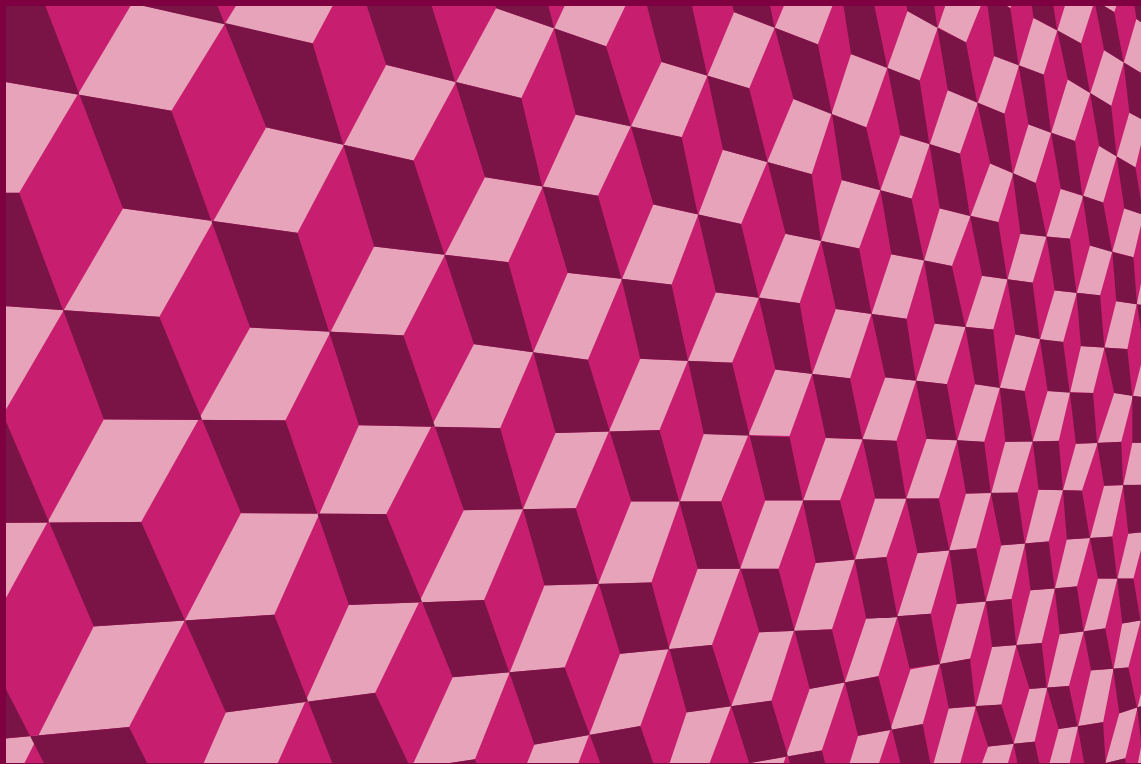




İSTATİSTİK ARAŞTIRMA DERGİSİ Journal of Statistical Research

Cilt-Volume: 07 Sayı-Number: 02
Aralık-December 2010

ISSN 1303-6319



TÜRKİYE İSTATİSTİK KURUMU
Turkish Statistical Institute



İSTATİSTİK ARAŞTIRMA DERGİSİ

Journal of Statistical Research

Cilt-Volume: 07 Sayı-Number: 02
Aralık-December 2010

Yayın istekleri için For publication order

Döner Sermaye İşletmesi Revolving Fund Management

Tel: + (312) 425 34 23 - 410 05 96 - 410 02 85

Fax: + (312) 417 58 86

Yayın içeriğine yönelik sorularınız için For questions about contents of the publication

Dergi Editörlüğü Journal Editorship

Tel: + (312) 410 03 67 - 284 45 00/171

Fax: + (312) 425 34 05

İnternet Internet
http://www.tuik.gov.tr http://www.turkstat.gov.tr

E-posta E-mail
dergi@tuik.gov.tr journal@tuik.gov.tr

Yayın No Publication Number
3507

ISSN
1303-6319

Türkiye İstatistik Kurumu Turkish Statistical Institute

Yücetepe Mah. Necatibey Cad. No: 114 06100 Çankaya-ANKARA / TÜRKİYE

Bu yayının 5846 Sayılı Fikir ve Sanat Eserleri Kanunu'na göre her hakkı Türkiye İstatistik Kurumu Başkanlığına aittir. Gerçek veya tüzel kişiler tarafından izinsiz çoğaltılamaz ve dağıtılamaz.

Turkish Statistical Institute reserves all the rights of this publication. Unauthorised duplication and distribution of this publication is prohibited under Law No: 5846.

Türkiye İstatistik Kurumu Matbaası, Ankara Turkish Statistical Institute, Printing Division, Ankara

Tel: 0312 410 01 64 * Fax: 0312 418 50 82

Şubat 2011 February 2011

MTB: 2011-089 - 475 Adet-Copies

Editör'den

Değerli okuyucular,

İstatistik Araştırma Dergisi'nin yedinci cildi ikinci sayısını tamamlamış olmanın mutluluğunu yaşamaktayız. Dergimizde yer alan makalelerin nitelik ve nicelik olarak arttığı görülmektedir. Bu sayı ile birlikte dergimizdeki çalışmaların bundan böyle daha geniş kitlelere ulaşması hedeflenmektedir.

Dergimize olan talebin artan bir eğilim sürecinde olması nedeniyle, hatırlayacağınız gibi, 2007 yılı sayısında başlatılan yılda bir sayı yayımlanması uygulamasına son verilerek, 2010 yılından itibaren yılda iki sayı yayımlanma geleneğine dönmüştür. Akademisyen, araştırmacı ve okuyucuların artan ilgisine paralel olarak bizlerin çabası, azmi ve kararlılığı da artacak olup, dergimiz daha üst seviyelere taşınacaktır.

Dergimizin ulusal ve uluslararası endekslerde taranması çalışmaları da devam etmektedir. Bu konuya ilişkin olarak alınacak sonuçlar sizlerle paylaşılacaktır.

Bu sayımızda, kavramsal, kuramsal ve uygulamalı çalışmalar olmak üzere toplam dokuz adet çalışmayı siz değerli okuyucularımızla paylaşmanın gururunu taşıyoruz. Bu değerli çalışmaları, bizlerle ve siz değerli okuyucularımız ile paylaşan sayın yazarlara teşekkürü bir borç biliriz. Çalışmaların daha nitelikli hale gelmesinde çok değerli öneri, eleştiri ve katkılarını esirgemeyen sayın hakemlere de şükranlarımızı sunuyoruz.

Dergi'nin basım aşamasına gelmesinde emeğini ve desteklerini esirgemeyen TÜİK Başkan Vekili Sayın A. Ömer TOPRAK'a, derginin her aşamasında emeği geçen Editör Yardımcısı Sayın Yrd. Doç. Dr. Özlem İLK'e, dergi çalışmalarını içtenlikle ve azimle yürüten Dergi Sekreteryası'na ve son olarak da emeği geçen diğer tüm TÜİK çalışanlarına teşekkürlerimi iletmek isterim.

Bu sayımızın da akademisyenler ile araştırmacılara faydalı olması temennisi ve gelecek sayılarda hedeflenenler ölçüsünde tekrar buluşmak dileği ile saygılar sunarım.

Prof. Dr. Fetih YILDIRIM
Dergi Editörü

TÜRKİYE İSTATİSTİK KURUMU **TURKISH STATISTICAL INSTITUTE**
İSTATİSTİK ARAŞTIRMA DERGİSİ **JOURNAL OF STATISTICAL RESEARCH**

Sahibi **Owner**

Türkiye İstatistik Kurumu Adına On Behalf of Turkish Statistical Institute
A. Ömer TOPRAK A. Ömer TOPRAK
Türkiye İstatistik Kurumu Başkan Vekili Acting President, Turkish Statistical Institute

Editör **Editor**

Prof. Dr. Fetih YILDIRIM Prof. Dr. Fetih YILDIRIM

Editör Yardımcısı **Assistant Editor**

Yrd. Doç. Dr. Özlem İLK Assoc. Prof. Özlem İLK

Sekreteryaya **Secretariat**

Buket AKGÜN
Mehmet Arif ŞAHİNLİ
Z. Nur EMRE

İÇİNDEKİLER	Sayfa Page	CONTENTS
ÖNSÖZ	III	FOREWORD
İÇİNDEKİLER	VII	CONTENTS
AMAÇ ve KAPSAM	IX	AIM and SCOPE
HAKEM LİSTESİ	XI	REFEREE LIST
Yarıparametrik Regresyon Modellerinde Tahmin Etme <i>Fikri AKDENİZ</i>	1	Estimation in Semiparametric Regression Models <i>Fikri AKDENİZ</i>
Bulanık Sistem Modellerinin Gelişimi <i>İ. Burhan TÜRKŞEN</i>	11	Development of Fuzzy System Models <i>İ. Burhan TÜRKŞEN</i>
Doğrudan ve Dolaylı Mevsim Düzeltilme Yaklaşımları: GSYİH Serileri İçin Bir Uygulama <i>Özlem YİĞİT</i> <i>Bedriye SARAÇOĞLU</i>	25	Direct and Indirect Seasonal Adjustment Approaches: An Application for GDP Series <i>Özlem YİĞİT</i> <i>Bedriye SARAÇOĞLU</i>
Sağa Çarpık Dağılım Ortalamaları İçin Bazı Testlerin Kullanımı ve Karşılaştırmaları <i>Emre Erçin SARISOY</i> <i>Hamza GAMGAM</i>	45	Use of Some Tests for Means of Positively Skewed Distribution and Comparisons <i>Emre Erçin SARISOY</i> <i>Hamza GAMGAM</i>
Gıda Tüketim Analizi: Genişletilmiş Doğrusal Harcama Sistemi <i>Mehmet Arif ŞAHİNLİ</i> <i>Ahmet ÖZÇELİK</i>	58	Food Consumption Analysis: An Extended Linear Expenditure System <i>Mehmet Arif ŞAHİNLİ</i> <i>Ahmet ÖZÇELİK</i>
Kişi Başına Düşen Milli Gelire Göre Ülkelerin Gıda Tüketim Eğilimlerinin Tespit Edilmesi – Veri Madenciliği Uygulaması <i>Fethi Ş. ÖZBEK</i>	66	Determining Countries' Food Consumption Patterns With Respect To National Income Per Capita – A Data Mining Application <i>Fethi Ş. ÖZBEK</i>

- Histogramlarda ve Sıklık Tablolarında Kullanılan Sütun / Sınıf Sayılarının Belirlenmesi: Kısa Bir Bibliyografya**
Nurhan DOĞAN
İsmet DOĞAN 77 **Determination of the Number of Bins / Classes Used in Histograms and Frequency Tables: A Short Bibliography**
Nurhan DOĞAN
İsmet DOĞAN
- Türkiye’de İstatistik Bölümlerinde Okuyan Üniversite Öğrencilerinin Sosyo-Ekonomik Statü (SES) Puanları**
Doğan Yıldız
Atıf Evren 87 **Socioeconomic Status (SES) Scores of Turkish Statistics Students**
Doğan Yıldız
Atıf Evren
- Türkiye’deki İllerin Sosyo-Kültürel Gelişmişlik Derecelerine Göre Değerlendirilmesi**
Esin FİRUZAN
Yusuf Yüksel AYVAZ
Ersen UZUN 101 **Assessment of the Cities in Turkey According to Their Socio-Cultural Development Level**
Esin FİRUZAN
Yusuf Yüksel AYVAZ
Ersen UZUN

AMAÇ VE KAPSAM

“İstatistik Araştırma Dergisi (İAD)”, istatistik araştırmaların niteliğinin yükseltilmesi, istatistik yöntem ve uygulamalarının geliştirilmesi, literatürde yer alan çalışmaların tartışılması, istatistik uygulamalarıyla ilgili anket çalışmalarının ele alınması, kuramsal ve uygulama alanındaki araştırmacılar arasında iletişimin ortak çalışma ve yayınlarla güçlendirilmesi amacıyla, yayımlanan hakemli bir dergidir.

“İstatistik Araştırma Dergisi”nin kapsamında yer alan tematik konular aşağıda özet olarak verilmiştir:

- Bankacılık, Finans, Sigortacılık, Aktüerya ve Risk Yönetimi; Bayesci İstatistik; Benzetim Teknikleri; Bilgi Sistemleri; Biyoistatistik; Bulanık Teori; Demografi; Deney Tasarımı ve Varyans Analizi; Ekonometri; Genel Sayımlar ve Değerlendirmeleri; İstatistik Eğitimi; İstatistik Etiği; İstatistik Kuramı; İstatistiksel Kalite Kontrolü; Kamuoyu ve Piyasa Araştırmaları; Klinik Denemeler; Mühendislikte İstatistik Uygulamaları; Olasılık ve Stokastik Süreçler; Optimizasyon; Örnekleme ve Araştırma Tasarımları; Parametrik Olmayan İstatistiksel Yöntemler; Resmi İstatistikler; Toplum Bilimlerinde İstatistik; Veri Analizi ve Modelleme; Veri Madenciliği; Veri Yönetimi ve Karar Destek Sistemleri; Verimlilikte İstatistiksel Yaklaşımlar; Yönetimsel Süreçlerde Performans Analizi; Yöneylem Araştırması; Zaman Serileri; Diğer İstatistiksel Yöntemler gibi istatistiğin her dalında yeni bilgi üretimine yönelik tüm araştırmalar.

Makale Dili ve Genel Kurallar

- Bu yayının 5846 Sayılı Fikir ve Sanat Eserleri Kanunu’na göre her hakkı Başbakanlık Türkiye İstatistik Kurumu Başkanlığı’na aittir. Gerçek veya tüzel kişiler tarafından izinsiz çoğaltılamaz ve dağıtılamaz.
- Makale taslakları WORD yazım dilinde, Times New Roman yazı tipinde, 12 punto büyüklükte, satırlar arasında bir satır boşluk bırakılarak yazılmalı, şekil ve grafikler JPG dosyaları olarak hazırlanmalıdır.
- A4 sayfa boyutunda; soldan 3,5 cm, sağdan, yukarıdan ve aşağıdan 2,5 cm boşluk bırakılmalıdır.
- Ana bölüm başlıklarının tümü büyük harf, 12 punto büyüklükte, koyu, ortalı ve Arap rakamları ile numaralandırılarak; alt bölüm başlıklarında ise sadece kelimelerin baş harfleri büyük diğerleri küçük harfle, 12 punto büyüklükte, koyu, sola dayalı ve ana bölüm başlığına endekslilikle Arap rakamları ile numaralandırılarak yazılmalıdır.
- Makale taslağı yazımında, okuyucunun, çalışmanın her aşamasını anlama ve değerlendirmesine olanak verecek bir anlatım ve plana uyulmalıdır.
- Anlatım olabildiğince sade, anlaşılabilir, öz ve kısa olmalıdır. Gereksiz tekrarlardan, desteklenmemiş ifadelerden ve konu ile doğrudan ilişkisi olmayan açıklamalardan kaçınılmalıdır.
- Yazımda çok genel ifadeler kullanılmamalıdır. Yargı veya kesinlik içeren ifadeler mutlaka verilere/referanslara dayandırılmalıdır.
- Araştırmacı/araştırmacılar tarafından probleme, hangi kuramsal/kavramsal açıdan yaklaşıldığı, gerekçeleri ile birlikte belirtilmelidir.
- Kullanılan araştırma yönteminin seçilme gerekçesi açıklanmalıdır. Bütün veri toplama araçlarının geçerliliği ve güvenilirliği belirtilmelidir.
- Araştırma sonucunda elde edilen veriler bir bütünlük içinde sunulmalıdır.
- Sadece elde edilen verilere dayanan sonuçlar sunulmalıdır.
- Sonuçların yorumları, varsa, literatürdeki diğer kaynaklarla desteklenerek, değerlendirilmelidir.
- Yararlanılan kaynaklar, çalışmanın kapsamını yansıtacak zenginlik ve yeterlikte olmalıdır.
- Türkçe ve İngilizce özetler; çalışmanın amacı, yöntemi, kapsamı ve temel bulgularını içermelidir.

Ayrıntılı bilgi için, <http://www.tuik.gov.tr> adresinden “İstatistik Araştırma Dergisi Kılavuzu”na bakınız.

AIM AND SCOPE

“*Journal of Statistical Research (JSR)*” is a refereed journal published with the aim to raise the quality of statistical researches, improve the statistical methodology and applications, discuss the studies included in literature, consider survey studies regarding the statistical application, and strengthen the communication between researchers in the field of theory and application by joint studies and publications.

The contents of the “*Journal of Statistical Research*” are summarized below:

- Researches aimed at producing new knowledge in every field of statistics such as Banking, Finance, Insurance Trade, Actuarial and Risk Management; Bayesian Statistics; Biostatistics; Clinic Tests; Data Analysis and Modeling; Data Management and Decision Support Systems; Data Mining; Demography; Econometrics; Experimental Design and Variance Analysis; Fuzzy Theory; General Census and Evaluation; Information Systems; Non-Parametric Statistical Methods; Official Statistics; Operational Research; Optimization; Sampling and Research Designs; Performance Analysis in Managerial Process; Probability and Stochastic Processes; Public Opinion and Market Researches; Statistical Applications in Engineering; Statistical Approaches in Efficiency; Statistical Ethics; Statistical Quality Control; Statistical Training; Statistics in Social Science; Statistics Theory; Simulation Techniques; Time Series; Other Statistical Methods.

Article Language and General Rules

- Prime Ministry, Turkish Statistical Institute reserves all the rights of this publication. Unauthorized duplication and distribution of this publication is prohibited under Law No: 5846.
- Article drafts should be prepared in WORD, using Times New Roman font, in 12 point size, with a blank line in between lines. Figures and tables should be prepared as JPG files.
- On A4 paper size; margins should be set as: left 3,5 cm; right, top and bottom 2,5 cm.
- Titles of the main sections should be all capitalized, in 12 point size, bold, centered and numbered with Arabic numerals; only the first letter of the words in the titles of the subsections should be capitalized, with 12 point size, bold, left justified and numbered with Arabic numerals indexed to the titles of the main sections.
- In article draft writing, writer should follow such a plan that reader should be able to understand and evaluate all the steps of the study.
- Narration should be as plain as possible, as well as comprehensible, compact and short. Unnecessary repetitions, unsupported declarations and explanations that are not in direct relation to the topic should be avoided.
- General statements should be avoided in writing. Statements that include judgment or facts must be supported by data/references.
- It should be stated, with justifications, from which theoretical/conceptual aspect the researcher/researchers have approached the problem.
- The reason of choosing the research methodology that is used should be explained. The validity and reliability of all the data collection tools should be presented.
- Data obtained as the result of the research should be presented in unity.
- Results that only rely on the obtained data should be presented.
- The interpretation of the results should be supported and evaluated by the other resources, if any, in the literature.
- Used resources should be in good wealth and proficiency that reflect the scope of the study.
- Turkish and English abstracts should include the goal, methodology, scope and main findings of the study.

For detailed information, please see “A Guide for Journal of Statistical Research” at <http://www.tuik.gov.tr>.

DERGİNİN BU SAYISINA BİLİMSEL KATKI SAĞLAYAN HAKEMLER
REFEREES WHO PROVIDED SCIENTIFIC CONTRIBUTIONS FOR THIS
VOLUME OF THE JOURNAL

Prof. Dr.	Ahmet YALNIZ	Çankaya Üniversitesi
Prof. Dr.	Bedriye SARAÇOĞLU	Gazi Üniversitesi
Prof. Dr.	Cem KADILAR	Hacettepe Üniversitesi
Prof. Dr.	Ergun KARAAĞAOĞLU	Hacettepe Üniversitesi
Prof. Dr.	Halis AKDER	Orta Doğu Teknik Üniversitesi
Prof. Dr.	İsmet KOÇ	Hacettepe Üniversitesi
Prof. Dr.	Melek BAŞAK	Yeditepe Üniversitesi
Prof. Dr.	Olca ARSLAN	Ankara Üniversitesi
Prof. Dr.	Onur BAŞKAN	Ege Üniversitesi
Prof. Dr.	Özkan ÜNVER	Ufuk Üniversitesi
Prof. Dr.	Öztaş AYHAN	Orta Doğu Teknik Üniversitesi
Prof. Dr.	Süleyman GÜNAY	Hacettepe Üniversitesi
Prof. Dr.	Ümit OKTAY FIRAT	Marmara Üniversitesi
Doç. Dr.	Ayşe Gündüz HOŞGÖR	Orta Doğu Teknik Üniversitesi
Doç. Dr.	Banu Akadlı ERGÖÇMEN	Hacettepe Üniversitesi
Doç. Dr.	Birdal ŞENOĞLU	Ankara Üniversitesi
Doç. Dr.	Cemal ATAKAN	Ankara Üniversitesi
Doç. Dr.	Filiz KARDAM	Çankaya Üniversitesi
Doç. Dr.	İnci BATMAZ	Orta Doğu Teknik Üniversitesi
Doç. Dr.	Mahmude Revan ÖZKALE	Çukurova Üniversitesi
Doç. Dr.	Saim YOLOĞLU	İnönü Üniversitesi
Yrd. Doç. Dr.	Ceylan YOZGATLIGİL	Orta Doğu Teknik Üniversitesi
Yrd. Doç. Dr.	Hakan Savaş SAZAK	Ege Üniversitesi
Yrd. Doç. Dr.	Özlem İLK	Orta Doğu Teknik Üniversitesi
Yrd. Doç. Dr.	Özlem TÜRKER BAYRAK	Çankaya Üniversitesi
Dr.	Dilek GÜVENÇ	Bilkent Üniversitesi
Uzman	Cengiz ERDOĞAN	Türkiye İstatistik Kurumu
Ekonomist	A. Atabek DEMİRHAN	T.C. Merkez Bankası

ESTIMATION IN SEMIPARAMETRIC REGRESSION MODELS

Fikri AKDENİZ*

ABSTRACT

In this paper we consider the semiparametric regression model, $y = X\beta + f + \varepsilon$. We introduce a Liu-type estimator (LTE) in a semiparametric regression model. We obtained the semiparametric restricted ridge regression and Liu-type estimators for the parametric component in semiparametric regression model. We also introduced difference-based ridge regression and Liu-type estimators in semiparametric regression model. Difference-based estimator and difference-based Liu-type estimator are compared in the sense of mean-squared error criterion.

Keywords: Difference-based estimator, Differencing matrix, Liu-type estimator, Ridge estimator, Multicollinearity, Semiparametric regression model.

1. INTRODUCTION

Let us consider the following semiparametric regression model (partially linear model):

$$y_i = x_i^T \beta + f(t_i) + \varepsilon_i, \quad i = 1, 2, \dots, n \quad (1)$$

where x_i (p-dimensional vectors) and t_i (real numbers) are known. In matrix form, it can be written as:

$$y = X\beta + f(t) + \varepsilon, \quad (2)$$

where $X = (x_1, \dots, x_n)^T$ is an $n \times p$ -dimensional matrix, $\beta = (\beta_1, \dots, \beta_p)^T$ is p-dimensional parameter vector, f is unknown function and to be estimated, ε is independent and identically distributed (i.i.d.) error terms with $E[\varepsilon | X, t] = 0$, $\text{Var}[\varepsilon | X, t] = \sigma_\varepsilon^2 I$. Without loss of generality, data is rearranged so that $0 \leq t_1 \leq t_2 \leq \dots \leq t_n \leq 1$. The explanatory variables are represented separately in two parts: the nonparametric part $f(t)$ and the parametric linear part $(X\beta)$.

Semiparametric regression model (SPRM) generalizes both the parametric linear regression model $y = X\beta + \varepsilon$ and nonparametric regression model $y = f(t) + \varepsilon$ which correspond to the cases $f=0$ and $\beta=0$, respectively. Some of the relations are believed to be of certain parametric form while others are not easily parameterized. SPRM is more flexible than the standard linear regression model since it combines both parametric and nonparametric components. Due to its flexibility, SPRM has been widely used in econometrics, finance, biology, sociology and so on. It allows easier interpretation of the effect of each variable compared to a completely nonparametric regression.

* Prof. Dr., Department of Statistics, Faculty of Arts and Sciences, Çukurova University, Adana, Turkey, e-posta: akdeniz@cu.edu.tr

This article was presented in the 7th Statistics Days Symposium organized by METU during 28-30 June 2010 and it has not been published in any other publications.

All the existing approaches for the partially linear model are based on different nonparametric regression procedures (see Härdle et al. (2000) and Härdle et al. (2004)). There have been several approaches for estimating β and f . Among the most important approaches are the spline methods used by Engle et al. (1986); Rice (1986); Speckman (1988); Green and Silverman (1994); Eubank et al. (1998) and Eubank (1999). Speckman (1988) introduced the idea of a profile least squares method and studied local constant smoother.

The paper is organized as follows. In Section 2, we introduce penalized least squares and Speckman’s methods in semiparametric regression models under multicollinearity. Section 3 gives restricted estimators in semiparametric regression models. The semiparametric restricted Liu-type estimator is given in Section 4. In Section 5, the differencing estimator proposed by Yatchew (2003) is given. The difference-based Liu estimator is explained in Section 6. Section 7 compares the difference based Liu-type estimator and difference-based estimator in terms of the mean squared error.

2. ESTIMATION IN SEMIPARAMETRIC REGRESSION MODELS UNDER MULTICOLLINEARITY

The problem of multicollinearity and its statistical consequences on a linear regression model are very well-known in statistics. A popular numerical technique to deal with multicollinearity is the use of ridge regression estimator: $\hat{\beta}(k) = (X^T X + kI)^{-1} X^T y, k > 0$ (see Hoerl and Kennard (1970)). An alternative is the use of Liu-type estimator (or Liu estimator): $\hat{\beta}_d = (X^T X + I)^{-1} (X^T y + \eta \hat{\beta}_{OLS}), 0 < \eta < 1$ (see Liu (1993); Akdeniz and Kaçıranlar (1995)).

2.1 Penalized Least Squares Method

We consider a penalized least squares method. Suppose we have the model provided in (2), then we estimate β and f by minimizing with respect to β and f simultaneously

$$S(f, \beta) = (y - X\beta - f)^T (y - X\beta - f) + \alpha f^T K f \tag{3}$$

(Buja *et al.* 1989), where K is a symmetric non-negative definite matrix, $\alpha K = S^{-1} - I$, $S = (I + \alpha K)^{-1}$, smoother matrix, and α is smoothing parameter. For model (2) the penalized least squares solution is given as below:

$$\hat{\beta} = [X^T (I - S) X]^{-1} X^T (I - S) y, \tag{4}$$

$$\hat{f} = S(y - X\hat{\beta}),$$

and

$$\mu_p = X\hat{\beta} + \hat{f} = H_p y$$

for

$$H_p = S + (I - S) X [X^T (I - S) X]^{-1} X^T (I - S)^T,$$

provided that $X^T (I - S) X$ is invertible (Buja *et al.* 1989, pp. 500).

2.2 Speckman's Method

If we wish to fit a SPRM given in (2) to data, we may discuss an alternative approach proposed by Speckman (1988) which minimizes

$$(\tilde{y} - \tilde{X}\beta)^T (\tilde{y} - \tilde{X}\beta) = \|(I - S)(y - X\beta)\|^2 \quad (5)$$

where $\tilde{y} = (I - S)y$, $\tilde{X} = (I - S)X$ and S is an arbitrary smoother matrix. We obtain a solution different from (4), namely the solution is now given in (6).

$$\tilde{\beta} = [X^T (I - S)^T (I - S)X]^{-1} X^T (I - S)^T (I - S)y, \quad (6)$$

$$\tilde{f} = S(y - X\tilde{\beta}),$$

$$\mu_s = X\tilde{\beta} + \tilde{f} = H_s y$$

for

$$H_s = S + (I - S)X[X^T (I - S)^T (I - S)X]^{-1} X^T (I - S)^T (I - S).$$

As it can be seen, the estimates are obtained by regression on partial residuals. $\hat{\beta}$ and $\tilde{\beta}$ are identical for symmetric and idempotent S matrix. The Speckman approach was investigated by Buja et al. (1989); Eubank et. al. (1998) and Schimek (2000).

a) Consider the following objective function:

$$L^* = (\tilde{y} - \tilde{X}\beta)^T (\tilde{y} - \tilde{X}\beta) + k\beta^T \beta \quad (7)$$

The first order condition of this objective function minimized by the vector β is

$$\frac{\partial L^*}{\partial \beta} = 0. \quad (8)$$

From condition (8) we obtain the following equation:

$$\tilde{\beta}_k = (\tilde{X}^T \tilde{X} + kI)^{-1} \tilde{X}^T \tilde{y}, \quad k > 0 \quad (9)$$

We call it a *ridge regression estimator of the semiparametric regression model*.

b) Consider the following conditional objective function:

$$F(\beta) = \arg \min_{\beta} \{(\tilde{y} - \tilde{X}\beta)^T (\tilde{y} - \tilde{X}\beta) + (\eta\tilde{\beta} - \beta)^T (\eta\tilde{\beta} - \beta)\} \quad (10)$$

By differentiating function F with respect to β , we obtain

$$\frac{\partial F}{\partial \beta} = 0.$$

The solution of this system is

$$\tilde{\beta}_\eta = (\tilde{X}^T \tilde{X} + I)^{-1} (\tilde{X}^T \tilde{y} + \eta\tilde{\beta}), \quad 0 < \eta < 1, \quad (11)$$

where η is a biasing parameter. Since there is a formal resemblance between (11) and the Liu estimator of the linear model, we call it a *Liu-type estimator of the semiparametric regression model*.

Combining $\tilde{\beta}_\eta$ and $f = S(y - X\beta)$, we obtain Liu-type estimator of f :

$$\tilde{f} = S(y - X\tilde{\beta}_\eta).$$

3. RESTRICTED ESTIMATORS IN SEMIPARAMETRIC REGRESSION MODELS

We will call an estimator $\beta^*(y)$ for β a restricted estimator with respect to $R\beta = r$, if it satisfies $R\beta^*(y) = r$ for all $n \times 1$ vectors y (Gross, 2003).

If the linear restrictions $R\beta = r$ are assumed to hold, we shall give the restricted estimator of β in semiparametric regression model. Here, R is an $m \times p$ known matrix with rank $m < p$ and r is an $m \times 1$ known vector.

Theorem 1. *If the smoother matrix S is symmetric, then the estimator of β obtained by using the penalized least-squares criterion, under the restriction $R\beta = r$, is*

$$\hat{\beta}_R = \hat{\beta} + D^{-1}R^T(RD^{-1}R^T)^{-1}(r - R\hat{\beta}) \tag{12}$$

with $D = X^T(I - S)X$ and $\hat{\beta} = D^{-1}X^T(I - S)y$.

Proof: (see Akdeniz and Tabakan (2009))

Theorem 2. *If S is an arbitrary smoother matrix, then the estimate of β , under $R\beta = r$, is*

$$\tilde{\beta}_R = \tilde{\beta} + Z^{-1}R^T(RZ^{-1}R^T)^{-1}(r - R\tilde{\beta}) \tag{13}$$

with $Z = X^T(I - S)^T(I - S)X$ and $\tilde{\beta} = (\tilde{X}^T\tilde{X})^{-1}\tilde{X}^T\tilde{y}$.

Let $\tilde{y} = (I - S)y$, $\tilde{X} = (I - S)X$.

Proof: (see Akdeniz and Tabakan (2009))

4. RESTRICTED LIU-TYPE ESTIMATORS IN SEMIPARAMETRIC REGRESSION MODELS

The semiparametric restricted ridge regression estimator for the parametric component in semiparametric regression model is established by Akdeniz and Tabakan (2009). We shall give the restricted Liu-type estimator $\tilde{\beta}_{R\eta}$ of β in semiparametric regression model such that condition $R\beta = r$ is satisfied. Przystalski and Krajewski (2007) presented a method of estimation of treatment effects and its application to hypothesis testing in a semiparametric model. Assuming the linear restrictions $R\beta = r$ hold, we

shall give the restricted estimator of β in the semiparametric regression model (2) in the following two theorems.

Theorem 3. *If S is an arbitrary smoother matrix, then the estimate of β in semiparametric model under the penalized least squares criterion such that condition $R\beta = r$ is satisfied, is equal to*

$$\tilde{\beta}_{R\eta} = \tilde{\beta}_\eta + \tilde{Z}^{-1}R^T(R\tilde{Z}^{-1}R^T)^{-1}(r - R\tilde{\beta}_\eta). \tag{14}$$

with

$$\begin{aligned} \tilde{\beta}_\eta &= (\tilde{X}^T \tilde{X} + I)^{-1}(\tilde{X}^T \tilde{y} + \eta \tilde{\beta}); & \tilde{Z} &= \tilde{X}^T \tilde{X} + I = X^T(I - S)^T(I - S)X + I & \text{and} \\ \tilde{\beta} &= (\tilde{X}^T \tilde{X})^{-1} \tilde{X}^T \tilde{y} = [X^T(I - S)^T(I - S)X]^{-1} X^T(I - S)^T(I - S)y. \end{aligned}$$

Proof: (Akdeniz and Akdeniz Duran (2010))

Theorem 4. *If the smoother matrix S is symmetric, then the estimate of β in semiparametric model, obtained by using the penalized least squares criterion*

$$\|y - X\beta - f\|^2 + \alpha f^T Kf + (\eta \hat{\beta}_\eta - \beta)^T (\eta \hat{\beta}_\eta - \beta) \tag{15}$$

satisfying condition $R\beta = r$, is equal to

$$\hat{\beta}_{R\eta} = \hat{\beta}_\eta + V^{-1}R^T(RV^{-1}R^T)^{-1}(r - R\hat{\beta}_\eta). \tag{16}$$

Proof: (Akdeniz and Akdeniz Duran (2010)) Let us assume that the smoother matrix S is symmetric and we want to find an estimate of β in semiparametric model satisfying condition $R\beta = r$ using the criterion (15). For this purpose we construct the following Lagrange function

$$L(\beta, f, \lambda) = \|y - X\beta - f\|^2 + \alpha f^T Kf + \|\eta \hat{\beta} - \beta\|^2 - 2\lambda^T(R\beta - r).$$

By differentiating function L with respect to β , f and λ we obtain

$$\hat{\beta}_{R\eta} = \hat{\beta}_\eta + V^{-1}R^T(RV^{-1}R^T)^{-1}(r - R\hat{\beta}_\eta). \tag{17}$$

This estimator satisfies $R\hat{\beta}_{R\eta} = r$. Thus $\hat{\beta}_{R\eta}$ is the restricted Liu-type estimator of β in semiparametric regression model.

5. DIFFERENCE-BASED ESTIMATORS

Remember that the model is given as the following:

$$y_i = x_i^T \beta + f(t_i) + \varepsilon_i, \quad i = 1, 2, \dots, n$$

In the matrix form, we can write

$$y = X\beta + f(t) + \varepsilon, \quad \dim(t) \leq 3 \tag{18}$$

where $X = (x_1, \dots, x_n)^T$ with $x_i, i \in (1, n)$ non-stochastic, p -dimensional vector; $\beta = (\beta_1, \dots, \beta_p)^T$ is parameter vector; f is smooth function whose first derivative is bounded by a constant; t has compact support, i.e. $[0, 1]$; ε is i.i.d. with $E[\varepsilon | X, t] = 0$, $\text{Var}[\varepsilon | X, t] = \sigma_\varepsilon^2 I$. t_i are equally spaced on the unit interval. Data is rearranged so that $0 \leq t_1 \leq t_2 \leq \dots \leq t_n \leq 1$.

Applying the differencing matrix to model (18) permits direct estimation of the parametric effect. In particular, take

$$Dy = DX\beta + Df(t) + D\varepsilon \doteq DX\beta + D\varepsilon$$

or

$$\ddot{y} \doteq \ddot{X}\beta + \ddot{\varepsilon} \tag{19}$$

where $\ddot{y} = Dy$, $\ddot{X} = DX$ and $\ddot{\varepsilon} = D\varepsilon$. We obtain the simple differencing estimator of the parameter β in semiparametric regression model as

$$\begin{aligned} \hat{\beta}_{diff} &= [(DX)^T DX]^{-1} (DX)^T Dy \\ &= (\ddot{X}^T \ddot{X})^{-1} \ddot{X}^T \ddot{y}. \end{aligned} \tag{20}$$

where differencing matrix D is $(n - m) \times n$ matrix

$$D = \begin{bmatrix} d_0 & d_1 & d_2 & \cdot & \cdot & \cdot & d_m & 0 & 0 & \cdot & \cdot & \cdot & \cdot & 0 \\ 0 & d_0 & d_1 & d_2 & \cdot & \cdot & \cdot & d_m & 0 & 0 & \cdot & \cdot & \cdot & 0 \\ \cdot & \cdot & \cdot & & & & & & & & & & & \\ \cdot & \cdot & \cdot & & & & & & & & & & & \\ 0 & 0 & 0 & \cdot & \cdot & \cdot & & d_0 & d_1 & \cdot & \cdot & d_m & 0 \\ 0 & 0 & 0 & \cdot & \cdot & \cdot & & d_0 & d_1 & \cdot & \cdot & d_m \end{bmatrix}.$$

and m order d_0, d_1, \dots, d_m differencing weights minimize

$$\min_{d_0, \dots, d_m} \delta = \sum_{k=1}^m (\sum_{j=1}^{m-k} d_j d_{k+j})^2 \text{ such that } \sum_{j=0}^m d_j = 0 \text{ and } \sum_{j=0}^m d_j^2 = 1 \text{ (Yatchew, 2003).}$$

5.1 What is the advantage of differencing?

An important advantage of differencing procedures is their simplicity. Increasing the order of differencing as sample size increases, the estimator of the linear component becomes asymptotically efficient (Yatchew 2003, p.72). Differencing allows one to perform inferences as if there were no nonparametric component f in the model (Yatchew, 2003). The m -th order differencing estimator of the residual variance is

$$S_{diff}^2 = \frac{1}{n} (Dy - DX \hat{\beta}_{diff})^T (Dy - DX \hat{\beta}_{diff}).$$

6. DIFFERENCE-BASED RIDGE AND LIU-TYPE ESTIMATORS

As an alternative to (20) we propose the use of

$$\begin{aligned} \hat{\beta}_D(k) &= [(DX)^T DX + kI]^{-1} (DX)^T Dy, \\ &= (\ddot{X}\ddot{X} + kI)^{-1} \ddot{X}\ddot{y} \end{aligned} \tag{21}$$

where I is $p \times p$ identity matrix, $k \geq 0$ is biasing parameter. We call it a difference-based ridge estimator of semiparametric regression model (Tabakan and Akdeniz, (2010)). If $k=0$, then $\hat{\beta}_D(k) = \hat{\beta}_{diff}$.

From the least squares perspective, the coefficients β are chosen to minimize

$$(Dy - DX\beta)'(Dy - DX\beta). \tag{22}$$

Adding to least squares objective (22), a penalizing function of the squared norm $\|\eta\hat{\beta}_{diff} - \beta\|^2$ for the vector of regression coefficients yields a conditional objective:

$$L(\beta) = \arg \min_{\beta} \left[(\ddot{y} - \ddot{X}\beta)^T (\ddot{y} - \ddot{X}\beta) + \|\eta\hat{\beta}_D - \beta\|^2 \right]. \tag{23}$$

Minimizing $L(\beta)$ by vector β , we obtain another alternative estimator $\hat{\beta}_D(\eta)$ to $\hat{\beta}_{diff}$ in (20):

$$\hat{\beta}_D(\eta) = (\ddot{X}^T \ddot{X} + I)^{-1} (\ddot{X}^T \ddot{y} + \eta\hat{\beta}_{diff}) \tag{24}$$

where $0 < \eta < 1$ is a biasing parameter for $\eta = 1$, $\hat{\beta}_D(\eta) = \hat{\beta}_{diff}$. We call $\hat{\beta}_D(\eta)$ differenced-based Liu-type estimator.

7. MSEM SUPERIORITY OF DIFFERENCE-BASED LIU-TYPE ESTIMATOR, $\hat{\beta}_D(\eta)$ OVER DIFFERENCING ESTIMATOR, $\hat{\beta}_{diff}$

7.1 Bias and Covariance of $\hat{\beta}_D(\eta)$

Putting $F_{\eta} = (\ddot{X}^T \ddot{X} + I)^{-1} (\ddot{X}^T \ddot{X} + \eta I)$ and observing F_{η} and $(\ddot{X}^T \ddot{X})^{-1}$ are commutative, we may write $\hat{\beta}_D(\eta)$ as

$$\hat{\beta}_D(\eta) = F_{\eta} \hat{\beta}_{diff} = F_{\eta} (\ddot{X}^T \ddot{X})^{-1} \ddot{X}^T \ddot{y} = (\ddot{X}^T \ddot{X})^{-1} F_{\eta} \ddot{X}^T \ddot{y}. \tag{25}$$

The expected value of $\hat{\beta}_D(\eta)$ can be written as

$$E[\hat{\beta}_D(\eta)] = \beta - (1-\eta)(\ddot{X}^T \ddot{X} + I)^{-1} \beta. \tag{26}$$

The bias of the $\hat{\beta}_D(\eta)$ estimator is given as

$$\text{Bias}[\hat{\beta}_D(\eta)] = -(1-\eta)(\ddot{X}^T \ddot{X} + I)^{-1} \beta. \tag{27}$$

Setting $S = (D^T \ddot{X})^T (D^T \ddot{X})$ and $U = (\ddot{X}^T \ddot{X})^{-1}$ we may write $\text{Cov}[\hat{\beta}_D(\eta)]$ as

$$\text{Cov}[\hat{\beta}_D(\eta)] = \sigma^2 U F_\eta S F_\eta^T U. \tag{28}$$

Similarly, we may obtain $\text{Cov}[\hat{\beta}_{diff}]$ as

$$\text{Cov}[\hat{\beta}_{diff}] = \sigma^2 U S U. \tag{29}$$

7.2 Variance Comparison

Theorem 5. *The sampling variance of the difference-based Liu-type estimator, $\hat{\beta}_D(\eta)$ is always less than or equal to that of differencing estimator, $\hat{\beta}_{diff}$ if and only if*

$$\lambda_{\max} \left(\frac{1}{\eta+1} M^{-1} N \right) \leq 1.$$

Proof : (Akdeniz Duran (2010))

When is difference-based Liu-type estimator superior to the differencing estimator in MSE sense?

Theorem 6. Consider two competing estimators $\hat{\beta}_D(\eta)$ and $\hat{\beta}_{diff}$ of β . Let

$$W = \left(\frac{1+\eta}{1-\eta} \right) (M + tN)$$

be a positive definite matrix. Then, the biased estimator $\hat{\beta}_D(\eta)$ is MSE- superior over the $\hat{\beta}_{diff}$ if and only if

$$\beta' W^{-1} \beta \leq \sigma^2.$$

Proof: (Akdeniz Duran (2010))

8. CONCLUSION

In this paper, we presented Liu-type Speckman and Liu-type penalized least squares estimators for the semiparametric regression model when linear restrictions on the parameter vector are assumed to hold. We also investigated ridge and Liu-type difference based regression estimators.

Acknowledgements

The author wish to thank to Esra Akdeniz Duran for her helpful suggestions and comments

9. REFERENCES

- Akdeniz D. E., 2010. Ph D. Thesis, Gazi University, Ankara (unpublished).
- Akdeniz, F., Kaçiranlar, S., 1995. On the almost unbiased generalized Liu estimator and unbiased estimation of the Bias and MSE. *Commun. Statist.-Theory and Meth.* 24(7), 1789-1797.
- Akdeniz, F., Tabakan, G., 2009. Restricted ridge estimators of the parameter in semiparametric regression model. *Communications in Statistics- Theory and Methods* 38(11):1852-1869.
- Akdeniz, F., Akdeniz D. E., 2010. Liu-type estimator in semiparametric regression models. *Journal of Statistical Computation and Simulation* 80(8):853-871.
- Akdeniz D. E., Akdeniz, F., 2010. Efficiency of Modified Jacknife Liu-type estimator. *Statistical Papers* DOI: 10.1007/s00362-010-0334-5).
- Akdeniz D, E., Akdeniz, F., Hongchang Hu., 2011. Efficiency of estimators in semiparametric regression models. *Journal of Computational and Applied Mathematics* 235:1418-1428.
- Buja, A., Hastie, T., Tibshirani, R., 1989. Linear smoothers and additive models. *The Annals of Statistics*, 17(2):453-555.
- Engle, R. F., Granger, C. W. J., Rice, J., Weiss, A., 1986. Semiparametric estimates of the relation between weather and electricity sales. *J. Amer, Statist. Assoc.* 81:310-320.
- Eubank, R. L., 1999. *Nonparametric regression and Spline Smoothing.* Marcel Dekker, New York.
- Eubank, R. L., Kambour, E. L., Kim, J. T., K.Klipple, K., Reese , C. S., Schimek, M. G., 1998. Estimation in partially linear models. *Computational Statistics and Data Analysis* 29:27-34.
- Green, P. J., Silverman, B. W., 1994. *Nonparametric Regression and Generalized Linear Models.* Chapman and Hall, New York.
- Gross, J., 2003. Restricted ridge estimation. *Statistics and Probability Letters* 65:57-64.
- Härdle, W., Liang, H., Gao, J., 2000. *Partially Linear Models* Springer-Verlag Co., New York.
- Härdle, W., Müller, M., Sperlich, S., Werwatz, A., 2004. *Nonparametric and Semiparametric Models.* Springer, New York.

- Hoerl, A. E., Kennard, R.W., 1970. Ridge regression: biased estimation for orthogonal problems. *Technometrics* 12:55-67.
- Liu, K. J., 1993. A new class of biased estimate in linear regression. *Communications in Statistics–Theory and Methods* 22:393-402.
- Przystalski, M., Krajewski, P., 2007. Constrained estimators of treatment parameters in semiparametric models. *Statistics and Probability Letters* 77(10):914-919.
- Rice, J. A., 1986. Convergence rates for partially spline models. *Statistics and Probability Letters* 4:203-208.
- Schimek, M. G., 2000. Estimation and inference in partially linear models with smoothing splines. *Journal of Statistical Planning and Inference* 91:525-540.
- Speckman, P., 1988. Kernel smoothing in partial linear models. *J. Roy, Statist.Soc. Ser. B.* 50(3):413-436.
- Tabakan, G., Akdeniz, F., 2010. Difference –based ridge estimator of parameters in partial linear model. *Stat. Papers*, 357-368.
- Yatchew, A., 2003. *Semiparametric Regression for the Applied Econometrician*. Cambridge University Press.

YARI PARAMETRİK REGRESYON MODELLERİNDE TAHMİN ETME

ÖZET

Bu makalede $y = X\beta + f + \varepsilon$ yarı parametrik regresyon modeli düşünüldü; yarı parametrik regresyon modelinde Liu-tip tahmin edici (LTE) önerildi. Ayrıca, yarı parametrik regresyon modelinde parametrik bileşen için yarı parametrik kısıtlı ridge regresyon ve Liu-tip tahmin edicileri de tanıtıldı. Farka dayalı tahmin edici ve farka dayalı Liu-tip tahmin edici hata kareleri ortalaması ölçütüne göre karşılaştırıldı.

Anahtar Kelimeler: Farka dayalı tahmin edici, Fark alma matrisi, Liu-tip tahmin edici, Ridge tahmin edici, Çoklu içilişki, Yarı parametrik regresyon modeli.

BULANIK SİSTEM MODELLERİNİN GELİŞİMİ

İ. Burhan TÜRKŞEN*

ÖZET

Bu çalışmamızda, Bulanık Sistem Modelleri'nin tarihi gelişimi gözden geçirilecektir. Zadeh (1975) "Bulanık Kural Tabanları" teklif etmiş, ve bunları Sugeno-Yasukawa (1993) geliştirmiştir. Diğer bir yönden, Tagaki-Sugeno (1985) sağ tarafı "Kural Tabanı" sol tarafı "Regresyon Denklemi" olan modellerin gelişimi yönünde tekliflerde bulunmuşlardır. Daha sonraki yıllarda, "Kural Tabanları" yerine "Bulanık Regresyon" modelleri teklif edilmiştir. Bunların ilki Hathaway ve Bezdek, (1993) tarafından doğrusal "Bulanık C-Regresyon Modeli" [Fuzzy C-Regression Model, (FCRM)] olarak sunulmuştur ve Höppner ve Klawonn (2003), Hathaway ve Bezdek (1993) modelinin doğrusal olmayan (nonlinear) gelişimini teklif etmişlerdir. Bu çalışmalar ötesinde, Türkşen (2008) "Bulanık Fonksiyonlar" modelini Hathaway ve Bezdek (1993) ve Höppner ve Klawonn (2003) modelleri yerine yeni bir model yapısı olarak sunmuştur. Daha sonra Türkşen (2008)'in sunduğu "Bulanık Fonksiyonlar" çeşitli yönlerde Çelikiyılmaz ve Türkşen (2008-2009) tarafından geliştirilmiştir.

Anahtar Kelimeler: Bulanık kural tabanları, Bulanık fonksiyonlar, Bulanık regresyon, Bulanık öbekleme.

1. GİRİŞ

Bu çalışmamızda tarihi gelişim içinde, önce Zadeh (1975) tarafından teklif edilen "Bulanık Sistem Modelleri"nden "Bulanık Kural Tabanları"nın gözden geçireceğiz. Bunların genelleştirilmiş modelleri Sugeno-Yasukawa (1993) tarafından incelenmiştir. Daha sonra Tagaki-Sugeno (1985) sağ tarafı "Kural Tabanı" ve sol tarafı "Regresyon" olan bir model sunmuşlardır.

İkinci olarak, doğrusal yaklaşımı Hathaway ve Bezdek (1993) tarafından önerilen "Bulanık C-Regresyon Modeli" (Fuzzy C-Regression Model) ve daha sonra doğrusal olmayan yaklaşımı Höppner ve Klawonn (2003) tarafından önerilen "Birleştirilmiş Bulanık Orta Çıkarma" (Combined FCM, ve FCRM) algoritmalarını da içeren gelişmeler inceleneceklerdir.

Üçüncü olarak Türkşen (2008) tarafından sunulan "Bulanık Fonksiyonlar" (Fuzzy Functions) ve daha sonra Çelikiyılmaz ve Türkşen (2008-2009) tarafından çeşitli yönlerden geliştirilen model tipleri irdelenecektir. Bezdek ve Türkşen tekliflerinin ana yapısı "Bulanık Kural Tabanları" yerine "Bulanık Orta Öbekleme" (FCM, Fuzzy C-Means, Bezdek (1981)) veya "Geliştirilmiş Bulanık Orta Öbekleme" (IFC, Improved Fuzzy Clustering, Çelikiyılmaz ve Türkşen (2008)) algoritmaları ile ortaya çıkan her bir "Bulanık Öbek"e ait bir fonksiyon yapısının sistem incelemelerinde kullanılmasıdır. Bu yazımızda, bütün bu modellerin değerlendirilmesi yapılacaktır.

* Prof. Dr., TOBB Ekonomi ve Teknoloji Üniversitesi, Endüstri Mühendisliği Bölümü, e-posta: bturksen@edu.edu.tr

Makale, 28-30 Haziran 2010 tarihlerinde ODTÜ tarafından düzenlenmiş olan 7. İstatistik Günleri Sempozyumu'nda sunulmuş olup, herhangi bir yerde yayımlanmamıştır.

1.1 Bulanık Fonksiyonların Temeli

Tarihi gelişim içinde, Bulanık Fonksiyonlar John Grinder ve Richard Bandler tarafından duyumsal temsili sistemlerin birleştirilmesi veya üst üste gelmesi olarak tanımlanmışlardır (*The Structure of Magic Volume II (1976)*). Teknik olarak, Grinder ve Bandler “bulanık fonksiyonları” şu şekilde ifade etmektedir:

“Bulanık Fonksiyonlar temsili bir sistemi modellemede ya girdi kanallarında ya da çıktı kanallarında girdi ve çıktılarının değişik tarzlarda yapılması durumunda sistemin temsil edilmesidir.” Geleneksel psikofizikte, “Bulanık Fonksiyon” terimine en yakın terim ‘Sinestezi’dir.

