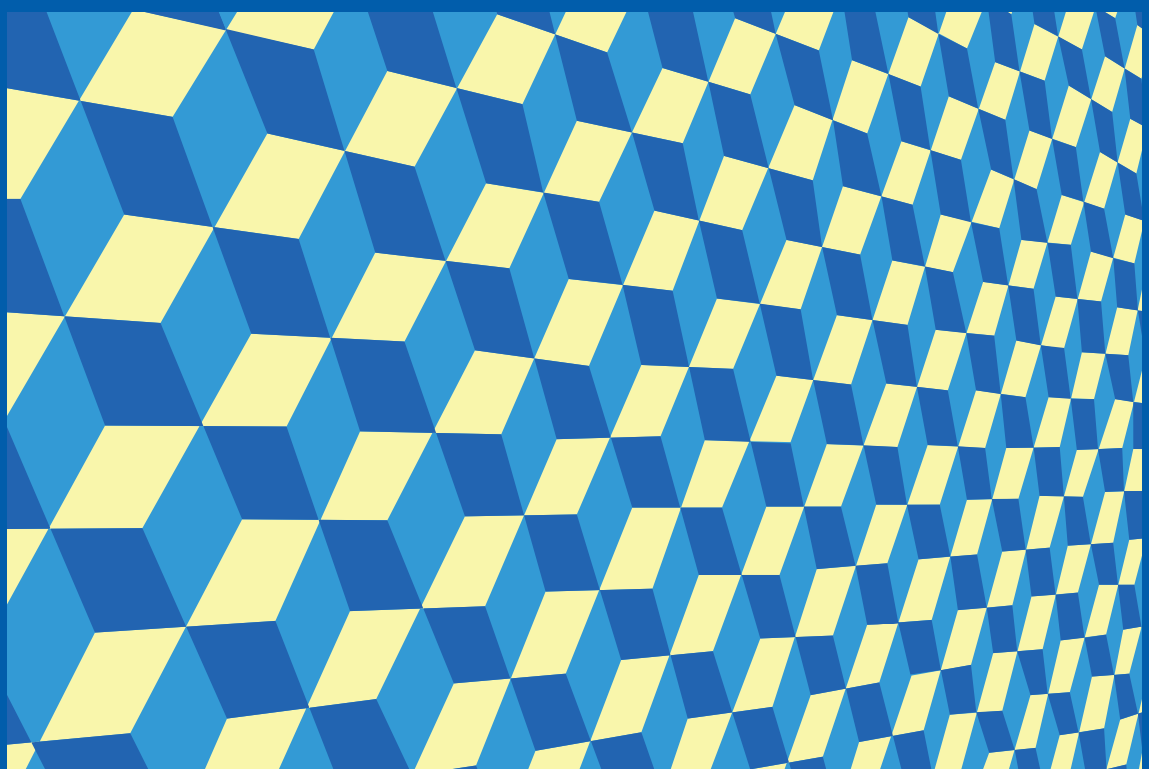
İSTATİSTİK ARAŞTIRMA DERGİSİ

Journal of Statistical Research

Cilt-Volume: 06 Sayı-Number: 01
Temmuz-July 2008

ISSN 1303-6319



TÜRKİYE İSTATİSTİK KURUMU
Turkish Statistical Institute

İSTATİSTİK ARAŞTIRMA DERGİSİ

Journal of Statistical Research

Cilt-Volume: 06 Sayı-Number: 01
Temmuz-July 2008

TÜRKİYE İSTATİSTİK KURUMU
Turkish Statistical Institute

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 75 - 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
3326

ISSN
1303-6319

Türkiye İstatistik Kurumu Turkish Statistical Institute

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

Türkiye İstatistik Kurumu Matbaası, Ankara Turkish Statistical Institute, Printing Division, Ankara
Tel: 0312 410 01 64 * Fax: 0312 418 50 82
Ağustos 2009 August 2009
MTB: 2009-0635 - 475 **Adet**-Copies

İstatistik Araştırma Dergisinin Değerli Okuyucuları,

Hakemlik sürecinin çok yavaş işlemesi nedeniyle yılda bir sayıya indirilmiş olan İstatistik Araştırma Dergisi'nin altıncı cildi yayımlanmış bulunmaktadır. Dergi'mizin bu cildinde yayımlanmak üzere gönderilmiş olan 31 adet makale taslağından, hakem sürecine girmiş olanların sayısı 16, sürece sokulmayanların sayısı 4, hakem süreci sonucunda basımı ret edilmişlerin sayısı 2, kabul edilmiş olanların sayısı ise 9'dur. Bu istatistiklerden de görüleceği gibi, oldukça çok sayıda makale taslağı hakem sürecinde beklemektedir. Bunun bir çok nedeni vardır. Başlıca nedenler arasında; hakemlerin değerlendirme sürecinde yurt dışına çıkmaları, bir şekilde adres değiştirmeleri, Dergi Editörlüğü'nde bu durumda olan hakemlerin güncel adres bilgilerinin mevcut olmaması ve sağlık nedenleri sıralanabilir. Sözlü ve yazılı hatırlatmalara rağmen değerlendirmelerini zamanında gönderemeyen hakemlerin yerine yeni hakemler belirlenmekte ise de, bazı durumlarda bunun da tek çözüm olmadığı Editörlüğümüzce gözlenmektedir. Dolayısıyla, hemen her sayının önsözünde dile getirildiği biçimde, makale inceleme sürecindeki gecikmelerin önüne geçilmesinde zorlanılmaktadır.

Hakemlik sürecindeki bu istenmeyen duruma rağmen, ileriki yıllarda daha iyiye doğru yol alacağımıza inancımızı korumaya çalışıyoruz. Dergi'mizin Journal Lists of Current Index to Statistics and Mathematical Reviews gibi bazı uluslararası endekslerce taranması için çalışmalarımızı sürdürüyoruz. Umarız yakın zamanda bununla ilgili olumlu haberleri de sizlere ulaştırabilme olanağına sahip olabiliriz.

Bu sayının önsözüne yetiştirilemeyen, ilk sayısından en son sayısına kadar Dergi'mizde yayımlanmış makalelere ait tüm ayrıntılı istatistikleri içeren bilgileri ve bunların genel analizlerini gelecek sayımızda bulacak ve bunlarla ilgili görüşlerinize de daha sonraki sayılarımızda yer vereceğiz.

Bu ve bundan sonraki sayılarımızda makaleler, niteliklerine göre sınıflandırılarak yayımlanacaktır. Makalelerin sınıflandırılmasında Dergi Kılavuzu'nda da belirtildiği üzere 6 grupta esas alınmıştır. Bunlar; özgün araştırma makaleleri, gözden geçirme makaleleri, teknik notlar, eleştirel derleme makaleleri, tartışma makaleleri ve güncel çeviri makaleleridir. Bu sınıflandırmalarla ilgili ayrıntılı bilgiyi, www.tuik.gov.tr adresinde yayımlanmış olan Dergi Kılavuzunda bulabilirsiniz.

Dergi'mizin bu sayısında makalelerin bilimsel yönden değerlendirilmesinde büyük özveriyle katkı sağlamış olan tüm hakemlere minnet ve şükranlarımı sunmayı bir borç bilirim. Dergi'nin her aşamasında vermiş olduğu destek ve katkılar için TÜİK Başkan Vekili Sayın A. Ömer TOPRAK'a, Dergi'nin basım sürecinin her aşamasında sağlamış olduğu katkılarından dolayı Editör Yardımcıları Sayın TÜİK Uzmanı Sevil UYGUR'a, Sayın Dr. Özlem İLK'e ve ayrıca, emeği geçen tüm TÜİK çalışanlarına içtenlikle teşekkür ederim.

Diğer sayılarda buluşmak dileğiyle saygılar sunarım.

Prof. Dr. Fetih YILDIRIM
Dergi Editörü

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
Fetih YILDIRIM Fetih YILDIRIM

Editör Yardımcısı Assistant Editor
Sevil UYGUR Sevil UYGUR

Editör Yardımcısı Assistant Editor
Özlem İLK Özlem İLK

Amaç ve Kapsam

İstatistik Araştırma Dergisi (İAD), istatistiki 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 bir dergidir.

İAD'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; Yönetsel 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 Başkanlığına 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.
- 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 endeksli olarak Arap rakamları ile numaralandırılarak yazılmalıdır.
- Makale taslağı yazımında, okuyucunun, çalışmanın her aşamasını anlama ve değerlendirmesine olanak verecek bir anlatım ve plâna 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 verilerek/ 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, www.tuik.gov.tr adresinden *"İstatistik Araştırma Dergisi Kılavuzu"*na bakınız.

Aim and Scope

"*Journal of Statistical Research" (JSR)* is a refereed journal with a view to raise the quality of statistical researches, improve the statistical methodology and applications, discuss the related studies in literature, consider survey studies regarding 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 an A4 paper size; from left 3,5 cm, from right, top and bottom 2,5 cm margins should be set.
- **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 centered 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 angle the researcher/researchers have approached the problem.
- The reason of why the employed research methodology is chosen should be explained. The validity and reliability of all the data collection tools should be presented.
- Data obtained in conclusion 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 will reflect the scope of the study.
- The Turkish and English abstracts should include; the goal, methodology, scope and main findings of the study.

Note: For detailed information, please see "A Guide for Journal of Statistical Research" at www.tuik.gov.tr web site.

