asset purchases by government facilities in the context of quantitative easing (QE) are necessary to avoid a downward spiral. The QE, which is used as a last resort by central banks when there are no other tools to restore the economy, was used by FED as a monetary tool during the 2008-09 Global Financial Crisis and by ECB 2010-2014 Euro-Area Debt Crisis. Firstly, QE promotes economic growth, supports businesses in recession, enhances consumer confidence, and makes imports more costly as a re-

sult of the decrease in the currency value owing to the increased money supply (Reddy, 2020). Central banks typically use the QE when other monetary policy tools, including interest rate decreases, are not efficient or sufficient, as mentioned above. In response to the severe economic downturn arising from the COVID-19 pandemic, several central banks, including the FED, the ECB, the BOJ, and other major central banks, are engaged in unprecedented programs for quantitative easing. These programs include major acquisitions of assets, namely central banks purchasing financial instruments like government and corporate bonds. The aim is to supply liquidity on money and capital markets and to promote economic activity. As of April, the U.S Federal Reserve and the other central banks in G7 purchased assets massively. Large-scale acquisitions are expected to extend to the following period depending on the severity of the economic crisis. In view of the current financial crisis, monetary authorities are planning to purchase more and more assets than ever. The Federal Reserve acquires corporate bonds of certain companies for the first time, including bonds that have recently fallen below investment grade (Martinez-Diaz and Christianson, 2020).

How European Central Bank React to the COVID-19 Pandemic?

With the emergence and growing spread of the COVID-19 pandemic, the Euro-Area has been placed under a health emergency that poses serious challenges to the economy and the financial system of the eurozone as well as to the transmission of single monetary policy in addition to the humanitarian crisis (ECB, 2020). The significant economic and financial consequences of the COVID-19 pandemic have become increasingly evident since March 2020. As the pandemic grew, investors readjusted their portfolios, pushing liquidity to crowd out and rising demand for more stable assets in many financial markets. In conjunction with a rise in market-based borrowing costs for companies, the sharp downturn in stock and bond market indices led to a substantial tightening of financial terms in the period from February to April 2020. In this atmosphere, financial markets witnessed a significant risk of adverse liquidity spirals, and excessive asset price changes in several markets could jeopardize financial stability and prevent monetary policy transmission. The monetary policy response of the ECB has centered on main issues as market stabilization, monetary policy safeguarding, providing sufficient central bank liquidity to promote credit provision to the real economy, and assuring that the general stance is effectively accommodative (ECB, 2020b). ECB implemented asset purchases, lending programs, swap-repo lines and taken supervisory measures (ECB, 2020c). Figure 2 shows the extensive package of crisis measures taken by the ECB over the March-August period.

	Asset Purchases	Lending Programs	Swap/Repo Lines	Supervisory Measures
March 2020	Asset Purchase Program (APP) was extended by 120 bn EUR in 2020. Pandemic emergency Purchase Program (PEPP) was launched.	Conditions for Targeted Lending Program (TLTRO-III) was eased. Additional longer-term refinancing operations (LTROs) were launched-facilitating switch into TLTRO-III	EUR Swap Lines were activated with additional central banks. USD Dolar Swap Lines were reactivated with Federal Reserve and other major central banks.	Temporary capital, liquidity and operational assistance: easing use of capital and liquidity buffers, proactive prudential treatment of loans backed by government support measures and prevention of procyclicality in accounting, recommendation against dividend payments.
April 2020		Further easing of TLRO-III conditions, depending on lending performance. Pandemic Emergency Longer-Term Operations (PELTROs) were introduced. Temporary easing of collateral requirements.	EUR Swap Lines were set up with additional central banks.	Temporary reduction in capital requirements for market risk.
June 2020	PEPP was extended by 600 bn EUR to 1,350 bn EUR.		EUREP Repo Facility to supply Euro liquidity to non-Area central banks and EUR Repo Line was set up with additional central banks. Frequency of 7 day USD operations was reduced.	
July and August 2020			EUR repo lines were set up with additional central banks. Frequency of 7 day USD operations was reduced again.	Clarification of capital/liquidity buffers restructuring and supervisory requirements on addressing debtor stress.

Fig. 2: The Crisis Measures of the European Central Bank: March-August Period
Source: ECB (2020b, 2020c).

When reviewed the interest rate policy implemented by the ECB, we see the key interest rates have been left unchanged in March, April, June, July, and September 2020. Accordingly, the main refinancing operation rate, the marginal lending facility, and the deposit facility are 0 percent (since March 2016), 0,25 percent (since March 2016), and -0,50 percent (since September 2019), respectively. Although the ECB has not adjusted the key short-term interest rates like the Federal Reserve and the Bank of England, by adapting its longer-term refinancing operations, it has greatly reduced the long-term funding costs. In this context, in March 2020, the ECB has launched a significant monetary stimulus through asset acquisition schemes. The approved additional net purchases under the new pandemic emergency purchase program (PEPP) and the increased asset purchase program (APP) amount to 7.3 percent of Euro-Area GDP (European Parliament, 2020). Figure 3 shows net purchases of the Euro system.

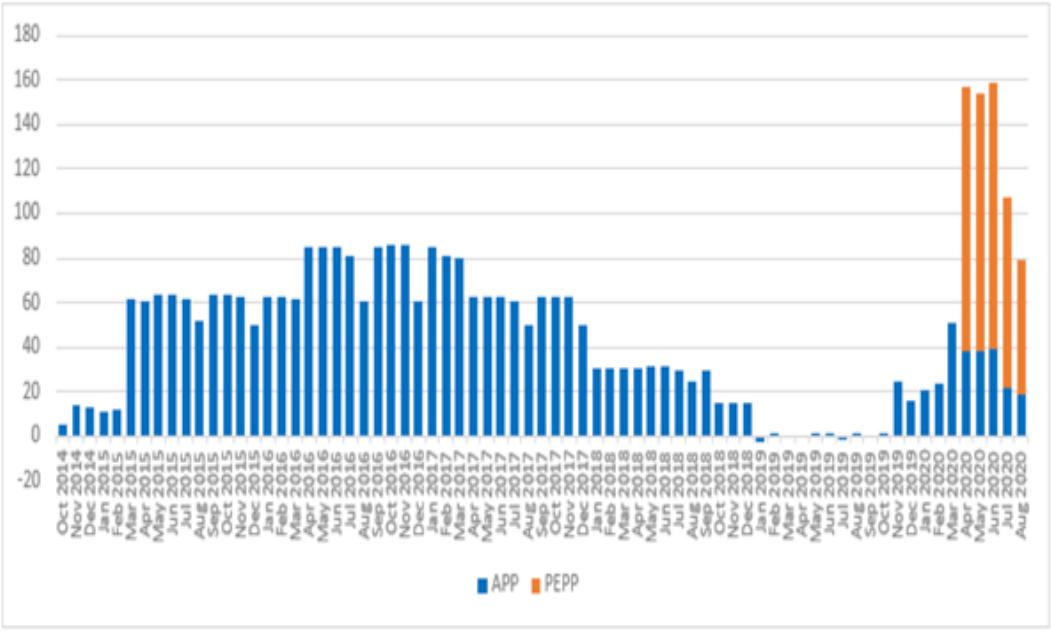


Fig. 3: Euro-system Net Asset Purchases by Month (EUR billion)
Source: EU (2020). Available from:
https://www.europarl.europa.eu/RegData/etudes/BRIE/2020/648787/IPOL_BRI(2020)648787_EN.pdf?cv=1
Access date: 19.10.2020.

The European Central Bank, which has agreed to act quickly and decisively in response to these challenges, has carried out some activities, including providing additional bank liquidity measures (targeted or non-targeted), promoting collateral easing initiatives, and large supplementary acquisitions of government and private sector assets under the APP and the PEPP (ECB, 2020a). In order to address the serious risks to the monetary policy transmission process and outlook for the Eurozone posed by the growing spread of the COVID-19, the ECB has agreed to introduce a new temporary APP of the private and public sector securities. Under the PEPP, purchases have been planned to be implemented until the end of 2020 and to include all the asset categories eligible under the APP (ECB, 2020d). The acquisition of private assets has been a major and focused part of the ECB's policy response to the economic uncertainty caused by the health crisis of COVID-19 (ECB, 2020e).

There are three reasons for extending the coverage of non-financial commercial paper under the CSPP. First of all, by easing funding conditions, it increases the transmission of the ECB monetary policy measures to the real economy. Secondly, it promotes liquidity provision through capital markets, which helps firms handle their needs for short-term financing. Thirdly, commercial paper purchases offer further opportunities for companies to provide access to capital markets (ECB, 2020a). The initiatives of the European Central Bank have led to stabilization of specific conditions of funding in the Euro-Area, improvement of market liquidity, and a decline in uncertainty, thus safeguarding the financial conditions required to achieve the price stability objective. In the second quarter of 2020, financial conditions indicate an unprecedentedly sharp and abrupt tightening in prices of equities, bond markets, foreign exchange, and money markets. Therefore, quick and decisive action was required to assure that what began as a health and economic crisis did not become a full-scale financial crisis with self-fulfilling price spirals and fire sales. In this context, the ECB has tried to stabilize the financial markets that was subject to an extraordinary uncertainty and low levels of liquidity (ECB, 2020e).