Ayrıca bulanık fonksiyon çalışmalarına ışık tutan bazı önemli yayınlar şunlardır: Sasaki, M., 1993, Fuzzy Functions, *Fuzzy Sets ve Systems*, 55, 295-301 ve Demirci, M., 1999, Fuzzy functions and their fundamental properties, *Fuzzy Sets ve Systems*, 106, 239-246.

1.2 Geleneksel Bulanık Sistem Modelleri (FRB, Fuzzy Rule Bases)

Geleneksel Bulanık Sistem Modelleri (FSM, Fuzzy System Models) bulanık öbeklerle tanımlanmış sistem değişkenlerinin arasındaki önemli ilişkilerin ortaya çıkarılmasıyla geliştirilmişlerdir.

En çok uygulanan bulanık kural tabanları, bir sistemin girdi ve çıktı değişkenleri arasındaki temel ilişkileri bulanık kümelerle tanımlamaya çalışan bir bulanık sistem modelidir. Bu yazımızda sadece Çok Girdili ve Tek Çıktılı (Multi-Input Single Output (MISO)) sistemlerden bahsedeceğiz. Genel olarak bulanık kural tabanlı sistem modelleri, girdi ve çıktı değişkenleri arasındaki ilişkileri bulanık kümelerle temsil edilen dil bilimsel yapılar kullanımıyla, diğer bir ifadeyle, “EĞER...İSE” yapılarıyla tanımlanan kurallar topluluğu olarak ifade edilirler.

Genel olarak her bulanık kural tabanı aşağıdaki gibi ifade edilir:

$$R : \text{ALSO}_{i=1}^{c^*} (\text{IF } antecedent_i \text{ THEN } consequent_i), \quad (1)$$

burada c^* bilirkişiler veya FCM veya IFCM gibi bir bulanık öbekleme algoritması tarafından belirlenen kural tabanının kural sayısıdır.

Yukarıda gösterilen formülüyle genel olarak tanımlanan bulanık kural tabanlarının bazı önemli yapıları şunlardır:

Eğer girdi ve çıktılar bulanık kümelerle temsil edilirse o zaman bulanık kural tabanı aşağıda görünen formüllerle yapılır:

$$R_i : \text{IF } \bigwedge_{j=1}^{m_i} (x_j \in X_j \text{ isr } A_{ij}) \text{ THEN } y \in Y \text{ isr } B_i, \forall i = 1, \dots, c^* \quad (2)$$

$$R : \text{ALSO}_{i=1}^{c^*} \left(\text{IF } \bigwedge_{j=1}^{m_i} (x_j \in X_j \text{ isr } A_{ij}) \text{ THEN } y \in Y \text{ isr } B_i \right) \quad (3)$$

Bu formüllerde “isr” X değişkenlerinin A kümeleriyle ve Y değişkeninin B kümeleriyle olan ilişkilerini belirler.

Başlangıçta, Zadeh [1975] yukarıda görülen yapıyı teklif etmiştir. Bu yapının genelleştirilmiş bir çevirisi Sugeno ve Yasukawa Bulanık Kural Tabanı [1993] olarak bilinir. Eğer çıktılar girdi değişkenlerinin doğrusal bir denklemi olarak ifade edilirse o zaman bunlara Takagi-Sugeno Bulanık Kural Tabanı [1985] denir. Bu Zadeh'in teklif ettiği Bulanık Kural Taban yapısının özel bir durumudur. Bu özel durum yapısı aşağıda gösterilmektedir. Başka bir deyişle, Tagaki-Sugeno [1985] sağ tarafı bir regresyon denklemi ve sol tarafı bir bulanık kural tabanı olan bir sistem yapısı sunarlar.

$$R: \text{ALSO } (IF \text{ antecedent}_i \text{ THEN } y_i = a_i x^T + b_i) \quad (4)$$

Burada c^* kural sayısını belirtmektedir.

1.3 Bulanık Fonksiyonlar

Yukarıda gözden geçirilen kural tabanlı model yapılarına karşılık, Bulanık Regresyon modelleri ayrıcalık taşıyan bir gelişmedir. Bunların:

- (a) İlk modeli, Hathaway ve Bezdek (1993) tarafından teklif edilen “Bulanık C-Regresyon” modelidir. Bu yapılanmada “Bulanık Orta Öbekleme” (FCM) yöntemiyle bulanık öbekler belirlenir. Böylece kaç regresyon denklemi kurulacağı tespit edilmiş olur. Diğer bir ifadeyle her bir öbek için bir regresyon modeli tanımlanır. Bundan sonra her bir öbek için FCRM algoritması kullanılarak bir “Bulanık C-Regresyon” modeli belirlenir.
- (b) İkinci modeli, Höppner ve Klawonn (2003) “Birleştirilmiş Bulanık Orta Çıkarma” (Combined FCM, ve FCRM) algoritması olarak sunarlar. Onların hedefi amaç fonksiyonunda yer alan Bulanık Orta Öbekleme (FCM) algoritmasını geliştirerek uyumlu yapıların etkisini önlemek olmuştur. Dikkate alınması gereken, nokta-odaklı öbekleme algoritması olan “Bulanık Orta Öbekleme (FCM)” ile “Bulanık C-Regresyon Modeli”nin öbekleme algoritmasının (FCRM) beraberce uygulamasının teklif edilmiş olmasıdır.

Tekrar vurgulamak gerekirse, çok iyi bilinen bir özellik şudur: Hathaway ve Bezdek (1993) doğrusal regresyon modelleri teklif etmişlerdir. Buna karşın, Höppner ve Klawonn (2003) doğrusal olmayan regresyon modelleri sunmuşlardır.

- (c) “Bulanık Fonksiyonlar”, Türkşen (2008) tarafından teklif edilmiş ve Çelikyılmaz ve Türkşen (2008-2009) tarafından birkaç yönde geliştirilmişlerdir. Bu fonksiyonlar Hathaway ve Bezdek (1993) ve Höppner ve Klawonn (2003) modellerinden yapısal olarak farklıdır. “Bulanık Fonksiyonlar”ın sağ tarafında esas girdi değişkenlerinin yanı sıra üyelik değerlerinin çeşitli dönüşümleri de yer almaktadır. Bu üyelik değerlerinin çeşitli dönüşümleri regresyon denkleminin tahmin etme gücünü artırır.

1.3.1 Bulanık C-Regresyon

Hathaway ve Bezdek (1993)'in “Bulanık C-Regresyon Modeli” (FCRM) başlangıçta nesnelere benzer öbeklere atamak için sunulmuştur.

Bulanık C-Regresyon Modeli (FCRM), verilen bir veri kümesini bulanık öbeklere atarken, eş zamanlı olarak model parametrelerini de tahmin eder. Hatırlamak gerekir ki “Bulanık Orta Öbekleme (FCM)” (Bezdek, 1981), nokta-odaklı bir öbekleme algoritmasıdır. Diğer bir yönden, Bulanık Orta Öbekleme (FCM)” öbekleri hiper-küre şeklindedirler. Karşıt olarak “Bulanık C-Regresyon Modeli” (FCRM) öbek prototiplerini geometric nesnelere yerine fonksiyonlarla belirler. Özel olarak Bulanık C-Regresyon Modeli (FCRM), öbek prototiplerini birbirinden ayrı doğrusal öbekler olarak belirler. Diğer bir deyişle, her bir öbek bir doğrusal fonksiyonla tanımlanır. FCRM (Hathaway ve Bezdek, 1993), öbekleri hiper-düzlem şeklindedir.

FCM ve FCRM Arasındaki Farklar:

FCM öbeklerini öbek merkezleri temsil etmektedir. Karşıt olarak, FCRM’in öbeklerini hiper-düzlemler temsil ederler ve aşağıdaki gibi ifade edilirler:

$$y_i = \beta_i^0 + \beta_i^1 x_1 + \dots + \beta_i^{mv} x_{mv}, \quad (5)$$

Her bir β_i ‘lar regresyon katsayılarıdır, $i=1 \dots c$, ve c öbek sayısını belirtir.

FCM algoritması, öbek merkezlerini her bir veri vektörünün üyelik değerleriyle ağırlıklandırılmış ortalaması olarak hesaplar. Buna karşılık FCRM her bir öbeği temsil eden fonksiyonları ağırlıklandırılmış “En Küçük Kareler” regresyon algoritmasıyla hesaplar:

$$y_k = f_i(\mathbf{x}_k, \beta_i), \quad \mathbf{x}_k = [x_{1,k}, \dots, x_{mv,k}]^T \in \mathfrak{R}^{nv}, \quad (6)$$

burada y_k , k ncı *çıktı* verisi, ve $\beta_i \in \mathfrak{R}^{nv}$, $i=1, \dots, c$ dir. Bu fonksiyonlar genel olarak aşağıda görülen denklemle ölçülürler:

$$E_{ik}(\beta_i) = (y_k - f_i(\mathbf{x}_k, \beta_i))^2 \quad (7)$$

Amaç fonksiyonu bu yaklaşık fonksiyonların tüm hatalarını en küçüğe indirmeyi öngörür:

$$E(U, \beta_i) = \sum_{i=1}^c \sum_{k=1}^n (\mu_{ik})^m E_{ik}(\beta_i) \quad (8)$$

FCRM’de, μ_{ik} ’ler, $f_i(\mathbf{x}_k, \beta_i)$ fonksiyonlarıyla tahmin edilen değerlerinin y_k değerine ne kadar yakın olduğunu ifade ederler.

FCM’den hatırlamamız gerekir ki:

$$\mu_{ik}^{(t)} = \left[\sum_{j=1}^c \left(\frac{d(x_k, v_j^{(t-1)})}{d(x_k, v_i^{(t-1)})} \right)^{\frac{2}{m-1}} \right]^{-1} \quad (9)$$

Buna karşı FCRM’den alınan sonuç:

$$\mu_{jk} = \left[\sum_{i=1}^c \left(\frac{E_{ik}}{E_{jk}} \right)^{\frac{1}{m-1}} \right]^{-1}, \quad \forall i, j=1, \dots, c < n \quad (10)$$

Özel olarak FCRM, verilen bir veri kümesinde var olan gizli yapıyı bulmaya çalışır. Bu aşağıda görülen yapı içinde geliştirilir:

$$\text{Min } E(U, \beta_i) = \sum_{i=1}^c \sum_{k=1}^n (\mu_{ik})^m E_{ik}(\beta_i) \quad (11)$$

burada $\beta_i = [X^T U_i X]^{-1} X^T U_i y$ ve

$$X_i = \begin{bmatrix} x_{i,1}^T \\ x_{i,2}^T \\ \vdots \\ x_{i,n}^T \end{bmatrix}, y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, U_i = \begin{bmatrix} \mu_{i1} & 0 & \cdots & 0 \\ 0 & \mu_{i2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \mu_{i,n} \end{bmatrix}$$

1.3.2 Doğrusal Olmayan Bulanık Fonksiyonlar

Höppner ve Klawonn (2003), FCM (Bezdek, 1981) ve FCRM (Hathaway ve Bezdek, 1993) algoritmalarını bir öbikleme kriteri içinde birleştirerek yeni bir öbikleme yapısı teklif ederler. Hedefleri sezgi yoluyla bulunan üyelik değerlerini yok etmektir. Bu nedenle FCM hedef fonksiyonu FCRM'le birleştirerek yeniden yapılandırılmıştır. Höppner ve Klawonn (2003)'da üyelik değerleri aşağıdaki gibi ifade edilmektedir.

$$\mu_{ik} = \left[\frac{d_{ik}^2 - \left(\min_{i=1..c} d_{ik}^2 - \eta \right)}{\sum_{j=1}^c \frac{d_{jk}^2 - \left(\min_{i=1..c} d_{jk}^2 - \eta \right)}{d_{jk}^2 - \left(\min_{i=1..c} d_{jk}^2 - \eta \right)}} \right]^{-1}, 0 < \eta \quad (12)$$

burada $\eta > 0$ araştırmacı tarafından belirlenen bir sabittir.

Höppner ve Klawonn (2003)'da her bir öbek için tanımlanan fonksiyon yapısı aşağıdaki gibidir:

$$\hat{y}_i = \hat{\beta}_i^T x_i \quad (13)$$

Bunlar Takagi-Sugeno (1985) modelinde bir kural gibi yorumlanırlar.

Bu amaçla, Höppner ve Klawonn (2003) yeni bir uzaklık fonksiyonu sunar. Bu uzaklık fonksiyonu aşağıda görüldüğü şekilde her iki yöntemin birleşimidir:

$$d_{ik}^2((x_k, y_k), (v_i, \hat{\beta}_i)) = \underbrace{\|x_k - v_i(x)\|^2}_{FCM \text{ distance}} + \underbrace{(y_k - \hat{\beta}_i^T \hat{x}_k)^2}_{FCRM \text{ distance}} \quad (14)$$

Burada \hat{x} , örneğin, değişkenleri $(1, x_1, x_2, x_1 x_2, x_1^2, x_2^2)$ olan iki boyutlu polinom gibi şekillendirilebilir. Modeldeki katsayılar aşağıda görülen formülle hesaplanırlar.

$$\hat{\beta}_i = \left(\sum_{k=1}^n (\mu_{ik})^m (y_k \hat{x}_k) \right) / \sum_{k=1}^n (\mu_{ik})^m (\hat{x}_k \hat{x}_k^T), \forall i = 1, \dots, c \quad (15)$$

1.3.3 Bulanık Fonksiyonlar

Başlangıçta Türkşen (2008)'in teklif ettiği ve sonra Çelikyılmaz ve Türkşen (2008-2009) tarafından çeşitli yönleri geliştirilen "Bulanık Fonksiyon" yapılarının ana özellikleri gözden geçirilecektir.

Örneğin, (X_k, Y_k) , $k=1, \dots, nd$, eğitim veri kümesinin gözlemleri ve $X_k = (x_{jk} | j = 1, \dots, nv, k = 1, \dots, nd)$ girdi değişkenleri olsunlar. Bu yöntemde en iyi (m^*, c^*) ikilisinin bir değerlendirme ölçüsü, “Öbek Geçerlilik Ölçüsü” dür. Bu ölçü, FCM veya IFC (Çelikyılmaz ve Türkşen (2009)) algoritmalarının tekrarlanan araştırmalarıyla tesbit edilirler. Bizim araştırmamızda “Bulanıklık Seviyesi”, m , genelde $m = 1.4, \dots, 2.6$ ve “Öbek sayısı”, c , genelde $c = 2, \dots, 10$ olarak alınır.

FCM, “Bulanık Orta Öbeleme” Algoritması:

$$\begin{aligned} \min J(U, V) &= \sum_{k=1}^{nd} \sum_{i=1}^c (u_{ik})^m (\|x_k - v_i\|)^2 \\ \text{s.t.} \quad 0 &\leq u_{ik} \leq 1, \forall i, k \\ \sum_{i=1}^c u_{ik} &= 1, \forall k \\ 0 &\leq \sum_{k=1}^{nd} u_{ik} \leq nd, \forall i \end{aligned} \tag{16}$$

Çalışmalarda A, Euclid Normu veya Mahalonobis Normu olarak ifade edilir.

Araştırmalarda FCM algoritması uygulanarak $m=m^*$ ve $c=1, \dots, c^*$ değerleri ve öbek merkezleri tespit edilir.

$$v_{X,j}^{m^*} = (x_{1,j}^{c^*}, x_{2,j}^{c^*}, \dots, x_{nv,j}^{c^*}) \quad v_{X|Y,j}^{m^*} = (x_{1,j}^{c^*}, x_{2,j}^{c^*}, \dots, x_{nv,j}^{c^*}, y_j^{c^*}) \tag{17}$$

En iyi Üyelik Değerleri ve “Artırılmış Girdi Matrisi” aşağıda görülen formüllerle tespit edilir.

$$u_{ik} = \left(\sum_{j=1}^c \left(\frac{\|x_k - v_{X,i}\|}{\|x_k - v_{X,j}\|} \right)^{\frac{2}{m-1}} \right)^{-1}, \quad \mu_{ik} = \{u_{ik} \geq \alpha\}, \tag{18}$$

Burada seçilen bir “alfa-kesiti” ile arzu edilmeyen “harmoni” değerleri yok edilerek değiştirilmiş üyelik değerleri aşağıda görülen formülle temin edilirler.

$$\gamma_{ij}(x_j) = \frac{\mu_{ij}(x_j)}{\sum_{i'=1}^c \mu_{i'j}(x_j)} \tag{19}$$

Böylece, aşağıdaki gibi ifade edilen X verilerinin i-inci öbektaki üyelik değerler matrisi temin edilmiş olur.

$$\Gamma_i = (\gamma_{ij} | i = 1, \dots, c^*; j = 1, \dots, nd) \tag{20}$$

Bazı “Artırılmış Girdi Matrisi” örnekleri aşağıda gösterilmiştir.

$$X_i' = [1, \Gamma_i, X] \quad X_i'' = [1, \Gamma_i^2, \Gamma_i^m, \exp(\Gamma_i), X] \tag{21}$$

Örneğin, birincisinin açık yazılımı şöyle olur:

$$X_{ij}' = [1, \Gamma_i, X_{ij}] = \begin{bmatrix} 1 & \gamma_{i1} & x_{i1} \\ \vdots & \vdots & \vdots \\ 1 & \gamma_{ind} & x_{ind} \end{bmatrix} \tag{22}$$

Böylece, her bir öbek için çıktının en küçük kareler tahmini, üyelik değerleri ve esas girdi değerlerinin gereken dönüşümleriyle birlikte belirlenir.

Bu yöntemle, bir girdisi ve ona ait üyelik değerleri olan birinci örnek için i-inci öbeğe ait **Bulanık Fonksiyonlar** aşağıdaki gibi ifade edilir.

$$Y_i = \beta_{i0} + \beta_{i1}\Gamma_i + \beta_{i2}X_{ij} \quad (23)$$

Burada değişkenler etkileşim içinde oldukları i-inci öbeğin j-inci kuralını temsil ederler. Modelin katsayıları aşağıda görülen formülle belirlenirler.

$$\beta_i^* = (X_{ij}^{*T} X_{ij}^*)^{-1} (X_{ij}^{*T} Y_i) \quad (24)$$

$$\beta_i^* = (\beta_{i0}^*, \beta_{i1}^*, \beta_{i2}^*)$$

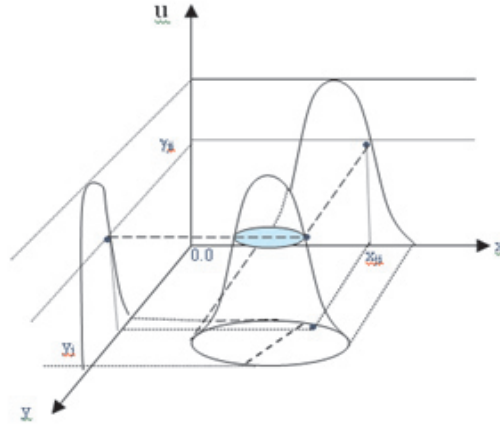
Bununla birlikte Y_i 'nin tahmin değerleri aşağıda görünen formülle belirlenirler.

$$Y_i^* = \beta_{i0}^* + \beta_{i1}^*\Gamma_i + \beta_{i2}^*X_{ij} \quad (25)$$

Son olarak, sistem çıktısının hesabı aşağıda gösterilmiştir.

$$Y^*_i = \frac{\sum_{i=1}^{c^*} \gamma_i Y_i^*}{\sum_{i=1}^{c^*} \gamma_i} \quad (26)$$

Bir boyutlu analizlerdeki ideal bir öbeğin gösterimi aşağıdaki grafikte verilmiştir.



Şekil 1. (UxXxY) uzayında tanımlanan bulanık öbek

Çelikyılmaz ve Türkşen tarafından geliştirilen bulanık fonksiyonlar hakkında daha detaylı bilgiye aşağıdaki kaynaklardan ulaşabilirsiniz.

“Enhanced Fuzzy System Models with Improved Fuzzy Clustering Algorithm”, (2008). “Uncertainty Modeling with Evolutionary Improved Fuzzy Functions Approach”, (2008). “Industrial Applications of Evolutionary Improved Fuzzy Functions”, Journal of Computers, (2008). “Increasing Accuracy of Two Class Pattern Recognition with Improved Fuzzy Functions”, (2009) ve diğerleri.

1.3.4 Geliştirilmiş Bulanık Öbeleme Algoritması (Improved Fuzzy Clustering Algorithm (IFC))

Bezdek (1981) tarafından sunulan Bulanık Öbeleme Algoritması (FCM) yerine, Çelikyılmaz, Türkşen (2007) Geliştirilmiş Bulanık Öbeleme Algoritması (*Improved Fuzzy Clustering (IFC)*) teklif etmiştir.

Yeni amaç fonksiyonu (27) iki amacı gerçekleştirir:

- (i) Bölme matrisinin iyi bir temsilini bulmak;
- (ii) Modellerin hatasını en küçüğe indirgeyen Bulanık Fonksiyon (*Fuzzy Fonksiyonları*)'ları bulmak.

$$J_m^{IFC} = \sum_{i=1}^c \sum_{k=1}^n (\mu_{ik}^{imp})^m d_{ik}^2 + \sum_{i=1}^c \sum_{k=1}^n (\mu_{ik}^{imp})^m (y_k - h_i(\tau_{ik}, \hat{w}_i))^2 \quad (27)$$

$d^2 = \|x_k y_k - v_i(x)\|^2$, her girdi-çıkıtlı veri vektörünün kesinliğini kontrol eder.

Geçici Bulanık Fonksiyonlar:

Üyelik değerlerinin çeşitli güçlerini ifade eden çeşitli geçici bulanık fonksiyonlar tanımlanabilir.

$$\text{Örneğin, } \tau_i = [\mu_i \log((1 - \mu_i^{imp}) / \mu_i^{imp})]$$

Her bir i-inci öbek için \mathfrak{R}^2 içinde yer alan düzlemler kümesinin tanımı şöyle olur:

$$h_i = \hat{w}_{0i} + \hat{w}_{1i} \mu_i^{imp} + \hat{w}_{2i} \log((1 - \mu_i^{imp}) / \mu_i^{imp}), \text{ yada } h_i = \tau_i^T \hat{w}_i, \quad (28)$$

burada, $\hat{w}_i^T = [\hat{w}_{0i} \ \hat{w}_{1i} \ \hat{w}_{2i}]$ "Geçici Bulanık Fonksiyonlar"ın katsayılarıdır.

Örneğin, \mathfrak{R}^2 'de yer alan özel bir bulanık fonksiyonlar kümesinin tanımı aşağıda verilmiştir:

$$\hat{y}_i = h_i(\tau_i, \hat{w}_i) = \hat{w}_{0i} + \hat{w}_{1i} \mu_i^{imp} + \hat{w}_{2i} \log\left(\frac{1 - \mu_i^{imp}}{\mu_i^{imp}}\right) = \hat{w}_{0i} + \sum_{j=1}^2 \hat{w}_{ji} \tau_{ji} \quad (29)$$

IFC algoritmasının uzaklık fonksiyonu şöyle belirlenir:

$$d_{ik}^{IFC} = \|z_k - v_i(z)\|^2 + (y_k - h(\tau_{ik}, \hat{w}_i))^2 \quad (30)$$

Bu amaca uyum sağlayan çözüm fonksiyonu aşağıda yazılan amaç fonksiyonuna bağlanır.

$$\text{Min } L_m = \sum_{i=1}^c \sum_{k=1}^n (\mu_{ik}^{imp})^m d_{ik}^2 + \sum_{i=1}^c \sum_{k=1}^n (\mu_{ik}^{imp})^m (y_k - h_i(\tau_{ik}, \hat{w}_i))^2 - \lambda \left(\left(\sum_{i=1}^c (\mu_{ik}^{imp}) \right) - 1 \right) \quad (31)$$

Burada λ , Lagrange çarpanıdır. Amaç fonksiyonunun öbek merkezlerine ve üyelik değerlerine göre türevleri alınarak en iyi üyelik değerleri elde edilir.

$$\left(\mu_{ik}^{imp} \right)^{(t)} = \left(\sum_{j=1}^c \left(\frac{(d_{ik}^{(t-1)})^2 + (y_k - h_i(\tau_{ik}^{(t-1)}, \hat{w}_i))^2}{(d_{jk}^{(t-1)})^2 + (y_k - h_j(\tau_{jk}^{(t-1)}, \hat{w}_j))^2} \right)^{1/(m-1)} \right)^{-1} \quad (32)$$

$1 < i, j \leq c$
 $1 \leq k \leq n$

2. DENEYLER

Aşağıda Tablo 1’de gösterilen çeşitli sistem modelleriyle deneyler yapılmıştır.

Tablo 1. Deneylerde kullanılan veri tabanlarının özellikleri

No	Dataset	Type*	OBS**	#Vars [§]	OBS used in three-way cross validation			Performance Measure Used [¶]
					Training	Validation	Testing	
1	Friedman Artificial	R	9,791	5	500	250	9,000	▷ RMSE ▷ MAPE ▷ Robust Simulated Trading Benchmark (RSTB) ▷ R ² ▷ Ranking
2	Auto-Mileage-UCI	R	398	8	125	45	160	
3	StockPrice Predict.	R	389	16	120	90	160	
	▷ TD	R	445	16	200	144	160	
	▷ H&M	R	445	16	200	144	160	
	▷ E&M	R	445	16	200	144	160	
	▷ SunLife	R	445	16	200	144	160	
4	Demographic Process	R	10,000	11	250	750	9000	
	▷ Rangem1	R	10,000	11	250	750	9000	
	▷ Rangem2	R	10,000	11	250	750	9000	
5	Liver Disorders-UCI	C	345	6	175	75	50	▷ Accuracy ▷ ROC Curve/ AUC ▷ Several Ranking Methods
6	Lymphoma-UCI	C	340	34	150	120	80	
7	Breast Cancer - UCI	C	277	9	130	70	50	
8	Diabetes-UCI	C	768	8	125	75	50	
9	Credit Scoring UCI	C	690	15	130	75	50	
10	California Housing	C	20,640	9	500	500	12,600	

*R: Regresyon, C: Sınıflama tipi veri setleri.

** OBS: Toplam gözlem sayısı.

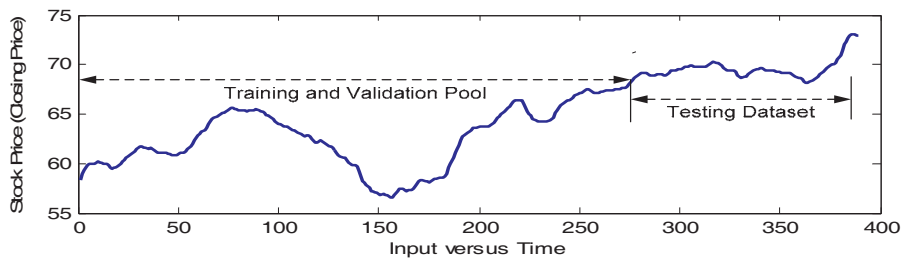
§ Var: Veri kümesinde yer alan toplam değişken sayısı.

UCI: Kaliforniya Üniversitesi, Irvine, Gerçek veri havuzu

2.1 Deneysel Sonuçları

Yukarıdaki Tablo 1’de belirtilen bütün veri kümeleri için eğitim (Training) ve onaylama (Validation) deneyleri yapılmıştır. Ayrıca Tip-1 ve Tip-2 modelleri kurularak karşılaştırma yapılmıştır.

Sonuçlar aşağıdaki şekilde gösterilmektedir. Hisse senedi fiyatlarının tahmin modelleri için kullanılan üç yönlü çapraz doğrulama işlemi



Şekil 2. Hisse senedi fiyatlarının tahmin modelleri için kullanılan üç yönlü çapraz doğrulama işlemi

2.2 Modellerin Tahmin Etme Gösterisi

Bu çalışmada çeşitli gösteri ölçümlerinin analizi yapılmıştır.

Tablo 1. Tip-1 ve Tip-2 bulanık fonksiyon modellerinin regresyon parametreleri

Parameters	T1FF	DIT2FF	ET1FF	EDIT2FF
FCM-Küme Sayısı*	[2,...,10]	[2,...,10]	min c = 2 max c = 10	min c = 2 max c = 10
FCM m-değerleri	{1.2,1.3,...2.6}	{1.2,1.3,...2.6}	m-alt = 1.2 m-üst = 2.6	m-alt = 1.2 m-üst = 2.6
α -kesiti	{0, 0.1}	{0, 0.1}	min- α -kesiti=0 max- α -kesiti=0.1	min- α -kesiti=0 max- α -kesiti=0.1
Regresyon için Bulanık Fonksiyon Tipleri	LSE ve SVM	LSE ve SVM	LSE ve SVM	LSE ve SVM
C-reg	{ $2^{-1}, 2^0, \dots, 2^7, 2^8$ }	{ $2^{-1}, 2^0, \dots, 2^7, 2^8$ }	minCreg = 2^{-1} maxCreg= 2^8	minCreg= 2^{-1} maxCreg= 2^8
Epsilon	{0.01,...,0.5}	{0.01,...,0.5}	min- ϵ =0.01 max- ϵ =0.5	min- ϵ =0.01 max- ϵ =0.5
Gamma	1	1	1	1
Kernel Tipi	Linear, Radial Basis Fonksiyonları	Linear, Radial Basis Fonksiyonları	Linear, Radial Basis Fonksiyonları	Linear, Radial Basis Fonksiyonları

Tablo 2. Tip-1 ve Tip-2 bulanık fonksiyon modellerinin klasifikasyon parametreleri

Parameters	T1FF-C	DIT2FF-C	ET1FF-C	EDIT2FF-C
FCM-Küme Sayısı*	[2,...,10]	[2,...,10]	min c = 2 max c = 10	min c = 2 max c = 10
FCM m-değerleri	{1.2,1.3,...2.6}	{1.2,1.3,...2.6}	m-alt = 1.2 m-üst = 2.6	m-alt = 1.2 m-üst = 2.6
α -kesiti	{0, 0.1}	{0, 0.1}	min- α -kesiti=0 max- α -kesiti=0.1	min- α -kesiti=0 max- α -kesiti=0.1
Sınıflama için Bulanık Fonksiyon Tipleri	LR, SVC	LR, SVC	LR, SVC	LR, SVC
C-reg	{ $2^{-1}, 2^0, \dots, 2^7, 2^8$ }	{ $2^{-1}, 2^0, \dots, 2^7, 2^8$ }	minCreg = 2^{-1} maxCreg= 2^8	minCreg = 2^{-1} maxCreg= 2^8
Gamma	1	1	1	1
Kernel Tipi	Linear,Radial Basis Fonksiyonları	Linear,Radial Basis Fonksiyonları	Linear,Radial Basis Fonksiyonları	Linear,Radial Basis Fonksiyonları

Tablo 3. Tip-1 ve Tip-2 geliştirilmiş bulanık fonksiyon modellerinin regresyon parametreleri

Parameters	T1IFF	DIT2IFF	ET1IFF	EDIT2IFF
FCM-Küme Sayısı*	[2,...,10]	[2,...,10]	min c = 2 max c = 10	min c = 2 max c = 10
IFC m-değerleri	{1.2,1.3,...2.6}	{1.2,1.3,...2.6}	m-alt = 1.2 m-üst = 2.6	m-alt = 1.2 m-üst = 2.6
α -kesiti	{0, 0.1}	{0, 0.1}	min- α -kesiti=0 max- α -kesiti=0.1	min- α -kesiti=0 max- α -kesiti=0.1
Regresyon için IFC Tipleri	LSE ve SVM	LSE ve SVM	LSE ve SVM	LSE ve SVM
C-reg	{ $2^{-1}, 2^0, \dots, 2^7, 2^8$ }	{ $2^{-1}, 2^0, \dots, 2^7, 2^8$ }	minCreg = 2^{-1} maxCreg= 2^8	minCreg = 2^{-1} maxCreg= 2^8
epsilon	{0.01,...,0.5}	{0.01,...,0.5}	min- ϵ =0.01 max- ϵ =0.5	min- ϵ =0.01 max- ϵ =0.5
Gamma	1	1	1	1
Kernel Type	Linear, Radial Basis Fonksiyonları	Linear, Radial Basis Fonksiyonları	Linear, Radial Basis Fonksiyonları	Linear, Radial Basis Fonksiyonları

Tablo 4. Freidman yapay veri tabanı uygulamalarında“Eğitim-Onaylama-Test” çalışmalarında Tip-1 bulanık fonksiyonlar yaklaşımında deneylerde elde edilen R^2 değerleri

R^2 (Stdev)	T1IFF	T1IFF	ET1IFF	ET1IFF
Eğitim- R^2	0.97 (0.008)	0.97 (0.009)	0.97 (0.003)	0.97 (0.003)
Doğrulama- R^2	0.94 (0.007)	0.94 (0.006)	0.94 (0.006)	0.94 (0.009)
Test- R^2	0.939 (0.001)	0.939 (0.001)	0.942 (0.001)*	0.942.004)

*En iyi model sonuçlar koyu renklerle gösterilmiştir.

Tablo-5. Tip-1 bulanık fonksiyon yaklaşımlarında kullanılan parametre değerleri

	T1IFF	T1IFF	ET1IFF	ET1IFF
Regresyon Tipi	Doğrusal Olmayan SVM	Doğrusal Olmayan SVM	Doğrusal Olmayan SVM	Doğrusal Olmayan SVM
Küme Sayısı	{8,10}	{8}	{7,8}	{6,7,8}
Alpha-kesiti	{0, 0.1}	{0, 0.1}	[0,0.1]	[0,0.1]
Bulanıklık Derecesi	{1.5,1.8}	{1.8}	{1.6,1.8}	{1.3,1.6,1.85}
C-reg	{2,24}	{2}	{6.2,8.3}	[3.5,55]
Epsilon	{0.05}	{0.05}	{0.02,0.08}	[0.014,0.1]

Tablo 6. Freidman yapay Veri Tabanı “Eğitim-Onaylama- Test” çalışmalarında elde edilen R^2 değerleri ve en iyi model parametreleri

R^2 (Stdev)	ANFIS	DENFIS	NN	GFS	SVM
Eğitim- R^2	0.999 (0.001)	0.917 (0.010)	0.887 (0.04)	0.985(0.002)	0.966 (0.004)
Onaylama- R^2	0.486 (0.211)	0.863 (0.006)	0.870(0.053)	0.735(0.156)	0.939 (0.007)
Test- R^2	0.444 (0.23)	0.855 (0.007)	0.873 (0.004)	0.728(0.155)	0.938 (0.001)
Çapraz Doğrulama sonucu elde edilen parametrelerin ortalaması	<i>Kural Sayısı: {51}</i>	<i>Eşikdeğeri: 0.1 KuralSayısı: {39}</i>	<i>Nöron Sayısı: 50</i>	<i>Kural Sayısı: {49, 50}</i>	<i>RBF C-reg: {2}, Epsilon: {0.05} Destek Vektör Sayısı: 432</i>

3. SONUÇLAR

Araştırmamızda uyguladığımız yöntemler ve elde ettiğimiz bazı sonuçlar aşağıdaki gibidir:

A) Herhangi bir veri tabanındaki saklı yapıları keşfetmek için iki öbikleme algoritması uyguladık. Böylece bulanık öbikleme algoritmalarının herhangi bir veri tabanına uygulanmasıyla o veri kümesindeki gizli yapıların tanımlanabileceğini aşağıda belirtilen algoritmalarla göstermiş olduk.

- I) Bulanık C- Ortalama (Fuzzy C-Means (FCM)) öbikleme metodu [Bezdek, 1981]
- II) Bulanık C- Regresyon Öbikleme Metodu (FCRM) [Hathaway ve Bezdek ,1993]
- III) Geliştirilmiş Bulanık Öbikleme Metodu (IFC) [Çelikyılmaz, Türkşen, 2008]

B) Aşağıda gösterilen her iki yolla deneyler yaptık:

- I) Tip-1 Bulanık Fonksiyonlar ve Geliştirilmiş Bulanık Fonksiyonlar ve
- II) Tip-2 Bulanık Fonksiyonlar ve Geliştirilmiş Bulanık Fonksiyonlar

Burada “Geliştirilmiş Bulanık Fonksiyonların” başlıca özelliklerini irdelemek gerekir. Bunlar:

- a) Geliştirilmiş bulanık öbikleme algoritmalarıyla elde edilen üyelik değerleri ve onların dönüşümleri esas girdi değişkenlerine ilave edilerek belirleyici faktörler olarak kullanılırlar.
- b) Fonksiyonların yapısını tanımlayan iki parametre, m^* ve c^* 'nin ötesinde, sunulan yapılanma, geliştirilmiş bulanık fonksiyon tip ve yapılarının tanımlanmasını gerektirir.

Bunların ötesinde, sistem parametrelerini tanımlayan iki farklı strateji uygulanmıştır.

- I) Birincisi Tip-1 Geliştirilmiş Bulanık Fonksiyonlar (T1IFF) yöntemiyle ayrıntılı araştırma uygulayarak çikarsama model parametrelerini tanımlamak.

- II) İkinci yöntem Evrimsel Tip-1 Geliştirilmiş Bulanık Fonksiyonları (ET1IFF)'nda genetik algoritmalar kullanarak sistem parametrelerini en iyileştirmek. ET1IFF, T1IFF'e oranla hesaplarda daha az zaman harcar, çünkü bu daha az en iyileştirme adımları kullanır.
- III) Böylece ET1IFF yöntemleri üssel büyüme gösteren tarama alanlarını yönetilir ölçüde aza indirger. ET1IFF metoduyla çıkarsama parametreleri, parametrelerin belirlenen sınırları içinde otomatik olarak tanımlanırlar.

4. KAYNAKLAR

- Bezdek, J. C., 1981. Pattern Recognition with Fuzzy Objective Functions Algorithms. Kluwer Academic Publishers, Norwell, MA.
- Çelikyılmaz, A., Türkşen, I.B., 2009. Modeling Uncertainty with Fuzzy Logic. Elsevier.
- Çelikyılmaz, A., Türkşen, I.B., 2008. Enhanced Fuzzy System Models With Improved Fuzzy Clustering Algorithm. IEEE Transactions on Fuzzy Systems, 16(3):779-794.
- Çelikyılmaz, A., Türkşen, I.B., 2008. Uncertainty Modeling of Improved Fuzzy Functions With Evolutionary Systems. IEEE Trans. on Sys., Man & Cybern., 38(4):1098-1110.
- Çelikyılmaz, A., Türkşen, I.B., 2008. Validation Criteria for Enhanced Fuzzy Clustering. Pattern Recognition Letters, 29:97-108.
- Demirci, M., 1999. Fuzzy functions and their fundamental properties. Fuzzy Sets ve Systems, 106:239-246.
- Grinder, J., Bveler, R., 1976. Patterns of the Hypnotic Techniques of Milton H. Erickson. M.D. Volume I. Cupertino, CA: Meta Publications, ISBN 1555520529.
- Hathaway, R.J., Bezdek, J.C., 1993. Switching regression models and fuzzy clustering. IEEE Trans. Fuzzy Syst., 1:195-204.
- Höppner, F., Klawonn, F., 2003. A Contribution to Convergence Theory of Fuzzy c-Means and Derivatives. IEEE Trans. on Fuzzy Systems, 11(5):682-694.
- Sasaki, M., 1993. Fuzzy Functions. Fuzzy Sets and Systems. 55(3):295-301.
- Sugeno, M., Yasukawa, T., 1993. A Fuzzy-Logic- Based Approach to Qualitative Modeling. IEEE-Trans. on Fuzzy Systems, 1(1):7-31.
- Takagi T., Sugeno M., 1985. Fuzzy Identification of System and Its Application to Modeling and Control. IEEE, SMC-15:115-133.
- Türkşen, I.B., 2008. Fuzzy functions with LSE. Appl. Soft Comput., 8(3):1178-1188.
- Türkşen, I.B., 2009. Fuzzy System Models. Encyclopedia of Complexity and Systems Science, 4080-4094.

Zadeh. L. H., 1975. The concept of a linguistic variable and its application to approximate reasoning. Part I: 'Inf. Sci., 8:199-249; 'Part II: 'Inf. Sci., 8:301-357; 'Part III: 'Inf. Sci., 9:43-80.

DEVELOPMENT OF FUZZY SYSTEM MODELS

ABSTRACT

In this study, we first review the development of Fuzzy System Models in an historical perspective. "Fuzzy Rule Bases" proposed by Zadeh (1975), first were developed by Sugeno-Yasukawa (1993). Later Tagaki-Sugeno (1985) proposed models which has "Fuzzy Rulebases" on the left hand side and "Regression Equations" on the right hand side. Later on "Fuzzy Regression" models were proposed in place of "Fuzzy Rulebases". First of these were proposed by Hathaway and Bezdek (1993) as "Fuzzy C-Regression Model". Secondly, Höppner and Klawonn (2003) proposed nonlinear versions of Hathaway and Bezdek (1993) model. Beyond these works, "Fuzzy Functions" were proposed by Türkşen (2008) in place of Hathaway and Bezdek (1993) and Höppner and Klawonn (2003) models. Further developments on "Fuzzy Functions" were proposed by Çelikyılmaz and Türkşen (2008-2009) in a variety of versions. An experimental assessment of all these models are also discussed in this study.

Keywords: Fuzzy rule bases, Fuzzy functions, Fuzzy regression, Fuzzy clustering.

DOĞRUDAN VE DOLAYLI MEVSİM DÜZELTME YAKLAŞIMLARI: GSYİH SERİLERİ İÇİN BİR UYGULAMA

Özlem YİĞİT*

Bedriye SARAÇOĞLU**

ÖZET

Toplaştırılmış zaman serilerinde mevsimsel düzeltme, mevsimsel düzeltilmiş ana grupların toplulaştırılması yoluyla dolaylı olarak ya da toplulaştırılmış serinin mevsimsellikten arındırılması yoluyla doğrudan elde edilebilir. Ne yazık ki bu stratejiler oldukça farklı sonuçlar üretebilir. Bu çalışmada literatürde doğrudan ve dolaylı yaklaşım olarak adlandırılan alternatif yaklaşımlarla elde edilmiş mevsimsel düzeltme sonuçlarını karşılaştırabilmek ve mevsimsel düzeltmenin kalitesini değerlendirebilmek için kullanılan genel ölçütler ve diyagnostikler tartışılmıştır. Doğrudan ve dolaylı yaklaşımla mevsimsel düzeltme yaklaşımları, Türkiye İstatistik Kurumu (TÜİK) tarafından yayımlanan Gayri Safi Yurtiçi Hasıla (GSYİH) serileriyle elde edilen mevsimsel düzeltme sonuçları temelinde karşılaştırılmıştır. 1998:01–2009:02 dönemine ait üç aylık verilerin kullanıldığı çalışmada GSYİH'nin dolaylı yaklaşımla mevsimsel düzeltilmesine karar verilmiştir.

Anahtar Kelimeler: Doğrudan ve dolaylı yaklaşım, Mevsimsel düzeltme, Zaman serileri.

1. GİRİŞ

Hava durumu, takvim etkileri, idari kararların zamanlaması, iktisadi birimlerin üretim ve tüketim kararları gibi etkenler; doğrudan ya da dolaylı olarak zaman serileri üzerinde mevsimsel dalgalanmalar oluşturur. Bu tür dalgalanmalar yıl içinde sistematik olmakla birlikte her zaman düzenli nitelikte değildir (Hylleberg, 2006). Mevsimsel düzeltme ise en yalın ifadeyle ham verilerde gözlenen bu mevsimsel dalgalanmalardan yok oluncaya kadar serilerin filtrelenmesi işlemidir (Ghysels ve Osborne, 2001, 93).

Zaman serilerinde var olan mevsimsel etkiler geçici nitelikte olup bu etkilerin varlığı ekonomik zaman serilerinin genel eğiliminin izlenmesini engeller. Ekonomik zaman serileriyle ilgili kısa dönemli kestirim yapabilmek ve zaman serilerini daha sağlıklı karşılaştırabilmek için buna benzer geçici etkilerin serilerden arındırılması son derece önemlidir. Bu amaçla bir çok istatistik ofisi yayınladığı istatistikleri mevsimsellikten arındırarak yayınlamaktadır¹.

Bir zaman serisi iki ya da daha fazla alt bileşenin toplamından (genellikle ağırlıklı toplamından) oluşabilir. Alt bileşenlerin toplulaştırılması yoluyla elde edilmiş serilerin mevsimsellikten arındırılması konusunda doğrudan mevsimsel düzeltme yaklaşımı ve dolaylı mevsimsel düzeltme yaklaşımı olmak üzere iki temel yaklaşım kullanılmaktadır.

* TÜİK Uzmanı, Türkiye İstatistik Kurumu Ankara Bölge Müdürlüğü, e-posta: ozlemyigit@tuik.gov.tr

** Prof. Dr., Gazi Üniversitesi İİBF. Ekonometri Bölümü Öğretim Üyesi, e-posta: bedriye@gazi.edu.tr

¹ Türkiye İstatistik Kurumu 2009 yılı içinde Sanayi Üretim Endeksi ve GSYİH verilerini mevsim ve takvim etkilerinden arındırarak yayınlamaya başlamıştır.

Doğrudan yaklaşımda (Direct Approach) alt bileşenler için hiçbir mevsimsel düzeltme işlemi yapılmaz. Mevsimsel düzeltme işlemi toplulaştırılmış seriye uygulanır. Dolaylı yaklaşımda (Indirect Approach) ise önce alt bileşenler aynı mevsimsel düzeltme yöntemi ve programıyla düzeltilir. İkinci aşamada mevsimsellikten arındırılmış seriler toplulaştırılır (Ladiray ve Mazzi, 2003, 37). Fakat genellikle, toplulaştırılmış seri alt bileşenlerin doğrusal bir kombinasyonu olduğundan ve mevsimsel düzeltme ise doğrusal olmayan bir süreç olduğundan mevsimsel düzeltmede doğrudan ve dolaylı yaklaşım farklı sonuçların elde edilmesine neden olabilir. Bu durum bu iki yaklaşım arasında seçim problemini doğurur.

Bu yaklaşımlardan hangisinin seçileceği konusu literatürde pek çok çalışmada tartışılmalı bir konu olmuştur (Ghysels, 1997; Hood and Findley, 2001; Ladiray ve Mazzi, 2003; Otranto and Triacca, 2002; Peronaci, 2003; Rietzler vd, 2000). Tartışmalar bu yaklaşımların arasında önemli bir fark olup olmadığı, (eğer varsa) bu yaklaşımlar arasında seçim yapabilmek için ne gibi ölçütlere gerek duyulduğu, hangi koşullarda hangi yaklaşımın daha doğru tahminler ürettiği gibi konular üzerinde yoğunlaşmıştır.

Türkiye’de ise Atuk ve Ural (2002) dışında bu konuya değinen bir çalışma göze çarpmamaktadır. Atuk ve Ural, TRAMO/SEATS (Time Series Regression with Arima Noise, Missing Observations and Outliers /Signal Extraction in ARIMA Time Series) ve X-12-ARIMA mevsimsel düzeltme yöntemlerinin para arzları M1, M2 ve M2Y üzerindeki performanslarını karşılaştırdıkları çalışmalarında doğrudan ve dolaylı yaklaşım konusuna oldukça kısıtlı bir çerçevede değinmişlerdir. Alt bileşenlerin birbirinden farklı takvim etkileri taşıması ve alt bileşenlerin veri kaynaklarının farklı olması sebebiyle dolaylı yaklaşımın kullanılmasının daha uygun olacağı sonucuna varmışlardır. Bu çalışmada ise mevsimsel düzeltmede doğrudan ve dolaylı yaklaşım konusu daha kapsamlı bir şekilde ele alınmak suretiyle elde edilen sonuçlar bilimsel kriterle değerlendirilmiş, böylece literatüre katkı sağlamak amaçlanmıştır. Çalışmada öncelikle alt bileşenler takvim etkileri bağlamında incelenmiş, ardından dolaylı yaklaşımla ve doğrudan yaklaşımla mevsimsellikten arındırılan seriler arasında önemli bir fark olup olmadığı Johansen Eşitlikleşme² (Johansen Cointegration) testiyle araştırılmıştır. Son olarak elde edilen sonuçlar düzgünlük, kalıntı mevsimsellik ve revizyon geçmişi kriterleri doğrultusunda karşılaştırılmıştır.

Çalışmanın ilk bölümü olan yöntem bölümünde doğrudan ve dolaylı yaklaşım arasındaki farklar, bu yaklaşımlar arasında seçim yapabilmek için gerekli teorik ve istatistiksel kriterler tartışılmıştır. Çalışmanın ikinci bölümünde ise TÜİK tarafından yayımlanan GSYİH serisi kullanılarak elde edilen uygulama sonuçları sunulmuş ve yöntem bölümünde açıklanan kriterler doğrultusunda doğrudan ve dolaylı yaklaşımla mevsimsellikten arındırılan seriler karşılaştırılmıştır. Sonuç bölümünde ise elde edilen sonuçlar yorumlanmış ve önerilere yer verilmiştir.

2. YÖNTEM

Ladiray ve Mazzi’ye göre mevsimsel düzeltmede dolaylı yaklaşımın daha iyi olduğuna dair yeterince güçlü bir kanıt olmadıkça doğrudan yaklaşım tercih edilmelidir (Ladiray ve Mazzi, 2003).

² Eşitlikleşme sözcüğünün “Cointegration” yerine kullanımında Türk Dil Kurumu’nun yayımlanmış olduğu Ekonometri Terimleri Sözlüğü’ne uyulmuştur.

Bunun en önemli nedeni, dolaylı yaklaşımın doğrudan yaklaşıma göre oldukça zahmetli bir yöntem olmasıdır. Aynı gerekçeyle Eurostat, bu yaklaşımlar arasındaki fark kabul edilebilir bir düzeyde ise doğrudan yaklaşımın uygulanmasını aksi halde dolaylı yaklaşımın uygulanmasını önermektedir (Otranto, Triacca, 2002:512). O halde doğrudan ve dolaylı yaklaşım arasında karar verebilmek için öncelikle bu yaklaşımlar arasında fark olup olmadığı araştırılmalıdır. Fakat hangi koşullarda bu yaklaşımlar birbirine yakın sonuç vermektedir? Eğer fark varsa bu farklılıklar neden ortaya çıkmaktadır? Bu yaklaşımlar arasında karar verebilmek için gerekli kriterler nelerdir? Bu sorulara yanıt verebilmek için yöntem bölümünün ilk alt bölümünde iki yaklaşım arasındaki farklılıkların hangi koşullarda oluştuğu ikinci alt bölümünde ise literatürde yer alan seçim kriterleri üzerinde durulmuştur.

2.1 Doğrudan ve Dolaylı Yaklaşım Arasında Fark Var mıdır?

Herhangi bir mevsimsel düzeltme işlemi zaman serilerini trend-çevrim, mevsim ve düzensiz bileşenlere ayırmaktadır. Örneğin, gözlenen bir zaman serisi gözlenemeyen bileşenlerine toplamsal bir modelde ayrıştırılabilir;

$$Y_t \equiv Y_t^{tc} + Y_t^s + Y_t^i \quad (1)$$

Bu gösterimde Y_t^{tc} , Y_t^s , Y_t^i sırasıyla trend-çevrim bileşeni, mevsimsel bileşen, ve düzensiz bileşen olarak adlandırılır. Bir çok durumda mevsimsel bileşen; tatiller, ticaret günü etkisi gibi deterministik etkiler tarafından büyür (Ghysels ve Osborne, 2001, 94). Bu deterministik etkileri hesaba katarsak; Y_t^{td} ticaret günü etkisi, Y_t^h ise tatil etkisi olmak üzere (1) numaralı gösterim yeniden yazılabilir;

$$Y_t \equiv Y_t^{tc} + Y_t^s + Y_t^i + Y_t^{td} + Y_t^h \quad (2)$$

Mevsimsel bileşen Y_t^s ve takvim etkileri ($Y_t^{td} + Y_t^h$), gözlenen zaman serisinden dışlandıktan sonra mevsimsel düzeltilmiş seri $Y_t^{ns} = Y_t^{tc} + Y_t^i$ elde edilir.