DERGİ'NİN BU SAYISINA BİLİMSEL KATKI SAĞLAYAN HAKEMLER-
REFEREE WHO PROVIDE SCIENTIFIC CONTRIBUTIONS FOR THIS VOLUME

1	**Prof. Dr.**	**Aydın ERAR**	Mimar Sinan Güzel Sanatlar Üniversitesi
2	**Prof. Dr.**	**Ayşen APAYDIN**	Ankara Üniversitesi
3	**Yrd. Doç. Dr.**	**Barış SÜRÜCÜ**	Orta Doğu Teknik Üniversitesi
4	**Doç. Dr.**	**Birdal ŞENOĞLU**	Ankara Üniversitesi
5	**Doç. Dr.**	**Cem KADILAR**	Hacettepe Üniversitesi
6	**Dr.**	**Cevriye AYSOY**	TC Merkez Bankası
7	**Dr.**	**Ceylan YOZGATLIGİL**	Orta Doğu Teknik Üniversitesi
8	**Doç. Dr.**	**Elvan CEYHAN**	Koç Üniversitesi
9	**Yrd. Doç. Dr.**	**Fazıl GÖKGÖZ**	Ankara Üniversitesi
10	**Prof. Dr.**	**Fetih YILDIRIM**	Çankaya Üniversitesi
11	**Yrd. Doç. Dr.**	**Funda SEZGİN**	Mimar Sinan Güzel Sanatlar Üniversitesi
12	**Doç. Dr.**	**Galip YÜKSEL**	Gazi Üniversitesi
13	**Prof. Dr.**	**Gülay BAŞARIR KIROĞLU**	Mimar Sinan Güzel Sanatlar Üniversitesi
14	**Prof. Dr.**	**Hamza GAMGAM**	Gazi Üniversitesi
15	**Prof. Dr.**	**Hüseyin TATLIDİL**	Hacettepe Üniversitesi
16	**Yrd. Doç. Dr.**	**Işıl ÜREDİ**	Mersin Üniversitesi
17	**Yrd. Doç. Dr.**	**Nevin UZGÖREN**	Dumlupınar Üniversitesi
18	**Dr.**	**Özlem İLK**	Orta Doğu Teknik Üniversitesi
19	**Dr.**	**Özlem TÜRKER BAYRAK**	Çankaya Üniversitesi
20	**Prof. Dr.**	**Sadullah SAKALLIOĞLU**	Çukurova Üniversitesi
21	**Yrd. Doç. Dr.**	**Seza DANIŞOĞLU**	Orta Doğu Teknik Üniversitesi
22	**Yrd. Doç. Dr.**	**Suat KASAP**	Hacettepe Üniversitesi
23	**Prof. Dr.**	**Süleyman GÜNAY**	Hacettepe Üniversitesi
24	**Dr.**	**Vilda PURUTÇUOĞU**	Orta Doğu Teknik Üniversitesi

Volume: 06 Number: 01 Category: 01 Page: 1-17 ISSN No: 1303-6319 Cilt: 06 Sayı: 01 Kategori: 01 Sayfa: 1-17

BAYESIAN INFERENCE OF THE COMPLEX MAPK PATHWAY UNDER THE STRUCTURAL DEPENDENCY

Vilda PURUTÇUOĞLU* **Ernst WIT****

ABSTRACT

The MAPK pathway is one of the main signal transaction system in all eukaryotes which regulates the cellular growth control. Because of its vital role, the regulation of the pathway is conducted via many proteins, thereby constitutes a complex structure. In inference of this system via MCMC techniques based on the Euler approximation, we have observed that there are many proteins which indicate high structural dependencies on other proteins and these species have caused singular diffusion matrices, hereby resulted in infeasible acceptance probabilities. Therefore, we have discarded these problematic substrates at the beginning of the inference and estimated the parameters by using merely linearly independent species in the system. However in that case, the accuracy of the estimation has been highly affected by the underlying exclusion, particularly, when the number of dependent species was big. The elimination of those proteins has led to a significant rise in the number of current missing components in MCMC. In this study, we implicitly include these proteins in our computation via an alternative approach which simulates dependent terms as a linear combination of linearly independent species. In that way, we can add the effect of dependent species in the calculation of acceptance probabilities of reaction rates and states. From the analysis, we conclude that the highlighted innovation decreases the average error of estimates and suggests less computational cost in inference of the MAPK pathway.

Keywords: Bayesian inference, Diffusion approximation, MAPK pathway.

1. INTRODUCTION

All cellular activations are regulated by various signal transduction pathways. The MAPK (mitogen-activated protein kinase) pathway is one of the main pathway structure which regulates the growth control in all eukaryotes, i.e. the organisms whose cells contain a nucleus, thereby it is the system of interest, particularly, in oncogene researches (Kolch, 2005; Orton et al., 2005).

Coming from the importance of the pathway in the cellular life cycle from the cell proliferation, i.e. the reproduction of the cell, to the apoptosis, i.e. the cell death, the activation of the MAPK pathway uses a number of proteins whose main components are Ras, Raf, MEK, and ERK proteins (Figure 1). This activation begins by an external stimulus which causes the binding of the signal to the Epidermal Growth Factor (EGF) receptor and is ended up by the production of the target c-Fos gene after a sequence of recruitments, phosphorylations, and inhibitions.

* Dr., Middle East Technical University, Faculty of Art and Science, Department of Statistics, e.mail: vpurutcu@metu.edu.tr

** Prof. Dr., University of Groningen, Institute of Mathematics and Computing Science, e.mail: e.c.wit@rug.nl

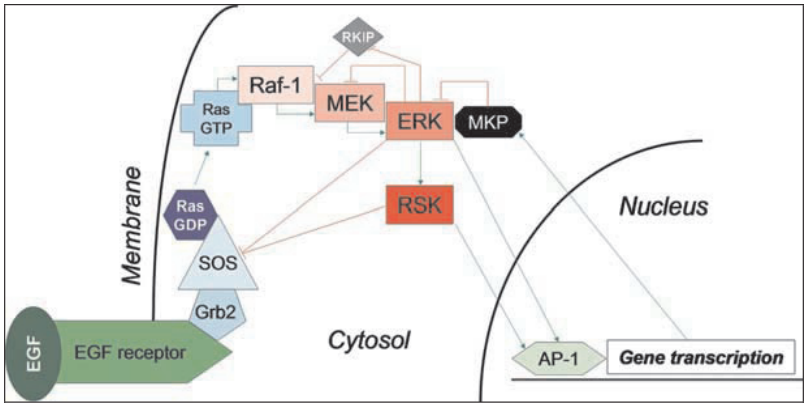


Figure 1. Main components of the MAPK pathway (Kolch, 2005)

In this study, we estimate the stochastic rate constants of quasi reactions of the MAPK pathway which is described by 51 proteins and 66 reactions (Purutçuoğlu and Wit, 2006 and 2008b). In the inference of the model parameters, i.e. reaction rates, from a simulated dataset, we implement the discretized version of the diffusion approximation known as the Euler-Maruyama approximation (Eraker, 2001; Golightly and Wilkinson, 2005).

In the estimation via the Euler technique, we overcome the problems of the missing data and sparse measurements, which are typical challenges in complex systems, by using the MCMC (Markov Chain Monte Carlo) framework. Accordingly we choose the Metropolis-within-Gibbs algorithm with the data augmentation technique (Golightly and Wilkinson, 2005; Purutçuoğlu and Wit, 2008a and 2008b) for the computation. From our previous analysis (Purutçuoğlu and Wit, 2008a and 2008b), we have seen that although the underlying MCMC methods are promising to estimate the reaction rates, the dependency between proteins causes singular diffusion matrices in implementations. Therefore, we have eliminated the proteins which lead to singularities and the algorithms have run by merely linearly independent terms. However, when the total number of excluded species became bigger, the estimation had to be conducted under a large number of missing information. In this study, to unravel the challenges caused by those large missing data, we develop an innovation in the current scheme such that the new plan uses these problematic substrates in the estimation.

We present a brief explanation about biochemical reactions and the diffusion approximation in Section 2. The details of MCMC updates and the new plan are given in Section 3.1 and Section 3.2, respectively. We evaluate the performance of the algorithm in Section 4 by comparing our outcomes with previous findings. Finally Section 5 concludes the results and discusses possible extensions.

2. BIOCHEMICAL PROCESS AND STOCHASTIC MODELLING

A biochemical reaction is a quantitative and qualitative description of a biochemical process. If we have *r* number of equations which explain a biochemical activation, this set of reactions presents a system. A simple biochemical system can be described as the following:

$$\begin{aligned}
&\text{Reaction 1: } m_{11}Y_1 + m_{12}Y_2 + \ldots + m_{1n}Y_n \xrightarrow{c_1} s_{12}Y_1 + s_{12}Y_2 + \ldots + s_{1n}Y_n \\
&\text{Reaction 2: } m_{21}Y_1 + m_{22}Y_2 + \ldots + m_{2n}Y_n \xrightarrow{c_2} s_{22}Y_1 + s_{22}Y_2 + \ldots + s_{2n}Y_n \\
&\qquad\qquad \vdots \;\; \vdots \;\; \vdots \;\; \vdots \\
&\text{Reaction } r\text{: } m_{r1}Y_1 + m_{r2}Y_2 + \ldots + m_{rn}Y_n \xrightarrow{c_r} s_{r2}Y_1 + s_{r2}Y_2 + \ldots + s_{rn}Y_n \qquad (1)
\end{aligned}$$

In that expression, $Y = (Y_1,...,Y_n)$ denotes the *n*-dimensional vector of current states of the system and *n* indicates the total number of species. The coefficients m_{ji} and s_{ji} display the stoichiometric coefficients associated with the *i*th reactant of the *j*th reaction and the *i*th product of the *j*th reaction, respectively, for $i = 1,...,n$ and $j = 1,...,r$. Finally c_j is the reaction rate constant which denotes the speed of the reaction dependent on the temperature of the system and physical properties of reactants.