Literature Review

In just eight months, we see that primary databases have been filled with research papers, reports, reviews, notes, and editorials focusing on the COVID-19 pandemic. On the other hand, there are a few studies focusing on the macroeconomic implications of the pandemic in countries, particularly in the Euro-Area. There is limited data as the macroeconomic variables are published with a lag. Of these studies on the COVID-19 crisis and economic fundamentals, Jinjarak et al. (2020), Bonatti et al. (2020), Delatte and Guillaume (2020), Aguilar et al. (2020), Masciandaro (2020), Andries et al. (2020), Ortmans and Tripier (2020) analyze the developments in Euro-Area. In attempting to compare the significance of the dominating market conditions and the policy responses to the

COVID-19 in addressing the growth of Eurozone (EZ) sovereign spreads throughout the first half of 2020, Jinjarak et al. (2020) concentrate on CDS premiums in Eurozone by employing a multi-stage econometric approach. In this regard, they aim to generalize model-implied changes in the CDS premiums for the period of 2019:07-2020:06 following estimating a multi-factor model for changes in CDS premiums in the pre-pandemic period of 2014:01-2019:07. They conclude that their model is doing well to measure the realization of the sovereign spreads over the remainder of 2019 but falls throughout the pandemic period, which varies dramatically in 2020:03. In the second phase, with an emphasis particular to the 2020 period, they see that the COVID-19 specific threats and related policies are well represented in the deviation experienced in 2020:03.

The COVID-19 pandemic deeply transformed the economic systems, economic procedures, and policymaking frameworks of the Eurozone. Owing to its deteriorating impact on the supply and demand side, much of the conventional thinking in economic policymaking tends to be inadequate. The COVID-19 pandemic has taken the ECB into a trilemma in maintaining the integrity of Euro-Area against major flight into safety phenomenon and self-fulfilling predictions, on the one-part, monetary orthodoxy and fiscal orthodoxy on the others. By illustrating the fundamental effect of COVID-19 on the economic system and by emphasizing the discrepancies between the new policy framework and past ECB programs, Bonatti et al. (2020) introduce the package that the ECB has taken to deal with the pandemic crisis. Furthermore, they address the medium to long-term challenges facing the ECB, which depend on the various post-COVID scenarios relating to economic growth and inflation, given its specific multinational competence.

Although the pandemic has represented the exogenous shock that contributed to the overall increase in the sovereign debt, the sovereign risk premiums in the Euro-Area have been heterogeneous. From this perspective, Delatte and Guillaume (2020) attempt to estimate the causal factors of sovereign bond spread in the Euro-Area in the period of 2020:01-2020:05. Their findings indicate that the resiliency to COVID-19 relies on the initial fiscal condition, the robustness of banks, and the potential for medical care. They also conclude that the ECB statements have become a game-changer during the recession and have a much greater contribution than the asset purchases programs. Moreover, the European Council's cooperation has also led to a reduction in the spread, but the consequence was largely compensated for by financial funding, focused on loans, which tended to increase spreads. Aguilar et al. (2020) intend to study the steps taken by the ECB and evaluate the economic and financial consequences of the major activities in both the Euro-Area and Spain. In this context, after they clearly define the monetary policy situation of the pre-virus Euro-Area, which is represented by low inflation and low-interest rates, they address the numerous steps taken by the ECB after the outbreak of pandemic in March 2020. Eventually, they examine using a range of mathematical methods how the major initiatives introduced during the present pandemic could theoretically influence economically and financially. These instruments imply that these will have positive first-order impacts on the GDP and inflation of the Euro-Area and Spain, considering the complexity of quantifying the resulting magnitude of the financial and economic downturn in the absence of this initiative. Masciandaro (2020) addresses the implementation of linkages between a fiscal backstop of currency transfers and a helicopter monetary strategy which creates losses in the central bank's balance sheet without substantial adjustment in the money base. Firstly, he finds that an optimal helicopter monetary policy could be developed when a central bank operates as a long-sighted policymaker. The characteristics of this strategy can be determined by taking into consideration fiscal risk, public debt costs, and macroeconomic characteristics overall. The policy composition might create distributional consequences if the responsible government consists of career-concerned officials and people are heterogeneous.

Globally, policymakers struggle with the COVID-19 pandemic by a combination of public health, fiscal, macroprudential, monetary, or market-oriented policies. In an event analysis approach, Andries et al. (2020) evaluate the impacts of the pandemic on sovereign CDS spreads throughout Euro-Area. They conclude that larger numbers of cases and deaths dramatically raise market volatility in European government bonds. From a different point of view, Ortmans and Tripier (2020) attempt to assess the response to the release of the COVID-19 case numbers and investigate the development it has taken around the ECB's facilities. They use national capital markets and country and time-specific impacts to assess the impartial influence of COVID-19 on sovereign risk in the Euro-Area. They examine how the ECB's monetary policy announcements have halted the spread of the COVID-19 pandemic in the European sovereign debt markets. Their findings demonstrate that new cases in Euro-Area have a major and enduring impact upon the sovereign bonds, and then this impact disappears with the ECB's press conference, implying that this change is a positive consequence of ECB's announcement.

Data and Econometric Methodology

Data

In our analysis, we use money supply (M3), employment rate (ER), inflation rate (INF), government debt (GD), and economic growth (EG). We try to analyze how the European Central Bank affects the real economy so we use money supply and government debt. Secondly, since we aim to see how money supply has impact on the real economy we utilize some macroeconomic variables as inflation, employment and economic growth.

To investigate the effects of money supply(M3) and government debt (GD) on the real economy (RE), we use the following model:

$$RE_t = \alpha_1 + \alpha_2 M3_t + \alpha_2 GD_t + e_t \quad (1)$$

Where RE_t shows an indicator of real economy. We employ employment rate (ER), inflation rate (INF), and economic growth (EG) as the proxy of the real economy. $M3_t$ is the broad money supply change while GD_t indicates of Government Debt (as a % of GDP). The data for these variables are obtained from the European Central Bank Statistical Data Warehouse, covering the period between 2000:Q1 and 2020:Q3.

ARDL Bootstrap Test

In this study, we employ an augmented version of one of the most popular cointegration tests. ARDL bounds test is developed by Pesaran et al. (2001), and due to its advantages over the existing cointegration tests, numerous practitioners have implemented the method. The ARDL bounds test is found as efficient in even small sample sizes (Narayan, 2005); the pre-condition that all of the regressors must be stationary at the first differences has been relaxed in the ARDL bounds test; to remedy autocorrelation and endogeneity problems, one can augment the test equation with appropriate lag orders (Nkoro and Uko, 2016), lastly, to reflect the specific effects of regressors, different lag orders of regressors are allowed unlike some other cointegration tests, such as Johansen cointegration test (Thao and Hua, 2016).

We employ Eq.1 in the unrestricted error correction form to implement the ARDL bounds test as follows:

$$\begin{aligned}\Delta RE_t = {} & \beta_1 + \beta_2 RE_{t-1} + \beta_3 M3_{t-1} + \beta_4 GD_{t-1} \\ & + \sum_{i=1}^{k} \gamma_i \Delta RE_{t-i} + \sum_{i=0}^{l} \phi_i \Delta M3_{t-i} + \sum_{i=0}^{m} \theta_i \Delta GD_{t-i} + u_t\end{aligned} \quad (2)$$

where Δ is the first difference operator. We use Akaike information criteria to determine the optimal lag length. To test the existence of a long-run relationship among the variables, the validity of the following hypotheses is investigated:

$$H_{0A} : \beta_2 = 0$$

$$H_{0B} : \beta_2 = \beta_3 = \beta_4 = 0$$

One can test the first hypothesis using t-test statistic (t) while the second using F-test statistic (F_I). However, the decision process of the ARDL bounds test is different from the traditional test. There are mainly two different sets of critical values. First is classified as lower bound (I(0)), second is defined as upper bound (I(1)). If the test statistics are found as higher than the upper critical values, then it is concluded that there is a cointegration relationship among the variables. Else, if the test statistics are lower than the lower bound, one can conclude that there is not any long-run relationship among the variables. The region between two bounds is called the zone of indifference, and if the test statistics are in the zone of indifference, one cannot decide whether there is a long-run relationship or not.