Gözlenen mevsimsel serinin Y_t olduğunu ve bu serinin mevsimsellik içeren X_t ve Z_t serilerinin toplamından oluştuğu ve gösterimde basitlik sağlamak için bu serilerin herhangi bir takvim etkisi taşımadığı varsayalım.

$$Y_t = X_t + Z_t, \quad t = 1, 2, \dots, T \quad (3)$$

Gözlenen Y_t , X_t ve Z_t serilerinin gözlenemeyen, mevsimsel ve mevsimsel olmayan iki bileşenden oluştuğu varsayalım:

$$Y_t = Y_t^{ns} + Y_t^s \quad (4)$$

$$X_t = X_t^{ns} + X_t^s \quad (5)$$

$$Z_t = Z_t^{ns} + Z_t^s \quad (6)$$

Bu gösterimde Y_t^{ns} , X_t^{ns} ve Z_t^{ns} ; trend-çevrim ve düzensiz bileşenleri içeren mevsimsel olmayan bileşenleri temsil ederken Y_t^s , X_t^s ve Z_t^s ise mevsimsel bileşenleri temsil etmektedir.

Mevsimsel düzeltme açısından arzu edilen özellik;

$$Y_t^{ns} = X_t^{ns} + Z_t^{ns} \quad (7)$$

olmasıdır (Otranto ve Triacca, 2002:513).

Fakat genellikle, toplulaştırılmış seri alt bileşenlerin doğrusal bir kombinasyonu olduğundan ve mevsimsel düzeltme ise doğrusal olmayan bir süreç olduğundan bu özellik sağlanamaz (Ladiray ve Mazzi, 2003:39). Diğer bir ifade ile mevsimsel düzeltmede doğrudan ve dolaylı yaklaşım farklı sonuçların elde edilmesine neden olabilir.

Mevsimsel düzeltmede kullanılan simetrik ve doğrusal filtre;

$$f(L) = \sum_{i=-n}^n f_i L^i \quad (8)$$

şeklinde gösterilsin. Bu gösterimde L ; $L^k z_t = z_{t-k}$, $k = \dots -1, 0, 1, \dots$ şeklinde işleyen gecikme işlemcisidir. Filtrenin bazı özelliklere sahip olması gerekir. Bunlar;

(i) $\sum_{i=-n}^n f_i = f(1) = 1$ olmalıdır. Yani filtrenin ağırlıklarının toplamının 1'e eşit olması gerekir. Bu özellik mevsimsel düzeltilmiş seri Y_t^{ns} ve orijinal serinin Y_t aynı seviyede olduğunu ima eder (Rietzler vd., 2000:4).

(ii) $f_{-i} = f_i$, $f = 1, 2, \dots, n$ olmalıdır. Bu özellik ise filtrenin simetrik olduğunu, yani Y_t^{ns} ve Y_t 'nin aynı fazda (evrede) olduğunu ima eder (Rietzler vd., 2000:4).

Doğrudan yaklaşımla mevsimsel düzeltme süreci ;

$$Y_t^{ns} = f(L)Y_t = \sum_{i=-n}^n f_i Y_{t-i} \quad (9)$$

$$\begin{aligned} Y_t^{ns} &= (X_t + Z_t)^{ns} = f(L)(X_t + Z_t) \\ &= f(L)X_t + f(L)Z_t \end{aligned} \quad (10)$$

şeklinde gösterilebilir.

Dolaylı yaklaşımla mevsimsel düzeltme süreci ise;

$$X_t^{ns} + Z_t^{ns} = f(L)X_t + f(L)Z_t \quad (11)$$

şeklinde gösterilebilir.

(10) ve (11) numaralı eşitlikler alt bileşenlerin tümü için aynı doğrusal filtrenin kullanılması durumunda doğrudan ve dolaylı yaklaşım arasında orijinal farklılık

gözlenmeyeceğini açık olarak göstermektedir. Rietzler vd. (2000) beş Avrupa ülkesine ait GSYİH verilerini kullandıkları çalışmalarında doğrudan ve dolaylı yaklaşım arasında uzun dönemde bir fark olup olmadığını eştümleşme analizi yaparak araştırmışlardır. Yazarlar seriler ister doğrudan ister dolaylı yaklaşımla düzeltilsin aynı doğrusal filtrenin kullanılması halinde uzun dönemde ve kısa dönemde bu yaklaşımlar arasında bir fark oluşmadığını göstermişlerdir.

Fakat doğrusal olmayan bir filtrenin kullanılması durumunda yukarıdaki sonuç geçerli olmayacaktır. $g(.)$ doğrusal olmayan bir dönüşümü göstermek üzere, doğrudan yaklaşımla mevsimsel düzeltme süreci;

$$\begin{aligned} Y_t^{ms} &= g(Y_t) \\ &= g(X_t + Z_t) \end{aligned} \quad (12)$$

şeklindedir. Dolaylı yaklaşımla mevsimsel düzeltme süreci ise;

$$X_t^{ms} + Z_t^{ms} = g(X_t) + g(Z_t) \quad (13)$$

şeklindedir.

$g(X_t + Z_t) \neq g(X_t) + g(Z_t)$ olduğundan doğrusal olmayan bir filtre kullanılması durumunda doğrudan ve dolaylı yaklaşım, farklı sonuçların elde edilmesine yol açabilir. Uygulamalı çalışmalar çok kısıtlı varsayımlar altında doğrudan ve dolaylı yaklaşımın birbirine yakın sonuçlar verdiğini göstermektedir. Bu varsayımlara göre serilerin herhangi bir uçdeğer içermediği, ayrıştırma işleminin toplamsal olduğu, tüm serilere aynı doğrusal filtrenin uygulandığı koşullarda sonuçlar arasında önemli bir farklılık oluşmamaktadır. Fakat bu koşullar uygulamada çok ender olarak sağlanabilmektedir. Ekonomik zaman serileri genellikle uç değerler içerir. Ayrıca toplamsal ayrıştırma bir çok zaman serisi için uygun bir ayrıştırma yöntemi değildir. En önemlisi model bazlı düzeltme yöntemlerinde uygulanan filtreler serilerin özelliklerine göre belirlenmektedir. Bu nedenle alt bileşenlerin tümü için aynı filtrenin uygulanması varsayımı çok güç sağlanabilmektedir (Ladiray ve Mazzi, 2003:39-40).

2.2 Doğrudan ve Dolaylı Yaklaşım: Seçim Problemi

Önceki bölümde anlatıldığı gibi mevsimsel düzeltmede doğrudan ve dolaylı yaklaşım farklı sonuçların elde edilmesine neden olabilir. O halde bu iki yaklaşım arasında nasıl seçim yapılacaktır? Mevsimsel bileşen gözlenemeyen bir bileşen olduğuna göre hangi yaklaşım daha doğru sonucu verecektir?

Findley ve Hood, eğer alt bileşenler birbirinden oldukça farklı mevsimsellik yapısı sergiliyorsa ve tüm alt bileşenler için yapılan mevsimsel düzeltme kabul edilebilir ise dolaylı yaklaşımla yapılan mevsimsel düzeltmenin kalitesinin doğrudan yaklaşıma göre daha yüksek olduğunu belirtmektedir (Hood ve Findley, 2001).

Ladiray ve Mazzi'ye göre, alt bileşenler birbirine benzer olmayan özellikler gösteriyorsa ve alt bileşenlerin göreceli önemi (ağırlıklar açısından) çok hızlı değişiyorsa dolaylı yaklaşım tercih edilmelidir (Ladiray ve Mazzi, 2003:40). Burada benzerlik kavramından kastedilen; bileşenlerin yatay ya da coğrafi (örneğin ülkeler) olarak toplulaştırılmasıdır. Bu durumda doğrudan yaklaşımın kullanılması daha uygun olur. Alternatif olarak bileşenler dikey ya da sektörel (örneğin sektörler, ürünler arasında) olarak toplulaştırılıyorsa dolaylı yaklaşım tercih edilmelidir.

Avrupa Merkez Bankası tarafından belirlenen kriterlere göre ise bu iki yaklaşım arasında önsel olarak seçim yaparken alt bileşenlerin özellikleri incelenmelidir (European Central Bank, 2000:11). Buna göre;

- Bileşenler birbirinden farklı mevsimsel yapı sergiliyorsa dolaylı yaklaşım kullanılmalıdır.
- Bileşenlerin veri kaynakları farklıysa dolaylı yaklaşım daha uygun olur.
- Bileşenler birbirinden farklı ticaret günü/çalışma günü etkileri gösteriyorsa dolaylı yaklaşım tercih edilmelidir.
- Bileşenler arasında yüksek korelasyon varsa doğrudan yaklaşım kullanılmalıdır.

Öte yandan sayılan kriterler önsel seçim kriterleri olup sadece bu kriterlere göre karar vermek her zaman yeterli olmayabilir. Örneğin Findley ve Hood, Amerikan Sayım Bürosu'nun verilerini kullanarak yapmış oldukları çalışmada tüm alt bileşenlere ait mevsimsel düzeltme kalite açısından kabul edilebilir olduğu halde toplulaştırılmış seride kalıntı mevsimsellik olabileceğini göstermişlerdir (Hood ve Findley, 2001). Bu nedenle her iki yaklaşımla elde edilen mevsimsel düzeltme sonuçlarının kalitesinin mutlaka değerlendirilmesi gerekir. Bununla ilgili literatürde pek çok kriter önerilmiştir. Dagum (1979) doğrudan ve dolaylı yaklaşım arasında karar verebilmek için düzgünlük (smoothness) kriterini önermiştir. Lothian ve Morry (1977) revizyon hatalarının küçüklüğünün bu seçimde önemli olduğunu vurgulamıştır. Planas ve Campolongo (2000) Euro Bölgesi'ndeki ülkelerin sanayi üretim endeksi serilerini kullanarak yapmış oldukları çalışmada eşzamanlı nihai tahmin hatalarını ve toplam revizyonları karşılaştırmışlardır. Peronaci (2003), Euro bölgesi M2 ve M3 para arzı göstergeleri için doğrudan ve dolaylı yaklaşımla elde ettiği mevsimsel düzeltme sonuçlarını serilerin düzgünlüğü, kalıntı mevsimsellik, revizyon geçmişi açısından karşılaştırmıştır.

2.2.1 Kalıntı Mevsimsellik

Mevsimsel düzeltmenin kalitesini değerlendirebilmek için gerekli en önemli koşul mevsimsellikten arındırılmış serilerde herhangi bir mevsimsel etkinin ya da takvim etkisinin gözlenmemesidir.

2.2.2 Düzgünlük

Mevsimsellikten arındırılmış serinin orijinal seriye göre daha az değişkenlik göstermesi gereklidir. Dagum (1979) serilerin düzgünlüğünü değerlendirebilmek için $L_2 - norm$ ölçümünü önermiştir.

$$L_2 = \frac{1}{N-1} \times \sum_{t=2}^T (SA_t - SA_{t-1})^2 = \frac{1}{N-1} \times \sum_{t=2}^T (\nabla SA_t)^2 \quad (14)$$

$\{SA_t : 1 \leq t \leq N\}$, burada SA_t mevsimsellikten arındırılmış seriyi, N ise serinin uzunluğunu göstermektedir. Daha küçük L_2 değeri daha düzgün bir seri anlamına gelmektedir.

2.2.3 Revizyon Geçmişi

Mevsimsel düzeltmenin kabul edilebilir olması, mevsimsellikten arındırılmış serinin kararlı bir yapı göstermesiyle mümkündür. Serinin kararlı olması demek yeni bir gözlemin eklenmesiyle mevsimsel düzeltme uygulanan serinin geçmişe yönelik

verilerindeki değişimin mümkün mertebe az olmasıdır. Revizyon hatalarının analizi kararlılık konusunda önemli bir göstergedir.

x_t 'nin sonlu bir seri olduğu ve $t=0, \dots, T$ sayıda gözlemin gerçekleşmiş olduğu varsayalım. Mevsimsel bileşen için tahmin edici $\hat{s}_{t/T}$ olsun. $t = T$ ise $\hat{s}_{t/t}$ 'ye eş zamanlı tahmin edici (Concurrent estimator) denir. $t < T$ ise $\hat{s}_{t/T}$ 'ye ön tahmin edici (Preliminary estimator) denir. $t > T$ ise T dönemindeki gözlemlen t dönemine kadar henüz gerçekleşmediğinden t-T dönem için öngörü değerleri kullanılır. $t \ll T$ ise (yani t, T'den çok küçükse) $\hat{s}_{t/T}$ yaklaşık olarak nihai tahmin ediciye (Final estimator) \hat{s}_t 'ye eşittir. T-t yeterince uzun bir zaman dilimi ise tüm tahminler için gerçekleşmiş gözlem değerleri kullanıldığından ön tahmin edici nihai tahmin ediciye \hat{s}_t 'ye yakınsar.

Mevsimsel düzeltmede simetrik filtreler kullanıldığından serilerin sonundaki gözlemler için öngörü değerleri kullanılmaktadır. Zaman serisine yeni gözlemlerin eklenmesiyle önceki öngörü değerlerinin yerini gerçekleşmiş gözlem değerleri almaktadır. Bu nedenle eş zamanlı tahmin edici $\hat{s}_{t/t}$, seriye her bir yeni gözlemin eklenmesiyle revize olur. Seriye her bir yeni gözlemin eklenmesiyle ön tahmin edicinin birbirinden farklı tahmin edilmesi revizyon problemine yol açar.

$$\hat{s}_{t/t} \neq \hat{s}_{t/t+1} \neq \hat{s}_{t/t+2} \dots \dots \dots \quad (15)$$

t döneminde nihai tahmin ediciden \hat{s}_t 'den eş zamanlı tahmin edici $\hat{s}_{t/t}$ 'ye kadar olan revizyon toplam revizyon hatası olarak adlandırılır ve;

$$r_t = \hat{s}_t - \hat{s}_{t/t} \quad (16)$$

şeklinde hesaplanır.

Toplam revizyon hatasının küçük olması ise mevsimsel düzeltmenin kalitesi konusunda son derece önemli bir göstergedir. Mevsimsel düzeltilmiş verilerde sık sık ve önemli ölçüde revizyonlar güvenilirliği zedeleyebilir.

3. BULGULAR

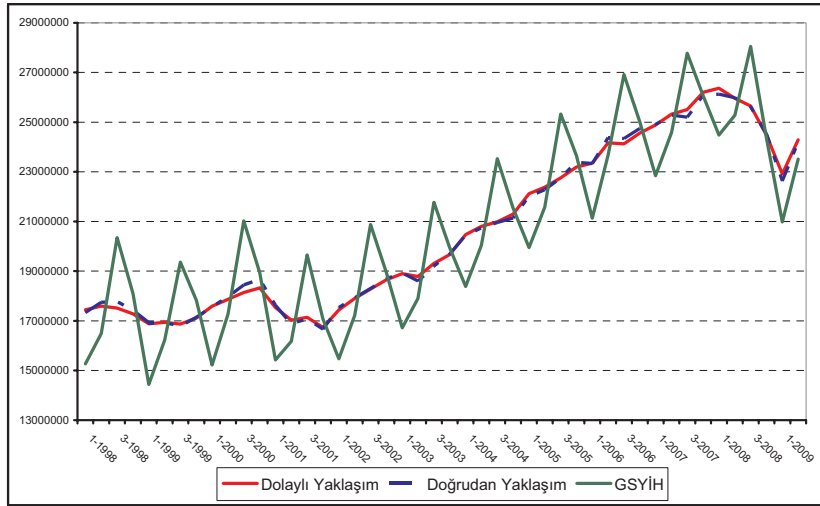
Bu bölümde doğrudan ve dolaylı mevsimsel düzeltme yaklaşımları, 1998 temel yıl fiyatlarıyla üçer aylık GSYİH serisi verileri kullanılarak TRAMO/SEATS mevsimsel düzeltme yöntemiyle karşılaştırılmıştır (Bkz, Ek 4). Gözlem dönemi 1998:1-2009:02 olup toplam gözlem sayısı 46'dır. Uygulamada Gianluca Caporello ve Agustin Maravall tarafından geliştirilen TSW Version Beta 1.04 (TRAMO/SEATS for Windows) paket programı kullanılmıştır.

Doğrudan ve dolaylı yaklaşım arasında önsel olarak karar verebilmek için öncelikle GSYİH serisinin alt bileşenlerinin mevsimsellik yapısı incelenmelidir. Alt bileşenler birbirinden farklı mevsim ve takvim etkileri taşıyorsa bu durum dolaylı yaklaşımın kullanılması için önemli bir gerekçedir. Bu nedenle bu bölümde önce alt bileşenlerin özellikleri incelenmiş daha sonra doğrudan ve dolaylı yaklaşımla mevsimsellikten arındırılan seriler arasında uzun dönemde bir fark olup olmadığını tespit edebilmek için eştümleşme analizi yapılmıştır. Son olarak elde edilen sonuçlar kalıntı mevsimsellik, düzgünlük ve revizyon geçmişi kriterleri doğrultusunda karşılaştırılmıştır.

İlk aşamada GSYİH serisi iktisadi faaliyet kollarına göre alt grupları itibariyle incelenmiş ve Konut sahipliği dışındaki alt grupların mevsimsellik taşıdığı tespit edilmiştir. Mevsimsellik içeren alt gruplar için uygun ARIMA modelleri³ seçildikten sonra ticari gün etkisi, hareketli tatil etkisi ve sabit tatil etkisi ayrı ayrı araştırılmıştır. Hareketli tatil etkisi için oluşturulan değişkenle ülkemizde Kurban ve Ramazan Bayramlarının; sabit tatil etkisi için oluşturulan değişkenle ise resmi tatillerin seriler üzerinde anlamlı etkisi olup olmadığı araştırılmıştır.⁴ Tatil etkileri için tek bir değişken oluşturulmamasının sebebi bunlardan ilkinin etkilerinin yıllara göre değişen tarihlerde ikincisinin ise sabit tarihlerde oluşmasıdır.

Ticari gün etkisi yalnızca Madencilik ve taşocakçılığı; İmalat sanayi ve Gayrimenkul kiralama ve iş faaliyetleri alt gruplarında anlamlı bulunmuştur (Bkz, Ek 2). Hareketli tatil etkisi İmalat sanayi; Elektrik, gaz, buhar ve sıcak su üretimi ve dağıtımı; Toptan ve perakende ticaret; Ulaştırma, depolama ve haberleşme; Kamu yönetimi ve savunma, zorunlu sosyal güvenlik; Sağlık işleri ve sosyal hizmetler; Vergi-Sübvansiyon alt gruplarında anlamlı bulunurken sabit tatil etkisi yalnızca İmalat sanayi alt grubunda anlamlı bulunmuştur (Bkz, Ek 2). Dolayısıyla İmalat sanayi alt grubunda hem ticari gün etkisi hem hareketli tatil etkisi hem de sabit tatil etkisi anlamlı bulunmuştur. Öte yandan toplululaştırılmış GSYİH serisinde yalnızca hareketli tatil etkisi anlamlı bulunmuştur. Bu sonuçlar GSYİH alt gruplarının birbirinden farklı takvim etkileri taşıdığını göstermektedir.

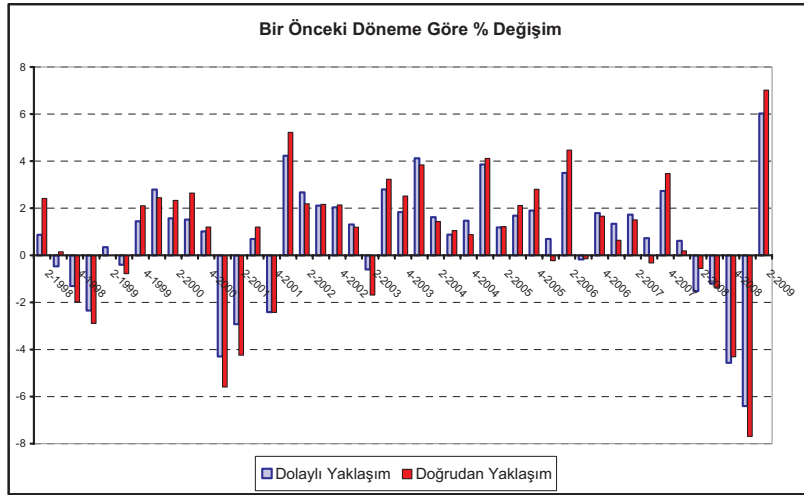
GSYİH serisi ve alt grupları için uygun modeller belirlendikten sonra uygulamanın bu aşamasında doğrudan ve dolaylı yaklaşımla mevsimsel düzeltme yapılmıştır. Sonuçlar Şekil 1 ve Şekil 2’de sunulmuştur.



Şekil 1. Dolaylı yaklaşımla ve doğrudan yaklaşımla düzeltilmiş GSYİH ve orijinal GSYİH serilerinin grafiği (Düzey)

³ ARIMA modelinde AR (p) ve MA (q) derecelerinin tespitinde BIC (Bayesyen Bilgi Kriteri) istatistiği kullanılmıştır. Uygun modelin belirlenmesinde ARIMA hatalarının diyagnostikleri ve ARIMA modelinin öngörü performansı dikkate alınmıştır.

⁴ Hareketli tatil için oluşturulan değişkende her bir çeyrek dönemde eğer varsa Kurban ve veya Ramazan bayramına denk gelen toplam gün sayısı kullanılmıştır. Aynı şekilde sabit tatil etkisi için oluşturulan değişkende her bir çeyrek dönemde eğer varsa resmi tatillere denk gelen gün sayısı dikkate alınmıştır.



Şekil 2. Doğrudan ve dolaylı yaklaşımla düzeltilmiş GSYİH serilerinin bir önceki döneme göre değişim oranları (%)

Doğrudan ve dolaylı yaklaşımla düzeltilmiş serileri düzey grafikleri incelendiğinde genel olarak birbirine çok yakın olmakla beraber 2000:3, 2000:4, 2006:2, 2008:1, 2009:1 gibi dönemlerde farkın açıldığı görülmektedir (Şekil 1). Doğrudan yaklaşım ve dolaylı yaklaşımla düzeltilmiş serilerin bir önceki döneme göre değişim oranları incelendiğinde önemli farklılıklar göze çarpmaktadır. Özellikle 1998:3, 1999:2, 2006:1 ve 2007:3 dönemlerindeki işaret farklılıkları dikkat çekmektedir (Şekil 2).

3.1 Eştümleşme Analizi

Alt bileşenlerin birbirinden oldukça farklı takvim etkileri taşıması dolaylı yaklaşımın seçilmesi için önemli bir gerekçedir. Diğer taraftan doğrudan yaklaşım dolaylı yaklaşıma göre oldukça kolay bir yöntemdir. Bu nedenle elde edilen sonuçlar arasında farkın önemli olmaması halinde doğrudan yaklaşımın kullanılması uygulama açısından daha avantajlıdır. Bu bölümde doğrudan ve dolaylı yaklaşımla elde edilmiş seriler arasında önemli bir fark olup olmadığını belirleyebilmek için Johansen Eştümleşme testi yapılmıştır. Doğrudan ve dolaylı yaklaşımla mevsimsellikten arındırılan seriler uzun dönemde birlikte hareket etmiyorsa yani aralarında bir eştümleşme ilişkisi mevcut değilse aralarındaki farkın önemli olduğu söylenebilir.

Johansen Eştümleşme analizinde iki test istatistiği kullanılmaktadır. Bunlardan ilki İz İstatistiği, diğeri ise Maksimum Özdeğer istatistiğidir. İz testinde “ r adet eşbütünleşik vektör vardır” yokluk hipotezi, “ n adet eşbütünleşik vektör vardır” alternatif hipotezine karşı test edilmektedir. Maksimum özdeğer testinde ise, “ r adet eşbütünleşik vektör vardır” yokluk hipotezi “ $r+1$ adet eşbütünleşik vektör vardır” alternatif hipotezine karşı test edilmektedir.

Johansen Eştümleşme testi sonuçları Tablo 1’de sunulmuştur. Eştümleşme testi için gecikme sayısı Olabilirlik Oranı (LR-Likelihood Ratio) kriteri yardımıyla beş olarak belirlenmiştir. $H_0 : r = 0$, doğrudan ve dolaylı yaklaşımla mevsimsellikten arındırılan seriler arasında eştümleşme ilişkisi yoktur hipotezi reddedilememiştir. Bu sonuç GSYİH serisi için uzun dönemde doğrudan ve dolaylı yaklaşım arasında farkın önemli olduğunu göstermektedir.

Tablo 1. Johansen eştümleşme testi sonuçları

Değişkenler	Hipotez H_0	Öz Değerler	İz İstatistiği*	Maksimum Özdeğer İstatistiği*
Doğrudan Yaklaşım	$r = 0$	0.179592	7.924244 [0.4735]	7.918154 [0.8341]
Dolaylı Yaklaşım	$r \leq 1$	0.000152	0.00609 [0.9371]	0.00609 [-0.9371]

Not: ADF test istatistiği sonuçlarına göre değişkenler birinci sıra fark durağan I (1) bulunmuştur (Bkz, Ek 3).

(*) Köşeli parantez içindeki değerler MacKinnon-Haug-Michelis (1999) p değerleridir.

3.2 Mevsimsel Düzeltmenin Kalitesininin Değerlendirilmesi

Bu bölümde doğrudan ve dolaylı yaklaşım arasında karar verebilmek için serilerin düzgünlüğü, kalıntı mevsimsellik olup olmadığı ve revizyon geçmişi incelenmiştir.

3.2.1 Kalıntı Mevsimsellik Açısından Değerlendirme

Mevsimsel düzeltmenin kalitesiyle ilgili en önemli koşul mevsimsellikten arındırılmış bir seride herhangi bir mevsimsel etkinin ya da takvim etkisinin gözlenmemesidir (Hood ve Findley, 2001:10; Peronaci, 2003:93). Bu amaçla doğrudan ve dolaylı yaklaşımla mevsimsellikten arındırılmış seriler yeniden TRAMO/SEATS yöntemiyle mevsimsel düzeltmeye tabi tutulmuş ve her iki seri için %95 anlamlılık düzeyinde anlamlı mevsimsellik içeren periyot sayısı sıfır bulunmuştur.

3.2.2 Düzgünlük (Smoothness) Açısından Değerlendirme

Mevsimsel dalgalanmalardan arındırılmış bir serinin orjinal seriye göre daha düzgün olması beklenir. Doğrudan ve dolaylı yaklaşım arasında düzgünlük kriteri açısından karar verebilmek için Dagum tarafından önerilen $L_2 - norm$ ölçümü hesaplanmıştır. Tablo 2’de sonuçlar sunulmuştur. Sonuçlar hem tüm gözlem dönemi için hem de son üç yıl analiz edildiğinde dolaylı yaklaşım lehinedir.

Tablo 2. Düzgünlük sonuçları

Analiz Dönemi	Doğrudan Mevsimsel Düzeltme	Dolaylı Mevsimsel Düzeltme
Tüm Gözlem Dönemi (N=46)	377.24×10^9	287.85×10^9
Son 3 yıl (N=12)	712.80×10^9	585.03×10^9

3.2.3 Revizyon Geçmişini Açısından Değerlendirme

Çalışmanın bu bölümünde doğrudan yaklaşımla ve dolaylı yaklaşımla mevsimsellikten arındırılmış serilerin revizyon geçmişini incelemiştir.

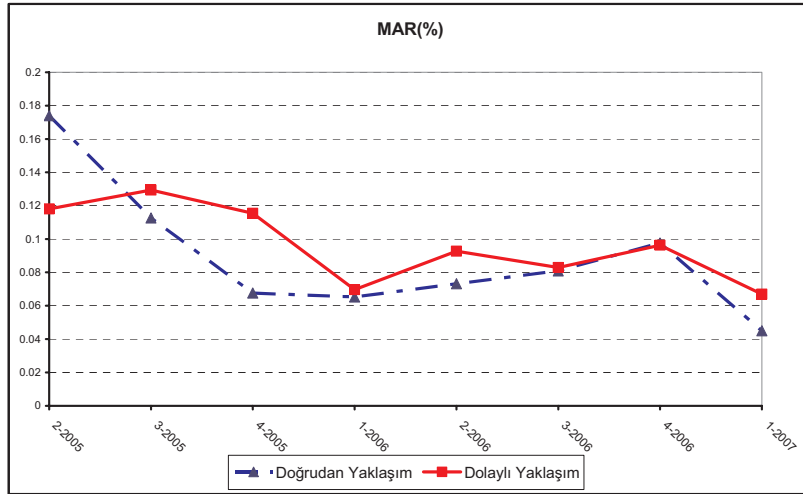
Mevsimsel düzeltmenin revizyonlar üzerindeki etkisini daha iyi değerlendirebilmek için uç değerler ve takvim etkilerinden arındırılmış olan doğrusallaştırılmış seriler kullanılmıştır. Analiz dönemi 1998:01-2009:2 tarihleri arası olup serilerden 16 gözlem çıkarılıp ARIMA modeli sabitlendikten sonra çıkarılan gözlemler bir bir eklenerek ardışık 16 tane doğrudan ve dolaylı yaklaşımla mevsimsel düzeltilmiş seri elde edilmiştir. Revizyon geçmişini analiz edebilmek için 2006:02-2007:01 alt dönemi

değerlendirmeye alınmıştır ve $MAR_t = \frac{1}{N} \sum_{k=1}^N AR_{t,k}$ formülü kullanılarak alt dönem içinde her bir üç aylık dönem için eşzamanlı düzeltmenin sekiz gecikme süresinde ortalama mutlak revizyon (Mean Absolute Revision-MAR) yüzdesi hesaplanmıştır. Burada $AR_{t,k}$ nihai tahmine göre mutlak revizyon yüzdesi olup;

$$AR_{t,k} = \left| 100 \times \frac{A_{t/T} - A_{t/t+k}}{A_{t/t+k}} \right| \quad (17)$$

şeklinde hesaplanmaktadır. $\{A_{t/t+k} : k = 0,1,\dots,N\}$ mevsimsellikten arındırılmış seriyi göstermekte olup $N=8$, $A_{t/t}$ eşzamanlı tahmin, $A_{t/t+k}$ ön tahmin ve $A_{t/T}$ nihai tahmindir. $T=2009:02$ dönemine kadar olan veri kullanılarak yapılan tahminin nihai tahmin olduğu varsayılmıştır. Eşzamanlı tahminin nihai tahmine yakınsama süreci serilerin ya da bileşenlerin stokastik yapısına bağlı olup bir çok seri için bu süre 2 ile 4 yıl arasında değişmektedir. 16 gözlemin çıkarılması yakınsamanın sağlanması için, geriye kalan 30 gözlem ise güvenilir bir mevsimsel düzeltme için gereklidir.

Sonuçlar Şekil 3'te sunulmuştur. Ortalama mutlak revizyonlar (MAR) her iki yaklaşım için de çok yüksek değildir. Üçüncü gecikmeden sonra revizyonlar %0,1'in altına düşmüştür. Revizyon geçmişi açısından değerlendirildiğinde doğrudan mevsimsel düzeltme ve dolaylı mevsimsel düzeltme yaklaşımları birbirleri üzerinde bir üstünlük sağlayamamıştır. Şekil 3'te görüldüğü üzere sonuçlar arasında önemli bir fark gözlenmemiştir.



Şekil 3. Doğrudan ve dolaylı yaklaşım MAR sonuçları

4. GENEL DEĞERLENDİRME VE SONUÇ

Bu çalışmada mevsimsel düzeltmede doğrudan yaklaşım ve dolaylı yaklaşım ele alınmıştır. Bu yaklaşımların arasında önemli bir fark olup olmadığı, (eğer varsa) bu yaklaşımlar arasında seçim yapabilmek için ne gibi ölçütler kullanıldığı konuları tartışılmıştır.

TÜİK tarafından yayınlanan 1998:01-2009:02 dönemine ait 1998 fiyatlarıyla GSYİH verileri kullanılarak doğrudan ve dolaylı yaklaşımla mevsimsel düzeltme yapılmış ve elde edilen sonuçlar karşılaştırılmıştır. Öncelikle GSYİH serisinin iktisadi faaliyet kollarına göre alt gruplarında mevsimsellik olup olmadığı incelenmiş ve 19 alt gruptan konut sahipliği dışında tüm alt gruplarda mevsimsellik tespit edilmiştir. Ayrıca serilerde takvim etkilerinden ticari gün etkisi, sabit tatil etkisi ve hareketli tatil etkisi araştırılmıştır. Hareketli tatil etkisinde Kurban ve Ramazan bayramlarının seriler üzerinde etkisi ve sabit tatil etkisinde resmi tatillerin seriler üzerindeki etkisi araştırılmıştır. Toplaştırılmış GSYİH serisinde yalnızca hareketli tatil etkisi anlamlı bulunurken toplam 3 alt grupta ticari gün etkisi, 7 alt grupta hareketli tatil etkisi, sadece 1 alt grupta ise sabit tatil etkisi anlamlı bulunmuştur. Özel olarak imalat sanayi alt grubunda bu etkilerin tamamı anlamlı bulunmuştur. Alt bileşenlerin birbirinden oldukça farklı takvim etkileri taşıdığı gözlenmiştir.

Özellikle ulusal istatistik ofisleri için mevsimsel düzeltmede kullanılan yöntemin teknik karmaşıklığından uzak ve kolay bir yöntem olması çok önemlidir. Bu çerçeveden bakıldığında iki yaklaşım arasında fark olmaması durumunda doğrudan yaklaşımın kullanılması önerilmektedir. Doğrudan ve dolaylı yaklaşımla mevsimsellikten arındırılan seriler arasında uzun dönemde önemli bir fark olup olmadığını tespit etmek için Johansen Eştleme testi yapılmıştır ve serilerin uzun dönemde birlikte hareket etmediği, aralarındaki farkın önemli olduğu sonucuna varılmıştır.

Son olarak elde edilen sonuçlar kalıntı mevsimsellik, düzensizlik ve revizyon geçmişi açısından karşılaştırılmıştır. Her iki yaklaşımla mevsimsellikten arındırılmış serilerde kalıntı mevsimsellik gözlenmemiştir. Serilerin düzensizliği açısından değerlendirme yapıldığında hem son üç yıl için hem de tüm gözlem dönemi için dolaylı yaklaşım öne

çıkmiştir. Son olarak doğrudan ve dolaylı yaklaşımla mevsimsellikten arındırılmış serilerin revizyon geçmişleri incelenmiştir. Her iki yaklaşım için de revizyonlar küçük olup aralarında önemli bir fark tespit edilmemiştir.

Doğrudan ve dolaylı yaklaşım sonucunda elde edilen GSYİH serileri arasındaki farkın önemli bulunması, alt bileşenlerin birbirinden oldukça farklı takvim etkileri taşıması, dolaylı yaklaşımla elde edilen sonuçların düzgünlük kriteri açısından daha tatmin edici olması nedeniyle GSYİH serisinin mevsimsellikten arındırılmasında dolaylı yaklaşımın kullanılması önerilmektedir.

Öte yandan mevsimsel düzeltmede doğrudan ve dolaylı yaklaşım arasındaki seçim problemi alt bileşenlerin toplulaştırılması yoluyla elde edilen tüm zaman serileri için geçerlidir. Araştırmacıların kullanacakları yöntemi belirlerken sadece önsel kriterlere göre değil belirtilen aşamaları takip edip, sonuçlarını değerlendirip karar vermelerinin daha faydalı olacağı düşünülmektedir.

5. KAYNAKLAR

Bell, W. R., Hillmer, S. C., 1984. Issues Involved with the Seasonal Adjustment of Economic Time Series. *Journal of Business & Economic Statistics*, American Statistical Association, 2(4): 291-320.

Burman, J. P., 1980. Seasonal Adjustment by Signal Extraction. *Journal of the Royal Statistical Society A*, 143:321-337.

European Central Bank, 2000. Seasonal Adjustment of Monetary Aggregates and HICP for the Euro Area. <http://www.ecb.int/pub/pdf/other/sama0008en.pdf>, 21 Mayıs 2009.

Dagum, E. B., 1979. On the Seasonal Adjustment of economic Time Series Aggregates: A Case Study of the Unemployment Rate, Counting the Labor Force, National Commion Employment and Unemployment Statistics. Appendix, 2:317-344.

Dosse, J., Planas, C., 1996. Pre-adjustment in Seasonal Adjustment Methods: A Comparison of REGARIMA and TRAMO. Eurostat Working Group Document, No: D3/SA/07.

Findley D. F., Soukup R. J., 2007. Detection And Modeling Of Trading Day Effects. U.S. Census Bureau Statistical Research Division, http://www.census.gov/ts/papers/ices00_td.pdf, 23 Aralık 2008.

Fischer, B., 1995. Decomposition of Time Series - Comparing Different Methods in Theory and Practice. Eurostat Working Paper, No 9/1998/A/8.

Ghysels, E., 1997. Seasonal adjustment and other data transformations. *Journal of Business and Economic Statistics* 15, 410-418.

Ghysels, E., Osborn, D. R., 2001. *The Econometric Analysis of Seasonal Time Series*. Cambridge University Press, 978-0-521-56588-2.

Gomez, V., Maravall, A., 1997. Program TRAMO and SEATS: Instructions for the User. Beta Version, (Online), Banco de Espana.

Gomez, V., Maravall A., 1998. Seasonal Adjustment and Signal Extraction in Economic Time Series. Banco de Espana, No 9809.

Hood, C. C. H., Findley, D. F., 2001. Comparing Direct and Indirect Seasonal Adjustment of Aggregate Series. 2001 Proceedings of the American Statistical Association.

Hylleberg, S., 2006. Seasonal Adjustment. University of Aarhus Department of Economics, Working Paper No. 2006-4.

Kaiser, R., Maravall, A., 2000. Notes on Time Series Analysis, ARIMA Models and Signal Extraction. Banco de Espana, No 0012.

Koçak, N. A., 2009. Takvim Etkileri: Ulusal Hesaplar Uygulaması. 17. İstatistik Araştırma Sempozyumu Bildiriler Kitabı, 2008, TÜİK, 154-168.

Ladiray, D., Quenneville, B., 2001. Seasonal Adjustment with the X-11 Method. V: 158, Springer , New York , 9780387951713.

Ladiray, D., Mazzi, G. L., 2003. Seasonal Adjustment of European Aggregates: Direct versus Indirect Approach. Proceedings of the Seminar on Seasonal Adjustment, Manna, M., Peronaci, R. editors, European Central Bank, 37-66.

Lothian, J., Morry, M., 1977. The Problem of Aggregation; Direct or Indirect. Working Paper, Time Series Research and Analysis Division, Statistics Canada, Ottawa ON, Canada.

Maravall, A., 2005. Brief Description of the Programs.
<http://www.bde.es/webbde/es/secciones/servicio/software/tramo/summprogs.pdf>.

Maravall, A., 2006. An application of the TRAMO-SEATS automatic procedure; direct versus indirect adjustment. Computational Statistics & Data Analysis, Vol.50, No.9, 2167-2190.

Maurin, L., 2003. Criteria to determine the optimal revision policy: A case study based on euro zone monetary aggregates data. Proceedings of the Seminar on Seasonal Adjustment, Manna, M., Peronaci, R. editors, European Central Bank, 67-84.

Otranto, E., Triacca, U., 2002. Measure to Evaluate the discrepancy Between Direct and Indirect Model-Based Seasonal Adjustment. Journal of Official Statistics, Vol 18, No 4, pp. 511-530.

Peronaci, R., 2003. The Seasonal Adjustment of Euro Area Monetary Aggregates: Direct Versus Indirect Approach. Proceedings of the Seminar on Seasonal Adjustment, Manna, M., Peronaci, R. editors, European Central Bank, 91-108.

Planas, C., 1997. Applied Time Series Analysis: Modeling, Forecasting, Unobserved Components Analysis and the Wiener-Kolmogorov Filter. Eurostat Working Paper, No:66630002.

Planas, C., Campolongo, F., 2000. The Seasonal Adjustment of Contemporaneously Aggregated Series. European Commission Working Paper, No:KS-AN-03-018-EN-N.

Rietzler, K., Stephan, S., Wolters, J., 2000. Aggregation and Seasonal Adjustment Empirical Results for EMU Quarterly National Accounts. Discussion Papers of DIW Berlin from DIW Berlin, No 228, German Institute for Economic Research.

DIRECT AND INDIRECT SEASONAL ADJUSTMENT APPROACHES: AN APPLICATION FOR GDP SERIES

ABSTRACT

The seasonal adjustment of aggregated time series can be obtained indirectly through the aggregation of the main components previously seasonally adjusted or directly seasonally adjusting the aggregated series. Unfortunately, these strategies could produce quite different results. In this paper we discussed the common criteria and diagnostics used to assess the quality of a seasonal adjustment and to compare seasonal adjustment results obtained with the alternative approaches known as direct or indirect in the literature. The direct and the indirect seasonal adjustment approaches are compared on the basis of the results obtained for the seasonal adjustment of the GDP series published by the Turkish Statistical Institute. In this study we have used the quarterly data pertaining to 1998:01–2009:02 period, and have concluded that the GDP is to be adjusted using the indirect approach.

Keywords: Direct and indirect approach, Seasonal adjustment, Time series.

Ek 1. İktisadi Faaliyet Kollarına Göre GSYİH

- 1 Tarım, avcılık ve ormancılık
- 2 Balıkçılık
- 3 Madencilik ve Taşocakçılığı
- 4 İmalat Sanayi
- 5 Elektrik, gaz, buhar ve sıcak su üretimi ve dağıtımı
- 6 İnşaat
- 7 Toptan ve perakende ticaret
- 8 Oteller ve Lokantalar
- 9 Ulaştırma, depolama ve haberleşme
- 10 Mali aracı kuruluşların faaliyetleri
- 11 Konut Sahipliği
- 12 Gayrimenkul, kiralama ve iş faaliyetleri
- 13 Kamu yönetimi ve savunma, zorunlu sosyal güvenlik
- 14 Eğitim
- 15 Sağlık işleri ve sosyal hizmetler
- 16 Diğer sosyal, toplumsal ve kişisel hizmet faaliyetleri
- 17 Eviçi personel çalıştıran hanehalkları
- 18 (Eksi) Dolaylı ölçülen mali aracılık hizmetleri
- 19 Vergi-Sübvansiyon

Ek 2. Deterministik Etkiler

Alt Grup	ARIMA Modeli** (p,d,q)(P,D,Q)s	Log Dönüşümü	Ticari Gün Etkisi	Hareketli Tatil Etkisi*	Sabit Tatil Etkisi*
Tarım, avcılık ve ormancılık	(1,0,0)(0,1,1)s	Evet	Hayır	Hayır	Hayır
Balıkçılık	(0,0,0)(0,1,0)s	Evet	Hayır	Hayır	Hayır
Madencilik ve Taşocakçılığı	(1,0,0)(0,1,0)s	Evet	Evet	Hayır	Hayır
İmalat Sanayi	(1,1,0)(0,1,0)s	Evet	Evet	-0.0502 [-4.90]	-0.0071 [-2.62]
Elektrik, gaz, buhar ve sıcak su üretimi ve dağıtımı	(0,1,0)(0,1,1)s	Evet	Hayır	-0.0039 [-1.97]	Hayır
İnşaat	(1,1,0)(0,1,1)s	Hayır	Hayır	Hayır	Hayır
Toptan ve perakende ticaret	(0,1,0)(0,1,1)s	Hayır	Hayır	-24.816 [-2.88]	Hayır
Oteller ve Lokantala	(0,1,0)(0,1,1)s	Evet	Hayır	Hayır	Hayır
Ulaştırma, depolama ve haberleşme	(0,1,0)(0,1,1)s	Evet	Hayır	-0.0071 [-2.92]	Hayır
Mali araç kuruluşların faaliyetleri	(0,1,0)(0,1,1)s	Hayır	Hayır	Hayır	Hayır
Konut Sahipliği	(0,1,1)(0,1,1)s	Hayır	Hayır	Hayır	Hayır
Gayrimenkul, kiralama ve iş faaliyetleri	(0,1,0)(0,1,1)s	Evet	Evet	Hayır	Hayır
Kamu yönetimi ve savunma, zorunlu sosyal güvenlik	(0,1,1)(0,1,1)s	Evet	Hayır	-0.0037 [-2.10]	Hayır
Eğitim	(1,0,0)(0,1,1)s	Evet	Hayır	Hayır	Hayır
Sağlık işleri ve sosyal hizmetler	(1,0,0)(0,1,1)s	Evet	Hayır	-0.0039 [-2.29]	Hayır
Diğer sosyal, toplumsal ve kişisel hizmet faaliyetleri	(1,0,0)(0,1,0)s	Evet	Hayır	Hayır	Hayır
Eviçi personel çalıştıran hanehalkları	(0,1,0)(0,1,1)s	Evet	Hayır	Hayır	Hayır
Dolaylı ölçülen mali aracılık hizmetleri	(0,1,0)(0,1,1)s	Evet	Hayır	Hayır	Hayır
Vergi-Sübvansiyon	(0,1,0)(0,1,1)s	Evet	Hayır	-0.0126 [-4.00]	Hayır
Gayri Safi Yurtiçi Hasıla	(0,1,0)(0,1,1)s	Hayır	Hayır	-0.1402 [-3.53]	Hayır

(*) Köşeli parantez içindeki değerler t istatistiği değerleridir.

(**) s=4 yıl içindeki gözlem sayısını; p, standart AR (Otoregresif) polinomunun; P, mevsimsel AR polinomunun; q, standart MA (Hareketli Ortalama) polinomunun; Q, mevsimsel MA polinomunun derecesini; d ve D serinin durağanlığının sağlanabilmesi için sırasıyla kaçınıcı dereceden standart ve mevsimsel farkların alındığını göstermektedir.

Ek 3. Genişletilmiş Dickey Fuller (Augmented Dickey Fuller-ADF) Testi Sonuçları

Değişkenler*	k	ADF Test İstatistiği (τ)	
		(Trendsiz)	(Trendli)
GSYİH(ID)	0	-0.336399 [0.9110]	0 -2.425400[0.36210]
GSYİH(D)	0	-0.529835 [0.8755]	0 -1.757545 [0.7085]
Δ GSYİH(ID)	0	-4.463746 [0.0008]	0 -4.397114 [0.0056]
Δ GSYİH(D)	0	-5.359640 [0.0001]	0 -5.258810 [0.0005]

Not: Parantez içindeki değerler MacKinnon p değerleridir. k, Schwartz bilgi kriterine göre uygun gecikme sayısıdır.

(*) GSYİH (ID), dolaylı yaklaşımla mevsimsel düzeltilmiş GSYİH'yı; GSYİH (D), doğrudan yaklaşımla mevsimsel düzeltilmiş GSYİH'yı temsil etmektedir.

Ek 4. TRAMO/SEATS Mevsimsel Düzeltme Yöntemi Hakkında Özet Bilgi

TRAMO/SEATS mevsimsel düzeltme yöntemi iki aşamalı olup birinci aşama olan TRAMO'da ham veriler doğrusallaştırılmakta, ikinci aşama olan SEATS aşamasında TRAMO'dan SEATS'e aktarılan doğrusallaştırılmış seriler bileşenlerine ayrıştırılmaktadır. Dolayısıyla TRAMO aşaması mevsimsel düzeltmeden önce yapılan bir ön düzeltme niteliği taşımaktadır. Ön düzeltmede serilere uygun dönüşümler yapıldıktan sonra otomatik olarak öncelikle uç değerler ya da diğer bozucu etmenler tespit edilip düzeltilmektedir. Kayıp gözlem varsa tahmin edildikten sonra takvim etkileri gibi önsel düzeltme seçenekleri arasında seçim yapılmaktadır. Ön düzeltmede, ayrıştırma aşamasında simetrik filtrelerin kullanılabilmesi için seriler ARIMA modelleriyle genişletilmektedir. Ön düzeltmede doğrusallaştırılan ve genişletilen seriler SEATS aşamasında bileşenlerine ayrıştırılmaktadır.

TRAMO'da gözlenen serilerin deterministik ve stokastik olmak üzere iki kısımdan oluştuğu varsayılmaktadır.

$$y_t = z_t' b + x_t \tag{1}$$

y_t modellenen ham seri, z_t ön düzeltmede kullanılan açıklayıcı değişkenler matrisi ve b regresyon katsayıları vektörüdür. (1) numaralı modelin ilk kısmı deterministik kısım olup deterministik etkilerden arındırılan x_t , doğrusallaştırılmış seridir ve ARIMA süreci izlemektedir.

$$\delta(B)\phi(B)\Phi_s(B^s)x_t = \theta(B)\Theta_s(B^s)a_t \tag{2}$$

Bu gösterimde B , $Bx_t = x_{t-1}$ şeklinde işleyen gecikme işlemcisidir. s yıl içindeki gözlem sayısını göstermek üzere $\phi(B)$ ve $\Phi_s(B^s)$ sırasıyla durağan olan p. dereceden düzenli ve P. dereceden mevsimsel AR (Otoregresif) polinomlarını, $\delta(B)$ durağan olmayan AR polinomunu, $\theta(B)$ ve $\Theta_s(B^s)$ sırasıyla tersi alınabilir q. dereceden düzenli ve Q. dereceden mevsimsel MA (Hareketli Ortalama) polinomlarını göstermektedir. a_t ise sıfır ortalama ve sabit varyansa sahip, bağımsız, özdeş dağılılan hata terimini göstermektedir. Bu özelliklere sahip hata terimi beyaz gürültü (white noise-wn) olarak adlandırılır ve

$$a_t \approx wn(0, V_a) \tag{3}$$

şeklinde gösterilir. (2) numaralı gösterimde;

$$\delta(B) = \Delta^d \Delta_s^D = (1-B)^d (1-B^s)^D \text{ 'dir.} \tag{4}$$

Δ^d düzenli fark işlemcisi ve Δ_s^D mevsimsel fark işlemcisidir. Gösterimde d ve D serinin durağanlığının sağlanabilmesi için sırasıyla kaçınıcı dereceden standart ve mevsimsel farkların alındığını göstermektedir. Ön düzeltmede yer alan açıklayıcı değişkenler dört temel başlık altında toplanabilir. Bunlar; sabit terim, uç değerler, takvim etkileri ve kullanıcı tanımlı etkilerdir.

TRAMO'da doğrusallaştırılan x_t serisi SEATS'e aktarılarak gerçek ayrıştırma yapılır. SEATS, ARIMA serilerinde sinyal çıkarma (signal extraction) anlamına gelmektedir. 1980'de Burman tarafından geliştirilen sinyal çıkarma yaklaşımı ARIMA model bazlı (AMB) yaklaşım olarak bilinir. AMB yaklaşımında gözlenemeyen bileşenler için yapılan varsayımlar şunlardır:

- (i) Gözlenemeyen bileşenler ortogondur. Bu varsayım farklı bileşenlerin gelişiminin farklı kaynaklardan yaratılan etkilere bağlı olduğunu göstermektedir.
- (ii) İki farklı bileşen aynı AR köküne sahip değildir.
- (iii) Gözlenen serilerin (x_t) izlediği ARIMA modeli bilinmekteyken gözlenemeyen bileşenler için modeller bilinmemektedir (Planas,1997).

Doğrusallaştırılmış x_t serisinin toplamsal olarak modellendiği varsayılın

$$x_t = s_t + p_t + c_t + u_t \quad (5)$$

Gösterimde kolaylık sağlamak için x_t serisi daha genel olarak iki bileşene ayrıştırılın:

$$x_t = s_t + n_t \quad (6)$$

n_t mevsimsel düzeltilmiş seri olup s_t mevsimsel bileşendir x_t kapalı ARIMA notasyonuyla;

$$\Phi(B)x_t = \theta(B)a_t, \quad a_t \approx w.n(0, V_a) \quad (7)$$

Burada $\Phi(B) = \phi_n(B)\phi_s(B)$ durağan olmayan AR köklerini de içermektedir. Hareketli ortalama gösterimi;

$$x_t = \frac{\theta(B)}{\Phi(B)} a_t \quad (8)$$

$$x_t = \frac{\theta(B)}{\Phi(B)} a_t = \frac{\theta_n(B)}{\phi_n(B)} a_{nt} + \frac{\theta_s(B)}{\phi_s(B)} a_{st} + u_t \quad (9)$$

$$u_t \approx w.n(0, V_u) \quad (10)$$

Dolayısıyla bileşenler için modeller ;

$$\phi_s(B)s_t = \theta_s(B)a_{st} \quad a_{st} \approx w.n(0, V_s) \quad (11)$$

$$\phi_p(B)n_t = \theta_n(B)a_{nt} \quad a_{nt} \approx w.n(0, V_p) \quad (12)$$

SEATS doğrusallaştırılmış x_t serisini bileşenlerine ayrıştırırken sonsuz sayıda seçeneğe sahiptir. Seçenekler arasındaki fark beyaz gürültünün bileşenler arasında tahsisine göre belirlenir. Kanonik ayrıştırma olarak adlandırılan yönteme göre düzensiz bileşenin varyansı enbüyüklenirken diğer bileşenlerin varyansı enküçüklenir. Belirlenme problemi bu şekilde aşılar. Böylece daha düzgün (smooth) trend ve mevsim bileşeni elde edilir.