Equation (1) can be also shown by a matrix form such that $MY \rightarrow SY$ where,

$$M = \begin{bmatrix} m_{11} & \cdots & m_{1n} \\ \vdots & \vdots & \vdots \\ m_{r1} & \cdots & m_{rn} \end{bmatrix} \quad \text{and} \quad S = \begin{bmatrix} s_{11} & \cdots & s_{1n} \\ \vdots & \vdots & \vdots \\ s_{r1} & \cdots & s_{rn} \end{bmatrix}$$

are the (*r*x*n*) - dimensional matrix of stoichiometries of reactants and the (*r*x*n*) - dimensional matrix of stoichiometries of products, respectively. The interpretation of this description is that when the *r*th reaction occurs, the number of molecules of Y_i $(i = 1,...,n)$ decreases by m_{ri} and increases by s_{ri} amount. As a result the molecular transfer causes a net change in the system with $V_{ri} = s_{ri} - m_{ri}$ where $V = S - M$ is called the (*r*x*n*) - dimensional matrix of net effects and V_{ji} is the corresponding net chance of the *i*th species after the execution of the *r*th reaction. More details about the formulation of biochemical processes and the network structure can be found in Wilkinson (2006) and Bower and Bolouri (2001). On the other side, the implementation of this description in a prokaryotic autoregulation gene network and in the MAPK pathway are given in Golightly and Wilkinson (2005) and Purutçuoğlu and Wit (2008b), respectively.

There are several approaches in order to capture the stochastic behaviour of the biochemical system (Gillespie, 1977; Gibson and Bruck, 2000; Turner et al., 2004). The Gillespie algorithm (Gillespie, 1977; Gillespie, 1992) is the most common exact method to simulate a biochemical network, whereas, it is computationally inefficient in inference of the realistic complexity (Golightly and Wilkinson, 2005; Wilkinson, 2006; Boys et al., 2008). The diffusion approximation is an efficient technique as an alternative estimation in place of Gillespie (Golightly and Wilkinson, 2005). In this research, we use the discretized version of the diffusion approximation, known as the Euler-Maruyama approximation, since the observed measurements are collected in discrete time. The Euler method explaines the change of states at time *t* by the following equation.

$$\Delta Y_t = \mu(Y_t, \theta)\Delta t + \beta^{\frac{1}{2}}(Y_t, \theta)\Delta W_t \qquad (2)$$

here ΔY_t stands for the change in state $Y = (Y_1, Y_2 ..., Y_n)$ at time t to $[t + \Delta t]$. $\theta = (c_1, c_2, ..., c_r)$ represents the parameter vector while n and r are the total number of substrates and the total number of reactions in the system, respectively, as mentioned beforehand. $\mu(Y_t, \theta)$ displays an n-dimensional mean or drift vector and is computed by $\mu(Y_t, \theta) = V'h(Y_t, \theta)$. On the other hand, $\beta(Y_t, \theta)$ shows an (nxn) diffusion or variance matrix and is found via $\beta(Y_t, \theta) = V'diag\{h(Y_t, \theta)\}V$. Both μ and β terms are the functions of Y and θ, and are calculated from the hazard $h(Y_t, \theta)$ as well as the net effect matrix V in which V' implies the transpose of V. $diag\{h(Y_t, \theta)\}$ in β is an (rxr) dimensional matrix whose diagonal terms set to $h(Y_t, \theta)$ and off-diagonals are equated to zero (Wilkinson, 2006). Finally ΔW_t denotes an n-dimensional independent identically distributed Brownian random vector generated from the normal distribution with mean zero and covariance-variance as the product of the identity matrix I and the discrete time interval Δt, i.e. $\Delta W_t \sim N(0, I\Delta t)$.

3. INFERENCE OF THE SYSTEM

In the inference of the reaction rates, we consider that the observation matrix Y is composed of both observed and unobserved measurements as used in the studies of Eraker (2001); Golightly and Wilkinson (2005). We denote observed and unobserved terms by n-dimensional X and Z vectors, respectively. Moreover in order to get more precise estimates from the Euler, we use the data augmentation by putting latent states within each pair of time-course measurements. More details about the implementation of the data augmentation can be found in Roberts and Stramer (2001) and Elerian et al. (2001). So every time state of the system Y_i ($i = 1,...,T$), where $i = 1$ indicates the initial time point and $i = T$ is the final time point after the data augmentation, is presented as $Y_i \equiv (X_i, Z_i)'$. Here $(A)'$ stands for the transpose of any vector (A). If the state has observed measurements, then X_i is set to x_i, which means the observed data by observed components.

In the update of the system via MCMC techniques we implement the Gibbs sampling seeing that the number of unobservable values, i.e. the number of reaction rates and missing data, are large. However as the dimension of the system for every time point is high and each state Y_i is updated via a different Gibbs sampler given the previous Y_{i-1} and the next Y_{i+1} state, we use the Metropolis-within-Gibbs (M-W-G) algorithm. Accordingly the candidate value for the ith state Y_i^* is proposed from the following multivariate normal distribution N. In this expression $\beta(Y_{i-1}, \theta)$ displays the diffusion matrix of the previous state Y_{i-1} for the given θ.

$$Y_i^* \sim N\left(\frac{1}{2}(Y_{i-1} + Y_{i+1}), \quad \frac{1}{2}\Delta t\beta(Y_{i-1}, \theta)\right) \tag{3}$$

Eraker (2001) shows that the transiton kernel, $q(Y_i \mid Y_{i-1}, Y_{i+1}, \theta)$, formulated in Equation (3) converges to the true distribution of Y_i, $\pi(Y_i \mid Y_{i-1}, Y_{i+1}, \theta)$, when $\Delta t \to 0$. If the state has additional observed measurements x_i, we consider to generate merely the candidate Z_i, Z_i^*, by further conditioning Y_i^* on $X_i = x_i$ since each Y_i^* can be decomposed as

$$Y_i^* \equiv \begin{pmatrix} X_i \\ Z_i^* \end{pmatrix} \tag{4}$$

Then for each state, the acceptance probability is computed for the candidate Y_i^* by

$$\alpha(Y_i^* \mid Y_i) = \min\left\{1, \quad \frac{p(Y_i^* \mid Y_{i-1}, Y_{i+1}, \theta) q(Y_i \mid Y_{i-1}, Y_{i+1}, \theta)}{p(Y_i \mid Y_{i-1}, Y_{i+1}, \theta) q(Y_i^* \mid Y_{i-1}, Y_{i+1}, \theta)}\right\} \tag{5}$$

where

$$\begin{aligned} p(Y_i \mid Y_{i-1}, Y_{i+1}, \theta) = & \mid \beta(Y_{i-1}, \theta)^{-1} \mid^{\frac{1}{2}} \mid \beta(Y_i, \theta)^{-1} \mid^{\frac{1}{2}} \\ & \exp\left\{-\frac{1}{2}(Y_i - Y_{i-1} - \mu(Y_{i-1}, \theta)\Delta t)'(\Delta t \beta(Y_{i-1}, \theta))^{-1}(Y_i - Y_{i-1} - \mu(Y_{i-1}, \theta)\Delta t)\right\} \\ & \exp\left\{-\frac{1}{2}(Y_{i+1} - Y_i - \mu(Y_i, \theta)\Delta t)'(\Delta t \beta(Y_i, \theta))^{-1}(Y_{i+1} - Y_i - \mu(Y_i, \theta)\Delta t)\right\} \end{aligned} \tag{6}$$

Here $p(Y_i \mid Y_{i-1}, Y_{i+1}, \theta)$ is directly proportional to $\pi(Y_i \mid Y_{i-1}, Y_{i+1}, \theta)$. More details about candidate generators and associated acceptance probabilities can be found in Golightly and Wilkinson (2005).

Once the updates of missing states are completed, the system executes the updates of reaction rates by the random walk algorithm. In this method the candidate rates are generated from the normal distribution and the acceptance probability is calculated by

$$\alpha(\theta, \theta^* \mid Y) = \min\left\{1, \frac{L(\theta^* \mid Y)}{L(\theta \mid Y)}\right\} \tag{7}$$

in which

$$L(\theta \mid Y) = \prod_{i=1}^{T} \pi(\theta) f(Y_i \mid Y_{i-1}, \theta) \tag{8}$$

In Equation (7), θ^* indicates the proposal rates which are produced via $\theta_j^* = \theta_j + \varphi_j$ ($j = 1, ..., r$) where $\varphi_j \sim N(0, \delta_j)$. The variance of each rate δ_j is called the "tuning parameter" and significantly affects the mixing property of the algorithm (Golightly and Wilkinson, 2005). For a good mixing in univariate random walk chains it is suggested that an acceptance ratio p of around 24% is optimal (Roberts et al., 1997). On the other hand for the multivariate inference, the optimal p is found as 0.574 (Roberts and Rosenthal, 1998). However, when the complexity of the network structure increases, very low ratios such as 5% can be tolerable since it is difficult to produce a candidate value for particular reaction rates. Thus, in our estimation to get a sensible value for the

variance of each rate δ_j, we define δ_j adaptively during the burn-in period of MCMC runs. We multiply δ_j by 1.1 if the acceptance ratio p at every 100th iteration in the burn-in is greater than 60% and we divide δ_j by 1.1 if p is less than 5%. Whereas if p lies between 5% and 60%, we keep the current δ_j. At the end of the burn-in, the final set of δ's is taken as constants and used until the end of the inference.