McNown et al. (2018) have improved the ARDL bounds test in two ways. First, they suggest complementing H_{0A} and H_{0B} hypotheses by testing the following hypothesis to reveal the degenerate case:

$$H_{0C} : \beta_3 = \beta_4 = 0$$

This hypothesis can be tested by the F-test (F_2). McNown et al. (2018) also suggest obtaining the critical values via bootstrap simulations to remedy the inconclusive results. To conclude that there is a cointegration among the variables, all of three statistics (t, F_1, and F_2) must be statistically significant.

Empirical Results

As a first step of the analysis, we test the unit root properties of the variables using augmented Dickey-Fuller (ADF) and Phillips-Perron (PP) unit root tests. Table 1 presents the results:

Table 1: The Results of Unit Root Tests

Series	ADF Unit Root Test	PP Unit Root Test
EG	-0.864 (0.794)	-1.361 (0.597)
ER	-2.389 (0.149)	-1.283 (0.634)
GD	-0.062 (0.949)	-0.427 (0.899)
M3	-1.374 (0.591)	-1.822 (0.368)
INF	-1.471 (0.543)	-2.506 (0.118)

Note: The numbers in the parentheses show the p-values.

The results in Table 1 provide evidence that all variables are not stationary at level[1]. So, we can apply the bootstrap ARDL bounds test to examine the long-run relationship between the real economic variables and M3, GDEBT. Table 2 contains the results of the bounds test:

Table 2: Results of the Bounds Test

Selected Model: ARDL(4, 1, 3)	Test Statistics	Critical Values		
Dependent Variable: EG		0.9	0.95	0.99
F_I	1.343	3.156	4.040	6.314
T	1.783	-1.576	-1.947	-2.790
F_2	0.237	3.591	4.850	7.839
Selected Model: ARDL(2, 1, 0)	Test Statistics	Critical Values		
Dependent Variable: ER		0.9	0.95	0.99
F_I	1.162	4.251	5.091	6.904
T	-1.238	-2.818	-3.281	-3.865
F_2	0.251	3.300	4.302	6.000
Selected Model: ARDL(3, 1, 0)	Test Statistics	Critical Values		
Dependent Variable: INF		0.9	0.95	0.99
F_I	4.386***	4.191	5.098	6.980
T	-3.356***	-3.067	-3.411	-4.142
F_2	4.604***	4.524	5.598	8.957

Note: *** shows the significance at the 10% level. We run 2000 simulations to obtain bootstrap critical values.

[1] Results of the tests for the difference data confirm that all variables become stationary at the first difference.

The findings from the ARDL bounds test, support the evidence of the existence of a long-run relationship for the $INF = f(M3, GD)$ equation since all test statistics are higher than the bootstrap critical values, so we estimate the long-run and short-run coefficients and tabulate the results in Table 3:

Table 3: Short-run and Long-run Coefficients

Variables	Coefficients	p-values
Short-run coefficients		
$D(INF_{t-1})$)	0.290*	0.006
$D(INF_{t-1})$	0.181	0.101
D(GD)	-0.101*	0.001
ECT	-0.252*	0.000
Long-run coefficients		
M3	-0.120	0.151
GDEBT	-0.073*	0.002
Constant	8.228*	0.000

Note: * shows the significance at the 1% level.

The error correction term (ECT) is found as negative and significant, which indicates that deviations will be corrected in the long-run. Besides, the GD find as significant and has a decreasing effect on the INF in both the short and long-run. Furthermore, the findings from the ARDLS bounds test indicate that money supply growth and government debt increase have a long-term impact on inflation, indeed, it is an expected effect theoretically.

Conclusion

In the first half of 2020, the COVID-19 outbreak and subsequent steps resulted in the closure of a number of economic activities and also adverse repercussions on the financial sector in those countries. The fiscal authorities and the central banks have responded quickly in order to mitigate the adverse impacts of the COVID-19 pandemic. There is growing evidence that the swift and decisive policy implementations of the central banks have already contributed to the improvement of the macroeconomic outlook, although it seems early to evaluate the full effect of these measures since the effects of the COVID-19 pandemic still continues. In this paper, we attempt to analyze the impact of the policy response of the European Central Bank on the economic activity between 2000:Q1 and 2020:Q3 following giving brief information about how the ECB has reacted to the economic and financial crisis resulting from the COVID-19 pandemic. In the empirical analysis, we employ the ARDL bounds test developed by Pesaran et al. (2001) to test the impact of money supply (M3) growth and government debt to GDP ratio on inflation, economic growth, industrial production index, and employment rate. We find that there are long-run relationships between money supply (M3) growth and inflation as well as government debt to GDP ratio and inflation. These results are in line with the other studies including Reddy (2020), Bonatti et al. (2020) and Aguilar et al. (2020) that confirms the expansionary monetary policy practices do have impacts on some macro-economic variables like inflation. On the other hand, we could not find a long-term impact of money supply growth and government debt increase on economic growth and employment rate. According to several researches, the power of the monetary policy implemented by the ECB in the pandemic period has seemed limited. For instance, as indicated by Lepetit and Fuentes-Albero (2022) flexible monetary measures, such as forward guidance, cause huge rises in inflation but have relatively modest effects on real economic activity as long as the risk of infection remains high. In the analysis which they examine the role of monetary policy in the COVID-19 pandemic, their conclusions concern the efficiency of monetary policy, implying that monetary policy is less successful in a pandemic than in normal times.

Our study that tries to analyze the key actions taken by the European Central Bank in 2020 against the COVID-19 pandemic yields the conclusion that these responses had a positive impact on the Euro-Area macroeconomic performance by increasing money supply, enhancing bank lending activity, easing collateral requirements and by indirectly paving the way for expansionary fiscal policies in the Euro-Area. The research on the impact of the policies of the ECB since the pandemic has increased in the last period. In this context, Benigno et al. (2022) emphasize that it is still unclear how this new European "policy mix" will develop. It is important to note that

the centralized fiscal policy, which was introduced in the EU and EA in 2020 and primarily implemented in 2021, is exceptional and transient. Furthermore, the projects, or a mix of reforms, public investments, and private investments, should be focused on three pillars: the "green" transition, digital innovation, and social inclusion. Finally, in terms of financial inclusion, the ECB implemented several programs including purchasing financial assets, supporting the balance sheets of financial institutions, decreasing the pressure in the financial markets since the onset of the pandemic to stabilize the financial system and these programs are expected to continue in the future by focusing particularly on the green transition and digital innovation.

References

Acharya, V and S Steffen (2020), "Stress tests' for banks as liquidity insurers in a time of COVID", VoxEU.org, 22 March. Available from: https://voxeu.org/article/stress-tests-banks-liquidity-insurers-time-covid Access date: 18.10.2020.

Aguilar, P., Arce, Ó., Hurtado, S., Mertinez-Martin, J., Nuño, G. & Thomas, C. (2020). The ECB Monetary Policy Response to the COVID-19 Crisis. Banco de Espana, Documentos Ocasionales, No: 2026. Available from: https://www.bde.es/f/webbde/SES/Secciones/Publicaciones/PublicacionesSeriadas/DocumentosOcasionales/20/Files/do2026e.pdf Access date: 01.11.2020.

Altavilla, C., Barbiero, F., Boucinha, M. & Burlon, L. (2020). The COVID-19 policy response and bank lending, VoxEU.org, 03 October. Available from: https://voxeu.org/article/covid-19-policy-response-and-bank-lending Access date: 18.10.2020.

Andries, Alin Marius and Ongena, Steven R. G. and Sprincean, Nicu, The COVID-19 Pandemic and Sovereign Bond Risk (May 17, 2020). Swiss Finance Institute Research Paper, No: 20-42.

Baker, S.R., Bloom, N., Davis, S.J, Kost, K.J., Sammon, M.C. & Viratyosin, T. (2020). The Unprecedented Stock Market Impact of Covid-19. NBER Working Paper 26945. Available from: https://www.nber.org/system/files/working_papers/w26945/w26945.pdf Access date: 18.10.2020.

Baldwin, R. & Weder di Mauro, B. (2020), Economics in the time of COVID-19, a CEPR press eBook. Available from: https://voxeu.org/content/economics-time-covid-19 Access date: 18.10.2020.

Benigno, P. Canofari, P., Bartolomeo, G. & Messori, M. (2022). The European Monetary Policy Responses During the Pandemic Crisis. Open Econ Rev, 33(4).

Bonatti, L., Fracasso, A. & Tamborini, R. (2020). COVID-19 and the Future of Quantitative Easing in the Euro Area: Three Scenarios with a Trilemma. European Parliament, Monetary Dialogue Papers, September 2020. https://movimentoeuropeo.it/images/documenti/IPOL_IDA2020652740_EN.pdf Access date: 19.10.2020.

Caballero, R. & Simsek, A. (2020). A risk-centric perspective on the central banks' Covid-19 policy response. VoxEU.org, 30 April. Avaialble from: https://voxeu.org/article/central-banks-covid-19-policy-response Access date: 18.10.2020.