Bileşenler için belirlenen modeller teorik modeller olup tahmin edilmesi gerekmektedir. Gözlenemeyen bileşenler için modeller belirlendikten sonra serilerin yapısına göre optimal bir filtre oluşturularak bileşenler tahmin edilir. Optimalden kasıt hata kareler ortalamasının minimizasyonudur. Minimum hata kare ortalaması (MMSE) tahmini, gözlenemeyen s_t 'nin koşullu beklenen değeri alınarak elde edilir. s_t 'nin MMSE tahmincisi ;

$$\hat{s}_t = E(s_t / X) \tag{13}$$

Burada s_t gerçek sinyal \hat{s}_t ise onun tahminidir. Koşullu beklenen değer Wiener Kolmogorov (WK) filtresiyle hesaplanabilir. Wiener Kolmogorov filtresinin işleyişi aşağıdaki gibidir.

$$X = [x_{-\infty} \dots x_t \dots x_{\infty}] \text{ ve } (F = B^{-1}; F^j x_t = x_{t+j}) \text{ ise} \tag{14}$$

$$\hat{s}_t = \left[\nu_0 + \sum_{j=1}^{\infty} \nu_j (B^j + F^j) \right] x_t \tag{15}$$

$$\hat{s}_t = \nu_0 x_t + \nu_1 (x_{t+1} + x_{t-1}) + \nu_2 (x_{t+2} + x_{t-2}) + \dots \tag{16}$$

$$\hat{s}_t = \nu(B, F)x_t \tag{17}$$

$\nu(B, F)$ Wiener-Kolmogorov Filtresidir (bundan sonra WK). WK filtresi tutarlı, merkezi, x_{t+j} ve x_{t-j} için aynı ağırlıklar kullanıldığından simetriktr. Örneğin mevsimsel bileşen s_t tahmin edilmek istensin. Bu durumda s_t sinyal olup s_t 'nin tahmini \hat{s}_t , bileşenlerin hareketli ortalama gösteriminden elde edilir. Gözlenen seri x_t ve sinyal s_t için modellerin MA gösterimi;

$$s_t = \Psi_s(B)a_{st}; \quad \Psi_s(B) = \frac{\theta_s(B)}{\phi_s(B)}; \tag{18}$$

$$x_t = \Psi(B)a_t; \quad \Psi(B) = \frac{\theta(B)}{\phi(B)} \tag{19}$$

şeklindedir.

$$\hat{s}_t = \underbrace{\left[\frac{V_s}{V_a} \frac{\Psi_s(B)\Psi_s(F)}{\Psi(B)\Psi(F)} \right]}_{WK \text{ Filtresi}} x_t \tag{20}$$

şeklinde elde edilir.

$\Phi(B) = \phi_n(B)\phi_s(B)$ olduğundan (20) numaralı eşitlik yeniden düzenlendiğinde mevsim bileşeni için WK filtresi;

$$\nu(B, F) = \frac{V_s}{V_a} \frac{\theta_s(B)\phi_n(B)}{\theta(B)} \frac{\theta_s(F)\phi_n(B)}{\theta(F)} \tag{21}$$

Dolayısıyla WK filtresi bileşenlerin özelliklerine göre belirlenir. WK filtresi gözlem sayısının sonsuz olduğu varsayımına dayanmakla beraber belirli bir gecikmeden sonra sifıra yakınsamaktadır. Filtrenin uzunluğu paydadaki θ parametreleri tarafından belirlenir. Bu parametreler sifıra ne kadar yakınsa yakınsama o kadar çabuk gerçekleşir.

SAĞA ÇARPIK DAĞILIM ORTALAMALARI İÇİN BAZI TESTLERİN KULLANIMI VE KARŞILAŞTIRMALARI

Emre Erçin SARISOY*

Hamza GAMGAM**

ÖZET

Bu makalenin amacı, sağa çarpık dağılımlardan çekilen örneklemeler ile kitle ortalamasına ilişkin yokluk hipotezini test etmek için kullanılan Johnson'ın düzeltilmiş karesel t testini, Sutton'ın bileşik testini ve Chen'in testini tanıtmak, ayrıca bunları 1. tip hata oranları ve testin gücü bakımından karşılaştırmaktır. Bu amaçla bazı sağa çarpık dağılımlardan üretilen veri kümeleri üzerinden simülasyon çalışması yapılmış ve Chen'in testinin daha güçlü olduğu tespit edilmiştir.

Anahtar Kelimeler: Chen'in testi, Çarpıklık, Johnson'ın düzeltilmiş karesel t testi, Sutton'ın bileşik testi, Testin gücü.

1. GİRİŞ

Kitle ortalamasına ilişkin yokluk hipotezinin test edilmesinde kullanılan klasik yöntemler Z ve Student t testleridir. Bu test istatistiklerini kullanmak için örneklemin normal dağılıma sahip bir kitleden geldiği varsayımı geçerli olmalıdır.

Merkezi Limit Teoremi'ne göre normal dağılıma sahip olmayan bir kitleden rasgele örneklem seçildiğinde, örneklem büyüklüğü (n) yeteri kadar büyük iken örneklem ortalamasının dağılımı μ ortalama ve σ^2/n varyans ile normal dağılıma yakınsar. Dolayısıyla, dağılımı ne olursa olsun herhangi bir kitleden seçilen örneklem büyüklüğü yeterince büyük olan örneklem ile kitle ortalamasına ilişkin yokluk hipotezinin testinde $Z \sim N(0,1)$ istatistiği kullanılır. Eğer örneklemin geldiği kitle normal dağılıma sahip ve kitle varyansı (σ^2) bilinmiyor ise, bu istatistik yerine Student (1908)'in önermiş olduğu $t \sim t_{n-1}$ istatistiği kullanılır. Örnek hacminin küçük ve örneklemin çekildiği kitle normal dağılımı çarpık iken, kitle ortalamasına ilişkin yokluk hipotezinin testinde bu test istatistiklerini kullanmak doğru olmaz.

Örnek hacmi yeteri kadar büyük değilken, normal dağılıma sahip olmayan kitleler için ortalamaya ilişkin hipotez testleri ve güven aralıkları hep tartışma konusu olmuştur. Sophister (1928), Neyman ve Pearson (1928), Nair (1941) çarpıklığın basıklığa göre t istatistiğini daha çok etkilediğini göstermişlerdir. Rider (1929), Geary (1936), Laderman (1939), Perlo (1933) ve Anscombe (1950)'un çarpık dağılımlardan çekilen örneklemeler ile kitle ortalamasına ilişkin yokluk hipotezini test etmeye yönelik çalışmaları, t istatistiğinde bir düzeltme yapmaya yöneliktir (Johnson, 1978).

Box ve Andersen (1955) parametrik olmayan yöntemlerin kullanımına dayanan çalışmalar yapmışlardır. Ayrıca, Tukey (1964), Andrews vd. (1972) ve Yuen (1974)'in budama yöntemine dayanan çalışmaları vardır (Johnson, 1978).

* TÜİK Uzmanı, Türkiye İstatistik Kurumu, Ankara, e-posta: emresarisoy@tuik.gov.tr

**Prof. Dr., Gazi Üniversitesi, GÜFEF, İstatistik Bölümü, Ankara, e-posta: gamgam@gazi.edu.tr

Kitle normal dağılıma sahip değil ve örnek hacmi küçük iken Bartlett (1935), Chung (1946) ve Gayen (1949) Edgeworth ve Gram-Charlier açılımlarını kullanarak çeşitli düzeltilmiş t istatistikleri tanımlamışlardır.

Bu çalışmanın ikinci bölümünde sağa çarpık dağılımların ortalaması ile ilgili yokluk hipotezlerinin testi için Johnson'ın düzeltilmiş karesel t testi, Sutton'ın bileşik testi ve Chen'in testi tanıtılacak ve bu testlerin "deneysel 1. tip hata oranları" ve "testin gücü değerleri" bakımından karşılaştırılması için yapılan simülasyon çalışması hakkında bilgi verilecektir. Üçüncü bölümde ise simülasyon çalışması sonucunda elde edilen bulgulara yer verilecektir. Dördüncü bölümde, çalışmanın genel sonuçları değerlendirilecek ve bazı önerilerde bulunulacaktır.

2. YÖNTEM

2.1 Johnson'ın Düzeltilmiş Karesel t Testi

Bir X rastgele değişkeni için Cornish ve Fisher (1937) Açılımı, Eşitlik 1'de verilen biçimde ifade edilir:

$$CF(X) = \mu + \sigma\zeta + (\mu_3 / \sigma^2)(\mu_3\zeta^2 - 1) + \dots \quad (1)$$

Burada ζ standart normal rastgele değişkeni, μ parametresi X rastgele değişkeninin ortalamasını, σ^2 bu değişkenin varyansını ve μ_3 ise bu değişkenin üçüncü merkezl momentini ifade etmektedir. Johnson'ın düzeltilmiş karesel t istatistiğinin türetilmesi Wallace (1958)'in CF yaklaşımı ile ilişkilidir (Johnson, 1978).

Johnson (1978)'in çalışmasında Eşitlik 1'deki tüm momentler ele alınmamıştır ve seri belirli bir terimden sonra devam ettirilmemiştir. Eşitlik 1 ifadesi \bar{X} 'ya göre düzenlenirse;

$$CF(\bar{X}) = \mu + \frac{\sigma}{\sqrt{n}}\zeta + (\frac{\mu_3}{6n\sigma^2})(\zeta^2 - 1) + o(n^{-\frac{3}{2}}) \quad (2)$$

eşitliği elde edilir (Johnson, 1978). Burada $o(n^{-\frac{3}{2}})$ terimi, ilk iki terimden sonraki terimlerin $n^{-\frac{3}{2}}$ terimine bölümü ile elde edilen serinin sıfıra yakınsadığını göstermektedir. Eşitlik 2'deki $CF(\bar{X})$ ifadesi, Cornish-Fisher açılımının iki terimli bir gösterimidir.

Johnson (1978), \bar{X} 'nin momentlerinden faydalanarak yeni bir test istatistiği türetmiştir. Bunun için ilk adım olarak Eşitlik 1'den Eşitlik 2 elde edilmiştir. t istatistiğine ilişkin düzeltmeyi türetmek için bir başka ifade ile, düzeltilmiş t istatistiğini elde etmek için bu açımdan $\zeta^2 - 1$ terimi çıkartılmıştır. Daha sonraki terimlerin çarpıklığı azaltmadaki etkisi çok az olduğundan, bu terimler etkisiz olarak kabul edilmiştir. Bu şekilde elde edilen t_1 değişkeni Eşitlik 3'teki gibi ifade edilebilir (Johnson, 1978).

$$t_1 = \left\{ (\bar{X} - \mu) + \lambda + \gamma \left[(\bar{X} - \mu)^2 - (\sigma^2 / n) \right] \right\} (\sigma^2 / n)^{-1/2} \quad (3)$$

Burada λ , n 'nin bir fonksiyonu ve γ momentsel bir orandır ve $\gamma = \mu_3 / 3\sigma^4$, $\lambda = \mu_3 / 2n\sigma^2$ olarak tanımlanır. Buradan

$$t_1 = \left[(\bar{X} - \mu) + \frac{\mu_3}{6\sigma^2 n} + \frac{\mu_3}{3\sigma^4} (\bar{X} - \mu)^2 \right] \left[\frac{\sigma^2}{n} \right]^{-1/2} \quad (4)$$

olarak elde edilir. Bu istatistik $n-1$ serbestlik dereceli t dağılımına sahiptir.

t_1 istatistiğine ilişkin yan, payındaki $\mu_3 / 2\sigma^2 n$ ve 3. merkezsel momenti olan $4\mu_3 / n$ terimine bağlıdır. t_1 istatistiğinin payı ile paydası arasındaki kovaryans $4\mu_3 / n^2$ 'den daha küçüktür. Aynı zamanda bu istatistiğin γ ve λ terimlerini içermesi, çarpıklıktan kaynaklanan yüksek yanlılığın etkisini azaltmaktadır. Uygulamada μ_3 ve σ^2 bilinmediğinden bu parametreler örnekten tahmin edilir. Buna göre $H_0 : \mu = \mu_0$ yokluk hipotezinin, $H_1 : \mu > \mu_0$ seçenek hipotezine karşı testinde

$$\left[(\bar{X} - \mu_0) + \frac{\hat{\mu}_3}{6S^2 n} + \frac{\hat{\mu}_3}{3S^4} (\bar{X} - \mu_0)^2 \right] > t_{n-1, \alpha} \frac{S}{\sqrt{n}} \quad (5)$$

ise $1-\alpha$ güven düzeyinde H_0 red edilir. Burada, $t_{n-1, \alpha}$ değeri $n-1$ serbestlik dereceli t dağılımından elde edilen kritik değerdir.

t_1 istatistiğinin payı doğrusal olmadığından güven aralığı için basit bir tanımlama yapılamaz. Ama γ ve λ yan miktarını ve çarpıklığın etkisini azalttığından dolayı karesel yapıyı içermeyen yeni bir istatistik tanımlanabilir. Johnson (1978), kitle ortalamasına ilişkin güven aralığını elde edebilmek için $(\bar{X} - \mu_0)^2$ terimini seriden çıkarmıştır. Buna göre, $1-\alpha$ güven düzeyinde kitle ortalamasına ilişkin güven aralığı

$$\left[(\bar{X} + \frac{\hat{\mu}_3}{6S^2 n}) \right] \pm t_{n-1, \alpha/2} S / \sqrt{n} \quad (6)$$

ile bulunabilir (Johnson, 1978).

Düzeltilmiş güven aralığı çok kullanışlıdır. Aralık bu düzeltmeye göre elde ediliyorsa, aralığın genişliği çok artmaz. Ayrıca, çarpık dağılımlarda medyan ve ortalama arasında belirli bir farkın olmasından dolayı CF açılımlarının kullanılarak bu farkın minimuma indirilmesi ve bu şekilde güven aralığının oluşturulması daha iyi sonuçlar elde edilmesini sağlar (Johnson, 1978).

2.2 Sutton'ın Bileşik Testi

Kitle ortalamasına ilişkin yokluk hipotezini test etmek için varyansı ve üçüncü merkezi momenti bilinmeyen bir kitleden, örneklem büyüklüğü n olan rastgele örneklem seçildiğini varsayalım. Bu örneklemde, örneklem ortalaması ve örneklem varyansı hesaplınsın. Johnson (1978) üçüncü merkezi momenti sıfırdan farklı bir değer varsayarak $n-1$ serbestlik dereceli Student t dağılımına sahip t_1 istatistiğini tanımlamıştır. Çarpıklık çok büyük olmadığında, bu test istatistiğinin Student t istatistiğine göre daha güçlü olduğu, Sutton (1993)'in yaptığı simülasyon çalışmaları ile gösterilmiştir. Ayrıca t_1 testinin kitlenin dağılımı sağa çarpık iken daha etkili olduğu Sutton (1993) tarafından vurgulanmıştır.

Sutton (1993) Üstel, Gamma, Ki-Kare, Weibull ve Lognormal gibi dağılımların ortalamalarına ilişkin yokluk hipotezini test etmek amacıyla hem Student t testini, hem de Johnson'ın Düzeltmiş Karesel t testini kullanmıştır (Sutton, 1993). Eğer çarpıklık Johnson (1978)'in incelediği duruma göre daha fazla ve örnek hacmi yeterince büyük değil ise deneysel 1. tip hata oranının nominal değerden farklı olacağını yine Sutton (1993) çalışmasında göstermiştir. Böylece her örneklem büyüklüğü için en az Johnson'ın Düzeltmiş Karesel Testi kadar güçlü birleşik bir test elde etmiştir. Bu test yine Johnson'ın düzeltilmiş karesel t istatistiğine dayanmaktadır.

$t_1 > t_{n-1,\alpha}$ veya $t_1 / s_{t_{1sb}} > z_\alpha$ veya $\left\{ \frac{(n-1)}{(n-3)} \right\}^{\frac{1}{2}} t_1 / s_{t_{1sb}} > t_{n-1,\alpha}$ ise $H_0 : \mu = \mu_0$ yokluk hipotezi $H_1 : \mu > \mu_0$ hipotezine karşı $1-\alpha$ güven düzeyinde red edilir. Burada, $s_{t_{1sb}}$ değeri t_1 istatistiğinin bootstrap standart hatasıdır. Bu bileşik test için red bölgesi $\alpha=0.01$ ve $\alpha=0.05$ için oluşturulmuş ve Monte Carlo çalışmaları ile bileşik testin güç bakımından iyi özelliklere sahip olduğu sonucuna varılmıştır (Sutton, 1993).

2.3 Chen'in Testi

Sağa çarpık dağılımların ortalamalarına ilişkin yokluk hipotezinin testi için önerilen bu istatistik, herhangi bir X rastgele değişkeni için aşağıda ifade edilen Edgeworth Açılımları kullanılarak elde edilmiştir (Chen, 1995).

$$P \left\{ \frac{\sqrt{n}(\bar{X} - \mu)}{S} \leq x - \frac{1}{6\sqrt{n}} \hat{\beta}_1 (1 + 2x^2) \right\} = \Phi(x) + o(n^{-\frac{1}{2}}) \quad (7)$$

Burada $\Phi(x)$ standart normal olasılık yoğunluk fonksiyonudur ve

$$\hat{\beta}_1 = \frac{\sum (X_i - \bar{X})^3 / n}{S^3}$$

olarak tanımlanmıştır (Hall, 1983). $o(n^{-\frac{1}{2}})$ ise ilk iki terimden sonraki terimlerin $n^{-\frac{1}{2}}$, ye bölümü ile elde edilen serinin sıfıra yakınsadığını göstermektedir. $H_0 : \mu = \mu_0$ hipotezinin $H_1 : \mu > \mu_0$ hipotezine karşı testinde

$$\frac{\sqrt{n}(\bar{X} - \mu_0)}{S} > Z_\alpha - \frac{1}{6\sqrt{n}} \hat{\beta}_1 (2Z_\alpha^2 + 1) \quad (8)$$

ise yokluk hipotezi $1 - \alpha$ güven düzeyinde red edilir (Chen, 1995). Burada Z_α standart normal dağılımın tablo değeridir. Karar kuralı aşağıdaki yaklaşım kullanılarak elde edilir.

$$\frac{\sqrt{n}(\bar{X} - \mu_0)}{S} > x - \frac{1}{6\sqrt{n}} \hat{\beta}_1 (2x^2 + 1), (\forall x) \quad (9)$$

Burada $a = \hat{\beta}_1 / 6\sqrt{n}$ ve $t = \sqrt{n}(\bar{X} - \mu_0) / S$ olmak üzere, $n \rightarrow \infty$ iken $(1 - 8a(t+a)) \geq 0$ varsayımı altında Eşitlik 8 kullanılarak

$$P \left\{ \frac{1 - \sqrt{1 - 8a(t+a)}}{4a} > Z_\alpha \right\} \leq \alpha \quad (10)$$

ifadesi yazılabilir. Taylor açılımı kullanılarak

$$\frac{1 - \sqrt{1 - 8a(t+a)}}{4a} = t + a + 2at^2 + 4a^2(t + 2t^3) + o(n^{-1}) \quad (11)$$

elde edilir [Chen, 1995].

Böylece test istatistiği

$$t_2 = t + a + 2at^2 + 4a^2(t + 2t^3) \quad (12)$$

olur. Bu istatistik, standart normal dağılıma sahiptir. Bu açılım, orijinal değişkenler ile ifade edilirse,

$$t_2 = \frac{\sqrt{n}(\bar{X} - \mu_0)}{S} + \frac{1}{6\sqrt{n}} \hat{\beta}_1 \left[2 \left(\frac{\sqrt{n}(\bar{X} - \mu_0)}{S} \right)^2 + 1 \right] + \frac{1}{9n} \hat{\beta}_1^2 \left[\frac{\sqrt{n}(\bar{X} - \mu_0)}{S} + 2 \left(\frac{\sqrt{n}(\bar{X} - \mu_0)}{S} \right)^2 \right] \quad (13)$$

elde edilir (Chen, 1995). Bu test istatistiği örnek hacmi ne kadar küçük olursa olsun kullanılabilir (Chen, 1995). Eğer $t_2 > Z_\alpha$ ise yokluk hipotezi α anlamlılık düzeyinde red edilir. Bu istatistik Student t istatistiği ve Johnson'ın düzeltilmiş karesel t istatistiğinin geliştirilmiş bir halidir. Chen (1995), yaptığı simülasyon çalışmaları sonucunda, test için Z_α kritik değerinin kullanılmasının küçük örneklem büyüklüğüne sahip örneklerde çok iyi sonuçlar verdiğini tespit etmiştir. Diğer taraftan, bu kritik değer kullanıldığında örneklem büyüklüğü 20'den fazla iken deneysel 1. tip hata oranlarının, nominal 1. tip hata değerini aştığı gözlemlenmiştir. Ayrıca, Chen (1995)

çalışmasında küçük örnek hacimlerinde bu test için $t_{n-1,\alpha}$ kritik değerinin kullanılabileceğini belirtmiştir. Dolayısıyla, bu çalışmada da Chen'in testi için $t_{n-1,\alpha}$ kritik değeri kullanılmıştır.

2.4 Simülasyon Çalışması

Simülasyon çalışmasında, $\lambda=0.25$ parametrelili üstel dağılım ($\mu=4$, $\sigma^2=16$ ve $\text{Çarpıklık}=2$), $k=0,50$, $\lambda=1,00$ parametrelerine sahip Weibull dağılımı ($\mu=2.00$, $\sigma^2=20.00$ ve $\text{Çarpıklık}=6.62$) ve $\alpha=1$, $\beta=2$ parametrelerine sahip Beta dağılımı ($\mu=0.33$, $\sigma^2=20.00$ ve $\text{Çarpıklık}=0.57$) kullanılmıştır. Belirtilen parametreler için ele alınan üç dağılım sağa çarpıktır.

Eğer X rastgele değişkeni Üstel dağılıma sahipse, bu değişkenin olasılık yoğunluk fonksiyonu

$$f(x; \lambda) = \begin{cases} \lambda e^{-\lambda x}, & x \geq 0 \\ 0, & x < 0 \end{cases} \quad (14)$$

olarak ifade edilir ve X : $\text{Üstel}(\lambda)$ ile gösterilir. Burada λ dağılımın tek parametresi olmak üzere pozitif bir sayıdır.

Weibull dağılımı bir sürekli olasılık dağılımı olup, olasılık yoğunluk fonksiyonu

$$f(x; k, \lambda) = \begin{cases} \frac{k}{\lambda} \left(\frac{x}{\lambda} \right)^{k-1} e^{-(x/\lambda)^k}, & x \geq 0; k > 0; \lambda > 0 \\ 0, & d.h. \end{cases} \quad (15)$$

olarak ifade edilir ve X : $\text{Weibull}(k, \lambda)$ ile gösterilir. Burada k şekil parametresi ve λ ölçek parametresidir.

Beta dağılımı $[0,1]$ aralığında α ve β gibi iki tane pozitif şekil parametresi ile normalize edilmiş bir sürekli olasılık dağılımı ailesidir. Beta dağılımının olasılık yoğunluk fonksiyonu

$$f(x; \alpha, \beta) = \begin{cases} \frac{x^{\alpha-1} (1-x)^{\beta-1}}{B(\alpha, \beta)}, & \alpha > 0, \beta > 0, x \in [0,1] \\ 0, & d.h. \end{cases} \quad (16)$$

olarak ifade edilir ve $X : Beta(\alpha, \beta)$ ile gösterilir. Olasılık yoğunluk fonksiyonundaki $B(\alpha, \beta)$ ifadesi Gamma fonksiyonlarından faydalanarak, aşağıdaki gibi elde edilir.

$$B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha + \beta)} \quad (17)$$

Simülasyon çalışmasında algoritmalar Matlab R2006a paket programı ile yazılmıştır. Simülasyon programları hem deneysel 1. tip hata oranlarını, hem de testin gücü değerlerini elde etmeye yöneliktir. Bu oranları elde etmek için kullanılan kayma miktarı (kym), kitle ortalamasında meydana gelen kaymayı göstermektedir. Simülasyon çalışmasında; aşağıda verilen örneklem büyüklükleri, nominal 1. tip hata oranları ve kayma miktarları kullanılmıştır.

Örneklem Büyüklükleri (n)= 7, 13, 20, 30, 50
Nominal 1. tip hata oranları (α) = 0.01, 0.05 ve 0.10
Kayma Miktarları (kym)= 0.50; 1.00; 1.50

Kayma miktarları, Beta dağılımı için 0.05, 0.10, 0.15 olarak alınmıştır. Bunun nedeni bu dağılımın 0 ile 1 arasında değer almasıdır. Çalışmada 10000 iterasyon kullanılmıştır.

Test isimleri uzun olduğu için tabloları elde etmede kolaylık olması açısından bazı kısaltmalar yapılmıştır. Bu kısaltmalar; Student t testi için **ST**, Johnson'ın düzeltilmiş karesel t testi için **JDKT**, Sutton'ın bileşik testi için **SBT** ve Chen'in testi için **CT** şeklindedir.

3. BULGULAR

Tablo 1'de, Üstel, Beta ve Weibull dağılımlarından 10000 iterasyona dayalı 7, 13, 20, 30 ve 50 örneklem büyüklükleri ile $H_1 : \mu > \mu_0$ hipotezine karşı $H_0 : \mu = \mu_0$ hipotezinin test edilmesinde her bir test istatistiği için elde edilen deneysel 1. tip hata oranları verilmiştir.

Tablo 1'de Üstel dağılım incelendiğinde, nominal $\alpha=0.01$ ve örneklem büyüklüğü 7 iken Chen'in testine ilişkin deneysel 1. tip hata oranının 0.009 olduğu görülmektedir. Aynı koşullarda Sutton'ın bileşik testi için deneysel 1. tip hata oranı 0.004 olarak gerçekleşmiştir. Diğer taraftan Weibull dağılımı için yine aynı koşullar incelendiğinde ise, Chen'in testine ilişkin deneysel 1. tip hata oranının 0.005 ve Sutton'ın bileşik testine ilişkin deneysel 1. tip hata oranının ise 0.003 olduğu görülmektedir. Diğer testler için Üstel dağılımdan ve Weibull dağılımından çekilen örneklemelerden hesaplanan deneysel 1. tip hata oranları ise nominal değerlere yaklaşmamıştır.

İlgili parametreler için Beta dağılımı, diğer dağılımlara göre daha düşük çarpıklık değerine sahiptir. Bu durumda, Chen'in testi ve Sutton'ın bileşik testi için deneysel 1. tip hata oranları, nominal α 'ya, diğer testler için elde edilen deneysel 1. tip hata oranlarına göre daha çok yaklaşmıştır. Beta dağılımı için hem Sutton'ın bileşik testi, hem de Chen'in testi deneysel 1. tip hata oranları bakımından hemen hemen aynı sonuçları vermiştir. Diğer testler için deneysel 1. tip hata oranları çarpıklıkların düşük olmasından dolayı belirli bir artış göstermişlerdir. Fakat, bu oranlar çoğu durumda nominal α 'ya yaklaşmamıştır.

Üstel dağılım ve Weibull dağılımı için örneklem büyüklüğünün 20, 30 veya 50 olduğu durumlarda, Chen'in testi deneysel 1. tip hata oranları bakımından nominal α 'ya en yakın sonuçlara sahiptir. Aynı koşullarda Beta dağılımı incelendiğinde ise Sutton'ın bileşik testi ve Chen'in testi nominal α 'ya en yakın sonuçlara sahiptir. Genel olarak düşünüldüğünde, Chen'in testi ele alınan dağılımlardan çekilen örneklem ile kitle ortalamasına ilişkin yokluk hipotezinin test edilmesinde diğer testlerden daha iyi sonuçlara sahiptir.

Doğal bir sonuç olarak, örneklem büyüklüğü arttıkça, bütün testler için deneysel 1. tip hata oranlarının nominal α 'ya yaklaştığı gözlemlenmiştir.

Tablo 1. Beta, Üstel ve Weibull dağılımları için deneysel 1. tip hata oranları

Örneklemin Geldiği Kitlenin Dağılımı	Nominal α	Test İstatistiği	$n=7$	$n=13$	$n=20$	$n=30$	$n=50$
Beta Dağılımı $\alpha=1$ $\beta=2$ (0.57)	0.01	ST	0.005	0.005	0.006	0.006	0.007
		JDKT	0.006	0.006	0.007	0.007	0.008
		SBT	0.008	0.008	0.008	0.009	0.009
		CT	0.008	0.008	0.008	0.009	0.009
	0.05	ST	0.031	0.033	0.035	0.040	0.041
		JDKT	0.031	0.033	0.037	0.043	0.045
		SBT	0.033	0.037	0.041	0.047	0.048
		CT	0.032	0.040	0.048	0.049	0.050
	0.10	ST	0.071	0.077	0.080	0.087	0.092
		JDKT	0.072	0.079	0.084	0.094	0.098
		SBT	0.074	0.084	0.088	0.097	0.101
		CT	0.074	0.084	0.089	0.098	0.101
Üstel Dağılım $\lambda=0.25$ (2.00)	0.01	ST	0.000	0.000	0.000	0.002	0.003
		JDKT	0.001	0.002	0.004	0.005	0.008
		SBT	0.004	0.005	0.005	0.007	0.009
		CT	0.009	0.010	0.010	0.011	0.011
	0.05	ST	0.012	0.015	0.018	0.023	0.029
		JDKT	0.014	0.031	0.033	0.041	0.048
		SBT	0.022	0.034	0.036	0.044	0.050
		CT	0.040	0.048	0.052	0.053	0.055
	0.10	ST	0.040	0.053	0.054	0.065	0.071
		JDKT	0.053	0.079	0.080	0.093	0.097
		SBT	0.062	0.081	0.083	0.096	0.101
		CT	0.090	0.095	0.102	0.104	0.106
Weibull Dağılımı $k=0.50$ $\lambda=1.00$ (6.62)	0.01	ST	0.000	0.000	0.000	0.000	0.000
		JDKT	0.001	0.002	0.002	0.003	0.007
		SBT	0.003	0.004	0.004	0.006	0.008
		CT	0.005	0.005	0.005	0.008	0.010
	0.05	ST	0.002	0.002	0.004	0.006	0.011
		JDKT	0.013	0.014	0.017	0.018	0.022
		SBT	0.014	0.020	0.024	0.035	0.041
		CT	0.022	0.031	0.041	0.044	0.050
	0.10	ST	0.013	0.019	0.023	0.030	0.042
		JDKT	0.032	0.053	0.077	0.078	0.085
		SBT	0.041	0.072	0.080	0.086	0.087
		CT	0.055	0.082	0.088	0.090	0.100

Tablo 2, Tablo 3 ve Tablo 4'te Beta, Üstel ve Weibull dağılımlarından 10000 tekrar ile seçilen 7, 13 ve 30 büyüklüğündeki örnekler ile $H_1: \mu > \mu_0$ hipotezine karşı $H_0: \mu = \mu_0$ hipotezinin test edilmesinde her bir test istatistiği için testin gücü değerleri verilmiştir.

Tablo 2’de yer alan Beta dağılımından çekilen örneklerle elde edilen testin gücü değerleri çarpıklığının düşük olması nedeniyle bütün testler için birbirine yakın çıkmıştır. Fakat Chen’in testi ve Sutton’ın bileşik testi, diğer testlere oranla daha yüksek testin gücü değerlerine sahiptir.

Tablo 2. Beta dağılımı için testin gücü değerleri

Örneklemin Geldiği Kitlenin Dağılımı	Nominal α	Test İstatistiği	$n=7$			$n=13$			$n=30$		
			$kym=$	$kym=$	$kym=$	$kym=$	$kym=$	$kym=$	$kym=$	$kym=$	$kym=$
			0.05	0.10	0.15	0.05	0.10	0.15	0.05	0.10	0.15
Beta Dağılımı $\alpha=1$ $\beta=2$ (0.57)	0.01	ST	0.014	0.037	0.084	0.029	0.110	0.325	0.083	0.427	0.881
		JDKT	0.015	0.038	0.085	0.029	0.126	0.401	0.100	0.502	0.931
		SBT	0.025	0.060	0.135	0.041	0.168	0.480	0.118	0.534	0.938
	0.05	CT	0.025	0.060	0.140	0.041	0.169	0.480	0.119	0.538	0.940
		ST	0.086	0.192	0.367	0.134	0.372	0.711	0.276	0.745	0.980
		JDKT	0.088	0.194	0.394	0.146	0.415	0.773	0.303	0.789	0.987
	0.10	SBT	0.098	0.220	0.446	0.160	0.437	0.789	0.315	0.797	0.989
		CT	0.100	0.230	0.448	0.161	0.437	0.791	0.317	0.798	0.989
		ST	0.182	0.347	0.590	0.250	0.559	0.855	0.423	0.864	0.994
		JDKT	0.183	0.366	0.636	0.270	0.597	0.885	0.451	0.883	0.996
		SBT	0.200	0.390	0.660	0.281	0.608	0.889	0.461	0.888	0.997
		CT	0.200	0.391	0.660	0.285	0.608	0.893	0.465	0.889	0.997

Tablo 3’te, üstel dağılım incelendiğinde, testin gücü değerlerinin düşük gerçekleştiği görülmektedir. Bütün örneklem büyüklüklerinde, kayma miktarı (kym) 0,50 iken en yüksek testin gücü değerlerine Chen’in testi sahiptir.

Üstel dağılımdan seçilen örneklem ile kitle ortalamasına ilişkin yokluk hipotezinin test edilmesinde, Chen’in testinin daha güçlü bir test olduğu aşikardır.

Tablo 3. Üstel dağılım için testin gücü değerleri

Örneklemin Geldiği Kitlenin Dağılımı	Nominal α	Test İstatistiği	$n=7$			$n=13$			$n=30$		
			$kym=$	$kym=$	$kym=$	$kym=$	$kym=$	$kym=$	$kym=$	$kym=$	$kym=$
			0.50	1.00	1.50	0.50	1.00	1.50	0.50	1.00	1.50
Üstel Dağılım $\lambda=0.25$ (2.00)	0.01	ST	0.002	0.003	0.006	0.003	0.006	0.018	0.011	0.042	0.101
		JDKT	0.002	0.005	0.008	0.008	0.019	0.044	0.032	0.100	0.220
		SBT	0.009	0.013	0.025	0.015	0.035	0.072	0.041	0.118	0.249
	0.05	CT	0.019	0.035	0.062	0.029	0.066	0.125	0.059	0.159	0.314
		ST	0.022	0.041	0.066	0.041	0.086	0.153	0.093	0.220	0.395
		JDKT	0.032	0.058	0.095	0.073	0.142	0.232	0.142	0.310	0.512
	0.10	SBT	0.045	0.078	0.124	0.082	0.160	0.253	0.149	0.321	0.524
		CT	0.077	0.129	0.193	0.111	0.207	0.319	0.174	0.356	0.562
		ST	0.072	0.121	0.177	0.115	0.206	0.311	0.197	0.385	0.589
		JDKT	0.098	0.158	0.227	0.159	0.274	0.396	0.249	0.466	0.659
		SBT	0.111	0.177	0.247	0.167	0.283	0.408	0.255	0.473	0.667
		CT	0.155	0.235	0.315	0.196	0.328	0.457	0.270	0.495	0.688

Tablo 4 incelendiğinde, Weibull dağılımı için Sutton’ın bileşik ve Chen’in testlerinin örneklem büyüklüğü 7, nominal $\alpha=0.01$ ve kayma miktarı 0,50 olduğunda aynı güce sahip olduğu görülmektedir. Diğer durumlarda Chen’in testine ilişkin testin gücü değerleri daha yüksek gerçekleşmiştir.

Örneklem büyüklüğünün 13 ve 30 olduğu durumda Weibull dağılımı için testlere ilişkin testin gücü değerleri incelendiğinde, Chen'in testinin daha güçlü bir test olduğu anlaşılmaktadır. Örneklem büyüklüğü arttıkça, Weibull dağılımdan çekilen örneklem ile kitle ortalamasına ilişkin yokluk hipotezinin test edilmesinde tüm testler için testin gücü değerleri artmıştır. Kayma miktarı arttıkça, tüm testler için testin gücü değerlerinde yine artış meydana gelmiştir.

Tablo 4. Weibull dağılımı için testin gücü değerleri

Örneklem Geldiği Kitlenin Dağılımı	Nominal α	Test İstatistiği	$n=7$			$n=13$			$n=30$		
			$kym=$ 0.50	$kym=$ 1.00	$kym=$ 1.50	$kym=$ 0.50	$kym=$ 1.00	$kym=$ 1.50	$kym=$ 0.50	$kym=$ 1.00	$kym=$ 1.50
Weibull	0.01	ST	0.000	0.001	0.034	0.000	0.002	0.179	0.008	0.026	0.674
Dağılımı		JDKT	0.005	0.007	0.049	0.008	0.017	0.345	0.022	0.179	0.726
$k=0.50$		SBT	0.024	0.030	0.125	0.028	0.066	0.416	0.042	0.269	0.729
$\lambda=1.00$		CT	0.024	0.031	0.258	0.030	0.079	0.506	0.050	0.306	0.733
(6.62)	0.05	ST	0.006	0.020	0.231	0.013	0.064	0.492	0.050	0.271	0.733
		JDKT	0.035	0.058	0.318	0.058	0.199	0.567	0.147	0.523	0.737
		SBT	0.054	0.109	0.363	0.075	0.237	0.574	0.160	0.545	0.737
		CT	0.060	0.139	0.456	0.090	0.291	0.595	0.194	0.595	0.738
	0.10	ST	0.035	0.101	0.408	0.067	0.226	0.587	0.157	0.524	0.737
		JDKT	0.096	0.202	0.480	0.157	0.396	0.610	0.272	0.685	0.738
		SBT	0.113	0.230	0.489	0.167	0.409	0.612	0.279	0.692	0.738
		CT	0.115	0.268	0.532	0.197	0.457	0.619	0.308	0.721	0.749

Beta dağılımı, Üstel dağılım ve Weibull dağılımından çekilen örneklem ile kitle ortalamasına ilişkin yokluk hipotezini test etmede testin gücü değerleri bakımından Chen'in testinin daha iyi bir test olduğu söylenebilir.

4. TARTIŞMA VE SONUÇ

Bu çalışmada, sağa çarpık kitlelerden seçilen örneklem ile kitle ortalamasının testi için kullanılan Johnson'ın düzeltilmiş karesel t testi, Sutton'ın bileşik testi, Chen'in testi deneysel 1. tip hata oranları ve testin gücü değerleri bakımından karşılaştırılmıştır. Bu amaçla, Üstel (0.25), Weibull (0.50, 1) ve Beta (1, 2) dağılımlarından örneklem çekilmiş ve her bir test için deneysel 1. tip hata oranları ve testin gücü değerleri hesaplanmıştır.

Gerçekte doğru olan yokluk hipotezini reddetme (1. tip hata) oranları için Tablo 1'e bakıldığında, Johnson'ın düzeltilmiş karesel testine ilişkin deneysel 1. tip hata oranları nominal α 'ya yaklaşırsa da, bu genelde çarpıklığın düşük ve örneklem büyüklüğünün büyük olduğu durumlarda gerçekleşmiştir. Sutton'ın bileşik testi ve Chen'in testi özellikle küçük örnekler için deneysel 1. tip hata oranları bakımından hemen hemen yakın sonuçlara sahiptir. Fakat küçük örneklerde başarılı gibi görünen Sutton'ın bileşik testi, aynı başarıyı büyük örneklerde gösterememiştir. Chen'in testi ise çoğu durumda, deneysel 1. tip hata oranları bakımından nominal α 'ya yaklaşımıştır. Bu sonuçlar, Chen'in testinin daha iyi bir test olduğunun göstergesidir.

Gerçekte yanlış olan yokluk hipotezini reddetme (testin gücü) değerleri için Tablo 2 ve sonraki tablolara bakıldığında, Johnson'ın düzeltilmiş karesel testi bazı durumlarda güçlü bir test olarak görünse de, bu sadece çarpıklığın düşük ve örneklem büyüklüğü olduğu durumlarda söz konusudur. Testin gücü değerleri için elde edilen çoğu tabloda,

Chen'in testinin gücü en yüksektir. Sutton'ın bileşik testi bazı durumlarda testin gücü açısından, Chen'in testine yaklaşmış olsa da, bu durum sadece küçük örneklerde veya yüksek kayma miktarlarında elde edilmiştir. Beta dağılımının çarpıklığı düşük olduğundan, Sutton'ın bileşik testi testin gücü değerleri bakımından, Chen'in testine ilişkin değerler ile yakın sonuçlara sahiptir.

Bazı durumlarda, Sutton'ın bileşik testi ile Chen'in testi için testin gücü değerleri ve deneysel 1. tip hata oranları bakımından yakın sonuçlar elde edilse de, Sutton'ın bileşik testi Bootstrap örneği çekmek gerektirdiğinden ve üç aşamalı bir karar mekanizmasına sahip olduğundan, uygulamada kullanılması zor bir yöntemdir. Chen'in testi ise tek bir karar kuralına sahip olduğundan daha kullanışlıdır.

Chen'in testi, ele alınan dağılımlar için gerçekte doğru olan yokluk hipotezini reddetme oranları ve gerçekte yanlış olan yokluk hipotezini reddetme değerleri bakımından diğer testlere göre daha iyi bir testtir.

5. KAYNAKLAR

Andrews, D. F., Bickel, P. J., Hampel, F. R., Huber, P. J., Rogers, W. H., Tukey, J. W., 1972. Robust estimates of location: Survey and advances. Princeton University Press, Princeton.

Anscombe, F. J., 1950. Tables of hyperbolic transformation $\sinh. \sqrt{X}$. Journal of the Royal Statistical Society, A/113, 228-229.

Bartlett, M. S., 1935. The effect of non-normality on the t distribution. Proceedings of the Cambridge Philosophical Society. 31, 223-231.

Box, G. E. P., Andersen, S. L., 1955. Permutation theory in the derivation of robust criteria and the study of departures from assumption. Journal of the Royal Statistical Society, B/17, 1-26.

Chen, L., 1995. Testing the mean of skewed distributions. Journal of The American Statistical Association. Vol. 90, No. 430, 767-772.

Chung, K., 1946. The approximate distribution of Student's statistics. Annals of Mathematical Statistics, 17, 447-465.

Cornish, E. A., Fisher, R. A., 1937. Moments and cumulants in the specification of distributions. Revue of the International Statistics Institute, 5, 307-327.

Gayen, A. K., 1949. The distribution of 'Student's' t in random samples of any size drawn from non-normal universes. Biometrika, 36, 353-369.

Geary, R. C., 1936. The distribution of Student's ratio for non-normal samples. Journal of the Royal Statistical Society, 3/2, 178-184.

Hall, P., 1983. Inverting an Edgeworth expansion. The Annals of Statistics, 11: 569-576.

- Johnson, N. J., 1978. Modified t tests and confidence intervals for asymmetrical populations. *Journal of the American Statistical Association*, Vol. 73, No. 363, 536-544.
- Laderman, J., 1939. The distribution of Student's ratio for samples of two items drawn from non-normal universes. *Annals of Mathematical Statistics*, 10, 376-379.
- Nair, A. K. N., 1941. Distribution of Student's t in the correlation coefficient in sample from non-normal population. *Sankhya*, 5, 383-400.
- Neyman, J., Pearson, E. S., 1928. On the use and interpretation of certain test criteria for purposes of statistical inference part I. *Biometrika*, 20A, 175-240.
- Perlo, V., 1933. On the distribution of Student's ratio for samples of three drawn from a rectangular distribution. *Biometrika*, 25, 203-204.
- Rider, P. R., 1929. On the distribution of the ratio of mean to standard deviation in small samples from non-normal universes. *Biometrika*, 21, 124-143.
- Sophister, 1928. Discussion of small samples drawn from an infinite skew population. *Biometrika*, 20A, 389-423.
- Student, 1908. The probable error of a mean. *Biometrika*, 6, 1-25.
- Sutton, C. D., 1993. Computer-intensive methods for tests about the mean of an asymmetrical distribution. *Journal of the American Statistical Association*, 88, 802-810.
- Tukey, J. W., 1964. Data analysis and behavioral sciences. Unpublished Manuscript.
- Wallace, D. L., 1958. Asymptotic approximations to distributions. *Annals of Mathematical Statistics*, 29, 635-654.
- Yuen, K. K., 1974. The two sample trimmed t for unequal population variances. *Biometrika*, 61, 165-170.

USE OF SOME TESTS FOR MEANS OF POSITIVELY SKEWED DISTRIBUTION AND COMPARISONS

ABSTRACT

The purpose of this article is to introduce Johnson's modified square t test, Sutton's composite test and Chen's test applied to test the null hypotheses of population mean with samples drawn from positively skewed distributions, also compare them on the account of type I error rates and power of the test. In this context, a simulation study is carried out via data group which is produced from some positively skewed distributions and it is determined that Chen's test is more powerful than other tests.

Keywords: Chen's test, Johnson's modified square t test, Power of the test, Skewness, Sutton's composite test.

GIDA TÜKETİM ANALİZİ: GENİŞLETİLMİŞ DOĞRUSAL HARCAMA SİSTEMİ

Mehmet Arif ŞAHİNLİ*

Ahmet ÖZÇELİK**

ÖZET

Hanehalkı bütçe anketinden elde edilen tüketim harcamaları bilgileri ile; tüketici fiyat endekslerinde kullanılacak maddelerin seçimi ve temel yıl ağırlıklarının elde edilmesi, hanelerin tüketim kalıplarında zaman içinde meydana gelen değişikliklerin izlenmesi, milli gelir hesaplamalarında özel nihai tüketim harcamaları tahminlerine yardımcı olacak verilerin derlenmesi, yoksulluk sınırının belirlenmesi, asgari ücret tespit çalışmaları, vb. diğer sosyo-ekonomik analizler için gerekli verilerin elde edilmesi amaçlanmaktadır. Bu çalışmada, Türkiye İstatistik Kurumu 2003 yılı Hanehalkı Gelir ve Tüketim Harcamaları anket verileri kullanılmıştır. Gıda alt grupları harcaması için marjinal bütçe payı ve gelir esnekliği, geliştirilmiş doğrusal harcama sistemi yaklaşımı çerçevesinde tahmin edilmiştir. Elde edilen bulgulara göre, hanehalkı tüketici davranışı analiz edilmektedir.

Anahtar Kelimeler: Genelleştirilmiş doğrusal harcama sistemi, Marjinal bütçe payı, Gelir esnekliği.

1. GİRİŞ

Belirli gelire sahip olan bireyler harcamalarını yaparken ilk önce ana mal grupları olarak bilinen gıda, giyim, konut, eğitim, ulaşırma vs. arasında paylaştırmaktadır. İkinci aşamada ise bireyler, alt gruplar bazında gelirini paylaşırır iken daha sonra kalan kısmını tasarrufa ayırmaktadır. Tüketicilerin toplam harcamasının mal grupları arasındaki dağılımı ve tüketim / tasarruf kararının çözümünde kullanılan talep denklem sisteminde tasarruf / tüketimin içsel olarak alındığı bir model gerekmektedir (Nişancı, 2002). Bu çalışmada ise, toplam gıda tüketim harcaması modele içsel bir değişken olarak katılmış olup, Lluch (1973)'ün geliştirmiş olduğu Genişletilmiş Doğrusal Harcama Sistemi (Extended Linear Expenditure System – ELES) kullanılmıştır.

Talep analizine ilişkin ekonometrik boyutta gerçekleştirilen bir çok araştırma yapılmıştır. Bunlardan Türkiye dışındaki ülke verileriyle gerçekleştirilenlerden bazıları Howe (1977), Sasaki ve Saegusa (1974), Narayana ve Vani (1996), Huang ve Bouis (2001), Burney ve Akmal (1991), Lee, Brown ve Seale (1994), Lewis ve Andrews (1989), Deaton, Castillo ve Thomas (1989), Eastwood ve Craven (1981), McConnell (1978), Lluch ve Williams (1975), Lluch (1973), Philips (1972), Stone (1954), Pollak ve Wales (1969)'dur.

Ülkemizde ise talep analizine dönük ekonometrik çalışmalardan bazıları ise Koç (1995), Ekinci (1996), Nişancı (2002), Özer (1992), Tansel (1986) ve Özer (2001)'dir.

* İstatistikçi, Türkiye İstatistik Kurumu, Eğitim ve Araştırma Merkezi Müdürlüğü, e-posta: arifsahinli@tuik.gov.tr

** Prof. Dr., Ankara Üniversitesi Tarım Ekonomisi Bölümü, e-posta: aoczelik@agri.ankara.edu.tr

Bu çalışmada, ELES modeli kullanılarak, Türkiye’de yaşayan hanehalklarının gıda alt gruplarına ilişkin tüketim kalıpları ortaya çıkarılmak istenmiştir. Bu tüketim kalıpları sayesinde hanehalklarının gelire bağlı tasarruf ve tüketim davranışları analiz edilmiştir. Çalışmada mevcut verilerin elde edilebilirliği düşünülerek 16 il çalışmaya dahil edilmiştir.

Çalışmanın ana planı şu şekilde özetlenebilir: Model ile ilgili teorik bilgiler Bölüm 2’de Yöntem kısmında, analizde kullanılan veriler Bölüm 3’de, elde edilen analiz sonuçları Bölüm 4’de değerlendirilmiş ve Bölüm 5’de sonuç kısmı kapsamlı olarak verilmiştir.

2. YÖNTEM

Doğrusal harcama sisteminde nispi fiyatların tasarruflar üzerindeki etkisinin ölçülememesi modelin en önemli eksikliğini oluşturmuştur. Bu eksikliği gidermek için model Lluch tarafından genişletilmiştir. Mal harcama sisteminin eşitliği ile içsel toplam tüketim harcamaları ve Stone-Geary fayda özelliği burada Genişletilmiş Doğrusal Harcama Sistemi (ELES) olarak adlandırılır. ELES ile ilgili toplam tüketim harcaması, mal harcama eşitliklerine ilave yapılarak elde edilmektedir. Fiyatlar, gelir ve gelir beklentileri sistemde dışsal değişkenlerdir (Lluch, 1973). Ancak bu çalışmada toplam gıda tüketim harcaması dikkate alınmıştır.

LES’in γ parametreleri geleneksel olarak “yaşam (geçim) miktarları” diye yorumlanmaktadır. Alternatif olarak toplam tüketim fonksiyonu sadece servet kavramları ile ilgili formüle edilmektedir. Genişletilmiş doğrusal harcama sistemi, LES ve toplam tüketim fonksiyonuna bölünebilmektedir. Fakat uygulamalarda ELES’ in stokastik özelliği, şu avantajlara sahiptir: a) Toplam tüketim fonksiyonu, fiyat verisinin yokluğunda, harcama sistemindeki parametreleri tanımlayabilmektedir. Özellikle, $\sum p_i \gamma_i$ ortalaması tahmin edilebilmektedir. b) Toplam tüketim harcamaları dışsal olarak incelendiği zaman, LES’ de parametrelerin tahmininde yanlılığı ölçmek mümkündür. c) Maksimum olabilirlik metotları ELES’de direkt olarak uygulanabilmektedir.