Indeed, apart from these highlighted optimal acceptance ratios, there are a number of other methods which can assess the convergence of the chain. For instance the sample autocorrelation function (Golightly and Wilkinson, 2005 and 2006b) and the posterior density of each parameter (Gelman et al., 2004), the potential scale reduction (Gelman et al., 2004), and the value of the convergence diagnostic (Geweke, 1992) are some of the methods used for monitoring the convergence. In Section 4 to control the convergence of our estimates, we choose the autocorrelation function and the posterior density besides the evaluation of results via acceptance ratios.

On the other hand $\pi(\theta)$ in Equation (8) shows the prior distribution of reaction rates which is taken as exponential with rate 1 seeing that it satisfies the positivity condition of our model parameters and $f(Y_i \mid Y_{i-1}, \theta)$ displays the transition density of the ith state given the previous state and reaction rates. Therefore $L(\theta \mid Y)$ in Equation (8) can be formulated as

$$L(\theta \mid Y) = \prod_{i=1}^{T} \exp\left\{-\sum_{k=1}^{r} \theta_k\right\} \mid \beta(Y_{i-1}, \theta) \mid^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(Y_i - Y_{i-1} - \mu(Y_{i-1}, \theta)\Delta t)'(\Delta t \beta(Y_{i-1}, \theta))^{-1} (Y_i - Y_{i-1} - \mu(Y_{i-1}, \theta)\Delta t)\right\} \quad (9)$$

Further discussion about the updates of rates can be found in Purutçuoğlu and Wit (2008b).

3.1 MCMC Steps for the MAPK Pathway

In the update of the system via MCMC algorithms and data augmentation schemes, we observe that the singularity of diffusion matrices and the dependency between substrates are the main challenges. In order to unravel the first problem, we suggest to work with only nonsingular matrices. Therefore, in every stage of updates we check the corresponding candidate diffusion terms whether they would cause singularities in the system if they are accepted. If the candidate generator leads to singular diffusions, then we simply reject the candidate states and rates without computing the acceptance probability α. Otherwise, α is calculated in order to decide on the next step. We call this kind of dependency the "incidental dependence" (Purutçuoğlu and Wit, 2008b). The second problem, on the other side, is originated from the dependency of V matrix where V is the (rxn) - dimensional net effect matrix. In that case, the substrates are dependent on each other from the description of the system since V is directly produced from the quasi list of systems' reactions. We name this dependency as the "structural dependence". In the update of the system under the structural dependence, we can exclude the problematic species at the beginning of MCMC algorithms (Purutçuoğlu

and Wit, 2008b). Because any linear dependence in this matrix, *V*, affects the rank of $V'V$, which is used in the computation of $\beta(Y_t, \theta)$ as stated in Section 2 and leads to singular diffusions. On the other hand, this elimination implies that we lose some of the observed *X* and unobserved *Z* components. Thus the exclusion can rise the average errors of estimates as the number of dependent substrates increases. For the MAPK pathway the number of structurally dependent proteins is 17 over 51 proteins, which correspond to a large proportion (around 33%) with respect to the total number of substrates in the system.

As an alternative approach for the elimination of these species, we consider to include them implicitly in our computation. We suggest that if we preserve the linear relationships between dependent and independent proteins, these relations can be used to generate dependent substrates after the updates of linearly independent terms. Then if these values are added in the calculations of diffusion and drift terms, then the system can be updated under both incidental and structural dependencies. Indeed we have implemented this idea in the simulation of the complex MAPK pathway via the diffusion approximation and we have observed that the method successfully deals with the singularity, which is particularly seen in the steady-state phase of the simulation (Purutçuoğlu and Wit, 2006). Therefore, we develop a new updating scheme for the inference which is based on that implicit computation. We list the steps of the underlying plan as follows:

1. The system is initialized by assigning values for missing states and reaction rates and the counter of iterations *g* is set to zero.
2. After the initialization, all *n* columns of *V* are checked from left to right whether there is any linearly dependent column, denoted as s, which indicates structurally dependent species. For simplicity we assume that we have totally |*s*| dependent columns, thereby (*n*-|*s*|) independent proteins. Then for each dependent species, the vector, which displays the coefficients of the linear relationship between dependent and independent substrates, λ_{jl} ($j \in s$ and $l = 1,...,n-|s|$), is preserved.
3. The system begins the updates from the states, whose substrates are linearly independent, Y_i^{indep} ($i = 1,...,T$). The candidate value of Y_i^{indep}, Y_i^{indep*}, is generated from the multivariate normal distribution given in Equation (3) and Equation (4). If the proposal state maintains the singularity of the candidate diffusion matrix β_i^*, that is the incidental dependence is not observed in β_i^*, as well as the positivity of the state is satisfied, then it is accepted as the generator for the linearly dependent proteins. Here as the candidate generators (Equation (3) and Equation (4)) are used for the linearly independent terms, which do not indicate neither the incidental nor structural dependence, the transiton kernels given in the study of Eraker (2001) and performed in our research still maintain the convergent properties to the true distribution. Indeed, from our reference study of Golightly and Wilkinson (2005), we also observe a structural dependence in the net effect matrix of a small prokaryotic autoregulation system. In order to unravel the singularity of the diffusion term, that particular dependent subsrate is excluded from the beginning of the estimation and the generators are produced from the remaining linearly independent species. With respect to that system of interest, our network is significantly complex, accordingly, the dependency is observed very often. Although we believe that

the missing data and the underlying high dependency between species can lead to biased estimates, the problems of inaccuracies of estimates can be improved by alternative approaches. More details about the possible solutions of the problems by using the same transition kernels of Eraker (2001) can be found in Section 5 and Purutçuoğlu and Wit (2008b). On the other hand, other alternative solutions to decline the dependency on the estimates can be seen in the studies of Golightly and Wilkinson (2006a and 2006b). In those works, basically, they suggest to update the missing data in block of random size and to implement the method of particle filterings.

4. To produce totally |s| linearly dependent species, initially (*n*-|*s*|) increments ψ are generated from the Brownian motion, i.e. the normal distribution with mean zero and variance Δt. These increments are multiplied by the square root of the diffusion term obtained from the previous time step $\beta_{i-1}^{1/2}$ of linearly independent species. Therefore, our computed $\beta_{i-1}^{1/2}$ matrix has the dimension of (*n*-|*s*|)x(*n*-|*s*|). In that way, we get the error term $\varepsilon = \psi\beta_{i-1}^{1/2}$ for linearly independent substrates which corresponds to $\beta^{1/2}(Y_t,\theta)\Delta Wt$ in Equation (2). Then the change in the state of new dependent substrates from $i = t$ to $i = t + \Delta t$ is simulated via $\Delta Y_i^{dep*} = \mu(Y_{i-1}^{indep},\theta)\Delta t + \varepsilon$ similar to Equation (2) in which $\mu(Y_{i-1}^{indep},\theta)$ refers to the (*n*-|*s*|)-dimensional drift vector of the previous state whose substrates are linearly independent. Hence, ΔY_i^{dep*} gives an (*n*-|*s*|) - dimensional vector. Accordingly the candidate state for dependent species, Y_i^{dep*}, is generated as $\Lambda^{dep*} = \sum_{l\notin s, l<j}^{n-|s|} \Delta Y_l^{dep^*}\lambda_{jl}$ and $Y_i^{dep*} = Y_{i-1}^{dep} + \Lambda^{dep*}$ when $j \notin s$, $l \in s$, and $l = 1,...,n-|s|$. Λ^{dep*} corresponds to an |*s*|-dimensional vector and represents the change in the state Y_i^{dep*} that is computed by the linear relation within dependent and independent proteins. On the other hand, Y_{i-1}^{dep} stands for the updated state Y at time $t = i-1$, whose proteins are linearly dependent. Finally, a complete proposal state Y_i^* is produced by combining Y_i^{indep*} with Y_i^{dep*} as a vectoral form.

5. The drift μ_i and the diffusion β_i of the updated state are computed from the hazard function of each reaction based on Y_i^*, i.e. $h(Y_i^*,\theta)$. If we do not observe a new inner dependence between linearly dependent substrates from the computation of the recent β_i, in other words, if we do not write any of the linearly dependent substrate in terms of other linearly dependent substrates, then the acceptance probability $\alpha(Y_i^* \mid Y_i)$ is calculated by (*n*x*n*) - dimensional diffusion matrices of Y_{i-1} and Y_i^*. Y_{i-1} shows the updated state at time $t = i-1$. Otherwise, $\alpha(Y_i^* \mid Y_i)$ is found from only linearly independent species. For the MAPK pathway, since we riddle with an inner linear dependence within linearly dependent substrates, α is derived from lower dimensional diffusion matrices whose components are linearly independent proteins. Whereas the computation of hazards is executed on both dependent and independent species as described beforehand.