CEIC (2020). The Global Impact of COVID-19 on Financial Markets. Available from: https://info.ceicdata.com/the-global-economic-impact-of-covid-19-on-financial-markets Access date: 23.10.2020.

Delatte, A L and A Guillaume (2020), "Covid 19: a new challenge for the EMU", CEPR Discussion Paper 14848. Available from: https://cepr.org/active/publications/discussion_papers/dp.php?dpno=14848 Access date: 25.10.2020.

European Central Bank (2021). Statistical Data Warehouse, Available from: https://sdw.ecb.europa.eu Access date: 14.01.2021.

ECB (2020a). The ECB's commercial paper purchases:
A targeted response to the economic disturbances caused by COVID-19. The ECB Blog. Blog post by Luis de Guindos, Vice-President of the ECB, and Isabel Schnabel, Member of the Executive Board of the ECB. 3 April. Available from: https://www.ecb.europa.eu/press/blog/date/2020/html/ecb.blog200403~54ecc5988b.en.html. Access date: 07.10.2020.

ECB (2020b). The impact of the ECB's monetary policy measures taken in response to the COVID-19 crisis. Published as part of the ECB Economic Bulletin, Issue 5/2020. Available from: https://www.ecb.europa.eu/pub/economic-bulletin/focus/2020/html/ecb.ebbox202005_03~12b5ff68bf.en.html Access date: 18.10.2020.

ECB (2020c). The pandemic emergency: the three challenges for the ECB. Speech by Philip R. Lane, Member of the Executive Board of the ECB, at the Jackson Hole Economic Policy Symposium, Federal Reserve Bank of Kansas City "Navigating the Decade Ahead: Implications for Monetary Policy". Available from: https://www.ecb.europa.eu/press/key/date/2020/html/ecb.sp200827~1957819fff.en.html Access date: 14.09.2020.

ECB (2020d). ECB announces €750 billion Pandemic Emergency Purchase Programme (PEPP). Press Release. Posted 18 March. Available from: https://www.ecb.europa.eu/press/pr/date/2020/html/ecb.pr200318_1~3949d6f266.en.html Access date: 17.10.2020.

ECB (2020e). The ECB's response to the COVID-19 pandemic. Speeches. Remarks by Isabel Schnabel, Member of the Executive Board of the ECB, at a 24-Hour Global Webinar co-organised by the SAFE Policy Center on "The COVID-19 Crisis and Its Aftermath: Corporate Governance Implications and Policy Challenges".

European Council (2020). 10 things the EU is doing to fight COVID-19 and ensure recovery. Policies. COVID-19 Coronavirus Pandemic. Available from: https://www.consilium.europa.eu/en/policies/coronavirus/10-things-against-covid-19/ Access date: 21.10.2020.

European Parliament (2020). The ECB's Monetary Policy Response to the COVID-19 Crisis. Briefing ECON in Focus. Updated 25 September. Available from: https://www.europarl.europa.eu/RegData/etudes/BRIE/2020/648787/IPOL_BRI(2020)648787_EN.pdf?cv=1 Access date: 19.10.2020.

Federal Reserve (2020). Recent Balance-sheet Trends. Available from: https://www.federalreserve.gov/monetarypolicy/bst_recenttrends.htm Access date: 20.11.2020.

Federal Reserve Bank of St. Louis (2020). Economic Data-Financial Indicators. Available from: https://fred.stlouisfed.org/series/VIXCLS Access date: 15.12.2020.

Jinjarak, Y., Ahmed, R., Nair-Desai, S., Xin, W. & Aizenman, J. (2020). Pandemic Shocks and Fiscal-Monetary Policies in the Eurozone: COVID-19 Dominance during January – June 2020. NBER 27451. Available from: https://www.nber.org/system/files/working_papers/w27451/w27451.pdf Access date: 24.10.2020.

IISS (2020). The European Central Bank and the COVID-19 recession, 26 (19), Publications, Available from: https://www.iiss.org/publications/strategic-comments/2020/covid-19-ecb Access date: 20.10.2020

Lepetit, A. & Fuentes-Albero, C. (2022). The Limited Power of Monetary Policy in a Pandemic, European Economic Review, 147, 104168.

Lu, M. (2020). How Global Central Banks are Responding to COVID-19, in One Chart. Visual Capitalist. 20 May. Available from: https://www.visualcapitalist.com/global-central-banks-policy-response-covid-19/ Access date: 08.10.2020.

Martinez-Diaz & Christianson, G. (2020). Quantitative Easing for Economic Recovery Must Consider Climate Change. 11 May. Available from: https://www.wri.org/blog/2020/05/coronavirus-responsible-quantitative-easing Access date: 14.10.2020.

Masciandaro, D. (2020). Shh, don't say it! ECB Helicopter Money: Economics and Politics. SUERF Policy Note. The European Money and Finance Forum. Issue No: 161. Available from: https://www.suerf.org/docx/f_f1b9528d5fb5c272d2f05a5b82611b3c_12871_suerf.pdf Access date: 18.10.2020.

McNown, R., Sam, C. Y., & Goh, S. K. (2018). Bootstrapping the autoregressive distributed lag test for cointegration. Applied Economics, 50(13), 1509-1521.

Narayan, P. K. (2005). The saving and investment nexus for China: evidence from cointegration tests. Applied economics, 37(17), 1979-1990.

Nkoro, E., & Uko, A. K. (2016). Autoregressive Distributed Lag (ARDL) cointegration technique: application and interpretation. Journal of Statistical and Econometric methods, 5(4), 63-91.

Ortmans, A. & Tripier, F. (2020). COVID-Induced Sovereign Risk in the Euro Area: When Did the ECB Stop the Contagion? CEPII Working Paper, No: 2020-11. Available from: http://www.cepii.fr/PDF_PUB/wp/2020/wp2020-11.pdf Access date: 5.11.2020.

Pesaran, M. H., Shin, Y., & Smith, R. J. (2001). Bounds testing approaches to the analysis of level relationships. Journal of applied econometrics, 16(3), 289-326.

Reddy, S. (2020). Quantitative easing: Is it a solution for economic crisis due to Covid19. Deccan Chronicle. 16 May. Available from: https://www.deccanchronicle.com/opinion/op-ed/160520/quantitative-easing-is-it-a-solution-for-economic-crisis-due-to-covid.html Access date: 06.10.2020.

Thao, D. T. & Hua, Z. J. (2016). ARDL bounds testing approach to cointegration: Relationship international trade policy reform and foreign trade in Vietnam. International Journal of Economics and Finance, 8(8), 1-84.

İstatistik Araştırma Dergisi
Journal of Statistical Research
ISSN: 2791-7614

(2023), 13 (1)
Başvuru/Received: 15/06/2023
Kabul/Accepted: 05/07/2023

https://journal.tuik.gov.tr/

ARAŞTIRMA MAKALESİ RESEARCH ARTICLE

The Effect of Corruption on the Tax Revenues: Case of EU Transition Economies

Ceyda TUNÇ YILANCI
Çanakkale Onsekiz Mart University / M.A.
ceyda.yilanci@comu.edu.tr
Orcid No: 0000-0003-3253-1693

Mahmut Ünsal ŞAŞMAZ
Uşak University / PhD
mahmut.sasmaz@usak.edu.tr
Orcid No: 0000-0001-9485-3933

Abstract

Taxation has historically been one of the primary sources of revenue for governments to meet their public needs. However, social, political, and economic factors have continued to shape the role of taxes in society over time. These factors directly or indirectly impact the determination of tax revenues. Economic factors, such as financial development, growth, economic freedom, tax rates, and globalization, also significantly influence taxation. Corruption, which involves the abuse of public power or resources for private gain, can have numerous economic effects. This study aimed to examine the relationship between corruption and tax revenues in 11 EU transition economies during the 2003-2015 period, utilizing the Westerlund and Edgerton (2007) panel cointegration test and the Emirmahmutoglu and Kose (2011) panel causality tests. The results indicate a long-run relationship between corruption and tax revenue, as well as a bidirectional causality relationship between the two variables.

Keywords: Corruption, Tax Revenues, Panel Data Analysis

Corresponding Author / Sorumlu Yazar: 1- Ceyda TUNÇ YILANCI, Çanakkale Onsekiz Mart University Çanakkale Vocational School of Social Sciences Accounting and Tax Applications Department

2- Mahmut Ünsal ŞAŞMAZ, Uşak University, Faculty of Economics and Administrative Sciences, Department of Finance

Citation / Atıf: TUNÇ YILANCI C., ŞAŞMAZ M. Ü. (2023). The Effect of Corruption on the Tax Revenues: Case of EU Transition Economies. İstatistik Araştırma Dergisi, 13 (1), 73-84.

This paper is taken from the thesis "Yolsuzluğun vergi gelirleri üzerindeki etkisi: Avrupa Birliği geçiş ekonomileri örneği" conducted in 2018 at Usak University under the direction of Mahmut Ünsal Şaşmaz.