Hanehalklarının verdiği kararların kişi başına harcama ile ilişkili olduğu, gelir ve fiyatların sadece tüketimi etkileyen değişkenler olduğu varsayımı altında, hanehalkı tüketim davranışı aşağıdaki formül ile elde edilmektedir.

$$e_i = p_i x_i = p_i \gamma_i + \beta_i (y - \sum_j p_j \gamma_j) \quad (1)$$

Burada $i, j = 1, 2, \dots, n$ malları, $e_i = i$ malına hanehalklarının yaptığı kişi başına tüketimi, $p_i = i$ malının fiyatı, $x_i = i$ malından hanehalklarının kişi başına tükettiği miktar, $y =$ hanehalkının kişi başına geliri, (γ_i, β_i) tahmin edilecek parametreleri göstermektedir. β_i ’ler farklı malların marjinal tüketim eğilimini göstermektedir. $\sum \beta_i = \mu$ toplam marjinal tüketim eğilimidir. $\sum p_j \gamma_j$ toplam geçimlik harcamayı göstermektedir. $(y - \sum p_j \gamma_j)$ artan (ekstra) geliri ifade etmektedir. Harcama eşitliklerinin bütününe toplanması ile toplam tüketim fonksiyonu şu şekilde elde edilir:

$$E = (1 - \mu) \sum_i p_i \gamma_i + \mu y \quad (2)$$

Burada E , toplam hanehalkı tüketim harcamasını göstermektedir. Eşitlik (2) fiyat verisinin yokluğunda $\sum p_j \gamma_j$ tanımlanmasına imkan verir ve yatay kesit verilerden fiyat esnekliklerinin elde edilmesine yardım etmektedir (Burney ve Akmal, 1991). γ_i bütün eşitliklerde görüldüğü için, Eşitlik (1) tarafından tanımlanan eşitlikler sistemi eş zamanlı olarak tahmin edilmelidir. Bu yatay-kesit sınırlamalara etki etmekte olup, olabilirlik fonksiyonun maksimumizasyonuna ihtiyaç duymaktadır. ELES'in stokastik özelliği aşağıdaki gibi yazılabilir:

$$e_{ih} = \alpha_i + \beta_i y_h + \varepsilon_{ih} \quad (3)$$

Burada $h = 1, 2, \dots, H$ hanehalklarını göstermektedir. $\alpha_i = \gamma_i^* - \beta_i \sum_j \gamma_j^*$, ve ε_{ih} , genel klasik özelliklere sahip hata terimidir. Eşitlik (3) ile tanımlanan denklemler sistemi tanımlı regresyon denklemlerinden biridir. Eşitlikteki denklemlerden her birinin sıradan en küçük kareler yöntemi ile ayrı ayrı tahmini sistemin maksimum olabilirlik tahminine özdeş olacaktır. μ , γ_i^* ve $\sum \gamma_i^*$ 'nin maksimum olabilirlik tahminleri α_i ve β_i 'nin sıradan en küçük kareler tahminlerinden aşağıdaki ilişkiler kullanılarak tahmin edilebilmektedir.

$$\begin{aligned} \mu &= \sum \beta_i \\ \sum \gamma_i^* &= \sum \alpha_i / (1 - \mu) \\ \gamma_i^* &= \alpha_i + \beta_i \sum \gamma_j^* \end{aligned} \quad (4)$$

İlgili talep esneklikleri aşağıdaki gibi hesaplanabilmektedir (Nişancı, 2002).

$$i \text{ malının marjinal bütçe payı} : \beta_i^* = \beta_i / \mu \quad (5)$$

$$i \text{ malının gelir esnekliği} : \eta_{iy} = \beta_i (y / e_i) \quad (6)$$

$$\text{gelire göre tasarruf esnekliği} : \eta_{sy} = y / (y - \sum \gamma_i^*) \quad (7)$$

3. VERİ

Türkiye İstatistik Kurumu (TÜİK), 2002 yılından itibaren yıllık olarak hanehalkı bütçe anketleri uygulamaktadır. 2002 Hanehalkı Bütçe Anketi (HBA), Türkiye genelinde her ay değişen 800, yıl boyunca toplam 9 600 örnek hanehalkına uygulanmıştır. 2003 Hanehalkı Bütçe Anketi, 1 Ocak -31 Aralık 2003 tarihleri arasında her ay değişen 2160, yıl boyunca toplam 25.920 örnek hanehalkına uygulanmıştır. 2004, 2005 ve 2006 Hanehalkı Bütçe Anketleri ise her ay değişen 720, yıl boyunca toplam 8.640 örnek hanehalkına uygulanmıştır.

2002 Hanehalkı Bütçe Anketi'nin tahmin düzeyi Türkiye, kent ve kırsal yerler ayrımıdır. 2003 anketinden İstatistik Bölge Birimleri Sınıflamasında kır/kent ayrımında 12 Düzey-1 bölgesi ve 26 Düzey-2 bölgesi bazında tahminler üretilmiştir. 2004 yılından itibaren ise, yıllık anket sonuçlarından Türkiye geneli ile kırsal ve kentsel yerler ayrımında; her yılın anket sonuçlarının önceki anket bilgileriyle birleştirilmesiyle de

bölgesel düzeyde tahminler üretilmesi mümkün olmaktadır (TÜİK, 2008). Bu çalışmada, Türkiye İstatistik Kurumu 2003 yılı HBA verilerinden yararlanılmıştır.

2003 Hanehalkı Bütçe Anketi (HBA) verilerinin kullanılmasının nedeni, 2002 yılında yapılmış olan aylık Hanehalkı Bütçe Anketi'nin hemen arkasından yapılmış kapsamlı bir çalışma olmasıdır. Analizde kullanılan veriler 1994=100 bazlı olarak kullanılmıştır. Model tahminine dahil edilen iller Adana, Antalya, Ankara, Bursa, Diyarbakır, Eskişehir, Erzurum, Gaziantep, İstanbul, İzmir, Kayseri, Konya, Malatya, Samsun, Trabzon ve Zonguldak'tır.

Aylık oluşan perakende fiyatlar verisi ve 12 aylık tüketici fiyatları indeks rakamları, Türkiye İstatistik Kurumu Fiyat İstatistikleri veri tabanından temin edilmiştir.

4. BULGULAR

Hanehalkları birinci aşamada, harcamalarını ana mal grupları (gıda, giyim, konut, ulaştırma, vb.) arasında pay ederken, ikinci aşamada ise, harcamalarını alt gruplara (ekmek ve tahıllar, sebze ve meyveler, et ve balık, vb.) şeklinde dağıtmaktadırlar. Bu çalışmada gıda alt gruplarına ilişkin toplulaştırılmış 6 mal grubu incelenmiştir.

Model parametrelerinin tahmininde Sıradan En Küçük Kareler metodu kullanılmıştır. Modelin tahmininde, SAS Enterprise Guide 3.0 paket programı kullanılmıştır (Sas, 2004). ELES modeli ile yapılan tahminde, gıda alt gruplarına ait α ve β katsayıları elde edilmiş olup, sonuçlar Tablo 1'de verilmiştir.

Tablo 1. ELES'e göre mal gruplarına ait parametre tahmin değerleri

Mal Grupları	α_i	β_i
Ekmek ve tahıllar	79809,39 (0,46)	16,28052 (1,17)
Et, balık ve kümes hayvanları	5107,27 (0,02)	20,17219 (0,96)
Yağ, süt ve günlük yiyecekler	636787,2 (2,20)	15,31129 (0,66)
Taze, kuru meyve ve sebzeler	117117,6 (0,84)	12,73585 (1,14)
Çeşitli hazır yiyecekler	75603,61 (0,69)	1,75768 (0,20)
Alkolsüz içecekler	49226,66 (0,33)	7,48267 (0,63)

Not: Parantez içindeki değerler t-istatistik değerleridir.

R^2 ve Durbin-Watson (DW) değerleri, Tablo 2'de verilmiştir. Modelden elde edilen R^2 değerlerine göre; en düşük R^2 değeri 0,5497 ile yağ, süt ve günlük yiyeceklerde yer alırken, en yüksek R^2 değeri 0,6681 ile ekmek ve tahıllar grubunda yer almaktadır.

Tablo 2. Mal gruplarına göre istatistik değerleri

Mal Grupları	R^2	DW
Ekmek ve tahıllar	0,6681	1,269
Et, balık ve kümes hayvanları	0,6009	2,085
Yağ, süt ve günlük yiyecekler	0,5497	2,019
Taze, kuru meyve ve sebzeler	0,6412	1,555
Çeşitli hazır yiyecekler	0,5733	2,754
Alkolsüz içecekler	0,5774	1,827

DW çizelgesinden %5 düzeyinde 16 gözlem ile altı açıklayıcı değişken için $d_L=0.502$ ve $d_U=2.388$ 'dir. Tablo değerlerine göre, mal gruplarının kararsızlık bölgesinde olduğundan söz edilebilmektedir.

Gıda alt gruplarına ilişkin esneklik değerleri hesaplanmış olup, elde edilen değerler Tablo 3'de verilmiştir.

Tablo 3. Mal gruplarına göre marjinal bütçe payları, gelir esnekliği ve gelire göre tasarruf esnekliği

Mal Grupları	Marjinal Bütçe Payı	Gelir Esnekliği
Ekmek ve tahıllar	0,2208	0,5115
Et, balık ve kümes hayvanları	0,2736	0,8104
Yağ, süt ve günlük yiyecekler	0,2076	0,4324
Taze, kuru meyve ve sebzeler	0,1727	0,7682
Çeşitli hazır yiyecekler	0,0238	0,2017
Alkolsüz içecekler	0,1015	0,3083
	Toplam=1,000	Gelire göre tasarruf esnekliği=1,7446

Temel fayda fonksiyonuna göre, genişletilmiş doğrusal harcama sistemi modelinden elde edilen marjinal bütçe paylarının daima pozitif çıkması beklenir. Değerler beklentiye uygun olarak elde edilmiştir. Diğer her şey sabitken, hanelerin kişi başına harcamalarında bir birim artış olduğu zaman, haneler bu artışın %22,08'ni ekmek ve tahıllara, %27,36'nı et, balık ve kümes hayvanlarına, %20,76'nı yağ, süt ve günlük yiyeceklere, %17,27'ni taze, kuru meyve ve sebzelere, %2,38'ni çeşitli hazır yiyeceklere, %10,15'ni alkolsüz içeceklerle dağıtmıştır.

Tablo 3 incelendiğinde, gıda alt grubuna ilişkin gelir esneklikleri sırasıyla 0,5115, 0,8104, 0,4324, 0,7682, 0,2017 ve 0,3083 olarak bulunmuştur. Engel yasasına göre, gelir arttığında gıdaya yapılan harcama oranı düşmektedir. Gıda alt gruplarına ilişkin harcamaların gelir esneklikleri, Engel yasasına uygun olarak birden küçük çıkmaktadır. Gelir esnekliğine göre, gıda alt gruplarında yer alan mallar normal mal olarak bulunmuştur.

Tasarrufların gelir esnekliği 1,7446 olarak hesaplanmıştır. Gelirdeki %1 değişme, tasarruflarda %1,7446'lık bir değişime yol açmaktadır.

5. SONUÇ

Gıda harcamasının toplam harcama içerisindeki payı 2003 yılında Türkiye için %27,5, kent için %24,1 ve kırsal için %36,4 iken, 2009 yılında bu oranlar sırasıyla Türkiye için %23,0, kent için %20,3 ve kırsal için %33,9 olarak hesaplanmıştır. Bu verilere göre, 2009 yılında Türkiye, kent ve kırsal yaşayan hanehalklarının gıda harcamalarında ciddi bir azalma söz konusudur.

Bu çalışmada, ELES modeli kullanılarak, Türkiye'de yaşayan hanehalklarının gıda alt gruplarına ilişkin tüketim kalıpları ortaya çıkarılmak istenmiştir. Bu tüketim kalıpları sayesinde hanehalklarının gelire bağlı tasarruf ve tüketim davranışları analiz edilmiştir. Çalışmada mevcut verilerin elde edilebilirliği düşünülerek 16 il çalışmaya dahil edilmiş ve analiz yapılmıştır. Tüketim kalıpları arasındaki farklılığı fiyat, gelir ve tasarruf esneklik değerleri daha iyi yansıtmaktadır. Bu nedenle, çalışmada gelir ve tasarruf esneklik değerleri hesaplanmıştır.

Marjinal bütçe paylarına ait değerler, beklentiye uygun olarak pozitif elde edilmiştir. Engel yasası, gelir arttığında gıdaya yapılan harcama oranının düştüğünü söylemektedir. Gıda alt gruplarına ilişkin harcamaların gelir esneklikleri, Engel yasasına uygun olarak birden küçük çıkmaları beklenmektedir. Gelir esnekliğine göre, gıda alt gruplarında yer alan mallar normal mal olarak bulunmuştur.

Tasarruf gelir esneklik değerlerinin yüksek olması, kişi başına gelir artışları sonucunda ekonominin yüksek miktarda tasarruf doğurma potansiyelinin olduğunu göstermektedir.

Marjinal tüketim eğiliminin azalması, gelir düzeyinin yükselmesi ile açıklanabilmektedir. Bu çalışmada gelir %20'lik dilimlere göre değerlendirilmediği için, marjinal tüketim eğiliminin gelir değişimleri karşısında nasıl bir yol izleyeceği hakkında bir fikir yürütülememektedir.

Yerli ve yabancı literatürde yapılan taramaya göre, genel itibariyle talep modelleri ile ilgili çalışmalarda gelir yerine toplam tüketim harcaması kullanılmıştır. Bu çalışmalara paralel olarak, yaptığımız çalışmada da toplam gıda tüketim harcaması kullanılmıştır. Ayrıca, gelir verilerini de kullanarak analizleri bu şekilde elde etmek faydalı olabilecektir. Türkiye için ileride yapılacak çalışmalarda, sosyo-demografik değişkenlerin etkisinin de araştırılması bu alanda yapılacak olan çalışmalara ışık tutacaktır.

6. KAYNAKLAR

Burney, N. A., Akmal, M., 1991. Food demand in Pakistan: An application of the extended linear expenditure system. *Journal of Agricultural Economics*, 42(2):185-95.

Deaton, A., Castillo, J. R., Thomas, D., 1989. The influence of household composition on household expenditure patterns: Theory and Spanish evidence. *The Journal of Political Economy*, 97: 179-200.

Eastwood, D. B., Craven, J. A., 1981. Food demand and savings in a complete, extended, linear expenditure system. *American Agricultural Economics Association*, 63(3): 544-549.

Ekinci, S., 1996. Türkiye'de bazı gıda maddelerinin talep analizi. Çukurova Üniversitesi (Yüksek Lisans Tezi), 65.

Howe, H., 1977. Cross-section application of linear expenditure systems: Responses to sociodemographic effects. *American Journal of Agricultural Economics*, 59(1):141-148.

Huang, J., Bouis, H., 2001. Structural changes in the demand for food in Asia: Empirical evidence from Taiwan. *Agricultural Economics*, 26:57-69.

Koç, A., 1995. Türkiye'de kırmızı et arz ve talebinin ekonometrik analizi ve kırmızı et sanayi yapısı ile işleyişinin incelenmesi. Çukurova Üniversitesi (Doktora Tezi).

Lee, J., Brown, M. G., Seale, J. L., 1994. Model choice in consumer analysis: Taiwan, 1970-89. *American Journal of Agricultural Economics*, 76(3):504-512.

Lewis, P., Andrews, N., 1989. Household demand in China. *Applied Economics*, 21(6):793-807.

Lluch, C., Williams, R., 1975. Cross country demand and savings patterns: An application of the extended linear expenditure system. *The Review of Economics and Statistics*, 57(3):320-328.

Lluch, C., 1973. The extended linear expenditure system. *European Economic Review*, 4(1):21-32.

Mcconnell, K., 1978. Estimating and testing a complete system of demand functions from regional data. *Applied Economics*, 10(2):93-104.

Narayana, N. S. S., Vani, B. P., 1996. Earnings and consumption by Indian rural laborers: Analysis with an extended linear expenditure system. *Journal of Policy Modeling*, 22(2):255-273.

Nişancı, M., 2002. Kırsal ve kentsel kesimlerde tüketim ve tasarruf kalıpları genişletilmiş doğrusal harcama sistemi uygulaması. *İktisadi ve İdari Bilimler Dergisi*, 16(3):60-73.

Özer, H., 1992. Erzurum'da tüketim harcamalarının ekonometrik analizi. Erzurum: Atatürk Üniversitesi Sosyal Bilimler Enstitüsü (Yayınlanmamış Yüksek Lisans Tezi).

Özer, H., 2001. Türkiye'de hanehalkı tüketim harcamalarının doğrusal harcama sistemi yaklaşımıyla analizi.

Philips, L., 1972. A dynamic version of the linear expenditure model. *The Review of Economics and Statistics*, 54(4):450-458.

Pollak, R. A., Wales, T. J., 1969. Estimation of the linear expenditure system. *Econometrica*, 37(4):611-628.

Sas, 2004. Sas enterprise guide. Sas Institute Inc., Cary, NC, USA.

Sasaki, K., Saegusa Y., 1974. Food demand matrix in an approximate linear expenditure system. *American Journal of Agricultural Economics*, 56(2):263-270.

Stone, R., 1954. Linear expenditure systems and demand analysis: An application to the pattern of British demand. *The Economic Journal*, 64:511-527.

Tansel, A., 1986. An engel curve analysis of household expenditure in Turkey. *METU Studies in Development*, 239-257.

Türkiye İstatistik Kurumu, 2008.

http://www.turkstat.gov.tr/MetaVeri.do?tb_id=22&ust_id=7, 27 Kasım 2008.

FOOD CONSUMPTION ANALYSIS: AN EXTENDED LINEAR EXPENDITURE SYSTEM

ABSTRACT

With information about consumption expenditures obtained from Household Budget Survey it is aimed to determine the items to be included in consumer price indices and the base year weights; to monitor the changes of households' consumption patterns within the time; to compile the data which will help in forecasting private final consumption expenditures used in the national income calculation; to obtain required data for the determination of poverty line and other socio-economic analysis; to obtain necessary information to enable the minimum wage to be determined according to recent conditions. In this study, data of the household income and consumption expenditure surveys conducted by the Turkish Statistical Institute for 2003 year is used. Marginal budget share and income elasticity for food sub groups expenditure are estimated within the framework of the extended linear expenditure system. According to the findings obtained, household consumer behavior is analyzed.

Keywords: An extended linear expenditure system, Marginal budget shares, Income elasticity.

KİŞİ BAŞINA DÜŞEN MİLLİ GELİRE GÖRE ÜLKELERİN GIDA TÜKETİM EĞİLİMLERİNİN TESPİT EDİLMESİ – VERİ MADENCİLİĞİ UYGULAMASI

Fethi Ş. ÖZBEK*

ÖZET

Bu çalışmada, kişi başına düşen milli gelir (GNI) verileri ile günlük kişi başı gıda tüketim verileri kullanılarak ülkelere ait GNI gruplarına göre gıda tüketim eğilimleri, veritabanlarında bilgi keşfi (VTBK) sürecinin en önemli aşaması olan veri madenciliği uygulaması kullanılarak oluşturulan bir model yardımıyla tespit edilmiştir. Modelin oluşturulmasında, SPSS Modeller programı içerisinde yer alan karar ağacı algoritmalarından güçlü sonuçlar veren C5.0 algoritması kullanılmıştır. Oluşturulan model ile GNI gruplarına göre; et, şeker ve tatlandırıcılar, alkollü içecekler, yumurta, kuru baklagiller, sebzeler, diğer su ürünleri, balık, diğer nişastalı gıdalar ve meyveler tüketimi önemli değişkenler olarak bulunmuştur. Model sonuçlarına göre; GNI arttıkça, günlük kişi başı ortalama et, alkollü içecekler, bitkisel yağlar, balık, gıda benzerleri, hayvansal yağlar, meyveler, sakatat, süt, şeker, tatlandırıcılar ve yumurta tüketiminde artış; tahıl ve kuru baklagil tüketiminde ise düşüş olduğu tespit edilmiştir.

Anahtar kelimeler: Gıda tüketimi, GNI, Karar ağaçları, Veri madenciliği.

1. GİRİŞ

Bu çalışma ile ülkelere ait 2005, 2006 ve 2007 yıllarına ait kişi başına düşen milli gelir (GNI) verileri ile aynı ülkelere ve yıllara ait günlük kişi başı gıda tüketim verileri kullanılarak ülkelere ait GNI gruplarına göre gıda tüketim eğilimleri, veritabanlarında bilgi keşfi (VTBK) süreci kullanılarak tespit edilmiştir. Böylelikle; yüksek, orta ve düşük gelire sahip ülkelere ait gıda tüketim eğilimleri ortaya konmuştur.

Verilerden doğru, faydalı, yeni ve anlaşılır modeller ve kalıplar elde etmekte kullanılan VTBK; problemin tanımlanması, verilerin hazırlanması, modelin kurulması ve değerlendirilmesi, modelin kullanılması ve modelin izlenmesi aşamalarından oluşmaktadır. Verilerin hazırlanması aşaması; farklı kaynaklardan verilerin toplanması, zaman, kod, ölçü, vb. farklılıkları ortadan kaldırmak amacıyla verilere değer biçilmesi, farklı veri kaynaklarının birleştirilmesi, ihtiyaç duyulmayan verilerin temizlenmesi, modelde kullanılacak bağımlı ve bağımsız verilerin belirlenmesi, ihtiyaç duyulduğunda veri yığınının örnek seçilmesi ve veriler arasında gerekli dönüşümlerin yapılması kısımlarından oluşmaktadır (Batmaz ve Köksal, 2010; Çingı, 2010; Han ve Kamber 2006).

VTBK aşamalarından en önemlisini, modelin kurulması ve değerlendirilmesi kısımlarını içeren veri madenciliği aşaması oluşturmaktadır. Önceden bilinmeyen geçerli ve uygulanabilir bilginin veri yığınlarından dinamik bir süreç ile elde edilmesi olarak tanımlanan veri madenciliği, büyük veri yığınlarından oluşan veritabanları içerisinde gizli kalmış bilgilerin tespit edilmesine olanak sağlamaktadır (Baykal, 2006).

* TÜİK Uzmanı, Türkiye İstatistik Kurumu, e-posta: fethiozbek@tuik.gov.tr

Bu çalışmada, GNI gruplarına göre gıda tüketim eğilimlerinin tespit edilmesinde, veri madenciliği uygulamasında temel sınıflandırma tekniklerinden olan karar ağacı uygulaması kullanılmıştır. Karar ağacı ile veri kümesi, belirlenen girdi değişkenlerine göre alt gruplara ayrılmaktadır.

Daha önce yapılan çeşitli araştırmalarda, ekonomik gelişmişlik ile tüketilen gıda türleri arasındaki güçlü ilişki araştırmacılar tarafından ortaya konmuştur. Bu araştırmalardan Leenes (2006) tarafından yapılan çalışmada, gelişmiş ülkelerin hayvansal ürün tüketimlerinin ekonomisi güçsüz olan ülkelere göre daha fazla olduğu tespit edilmiştir. Mancini vd. (2009) tarafından yapılan çalışmada; Almanya, Belçika, Avusturya gibi GNI'si yüksek olan ülkelere göre kişi başına düşen hayvansal ürün tüketiminin, Romanya, Bulgaristan, Polonya gibi GNI'si düşük olan ülkelere göre daha yüksek; tahıl ve patates tüketiminin ise bu ülkelere göre daha düşük olduğu tespit edilmiştir.

2. MATERYAL VE YÖNTEM

2.1 Materyal

Bu çalışmanın ana materyalini, Dünya Tarım Örgütü (FAO) veritabanından elde edilen ülkelere ait kişi başı günlük gıda arzı verileri (FAO, 2010) ile Dünya Bankası (WB) veritabanından elde edilen ülkelere ait GNI verileri (WB, 2010b) oluşturmaktadır. FAO veritabanında yer alan gıda arzı verileri birincil bitkisel ve hayvansal ürün eşdeğerlerinden oluşmaktadır. Bu kısımda yer alan veriler, ülkelere ait ürün bazında oluşturulan gıda denge tablolarından elde edilmektedir. Belirli bir ürüne ait denge tablosundan elde edilen gıda arzı, birincil ürün ile bu üründen türetilen işlenmiş ürünlerin hammadde eşdeğeri toplamına karşılık gelmektedir. Örneğin, elma için gıda arzı yalnızca birincil ürün olarak tüketime sunulan elmayı değil elmanın işlenmesi ile elde edilen işlenmiş üründeki elma miktarını da kapsamaktadır.

Ülkelere ait gıda tüketim eğilimlerinin tespit edilmesi amacıyla FAO veritabanında yer alan gıda tüketim gruplarının tamamı çalışma kapsamına dahil edilmiştir. Bu gruplar Tablo 1'de sunulmaktadır.

Tablo 1. Çalışma kapsamında kullanılan gıda tüketim grupları

Bitkisel ürünler		Hayvansal ürünler
Alkollü içecekler	Gıda benzerleri	Hayvansal yağlar
Tahıllar	Şeker ve tatlandırıcılar	Su ürünleri, diğer
Meyveler	Sert kabuklular	Yumurta
Yağlı tohumlar	Bitkisel yağlar	Balık
Kuru baklagiller	Sebzeler	Et
Baharatlar		Süt
Niştastalı gıdalar, diğer		Sakatat

Çalışmada, ilgili veritabanlarında tüm gıda gruplarına ait bitkisel ve hayvansal gıda verileri mevcut olan ülkeler kapsanmıştır. Buna göre, yüksek GNI grubunda 26, yüksek orta GNI grubunda 23, düşük orta GNI grubunda 28 ve düşük GNI grubunda 13 olmak

üzere toplam 90 ülkeye ait 2005, 2006 ve 2007 verileri çalışma kapsamında kullanılmıştır.

2.2 Yöntem

Çalışmada kullanılan model, problemin tanımlanması, verilerin hazırlanması, modelin kurulması ve değerlendirilmesi, modelin kullanılması ve modelin izlenmesi aşamalarından oluşan VTBK süreci izlenerek oluşturulmuştur. VTBK sürecinde izlenen aşamalar şu şekildedir:

2.2.1 Problemin tanımlanması

VTBK sürecinin ilk aşaması olan bu aşamada, çalışma konusu; kişi başı milli gelir gruplarına göre ülkelerin gıda tüketim eğilimlerinin tespit edilmesi olarak belirlenmiştir.

2.2.2 Verilerin hazırlanması

Verilerin hazırlanması aşaması; verilerin toplanması, verilere değer biçilmesi, birleştirme ve temizleme, seçim ve dönüştürme kısımlarından oluşmaktadır. Verilerin toplanması aşamasında belirlenen amaca ulaşmak için gerekli veriler ve bu verilerin toplanacağı veri kaynakları tespit edilmektedir. Çalışmada kullanılan veriler; ülkelere ait gıda tüketim verileri ve GNI olarak belirlenmiştir. Bu veriler, FAO ve WB veri tabanlarından elde edilmiştir.

Araştırmada kullanılan verilerde zaman, kodlama ve ölçü birimi farklılıkları olmaması nedeniyle verilere değer biçme işlemi uygulanmamıştır.

Birleştirme ve temizleme aşamasında, FAO veri tabanından elde edilen gıda tüketim verileri, WB veritabanından elde edilen GNI verileri ile birleştirilmiştir. Bu birleştirme sonrasında eksik verileri olan ülkelere ait kayıtlar veri kümesinden çıkartılmıştır.

Seçim aşamasında, kurulan modelde kullanılan hedef değişken GNI, girdi değişkenleri ise gıda tüketim grupları olarak belirlenmiştir. Modelde kullanılan veri kümesinden herhangi bir örnek seçimi yapılmamıştır.

Çalışmada kullanılan C5.0 algoritmasında sembolik tipte hedef değişken kullanılmaktadır. Bu nedenle dönüştürme aşamasında, sayısal tipteki hedef değişken değerleri gruplandırılarak sembolik hale dönüştürülmüştür. Hedef değişken grubu olan GNI; yüksek (H), yüksek orta (UM), düşük orta (LM) ve düşük (L) olarak sınıflandırılmıştır. Bu sınıflandırmada, Dünya Bankası'nın gelire göre ülke sınıflandırması değerleri kullanılmıştır (WB, 2010a).

2.2.3 Modelin kurulması ve değerlendirilmesi

Bazı sosyo-ekonomik değişkenlere göre ülkelerin gıda tüketim eğilimlerinin tespit edilmesinde SPSS Modeler programı içerisinde yer alan karar ağacı algoritmalarından güçlü sonuçlar veren C5.0 algoritması kullanılmıştır. Tahmin edici ve tanımlayıcı özelliklere sahip olan karar ağaçları, veri madenciliğinde kuruluşlarının ucuz olması, yorumlanmalarının kolay olması, veri tabanı sistemleri ile kolayca entegre

edilebilmeleri ve güvenilirliklerinin daha iyi olması nedenlerinden dolayı sınıflama modelleri içerisinde yaygın bir kullanım alanına sahiptir (Akpınar, 2000).

1998 yılında Ross Quinlan tarafından önerilen C5.0 algoritması, diğer bir karar ağacı algoritması olan C4.5 algoritmasının geliştirilmiş bir halidir. Bilgi kazanımı kavramını kullanarak ağaç oluşumu sağlayan bu algoritma; hız, etkin hafıza kullanımı, daha küçük ağaç oluşumu, budama, ağırlıklandırma ve daha kesin sonuçlar verme gibi özellikleri sayesinde C4.5 algoritmasına göre daha gelişmiş bir algoritmadır (Güntürkün, 2007).

Örneklerin kökten yapraklara doğru sıralanarak sınıflandırıldığı karar ağacı algoritmasında, ilk olarak ağacın kökünde hangi değişken ile test edilmesi belirlenmekte daha sonra yukarıdan aşağı doğru ağaç oluşturulmaktadır. Karar ağacı algoritmasında her düğümde hangi değişkenin yer alacağı, değişkenlere ait bilgi kazanımı (information gain) ölçüsü ile belirlenmekte olup en yüksek bilgi kazanımı değerine sahip değişken ağacın kökünde yer almaktadır. Bilgi kazanımı entropi kullanılarak hesaplanmaktadır. Entropi, örneklerin keyfi olarak toplanmasının kirliliğini tanımlamaktadır (Kök ve Kuloğlu, 2005). Bir değişkene ait entropinin yüksek olması; ilgili değişkenin tekdüze dağılım gösterdiği anlamına gelmektedir. Hedef değişkenin alabileceği değerleri içeren S topluluğunun entropi değeri (1)'de yer alan formül yardımıyla aşağıdaki gibi hesaplanmaktadır (Moore, 2003):

$$\text{Entropi}(S) = - \sum_{j=1}^m p_j \log_2 p_j = -p_1 \log_2 p_1 - p_2 \log_2 p_2 - \dots - p_m \log_2 p_m \quad (1)$$

V: X'in alabileceği değerler

$$p_1 = P(X=V_1)$$

$$p_2 = P(X=V_2)$$

$$p_3 = P(X=V_3)$$

...

$$p_m = P(X=V_m)$$

Bilgi kazanımı değeri ise (2)'de yer alan formül yardımıyla aşağıdaki gibi hesaplanmaktadır.

$$\text{KAZANIM}(S,A) = \text{Entropi}(S) - \sum_v (|S_v|/|S|) \times \text{Entropi}(S_v) \quad (2)$$

Kazanım (S,A): A değişkenine ait bilgi kazanımı değeri

v : A değişkeninin alabileceği tüm değerler

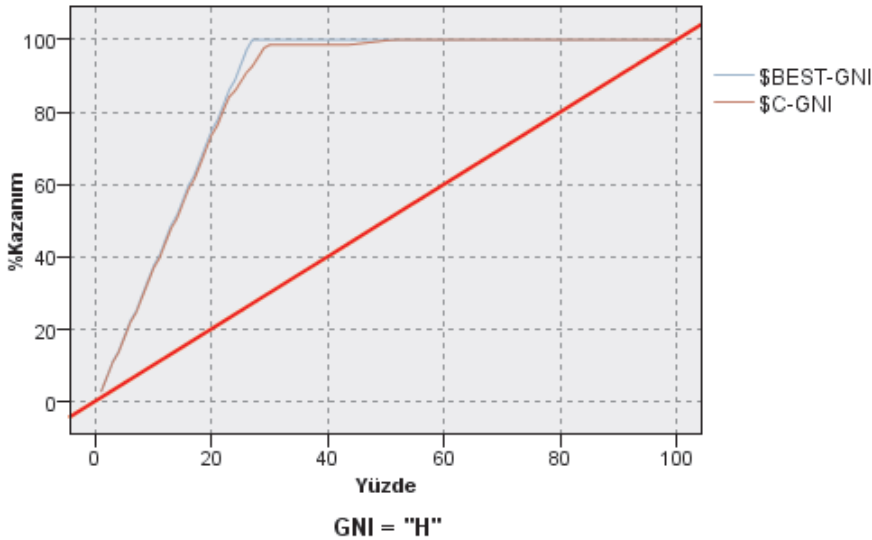
Karar ağacı algoritması kullanılarak kurulan modelin değerlendirilmesinde, değerlendirme grafikleri (kazanım grafiği), uygunluk matrisleri ve çapraz geçerlilik uygulamaları kullanılmıştır.

2.2.4 Modelin kullanılması

Model sonuçlarına göre GNI ile ülkelerin gıda tüketim eğilimleri tespit edilmiştir. Böylelikle, GNI gruplarına göre tüketimi fazla ve az gıda grupları belirlenmiştir. Oluşturulan model, gelir seviyesine göre tüketilen gıda gruplarının analiz edilmesi amacı ile kullanılabilir.

3. BULGULAR

SPSS Modeller kullanılarak C5.0 algoritması ile oluşturulan modelin doğruluğunun incelenmesi amacıyla oluşturulan değerlendirme grafiği Şekil 1'de sunulmaktadır. Grafikteki dikey eksen, hedef olarak seçilen değerlerin birikimli yüzdelerini gösterirken, yatay eksen model tahmini ve güven oranına göre sıralanmış yüzdeleri göstermektedir. Grafikte, köşegen doğrusu; hedef değişken rastgele bir seviyede tahmin edildiğinde beklenen temel oranı, en üstte bulunan eğri en iyi model veri kümesine uygulandığındaki sonuçları, ortadaki eğri ise model sonuçlarını göstermektedir. Grafikten de görüleceği gibi en iyi eğri ile model eğrisi birbirine oldukça yakındır, bu durum oluşturulan modelin doğruluğunun oldukça iyi olduğunu ifade etmektedir.



Şekil 1. Oluşturulan modele ait değerlendirme grafiği

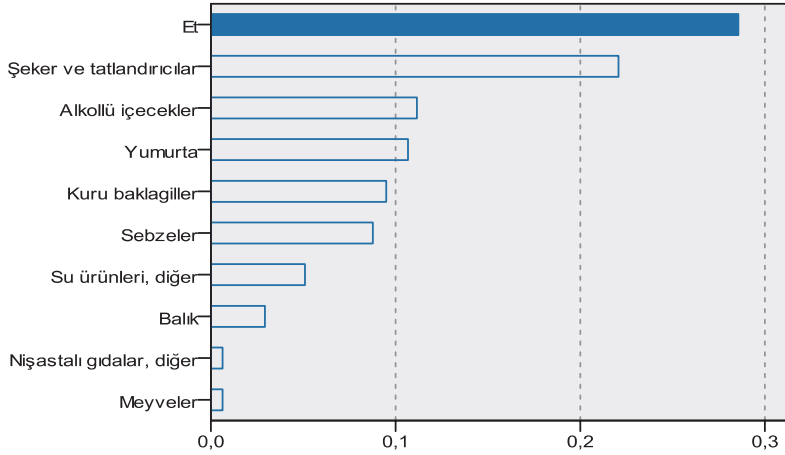
Analiz doğruluk oranı %97,28 olan modelin, gerçek değerlerle tahmin değerlerinin tutarlılığının kontrol edilmesi amacıyla oluşturulan çapraz tablo sonuçlarına göre yüksek ve düşük GNI gruplarının tamamı, yüksek orta ve düşük orta gruplarının ise tamamına yakını model tarafından doğru olarak tahmin edilmiştir. GNI gruplarına göre model tahmin sonuçları Tablo 2'de sunulmaktadır.

Tablo 2. GNI gruplarına göre model tahmin sonuçları (%)

GNI	H	L	LM	UM
H	100,0	0,0	0,0	0,0
L	0,0	100,0	0,0	0,0
LM	0,0	1,6	96,8	1,6
UM	2,7	2,7	2,7	91,9

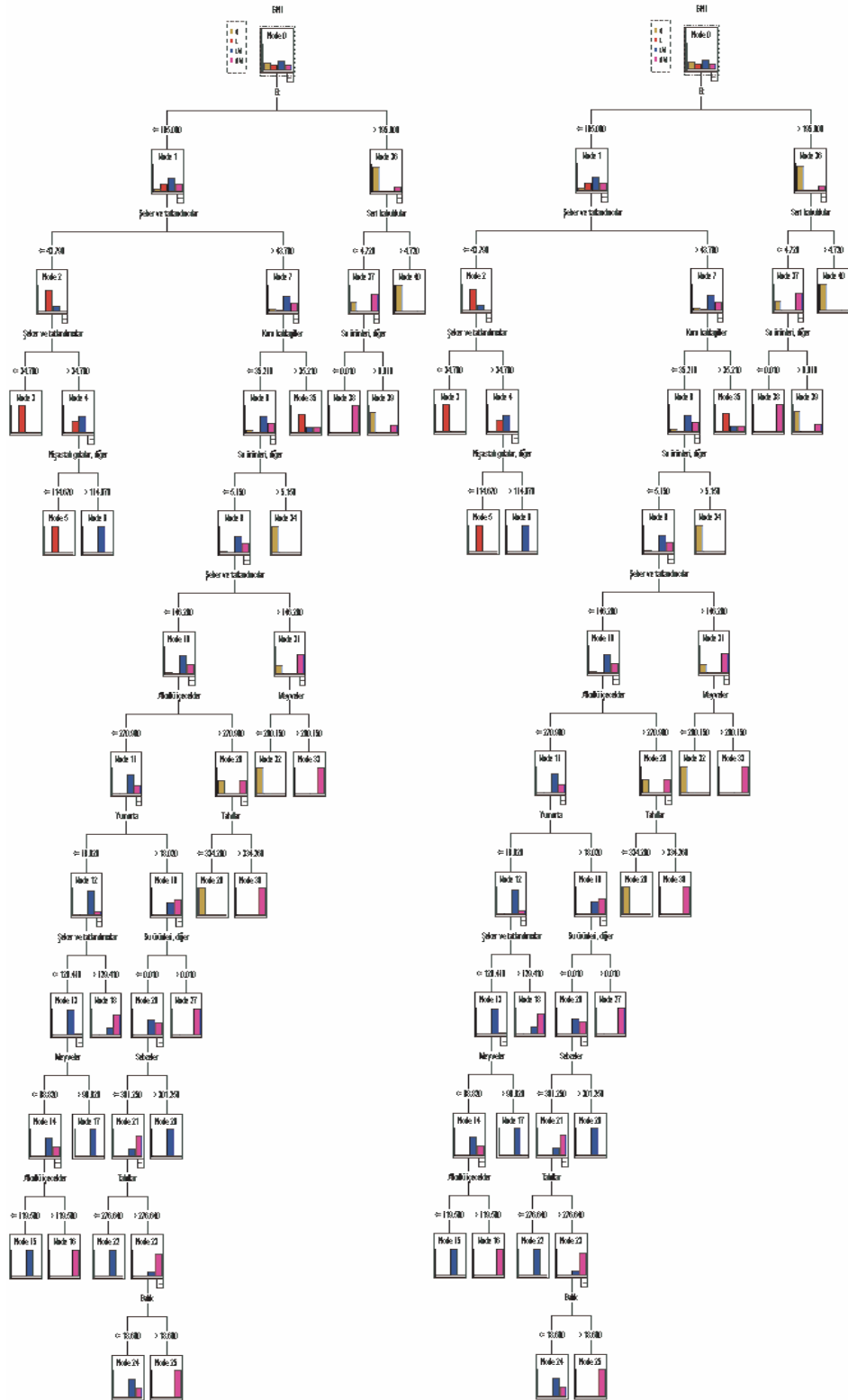
Oluşturulan model ile GNI gruplarına göre et, şeker ve tatlandırıcılar, alkollü içecekler, yumurta, kuru baklagiller, sebzeler, diğer su ürünleri, balık, diğer nişastalı gıdalar ve meyve tüketimi önemli değişkenler olarak bulunmuştur. GNI gruplarına göre önem arz eden değişkenler Şekil 2’de sunulmaktadır.

Önemli değişkenler



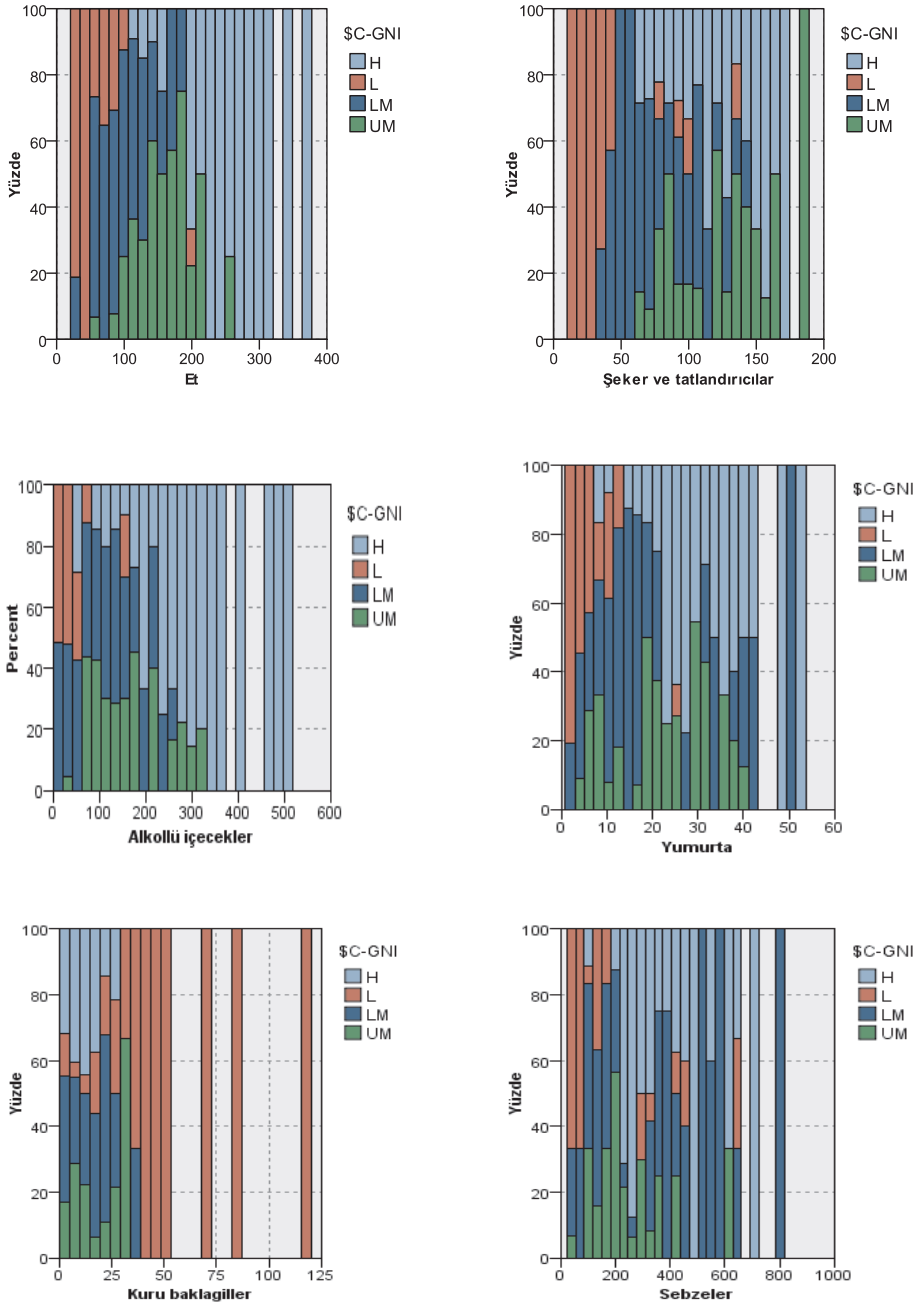
Şekil 2. GNI gruplarına göre tüketimi önem arz eden değişkenler

Oluşturulan ağaçta ilk bölünme “Et” değişkeninde olmaktadır. Et tüketim miktarının belirli bir değerin üstünde olması durumunda sonraki bölünmeler sert kabuklular ve diğer su ürünleri şeklinde olmaktadır. Et tüketim miktarının belirli bir değerin altında olması durumunda ise sonraki bölünmeler şeker ve tatlandırıcılar, kuru baklagiller, diğer su ürünleri ve diğer nişastalı gıdalar olarak devam etmektedir. Model çıktılarının yer aldığı Şekil 3’den de görüldüğü gibi günlük kişi başı 195 gramın üstünde et tüketimi olan ülkeler, yüksek ya da yüksek orta GNI grubunda yer almaktadır. Günlük kişi başı et tüketimi 195 gramın altında olan ülkelerden günlük kişi başı şeker ve tatlandırıcı tüketimi 34,76 gramın altında olan ülkeler düşük GNI grubunda yer almaktadır. Günlük kişi başı şeker ve tatlandırıcı tüketimi 34,76 ve 43,79 arasında olan ülkeler ise günlük diğer nişastalı gıdalar tüketimine göre düşük veya düşük orta GNI grubunda yer almaktadır. Günlük kişi başı şeker ve tatlandırıcı tüketimi 43,79 gramın üzerinde olan ülkeler ise kuru baklagiller ve diğer su ürünleri tüketimine göre gruplandırılmaktadır.



Şekil 3. Oluşturulan karar ağacının gösterimi

GNI'ye göre gruplanan ülkelerde önemli gıda tüketim değişkenlerinden olan et, şeker ve tatlandırıcılar, alkollü içecekler, yumurta, kuru baklagiller ve sebze tüketiminin GNI gruplarına göre histogram grafikleri Şekil 4'de sunulmaktadır. Grafikten de görüleceği gibi yüksek GNI gruplarında et, şeker ve tatlandırıcılar, alkollü içecekler, yumurta, tüketimi fazla, kuru baklagiller tüketimi ise düşüktür.



Şekil 4. Gıda tüketimlerinin GNI gruplarına göre histogram grafikleri

Ülkelerin gıda tüketim gruplarının GNI grupları ayrımında ortalamaları Tablo 3'de sunulmaktadır. Buna göre et, alkollü içecekler, bitkisel yağlar, balık, gıda benzerleri, hayvansal yağlar, meyveler, sakatat, süt, şeker, tatlandırıcılar ve yumurtada GNI azaldıkça günlük kişi başı ortalama tüketim azalmaktadır. GNI azaldıkça, tahıl ve kuru baklagil tüketiminde ise artış gözlemlenmektedir. Düşük GNI grubunda günlük kişi başı ortalama diğer su ürünleri, hayvansal yağlar, sert kabuklular, alkollü içecekler, gıda benzerleri, yumurta, süt, et ve meyve tüketimi; yüksek GNI grubundaki tüketime göre sırasıyla %95, %92, %92, %89, %87, %86, %79, %78 ve %77 azalma göstermektedir.

Tablo 3. Ülkelerin gıda tüketim gruplarının GNI grupları ayrımında ortalamaları (g/gün/kişi)

Değişken adı	H	UM	LM	L
Alkollü içecekler	273,857	160,438	78,175	28,878
Baharatlar	1,880	3,368	3,154	0,895
Balık	77,931	52,576	53,213	19,501
Bitkisel yağlar	47,466	36,006	25,798	21,934
Et	241,541	147,958	97,111	53,118
Gıda benzerleri	24,766	13,672	9,471	3,133
Hayvansal yağlar	30,295	12,085	7,302	2,288
Kuru baklagiller	9,648	13,400	13,183	27,813
Meyveler	313,808	230,566	210,939	70,976
Niştastalı gıdalar, diğer	183,421	156,530	200,529	248,133
Sakatat	11,168	7,730	7,585	6,210
Sebzeler	325,058	223,399	252,421	125,675
Sert kabuklular	13,077	3,959	5,782	1,073
Su ürünleri, diğer	0,600	0,172	0,119	0,031
Süt	607,766	390,595	256,063	128,126
Şeker ve tatlandırıcılar	119,071	108,989	83,121	34,650
Tahıllar	322,386	377,939	391,670	432,508
Yağlı tohumlar	13,214	16,876	43,770	26,009
Yumurta	30,228	22,634	16,195	4,238

4. TARTIŞMA VE SONUÇ

Bu çalışmada kullanım alanı gün geçtikçe yaygınlaşan veri madenciliği uygulaması kullanılarak karar ağacı algoritması ile oluşturulan bir model yardımıyla ülkelere ait GNI gruplarına göre gıda tüketim eğilimleri tespit edilmiştir. Analiz doğruluk oranı %97,28 olan modelde, hedef değişkenin tamamına yakını doğru olarak tahmin edilmiştir.

Oluşturulan modelde; şeker ve tatlandırıcılar, alkollü içecekler, yumurta, kuru baklagiller, sebzeler, diğer su ürünleri, balık, diğer nişastalı gıdalar ve meyve tüketimi önemli değişkenler olarak bulunmuştur. GNI'si yüksek olan ülkelerde, aynı zamanda karar ağacının ilk bölünmesi olan et tüketiminin belirli bir değerin üstünde olduğu tespit edilmiştir.

Model sonuçlarına göre; GNI'si düşük ülkelerde, günlük kişi başı ortalama et, alkollü içecekler, bitkisel yağlar, balık, gıda benzerleri, hayvansal yağlar, meyveler, sakatat, süt, şeker, tatlandırıcılar ve yumurta tüketiminin düşük; tahıl ve kuru baklagil tüketiminin ise yüksek olduğu tespit edilmiştir. Çalışmadan elde edilen bu sonuçlar, ayrıntıları "Giriş" kısmında verilen Leenes (2006) ve Mancini vd. (2009) tarafından yapılan çalışmalardan elde edilen sonuçlarla tutarlılık göstermektedir.

Bu sonuçlar ışığında, gelir seviyesi düşük olan ülkelerde yüksek fiyatları nedeniyle et, süt, meyve, yumurta gibi besleyici yönü güçlü gıdaların tüketiminin düşük olduğu, tahıl ve kuru baklagil gibi fiyatı düşük ürünlerin tüketiminin ise yüksek olduğu anlaşılmaktadır. Bu durum, gelir seviyesi düşük olan ülkelerde beslenme yetersizliğinden kaynaklı sorunlara neden olmaktadır.

Çalışma sonuçlarına göre gelir seviyesi yüksek olan ülkelerde; şeker, tatlandırıcılar, alkollü içecekler, bitkisel ve hayvansal yağ tüketiminin yüksek olduğu tespit edilmiştir. Bu gıda gruplarının fazla tüketiminin gelir seviyesi yüksek olan ülkelerde başta obezite olmak üzere dengesiz beslenmeden kaynaklanan çeşitli sorunlara yol açtığı bilinmektedir. Ayrıca, obezite ve fazla alkol tüketiminin, gelişmiş ülkelerde en sık rastlanan ölüm nedeni olan kardiyovasküler hastalıklar için risk faktörleri arasında yer aldığı bilinmektedir (Samur ve Yıldız, 2008) .

5. KAYNAKLAR

Akpınar, H., 2000. Veri tabanlarında bilgi keşfi ve veri madenciliği. İ.Ü. İşletme Fakültesi Dergisi, 29, 1-22.

Batmaz, İ., Köksal, G., 2010. "Overview of knowledge discovery in databases process and data mining for surveillance technologies and early warning systems", in *Surveillance Technologies and Early Warning Systems: Data Mining Applications for Risk Detection*. Koyuncugil, A. S., Özgülbaş, N. (editors), IGI Global Publisher, (basım aşamasında).