6. From the result of $\alpha(Y_i^* \mid Y_i)$, if the move is accepted, $Y_i^{(g)} = Y_i^*$ at the *g*th iteration. Otherwise, the system preserves the current state. Then we return to Step 3 to update the $(i+1)$th state by M-W-G algorithm and repeat the process until $i = T-1$. In the final column, i.e. when $i = T$, we perform the Gibbs sampling in place of M-W-G and directly accept the proposal state Y_T^* without computing $\alpha(Y_T^* \mid Y_T)$.
7. The model parameters of the system, i.e. reaction rates, are updated via the random walk algorithm by *d*-dimensional blocks. The *d*-number of deviance terms is generated from the normal distribution with mean zero and variance δ_j $(j = 1,...,r)$ and is added to the current θ to produce a candidate θ, θ^*. The new θ^* for each *d*-dimensional group is controlled whether it causes a new source of incidental dependences when it is used in the diffusions of Y_i $(i = 1,...,T)$. If θ^* does not lead to any singularity, the acceptance probability given in Equation (7), $\alpha(\theta, \theta^* \mid Y)$, is computed. If the candidate reaction rates increase the likelihood, the move is accepted and $\theta^{(g)} = \theta^*$ at the *g*th iteration, otherwise, the chain does not move. On the other hand, if θ^* results in an incidental dependence, then a new θ^* is proposed until the nonsingularity of all diffusion terms is satisfied for every state.
8. When all states and reaction rates are updated, the counter of the algorithm goes from g to $(g+1)$. The algorithm is repeated from Step 2 until the system converges to the stationary distribution.

4. APPLICATION OF THE METHOD

In order to evaluate the performance of MCMC algorithms, we use a simulated dataset which we previously applied in our analysis (Purutçuoğlu and Wit, 2008b). This dataset is generated from the Gillespie algorithm and has 28 observed and linearly independent substrates, and 23 unobserved substrates in which 6 of them are linearly independent and the remaining 17 terms are dependent species. We choose 50 time points from the underlying data and accept that these are our time-course measurements. Then we extent the dataset by adding 3 augmented states between each pair of 50 time points. Therefore, we generate an observation matrix *Y* which has 197 instead of 50 columns, i.e. $i = 1,...,197$. The complete list of observed and unobserved substrates and more details about the simulated data can be found in Purutçuoğlu and Wit (2008b).

In this study, all the computational work is carried out in the programme language R and our codes are executed on Dual Core Xeon 3.00 GHz processor. To estimate the reaction rates of the MAPK pathway, we iterate the algorithm 200,000 times and take the mean of the last 50,000 MCMC outputs as the estimated values of our model parameters. The lists of estimated rates with true values are presented in Table 1 and Table 2. The first table shows the results from the new algorithm and the second one illustrates the outputs obtained by MCMC algorithms which are conducted by merely linearly independent substrates. From both tables, it is found that most of the acceptance ratios of estimated values lie between 0.05 and 0.60 which display good mixing properties in the inference. Figure 2 and Figure 3 are drawn as an example from the posterior distributions of selected reaction rate constants and their autocorrelation

functions after the burn-in. From the figures it is seen that the selected parameters indicate convergent distributions supporting their acceptance ratios given in Table 1 and Table 2. But the new plan typically offers lower acceptance ratios than the old plan produces. On the other side, from the comparison of the average error of each estimate calculated by the following Equation (10),

$$\text{Average error} = |\text{True value} - \text{Estimated value}| \,/\, \text{True value} \qquad (10)$$

we observe that the new algorithm considerably decreases the error (Table 3). Moreover, from the evaluation of the CPU (Central Processing Unit) time, it is seen that the new scheme also offers a less computational cost (Table 3). Indeed, with respect to the complexity of algorithms, the new scheme has more computational steps, thereby it is expected that this scheme should be computationally more demanding. From our results although, at first sense it seems to be a contradiction, we explain this situation as follows: As stated in Section 3.1, the MAPK pathway can use the dependent substrates solely in the calculation of hazards functions, rather than during the calculation of acceptance probabilities of both rates and states. Hence, the complete computation of dependent substrates according to the new plan cannot be performed in our system. In other words, the steps of both the new and previous algorithms are run for $(n\text{-}|s|)$ terms

Table 1. Posterior means (μ), standard deviations (σ), and acceptance ratios (p) of estimated reaction rate constants found by the MCMC plan which includes structurally dependent substrates

Reaction	True rate	μ	σ	p	Reaction	True rate	μ	σ	p
c_2	0.010	0.020	0.000	0.291	c_{35}	0.010	4.853	0.255	0.450
c_3	0.010	0.051	0.007	0.303	c_{36}	0.010	0.235	0.002	0.468
c_4	0.010	0.130	0.002	0.271	c_{37}	0.010	0.576	0.004	0.510
c_5	1.000	0.596	0.007	0.294	c_{38}	1.000	0.130	0.002	0.476
c_6	1.000	0.996	0.001	0.029	c_{39}	1.000	0.130	0.002	0.456
c_7	1.000	1.001	0.001	0.021	c_{40}	1.000	0.015	0.000	0.499
c_8	1.000	1.028	0.001	0.021	c_{41}	1.000	0.001	0.000	0.014
c_9	0.010	0.001	0.000	0.029	c_{42}	0.010	0.000	0.000	0.014
c_{10}	0.010	0.000	0.000	0.029	c_{43}	0.010	0.252	0.005	0.014
c_{11}	1.000	2.761	0.060	0.548	c_{44}	1.000	0.257	0.001	0.014
c_{12}	0.015	1.619	0.077	0.554	c_{45}	0.015	0.354	0.006	0.014
c_{13}	0.010	0.060	0.001	0.574	c_{46}	0.010	0.002	0.000	0.058
c_{14}	0.010	0.082	0.001	0.595	c_{47}	0.010	0.024	0.006	0.058
c_{15}	0.010	0.083	0.001	0.613	c_{48}	0.010	0.319	0.040	0.058
c_{16}	0.010	4.456	0.117	0.820	c_{49}	0.010	0.119	0.039	0.058
c_{17}	1.000	0.294	0.004	0.776	c_{50}	1.000	0.001	0.000	0.058
c_{18}	0.010	5.404	0.175	0.848	c_{51}	0.010	3.643	0.100	0.782
c_{19}	1.000	0.337	0.006	0.817	c_{52}	1.000	0.126	0.001	0.533
c_{20}	1.000	3.334	0.224	0.866	c_{53}	1.000	0.097	0.001	0.821
c_{21}	0.010	0.041	0.004	0.426	c_{54}	0.010	0.078	0.003	0.850
c_{22}	0.010	4.913	0.235	0.420	c_{55}	0.010	3.875	0.130	0.801
c_{23}	0.015	1.264	0.007	0.283	c_{56}	0.015	0.013	0.000	0.019
c_{24}	0.010	0.010	0.000	0.390	c_{57}	0.010	0.000	0.000	0.020
c_{25}	0.010	0.066	0.001	0.420	c_{58}	0.010	0.004	0.000	0.019
c_{26}	0.010	0.003	0.000	0.345	c_{59}	0.010	0.637	0.007	0.019
c_{27}	0.010	0.416	0.002	0.324	c_{60}	0.010	0.000	0.000	0.020
c_{28}	0.010	0.058	0.001	0.355	c_{61}	0.010	0.000	0.000	0.428
c_{29}	0.010	0.090	0.001	0.353	c_{62}	0.010	0.004	0.000	0.422
c_{30}	0.010	0.016	0.000	0.333	c_{63}	0.010	1.028	0.011	0.317
c_{31}	0.010	0.014	0.010	0.461	c_{64}	0.010	0.705	0.034	0.420
c_{32}	0.010	0.019	0.000	0.415	c_{65}	0.010	0.418	0.006	0.404
c_{33}	1.000	0.223	0.003	0.415	c_{66}	1.000	9.448	0.484	0.775

Table 2. Posterior means (μ), standard deviations (σ), and acceptance ratios (p) of estimated reaction rate constants found by the MCMC plan which excludes structurally dependent substrates

Reaction	True rate	μ	σ	p	Reaction	True rate	μ	σ	p
c_2	0.010	0.050	0.001	0.540	c_{35}	0.010	2.465	0.096	0.489
c_3	0.010	0.040	0.001	0.535	c_{36}	0.010	0.269	0.004	0.457
c_4	0.010	0.031	0.001	0.547	c_{37}	0.010	0.422	0.006	0.478
c_5	1.000	5.228	0.187	0.541	c_{38}	1.000	0.720	0.019	0.492
c_6	1.000	1.350	0.034	0.247	c_{39}	1.000	1.011	0.015	0.410
c_7	1.000	1.245	0.020	0.232	c_{40}	1.000	0.019	0.000	0.460
c_8	1.000	0.956	0.017	0.265	c_{41}	1.000	0.002	0.000	0.153
c_9	0.010	0.252	0.005	0.257	c_{42}	0.010	0.002	0.004	0.156
c_{10}	0.010	0.000	0.000	0.273	c_{43}	0.010	0.008	0.005	0.156
c_{11}	1.000	2.165	0.040	0.532	c_{44}	1.000	0.483	0.003	0.149
c_{12}	0.015	1.120	0.045	0.544	c_{45}	0.015	1.053	0.007	0.125
c_{13}	0.010	4.719	0.168	0.597	c_{46}	0.010	0.890	0.034	0.779
c_{14}	0.010	0.040	0.000	0.582	c_{47}	0.010	0.052	0.002	0.777
c_{15}	0.010	0.056	0.001	0.593	c_{48}	0.010	9.416	0.205	0.761
c_{16}	0.010	3.723	0.140	0.807	c_{49}	0.010	5.577	0.308	0.773
c_{17}	1.000	0.266	0.004	0.730	c_{50}	1.000	1.155	0.016	0.572
c_{18}	0.010	3.589	0.139	0.816	c_{51}	0.010	7.070	0.280	0.803
c_{19}	1.000	0.288	0.004	0.773	c_{52}	1.000	0.160	0.006	0.671
c_{20}	1.000	1.789	0.090	0.790	c_{53}	1.000	0.175	0.003	0.673
c_{21}	0.010	0.003	0.002	0.437	c_{54}	0.010	0.217	0.005	0.738
c_{22}	0.010	1.317	0.044	0.381	c_{55}	0.010	4.301	0.209	0.794
c_{23}	0.015	0.638	0.004	0.403	c_{56}	0.015	0.014	0.001	0.228
c_{24}	0.010	0.004	0.000	0.424	c_{57}	0.010	1.968	0.053	0.220
c_{25}	0.010	0.181	0.006	0.380	c_{58}	0.010	1.964	0.008	0.188
c_{26}	0.010	4.244	0.151	0.459	c_{59}	0.010	0.239	0.016	0.229
c_{27}	0.010	0.449	0.004	0.433	c_{60}	0.010	0.000	0.000	0.232
c_{28}	0.010	0.070	0.001	0.467	c_{61}	0.010	0.033	0.032	0.578
c_{29}	0.010	0.116	0.002	0.412	c_{62}	0.010	0.012	0.001	0.535
c_{30}	0.010	0.008	0.000	0.459	c_{63}	0.010	1.198	0.010	0.368
c_{31}	0.010	0.008	0.005	0.526	c_{64}	0.010	0.632	0.039	0.570
c_{32}	0.010	0.008	0.000	0.506	c_{65}	0.010	1.119	0.010	0.374
c_{33}	1.000	4.102	0.065	0.473	c_{66}	1.000	9.286	0.475	0.796