Citation / Atıf: TUNÇ YILANCI C., ŞAŞMAZ M. Ü. (2023). The Effect of Corruption on the Tax Revenues: Case of EU Transition Economies. İstatistik Araştırma Dergisi, 13 (1), 73-84.

Yolsuzluğun Vergi Gelirleri Üzerindeki Etkisi: AB Geçiş Ülkeleri Örneği

Geçmişten günümüze, devletlerin kamusal ihtiyaçlarını karşılamada en önemli kaynaklardan biri olan vergiler, sosyal, siyasal ve ekonomik bir olgu olarak karşımıza çıkmaktadır. Vergi gelirlerinin belirlenmesinde kilit rol oynayan bu sosyal, siyasal ve ekonomik olgular, doğrudan ya da dolaylı olarak vergi üzerinde belirleyici olmaktadır. Vergi gelirlerinin belirlenmesinde önemli bir rol oynayan yolsuzluk, küreselleşen dünyada yaygın bir sorun haline gelmiştir. Kavramsal olarak, kamu gücünün veya kaynaklarının özel çıkarlar için kötüye kullanılması olarak tanımlanan yolsuzluğun birçok ekonomik sonucu bulunmaktadır. Bu çalışmada, 11 AB geçiş ekonomisinde 2003-2015 döneminde yolsuzluk ve vergi gelirleri arasındaki ilişki Westerlund ve Edgerton (2007) panel eşbütünleşme testi ve Emirmahmutoglu ve Kose (2011) panel nedensellik testi kullanılarak analiz edilmiştir. Bu çalışmanın sonuçları yolsuzluk ve vergi gelirleri arasında uzun dönemli bir ilişki olduğunu ve bu iki değişken arasında çift yönlü bir nedensellik ilişkisi olduğunu göstermektedir.

Anahtar Kelimeler: Yolsuzluk, Vergi Gelirleri, Panel Veri

Introduction

Corruption is defined as the abuse of public force or resources for private benefits (World Bank, 1997). With globalization, corruption has become a common problem worldwide. Corruption is more commonly observed in developing and transition economies than in developed economies. Corruption has serious repercussions for economic life. In particular, it negatively affects tax revenues, investments and savings, income distribution, the informal economy, foreign direct investments, efficiency of public expenditures, and economic growth and development.

The role of corruption in determining tax revenues has been studied since the 1990s. In recent years, much attention has been paid to the development of anticorruption strategies. To this end, the world's leading organizations, such as Transparency International, the World Bank, and the World Trade Organization, have begun to develop regulations to prevent corruption. Determining the effect of corruption on tax revenues is important to prevent corruption and facilitate efficient taxation.

In this context, panel data techniques are used in this study to determine the effect of corruption on tax revenues. Several studies have been carried out to determine the relationship between corruption and tax revenues, and in general, the studies found that corruption negatively affects tax revenues (see also Johnson et al., 1998; Tanzi & Davoodi, 2000; Imam & Jacobs, 2007; Brasoveanu & Brasoveanu, 2009; Hunady & Orviska, 2015; Ozekicioglu & Bayar, 2017; Epaphra & Massawe, 2017).

1. Theoretical and Empirical Literature Review

One of the most important costs that corruption imposes on economic life is the reduction in tax revenues. Tax revenues, which are the main source of public expenditures, decrease due to corruption and change the course of expenditures that should be made by public administration. A change in the course of public expenditure leads to a reduction in social welfare, and also leads to decreased efficiency and diversity in the goods and services offered by the state (Bakırtas, 2012, p.91).

The field in which public finances are most affected by corruption is the accrual and collection of taxes. Corruption has a significant impact on tax accruals and the collection process. Reductions in taxes are most commonly experienced in tax offices. When this takes the form of tax avoidance, tax evasion, improper tax exemption, and exception depending on corruption, income losses to the state become inevitable (Gedikli, 2011, p.180). In addition, corruption affects tax-free income, with the exception of tax income. This causes reductions in fees, betterments, tax-like incomes, real estate, and entity incomes. However, this leads to an increase in borrowing (Bagdigen and Dokmen, 2006, p.66).

The negative impact of corruption on tax revenues is generally evaluated in two aspects. Taxes are the main source of income for public finance. Depreciation in tax revenues as a result of corruption decreases the income elasticity by causing public borrowing. Additionally, corruption negatively affects the tax structure. Tax revenues decreasing as a result of corruption lead to additional taxes and an increase in tax rates and put pressure on taxpayers to fulfill their tax duties regularly (Asher, 2001, 2).

Corruptions can affect tax revenues indirectly, as follows (Bagdigen and Dokmen, 2006, 63):

- Corruption causes the tax base to reduce, and therefore, tax revenues decrease by increasing informality.
- The fact that corruption reduces investments and negatively influences economic growth and causes tax revenues to shrink.
- Taxpayers' unwillingness to satisfy the illegal demands of illegitimate public officials causes tax revenues to be reduced by leading taxpayers to transfer their financial activities to informal activities or to finish their activities.

The degree of tax revenue problems also varies according to the tax type. Each tax has different characteristics in terms of tax objects, tax causing events, imposition, accruals, and collection processes. For instance, corruption influences direct and indirect taxes in different ways. Frequent taxpayers' and tax authorities' confrontation with direct taxes increases the probability of corruption. However, taxpayers and tax offices do not frequently confront each other because indirect taxes are collected on sales. In addition, the auditing process works effectively because indirect taxes require an effective accounting registry and taxpayers consist of a few large companies. Therefore, indirect taxes are less affected by corruption than indirect taxes. However, both indirect and direct taxes decline in economies in which corruption is common (Dokmen, 2012, 44).

The fact that corruption due to the structure of the tax would both reduce tax revenues and cause inequity in the distribution of tax burden was also indicated in literature. For instance, Tanzi, in a study in 1998, analyzed the effect of corruption on tax revenues theoretically. As a result, he states that corruption has a negative effect on tax revenues. He stated that the devastating effect of corruption on tax revenues would decrease through a transparent taxation system and structural reforms. Bagdigen and Dokmen, in their study in 2006, aimed to theoretically analyze the relationship between corruption and public incomes and expenditures. Their study found that corruption may negatively affect public income.

2. Literature

Johnson et al. (1998) researched the relationship between corruption and tax revenues of 49 countries in three different parts of the world in the 1990's by using regression analysis. As a result, they identified a negative relationship between corruption and tax revenue.

Ghura (1998) attempted to identify the effects of the economic policies and corruption levels of 39 African countries between 1985 and 1996 on tax revenues and Gross National Product (GNP) rates by using panel data analysis. The results revealed that an increase in corruption rates decreased tax revenues and the proportion of tax revenues to GNP.

The study conducted by Fisman and Svensson (2000) aimed to investigate the impact of corruption and taxation on the growth rates of companies operating in various sectors in Uganda during the period of 1995-1997, utilizing time series analysis. The results of the study indicated that both corruption and taxation had a detrimental effect on the growth rate of the companies. Furthermore, the authors suggested that the negative impact of corruption on companies could potentially lead to a decrease in tax revenues.

Tanzi and Davoodi (2000) attempted to identify the effect of corruption levels in 97 countries between 1980 and 1997 on tax and tax-free revenues using regression analysis. The results show that corruption has a negative effect on tax revenue. In addition, they reveal that corruption and direct taxes in tax revenues have more negative effects than indirect taxes.

Imam and Jacobs (2007) conducted a study to examine the relationship between corruption and tax revenues in 12 Middle Eastern countries from 1990 to 2003, utilizing the Generalized Method of Moments (GMM) approach. Their analysis revealed that trade and individual taxes had a more pronounced negative impact on corruption levels.

Attila (2008) researched the relationship between corruption, economic growth and taxation based on the data between 1980 and 2002 in 90 countries using the GMM method. As a result of the study, he found out that corruption may affect economic growth positively under some conditions but may affect negatively through taxes. In addition, he states that corruption may lead to excessive tax rates that could be harmful to growth.

Brasoveanu and Brasoveanu (2009) aimed to identify the relationship between corruption and tax revenues in 27 European Union countries in 1995-2008 by using panel data analysis. As a result, they presented a negative relationship between corruption and tax revenues.

Ajaz and Ahmad (2010) tried to identify the effect of corporate and structural variables (corruption and governance) of 25 developing countries between 1990 and 2005 on tax revenues by using panel data set. They find that governance and corruption are the two main determinants of tax revenue. The findings suggest that corruption had a negative and significant effect on tax revenues, but governance had a positive and significant effect on tax revenues.

The study conducted by Potanlar et al. (2010) sought to investigate the relationship between corruption and tax revenues in 27 developing countries over the period from 2002 to 2006, employing panel data analysis. The findings of the research demonstrated a negative association between corruption and tax revenue.

Monteiro et al. (2011) tried to identify the relationship between corruption and corporate tax in 27 EU countries between 1998 and 2009 employing the least squares method. The analysis reveals that corruption has a decreasing effect on corporate tax revenues.