Baykal, A., 2006. Veri madenciliği uygulama alanları. D.Ü. Ziya Gökalp Eğitim Fakültesi Dergisi, 7, 95-107.

Çingir, H., 2010. Veri madenciliğine giriş–ders notları, <http://yunus.hacettepe.edu.tr/~hcingir/ist376a/6Bolum.doc>, 15 Ağustos 2010.

FAO, 2010. <http://faostat.fao.org/site/345/default.aspx>, 01 Ağustos 2010.

Güntürkün, F., 2007. A comprehensive review of data mining applications in quality improvement and a case study. O.D.T.Ü., Ankara (in English).

Han, J. Kamber M., 2006. Data Mining: Concepts and Techniques. 2nd ed., Morgan Kaufmann Publishers, USA.

Kök, B. V., Kuloğlu, N., 2005. Sollama esnasında taşıt ve yol ile ilgili faktörlerin karar ağacı yöntemiyle irdelenmesi. Erciyes Üniversitesi Fen Bilimleri Enstitüsü Dergisi, 21, 180-188.

Leenes, P. W., 2006. Natural resource use for food: land, water and energy in production and consumption systems. Groningen, Holland.

Mancini, L., Lettenmeier M., Rohn, H., Liedtke C., 2009. Using MIP methodology to evaluate food-farming systems sustainability. http://fromGNItoWellbeing.univpm.it/doc/papers/GNI2WB_034.pdf, 03 Eylül 2010.

Moore, A. W., 2003. Information gain. <http://www.autonlab.org/tutorials/infogain11.pdf>, 15 Eylül 2010.

Samur, G., Yıldız, E., 2008. Obezite ve kardiyovasküler hastalıklar/Hipertansiyon. http://www.beslenme.saglik.gov.tr/content/files/yayinlar/kitaplar/obezite_bilgi_serisi/D6.pdf, 08 Aralık 2010.

WB, 2010a. Ülke sınıflamaları. <http://data.worldbank.org/about/country-classifications>, 01 Eylül 2010.

WB, 2010b. <http://databank.worldbank.org/ddp/home.do?Step=2&id=4&DisplayAggregation=N&SdmxSupported=Y&CNO=2>, 14 Eylül 2010.

DETERMINING COUNTRIES' FOOD CONSUMPTION PATTERNS WITH RESPECT TO NATIONAL INCOME PER CAPITA - A DATA MINING APPLICATION

ABSTRACT

Food consumption patterns of countries by gross national income (GNI) groups are determined with a model by using GNI per capita and daily food consumption per capita data. The model was formed with data mining application, the most important stage of the knowledge discovery in databases. C5.0, one of the decision tree's algorithms giving powerful results in SPSS Modeler, was used to form the model. According to the model, the consumption of meat, sugar and sweeteners, alcoholic beverages, eggs, pulses, vegetables, other aquatic products, fish, other starchy foods and fruits are determined to be important variables by GNI groups. It is also determined that when GNI increases, daily average consumptions per capita of meat, alcoholic beverages, vegetable oils, fish, stimulants, animal fats, fruits, offal, milk, sugar and sweeteners and eggs increase; while those of cereals and pulses decrease.

Keywords: Data mining, Decision trees, Food consumption, GNI.

DETERMINATION OF THE NUMBER OF BINS/CLASSES USED IN HISTOGRAMS AND FREQUENCY TABLES: A SHORT BIBLIOGRAPHY

Nurhan DOĞAN*

İsmet DOĞAN**

ABSTRACT

The histogram is the oldest and most popular tool for graphical display of a univariate set of data. An important parameter that needs to be specified when constructing a histogram is the bin width (also called the interval width or class width). This is simply the length of the subintervals of the real line, sometimes called "bins" or "intervals" (also called "classes"), on which the histogram is based. Frequency distributions facilitate the organization and presentation of data. A major issue with all classifying techniques is how to select the number of classes. There is no "correct" answer for every set of data. Each case must be treated separately; each frequency table must be designed individually. The number of bins / classes increases as the sample size increases. Larson's (1975) formula had the lowest number of bins / classes and the Ishikawa's (1986) formula had the highest one for $n > 300$. The aim of this study is to give a short bibliography for bins/class numbers.

Keywords: Bin width, Class number, Frequency tables, Histogram.

1. INTRODUCTION

Statistics is the science of assembling, classifying, tabulating and analyzing numerical facts or data. A major issue with all tabulating and classifying techniques is how to select the number of classes. Questions are frequently asked about how many classes are useful and what size they should be. There are many criteria and guidelines for approaching the problem. Unfortunately, no standart, objective selection procedure exists. There is no "correct" answer for every set of data. Each case must be treated separately; each frequency table must be designed individually. More detail can be shown in larger number of classes; however, if the number of classes is too large, the classification loses its effectiveness as a means of summarizing data. On the other hand, if the number of classes is too small, the data will be condensed so much that little or no insight can be gained into the nature of the pattern of variation (Plane and Oppermann, 1981).

Frequency distributions facilitate organization and presentation of data. It is generally recognized that the process of tabulating data in classes may give rise to a considerable degree of distortion of the original data. Even when the class limits are chosen with the utmost care, so as to secure an adequate scatter in each class, the measures of central tendency, dispersion and the type of distribution may be none too dependable. Nevertheless, frequency classifications are indispensable, both as means of summing up a large array of data compactly and as a form of generalization.

* Assist.Prof., Afyon Kocatepe University, Medical Faculty, Biostatistics Department, Afyonkarahisar, Turkey, e-mail: ndogan@aku.edu.tr

** Prof., Afyon Kocatepe University, Medical Faculty, Biostatistics Department, Afyonkarahisar, Turkey, e-mail: dogan@aku.edu.tr

Very often it happens that the statistician must work with tabulations made up from original sources which he cannot conveniently consult. Hence, the analysis of frequency distributions is likely to remain one of the important branches of statistics (Davies, 1929).

The composition of a numerical random sample is conveniently pictured by its histogram. A histogram conveys visual information of both the frequency and relative frequencies of observations; that is the essence of a density function (Scott, 1992). For many classes of data one expects the underlying population to be approximately normal and hence the histogram of the sample also to be approximately normal. If so, it may be further convenient to smooth the histogram by approximating it by a suitable normal density curve (Brown and Hwang, 1993).

The framework of the classical histogram is useful for conveying the general flavor of nonparametric theory and practise. The histogram is most often displayed on a nondensity scale: either as bin counts or as a stem-and-leaf plot. The classical frequency histogram is formed by constructing a complete set of nonoverlapping intervals, called bins, and counting the number of points in each bin. In order for the bin counts to be comparable, the bins should all have the same width (Scott, 1992).

The histogram is the classical nonparametric density estimator, probably dating from the mortality studies of John Graunt in 1662. Today the histogram remains an important statistical tool for displaying and summarizing data. In addition, it provides a consistent estimate of the true underlying probability density function (Scott, 1979).

The histogram is the oldest and most popular tool for graphical display of a univariate set of data. It is taught in virtually all elementary data analysis courses and available in most statistical computing packages. An important parameter that needs to be specified when constructing a histogram is the bin width (also called the interval width or class width). This is simply the length of the subintervals of the real line, sometimes called “bins” or “intervals” (also called “classes”), on which the histogram is based. It is not very difficult to see that the choice of the bin width has an enormous effect on the appearance of the resulting histogram. The choice of a very small bin width results in a jagged histogram, with a separate block for each distinct observation. On the other hand, a very large bin width results in a histogram with a single block. Intermediate bin widths lead to a variety of histogram shapes between these two extremes. Ideally, the bin width should be chosen so that the histogram displays the essential structure of the data, without giving too much credence to the data set at hand. Essentially it amounts to choosing the bin width (h);

$$h = \frac{\text{range of data}}{\text{number of class}} \quad (\text{Wand, 1997}).$$

The choice of bin width is directly related to the number of classes. It is best if all bins are of the same width. However, in certain cases, unequal bin width must be used (Plane and Oppermann, 1981).

The histogram is a statistical technique with a long history. Unfortunately, there exist only a few explicit guidelines, which are based on statistical theory, for choosing the number of bins that appear in the histogram (He and Meeden, 1991).

Since the well-known formula by Sturges (1926), many authors have proposed rules for choosing the bin-width (class width) or the number of bins. These rules have always been regarding the histogram as an estimator of an underlying density and relying on large- n asymptotics. Histograms constructed according to these rules have been compared with other density estimators, notably kernel and spline estimators (Hirai, 1990; Denby and Mallows, 2009).

To construct a frequency table or a histogram for grouped data, most introductory statistics books agree on determining the four crucial elements. These four elements are: The range, the number of class intervals, the class interval width and the starting point (Lohaka, 2007).

How many classes? There are too many numbers being recommended for the number of classes. To start with, a number of textbooks recommended as a general rule that 10 classes to be taken as optimal and 30 as the maximum. A few others urged between 10 and 20 classes. According to some other authors, the number of classes can be situated either between 10 and 14, or 5 and 15. Besides these suggestions, some formulas have been developed in choosing an appropriate number of classes, which is denoted by the letter k .

Table 1. Formulas for the number of bins / classes

Order	k	Reference	Order	k	Reference
1	$1 + \lceil 3.3 \times \log_{10}(n) \rceil$	Sturges, 1926	13	Rudemo suggested a cross validation technique for selecting the number of bins.	Rudemo, 1982
2	$4 \times \sqrt[3]{2 \times (n-1)^2 / c^2}$ $c = \frac{1}{\sqrt{2\pi}} \int_c^\infty e^{-\frac{1}{2}t^2} dt$	Mann and Wald, 1942	14	Gislason and Goldfield proposed in their study a simple method for determining the optimum number of bins to be used in a histogrammic representation of a function of a continuous variable.	Gislason and Goldfield, 1984
3	$\sqrt{\frac{n}{5}}$	Cohran, 1954	15	Suzuki used a rounded number for the class interval and also the endpoint of a class.	Suzuki, 1985
4	$\sqrt[3]{n}$	Cencov, 1962	16		Terrel and Scott, 1985
5	$1.87 \times (n-1)^{0.4}$	Bendat and Piersol, 1966	17		Ishikawa, 1986
6	Mori proposed a method by minimizing the mean squared error of a histogram estimate \hat{f}_n of the true density $f(x)$.	Mori, 1974	18	Using minimum description length principle, Rissanen estimated the density.	Rissanen, 1992
7	$1 + \lceil 2.2 \times \log_{10}(n) \rceil$	Larson, 1975	19	In their note, they gave a decision theoretic approach (using loss function and stepwise Bayes rule based on the Bayesian bootstrap) to the problem of choosing the number of bins in a histogram.	He and Meeden, 1997
8	$1 + \log_2 n + \log_2 \gamma$ $\gamma = \sqrt{\frac{(n+1) \times (n+3) \times \sum_{i=1}^n (x_i - \bar{x})^2}{6 \times (n-2) \times \sum_{i=1}^n (x_i - \bar{x})^2}}$	Doane, 1976	20	Using Bayesian probability theory, Knuth derived a straightforward algorithm that computed the posterior probability of the number of bins for a given data set.	Knuth, 2006
9	$2 \times \sqrt{n}$ for $n \leq 100$ $10 \times \log_{10} n$ for $n \geq 100$	Velleman, 1976	21*	$2 \times \sqrt[3]{n}$	Rice
10	\sqrt{n}	Mosteller and Tukey, 1977	22*	$2.5 \times \sqrt[4]{n}$	Anonymous1
11	$\lceil (\text{range of data} \times \sqrt[3]{n}) / (3.49 \times \sigma) \rceil$ $\sigma = \text{standard deviation of sample}$	Scott, 1979	23*	$2^k \geq n$	Anonymous2
12	$\lceil \text{range of data} / 2 \times (IQ) \times n^{-1/3} \rceil$ $IQ(\text{Interquartile}) = Q_3 - Q_1$	Freedman and Diaconis, 1981		k is the smallest integer number of classes and n being the number of observations	

* Directly taken from Lohaka, 2007.

Sturges proposed a simple rule for classifying a series of n item. Sturges' rule probably survived as long as it did because, for moderate n (less than 200), it gives similar results to the alternative rules, and so produces reasonable histograms. However, it does not work for large n (Hyndman, 1995).

The expectation is that a normally distributed variable can be appropriately divided so that the class frequencies comprise a binomial series for all n which are even powers of 2. Sturges' unambiguous rule has become a guideline for researchers, even when it is inappropriate. Sample data are seldom symmetric, let alone normally distributed. Sturges' rule does not always provide enough classes to reveal the shape of a severely skewed distribution. At a minimum, the realworld researcher would want to modify Sturges' Rule to reflect skewness. The statistic;

$$\sqrt{b_1} = \frac{\sum (x - \bar{x})^2}{[\sum (x - \bar{x})^2]^{3/2}}$$

is a well known measure of departure from the symmetric normal distribution. Since Sturges' formula provides for translating continuous, symmetric, normal data into discrete, symmetric, binomial classes, it is appropriate to use $\sqrt{b_1}$ to modify his rule. If $\sqrt{b_1}$ for a particular sample is more than so-and-so many standart deviations away from zero, one would wish to reject the hypothesis upon which Sturges' rule is predicted. The standart deviation of $\sqrt{b_1}$ depends only upon sample size, becoming smaller as sample size increases:

$$\sqrt[3]{b_1} = \sqrt{\frac{6 \times (n - 2)}{(n + 1) \times (n + 3)}}$$

The proposed rule for adding extra classes is given below:

$$K_e = \log_2 \left(1 + \frac{\sqrt{b_1}}{\sqrt[3]{b_1}} \right)$$

If $\sqrt{b_1} = 0$, no extra classes are added. As departure from the symmetric normal distribution becomes more obvious, classes are added, but at a decreasing rate (Doane, 1976).

2. MATERIALS AND METHODS

In this article, 14 formulas are used to compute the number of bins / classes taken from Table 1. Data were generated to simulate a number of experimental scenarios. The simulation results were analysed by using table representations. All computations were performed with Excel software for Windows.

3. RESULTS AND DISCUSSION

After determining the range of scores and distributions, the main question before a set of data can be converted into a grouped-data frequency table is determining the number of classes. The number of classes to be used is primarily not only dependent on the number of observations in the data set, but also on the range of observed scores. That is, larger numbers of observations require a larger number of class groups. In general, however, the frequency distribution should have at least five class groupings, but no more than 15. If there are not enough class groupings or if there are too many, little information would be obtained (Berenson and Levine, 1992).

Table 2 and Table 3 present the results that are obtained with the Scott's (1979) formula and Freedman and Diaconis's (1981) formula, respectively. According to the results of both Scott's formula and Freedman and Diaconis's formula, the number of classes increase rapidly as both range and standart deviation / interquartile of data goes up, particularly when n is in thousands.

Table 2. General results of Scott's (1979) formula

n	Range	Standart Deviation	Number of Bins/ Classes
Increase	Increase	Increase	Increase
Increase	Increase	Decrease	Increase
Increase	Decrease	Increase	Decrease
Increase	Decrease	Decrease	Decrease
Decrease	Increase	Increase	Decrease
Decrease	Increase	Decrease	Decrease
Decrease	Decrease	Increase	Decrease
Decrease	Decrease	Decrease	Decrease

Table 3. General results of Freedman and Diaconis' (1981) formula

n	Range	Interquartile	Number of Bins/ Classes
Increase	Increase	Increase	Increase
Increase	Increase	Decrease	Decrease
Increase	Decrease	Increase	Decrease
Increase	Decrease	Decrease	Decrease
Decrease	Increase	Increase	Increase
Decrease	Increase	Decrease	Increase
Decrease	Decrease	Increase	Increase
Decrease	Decrease	Decrease	Decrease

Table 4 lists a few numbers of observed values, n , for each of the 12 formulas proposed to compute an appropriate or suitable number of classes, k . It can be observed that there is a great diversity of k values obtained.

Table 4. Value of k for selected n numbers

n	Sturges	Cohran	Cencov	Bendat and Piersol	Larson	Velleman	Mosteller and Tukey	Terrel and Scott	Ishikawa	Rice	Anonymous1	Anonymous2
10	5	2	3	5	4	7	4	3	7	5	5	4
20	6	2	3	7	4	9	5	4	7	6	6	5
30	6	3	4	8	5	11	6	4	7	7	6	5
40	7	3	4	9	5	13	7	5	7	7	7	6
50	7	4	4	9	5	15	8	5	7	8	7	6
60	7	4	4	10	5	16	8	5	8	8	7	6
70	8	4	5	11	6	17	9	6	8	9	8	7
80	8	4	5	11	6	18	9	6	8	9	8	7
90	8	5	5	12	6	19	10	6	8	9	8	7
100	8	5	5	12	6	20	10	6	8	10	8	7
125	8	5	5	13	6	21	12	7	9	10	9	7
150	9	6	6	14	6	22	13	7	9	11	9	8
175	9	6	6	15	6	23	14	8	10	12	10	8
200	9	7	6	16	7	24	15	8	10	12	10	8
250	9	8	7	17	7	24	16	8	11	13	10	8
300	10	8	7	19	7	25	18	9	12	14	11	9
350	10	9	8	20	7	26	19	9	13	15	11	9
400	10	9	8	21	7	27	20	10	14	15	12	9
450	10	10	8	22	7	27	22	10	15	16	12	9
500	10	10	8	23	7	27	23	10	16	16	12	9
1000	11	15	10	30	8	30	32	13	26	20	15	10
1500	12	18	12	35	8	32	39	15	36	23	16	11
2000	12	20	13	40	9	34	45	16	46	26	17	11
2500	13	23	14	43	9	34	50	18	56	28	18	12
5000	14	32	18	57	10	37	71	22	106	35	22	13
10000	15	45	22	75	10	40	100	28	206	44	25	14
20000	16	64	28	99	11	44	142	35	406	55	30	15
40000	17	90	35	130	12	47	200	44	806	69	36	16
50000	17	100	37	142	12	47	224	47	1006	74	38	16
100000	18	142	47	187	12	50	317	59	2006	93	45	17

When setting up a grouped-data frequency table, it makes a big difference how many classes are used. The choice of the number of class is quite arbitrary. It is even highly subjective as it is a matter of personal judgement. There exist guidelines to help the researchers with this, but they remain vague in that they are subject to personal interpretations, tastes and preferences.

It is important to remember that frequency distribution tables are employed to reveal or emphasize a group pattern. Either too many or too few classes may blur that pattern and thereby work against the researcher who seeks to add clarity to the analysis. In sum, the researcher generally makes a decision as to the number of classes based on the set of data and personal objectives, factors that may vary considerably from one research situation to another (Lohaka, 2007).

A considerable amount of work has been done on how to set up the best histogram. However, there is no agreement as to which method should work best (Gislason and Goldfield, 1984).

4. REFERENCES

Bendat, S. J., Piersol, A.G., 1966. *Measurements and Analysis of Random Data*. John Wiley & SONS, Inc., New York.

Berenson, M. L., Levine, D.M., 1992. *Basic Business Statistics: Concepts and Applications*. Prentice Hall Englewood Cliffs, New Jersey.

Brown, L.D., Hwang, J.T., 1993. How to Approximate a Histogram by a Normal Density. *The American Statistician*, 47 (4), 251-255.

Cencov, N.N., 1962. Evaluation of an Unknown Distribution Density From Observations. *Soviet Mathematics*, 3, 1559-1562.

Cohran, W.G., 1954. Some Methods for Strengthening the Common Chi Square Test. *Biometrics*, 10 (4), 417-451.

Denby, L., Mallows, C., 2009. Variations on the Histogram. *Journal of Computational and Graphical Statistics*, 18(1), 21-31.

Davies, G.R., 1929. The Analysis of Frequency Distributions. *Journal of the American Statistical Association*, 24 (168), 349-366.

Doane, D.P., 1976. Aesthetic Frequency Classifications. *The American Statistician*, 30 (4), 181-183.

Freedman, D., Diaconis, P., 1981. On The Histogram as a Density Estimator: L_2 Theory. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 57, 453-476.

Gislason, E. A., Goldfield, E. M., 1984. Determination of the Optimum Number of Bins to Use in a Histogrammic Representation of a Probability Density Function. *Journal of Chemical Physics*, 80(2), 701-704.

- He, K., Meeden, G., 1997. Selecting the Number of Bins in a Histogram: A Decision Theoretic Approach. *Journal of Statistical Planning and Inference*, 61, 59-69.
- Hirai, Y., Some Remarks on Class Interval of Histograms, 2009. http://eprints.lib.okayama-u.ac.jp/9696/1/082_0113_0117.pdf.
- Hyndman, R. J., The Problem with Sturges' Rule for Constructing Histograms, 1995, Monash University, www.robjhyndman.com/papers/sturges.pdf.
- Ishikawa, K., 1986. *Guide to Quality Control*. White Plains, New York: Unipub, Kraus International.
- Knuth, K.H., Optimal Data-Based Binning for Histograms, Draft Paper, 2006. <http://www.huginn.com/knuth/papers/knuth-histo-draft-060221.pdf>.
- Larson, H.J., 1975. *Statistics: An Introduction*. John Wiley & SONS, Inc., New York.
- Lohaka, H.O., 2007. Making a Grouped Data Frequency Table: Development and Examination of the Iteration Algorithm. PhD Thesis, College of Education, Ohio University (unpublished).
- Mann, H. B., Wald, A., 1942. On the Choice of the Number of Class Intervals in the Application of the Chi Square Test. *The Annals of Mathematical Statistics*, 13 (3), 306-317.
- Mori, T., 1974. An Optimal Length of Class Interval for Histogram. *Japan Journal of Applied Statistics*, 4(1), 17-24.
- Mosteller, F., Tukey, J.W., 1977. *Data Analysis and Regression, A Second Course in Statistics*. Addison-Wesley, Reading, MA.
- Plane, D. R., Oppermann, E.B., 1981. *Business and Economic Statistics*. Business Publications, Inc., Plano, Texas.
- Rissanen, J., 1992. *Stochastic Complexity in Statistical Inquiry*. World Scientific, Singapore.
- Rudemo, M., 1982. Empirical Choice of Histograms and Kernel Density Estimators. *Scandinavian Journal of Statistics*, 9 (2), 65-78.
- Scott, D. W., 1979. On Optimal and Data-Based Histograms. *Biometrika*, 66(3), 605-610.
- Scott, D. W., 1992. *Multivariate Density Estimation: Theory, Practice and Visualization*. John Wiley & SONS, Inc., New York.
- Sturges, H.A., 1926. The Choice of a Class Interval. *Journal of the American Statistical Association*, 21 (153), 65-66.

Suzuki, G., 1985. Effective Use of Graphical Representation in Statistics. Japan Journal of Applied Statistics, 14(1), 27-37.

Terrel, G.R., Scott, D.W., 1985. Oversmoothed Nonparametric Density Estimates. Journal of the American Statistical Association, 80 (389), 209-214.

Velleman, P.F., 1976. Interactive Computing for Exploratory Data Analysis I: Display Algorithms. 1975 Proceedings of the Statistical Computing Section, 142-147, Washington DC: American Statistical Association.

Wand, M.P., 1997. Data-Based Choice of Histogram Bin Width. The American Statistician, 51 (1), 59-64.

HİSTOGRAMLARDA VE SIKLIK TABLOLARINDA KULLANILAN SÜTUN / SINIF SAYILARININ BELİRLENMESİ: KISA BİR BİBLİYOGRAFYA

ÖZET

Tek değişkenli bir veri setinin grafiksel gösterimi için en eski ve en popüler yöntem histogramdır. Bir histogramın oluşturulmasında belirlenmesi gereken önemli bir parametre ise aralık genişliği veya sınıf genişliği olarak da bilinen sütun genişliğidir. Histogramların esas olan sütun genişliği basit olarak sütun, aralık veya sınıf olarak isimlendirilen alt grupların uzunluğudur. Sıklık dağılımları ise verinin düzenlenmesi ve sunulmasında yardımcı olur. Sınıflandırma tekniklerinin hemen tamamında temel sorun sınıf sayısının nasıl seçileceğidir. Sınıf sayısı ile ilgili her veri seti için geçerli doğru bir cevap bulunmamaktadır. Her bir durum ayrı ayrı dikkate alınmalı, her bir sıklık tablosu kendine özgü olarak düzenlenmelidir. Örnek büyüklüğü arttıkça sütun / sınıf sayısı artmaktadır. $n > 300$ olması durumunda en düşük sütun / sınıf sayısı Larson (1975) formülünden, en yüksek sütun / sınıf sayısı ise Ishikawa (1986) formülünden elde edilmektedir. Bu çalışmanın amacı, sütun / sınıf sayısının belirlenmesi ile ilgili kısa bir bibliyografya vermektir.

Anahtar Kelimeler: Sütun genişliği, Sınıf sayısı, Sıklık tablosu, Histogram.

SOCIOECONOMIC STATUS (SES) SCORES OF TURKISH STATISTICS STUDENTS

Doğan YILDIZ*

Atıf EVREN**

ABSTRACT

Based on the findings of the study which we submitted to TUBITAK in 2007, we tried to estimate SES (socioeconomic status) scores of statistics students' families using some answers on the questionnaire forms of 1794 students in our sample. By analysing SES scores, some sound "distinctions" between the profiles of students from different regions of Turkey (and between those of private university students and public university students as well) may be possible. This point should be emphasized especially in discussing some issues on future of university education. Within last years, opening new universities is on the current agenda of Turkey. Some people emphasize that this expansionary process will increase the lack of quality present in university education. All these criticisms have sound and logical bases. On the other hand, this process itself brings more students more opportunities in university education. The significant differences between SES scores of statistics students lead us to think that opening new universities can create some opportunities for a kind of social mobility.

Keywords: Socioeconomic status scores, Turkish universities.

1. INTRODUCTION

The discussions on the performance of statistics education were intensified especially at the end of nineteen nineties in United States of America. For more than 20 years, ASA (American Statisticians Association) has been organising meetings and symposiums to evaluate the quality of statistics education throughout the world and nowadays there is a vast literature on statistics education. A summary of this process can be found in Yıldız, D., Evren, A. (2009a), and Yıldız, D., Evren, A. (2009b). Besides, some contemporary trends and some important evaluations on the performance of Turkish university education may also be found in Akyol(2010). Statistics is a young discipline in Turkey as far as the establishment years of statistics departments are taken into account. In Turkey, the oldest statistics departments were established in 1960's. An increasing demand for statistical studies from business life is a fact. On the other hand; as a by-product of this rapid growth process; there are some serious educational problems that cannot be vanished easily.

In 2007, we tried to analyze some issues of statistics education in Turkey by valuable contributions of TUBITAK (The Scientific and Technical Research Council of Turkey). Our study was based on the information derived from the questionnaire forms of 1794 students from different statistics departments in Turkey. In this context, questionnaire forms were distributed in 19 universities from 10 different cities of Turkey. The questionnaire forms were analyzed by SPSS 11.0 and SAS 9.4.

* Assist. Prof. Dr., Yıldız Technical University, Faculty of Sciences and Literature, Department of Statistics, Istanbul, e-mail: dyildiz@yildiz.edu.tr

**Assist. Prof. Dr., Yıldız Technical University, Faculty of Sciences and Literature, Department of Statistics, Istanbul, e-mail: aevren@yildiz.edu.tr

Some of the questions present in the questionnaire forms were aimed at investigating the students' evaluations on university education, learning difficulties they encounter, statistics as a prospective profession, or the quality of current statistics education, etc. Besides, there were some questions on consumption patterns, free time activities, or some other questions which might indicate some aspects of socio economic positions of students' families as well. Some findings of this study can be found in Evren, A, Yıldız, D. (2009), Yıldız, D., Evren, A. (2008), Yıldız, N.Ç., et all (2009).

The frequency distribution of sampled students coming from statistics departments of different universities can be investigated by the following table:

Table 1. The frequency distribution of students participating this survey

University	Frequency	Percentage
Anadolu	117	6,522
Ankara	78	4,348
Başkent	37	2,062
Çukurova	58	3,233
Dokuz Eylül	91	5,072
Ege	182	10,145
Gazi	164	9,142
Hacettepe	128	7,135
Istanbul Ticaret	13	0,725
Karadeniz Teknik	72	4,013
Kırıkkale	79	4,404
Mimar Sinan	87	4,849
Muğla	50	2,787
Ondokuz Mayıs	295	16,444
Ortadoğu Teknik	37	2,062
Osmangazi	65	3,623
Selçuk	98	5,463
Yıldız Teknik	119	6,633
Fırat	24	1,338
Total	1794	100

The university with the maximum number of participating students contributes with 295 students (which is equal to the 16.4% of the sampled students) and the university with the minimum number of participating students contributes with 13 students (0.7% of the sample).

Another table may be helpful to analyze frequency distribution of students with respect to their class identifications.

Table 2. The frequency distribution of sampled students with respect to their classes

Group	Frequency	Percentage	Cumulative Percentage
Freshman	311	17,34	17,34
Sophomore	415	23,13	40,47
Junior	469	26,14	66,61
Senior	599	33,39	100
Total	1794	100	

We intended to include half of the total statistics students (2142 of 4282 students) in the sample. Besides some forms from 2142 students could not be used. So we could have evaluated 1794 of 2142.

We adopted a kind of quota sampling strategy. The quotas are as follows:

1-Quota on classes: It is observed that most of the students in the sample are especially from higher classes. This is not purely coincidental. Because we think that as the levels of engagement of the students in statistics departments increase, their evaluations on their departments or on their profession will inevitably become more objectivistic and more sound. For this reason most of the students are sampled from the third and fourth classes intentionally.

Table 3. Quota on classes

Classes	Planned		Realized	
	%	Fi	%	Fi
1	10	214	17,3	311
2	20	428	23,2	415
3	30	642	26,1	469
4	40	856	33,4	599
Total	100	2142	100,0	1794

2-Quota on gender: We planned that 52% of the students in the sample be female. Here the percentages of male and female students that would appear in the sample were based on the figures taken from statistics departments. So 52%, and 48% are average figures based on the data we got from statistics departments. The planned and realized figures on this issue are given below:

Table 4. Quota on gender

Gender	Planned		Realized	
	%	Fi	%	Fi
Female	52	1114	52	982
Male	48	1028	48	862
Total	100	2142	100,0	1794

3-Quota on education program: The percentages of students from the first educational program and those from the second educational program that would be in the sample were based on the figures taken from statistics departments. So 85%, and 15% are average figures based on the data we got from statistics departments. The planned and realized figures on this issue are given below:

Table 5. Quota on education program

Program	Planned		Realized	
	%	Fi	%	Fi
First module	85	1820	84,1	1509
Second module (night program)	15	322	15,9	285
Total	100	2142	100,0	1794

2. BASIC DISCUSSIONS ON DETERMINING THE SOCIO-ECONOMIC STATUS

While engaging in any reform effort on university education programs, one should consider the general cultural and economic level of students and instructors as a crucial starting point. Because current living conditions of these people inevitably affect their scientific background and cultural level.

There is not a unique formula for determining SES scores. In 2006, Turkish marketing researchers made an agreement on some formulations and further researches were developed via this agreement. In our study we also adopted their approach. The criteria defined and the conventional points given to each item or answer are as below. There are 5 basic criteria in calculating SES scores:

Criterion 1 or SES 1: The ownership of durable consumption goods

Table 6. Points related to ownerships of durable goods

	POINTS	
	PRESENT	ABSENT
REFRIGIRATOR	0	-13
TELEVISION	0	-10
WASHING MACHINE	0	0
DISH WASHER	11	0
DRYING MACHINE	14	-4
MUSIC SET	0	0
VIDEO	4	0
VIDEO CAMERA	12	0
DVD PLAYER	6	0
DESKTOP COMPUTER	10	0
PRINTER	10	0
LAPTOP COMPUTER	14	0
MICROWAVE OWEN	11	0
AIR CONDITION	15	0

Criterion 2 or SES 2: Real-Estate and Vehicle Ownership**Table 7. Points related to real-estate ownership**

	POINTS	
	PRESENT	ABSENT
REGULAR	25	0
SUMMER HOUSE	20	0
COLLECTIVE OWNERSHIP FOR SUMMER HOUSES	15	0
AUTOMOBILE	10	0
YACHT/SAILING- BOAT	30	0

Criterion 3 or SES 3: Education Level**Table 8: Points related to education level**

EDUCATION LEVEL	POINTS
LITERATE	-14
PRIMARY SCHOOL	-2
SECONDARY SCHOOL	3
HIGH SCHOOL	9
UNDERGRADUATE	18
GRADUATE	24

Criterion 4 or SES 4: Profession**Table 9. Points related to profession**

	POINTS
RETIRED	5
RETIERS	10
UNEMPLOYED	-11
BIG CAPITALIST (PRODUCTION / SERVICES, EMPLOYING MORE THAN 50 PEOPLE)	25
MIDDLE-SIZED CAPITALIST (PRODUCTION/SERVICES, EMPLOYING BETWEEN 10 AND 50)	20
SMALL-SIZED CAPITALIST (PRODUCTION/SERVICES, EMPLOYING LESS THAN 10 PEOPLE)	15
BIG MERCHANT / BIG WHOLESALER / BIG TRADER	23
MIDDLE- SIZED TRADER / MEDIUM-SIZED EMPLOYER	18
SMALL-TRADER / SMALL-SIZED EMPLOYER	10
DRIVER HAVING HIS/HER OWN VEHICLE	9
DOCTOR / PHARMACIST / DENTIST /ARCHITECT, ETC.	20
TOP MANAGER IN EITHER PUBLIC OR PRIVATE SECTOR (WITH EMPLOYEES MORE THAN 50)	19
HIGH MANAGER (WITH EMPLOYEES BETWEEN 10 AND 50)	16
HIGH MANAGER (WITH EMPLOYEES LESS THAN 10)	14
MEDIUM LEVEL MANAGER IN EITHER PUBLIC OR PRIVATE SECTOR (WITH MORE THAN 50 EMPLOYEES)	15
MEDIUM LEVEL MANAGER (WITH EMPLOYEES BETWEEN 10 AND 50)	14
MEDIUM LEVEL MANAGER (WITH EMPLOYEES LESS THAN 10)	12
LOW LEVEL MANAGER IN EITHER PUBLIC OR PRIVATE SECTOR	10
EMPLOYEE IN PUBLIC OR PRIVATE SECTOR /BANKING SECTOR EMPLOYEES, ETC.	9
UNSKILLED PUBLIC OR PRIVATE SECTOR EMPLOYEES	4
DOCTOR / PHARMACIST / DENTIST / ARCHITECT ETC.	18
SALES PERSON, MARKETER	9
NURSE / HOSTESS / WAITRESS / BARMEN, SECRETARY	10
TEACHER	10
ACADEMICIAN	12
OFFICER AND SUBOFFICER	12
HIGH OFFICER (CAPTAIN AND HIGHER)	16
DRIVER	8
QUALIFIED WORKERS	10
TECHNICIANS, PHARMACY REPRESENTERS, ETC.	8

Criterion 5 or SES 5: Monthly Total Income

The frequency distribution of the monthly incomes of 1794 students' families are given in the following table.

Table 10. Income distribution of 1794 students' families

Monthly Income	Frequency	Percentage	Cumulative Percentage
less than 750 YTL	170	9,476	9,476
between 750 and 1500 YTL	930	51,839	61,315
between 1501 and 2250 YTL	321	17,893	79,208
between 2251 and 3000 YTL	230	12,821	92,029
more than 3000 YTL	143	7,971	100
Total	1794	100	

Total income data are converted into scores ranging between 1 and 69. Then these scores are evaluated as the total income component of SES scores.

3. THE DETERMINATION OF STUDENTS' FAMILIES' SES SCORES

Total SES point is calculated by simply adding 5 SES score components as follows:

$$\text{SESTOTAL} = \text{SES1} + \text{SES2} + \text{SES3} + \text{SES4} + \text{SES5} \quad (1)$$

To be able to make comparisons between the SES scores of overall Turkish population and those of the students' families in the sample, we give the percentages of socioeconomic groups of Turkish population as follows:

Table 11. The percentages of SES groups of Turkish population

Groups	Urban Areas	Rural Areas	General
A	1,5	0	1,1
B	12,2	2,2	9,1
C1	23,3	8,9	18,9
C2	31,8	31,2	31,6
D	23,3	40,3	28,5
E	7,9	17,4	10,8

In our study, before calculating 5 components of SES scores of each student, we first tested the consistency of some of the answers with respect to some others by means of correlations and cross-tabulations. Besides, apart from the general evaluation on Turkish population, here the groups D and E are united as a single group intentionally. The reason for this is that the number of students coming from group E is so small that the students from this group can be simply added to group D. The class limits for each SES group can be seen through the following table:

Table 12. SES scores and groups

SES Groups	Points
A	201 and over
B	between 146 and 200
C1	between 105 and 145
C2	between 70 and 104
D and E	69 and lower

Then we have the frequency distribution of students with respect to their group identifications as follows:

Table 13. Frequency distribution of SES scores of 1794 students

Groups	Frequency	Percentage	Cumulative Percentage
A	168	9,36	9,36
B	538	29,99	39,35
C1	543	30,27	69,62
C2	353	19,68	89,3
D & E	192	10,7	100
Total	1794	100	

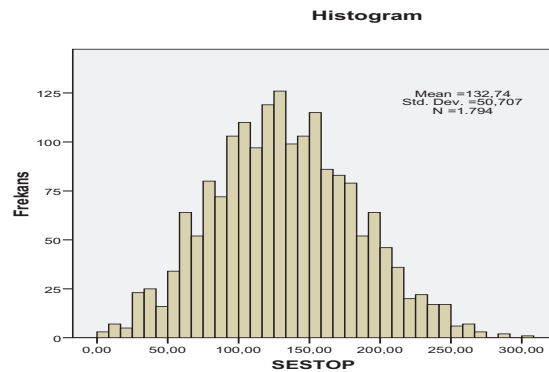


Figure 1. Frequency distribution of SESTOP scores

As can be seen from the previous histogram, total SES scores follow a symmetrical distribution roughly.

4. COMPARISONS OF THE PERCENTAGES OF SES GROUPS IN TURKISH POPULATION AND THOSE OF THE STUDENT POPULATION IN STATISTICS DEPARTMENTS

Besides, we wondered if each SES group contribute university populations proportionally or not. By uniting the entries of Table 8 and Table 10, we obtain Figure 2 from which we can easily conclude that all socioeconomic groups do not contribute equally to the student population in statistics departments. To make more sound comparisons, the following figure might be functional.

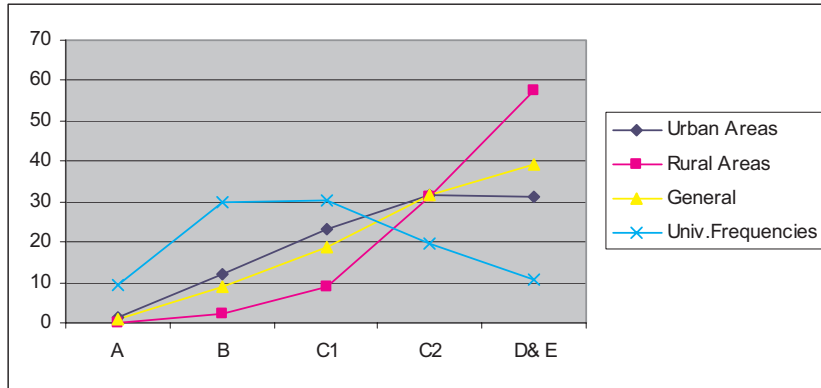


Figure 2. SES scores distributions within Turkish population and within statistics departments

In general, higher income groups A, B and C1 contribute to the population of statistics students much more than their relative weights in Turkish population. Approximately 10% of statistics students come from group A, while only 1% of the total Turkish population is from the same group. The figures for groups B, and C1 are not proportionate with what is expected under the social equality hypothesis either. For the groups C2, D and E the situation is worse. For example 39,3% of the population makes only 10,7% of the statistics students population in the universities. All these arguments will suffice to show that there are considerable inequalities between social groups in contributing to university populations based on the figures of sampled statistics students. It can be simply deduced that higher education is still a matter for higher classes because of high educational costs and maybe because of the existence of social inequalities .

5. COMPARISON OF THE AVERAGES OF TOTAL SES SCORES OF STUDENTS FROM DIFFERENT STATISTICS DEPARTMENTS

It is worthwhile to ask whether there are significant differences between the average total SES scores between different statistics departments. The following figure will show the variation between statistics departments of different universities.

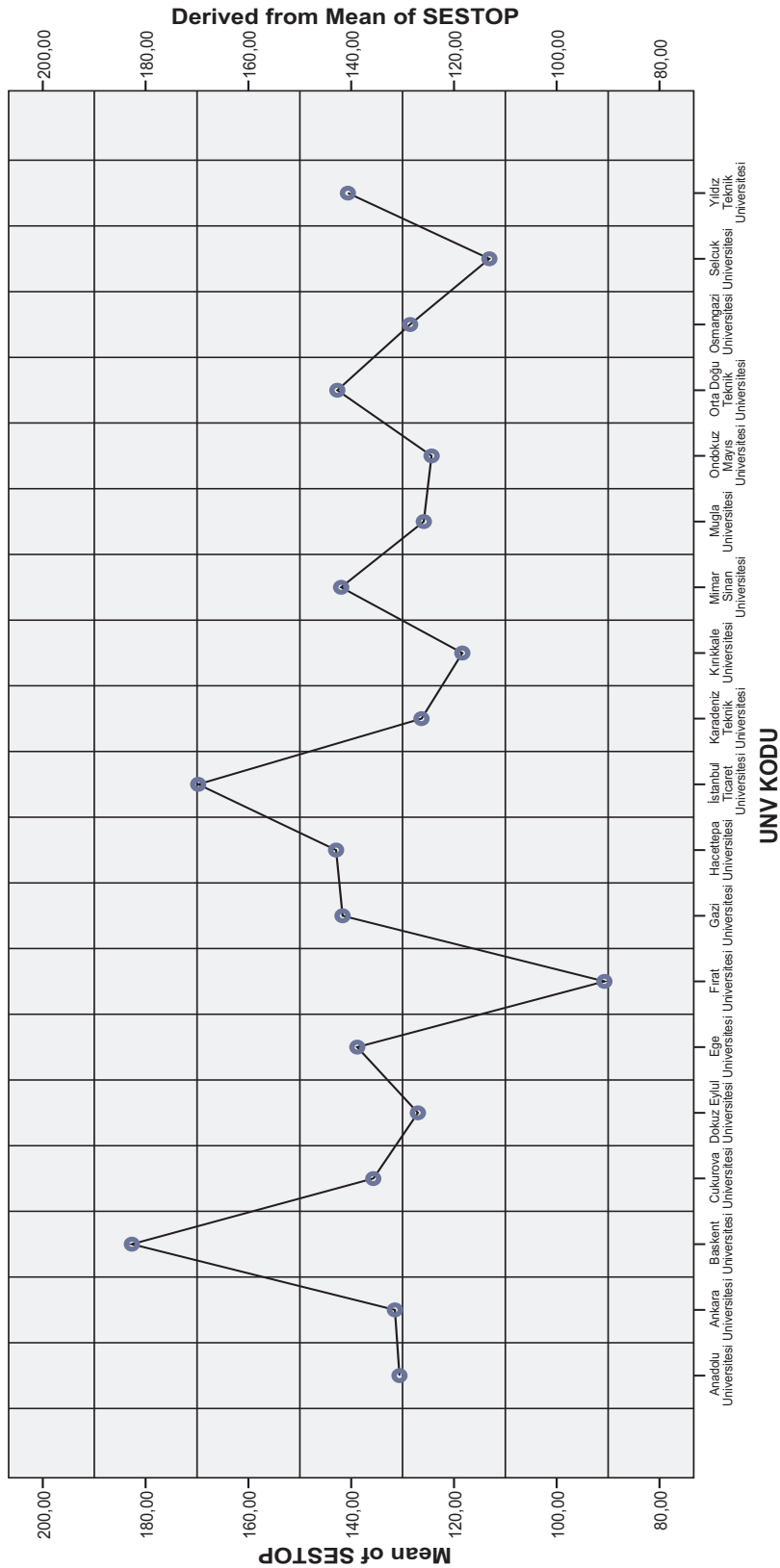


Figure 3. The average total SES scores of different statistics departments

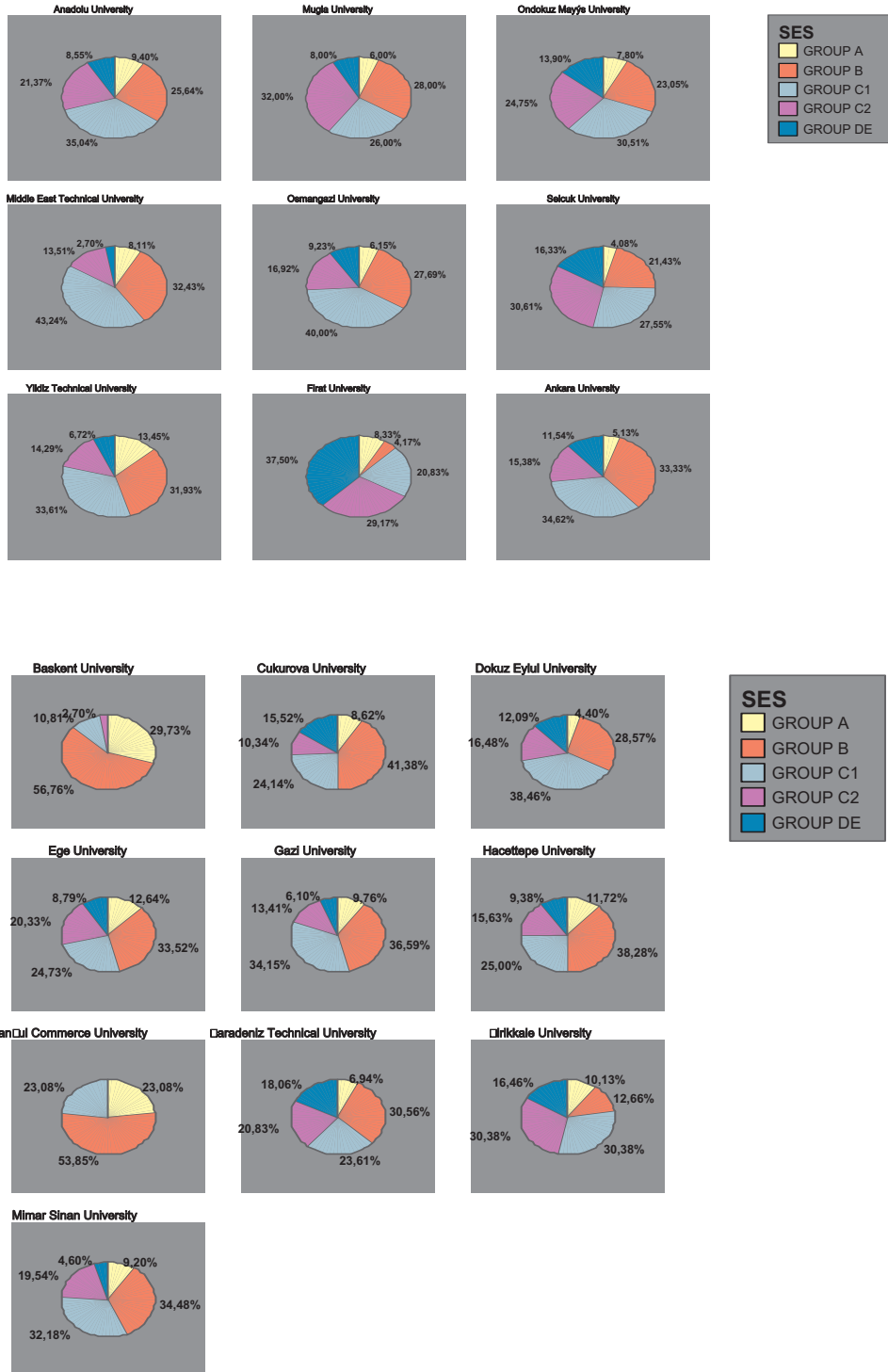


Figure 4. SES scores distributions within statistics departments*

*The exact figures are given in Appendix.

6. DISCUSSION

The findings of our empirical study show that there are significant differences between the SES scores of statistics students. The average total SES scores of two private universities take the two top positions as can be expected a priori. Then come the statistics departments in three metropolitan cities of Turkey (İstanbul, Ankara and İzmir). This result can also be predicted by considering the economic and cultural development levels of these cities. Finally, the students from the statistics departments established either in eastern (Fırat University) or central parts (Selçuk and Kırıkkale Universities) of Turkey have the lowest average total SES scores. These results are in harmony with the general economic positions or development levels of the various regions of Turkey.

Within last years, opening new universities is on the current political agenda in Turkey. Some people emphasize that this expansionary process will increase the lack of quality in present university education. All these criticisms have sound and logical bases. On the other hand, this process itself brings more students more opportunities in higher education. Therefore we think that this expansionary process supported by quality will be beneficial for all. This might be achieved by changing some priorities of the macro plans for allocations of resources by the governments.

7. REFERENCES

Akyol, T., 2010. Bilim ve Yanılgı, Dođan Kitap, Geniřletilmiş 7. Baskı, İstanbul.

Çađdař, C., 2007. IAA Power Point Sunumu.

<http://baybul.com/pazarlama-ve-satis/374618-turkiyenin-yeni-sosyo-ekonomik-statues-profil.html>.

http://www.mayak.com.tr/documents/MSESI_sunum.pdf.

ESOMAR Standard Demographic Classification, 1997. <http://www.esomar.com>.

Yıldız D. vd., 2007. Türkiye’de Tüm İstatistik Bölümleri Bazında İstatistik Eğitiminin Öğrenci ve Öğretim Üyesi Gözüyle Deđerlendirilmesi. TÜBİTAK Proje No:105K171, İstanbul.

Evren, A., Yıldız, D., 2009. İstatistik Bölümü Öğrencileri Anketi ve Öğrencilerin İstatistik Bölümlerinden Duyduđu Memnuniyet Üzerine Bir Faktör Regresyon Modeli. 18. İstatistik Arařtırma Sempozyumu, 2009.

Yıldız, D., Evren, A., 2008. Some Results from the Survey on Turkish Statistics Education. 6th International Conference Sustainable Development Culture and Education, Conference Proceeding CD, June 4-7 2008, Eskiřehir, Turkey, 816-830.

Yıldız, N. Ç., Yıldız, D., Evren, A., 2009. Türkiye’deki İstatistik Bölümlerinin Belirli Etkenler Açısından Karřılařtırılması. I. Ulusal Eređli Meslek Yüksek Okulu Tebliđ Günleri, 1(1):363-377.

Yıldız, D., Evren, A., 2009a. Yurtdışında İstatistik Programlarının Reforma Tabi Tutulma Çabaları. VI. İstatistik Günleri Sempozyumu Bildiriler Kitabı, 2009, Samsun, 59-65.

Yıldız, D., Evren, A., 2009b. On Some Reform Initiatives on Statistics Education throughout the World. Selçuk Journal of Applied Mathematics, 10(1):95-106.

TÜRKİYE'DE İSTATİSTİK BÖLÜMLERİNDE OKUYAN ÜNİVERSİTE ÖĞRENCİLERİNİN SOSYO-EKONOMİK STATÜ (SES) PUANLARI

ÖZET

2007 yılında TÜBİTAK'a sunduğumuz çalışmamızın verilerinden hareketle örnekleme dahil edilen 1794 öğrencinin ailelerinin sosyo ekonomik statü (SES) puanlarını hesapladık. SES puanlarının analiz edilmesi ile Türkiye'nin farklı üniversitelerindeki öğrencilerin profilleri arasında (aynı zamanda özel üniversite öğrencilerinin profilleri ile kamu üniversitelerinde okuyan öğrencilerin profilleri arasında) bazı anlamlı "farklılıkların" saptanmasının mümkün olduğunu düşünüyoruz. Bu noktanın vurgulanması özellikle üniversite eğitiminin geleceği ile ilgili bazı sorunların tartışılmasında yararlı olacaktır. Son yıllarda yeni üniversitelerin açılması Türkiye'de gündemdedir. Bazı insanlar bu genişleme sürecinin şimdiki üniversite eğitiminde bulunan kalite eksikliğini arttıracaklarını vurgulamaktadırlar. Bu türden eleştirilerin anlamlı ve mantıksal bir temeli vardır. Bununla birlikte bu süreç daha çok öğrenciye üniversite eğitiminde daha fazla olanaklar getirmektedir. İstatistik öğrencilerinin SES puanları arasında anlamlı farklılıkların gözlenmiş olması açılan yeni üniversitelerin bir tür sosyal hareketlilik olanağı sunabileceğini bizlere düşündürmüştür.