Table 3. Mean and standard deviation of average errors of results presented in Table 1 and Table 2 and corresponding CPU used in inference

	Mean of average errors	Standard deviation of average errors	CPU
Including structurally dependent proteins	60.848	144.652	404.74
Excluding tructurally dependent proteins	90.572	185.176	549.38

in place of n species for the new plan and (n-|s|) subsrates for the previous scheme. On the other hand, the inclusion of dependent species in hazards enables to produce more sensible drift and diffusion terms in the updates of rates and missing states. Thus, as understood from the findings, these highlighted improvements in hazards cause less number of singularities in the system, hereby accelerates the speed of computations by the new plan.

As a result, we consider that our innovated algorithm is more advantageous in the inference of complex systems in terms of the accuracy of estimates. However, it cannot be seen as a better algorithm regarding to the computational cost in the estimation of every complex structure, rather it can be evaluated as a computationally efficient method for the network having an inner dependence like the MAPK pathway.

5. CONCLUSION AND DISCUSSION

We have presented a new MCMC scheme which includes structurally dependent substrates in the estimation of reaction rates of a complex biochemical system. In the inference, we have implemented Bayesian methods based on the Euler approximation and data augmentation techniques due to the fact that the former is computationally more efficient than the exact algorithm and the latter can decrease the bias on estimates caused by the discretization of the diffusion approximation via the Euler method.

In our new algorithm, we have generated candidate values of structurally dependent substrates by using their linear relationships with linearly independent proteins. To capture the underlying linear links, we have investigated the singularity of the net effect matrix.

In our system, since we have observed an additional structural dependency within linearly dependent substrates, we have used all proteins in the calculation of hazard functions which are the base components of drifts and diffusion terms. Whereas all acceptance probabilities α have been computed solely by linearly independent terms, because of the fact that the highlighted inner-structural dependence within dependent substrates has led to infeasible likelihoods in α. However, this is a particular problem in the MAPK description. Therefore, we suggest that indeed thanks to this new algorithm, the calculation of α can be easily implemented by dividing it into two parts. In the first part of the calculation, we can compute the likelihood of linearly independent proteins and in the second part, we can only consider the likelihood found by linearly dependent proteins. Then we can multiply these two terms since the application of our new plan enables to factorize the likelihood. This process can be performed for both the update of reaction rates and missing states.

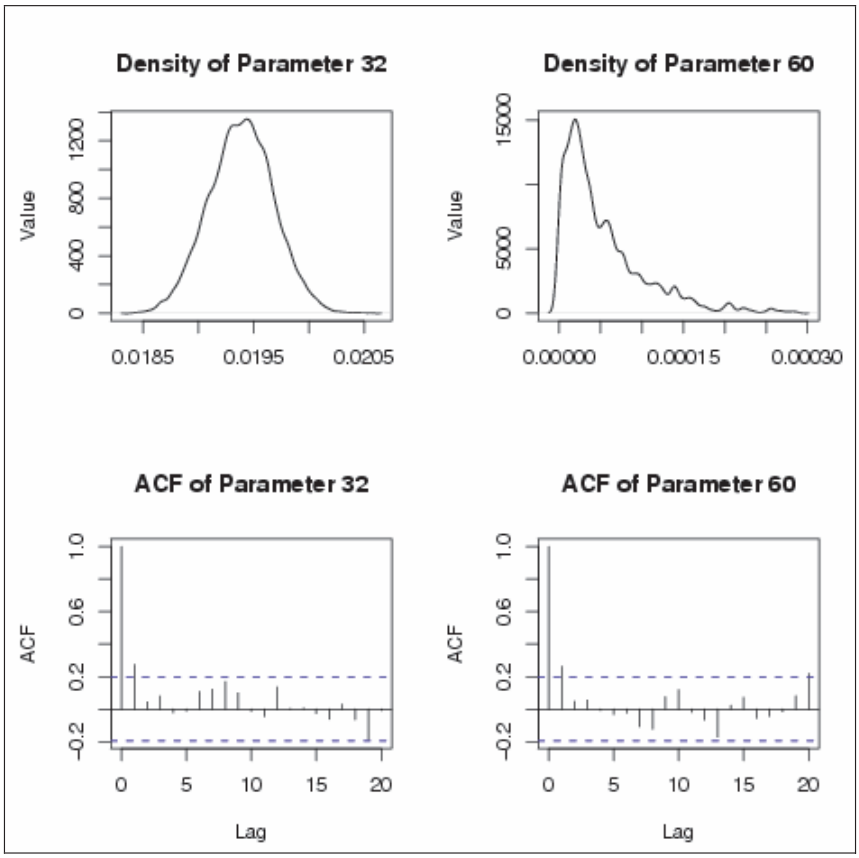


Figure 2. Posterior distributions and autocorrelation functions (ACF) of reaction rate constants 32 and 60 after burn-in via the MCMC plan which includes structurally dependent substrates

As an extension of our study, seeing that we have investigated an inner dependence within structurally dependent species, we propose to develop a sub-algorithm for merely linearly dependent substrates. In that plan, our new scheme can be repeated within these terms iteratively until each linearly dependent protein can be generated in terms of its associated linearly independent species within that particular group. Under this condition the complete likelihood is factorized as a number of independent parts, and thereby can be computed as the product of underlying independent pieces of information. We consider that such an iterative calculation can further improve the accuracy of estimates even though it can also increase the computational cost of the inference. However, we believe that this additional computational demand can considerably decline if the codes are executed on an efficient programme language.

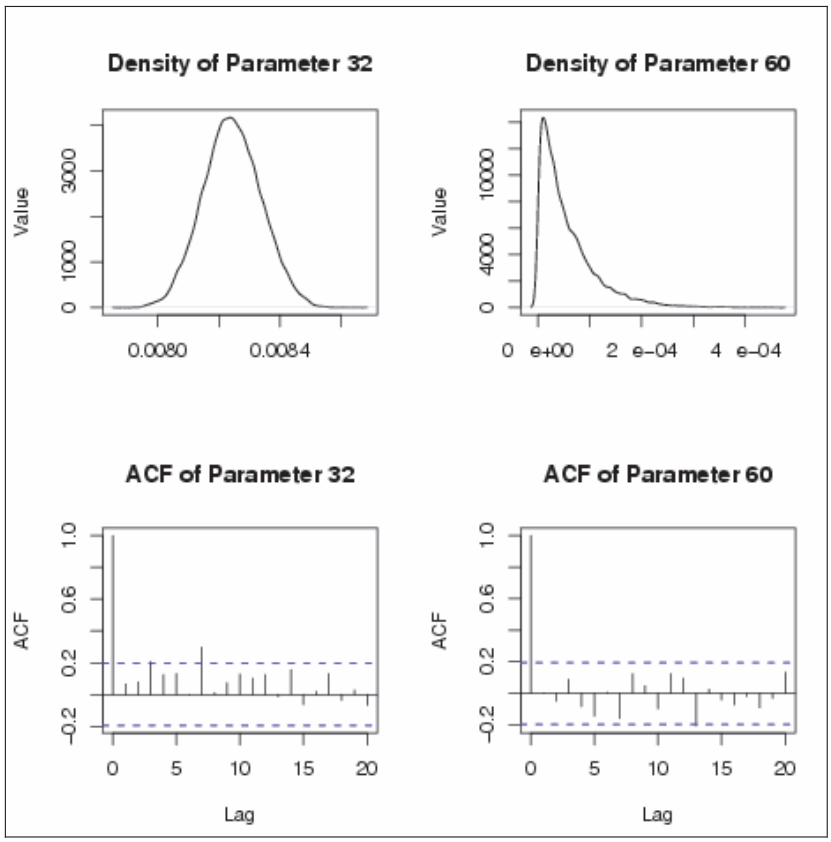


Figure 3. Posterior distributions and autocorrelation functions (ACF) of reaction rate constants 32 and 60 after burn-in via the MCMC plan which excludes structurally dependent substrates

6. REFERENCES

Bower, J.M., and Bolouri, H., 2001. Computational modelling of genetic and biochemical networks (Second edition). Massachusetts Institute of Technology. Cambridge. Massachusetts.