The research conducted by Hunady and Orviska (2015) aimed to examine the relationship between corruption and the total tax revenues of 46 OECD and Latin American countries between 1998 and 2013, utilizing panel data analysis. The results of the study indicated that corruption had a significant negative effect on the total tax revenues of the countries in question.

Ozekicioglu and Bayar (2017) attempted to identify the effect of corruption and public administration in 35 OECD countries between 2002 and 2015 on tax revenues using panel data analysis. As a result, they determined that improvements in corruption, government efficiency, legislation regulations, and the supremacy of law would have a positive effect on tax revenues.

In their study, Epaphra and Massawe (2017) employed panel data analysis to investigate the influence of corporate variables, including corruption and governance, on tax revenues across 30 African countries from 1996 to 2016. The findings of the analysis revealed that corruption adversely affected tax revenues, whereas good governance, efficient government, the rule of law, and regulatory accountability measures contributed to an increase in tax revenues. Furthermore, the study established that governance had a substantial impact on tax revenues.

3. Data

This study analyzes the effect of corruption on tax revenues in 11 EU transition economies—Bulgaria, the Czech Republic, Estonia, Croatia, Latvia, Lithuania, Hungary, Poland, Romania, Slovakia, and Slovenia— using annual data from 2003 to 2015. In this study, we use tax revenue as the dependent variable, corruption as the independent variable, and economic growth as the control variable. Table 1 presents the variables used in this study and their sources.

Table 1: Data Set

Variables	**Symbol**	**Source**
Corruption	CORRUP	World Bank (2018a)
Tax Revenues (%)	TAXR	World Bank (2018b)
Economic Growth (%)	GRW	World Bank (2018c)

In this study, the long-run relationship and causality relationship between corruption and tax revenues are analyzed using panel data techniques. In the following section we describe these methods.

4. Econometrical Methodology and Empirical Results

4.1. Cross-Sectional Dependence Tests

Cross-sectional dependence is generally identified as an effect of an economic shock occurring in another country on one country. In the case of cross-sectional dependence among variables, tests that consider cross-sectional dependence must be used to obtain reliable results. Therefore, cross-sectional dependence in the series

must be tested first (Pesaran, 2004).

The first study on cross-sectional dependence is the CDLM test by Breusch and Pagan (1980). Pesaran's (2004) Cross-Sectional Dependence (CD) test, and Pesaran et al. (2008) LM test (LM_{adj}) was used in conjunction with a CD test.

The CDLM test statistic proposed by Breusch and Pagan (1980) is computed as follows:

$$CDLM = T\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}\check{\rho}_{ij}^{2} \quad x^{2}\frac{N(N-1)}{2} \tag{1}$$

Pesaran's (2004) CD_{LM} test statistics can be obtained as the following;

$$CD = \sqrt{\frac{1}{n(n-1)}}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}(T\check{\rho}_{ij}^{2} = \pi r^{2} - 1) \tag{2}$$

The adjusted LM test statistic developed by Peseran et al. (2008) is as follows.

$$LM_{adj} = \sqrt{\frac{2}{N(N-1)}\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}\frac{(T-k)\hat{\rho}_{ij}^{2} - \mu_{Tij}}{v_{Tij}}} \tag{3}$$

By using the CD tests, we test the null hypothesis of no cross-sectional dependence against the alternative of cross-sectional dependency among the panel members.

If the time dimension of cross-sectional dependence is higher than the cross-section dimension (*T>N*), the Breusch-Pagan (1980) CDLM1 test is applied. Since the cross-sectional dimension (N=11) of the dataset in this study is lower than the time dimension (*T=13*), cross-sectional dependence between the variables was tested using Breusch-Pagan's (1980) CDLM1 test.

We apply several cross-section dependency tests and report the results in Table 2. As the results show, the null hypothesis of no cross-sectional dependence is rejected at the 10% level; thus, it is determined that a cross-sectional dependence exists between the series.

Table 2: Results of CD Tests

Variable	CD_{LM1}	CD_{LM2}	CD_{LM}	CD_{adj}
GRW	99.754 (0.000)	4.267 (0.000)	-1.934 (0.027)	-0.064 (0.526)
TAXR	244.991 (0.000)	18.115 (0.000)	-1.599 (0.055)	6.892 (0.000)
CORRUP	172.055 (0.000)	11.161 (0.000)	-1.393 (0.082)	2.388 (0.008)

4.2. Panel CIPS Unit Root Test

The panel CIPS unit root test which is developed by Peseran (2007) considers the cross-sectional dependency among the panel members.

The equation of the Panel CIPS test is presented as following:

$$y_{it} = (1-\phi_i)\mu_i + \beta_i y_{i,t-1} + u_{it}, i = 1, \dots., N; t = 1, \dots, T \quad u_{it} = \Upsilon_i f_t + \varepsilon_{it} \tag{4}$$

The null and alternative hypotheses of the panel CIPS unit root test are presented as follows (Pesaran, 2007, 268);

$$H_0 : \beta_i = 0 \qquad \text{(The series has a unit root.)} \tag{5}$$

$$H_1 : \beta_i < 0, i = 1,2, \dots, N_1, \beta_1 = 0, i = N_1 + 1, N_1 + 2, \dots, N \quad \text{(The series is stationary.)} \tag{6}$$

The CIPS test statistics can be computed by using the average of test statistics value (for β);

$$CIPS(N,T) = t - bar = N^{-1} \sum_{i=1}^{N} t_i(N,T) \quad (7)$$

Because the study confirmed the cross-sectional dependence between the series, we next employed the CIPS unit root test. The results, shown in Table 3, indicate that each variable is stationary in different ways.

Table 3. Panel CIPS Unit Root Test Results

LEVEL	**CORRUP**	**GRW**	**TAXR**
Constant	-1.413	-2.207	-1.914
Constant + Trend	-1.887	-2.545	-1.940
FIRST DIFFERENCE	**CORRUP**	**GRW**	**TAXR**
Constant	**-2.866*** **	**-2.941*** **	**-3.062*** **
Constant + Trend	**-2.415**	**-2.784***	**-2.803***

Note: *, **, and *** indicate the significance at the 10%, 5%, and 1% significance levels, respectively. Critical values were constant -2.97 (1%), -2.52 (5%), -2.31 (10%); constant + trend is -3.88 (1%), -3.27 (5%), -2.98 (10%). The critical values were obtained from Pesaran (2007).

4.3. Westerlund and Edgerton (2007) Cointegration Test

Prior to applying Westerlund and Edgerton's (2007) cointegration test, a homogeneity test should be used. The homogeneity test is employed to identify whether the cointegration slope coefficients are homogenous or heterogeneous (Pesaran and Yamagata, 2008, 56).

According to Pesaran and Yamagata (2008), the delta test is calculated as follows:

$$\tilde{\Delta} = \sqrt{N\left(\frac{N\hat{}(-1)S^{-}-k}{2k}\right)} \sim X_{k\hat{}2} \quad (8)$$

$$\tilde{\Delta}_{\text{adj}} = \sqrt{N\left(\frac{N\hat{}(-1)S^{-}-k}{\text{v}(T,k)}\right)} \sim N(0,1) \quad (9)$$

where N indicates the number of cross-sections, S, k, and v(T,k) indicate Swamy test statistic, number of explanatory variables, and the standard error. The null hypothesis of slope coefficients are homogeneous is tested against the alternative of slope coefficients are not heterogeneous.

The outcomes of the homogeneity tests are displayed in Table 4. Since the probability values of these tests, as calculated in Table 4, are higher than 0.10, the null hypothesis is not rejected. Consequently, the results support the evidence that the slope coefficients in the cointegration equations are homogeneous. So, to test the long-run relationship between the series, this homogeneity should be considered.

Table 4: Homogeneity Test Results

	Test Statistics	**Probability Value**
$\tilde{\Delta}$	0.517	0.302
$\tilde{\Delta}_{\text{adj}}$	0.612	0.270

Following the homogeneity test, we employed the panel bootstrap cointegration test suggested by Westerlund and Edgerton (2007) to investigate the cointegration relationship. This cointegration test, based on the Lagrange multiplier test suggested by McCoskey and Kao (1998), takes into account the dependence among cross-sectional units. Notably, Westerlund and Edgerton's (2007) cointegration test has been shown to produce reliable results in small samples (Westerlund and Edgerton, 2007, 185-190).

Westerlund (2007) cointegration test include four new panel cointegration tests based on structural dynamics. The first two tests in the panel cointegration tests indicate group-average statistics. The last two tests indicate panel statistics. Panel statistics are computed by coupling information about error correction in the cross-sectional dimension of the panel together. However, this information is not used in group-average statistics. The difference between the tests is based on alternative hypothesis tests (Westerlund, 2007, 710-712).