Anahtar Kelimeler: Sosyo-ekonomik statü puanları, Türkiye'deki üniversiteler.

APPENDIX: Exact Frequency Distributions of SES Scores within Universities

	SES				
	GROUP A	GROUP B	GROUP C1	GROUP C2	GROUP DE
Anadolu University	9,4%	25,6%	35,0%	21,4%	8,5%
Ankara University	5,1%	33,3%	34,6%	15,4%	11,5%
Baskent University	29,7%	56,8%	10,8%	2,7%	,0%
Cukurova University	8,6%	41,4%	24,1%	10,3%	15,5%
Dokuz Eylul University	4,4%	28,6%	38,5%	16,5%	12,1%
Ege University	12,6%	33,5%	24,7%	20,3%	8,8%
Gazi University	9,8%	36,6%	34,1%	13,4%	6,1%
Hacettepe University	11,7%	38,3%	25,0%	15,6%	9,4%
Istanbul Commerce	23,1%	53,8%	23,1%	,0%	,0%
Kadir Has University	,0%	,0%	,0%	,0%	,0%
Karadeniz Technical	6,9%	30,6%	23,6%	20,8%	18,1%
Afyon Kocatepe University	,0%	,0%	,0%	,0%	,0%
Kirikkale University	10,1%	12,7%	30,4%	30,4%	16,5%
Mimar Sinan University	9,2%	34,5%	32,2%	19,5%	4,6%
Mugla University	6,0%	28,0%	26,0%	32,0%	8,0%
Ondokuz Mayıs University	7,8%	23,1%	30,5%	24,7%	13,9%
Middle East Technical	8,1%	32,4%	43,2%	13,5%	2,7%
Osmangazi University	6,2%	27,7%	40,0%	16,9%	9,2%
Selcuk University	4,1%	21,4%	27,6%	30,6%	16,3%
Ufuk Universitesi	,0%	,0%	,0%	,0%	,0%
Yildiz Technical University	13,4%	31,9%	33,6%	14,3%	6,7%
Yasar Universitesi	,0%	,0%	,0%	,0%	,0%
Firat University	8,3%	4,2%	20,8%	29,2%	37,5%

TÜRKİYE'DEKİ İLLERİN SOSYO-KÜLTÜREL GELİŞİMİŞLİK DERECELERİNE GÖRE DEĞERLENDİRİLMESİ

Esin FİRUZAN* Yusuf Yüksel AYVAZ** Ersen UZUN***

ÖZET

Türkiye'nin bölgelerinde gelişme sağlanması için, alınacak tedbirlerin ve uygulanacak politikaların tutarlı olması ve yatırımların dağılımında bölgesel gelişmenin dikkate alınması kaçınılmaz olmuştur. Bu çalışmada, Türkiye'de tüm illerin gelişmişlik dereceleri, sosyo-kültürel çerçeveden ölçülmeye çalışılmıştır. Bu amaçla tüm illerin sosyo-kültürel göstergeleri, çok değişkenli istatistiksel analiz yöntemleri kullanılarak homojen il grupları oluşturulmuştur. Temel bileşenler analizi ile uygun boyut derecesi bulunmuş, faktör analizi ile hangi değişkenlerin belirli faktörler altında toplanabileceği belirlenmiştir. Faktör Analizi ile elde edilen değişkenler, sosyo-kültürel açıdan daha önce hiç sınıflandırılmamış olan illerin sınıflandırılması için kümeleme analizinde kullanılmıştır. Elde edilen küme sayısı, ayrıştırma analizi için bir ön bilgi olarak kabul edilip, kümeleme analizi ile elde edilen grupların sınıflandırma başarısı hakkında fikir sahibi olmaya çalışılmıştır. Ayrıca ayrıştırma analizi ile grupları birbirlerinden ayırmada önemli rol oynayan değişkenler de belirlenmiştir. Çalışma sonunda elde edilen illerin sosyo-kültürel gelişmişlik sıralaması ile Devlet Planlama Teşkilatı'nın (DPT) 2003 yılında yayınladığı sosyo-ekonomik gelişmişlik sıralaması karşılaştırılmıştır.

Anahtar Kelimeler: Ayrıştırma, Faktör, Kümeleme, Sosyo-kültürel gelişmişlik.

1. GİRİŞ

Gelişmekte olan ülkeler arasında yer alan Türkiye'nin coğrafyasında dengeli bir gelişme sağlanması hedefi, Türkiye ekonomisi için hedef alınan yüksek kalkınma hızı kadar önemlidir. Günümüzde dengeli gelişme amacı doğrultusunda alınması gereken tedbirlerin ve uygulanacak politikaların tutarlı olması ve kamu yatırımlarının dağılımında ekonomik coğrafyanın ve bölgesel gelişmenin dikkate alınması kaçınılmaz olmaktadır. 1970'lerde il sayısı 67 olan, 2000'li yıllarda il sayısı 81'e ulaşan Türkiye'de iller itibariyle dengeli sosyo-ekonomik gelişmişliğin amacı, iller ve bölgeler arasındaki gelişmişlik farklarının kabul edilebilir bir düzeye getirilmesi, göreceli olarak geri kalmış il veya bölgelerin geliştirilmesidir. Yoksa iller veya bölgeler arasında her zaman gelişmişlik düzeylerinde göreceli farklılıklar olacaktır.

Ekonomik ve sosyal yönleri ile bir bütün olan gelişmenin, ekonomik yönleri gelir arttırıcı, sosyal yönleri ise sosyo-kültürel değişim ile ilgilidir. Bu nedenle, iller arasındaki sosyo-ekonomik gelişmişlik farklılıklarının incelenmesinin yanı sıra sosyo-kültürel gelişmişlik farklılıklarının da ortaya konması ihtiyacı doğmaktadır. Nitekim bu araştırma, daha önce benzer şekilde yapılan çalışmalarla beraber, bölgesel gelişme politikaları oluşturma süreçlerine analitik bilgi sağlamayı amaçlamaktadır.

* Doç. Dr., Dokuz Eylül Üniversitesi, Fen Fakültesi, İstatistik Bölümü, e-posta: esin.firuzan@deu.edu.tr

** Yrd. Doç. Dr., Celal Bayar Üniversitesi, e-posta: yuksel.avvaz@bayar.edu.tr

*** Araş. Gör., Muğla Üniversitesi, İktisadi ve İdari Bilimler Fakültesi İşletme Bölümü, e-posta: eruzun@mu.edu.tr

Yapılan planlama çalışmalarında daha etkin ilke ve tedbirlerin alınmasına imkan sağlayan “İllerin Sosyo-Ekonomik Gelişmişlik Sıralaması” araştırmaları sayesinde, ülke genelinde illerin ekonomik ve sosyal sektörler açısından gelişme yönleri ve geldikleri son durumları izlenebilmektedir. İl bazında karşılaştırmalı olarak gelişmişlik değişkenlerinin belirlendiği, illerin mevcut sosyal ve ekonomik kaynak yapılarının, gelişmişlik durumlarının ve olası gelişme eğilimlerinin incelendiği pek çok çalışma ortaya çıkmıştır. Araştırma sonuçlarının iller ve bölgelerin gelişmesi açısından yapılacak olan çalışmalara ışık tutacağı düşünülmektedir.

İl olma ölçütleri içerisinde şehrin ulaştığı üretim ve ticaret kapasitesi, içinde bulundurduğu yapıların kalitesi, coğrafik özellikleri veya barındırdığı sosyo-kültürel fonksiyon alanlarının işlevselliği göz önünde bulundurulması gerekirken, göz önünde bulundurulmuş tek ölçüt ulaşılan nüfus olmuştur. Bu noktada nüfusun hem avantaj hem de dezavantaj olduğu söylenebilir. Nüfus arttıkça organizasyon zorlaşır, ama nüfus artışı toplum yaşamını canlı kılar ve ekonomiye yansıyan bu canlılık artı değer üretimini artırır. Bu sebeple nüfusu yükselen bir şehrin zenginleşme olasılığı daha fazladır. Eğer nüfus artışı zenginliği de beraberinde getiriyorsa, üretilen artı değer bir kısmının ortak mekan düzenlemesine ve organizasyon üretimine ayrılması sorunları çözebilecektir. İşte bu noktada kültür devreye girer. Yani bir toplumun doğa karşısında üretildiği değer noktasında zengin bir deneyimi olup olmadığı önem kazanır. Toplumun günlük yaşam araçlarını derinleştirebilmesi ve farklılaştırabilmesi gerekir. Bu ise ancak derinlikli bir kültürel zenginliğe sahip olup olmamakla başarılabilir bir şeydir. Niceliksel olarak il sayısının artması, nüfusun artması ve il sınırlarının genişlemesini ifade ederken, niteliksel olarak da kente özgü hayat tarzının oluşması ve sosyo-kültürel gelişmeyi ifade etmektedir. Ayrıca, il kurulmasında ve gelişmişlik tanımında herhangi bir ölçütün olmayışı, yerleşimlerin şehirleşme düzeylerinin sağlıklı olarak belirlenmeyişi rastgele ve düzensiz uygulamaların da yolunu açmaktadır.

Şehirleşmenin demografik, ekonomik ve sosyo-kültürel olmak üzere birbirleriyle bağlantılı pek çok yapısal özellikleri bulunması nedeniyle bu konuda ölçüt geliştirmek çok kolay olmayacaktır. Tüm dünya ülkeleri tarafından uygulanan şehirleşme ölçütleri incelendiğinde, illerin nüfusunu, büyüklüğünü, ekonomik yapısını, idari yapısını, şehirleşme fonksiyonlarının dağılımını ya da bunların birkaçını birlikte esas alan ölçütlere rastlamak mümkündür.

Literatürde, illerin gelişmişlik dereceleri; demografik, istihdam, eğitim, sağlık, alt yapı ve refah alt başlıklarından oluşan Sosyal Göstergeler; sanayi, inşaat, tarım ve mali yapı alt başlıklarından oluşan Ekonomik Göstergelerin birleşimi olan sosyo-ekonomik gelişmişlik dereceleri başlığı altında incelenmiştir.

Güveli ve Kılıçkaplan 2000’de, ekonomik ve sosyal olmak üzere 33 değişken kullandıkları çalışmada, İslam ülkelerinin sosyo-ekonomik gelişmişlik düzeylerini incelemişlerdir. Akyüz ve diğerleri 2001 yılında, belirli sosyo-ekonomik değişkenler kullanarak Türkiye’deki iller arasında gruplandırma çalışması yapmışlardır. Çalışmada, yedi sosyo-ekonomik değişken kullanarak, aynı yapıyı gösteren homojen il gruplarının belirlenmesini amaçlamışlardır. Araştırmada, sosyo-ekonomik yapıyı temsil etmek üzere demografik, finansal, ekonomik, eğitim, sağlık ve enerji birimlerinden oluşan yedi değişken kullanmışlardır. Karpat ve Açıkgöz 2001 yılında, ülkelerin kalkınmışlık düzeylerini ele aldıkları çalışmada 1992’de Saraçoğlu’nun yaptığı çalışmasından yola çıkarak ülkelerin kalkınmışlık düzeyleri üzerine sınıflandırma yapmışlardır. Çalışmada

29 değişken kullanılarak ülkeler gruplandırılmış ve aynı zamanda Saraçoğlu'nun çalışma sonuçları ile karşılaştırılmış ve zaman içerisinde yapılmış olan bu gruplandırmalarda farklılık olup olmadığını belirlemeye çalışmışlardır. Süer ve Şahin 2001 yılındaki çalışmalarında Türkiye'de bölgeler arasındaki farklılaşmada ekonomik ve sosyo-demografik değişkenlerin önemini incelemiştir. 1990-1994 yılları arasındaki Devlet Planlama Teşkilatı (DPT) ve o zamanki adı ile Devlet İstatistik Enstitüsü (DİE) verilerine dayanarak yaptıkları araştırmada sosyo-demografik değişkenlerin bölgesel farklılaşma üzerinde ekonomik değişkenlere nazaran daha önemli olduğunu vurgulamışlardır.

Albayrak 2005'te, Türkiye'de illerin sosyo-ekonomik gelişmişlik düzeylerini belirleyen hipotetik yapıları incelemiştir. Öner ve diğerleri 2006'da, Türkiye'de illerin sosyo-ekonomik olarak farklılaşmasını incelemiştir. Öner, çalışmasında 81 il içerisinde kümeleme analizi ile bir birine benzer olan illeri tespit etmiştir.

Bu araştırma kapsamında, literatürdeki diğer çalışmalardan farklı olarak, Türkiye'de tüm illerin gelişmişlik dereceleri, 2003 ve 2006'lı yıllara ait sayısal göstergeler kullanılarak, ilk defa sosyo-kültürel çerçeveden ölçülmeye çalışılmıştır. Bu çalışmada ele alınan bazı değişkenler için 2006 yılı verileri kullanılırken, bazı değişkenlerde de 2003 yılı verileri kullanılmıştır. Örneğin, illere göre federasyonlara bağlı lisanslı ve faal sporcu sayıları değişkeni en yakın tarihte sadece 2002-2003 sezonu için mevcut olduğundan dolayı o yılın değeri veri olarak alınmıştır. Diğer yandan, müze sayısı değişkeni için 2006 yılı mevcut olduğu için daha önceki yılların değerleri alınmamıştır. Bu sebeple, her ilin farklı yılları arasında oluşabilecek korelasyon oluşması söz konusu değildir. Bu doğrultuda, zaman ekseninde korelasyonlu olmayan illerin seçilen sosyo-kültürel göstergeleri ile çok değişkenli istatistiksel analiz yöntemleri kullanılarak homojen il grupları oluşturulmuştur. Böylelikle gelişmiş kabul edilen ve gelişme düzeyinde olan illerin sosyo-kültürel açıdan hangi konumda oldukları ve bu açıdan Türkiye'nin genel durumunu ortaya koyabilmek amaçlanmıştır.

2. YÖNTEM

Bu çalışmada, çok sayıda gösterge çerçevesinde illerin sosyo-kültürel gelişmişlik düzeyleri incelenmiştir. Amaç, illeri sosyo-kültürel açıdan sınıflandırmak olduğu için, sosyal göstergeler; demografik, eğitim, sağlık, kültür ve spor olmak üzere dört grupta sınıflandırılmıştır. Dört grupta sınıflandırılan bu göstergeleri tanımlayan pek çok değişken bulunmaktadır. Bu açıdan değerlendirildiğinde, çok değişkenli istatistiksel analiz yöntemlerinden temel bileşen analizi, faktör analizi, kümeleme ve ayırıştırma analizi kullanmanın uygun olacağı düşünülmüştür.

Çok değişkenli istatistiksel analiz yöntemlerinde değişken seçimi üzerinde titizlikle durulması gereken önemli bir aşamadır. Öncelikle sosyo-kültürel gelişmişlikle ilgili tüm değişkenler alınmalı, gereksiz görülenler elenmelidir. Araştırmada analiz öncesi sosyo-kültürel gelişmişlikle ilgili 62 değişken elde edilmiştir. Ancak çalışmanın bu aşamasında, Türkiye'nin bazı illerinin kültürel açıdan alt yapısının ne kadar zayıf olduğu görülmüştür. Çalışmada tiyatro sayısı (devlet/özel), sinema salonu sayısı ve oynanan oyun sayısı gibi kültürel faaliyetler içerisinde yer aldığı düşünülen değişkenler açısından 81 il arasında çok ciddi farklılıklar olduğu görülmüştür. İl olmasına rağmen, 69 ilde devlet tiyatrosu bulunmamaktadır. Özel tiyatro sayısı, üç büyük il hariç, yok denecek kadar azdır. Bu nedenle eksik olan veriye sahip bazı değişkenler, kayıp gözlem

sorununa sebep olmaması için analizlere dahil edilmemiştir. Bu nedenle değişken sayısı 50 olarak belirlenmiştir. Bu 50 değişkenin, Türkiye'deki 81 il için 2003 ve 2006 dönemlerini kapsayacak şekilde veri matrisi hazırlanmıştır. Veriler Türkiye İstatistik Kurumu (TÜİK) ve DPT internet sitelerinden elde edilmiştir. Çalışmada kullanılan değişkenler Ek 1'de yer alan Tablo 1'de başlıklar altında verilmektedir.

İller arasındaki karşılaştırmada hatayı en aza indirmek için değişkenler uygun ölçü birimleriyle tanımlanmaya çalışılmıştır. İllerin gelişmişlik düzeyleri, nüfus büyüklükleri ile orantılı olarak değil, büyük ölçüde göreceli veya kişi başına düşen büyüklükler olarak tanımlanmıştır. Kullanılan tüm göstergelerin kişi başına büyüklükler olarak tanımlanması, nüfusu fazla olan illerin aleyhine bir durum yaratmaktadır. Bu nedenle, duyarlılığın arttığı durumlarda yani iller arasındaki farklılığın çok büyük olduğu durumlarda (nüfusla orantılı olarak) göreceli veya kişi başına düşen, bunun dışında ise toplam büyüklükler olarak kullanılarak denge sağlanmaya çalışılmıştır.

Çalışmada yer alan 50 değişkenin büyük bir kısmı oransal ölçekten oluşmaktadır. Çok değişkenli istatistiksel analiz yöntemlerinin uygulanması için gerekli görülen en düşük ölçek sıralayıcı ölçek olduğundan dolayı, sınıflayıcı ölçek olan bazı değişkenler analize dahil edilmemiştir. Seçilen değişkenler arasında çoklu doğrusal bağlantı durumu ile karşılaşma olasılığı yüksek olduğu için öncelikle boyut indirgeme yöntemleri ile analize başlanmıştır. Bunun için önce temel bileşenler analizi ile uygun boyut derecesi bulunmuş, faktör analizi ile hangi değişkenlerin belirli faktörler altında toplanabileceği belirlenmiştir. Faktör analizi ile birbirlerinden bağımsız değişkenler elde edilmiştir. Elde edilen değişkenler, sosyo-kültürel açıdan daha önce hiç sınıflandırılmamış olan illerin sınıflandırılması için kümeleme analizinde kullanılmıştır. Sosyo-kültürel açıdan oluşabilecek sınıf sayısı ve sınıf yapısı hakkında önceden bir bilgi olmadığı için öncelikle ayrıştırma analizi yerine kümeleme analizi yapılması uygun görülmüştür. Yapılan kümeleme analizi ile birbirleriyle benzer yapıda olan iller bir gruba alınarak küme içi birbirinin aynı, küme dışı birbirinden farklı gruplar elde edilmesi amaçlanmıştır. Elde edilen küme sayısı, ayrıştırma analizi için bir ön bilgi olarak kabul edilip, kümeleme analizi ile elde edilen grupların sınıflandırma başarısı hakkında fikir sahibi olmaya çalışılmıştır. Ayrıca ayrıştırma analizi ile grupları birbirlerinden ayırmada önemli rol oynayan değişkenler de belirlenmiştir. Bütün çok değişkenli istatistiksel yöntemler SPSS 15 ve MINITAB 14 paket programları kullanılarak yapılmıştır.

2.1 Temel Bileşenler Analizi

Temel bileşenler analizi (TBA), değişkenlerdeki değişim oranını inceleyerek çok boyutlu bir veri setinin gerçek boyutunun bulunmasında kullanılır. TBA, orijinal p değişkenin varyans yapısını daha az sayıda değişkenle aralarında doğrusal bir ilişki olan p sayıda değişkenin açıkladığı yapıyı, aralarında ilişki olmayan ve sayıca orijinal değişken sayısından az olan değişkenlerle ifade etme yöntemidir.

TBA, verilere genel bir bakış sağladığı için değişkenlerin dağılımının Normal dağılıma uyup uymadığı ve veri setinde herhangi bir uç değer olup olmadığı bu analiz yardımıyla incelenebilir.

Temel bileşenler analizinde değişkenlerdeki değişim yapısı, korelasyon ya da kovaryans matrisleri üzerinden incelenebilir. Veri matrisinde, değişkenler arasında birim farklılığı yoksa ve değişkenlerin varyans ve kovaryansları birbirlerine yakın değerler alıyorsa

TBA kovaryans matrisi ile uygulanmalıdır. Tersine veri matrisinde, değişkenler arasında ölçü birimi açısından farklılıklar büyükse ya da değişkenlerin değişim aralıkları analizi etkileyecek kadar çok farklılık gösteriyorsa bu durumda verilerin standardize edilerek kullanılması uygundur. Verilerin kovaryans matrisini kullanmak yerine orijinal verilerin korelasyon matrisi kullanılarak TBA yapılır.

Temel bileşenler analizinde, özdeğerler her bir bileşenin sahip olduğu varyansı göstermektedir. Buna göre; ilk özdeğer barındırdığı en yüksek değişkenliği barındırmalı ve bunu izleyen özdeğerler de bu sırayla gitmelidir. Her bir bileşendeki değişimin toplam değişime oranıyla veri setindeki değişimin ne kadarının açıklanabildiği görülmekte ve bileşen sayısına karar verilebilmektedir.

Bileşen sayısı belirlenirken özdeğerler yardımıyla;

1. Korelasyon matrisi için, 1'den büyük olan özdeğer sayısı kadar temel bileşen seçilir,
2. Toplam değişimin en az %67'sini (Johnson, 1998) ya da toplam değişimin ne kadarının açıklanması hedefleniyorsa ona denk gelen özdeğer sayısı kadar temel bileşen seçilir,
3. Özdeğerler bulunduktan sonra, bu özdeğerler büyüklük sırasına göre çizgi grafiğinde gösterilir. Bu çizgi grafiğinin eğiminin sabitleştiği noktadaki (kırılma noktası) özdeğer sayısı kadar temel bileşen seçilir.

2.2 Faktör Analizi

Faktör analizi, temel bileşenler analizi gibi çok değişkenli veri setlerinde boyut indirgemesinde kullanılan analizlerden biridir. Bu iki analizin temel farkı, temel bileşenler analizinde; bileşenler, değişkenlerin doğrusal kombinasyonu şeklinde ifade edilirken; faktör analizinde değişkenlerin faktörlerin doğrusal kombinasyonu şeklinde ifade edilmesidir. Aynı şekilde temel bileşenler analizi, değişkenlerdeki değişim miktarlarıyla ilgilenirken, faktör analizi kovaryans ve korelasyon yapılarıyla ilgilenir. Ayrıca, faktör analizi değişkenleri gruplayarak ortak faktör tanımlama özelliğine sahiptir.

Faktör analizinin ana amacı, ilişkisiz yeni değişkenler türetmek ve bu yeni değişkenlere isimler vererek deney birimi bazında değerlendirme yapmaktır.

Faktör Analizi Modeli:

$$\mathbf{x}_j = \boldsymbol{\mu}_j + \lambda_{j1}f_1 + \lambda_{j2}f_2 + \dots + \lambda_{jm}f_m + \eta_j \quad j = 1, 2, \dots, p \quad (1)$$

\mathbf{x}_j : Cevap değişkeni vektörü

$\boldsymbol{\mu}$: İlgilenilen kitlenin ortalama vektörü

p : Orijinal değişken sayısı

m : Faktör sayısı

f : Faktörler

η_j : Spesifik varyans

Varsayımlar:

- f_k , ortalaması 0, varyans 1 olan birbirinden bağımsız, özdeş dağılıma sahiptirler.
 $k=1, 2, \dots, m$

- f_k ve η_j , k ve j 'nin bütün kombinasyonları için bağımsız dağılıma sahiptirler.

Faktör Döndürme:

Bazen orijinal faktör yüklerinden bilgi elde edilmesi zor olabilir. Bu nedenle faktör yapısını daha basit hale getirmek için belirli bir açı ile döndürmek uygun olacaktır.

Faktör döndürme, faktör yüklerinin dik hale getirilmesi için eksenlerin optimal açı ile döndürülmesi ve dikleştirilmesinin sağlanması olarak ifade edilebilir. Faktör döndürme ile faktörlere ait varyans, spesifik varyans, korelasyon ya da kovaryans matrisi değişmez. Yaygın olarak kullanılan faktör döndürme yöntemleri, Dik ve Eğik Döndürme yöntemleridir. TBA ile yorumlanamaz ya da anlamlı çıkmayan sonuçların, Faktör döndürme yöntemi ile yorumlanabilir sonuçlar olarak çıkmasının mümkün olması faktör analizinin TBA'ya üstünlüğüdür.

2.3 Kümeleme Analizi

Kümeleme analizi, veri matrisinde yer alan ve doğal gruplamaları kesin olarak bilinmeyen birimleri, değişkenleri ya da birim ve değişkenleri birbirleriyle benzer olan alt kümelere (grup, sınıf) ayırmaya yardımcı olan yöntemler topluluğudur. Kümeleme analizinin genel amacı, gruplanmamış verileri benzerliklerine göre bazı ölçütlerle (benzerlik, uzaklık vs.) sınıflandırmada ve araştırmacıya uygun, işe yarar özetleyici bilgiler elde etmede yardımcı olmaktır.

Kümeleme analizi, temel olarak dört değişik amaca yönelik uygulanan bir yöntemdir (Özdamar, 1999):

1. n sayıda birimi, p değişkene göre saptanan özelliklerine göre olabildiğince kendi içinde homojen ve kendi aralarında heterojen alt gruplara (kümelere) ayırmak,
2. p sayıda değişkeni, n sayıda birimde saptanan değerlere göre ortak özellikleri açıkladığı varsayılan alt kümelere ayırmak ve ortak faktör yapıları ortaya koymak,
3. Hem birimleri hem de değişkenleri birlikte ele alarak ortak n birimi p değişkene göre ortak özellikli alt kümelere ayırmak,
4. Birimleri, p değişkene göre saptanan değerlere göre, izledikleri biyolojik ve tipolojik sınıflamayı ortaya koymak.

Kümeleme analizinde ilk aşama, birim ya da değişkenlerin doğal gruplamaları hakkında kesin bilgilerin bulunmadığı kitlelerden alınan n sayıda birimin p sayıda değişkenine ilişkin gözlemlerin elde edilmesidir. İkinci aşama, birimlerin/değişkenlerin birbirleriyle olan benzerliklerini ya da farklılıklarını gösteren uygun bir benzerlik ölçüsü ile birbirlerine uzaklıklarının belirlenmesi aşamasıdır (Uzaklıklar matrisinin elde edilmesi).

Üçüncü aşamada, uygun kümeleme yöntemi seçilerek birim/değişkenlerin uygun sayıda kümeye ayrılması aşamasıdır. Son aşama, kümelerin yorumlanması, sonuçların duyarlılığının ve anlamlılığının tartışılması, kurulan hipotezlerin doğrulanması ve gerekli analitik yöntemlerin uygulanması aşamasıdır. Sonuçların uygun olmaması durumunda (değişkenlerin uygun olmaması ve/veya küme sayısının doğru belirlenmemiş olması, vs.) tekrar ikinci aşamaya dönülerek işlemler tekrarlanır (Tatlıdil, 1996).

Varsayımlar: Kümeleme analizi, kitleyi iyi temsil eden örneklemden yola çıkarak parametre hakkında istatistiksel çıkarım yapma yöntemi değildir. Aslında, kümeleme analizi, gözlemler grubunun özellik yapılarını ölçmeye çalışan bir yöntemdir. Bu nedenle, güçlü matematiksel özellikler içermesine rağmen istatistiksel bulgularda güçlü değildir. Diğer yöntemlerde çok önemli olan doğrusallık, normallik ve homojen varyanslılık varsayımları diğer yöntemlerde olduğu kadar vazgeçilmez değildir (Hair vd., 2006).

Kümeleme analizinde, ayırıştırma analizinde olduğu gibi verilerin normal dağılımlı olması gerektiği varsayımı olmakla birlikte bu varsayım prensipte kalmakta, uzaklık değerlerinin normalliği yeterli görülmektedir (Tatlıdil, 1996).

Ayrıca kümeleme analizinde, kovaryans matrisine ilişkin herhangi bir varsayım bulunmamaktadır.

Uzaklık/Farklılık Ölçütleri: Uzaklık ölçüleri ya da benzerlik ölçüleri veri matrisinde yer alan değişkenlerin ölçek türlerine göre farklılık göstermektedir. Eğer değişkenler oransal yada aralıklı ölçekle elde edilmiş değerler ise uzaklık ya da ilişki türü ölçütlerden yararlanılır. Ölçümler sayısal değerler olarak yapılmış ise tercih edilen ölçütler Ki-kare uzaklık ölçütüdür. Eğer sınıflayıcı ölçekli gözlemler alındıysa birimler arasındaki benzerlikleri belirlemede Öklid, Karesel Öklid gibi ölçütler kullanılır (Özdamar, 1999). Literatürde uygulanan pek çok uzaklık ölçütü yer almaktadır. Fakat burada sadece en çok kullanılan üç tanesi hakkında bilgi verilmiştir.

1. Öklid Uzaklığı: En basit uzaklık ölçüsüdür. İki gözlem arasındaki uzaklığı, p -boyutlu örneklem uzayında düz bir çizgi ile ölçülebiliyorsa bu uzaklık ölçütü kullanılır (Johnson 1998). Uzaklık matrisinde birimler arasındaki uzaklıklar simetriktir ve iki nokta arasındaki uzaklık dönüşümlüdür. Bu uzaklık kolaylıkla üç değişken için de genellenebilir (Hair vd., 2006).

\mathbf{x}_r ve \mathbf{x}_s gibi iki nokta arasındaki Öklid uzaklığı eşitlik 2 yardımıyla hesaplanır.

$$d_{rs} = [(\mathbf{x}_r - \mathbf{x}_s)'(\mathbf{x}_r - \mathbf{x}_s)]^{1/2} \quad (2)$$

2. Karesel Öklid Uzaklığı: Karesel Öklid uzaklığı ise, aynı Öklid uzaklığı gibi hesaplanır. Değişkenlere göre toplam uzaklığın karekökü alınmaksızın elde edilen uzaklıklar doğrudan matrise yazılır. Eşitlik 3 kullanılarak hesaplanır.

$$d_{rs}^2 = [(\mathbf{x}_r - \mathbf{x}_s)'(\mathbf{x}_r - \mathbf{x}_s)] \quad (3)$$

3. Mahalanobis Uzaklık (D^2): Her bir değişkene eşit ağırlık vererek değişkenler arasındaki korelasyonu dikkate alan bir uzaklık ölçütüdür. Standartlaştırılmış değişkenler içinde bu uzaklık ölçütü hesaplanabilir (Hair vd., 2006). Eşitlik 4 kullanılarak elde edilir.

$$d_{rs} = [(\mathbf{x}_r - \mathbf{x}_s)' \Sigma^{-1} (\mathbf{x}_r - \mathbf{x}_s)]^{1/2} \quad (4)$$

Kümeleme Yöntemleri: Kümeleme yöntemleri, birim ya da değişkenleri uygun gruplara ayırırken, grupları belirlemede izledikleri yaklaşımlara göre iki temel gruba ayrılırlar. Bu iki yöntem en yaygın kullanılan kümeleme yöntemleridir.

1. Hiyerarşik Kümeleme Yöntemleri: Hiyerarşik kümeleme yöntemleri, birimlerin benzerliklerini dikkate alarak belli düzeylerde birbirleri ile birleştirmeyi amaçlayan yöntemlerdir. Bu yöntemler, birimleri birbirleriyle değişik aşamalarda bir araya getirerek ardışık biçimde kümeler belirlemeyi ve

bu kümelere girecek elemanların hangi uzaklık düzeyinde küme elemanı olduğunu belirlemeye yöneliktirler. Hiyerarşik kümeleme yöntemleriyle bir dendogram üretilmesi, bir bireyin tüm birimlerine olan uzaklıklarının hesaplanması ile başlar. Tüm bireylerin ayrı birer grup oldukları kabul edilir. Daha sonra birbirine yakın olan bireyler birleştirilerek gruplar haline getirilir.

2. Hiyerarşik Olmayan Kümeleme Yöntemleri: Küme sayısı konusunda ön bilgi varsa ya da araştırmacı anlamlı olacak küme sayısına karar vermiş ise bu durumda çok uzun zaman alan hiyerarşik yöntemler yerine hiyerarşik olmayan yöntemler tercih edilir. Ayrıca bu yöntemlerin kuramsal dayanaklarının daha güçlü olması diğer bir tercih nedenidir (Tatlídil, 1996). Hiyerarşik olmayan kümeleme yöntemlerinde, tüm bireyler başlangıçta tek grup olarak düşünülüp daha sonra her bir birey grupta yer alacak şekilde n gruba bölünür (Mankly, 1990).

2.4 Ayrıştırma Analizi

Ayrıştırma analizi, veri setindeki değişkenlerin iki ya da daha fazla gerçek gruplara ayrılmasını sağlayan birimlerin p tane özelliğini ele alarak bu birimlerin doğal ortamdaki gerçek gruplarına, sınıflarına en uygun düzeyde atanmalarını sağlayacak fonksiyonlar türeten bir yöntemdir.

Ayrıştırma analizi, başlangıçta tanımlanan sınıflandırma değişkeninin incelenen bireylerin gruplandırılmasını ne ölçüde başardığını ortaya koyan, gruplar arasında ayrım sağlama konusunda en fazla etkisi olan değişken veya değişkenleri belirleyen ve aynı değişkenler ile yeni bir bireyin hangi grupta yer alabileceği konularının ortaya çıkarılabilmesini amaçlayan çok değişkenli bir analizdir (Gümüş, 1996).

Ayrıştırma fonksiyonu en iyi ayrılmış grup değişkenlerinin doğrusal bir fonksiyonudur (Rencher, 2001). Ayrıştırma analizi, bir fonksiyonda gruplar arası varyansın grup içi varyansa oranını maksimum kılarak en iyi ayrımın yapılmasını sağlar (Balakrishnama ve Ganapathiraju, 2008). Analiz sonucunda yapılan sınıflama ile orijinal grup üyeliklerinin karşılaştırılması bilinen fonksiyonun yeterliliğini test etmeye olanak sağlar (Erçetin, 1993).

Ayrıştırma analizinin iki temel görevi vardır. Birincisi grupları birbirinden ayırtırmayı sağlayan fonksiyonları bulmak ve hesaplanan fonksiyonlar aracılığıyla yeni gözlenen bir birimi sınıflama hatası minimum olacak şekilde m gruptan herhangi birine atamaktır.

Varsayımlar: Ayrıştırma analizinde, aşağıda belirtilen varsayımların sağlanamaması, sınıflandırmanın başarısını önemli derecede etkileyecektir. Varsayımlar:

- k tane bağımsız değişken çok değişkenli normal dağılıma sahiptir.
- k tane bağımsız değişken birbirlerinden bağımsızdırlar ve aralarında çoklu doğrusal bağlantı yoktur.
- m tane gruba ait k tane bağımsız değişkenin varyans-kovaryans matrisleri homojen bir yapıdadırlar.
- Ayrıştırma fonksiyonlarının parametreleri ile ilişkileri doğrusaldır.
- Veriler arasında uç değer yoktur.

Değişkenlerin çok değişkenli normal dağılıma sahip olamaması, bağımsız olmamaları veya çoklu doğrusal bağlantıya sahip olmaları gibi durumlarda çeşitli dönüşümler yapılarak varsayımlar sağlanmaya çalışılır. Varyans-kovaryans matrislerinin homojenliği sağlanamıyorsa, kovaryansı fazla olan birimlerin farklı bir sınıfa atama yapacağından ayrıştırma analizi için hatayı büyüten bir unsur olacaktır. Bu durumda da havuzlanmış varyans-kovaryans matrisi kullanılabilir ya da lojistik regresyon gibi başka yöntemlere geçiş yapılabilir (Hair vd., 2006). Veriler arasında uç değerlerin varlığı tespit edilirse ve gerçekten uç değer olduğuna testlerle karar verilirse bu uç değerlerin atılma işlemi de yapılabilir (Hair vd., 2006).

Ayrıştırma Kuralları: Ayrıştırma analizinde sınıflandırmalar dört kurala göre yapılır. Kurallar aşağıdaki gibidir:

- Olabilirlik Kuralı
- Doğrusal Ayrıştırma Fonksiyonu Kuralı
- Mahalanobis Uzaklık Kuralı,
- Ardıl Olasılık Kuralı

Yukarıda sayılan dört kurala göre ayrıştırma analizi yapıldıktan sonra yanlış sınıflandırma olasılıkları da çeşitli yollarla hesaplanabilir. Bunlardan en çok kullanılan yöntem Çapraz Geçerlilik kestirimidir. Bu yöntemle elde edilen sonuç, diğer yöntemlere göre yansız olmaktadır (Johnson, 1998). Bu yöntem, gözlemin veri setinden çıkarılıp kalan veriler üzerinden geliştirilen kuralla hangi gruba yerleştirildiğine dayanan bir yöntemdir. Veri, kuralın geliştirilmesinde kullanılmadığı için yanlış sınıflandırma olasılığını hesaplamada daha gerçekçi sonuçlar üretir.

3. BULGULAR

3.1 Temel Bileşenler Analizi İle İlgili Bulgular

Analize başlamadan önce, korelasyon veya kovaryans matrisinden hangisinin kullanılacağına karar vermek gerekir. Bunun için araştırmada kullanılan 50 adet değişkene ait korelasyon ve kovaryans matrisleri elde edilmiştir. Her iki matrisin de boyutu (50x50) çok büyük olduğundan dolayı çıktı olarak verilememiştir.

Korelasyon matrisi incelendiğinde, değişkenler arası korelasyonun yüksek ve istatistiksel olarak anlamlı olduğu görülmüştür. Bu matrislere ait korelasyon matrisinin determinantının 0'a çok yakın çıkması ($\text{Det}(R)=1.640E^{-66}$) ve varyans-kovaryans matrisinde değişkenler arası değişimin oldukça farklı çıkması nedeniyle temel bileşenler analizinin korelasyon matrisi üzerinden uygulanmasına karar verilmiştir. TBA ile ilgili bütün SPSS çıktıları Ek 2'de verilmektedir.

Değişkenler arasında bağımlılık yapısını görmek için Bartlett testi uygulanmıştır. Bu test, değişkenler arasındaki bağımlılık yapısının, Temel bileşen analizi veya Faktör analizinin (FA) yapılmasına uygunluğunu test etmektedir. Çünkü değişkenler arasında bağımlılık yoksa bu analizlerin yapılmasına da gerek yoktur.

Bu testte, sıfır hipotezi, "Değişkenler arasında ilişki yoktur. ($P = I$)" olarak kurulur. Tablo 2'de verilen çıktı sonucuna göre, sıfır hipotezinin reddedilmesi ile değişkenler arasındaki ilişki derecesinin TBA veya FA analizlerini yapmaya uygun olduğuna karar verilir.

İllere göre sosyo-kültürel gelişmişlik derecelerinin belirlenmesi için yapılan çalışmada, boyut indirgemek için uygulanan Temel bileşen analizinde bileşen sayısı belirlenirken, öncelikle özdeğerlerin açıklama oranlarına, sonra da çizgi grafiğe bakılmıştır. Tablo 3’de verilen sonuçlara göre, $\lambda \geq 1$ olan 7 bileşen (özdeğer) belirlenmiştir ve bu özdeğerler, değişimin %87’sini açıklamaktadır. Daha sonraki özdeğerlerde açıklama oranı artmakla beraber, bu özdeğerlerin 1’den düşük değerlere sahip olması nedeniyle analizin 7 bileşen üzerinden yapılmasına karar verilmiştir.

Şekil 1’de gösterilen çizgi grafiğine bakıldığında, 7 özdeğerden sonra özdeğerler 0’a çok yakın değerler aldıkları ve şekilde de 7. özdeğerden sonra oluşan kırılma nedeniyle bu çizimde de bileşen sayısı 7 olarak belirlenmiştir.

Uygulanan temel bileşenler analizi sonucunda yedi bileşen için elde edilen katsayılar Tablo 4’de verilmiştir. Bu katsayılar, temel bileşen denkleminin oluşturulmasında kullanılan katsayılardır. Aynı zamanda temel bileşen skorları elde edilirken de bu katsayılardan yararlanır.

Oluşturulan temel bileşenlerin normal dağılıma uygun olma durumları ve uç değerlerin varlığının tespiti için tanımlayıcı istatistikleri elde edilmiştir. İncelenen tanımlayıcı istatistiklerde bazı illerin skorlarının diğer değişkenlerden oldukça farklı olduğu, yani skorlar arasında uç değerlerin varlığı tespit edilmiştir. Uç değerlerin daha net görülebilmesi için, MINITAB 14’de çizdirilen, yedi temel bileşen için Şekil 2’de gösterilen serpm diyagramı çizdirilmiştir. Bazı skorlarda uç değer olabilecek değerler olduğu serpm diyagramından da görülmektedir.

3.2 Faktör Analizi İle İlgili Bulgular

Yedi faktör için en çok olabilirlik yöntemiyle faktör yükleri belirlenmiş ve elde edilen faktörlere Kaiser Normalizasyonu ile Varimax Rotasyonu uygulanmıştır. Ek 3’de düzenlenmiş faktör yükleri ve oransal varyans değerleri yer almaktadır. Uygulanan faktör analizi sonucunda, her bir faktörde yer alan değişkenler Tablo 5’de gösterilmiştir. Tablo 5’deki faktör yüklerine göre, her bir faktörde yer alan değişkenlere göre belirlenen faktör isimleri aşağıdaki gibidir:

Faktör 1: Sosyal ve Kültürel Göstergeler

Faktör 3: Eğitim Göstergeleri

Faktör 5: Sağlık Göstergeleri

Faktör 7: Diş hekimi Göstergesi

Faktör 2: Demografik Göstergeler

Faktör 4: Gelişmişlik Göstergeleri

Faktör 6: Tesis/Yatırım Göstergeleri

Faktör skorları illere göre değerlendirildiğinde, en yüksek ve en düşük skorların alındığı üç illerin bir arada gösterildiği Tablo 6 oluşturulmuştur. Bu tabloya göre, sosyal ve kültürel göstergeler açısından *Hakkari*, *Batman* ve *Bilecik* en düşük skorlara sahip ilken, *İzmir*, *Ankara* ve *İstanbul* en yüksek il durumundadırlar. İkinci Faktör olan demografik göstergelere göre *Tekirdağ*, *Antalya* ve *Muğla* en düşük skorlara sahip ilken *Ardahan*, *Siirt* ve *Mersin* en yüksek skorlara sahip il durumundadırlar. Burada en yüksek skorlara sahip olmak o illerin çok gelişmişliğini göstermez. Çünkü demografik özelliklerin içerisinde yer alan değişkenlere, onların faktör yüklerine ve işaretlerine (negatif/pozitif) bakıldığı takdirde en yüksek skorları alan illerde doğurganlık hızı, bebek ölüm hızı gibi değişkenlerin önemli ağırlıklara sahip olduğu görülmektedir. Yani bu 3 ilin diğer illere göre yüksek skor alması aslında adı geçen değişkenler açısından

olumsuz anlamda diğer illerden farklı olduğunu gösterir. Dikkat çekici bazı faktörler değerlendirilecek olursa, Sağlık göstergeleri açısından *Gaziantep*, *Kocaeli* ve *Mersin* illerinin diğer illere göre olumlu yönde daha yüksek skora sahip oldukları görülmektedir. Örneğin Tesis/yatırım göstergeleri açısından en iyi durumda olan il *İzmir*'dir. Burada, Üniversite oyunlarının İzmir'de gerçekleştirilmesinin önemli bir rolü olduğu düşünülmektedir.

3.3 Kümeleme Analizi İle İlgili Bulgular

81 il üzerinde yapılan faktör analizi sonuçlarında 50 değişken 7 değişkene indirilmiştir. Faktör analizi varsayımlarından dolayı, bu yeniden elde edilen değişkenler kesinlikle birbirleriyle bağımlı değildir. Çoklu doğrusallık sorunu ortadan kaldırılmıştır. Faktör skorlarına bakılarak da o değişkenler açısından hangi illerin birbirine yakın olduğu sonucuna sezgisel de olsa varılabilir. Fakat bu çalışmada, iller arasında bu değişkenler dikkate alınarak yapılan kümeleme analizi ile illerin adı geçen yedi değişken açısından kaç gruba ayrılacağı bilinmediği için kümeleme analizi yapılması uygun görülmüştür.

Literatürde, kümeleme analizi için, değişkenler arasında ölçek farklılığı olduğu durumda, verileri standartlaştırmanın kümeleme başarısını daha iyileştireceği söylenmektedir. Fakat bu çalışmada, faktör analizi sonucunda elde edilen yeni değişkenler ve onların skorları yeni veriler olarak ele alınmış ve kümeleme analizi bu değişkenler üzerinden gerçekleştirilmiştir. Sadece kontrol amacıyla, ham verilere standartlaştırma işlemi uygulanarak da kümeleme analizi yapılmıştır. Fakat bu durumda kümeleme başarısı oldukça düşmüştür. Kümeleme sayısı da oldukça fazla çıkmıştır.

Kümeleme analizi sadece iller bazında yapılacağı için, 7 değişken açısından illerin birbirlerine uzaklıklarına bakılmıştır. Bunun için Mahalanobis Uzaklık ölçütü kullanılması uygun görülmüştür. Bu çalışmada matris boyutları çok büyük olduğundan, sadece matrislerden çıkarılan sonuçlar tablolar halinde verilebilmiştir. Bu ölçüte göre, birbirlerine uzaklığı en fazla olan iller aynı küme içerisinde yer alamaz. Kümeleme analizi sonucunda elde edilen illerin grupları, küme-için ortalama uzaklıkları ile birlikte Tablo 1'de verilmiştir. Bu tabloya bakıldığında, ilk grupta yer alan illerin diğer gruplara göre daha fazla gelişmiş oldukları kesinlikle söylenemez. Bu tabloya bakılarak sadece, birbirlerine yakın olan illerin aynı grupta yer aldıkları, daha homojen bir yapıda olacak şekilde bir grupta toplandıkları söylenebilir.

Tablo 1. Kümeleme analizi sonucunda elde edilen gruplandırmalar

	İl Grupları	Küme-içi ortalama uzaklık
1.Grup	Adıyaman, Afyon, Ağrı, Aksaray, Amasya, Antalya, Ardahan, Artvin, Balıkesir, Bartın, Bayburt, Bingöl, Bitlis, Burdur, Bursa, Çanakkale, Çankırı, Çorum, Denizli, Düzce, Erzincan, Erzurum, Eskişehir, Giresun, Gümüşhane, Hatay, Iğdır, Isparta, Karaman, Kars, Kastamonu, Kayseri, Kırıkkale, Kırşehir, Kilis, Kocaeli, Konya, Kütahya, Malatya, Mersin, Muğla, Muş, Ordu, Osmaniye, Rize, Sakarya, Samsun, Siirt, Sinop, Sivas, Şırnak, Tekirdağ, Tokat, Trabzon, Tunceli, Uşak, Yozgat, Zonguldak	2.0759
2.Grup	Aydın, Bolu, Edirne, Elazığ, Manisa, Niğde,	1.9137
3.Grup	Kahramanmaraş, Mardin, Şanlıurfa, Van	1.7304
4.Grup	Bilecik, Karabük, Kırklareli, Nevşehir, Yalova	2.3199
5.Grup	Batman, Diyarbakır, Gaziantep	1.7736
6.Grup	Ankara, İzmir	3.5432
7.Grup	Adana, Hakkari, İstanbul	4.3019

3.4 Ayrıştırma Analizi İle İlgili Bulgular

Kümeleme analizinde yapılan 7 gruplandırmanın anlamlı olup olmadığı, ayırma analizi ile test edilmiş, en yüksek başarı düzeyin 7'li sınıflandırma sonucunda elde edildiği belirlenmiştir. Faktör Analizi ile belirlenen 7 değişken için gruplar arası farklılıkları ortaya çıkaran değişkenlerin belirlenmesi için Ayrıştırma Analizi yapılmıştır. Analiz sonuçlarına göre, sınıflandırma başarısı %95.1 olarak bulunmuştur.

Ayrıştırma Analizi'nde amaç, faktör analizi ile elde edilen gösterge değerlerine göre, iller arasındaki farklılıkları ortaya koymaktır. Ayrıca bu analizle 7 grup arasındaki farklılıkları ortaya çıkaran önemli değişkenler de tespit edilmeye çalışılmıştır.

Varsayımlar: Ayrıştırma Analizi'nde faktör analizi sonucunda elde edilen birbirinden bağımsız 7 değişken kullanıldığı için, bu değişkenler birbirlerinden bağımsız ve bağımlı değişkenle doğrusal bir ilişkiye sahiptirler.

Ek 4'de yer alan, Tablo 8'de SPSS çıktısı verilen testler, yedi grup arasında bağımsız değişkenlerin kovaryans matrislerinin benzerliklerini değerlendirmek için yapılır. Test sonucunda, 7 grubun 7 değişken açısından kovaryans matrislerinin birbirlerine eşit olmadığı görülmektedir. Bunun için analizde, havuzlanmış kovaryans matrisi (Pooled Covariance Matrix), öncül olasılıkları da Tablo 9'da gösterildiği gibi grup büyüklükleri ile ağırlıklandırılmış olarak, değişken seçim yönteminde de Aşamalı Değişken Seçme Yöntemi kullanılmıştır.

Tablo 10'da gösterilen Ayrıştırma Analizi tablosuna göre, demografik gösterge ve sağlık gösterge değişkenleri haricinde tüm diğer değişkenlerin ortalamaları birbirlerinden farklı çıkmıştır. Mahalanobis uzaklık değeri, gruplararası ayrıştırma derecesinin belirlenmesini sağlar. Her bir değişken için, minimum uzaklık değerine sahip olan gruplar en yakın gruplardır. Örneğin demografik göstergeler değişkeni en büyük minimum Mahalanobis uzaklık değerine sahiptir. Bu da bu değişken için gruplararası farkın büyük olduğunu gösterir.

Tablo 11 ve 12, aşamalı değişken seçim yönteminin çıktıları olarak Ek 4'te verilmiştir. Wilks' Lamda değeri, hem tek yönlü hem de çok yönlü varyans analizinde

kullanılabilen bir istatistiktir. İlk aşamada bu değerler birbirlerinden ayrı olarak kıyaslanırken bu aşamada değişkenlerin birbirleri arasındaki ilişkiler de gözönünde bulundurularak hangi değişkenleri ayırtmada önem kazandığı araştırılmıştır. Aşamalı değişken seçme yöntemi sürecinde, en yüksek Wilks' Lamda değerine ait değişken anlamlı ise ayırtma değişkeni olarak kullanılmaktadır. Daha sonra ise, bu değişkenden kaynaklanan grup farklılıkları hesaba katılarak ANCOVA tablosunda en yüksek F değeri veren değişken bulunur ve anlamlı ise bu değişken de ayırtma değişkeni olarak kullanılır, bu noktada diğer değişkenlerin anlamlılıklarını koruyup korumadıkları kontrol edilir. Bu süreç anlamlı değişken kalmayana kadar devam eder. Aşamalı değişken seçme yöntemine ait çıktılara göre, *Diş Hekimi, Sosyo-Kültürel, Eğitim, Tesis/Yatırım* ve *Gelişmişlik Göstergeleri* değişkenleri olmuştur. Her bir değişken eklendikçe ayırtma daha da artmıştır. Wilks' Lamda değerlerinin düşmesi ayırtmanın daha iyi durumda olduğunu göstermektedir. Bu nedenle ayırtma analizi sırasında bu beş değişken kullanılacaktır.

Kanonik ayırtma analizi, temel bileşen analizi ve faktör analizinde olduğu gibi değişkenlerin boyutunu azaltmak ve bu indirgenen boyut üzerinden sınıflandırma yapmak amaçlı uygulanan bir analizdir. Ancak diğer analizlerden farklı olarak varyans-kovaryans matrisi ya da korelasyon matrisi değil, grup içi değişim ve gruplararası değişim matrislerini kullanır.