Boys, R.J., Wilkinson, D.J., and Kirkwood, T.B.L., 2008. Bayesian inference for a discretely observed stochastic kinetic model. Statistical Computing, 18, 125-135.

Elerian, B.O., Chib, S., and Shephard, N., 2001. Likelihood inference for discretely observed nonlinear diffusions. Econometrica, 69 (4), 959-993.

Eraker, B., 2001. MCMC analysis of diffusion models with application to finance. Journal of Business and Economic Statistics, 19 (2), 177–191.

Gelman, A., Carlin, J.B., Stern, H.S., and Rubin, D.B., 2004. Bayesian data analysis. Chapman and Hall/CRC. Florida. U.S.A.

Geweke, J., 1992. Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. Bayesian Statistics 4 in Bernando, J.M., Berger, J.O., Dawid, A. P., and Smith, A.F.M. (Eds), 169-193. Oxford University Press. Oxford.

Gibson, M.A., and Bruck, J., 2000. Efficient exact stochastic simulation of chemical systems with many species and many channels. Journal of Physical Chemistry, A(104), 1876–1889.

Gillespie, D.T., 1977. Exact stochastic simulation of coupled chemical reactions. Journal of Physical Chemistry, 81 (25), 2340–2361.

Gillespie, D.T., 1992. A rigorous derivation of the chemical master equation. Physica, A 188, 404–425.

Golightly, A., and Wilkinson, D.J., 2005. Bayesian inference for stochastic kinetic models using a diffusion approximation. Biometrics, 61 (3), 781–788.

Golightly, A., and Wilkinson, D.J., 2006a. Bayesian sequential inference for nonlinear multivariate diffusions. Statistics and Computing, 16, 323-338.

Golightly, A., and Wilkinson, D.J., 2006b. Bayesian sequential inference for stochastic kinetic biochemical network models. Journal of Computational Biology, 13 (3), 838-851.

Kolch, W., Calder, M., and Gilbert, D., 2005. When kinases meet mathematics: the systems biology of MAPK signaling. FEBS Letters, 579, 1891–1895.

Orton, R., Sturm, O.E., Vyshemirsky, V., Calder, M., Gilbert, D.R., and Kolch, W., 2005. Computational modelling of the receptor-tyrosine-kinase-activated MAPK pathway. Biochemical Journal, 392, 249–261.

Purutçuoğlu, V., and Wit, E., 2006. Exact and approximate stochastic simulations of the MAPK pathway and comparisons of simulations' results. Journal of Integrative Bioinformatics, 3, 231-243.

Purutçuoğlu, V., and Wit, E., 2008a. Inclusion of convoluted measurements in Bayesian inference of the MAPK/ERK pathway via multivariate diffusion model. Proceeding of the Third International Symposium on Health, Informatics and Bioinformatics in Sezerman, U. (Ed), Sabancı University, İstanbul, Turkey, CD-Rom.

Purutçuoğlu, V., and Wit, E., 2008b. Bayesian inference for the MAPK/ERK pathway by considering the dependency of the kinetic parameters. Bayesian Analysis, 3 (4), 851-86.

Roberts, G.O., Gelman, A., and Gilks, W.R., 1997. Weak convergence and optimal scaling of random walk metropolis algorithms. The Annals of Applied Probability, 77 (1), 110-120.

Roberts, G.O., and Rosenthal, J.S., 1998. Optimal scaling of discrete approximations to Langevin diffusions. Journal of Royal Statistical Society, Series B, 60 (1), 255-268.

Roberts, G.O., and Stramer, O., 2001. On inference for partially observed nonlinear diffusion models using the Metropolis-Hastings algorithm. Biometrika, 88 (3), 603–621.

Turner, T.E., Schnell, S., and Burrage, K., 2004. Stochastic approaches for modelling in vivo reactions. Computational Biology and Chemistry, 28, 165–178.

Wilkinson, D.J., 2006. Stochastic modelling for systems biology. Chapman and Hall/CRC. Florida. U.S.A.

YAPISAL BAĞIMLILIK ALTINDA KARMAŞIK MAPK YOLUNUN BAYESCİ TAHMİNİ

ÖZET

MAPK yolu, tüm ökaryotlarda bulunan hücresel büyüme kontrolünü düzenleyen başlıca sinyal iletim sistemlerinden biridir. Hayati görevinden dolayı sistemin idaresi çok sayıda protein vasıtasıyla yürütülür, buna bağlı olarak karmaşık bir yapı oluşturur. Çalışmada, Euler yaklaşımına dayalı MCMC teknikleriyle bu sistemin tahmininde diğer proteinlerle yüksek yapısal bağımlılıklar gösteren bir çok proteinin varolduğu gözlenmiştir. Bu proteinler kabul edilme olasılıklarını imkansız yapan tekil difüzyon/varyans matrislerine neden olmuşlardır. Bu nedenle bu sorunlu proteinler tahmin hesabının başında çıkarılmış ve parametreler sadece sistemdeki doğrusal bağımsız türler kullanarak tahmin edilmiştir. Ancak bu durumda da özellikle bağımlı türlerin sayısı arttıkça, tahminin doğruluğu bahsedilen eliminasyondan oldukça etkilenmektedir. Bu proteinlerin elenmesi MCMC'deki mevcut kayıp terim sayısının belirgin derecede artmasına neden olmaktadır. Bu çalışmada dolaylı yoldan bu proteinler, bağımlı terimlerin bağımsız türlerin doğrusal kombinasyonu şeklinde simülasyon eden alternatif bir yaklaşımla hesaplamaların içine katılmaktadır. Bu şekilde reaksiyon oranlarının ve durumlarının kabul edilme olasılıklarını hesaplamada bağımlı türlerin etkileri ilave edilebilmektedir. Analizlerden, bahsedilen yeniliğin tahminlerin ortalama hatalarını azalttığı ve MAPK yolunun tahmininde daha az hesaplama maliyeti önerdiği sonucuna varılmıştır.

Anahtar Kelimeler: Bayesci tahmin, Difüzyon yaklaşımı, MAPK yolu.

Cilt: 06 Sayı: 01 Kategori: 01 Sayfa: 18-37 ISSN No: 1303-6319 Volume: 06 Sayı: 01 Category: 01 Page: 18-37

THE PERFORMANCE EVALUATION OF ROBUST PAIRWISE COVARIANCE ESTIMATOR

Özlem YORULMAZ*

ABSTRACT

Multivariate analysis and multidimensional outlier detection techniques necessitate using robust high breakdown covariance estimators, which have time saving algorithms in the presence of outliers in high dimensional data. The preference for robust estimators arises from the distortion effect of outliers when classical estimators are used. Orthogonalized Gnanadesikan-Kettering (OGK) estimator (Maronna and Zamar, 2002) was devised in order to address the computational challenge of high breakdown estimators. In this study the focus is on the evaluation of some covariance estimators in Principal Component Analysis (PCA). A comparison of the performance of OGK in PCA and Robust Principal Component Analysis (ROBPCA) (Hubert et al, 2005) has been carried out by way of simulations and with real data sets.

Key Words: Fast minimum covariance determinant estimator, Orthogonalized Gnanadesikan-Kettering estimator, Outliers, Principal components analysis, Robust principal component analysis.

1. INTRODUCTION

As is commonly known, the covariance matrix is one of the fundamental instruments of statistical analysis that is widely used for obtaining correlation coefficients between variables, reducing the number of variables and diagnosing multivariate outliers.

An observation whose pattern differs from the majority of data is generally called an outlier. Outliers may cause misleading estimations when classical empirical covariance matrices are used; therefore, statisticians directed their attention to robust techniques and different robust methods have been invented to estimate the covariance matrix. If there are some outliers in the data, the classical (maximum likelihood) estimator of the covariance matrix may not prevent masking (case when analysis suggests that one or more outliers are in fact good cases) and swamping (case when analysis suggests that one or more good cases are outliers) effects. For this reason it is much safer to use robust estimators instead.

* Istanbul University, Faculty of Economics, Department of Econometrics, e-mail: yorulmaz@istanbul.edu.tr

Here some estimators are defined briefly; throughout the definitions X_{nxp} notation is used which stands for nxp data matrices, where n indicates the number of objects and p indicates the number of variables.

OGK is a robust covariance matrix estimator for high dimensional data sets which has been proposed by Maronna and Zamar (2002) as an alternative to Fast Minimum Covariance Determinant (FMCD) estimator. FMCD (Rousseeuw and Van Driessen, 1999) is a high breakdown robust estimator, an improved form of the Minimum Covariance Determinant (MCD) high breakdown estimator. It has been stressed by Maronna and Zamar (2002) that the increase of cases (n) diminishes the high breakdown property of FMCD and it also has been emphasized that the increase of the dimension (p) requires immense computational time for FMCD although it is the quicker alternative of MCD.

Underlying purposes of the study are firstly using OGK covariance matrix in one of the dimension reduction technique PCA, secondly comparing and evaluating some properties of robust and classical matrices with several data sets and simulations. Through the comparison and evaluation step a matlab Library for Robust Analysis (LIBRA) was used and besides codes for OGK estimator were written in Matlab (See, Appendix-2).

2. PROPERTIES OF ROBUST ESTIMATORS

The properties of robust covariance estimator can be summarized as breakdown value, positive definiteness and affine equivariance. These properties allow a characterization of estimators as low breakdown, high breakdown, affine and not affine.