While H_0^p: $\alpha_i = 0$ for the null hypothesis and H_A^p: $\alpha_i = \alpha < 0$ for the alternative hypothesis is used in panel statistics, H_A^β: $\alpha_i = 0$ null hypothesis and H_A^β: $\alpha_i = \alpha < 0$ alternative hypotheses are used in group-average statistics. The rejection of the null hypothesis indicates that cointegration exists; however, the acceptance of the null hypothesis indicates that cointegration does not exist. It is suggested that group and panel tests conducted by Westerlund obtained very steady results if the cross-sectional dependence exists (Westerlund, 2007, 721-722).

Because all variables are *I*(1), Westerlund and Edgerton's (2007) cointegration test is used to identify the long-term relationship between variables. The null hypothesis of the existence of a cointegration relationship between the variables H_0: $\sigma_i^2 = 0$ (for all values of i), is tested against the alternative hypothesis, H_A: $\sigma_i^2 > 0$ (for some values of *i*), which proposes that there is no cointegration (Westerlund and Edgerton, 2007, 185-186). Table 5 presents the results.

Table 5: Cointegration Test Results

Model	LM- Statistics	Bootstrap Probability Value
Constant Model	14.740	0.427
Constant + Trend Model	9.784	0.553

Note: Bootstrap p-values were obtained from 10,000 simulations.

We cannot reject the null hypothesis of cointegration according to the findings in Table 5, that is, there is a long-run relationship between corruption (CORRUP), economic growth (GRW), and tax revenues (TAXR) since the bootstrap p-values are higher than the traditional significance levels.

4.4. FMOLS

In this study, to estimate the long-run coefficients we employ fully modified ordinary least squares method (FMOLS) which corrects the deviations in estimators (consisting of problems such as changing variance and autocorrelation). According to Pedroni (2000), the power of the FMOLS method in small samples is good (Kok and Simsek, 2006, 7-8).

After determining the cointegration relationship, we estimate the panel FMOLS to identify the degree of long-term relationships among the variables. Table 6 shows the estimation results.

Table 6: Long-run Coefficients

Variable	Coefficient	Std. Error	T-Statistic	Prob.
GRW	**0.111099*** **	0.028507	3.897202	0.0002
CORRUP	**-1.979744** **	0.907551	2.181415	0.0310

Note: ***, **, and * indicate statistical significance at the 1%, 5%, and 10% levels, respectively.

The results shown in Table 6 indicate that a one-unit increase in corruption causes tax revenues to decrease by 1.97 units. This result may be due to an increase in tax losses due to corruption. Tax losses cause the tax base to shrink, tax revenues to decrease, and the tax burden to be unfairly distributed. These negative effects also negatively affect social welfare by changing the state of public expenditures. For this reason, countries must consider the effects of corruption if they follow policies to increase tax revenues. In addition, states can positively contribute to the struggle against corruption by devising a tax system suited to the economic conditions of their own countries.

However, economic growth positively affects tax revenues. In other words, a one-unit increase in economic growth leads to a 0.11 unit increase in tax revenues. This increase is attributable to the expansion of the tax base because of an increase in economic growth, production, and consumption. The expansion of the tax base positively affects tax revenues.

4.5. Emirmahmutoglu and Kose (2011) Causality Test

Emirmahmutoglu and Kose (2011) developed a panel causality test which is based on Toda-Yamamoto causalit test and considers cross-sectional dependence and allows the examination of causality relationship for the panel members individually.

Emirmahmutoglu and Kose (2011) considered following VAR model to apply the panel causality test:

$$x_{i,t} = \mu_i^x + \sum_{j=1}^{k_i+d\ max_i} A_{11,ij}\, x_{i,t-j} + \sum_{j=1}^{k_i+d\ max_i} A_{12,ij}\, y_{i,t-j} + u_{i,t}^x \qquad (10)$$

$$y_{i,t} = \mu_i^y + \sum_{j=1}^{k_i+d\ max_i} A_{21,ij}\, x_{i,t-j} + \sum_{j=1}^{k_i+d\ max_i} A_{22,ij}\, y_{i,t-j} + u_{i,t}^y \qquad (11)$$

Where $dmax_i$, indicates the maximum integration level of variables. Tables 7 and 8 present the Emirmahmutoglu and Kose (2011) test results, respectively.

Table 7: Panel Causality Test Results

Countries	CORRUP-TAXR		TAXR-CORRUP	
	Test Statistics	**Prob. Value**	**Test Statistics**	**Prob. Value**
Bulgaria	**4.225****	0.040	0.141	0.708
Croatia	**4.419****	0.036	**6.180****	0.013
Czech Republic	0.423	0.515	1.538	0.215
Estonia	**3.410***	0.065	0.053	0.819
Hungary	0.943	0.332	0.000	0.994
Latvia	0.122	0.727	1.632	0.201
Lithuania	0.000	0.983	1.666	0.197
Poland	**7.551*****	0.006	**8.727*****	0.003
Romania	0.089	0.765	1.075	0.300
Slovakia	0.740	0.390	1.049	0.306
Slovenia	0.180	0.742	**6.719****	0.010
Panel	**36.050****	0.030	**44.946*****	0.003

Note: ***, **, and * indicate statistical significance at the 1%, 5%, and 10% levels, respectively.

When the test results shown in Table 7 are analyzed, we can identify a significant and bidirectional causality relationship between corruption and tax revenues, according to the panel test results. In other words, a change in corruption may lead to a change in tax revenue. Similarly, a change in tax revenue may lead to a change in

corruption. It can be thought that the fact that an increase in corruption in a country may affect tax revenues negatively may be the reason for it. Alternatively, starting to obtain more tax revenues in countries will increase the tax burden; this is thought to cause an increase in corruption by directing individuals to some behaviors such as tax evasion, tax refusal, bribery, and so on.

Table 8: Panel Causality Test Results

Countries	GRW-TAXR		TAXR-GRW	
	Test Statistics	**Prob. Value**	**Test Statistics**	**Prob. Value**
Bulgaria	**4.559****	0.033	**3.926****	0.048
Croatia	0.208	0.648	**3.186***	0.074
Czech Republic	0.284	0.594	0.116	0.734
Estonia	0.679	0.410	0.002	0.963
Hungary	**3.437***	0.064	0.023	0.878
Latvia	0.040	0.842	0.306	0.580
Lithuania	0.543	0.461	1.292	0.256
Poland	0.796	0.372	**5.888****	0.015
Romania	**2.769***	0.096	1.279	0.258
Slovakia	0.335	0.563	2.365	0.124
Slovenia	**3.485***	0.062	1.428	0.232
Panel	**31.303***	0.090	**34.235****	0.046

Note: ***, **, and * indicate statistical significance at the 1%, 5%, and 10% levels, respectively.

The findings presented in Table 8 indicate a statistically significant bidirectional causality relationship between economic growth and tax revenue, based on the panel average which suggests that changes in economic growth may result in changes in tax revenues, and vice versa. Similarly, a change in tax revenue may lead to a change in economic growth. It can be thought that the fact that an increase in economic growth in a country may affect tax revenues positively may be the reason for it.

Conclusion

Among the top issues surrounding corruption is its effect on tax revenue. This study empirically analyzes the implications of such an effect on EU transition economies. The analysis identifies a cointegration relationship between corruption and tax revenue. In other words, a long-term relationship is detected between the variables. In addition, according to Emirmahmutoglu and Kose's (2011) causality test, a significant and bidirectional causality relationship between corruption and tax revenues is identified.

When the results obtained from the study are generally evaluated, it can be seen that an increase in corruption in a country negatively affects the country's tax revenues. This negative effect on tax revenues leads to an unfair distribution of the tax burden among the population. In addition, the loss of revenue resulting from corruption causes the tax base to shrink, with a corresponding reduction in public expenditure.

Decreased tax revenues resulting from corruption lead to surtaxes or expensive public services by causing an increase in tax rates. This paves the way for increased activity in the informal economy and taxpayer resistance to tax. In addition, it causes public debt to increase public expenditure. These effects influence economic growth and employment negatively by causing fiscal gaps in the public sector. Therefore, a struggle against corruption is important for the development of countries.

As corruption is a phenomenon that is impossible to complete completely, it can be said that taking precautions to reduce corruption to a reasonable level is important for countries. The precautions to be taken are as follows.

- Education level is highly important for reducing corruption. Corruption is experienced more frequently in countries with low education levels. For this reason, it is highly important for countries to determine

policies to increase educational levels in the struggle against corruption. However, an increase in education level reduces negative behaviors. Therefore, it can be said that rules to prevent negative behaviors would reduce corruption.