Tablo 13 ve 14 çıktılara göre, ayırtma modelinin bütün olarak uyum iyiliğine bakıldığında, beş fonksiyonun istatistiksel olarak anlamlı olduğu görülmektedir. İlk fonksiyon, beş fonksiyon tarafından açıklanan varyansın %63'ünü açıklamaktadır. En az açıklama yüzdesine de 5.fonksiyon sahiptir. Toplamda, varyansın %100'ünü bu beş fonksiyon açıklamaktadır. Yani grup ortalamalarının beş boyutta ifade edilebildiğini ve beş kanonik fonksiyonun kullanılabileceğini gösterir. Bu fonksiyonlara ait Wilks' Lamda değerinin istatistiksel olarak anlamlı olması gruplar arasında farklılık olduğunu göstermektedir.

Tablo 15 ve 16'da standartlaştırılmış ve standartlaştırılmamış kanonik ayırtma fonksiyon katsayıları kanonik skorların elde edilmesinde kullanılmaktadır. Buna göre standartlaştırılmış beş kanonik fonksiyon eşitlik 5'te gösterildiği gibidir.

$$\begin{aligned}
 CANFONK_1 &= 1.182SosKul + 0.175Egit - 0.052GelGost + 0.231TesYat + 1.134Dis \\
 CANFONK_2 &= 0.151SosKul + 0.849Egit - 0.384GelGost + 0.725TesYat - 0.386Dis \\
 CANFONK_3 &= -0.062SosKul + 0.599Egit + 0.219GelGost - 0.652TesYat + 0.082Dis \\
 CANFONK_4 &= 0.128SosKul + 0.026Egit + 0.914GelGost + 0.248TesYat - 0.131Dis \\
 CANFONK_5 &= -0.549SosKul + 0.123Egit + 0.096GelGost + 0.283TesYat + 0.429Dis
 \end{aligned}
 \tag{5}$$

Tablo 17 ve 18'de yer alan beş değişkene ait ayırtma fonksiyonlarındaki ağırlıklara ve grup ortalamalarına bakıldığında, birinci ayırtma fonksiyonunda, Grup 6 ve Grup 7'yi *Sosyo-kültürel* ve *Diş hekimliği* gösterge değişkenleri açısından diğer gruplardan farklı olduğu görülmektedir. İkinci ayırtma fonksiyonu, Grup 3,5 ve 6'yı *Eğitim göstergeleri* ve tesis/yatırım göstergeleri açısından farklı ayırtmıştır. Üçüncü ayırtma fonksiyonuna göre, 4,5 ve 6 grupları *Eğitim göstergeleri* açısından farklı çıkmıştır. Dördüncü ayırtma fonksiyonuna göre, 5 ve 6 grupları *Gelişmişlik göstergeleri* açısından, beşinci ayırtma fonksiyonuna göre Grup 2'yi *Diş hekimliği* ve *tesis/yatırım* açısından diğer gruplardan ayırtmıştır. Yeni il olma durumu ile

karşılaştığında, değerlendirmeler üzerinden yeni il (yeni gözlem) bu merkezlerden hangisine yakınsa o gruba yerleşecektir.

Tüm bu ön analizler tamamlandıktan sonra belirlenen kanonik fonksiyon ile verilere ayrıştırma analizi uygulanmıştır. Tablo 19 çıktılarına göre, oluşturulan ayrıştırma kuralı, iller arasındaki sınıflandırmayı %95 doğru olarak sınıflandırmaktadır. Çapraz geçerlilik yöntemi ile yapılan sınıflandırma sonuçlarına göre, *Antalya* Grup 1’de yer almışken, oluşturulan ayrıştırma kuralı ile Grup 6’nın içerisinde yer aldığı, *Hakkari*’nin Grup 7 yerine Grup 6, *Konya*’nın Grup 1 yerine Grup 7’de, *Mersin*’in Grup 1 yerine Grup 5’te yer aldığı görülmüştür.

Son olarak, analizde beş tane ayrıştırma fonksiyonu oluşturulduğundan, örnek olması açısından Fonksiyon 1 ve fonksiyon 2 için sınıflandırmanın başarısı göstermek amacıyla Şekil 3 çizdirilmiştir. Bu şekilde göre, birinci ayrıştırma fonksiyonu, Grup 6 ve Grup 7’yi, ikinci ayrıştırma fonksiyonu, Grup 3, 5 ve 6’yı *Eğitim göstergeleri* ve *tesis/yatırım göstergeleri* açısından farklı ayırtmıştır.

4. TARTIŞMA VE SONUÇ

Araştırmada öncelikle temel bileşen analizi ile illerin sosyo-kültürel gelişmişliğinin temel boyutları, faktör analizi ile boyutlarda yer alan değişkenler belirlenmiş ve 7 adet birbirinden bağımsız yeni değişken elde edilmiştir. Bu değişkenler de *sosyo-kültürel*, *demografik*, *eğitim*, *gelişmişlik*, *sağlık*, *tesis/yatırım* ve *diş hekimliği* göstergeleri olarak belirlenmiştir.

Faktör analiziyle elde edilen sonuç, her bir il için sosyo-kültürel gelişmişlik endeksi olarak tanımlanabilecek sayısal değerler elde edilmiş ve bu değerlere göre illerin sıralaması yapılmıştır. Daha sonra, gelişmişlik endeksi değerlerinde önemli oranlarda farklılık gösteren noktalar saptanarak, illerin farklı gelişmişlik düzeylerine göre 7 ayrı homojen grup oluşturulmuştur. Bunda amaç, aynı özellikleri taşıyan homojen il gruplarının belirlenmesidir. Daha önce yapılan ve DPT’nin 2003 yılında yaptığı illeri sıralama çalışmaları sadece sosyo-ekonomik gelişmişlik derecelerine göre yapılmıştır. Bu çalışmada, bu sıralama sosyo-kültürel açıdan yapılmıştır. Ama sosyo-ekonomik gelişmişlik sıralaması ile sosyo-kültürel gelişmişlik sıralaması arasında bağlantı olacağı düşüncesiyle DPT sıralamasıyla bu çalışmada elde edilen sıralama karşılaştırılmıştır. Bu karşılaştırma, sadece çalışmada elde edilen sosyo-kültürel gelişmişlik sıralaması ile sosyo-ekonomik sıralaması arasındaki tutarlılığı görmek için yapılmıştır. Tabii ki DPT, bu karşılaştırmayı 2003 yılında ve sosyo-ekonomik göstergeler açısından yapmıştır. Yani hem zaman hem de gösterge açısından arada farklar bulunması gayet doğaldır. Burada karşılaştırmaktan kasıt, 2003 yılına ait de olsa sosyo-ekonomik ile sosyo-kültürel açıdan illerin sıralamasında anlamlı bir tutarlılık yakalanabilir mi sorusuna yanıt bulmaktır. Zaten karşılaştırma istatistiksel anlamlılıktan uzak basit bir karşılaştırmadan ibarettir. Buna göre, sosyo-ekonomik gelişmişlik derecesi ile sosyo-kültürel gelişmişlik derecesi sıralamalarında ilk sıralarda beklendiği üzere bir değişiklik olmamıştır. Sosyo-ekonomik gelişmişlik sıralamasında ilk sıralarda yer alan *İstanbul*, *Ankara*, *İzmir*, *Bursa*, *Kocaeli*, *Adana* ve *Antalya* sosyo-kültürel açıdan da ilk sırayı alan iller olmuşlardır. Ekonomik gelişmenin etkisi şehre kültürel anlamda da olumlu yansıdığına işaret olduğu düşünülmektedir. Halbuki bazı iller var ki sosyo-ekonomik gelişmişlik sıralamasında orta sıralarda yer alırken sosyo-kültürel açıdan çok farklı bir tablo sergilemişlerdir. Örneğin, *Eskişehir*, *Yalova*, *Bolu*, *Edirne* ve *Bilecik* sosyo-

ekonomik gelişmişlik derecesinde ilk 20 şehir arasına girerken sosyo-kültürel açıdan, son 10 şehir arasındadırlar. Tam tersine *Yozgat*, *Ardahan* ve *Şırnak* illeri sosyo-ekonomik gelişmişlik derecesinde son 20 şehir arasında iken sosyo-kültürel gelişmişlik derecesi açısından ilk 30 il arasında yer almışlardır. Diğer yandan *Hakkari*, *Batman*, *Iğdır*, *Mardin* ve *Kars* her iki gelişmişlik derecesi açısından Türkiye'de son 10 il arasındadırlar. Bu illerde ekonomik gelişme olmadığı gibi toplumda kültürel bir gelişme de sağlanamamıştır.

Saptanan genel sonuçlardan biri, Türkiye'deki alansal sosyo-kültürel gelişme eğilimlerinin sıçramalardan çok yayılma dinamiklerince belirlendiği görüşüdür. Öyle ki birinci derecede sosyo-kültürel gelişmişlik grubunda yer alan illerden *İstanbul*, *İzmir* ve *Ankara* Türkiye'nin ilk ve en önemli gelişme merkezleri iken, *Kocaeli* ve *Bursa* ise özellikle *İstanbul*'dan yayılan gelişme faaliyetleri ile gelişme sürecine girmişlerdir.

İllerin sosyo-kültürel gelişmişlik endeksleri incelendiğinde iller arasındaki dengesizliklerin arttığı görülmektedir. Bunun en temel ve basit göstergesi sosyo-kültürel endekslerin değişim aralığı olacaktır. En yüksek endeks ile en düşük endeks arasındaki fark 7.30'dur. Sosyo-kültürel gelişmişlik endeksi kötü olan illere bakıldığında, bu kötüye gidişlerde rol oynayan en önemli nedenler arasında terör ve siyasi istikrarsızlık, terörden kaynaklanan göç olgusu, olumsuz coğrafi yapı, politik faktörler ve yaşanan ekonomik krizler gibi pek çok etken sayılabilir.

Araştırmanın sonuçlarında da görüldüğü gibi, Türkiye'de bölgesel sosyo-kültürel gelişmişlik farklılıklarını azaltmaya yönelik önlemler uygulanmış olmakla (özellikle doğuya yönelik okuma programları, göçü engellemek için yapılan yatırımlar) beraber, alınan önlemlerin bölgesel dengesizliklerle ilgili sorunlara etkin ve kalıcı çözümler getiremediği görülmektedir. Zira, Türkiye'de iller arasındaki sosyo-gelişmişlik farklılıklar arttığı gibi sosyo-kültürel farklılıklar da artmaktadır.

Çalışmada sosyo-kültürel gelişmişlik derecesine göre illerin sıralaması incelenmiş olsa da sosyo-ekonomik açıdan gelişmişlik derecesi ile paralel bir sıralama çıkması bekleniyordu. Genel olarak toplumda paylaşılan ortak mekanların artmasının, insanların kültürel paylaşımında bulunmalarının doğrudan ekonomik zenginleşme ve refah düzeyinin artması ile ilişkili olduğu inkar edilemez. Sürdürülebilir bir kalkınmanın sağlanabilmesi için nüfus dinamikleri ile doğal kaynaklar, ekonomik faaliyetler, teknolojik gelişme, sosyo-kültürel yapı arasındaki dengenin her seviyedeki planlama ve politika geliştirme süreçlerinde göz önünde bulundurulması gerekmektedir. Bu hedefe ulaşabilmek için nüfusun eğitim, sağlık ve insan gücü yönünden niteliklerinin iyileştirilmesi, yaşam kalitesinin yükseltilmesi ve bu alanlarda bölgeler ve yerleşim yerleri arasındaki farklılıkların azaltılması temel ilke olmalıdır.

5. KAYNAKLAR

Akyüz, K. C., Cımdık, H., Serin, H., Akyüz, İ., 2001. Çok Boyutlu İstatistiksel Yöntemler Kullanılarak Belirli Sosyo-Ekonomik Değişkenlerin 80 İl Düzeyinde Analizi. V. Ulusal Ekonometri ve İstatistik Sempozyumu, Adana.

Albayrak, A. S., 2005. Türkiye'de İllerin Sosyoekonomik Gelişmişlik Düzeylerinin Çok Değişkenli İstatistik Yöntemlerle İncelenmesi, İktisadi Araştırmalar Vakfı Yayınları, İstanbul, HT 178.T8ALB.

Albayrak, A. S., 2004. Türkiye'de İllerin Sosyoekonomik Gelişmişlik Düzeylerini Belirleyen Hipotetik Yapıların Faktör Analiziyle İncelenmesi. İşletme İktisadi Enstitüsü-Yönetim Dergisi, Yıl:15, Sayı:48, sayfa: 53-79.

Arı, Y., 2003. Teksas'ta Yabancı Doğumlu Nüfusun Karşılaştırmalı Analizi. Coğrafi Bilimler Dergisi, Cilt:1, Sayı:2, sayfa: 85-101.

Balakrishnama, S., Ganapathirajua, 2008.. Linear Discriminant Analysis - A Brief Tutorial", Institute for Signal and Information Processing Department of Electrical and Computer Engineering Mississippi State University, "Theory of LDA" http://lcv.stat.fsu.edu/research/geometrical_representations_of_faces/PAPERS/lda_theory.pdf, page:1-8, (10.05.2008).

Diñçer, B., 1996. İlçelerin Sosyo-ekonomik Gelişmişlik Sıralaması. DPT, Bölgesel Gelişme ve Yapısal Uyum Genel Müdürlüğü, Ankara.

Diñçer, B., Özaslan M., Satılmış E., 1996. İllerin Sosyoekonomik Gelişmişlik Sıralamasının Araştırılması, Ankara: DPT.

D.P.T. 1999. İller ve Bölgeler İtibariyle Çeşitli Göstergeler. Bölgesel Gelişme ve Yapısal Uyum Müdürlüğü.

Erçetin, Y., 1993. Diskriminant Analizi ve Bankalar Üzerine Bir Uygulama. Türkiye Kalkınma Bankası A.Ş, APM/28 (KİG-26).

Güveli, A., Kılıçkaplan, S., 2000. A Ranking of Islamic Countries in Terms Level of Socio-Economic Development, Journal of Economic Cooperation Among Islamic Countries, 21, 1, 97-114.

Gümüş, C., 1996. Orman Köyleri Kalkınma Planlarında Çok Boyutlu Yöntemlerden Yararlanma Olanakları, Express Ofset, 80 syf., Trabzon.

Härdle ,W., Hlávka, Z., 2001. Multivariate Statistics: Exercises and Solutions, Springer, New York.

Hair, J. F., Anderson R. E., Tatham, R. L., Black W. C., 2006. Multivariate Data Analysis, Prentice Hall, New Jersey.

Johnson, D. E., 1998. Applied Multivariate Methods for Data Analysts. Duxbury Press, California.

Karpat, G., Açıköz, Ş., 2008. Ülkelerin Kalkınmışlık Faktörleri Açısından Değerlendirilmesi, V. Ulusal Ekonometri ve İstatistik Sempozyumu, 19-22 Eylül 2001 Çukurova Üniversitesi Adana, <http://idari.cu.edu.tr/sempozyum/bil16.htm>, 06.2008.

Mankly, B. F. J., 1990. Multivariate Statistical Methods. Primer Fourt Editor. J. W. Arros Smith Ltd., Bristol.

Oğuzhan, A., Aydın, D., 2000. Trakya ve Batı Anadolu'da Yaşayan Farklı Apodemus Türleri Arasındaki İlişkilerin Diskriminant Analizi, Trakya Üniversitesi Dergisi Sosyal Bilimler C Serisi, 1(1):1-10.

Öner, Y., Tunç T., Altun İ., Bayraktar R. ve Yurdem A. D., 2006. Türkiye'de İllerin Sosyo-ekonomik Yönden Farklılaşmasında Etkili Olan Faktörlerin İncelenmesi Trakya ve Batı Anadolu'da Yaşayan Farklı Apodemus Türleri Arasındaki İlişkilerin Diskriminant Analizi, TÜİK, İstatistik Araştırma Sempozyumu Bildiriler Kitabı, 326-342.

Özdamar, K., 1999. Paket Programlar ile İstatistiksel Veri Analizi. 2. Baskı, Kaan Kitabevi, Eskişehir.

Rencher, A. C., 2001. Methods of Multivariate Analysis. Second Edition, Wiley & Sons Inc. Publication, New York.

Saraçoğlu, B., 1992. Ülkelerin Ekonomik Kalkınmışlık Düzeyleri Açısından İncelenmesi. Hacettepe Üniversitesi İktisadi İdari Bilimler Fakültesi Dergisi, 10:18-54.

Süer, S., Şahin, N., 2002. Türkiye'de Bölgesel Farklılaşmada Sosyo-Demografik Değişkenlerin Önemi 1990-1994, İstatistik Araştırma Dergisi, 1(2):263 -276.

Tatlıdil, H., 1996. Uygulamalı Çok Değişkenli İstatistiksel Analiz. Cem Ofset, Ankara.

Tunç T., Öner Y., Altun İ., Bayraktar R., Yurdemi A., 2006, Türkiye'de İllerin Sosyo Ekonomik Yönden Farklılaşmasında Etkili Olan Faktörlerin İncelenmesi, 15. İstatistik Araştırma Sempozyumu'06, Ankara.

Yüksel, N., Aksoy, H., Ercanoğlu, M., 2008. Çok Değişkenli İstatistiksel Analiz Yöntemleri ile Heyelan Duyarlılık Haritalarının Oluşturulması, 61. Türkiye Jeoloji Kurultayı, Bildiri Özleri Kitabı, 286-287.

ASSESSMENT OF THE CITIES IN TURKEY ACCORDING TO THEIR SOCIO-CULTURAL DEVELOPMENT LEVEL

ABSTRACT

Applying consistent precautionary measures and politics and taking the regional development in the distribution of investments into consideration become inevitable for providing development in Turkey's regions. In this research, the development levels of all cities in Turkey are aimed to be measured from a socio-cultural framework. For this purpose, homogenous city groups are constructed by using multivariate statistical methods from the socio-cultural indices of all cities. Convenient dimension is found by principal component analysis and factor analysis is used to determine which variables can be grouped under certain factors. The variables which are obtained from factor analysis are used in cluster analysis for classifying the cities that have not been classified before in terms of socio-cultural framework. The obtained number of clusters is considered as a prior information for discriminant analysis, and it is used to gather an idea on the classification success of groups obtained by the cluster analysis. Moreover, by using discriminant analysis, important variables that discriminate groups from each other are determined. The socio-cultural development ranking of the cities, obtained at the end of this study, are compared with the ranking of the cities in terms of socio-economic development index, which was obtained by the State Planning Organization and published in 2003.

Keywords: Discriminant, Factor, Cluster, Socio-cultural development.

Ek 1. Araştırmada Kullanılan Değişkenler

Tablo 1. Araştırmada Kullanılan Değişkenler

Değişkenler	
<i>Demografik Göstergeler</i>	
1	Toplam Nüfus
2	Şehirleşme oranı
3	Yıllık Ortalama Nüfus Artış Hızı
4	Doğurganlık Hızı
5	Ortalama Hane halkı Büyüklüğü
6	Yıllık Net Göç Hızı
7	Sosyo-ekonomik gelişmişlik sıralaması 1996/2003
8	Bölgeler
9	İllerin Gelişmişlik Derece Bölgesi
10	İllerin Gelişmişlik Endeksi
<i>Eğitim Göstergeleri</i>	
11	Okur-yazar Nüfus oranı
12	Okur-yazar kadın nüfusun toplam kadın nüfusuna oranı
13	Üniversite bitirenlerin okul bitirenlere oranı
14	İlkokul Okullaşma oranı
15	Ortaokul Okullaşma oranı
16	Okul başına öğrenci (ilköğretimde)
17	Öğretmen başına öğrenci sayısı (ilköğretimde)
18	Derslik başına öğrenci sayısı (ilköğretimde)
19	Okul başına öğrenci sayısı (ortaöğretimde)
20	Öğretmen başına öğrenci sayısı(ortaöğretimde)
21	Derslik başına öğrenci sayısı (ortaöğretimde)
22	Anasınıfı okul sayısı
23	Anasınıfı öğrenci sayısı
24	Anasınıfı Öğretmen Sayısı
25	Anasınıfı derslik sayısı
26	İlköğretim Okul Sayısı
27	İlköğretim Şube Sayısı
28	Ortaöğretim Okul Sayısı
29	Ortaöğretim Şube Sayısı
30	İlkokul Öğrenci Sayısı
31	Ortaokul Öğrenci Sayısı
32	Mesleki Kurs Sayısı
33	Başlayan kursiyer sayısı
34	Bitiren kursiyer sayısı
35	Kurs Öğretmeni Sayısı
36	Kurs Sınıfı Sayısı
<i>Sağlık Göstergeleri</i>	
37	Bebek Ölüm Hızı
38	Onbin Kişiye Düşen Hekim Sayısı
39	Onbin Kişiye Düşen Diş Hekimi Sayısı
40	Onbin Kişiye Düşen Eczane Sayısı
41	Onbin Kişiye Düşen Hastane Yatağı Sayısı
<i>Kültür ve Spor Göstergeleri</i>	
42	İllere göre resmi ve özel kuruluşlara ait spor tesisleri sayıları
43	İllere göre federasyonlara bağlı lisanslı ve faal sporcu sayıları,2002-2003 sezonu
44	Faal antrenörlerin sayısı, 2002
45	Amatör sporcuların sayısı
46	Müze sayısı, 2006
47	Müzeleri Ziyaretçi Sayısı, 2006
48	Matbaa sayıları 2006
49	Sinema salonu sayısı
50	Sinemalarda gösterilen film sayısı 2006

Ek 2. Temel Bileşenler Analizi SPSS çıktıları

Tablo 2. Bartlett Testi SPSS çıktısı

KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		,851
Bartlett's Test of Sphericity	Approx. Chi-Square	4956,726
	df	1035
	Sig.	,000

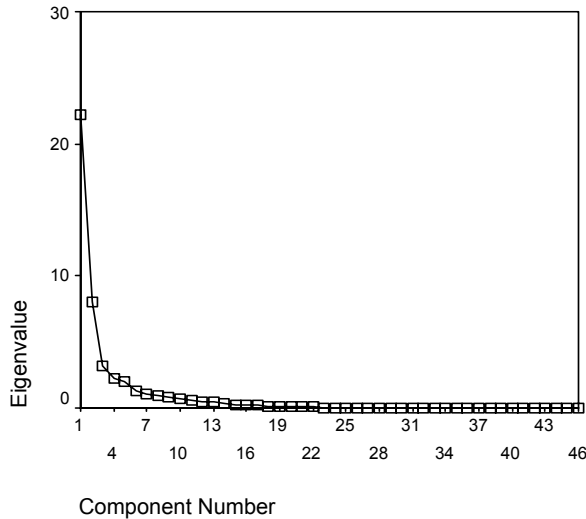
Tablo 3. Açıklanan toplam varyans oranı

Total Variance Explained

Component	Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %
1	22,170	48,196	48,196
2	8,068	17,539	65,735
3	3,236	7,034	72,769
4	2,269	4,932	77,701
5	2,021	4,393	82,093
6	1,333	2,898	84,992
7	1,063	2,311	87,303

Extraction Method: Principal Component Analysis.

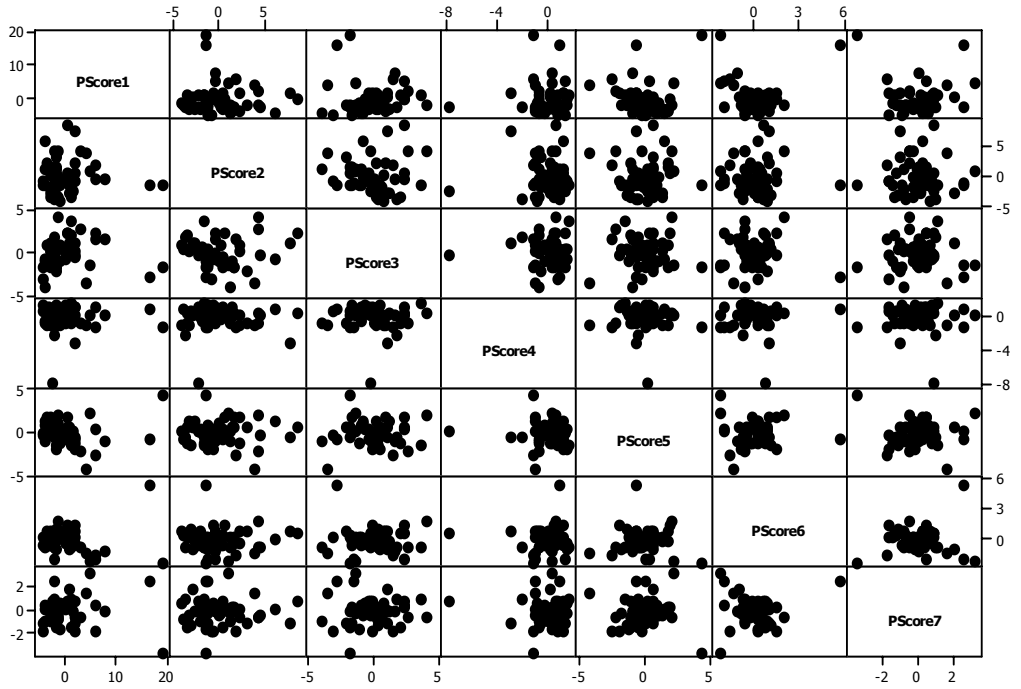
Scree Plot



Şekil 1. Çizgi Grafiği

Tablo 4. İllere göre elde edilen değişkenlere ait temel bileşen katsayıları

Değişkenler	PC1	PC2	PC3	PC4	PC5	PC6	PC7
SosEkoGelSira	-0,130	0,201	-0,107	-0,015	0,288	0,153	-0,030
GelisDerBol	-0,136	0,195	-0,053	-0,003	0,294	0,113	-0,006
GelisEndeks	0,182	-0,156	0,027	0,017	-0,118	-0,061	-0,053
TopNufus	0,210	0,065	-0,054	0,011	0,064	-0,015	0,010
SehOran	0,175	0,007	0,115	-0,075	-0,131	0,024	-0,135
RezOzSporSay	0,111	-0,072	-0,111	0,068	-0,135	0,509	0,170
LisSporSay	0,172	-0,044	-0,027	0,005	-0,203	0,275	0,025
AntrenorSay	0,074	0,004	-0,227	-0,038	-0,300	0,069	-0,085
AmatorSporSay	0,198	-0,052	-0,046	0,098	-0,057	0,065	0,083
MuzSay	0,092	0,011	-0,173	-0,100	-0,148	-0,189	0,394
MuzZiySay	0,107	-0,010	-0,111	-0,052	0,273	-0,323	0,204
MatbaaSay	0,206	-0,022	-0,087	-0,002	0,124	0,034	-0,079
SinSalSay	0,199	-0,027	-0,054	-0,019	0,144	-0,062	-0,147
GosFilmSay	0,198	-0,027	-0,055	-0,011	0,136	0,018	-0,115
İlkOkulBasOgr	0,179	-0,017	0,155	-0,010	-0,203	-0,153	-0,032
İlkOkulOgrtmn	0,064	0,241	0,300	0,031	-0,112	0,020	-0,018
BasDers							
İlkOkulDersOgr	0,107	0,236	0,245	-0,039	-0,069	0,055	-0,049
Say							
OrtaOkulBasOgr	0,113	0,212	0,231	-0,027	-0,156	0,024	-0,075
OrtaOkulOgrtmn	0,095	0,238	0,258	-0,033	-0,038	0,089	-0,080
BasOgr							
OrtaOkulOgrSay	0,029	0,253	0,305	0,004	-0,009	0,127	-0,028
AnaOkulSay	0,202	0,077	-0,060	0,008	0,088	-0,003	0,064
AnaOgrSay	0,207	0,057	-0,053	-0,026	0,107	-0,040	-0,050
AnaOgrtmnSay	0,210	-0,001	-0,069	-0,006	0,124	-0,057	-0,061
AnaDersSay	0,213	0,035	-0,041	0,017	0,059	-0,047	0,021
İlkOkulSay	0,126	0,227	0,012	0,006	0,103	0,092	0,035
İlkSubeSay	0,191	0,145	-0,002	-0,011	0,054	0,018	-0,018
OrtaOkulSay	0,210	-0,011	-0,101	0,013	0,087	-0,031	-0,025
OrtaSubSay	0,214	0,013	-0,067	0,007	0,040	-0,029	-0,039
İlkOkullasOran	0,104	-0,144	0,035	-0,086	-0,345	-0,150	-0,023
OrtaOkullasOran	0,044	-0,295	-0,019	-0,008	-0,196	-0,098	-0,054
MesKursSay	0,209	-0,008	-0,079	-0,012	0,115	-0,051	-0,104
MesKurBasSay	0,214	0,003	-0,072	0,002	0,031	0,037	-0,009
MesKurBitSay	0,212	0,011	-0,076	0,000	0,045	0,054	-0,008
KursOgrtmn	0,211	-0,014	-0,060	-0,009	0,088	-0,074	-0,108
KursSinifSay	0,117	0,009	-0,172	0,111	-0,109	0,366	0,401
NufArtHiz	0,118	0,121	0,199	0,059	0,061	-0,116	0,390
DogurHiz	-0,026	0,292	-0,146	-0,083	-0,127	-0,061	0,023
BebekOlumHiz	-0,058	0,148	-0,237	-0,171	-0,050	0,048	-0,197
HekimSay	0,034	-0,030	0,055	-0,582	0,004	0,147	0,111
DisHekSay	0,056	0,021	0,097	-0,009	-0,142	-0,242	-0,164
EczaciSay	0,038	-0,097	0,091	-0,489	0,129	0,064	-0,041
HasYatSay	0,044	-0,175	0,039	0,028	0,059	0,296	-0,373
OrthanBuy	-0,038	0,268	-0,214	-0,027	-0,123	-0,061	0,012
GocHiz	0,087	-0,128	0,329	0,034	0,112	-0,042	0,308
OkurYazNufOrani	-0,029	0,290	-0,203	-0,029	-0,186	-0,100	-0,041
OkurYazKadNuf	-0,028	0,267	-0,214	-0,062	-0,167	-0,119	-0,008
UnivBitOnulBit	-0,040	-0,062	0,001	-0,560	0,003	0,031	0,104
Orani							



Şekil 2. Yedi temel bileşen skorları için serpmeye diyagramı

Ek 3. Faktör Analizi SPSS Çıktıları

Tablo 5. Rotasyon yapılmış faktör yükleri ve oransal varyanslar

Rotated Factor Loadings and Communalities Varimax Rotation								
Değişkenler	Faktör 1	Faktör 2	Faktör 3	Faktör 4	Faktör 5	Faktör 6	Faktör 7	Oransal Varyans
Anasınıfı öğretmen sayısı	0,977	-0,099	0,110	0,074	0,010	0,049	0,010	0,932
Toplam Nüfus	0,976	-0,046	0,162	0,056	-0,018	0,076	-0,010	0,913
Ortaöğretim Şube Sayısı	0,975	-0,091	0,122	0,114	-0,005	0,079	0,015	0,959
Sinema salonu sayısı	0,974	-0,050	0,060	-0,009	0,029	-0,038	0,011	0,994
Başlayan kursiyer sayısı	0,973	-0,089	0,135	0,106	0,003	0,089	0,033	0,792
Mesleki Kurs Sayısı	0,971	-0,106	0,103	0,084	0,016	0,072	0,032	0,656
Bitiren kursiyer sayısı	0,971	-0,077	0,144	0,098	-0,003	0,085	0,011	0,859
Anasınıfı derslik sayısı	0,971	-0,082	0,164	0,093	-0,020	0,065	0,010	0,607
Ortaöğretim Okul Sayısı	0,964	-0,129	0,093	0,113	-0,004	0,116	0,016	0,959
Gösterilen film sayısı	0,963	-0,065	0,074	-0,006	0,034	0,019	0,011	0,608
Amatör sporcuların sayısı	0,957	-0,119	0,052	0,108	-0,051	-0,007	0,010	0,918
Anasınıfı öğrenci sayısı	0,954	-0,055	0,202	0,099	-0,002	0,091	-0,038	0,972
Kurs Öğretmeni Sayısı	0,951	-0,139	0,122	0,126	0,017	0,104	0,063	0,958
Anasınıfı okul sayısı	0,936	-0,061	0,209	0,090	-0,031	0,154	-0,042	0,913
İlköğretim Şube Sayısı	0,927	0,030	0,298	0,036	-0,028	0,146	-0,041	0,860
Müze Ziyaretçi sayısı	0,914	0,037	-0,037	-0,103	-0,004	-0,218	-0,049	0,929
Kurs Sınıfı Sayısı	0,913	0,031	0,000	0,026	-0,055	-0,017	-0,078	0,891
Federasyon lisanslı sporsay	0,820	-0,152	0,119	0,234	0,036	0,299	0,208	0,933
İlkokul başına öğrenci	0,768	-0,212	0,320	0,371	-0,032	-0,152	0,064	0,865
Gelişmişlik endeksi	0,740	-0,446	-0,011	0,434	-0,012	0,070	0,090	0,970
Müzelerin sayısı	0,702	0,123	-0,021	0,184	0,028	-0,119	0,143	0,977
İlköğretim Okul Sayısı	0,605	0,166	0,481	-0,081	-0,083	0,390	-0,117	0,984
Şehirleşme Oranı	0,574	-0,252	0,360	0,412	0,077	0,157	0,034	0,993
Okur-Yazar Nüfus Oranı	-0,032	0,894	0,333	-0,026	-0,019	-0,139	0,021	0,825
Okyazkadtöpkadın nüforan	-0,026	0,883	0,259	0,020	0,013	-0,195	0,028	0,986
Ort. Hane Halkı Büyüklüğü	-0,116	0,821	0,330	-0,170	-0,060	-0,183	0,026	0,984
Yıllık Net Göç Hızı	0,168	-0,774	0,339	0,024	0,083	-0,143	0,094	0,995
Bebek Olum Hız	-0,026	0,752	0,032	-0,139	-0,006	0,098	0,001	0,731
Doğurganlık Hızı	-0,104	0,750	0,476	-0,131	-0,046	-0,190	0,072	0,879
Orta Okullaşma Oranı	0,089	-0,618	-0,466	0,530	0,069	-0,013	0,067	0,981
İlk Okul DersBaşöğr. sayı	0,318	0,215	0,879	-0,007	-0,055	0,025	0,011	0,994
Orta ÖğretmenBaşöğr. sayı	0,201	0,315	0,864	-0,067	-0,026	0,025	0,055	0,990
İlk ÖğretmenBaş ögr. sayı	0,222	0,283	0,842	-0,125	-0,140	-0,066	0,031	0,969
Orta OkulBaş öğrenci sayısı	0,318	0,176	0,804	0,214	-0,009	0,103	-0,005	0,859
Ortalama Nüfus Artış Hızı	0,221	-0,138	0,755	-0,028	-0,009	-0,148	0,112	0,853
Orta DersBaş öğrenci sayısı	-0,008	0,439	0,722	-0,287	-0,047	-0,135	0,115	0,894
İlk Okullaşma Oranı	0,241	-0,398	-0,143	0,726	-0,103	0,080	-0,052	0,623
Gelişmişlik Derece Bölgesi	-0,389	0,572	0,071	-0,634	0,066	-0,110	-0,097	0,857
Soseko gelişmişlik sıra	-0,356	0,618	0,062	-0,632	0,044	-0,068	-0,132	0,432
Faal antrenörlerin sayısı	0,107	0,063	-0,063	0,591	-0,010	0,445	-0,109	0,760
Onbine düşen hekim sayısı	-0,009	-0,072	0,049	0,039	0,899	0,112	-0,030	0,785
Onbine düşen eczane sayısı	0,064	-0,078	-0,057	0,029	0,827	0,032	-0,035	0,865
ÜniBitOkulBitirenlereOran	-0,080	0,061	-0,196	-0,168	0,807	-0,133	0,074	0,814
OnbineDushast. yatak sayı	0,057	-0,468	-0,240	0,055	0,033	0,638	-0,055	0,961
ResÖzel Spor tesis sayıları	0,195	-0,137	0,002	0,233	0,026	0,612	0,148	0,869
OnbineDus dış hek. sayısı	0,048	0,006	0,173	-0,013	-0,006	0,060	0,931	0,810
Anasınıfı öğretmen sayısı	0,977	-0,099	0,110	0,074	0,010	0,049	0,010	0,932

Tablo 6. Faktör skorları en düşük ve en yüksek olan 3 il sıralaması

Faktör Skorları En Düşük olan 3 il													
	Faktör 1		Faktör 2		Faktör 3		Faktör 4		Faktör 5		Faktör 6		Faktör 7
HAKKARI	-0,54	TEKIRDAG	-2,05	TUNCELI	-2,04	G.HANE	-2,30	BARTIN	-1,35	KARABUK	-1,93	DIYARBAKIR	-1,41
BATMAN	-0,49	ANTALYA	-1,78	SINOP	-1,93	MARDIN	-2,13	AGRI	-1,12	YALOVA	-1,86	SANLIURFA	-1,28
BILECIK	-0,49	MUGLA	-1,66	ARDAHAN	-1,82	AGRI	-1,78	BITLIS	-1,03	ISTANBUL	-1,80	MARDIN	-1,08
Faktör Skorları En Yüksek olan 3 il													
	Faktör 1		Faktör 2		Faktör 3		Faktör 4		Faktör 5		Faktör 6		Faktör 7
IZMIR	1,95	ARDAHAN	2,11	DIYARBAKIR	2,28	GANTEP	1,57	GIRESUN	3,27	SIVAS	1,92	ANTALYA	2,46
ANKARA	2,59	SIIRT	2,26	BATMAN	2,28	KOCAELİ	2,03	G.HANE	3,86	SAMSUN	1,98	HAKKARI	3,75
ISTANBUL	7,84	MERSIN	2,33	SANLIURFA	2,70	MERSIN	3,92	AMASYA	4,57	IZMIR	4,15	ADANA	6,42

Ek 4. Ayrıştırma Analizi SPSS Çıktıları

Tablo 7. Box's M testi sonucu

Box's M	259,192
F Approx.	10,198
df1	15
df2	289,735
Sig	,000

Tablo 8. Öncül olasılıklar

Grup	Prior	Cases Used in Analysis	
		Unweighted	Weighted
		1,00	,716
2,00	,074	6	6,000
3,00	,049	4	4,000
4,00	,062	5	5,000
5,00	,037	3	3,000
6,00	,025	2	2,000
7,00	,037	3	3,000
Total	1,000	81	81,000

Tablo 9. Grup Ortalamaları, Ortalama Farkları, Wilks' Lambda Değerleri ve Mahalanobis Uzaklıkları

Bağımsız Değişkenler	Bağımlı Değişken Grup ortalamaları							Gruplararası ortalama farkı			Minimum Mahalanobis D ²	
	Grup1 (n=58)	Grup2 (n=6)	Grup3 (n=4)	Grup4 (n=5)	Grup5 (n=3)	Grup6 (n=2)	Grup7 (n=3)	Wilks' Lambda	F Değeri	Önem Seviyesi	Minimum D ²	Gruplararası
Sosyo-kültürel göstergeler	-.132	-.221	-.226	-.332	-.238	2.274	2.579	,593	8,456	,000	0.012	2 ve 3
Demografik Göstergeler	.059	-.499	-.245	-.530	1.047	-.962	.655	,878	1,719	,128	1.438	1 ve 2
Eğitim Göstergeleri	-.242	-.254	1.880	-.395	2.273	.583	.692	,545	10,277	,000	0.000	1 ve 2
Gelişmişlik Göstergeleri	-.046	.1887	-1.546	.893	1.086	.550	-.367	,769	3,696	,003	0.045	4 ve 5
Sağlık Göstergeleri	.015	-.599	-.186	-.107	.974	.793	-.168	,918	1,101	,370	0.000	3 ve 7
Tesis/Yatırım Göstergeleri	-.041	.572	.660	-1.51	-.002	2.580	-.404	,636	7,059	,000	0.000	3 ve 5
Diş Hekimi Göstergeleri	-.137	.456	-.894	.272	-.893	.257	3.209	,508	11,941	,000	0.000	3 ve 5

Tablo 10. Aşamalı değişken seçme yöntemi SPSS çıktısı

Variables in the Analysis					
Step		Tolerance	F to Remove	Min. D Squared	Between Groups
1	Tesis/Yatırım Gostergeleri	1,000	7,059		
2	Tesis/Yatırım Gostergeleri	,987	7,206	,000	3,00 and 5,00
	Dis Hekimi Gostergeleri	,987	12,083	,002	1,00 and 5,00
3	Tesis/Yatırım Gostergeleri	,972	7,416	,142	4,00 and 6,00
	Dis Hekimi Gostergeleri	,986	11,963	,282	1,00 and 7,00
	Gelismislik Gostergeleri	,983	3,862	,649	3,00 and 5,00
4	Tesis/Yatırım Gostergeleri	,942	7,928	,792	2,00 and 4,00
	Dis Hekimi Gostergeleri	,609	26,420	,578	1,00 and 2,00
	Gelismislik Gostergeleri	,983	3,818	,675	3,00 and 5,00
	Sosyo-kültürel Gostergeler	,590	21,953	1,112	1,00 and 2,00
5	Tesis/Yatırım Gostergeleri	,873	9,351	,876	2,00 and 4,00
	Dis Hekimi Gostergeleri	,609	26,067	,595	1,00 and 2,00
	Gelismislik Gostergeleri	,959	4,144	,778	3,00 and 5,00
	Sosyo-kültürel Gostergeler	,574	22,595	1,112	1,00 and 2,00
	Egitim Gostergeleri	,889	12,398	1,495	1,00 and 2,00

Tablo 11. Değişken seçiminde Wilks' Lambda değerleri

Wilks' Lambda						
Step	Number of Variables	Lambda	Exact F		Approximate F	
			Statistic	Sig.	Statistic	Sig.
1	1	,508	11,941	,000		
2	2	,184	16,223	,000		
3	3	,097			14,498	,000
4	4	,056			13,378	,000
5	5	,041			11,491	,000

Tablo 12. Ayrıştırma analizi uyum iyiliği test sonuçları

Wilks' Lambda				
Test of Function(s)	Wilks' Lambda	Chi-square	df	Sig.
1 through 5	,041	236,386	30	,000
2 through 5	,206	116,838	20	,000
3 through 5	,551	44,128	12	,000
4 through 5	,775	18,822	6	,004
5	,966	2,562	2	,278

Tablo 13. Kanonik fonksiyon Wilks' Lambda değerleri

Eigenvalues				
Function	Eigenvalue	% of Variance	Cumulative %	Canonical Correlation
1	4,031 ^a	63,1	63,1	,895
2	1,671 ^a	26,2	89,2	,791
3	,408 ^a	6,4	95,6	,538
4	,246 ^a	3,8	99,4	,444
5	,035 ^a	,6	100,0	,184

a. First 5 canonical discriminant functions were used in the analysis.

Tablo 14. Standartlaştırılmış kanonik ayrıştırma fonksiyon katsayıları

Standardized Canonical Discriminant Function Coefficients

	Function				
	1	2	3	4	5
Sosyo-kültürel Gostergeler	1,182	,151	-,062	,128	-,549
Eğitim Gostergeleri	,175	,849	,599	,026	,123
Gelişmişlik Gostergeleri	-,052	-,384	,219	,914	,096
Tesis/Yatırım Gostergeleri	,231	,725	-,652	,248	,283
Dis Hekimi Gostergeleri	1,134	-,386	,082	-,131	,429

Tablo 15. Standartlaştırılmamış kanonik ayrıştırma fonksiyon katsayıları

Canonical Discriminant Function Coefficients

	Function				
	1	2	3	4	5
Sosyo-kültürel Gostergeler	1,476	,189	-,078	,160	-,685
Eğitim Gostergeleri	,228	1,105	,779	,034	,160
Gelişmişlik Gostergeleri	-,057	-,421	,240	1,002	,105
Tesis/Yatırım Gostergeleri	,279	,875	-,786	,299	,341
Dis Hekimi Gostergeleri	1,530	-,521	,111	-,177	,579
(Constant)	,000	,000	,000	,000	,000

Unstandardized coefficients

Tablo 16. Ayrıştırma fonksiyonları grup ortalamaları (Centroid)

Functions at Group Centroids

Grup	Function				
	1	2	3	4	5
1,00	-,470	-,239	-,172	-,064	-,047
2,00	,462	-,140	-,534	,235	,590
3,00	-1,003	3,731	,493	-1,166	,002
4,00	-,638	-2,344	1,155	,327	-,101
5,00	-1,265	2,474	1,954	1,285	,123
6,00	4,572	2,965	-1,590	1,663	-,379
7,00	8,784	-,620	,926	-,621	,026

Unstandardized canonical discriminant functions evaluated at group means

Tablo 17. Ayrıştırma fonksiyonları ağırlıkları

Classification Function Coefficients

	Grup						
	1,00	2,00	3,00	4,00	5,00	6,00	7,00
Sosyo-kültürel Gostergeler	-,704	,331	-1,004	-1,352	-1,431	7,959	12,662
Eğitim Gostergeleri	-,515	-,363	4,240	-1,841	4,033	3,074	2,018
Gelişmişlik Gostergeleri	,017	,202	-2,563	1,617	,802	-,261	-,633
Tesis/Yatırım Gostergeleri	-,239	,698	2,248	-3,073	,701	5,485	1,000
Dis Hekimi Gostergeleri	-,630	1,021	-3,217	,259	-3,163	4,758	13,991
(Constant)	-,491	-3,064	-11,273	-6,462	-9,899	-21,263	-42,692

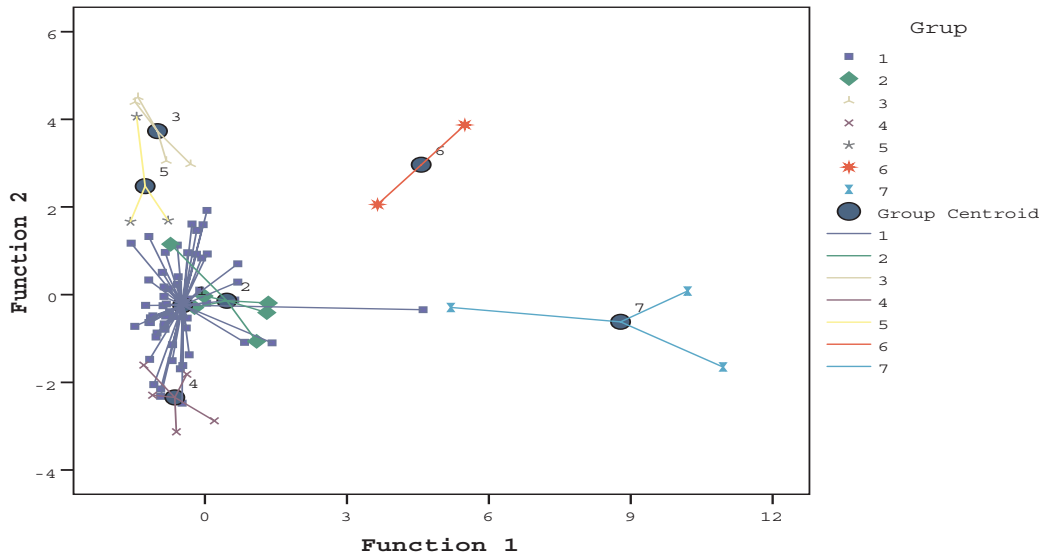
Fisher's linear discriminant functions

Tablo 18. Çapraz geçerlilik yöntemiyle sınıflandırma sonuçları

		Predicted Group Membership							Total
Original	Grup	1,00	2,00	3,00	4,00	5,00	6,00	7,00	
Count	1,00	55	0	0	0	1	1	1	58
	2,00	0	6	0	0	0	0	0	6
	3,00	0	0	4	0	0	0	0	4
	4,00	0	0	0	5	0	0	0	5
	5,00	0	0	0	0	3	0	0	3
	6,00	0	0	0	0	0	2	0	2
	7,00	0	0	0	0	0	1	2	3
%	1,00	94,8	,0	,0	,0	1,7	1,7	1,7	100,0
	2,00	,0	100,0	,0	,0	,0	,0	,0	100,0
	3,00	,0	,0	100,0	,0	,0	,0	,0	100,0
	4,00	,0	,0	,0	100,0	,0	,0	,0	100,0
	5,00	,0	,0	,0	,0	100,0	,0	,0	100,0
	6,00	,0	,0	,0	,0	,0	100,0	,0	100,0
	7,00	,0	,0	,0	,0	,0	33,3	66,7	100,0

a. 95,1% of original grouped cases correctly classified.

Canonical Discriminant Functions



Şekil 3. Kanonik ayırma fonksiyonları - Grup ortalamaları ve uzaklıkları

DANIŐMA KURULU ÜYELERİ - ADVISORY BOARD MEMBERS

Ahmet KARA	Fatih Üniversitesi
Ali YAZICI	Atılım Üniversitesi
Alper GÜVEL	Çukurova Üniversitesi
Asaf Savaş AKAT	Bilgi Üniversitesi
Aşır GENÇ	Selçuk Üniversitesi
Aydın ÖZTÜRK	Ege Üniversitesi
Ayşe GÜNDÜZ HOŐGÖR	Orta Dođu Teknik Üniversitesi
Bedriye SARAÇOĐLU	Gazi Üniversitesi
Ceyhan İNAL	Hacettepe Üniversitesi
Coşkun Can AKTAN	Dokuz Eylül Üniversitesi
Deniz GÖKÇE	Bahçeşehir Üniversitesi
Ekrem ERDEM	Erciyes Üniversitesi
Ercan UYGUR	Ankara Üniversitesi
Erdem BAŐCI	T.C. Merkez Bankası
Erinç YELDAN	Bilkent Üniversitesi
Erol TAYMAZ	Orta Dođu Teknik Üniversitesi
Eser KARAKAŐ	Bahçeşehir Üniversitesi
Fatih ÖZATAY	TOBB Ekonomi ve Teknoloji Üniversitesi
Fatin SEZGİN	Bilkent Üniversitesi
Fikri AKDENİZ	Çukurova Üniversitesi
Fikri ÖZTÜRK	Ankara Üniversitesi
Gülay BAŐARIR KIROĐLU	Mimar Sinan Güzel Sanatlar Üniversitesi
Güven SAK	TOBB Ekonomi ve Teknoloji Üniversitesi
Haluk LEVENT	Galatasaray Üniversitesi
Hamza EROL	Çukurova Üniversitesi
İbrahim DALMIŐ	Kırıkkale Üniversitesi
İlhan TEKELİ	Orta Dođu Teknik Üniversitesi
İmdat KARA	Başkent Üniversitesi
İnsan TUNALI	Koç Üniversitesi
Levent KANDİLLER	Çankaya Üniversitesi
Mehmet KAYTAZ	Işık Üniversitesi
Meltem DAYIOĐLU TAYFUR	Orta Dođu Teknik Üniversitesi
Metin TOPRAK	BDDK
Mustafa ACAR	Kırıkkale Üniversitesi
Mustafa AYTAÇ	Uludağ Üniversitesi
Nihat BOZDAĐ	Gazi Üniversitesi
Onur BASKAN	Ege Üniversitesi
Orhan GÜVENEN	Bilkent Üniversitesi
Ömer Faruk ÇOLAK	Gazi Üniversitesi
Ömer L. GEBİZLİOĐLU	Ankara Üniversitesi
Özkan ÜNVER	Ufuk Üniversitesi
Öztaş AYHAN	Orta Dođu Teknik Üniversitesi
Reşat KASAP	Gazi Üniversitesi
Savaş ALPAY	SESRTCIC
Seyfettin GÜRSOY	Galatasaray Üniversitesi
Süleyman GÜNAY	Hacettepe Üniversitesi
Turan EROL	THY Yönetim Kurulu
Ümit OKTAY FIRAT	Marmara Üniversitesi
Yasin AKTAY	Selçuk Üniversitesi
Yılmaz AKDİ	Ankara Üniversitesi
Yusuf Ziya ÖZCAN	YÖK