- NGOs, an important factor in reducing corruption, play a role in leading related issues and laws to wide ranges. NGOs are important for a healthy, sensitive, and strong society. It can be stated that corruption can be reduced by increasing the organizational awareness of countries through NGOs.
- Telling a lie and cheating underlying corruption may cause individuals' loss of trust in the state and the state to lose power by affecting the social system. In addition, the desire to earn money quickly and easily may increase the possibility of experiencing corruption because of financial inequality among individuals. For this reason, it may be important to reinforce public awareness of society in order to ensure fair salary levels and income distribution in the struggle against corruption.
- Corruptions cause significant fiscal problems for the public by negatively affecting tax revenue. Corruption must be considered if a tax revenue-increasing policy is followed. The fact that states determine a tax system that is suitable for their own conditions and economic conjuncture may positively contribute to the process of the struggle against corruption.
- Owing to the negative effects of corruption on the accrual and collection of taxes, reductions may occur in tax bases. To prevent these reductions, the negative effects of corruption can be minimized by applying restrictions through taxes in fields that lead to corruption. In addition, the tax revenues of countries can be increased by reducing corruption by increasing the number of tax auditing personnel working in the taxation system.
- Equivalent and fair distribution of the tax burden is important in the struggle against corruption. Tax burden is a phenomenon that affects the attitudes and behaviors of citizens towards taxes. A high tax burden causes citizens to act negatively by leading activities such as tax evasion, smuggling, tax refusal, and an informal economy. Therefore, these negative actions may lead to an increase in corruption. In this sense, an increase in corruption may lead to a reduction in tax revenue. Benefiting from the decision-making characteristics of the tax burden in tax policies, it can be utilized in the struggle against corruption.

Inflation, which is one of the reasons for corruption, negatively affects the economies of countries in many ways. Because inflation may lead individuals to take illegal actions by affecting their purchasing power, corruption may lead to an increase in informality and unethical actions but a decrease in tax revenues. Therefore, inflation may be reduced to a reasonable level in the struggle against corruption. Therefore, corruption can be reduced by preventing individuals from engaging in illegal action.

References

Ajaz, T., Ahmad, E. (2010). The effect of corruption and governance on tax revenues. *The Pakistan Development Review*, 49(4), 405-417.

Asher, M.G. (2001). Design of tax systems and corruption. *In Conference On "Fighting Corruption: Common Challenges and Shared Experiences"*, 1-19.

Attila, G. (2008). Corruption, taxaion and economic growth: theory and evidence. *Cerdi, Etudes et Documents.*

Bagdigen, M., Beskaya, A. (2005). The impact of corruption on government revenues: The turkish case. *Yapı Kredi Review*, 16(2), 31-54.

Bagdigen, M., Dokmen, G. (2006). Yolsuzluğun kamu gelir ve giderleri üzerine etkisi. *ZKÜ Sosyal Bilimler Dergisi*, 2(3), 53-69.

Bakırtas, D. (2012). Yolsuzluğun vergi gelirleri üzerindeki etkisi: Türkiye örneği. *Yönetim ve Ekonomi Dergisi*, 19(2), 87-98.

Brasoveanu, I.V., Brasoveanu, L.O. (2009). Correlation between corruption and tax revenues in EU 27. *Economic Computation and Economic Cybernetics Studies and Research*, 43(4), 133-142.

Breitung, J. (2005). A parametric approach to the estimation of cointegration vectors in panel data. *Econometric Reviews*, 24(2), 151-173.

Breusch, T.S., Pagan, A.R. (1980). The lagrange multiplier test and ıts applications to model specification in econometrics. *The Review of Economic Studies*, 47, 239-253.

Dokmen, G. (2012). Yolsuzlukların vergi gelirleri üzerindeki etkisi: Dinamik panel veri analizi. *Doğuş Üniversitesi Dergisi*, 13(1), 41-51.

Emirmahmutoglu, F., Kose, N. (2011). Testing for granger causality in heterogeneous mixed panels. *Economic Modelling*, 28, 870-876.

Epaphra, M., Massawe, J. (2017). Corruption, governance and tax revenues in Afrika. *Business and Economic Horizons*, 1(4), 439-467.

Fisman, R., Jakob, S. (2000). Are corruption and taxation really harmful to growth? firm –level evidence. *The World Bank Policy Research Working Paper*, 2485, 1-24.

Gedikli, A. (2011). Kamu hizmetlerinin yönetimi sürecinde yolsuzluğun derinleştirdiği ekonomik büyüme ve yoksulluk sorunu üzerine bir değerlendirme. *Öneri*, 9(36), 169-188.

Ghura, D. (1998). Tax revenue in Sub-saharan Africa: Effects of economic policies and corruption. *IMF Working Paper*, WP/98/135, 1-25.

Hunady, J., Orviska, M. (2015). The effect of corruption on tax revenue in OECD and Latin America countries. *The Oretical and Practical Aspects of Public Finance*, 20, 1-6.

Imam, P.A., Jacobs, D.F. (2007). Effect of corruption on tax revenues in the Middle East. *IMF Working Paper*, WP/07/270, 1-34.

Johnson, S., Kaufman, D., Lobaton, P.Z. (1998). Regulatory discretion and the unofficial economy. *American Economic Review*, 88(2), 387-392.

Kao, C., Chiang, M.H. (2000). On the estimation and inference of a cointegrated regression in panel data B. H. BALTAGI, T. B. FOMBY and R. C. HILLS (der), nonstationary panels, panel cointegration and dynamic panel advances in econometric. *Elsevier Science***,** Amsterdam, 179-222.

Kok, R., Simsek, N. (2006). Endüstri-içi dış ticaret, patentler ve uluslararası teknolojik yayılma. *Türkiye Ekonomi Kurumu Uluslararası Ekonomi Konferansı*, 11-13.

Kok, R., Ispir, M.S., Arı, A.A. (2010). Zengin ülkelerden azgelişmiş ülkelere kaynak aktarma mekanizmasının gerekliliği ve evrensel bölüşüm parametresi üzerine bir deneme. *Uluslararası Ekonomi Konferansı*, Türkiye Ekonomi Kurumu, Kıbrıs.

McCoskey, S., Kao, C. (1998). A residual-based test ff the null of cointegration in panel data. *Econometric Reviews*, 17(1), 57–84.

Monteiro, M.R., Brandao, E.F.M., Martins, F.V.S. (2011). A panel data econometric study of corporate tax revenue in European Union: structural, cyclical business and institutional determinants. *FEB Working Papers*, 437, 1-36.

Ozekicioglu, S., Bayar, Y. (2017). Tax revenues, corruption and governance in OECD countries: A panel regression analysis. *The Journal of Economic Sciences: Theory and Practice*, 74(2), 51-63.

Pedroni, P. (2000). Fully-modified ols for heterogeneous cointegrated panels. *Advances in Econometrics*, 15, 93-130.

Pesaran, M.H. (2004). General diagnostic tests for cross section dependence in panels. *Cesifo Working Paper*, 1229, 1-40.

Pesaran, M.H. (2007). A simple panel unit root test in the presence of cross section dependence. *Journal of Applied Econometrics*, 22(2), 265-312.

Pesaran, M.H., Ullah, A., Yamagata, T. (2008). A bias-bdjusted LM test of error cross-section independence. *Econometrics Journal*, 11(1), 105-127.

Pesaran, M.H., Yamagata, T. (2008). Testing slope homogeneity in large panels. *Journal of Econometrics*, 142(1), 50-93.

Potanlar, S.K., Samimi, A.J., Roshan, A.R. (2010). Corruption and tax revenues: New evidence from some developing countries. *Australian Journal of Basic and Applied Sciences*, 4(9), 4218-4222.

Tanzi, V. (1998). Corruption around the world; causes, consequences, scope, and cures. *IMF Staff Paper*, 45(4), 559-594.

Tanzi, V., Davoodi, H.R. (2000). Corruption, growth and public finances. *IMF Working Paper*, 182, 1-27.

Toda, H.Y., Yamamoto, T. (1995). Statistical inferences in vector autoregressions with possibly integrated processes. *Journal of Econometrics*, 66, 225-250.

Vergil, H., Ayas, N. (2009). Doğrudan yabancı yatırımların istihdam üzerindeki etkileri: Türkiye örneği. *Iktisat, Isletme ve Finans*, 24(275), 89-114.

Westerlund, J. (2007). Testing for error correction in panel data. *Oxford Bulletin of Economics and Statistics*, 69(6), 709-748.

Westerlund, J., Edgerton, E. (2007). *A* panel bootstrap cointegration test. *Economics Letters*, 97(3), 185-190.

World Bank. (1997). World development report 1997: The state in a changing world. *New York: Oxford University Press.* http://hdl.handle.net/10986/5980.

World Bank. (2018a). Control of Corruption, http://info.worldbank.org/governance/wgi/#home.

World Bank. (2018b). Tax Revenue (% of GDP), https://data.worldbank.org/indicator/GC.TAX.TOTL.GD.ZS?view=chart.

World Bank. (2018c). GDP Growth (Annual %), https://data.worldbank.org/indicator/NY.GDP.MKTP.KD.ZG?view=chart.