Breakdown value is a maximum amount of contamination that an estimator can carry. This value also measures the robustness of an estimator. As can be inferred from the following notations,

$\hat{\Sigma}$: Covariance matrix estimator
X : Data matrix,
X': Matrix obtained by replacing m points out of X

the breakdown value, ε_n^*, is the largest eigenvalue of $\hat{\Sigma}$ driven to ∞ or the smallest eigenvalue of $\hat{\Sigma}$ driven to zero:

$$\varepsilon_n^*(\hat{\Sigma}, X) = \min\left\{\frac{m}{n}\sup_{X'}\frac{\lambda_{\max(X')}}{\lambda_{\min(X')}} = \infty\right\}$$

Conventional wisdom tells that the covariance matrix yields multivariate scatter of data which is represented by an ellipsoid. The affine equivariance and positive definiteness properties that were mentioned above are strongly related to this ellipsoid because the eigenvectors of a covariance matrix determine the axes of an ellipsoid and the eigenvalues of this covariance matrix are equal to the length of these axes. Given this geometrical concept, the positive definiteness of a covariance matrix can be easily perceived.

Generally the location and scale estimators are expected to be affine equivariant, which means that after a linear transformation of the data the estimators will be transformed accordingly. If A_{pxp} is an orthogonal matrix ($A' = A^{-1}$) and the data matrix is transformed as $XA'+1_n v$, then the center $\hat{\mu}_x$ and the loading matrix $P_{p,k}$ of CPCA or ROBPCA are equal to $A\hat{\mu}_x + v$ and AP respectively. The eigenvalues of the defined ellipsoid and the scores remain the same under this transformation for CPCA and ROBPCA. If an orthogonal transformation is applied to the data as XA' and an estimator rotates accordingly, this estimator can be defined as an orthogonal equivariant. From the above it can be deduced that CPCA and ROBPCA estimators are location and orthogonal equivariant but, as will become clear from the simulation study, OGK is not. This can be rated as a disadvantage of OGK because the absence of the equivariance property makes it hard to predict the behavior of the OGK against outliers on rotated data.

3. ADVANCES ON ROBUST COVARIANCE ESTIMATORS

In the statistical literature, a substantial number of studies have been proposed about robust scatter matrix estimation. The M estimator is the initial robust scatter matrix estimator which was suggested by Hampel in 1973, then studied by Maronna (1976) and Huber (1981). This estimator is positive definite and affine equivariant, but its breakdown point, 1/p, is not satisfactory.

Subsequently, high breakdown affine equivariant and positive definite estimators have been studied. These are the Stahel-Donoho (SD) estimator by Stahel-Donoho (1981) and studied by Maronna and Yohai (1995), the Minimum Volume Estimator (MVE) and the Minimum Covariance Determinant (MCD) by Rousseeuw (1987, 254). Due to the efficiency in high dimensions Croux and Haesbroek recommend to use MCD instead of MVE (2000).

MCD is a highly robust estimator of multivariate location and scatter. Its objective is to find h observations out of n whose classical covariance matrix has the lowest determinant where h is defined as a default value (n+p+1)/2. The value for h is $\lfloor n+p \rfloor / 2 \le h \le n$. The estimation of MCD is time-consuming and therefore limited to a few hundred objects in a few dimensions since the exact solution has to be found among all possible subsets of n observations taken in h dimensional subgroups.

FMCD (Rousseeuw and Van Driesen, 1999) has been developed to address this shortcoming; the algorithm of this estimator is set up on a re-sampling scheme which is called the C-step. But it has to be stressed that FMCD still requires substantial computation time when n is large (Alqallaf et al, 2002).

4. OGK ESTIMATOR

As a result of giving up the requirements of affine equivariance and positive definiteness, one can get estimates much faster. A straightforward estimator for multivariate location is a coordinatewise one which can be calculated from a robust location estimator for each variable in the data. Similarly for a multivariate covariance matrix, pairwise estimators can be used by applying a robust covariance estimate to each pair of variables.

Because of the computational burden of affine equivariant and high breakdown estimators, Marrona and Zamar (2002) dropped the affine equivariance property and introduced the OGK estimator. OGK is based on the Gnanadesikan-Kettenring (G-K) robust pairwise covariance matrix estimate. The G-K estimator,

$$\mathrm{cov}(X,Y)=\frac{1}{4}(\sigma(X+Y)^2-\sigma(X-Y)^2),$$

was suggested by Gnanadesikan and Kettenring (1972). This estimator is not guaranteed to be positive definite whereas the OGK pairwise estimator preserves positive definiteness.

Before explaining the steps of the algorithm, some notations have to be defined as X_j refers to the columns of X data matrix where j is $j=1,.....p$ and $x_i^{'}$ refers to the rows where i is $i=1,...n$.

- For each variable MAD values and w_i weights are calculated. MAD stands for the median absolute deviation from the median and w_i values are obtained from $W_c(x)$ function.

$$\sigma_{0j}=MAD(X_j)=med(\left|X_j-1_n med(X_j)\right|)$$

$$W_j=W_{c1}\left((x_j-1_n med(X_j))/\sigma_{0j}\right),\ W_c(x)=\left(1-\left(x/c\right)^2\right)^2 I(\left|x\right|\le c) \qquad (1)$$

and $I(.)$ is the indicator function.

- Location and scale statistics are obtained from

$$\mu(X_j)=\sum_i x_{ij}w_{ij}/\sum_i w_{ij} \qquad \text{and}$$

$$\sigma(X_j)^2=(\sigma_{0j}^2/n)\sum_i \rho_{c_2}\left((x_{ij}-\mu(X_j))/\sigma_{0j}^2\right) \qquad (2)$$

where ρ_{c_2} can be obtained from $\rho_c(x)=\min(x^2,c^2)$.

Maronna and Zamar (2002) proposed to use $c_1=4.5$ and $c_2=3$ for combining the robustness and efficiency.

- A new diagonal matrix is defined by means of scale statistics that were obtained in the previous step

D = diag($\sigma(X_1),....,\sigma(X_p)$); using the inverse of D with the columns of X_j a new variable,

$$Y = D^{-1}X', \quad (3)$$

Y is defined.

This step makes the estimator scale equivariant.

- U=[u_{jk}] correlation matrix is computed by applying v to the columns of Y.

$$U_{jk} = v(Y_j, Y_k) = \begin{cases} \frac{1}{4}(\sigma(X_j + Y_k)^2 - \sigma(X_j - Y_k)^2) & j \neq k \\ 1 & j = k \end{cases} \quad (4)$$

- The eigenvalues λ_j and the eigenvectors e_j of U ($j = 1,....p$) are obtained and new matrices are defined as $\Lambda = diag(\lambda_1,..,\lambda_p)$ and E whose columns are the e_j's. Then U is decomposed as $U = E\Lambda E'$.
  $A = DE$, and $Z = (E'Y')' = (A^{-1}X')'$ are defined. (5)

- After the extraction of $\Gamma = diag(\sigma(Z_1)^2,.....,\sigma(Z_p)^2)$, the seeking Orthogonolized Gnanadesikan-Kettenring estimators $V(X) = A\Gamma A'$ and $t(X) = A\nu$, where and
  $\nu = (\mu(Z_1),....,\mu(Z_p))'$ are found. (6)

- Maronna and Zamar (2002) suggested using an improvement for the resulting estimator by a reweighting procedure.

$$t_{wj} = \sum_i w_i X_j / \sum_i w_i \ , \qquad V_{wj} = [\sum_i w_i (X_j - 1_n t_{wj})(X - 1_n t_{wj})']/\sum_i w_i \quad (7)$$

The weight function W, $W(d) = I(d \leq d_0)$, can be extracted from,

$$d_i = \sum_j \left((z_{ij} - \mu(Z_j))/\sigma(Z_j)\right)^2 \text{ and } \qquad d_0 = \chi_p^2(\beta) med(d_1,..,d_n)/\chi_p^2(.5)$$

This resulting estimator is called R-OGK (Reweighted Orthogonolized Gnanadesikan-Kettenring) estimator.

Maronna and Zamar discussed different β values with respect to their simulation results and they mentioned that $\beta = 0.90$ generally yielded the best results. Also the R-OGK procedure can be iterated by replacing U in step 5 by $E\Gamma E'$ until convergence but authors warned not to iterate beyond the second iteration.

5. CLASSICAL AND ROBUST PCA

Principal Component Analysis is a technique for explaining the covariance structure of the data by forming new orthogonal variables which are linear combinations of the original variables. These new variables are referred to as principal components which correspond to the eigenvectors of the covariance matrix. The first principal component accounts for the maximum variance of projected data points on it. The second principal component accounts for the maximum variance that has not been accounted for by the first principal component. The procedure continues in this way and it is expected to use few principal components for most of the variance in the data.

But as the principal components are the eigenvectors of classical covariance matrix, it is possible that the components have been adversely influenced by outliers. In this case it is preferable to use robust principal component approaches which can prevent outlier effects. These approaches can be categorized into three different groups:

- replacing classical covariance matrix with robust covariance matrix
- using projection pursuit method
- combining projecting pursuit and robust covariance matrix

Campbell (1980) used M estimators of covariance matrices but they are not resistant against many outliers. Croux and Haesbroeck (2000) used MCD by replacing the classical covariance matrix. However this method is limited to small, moderate samples. In this study, the OGK covariance matrix was replaced with the classical covariance matrix in a similar way and the results are presented through simulation and real data sets.

Li and Chen (1985) and Hubert et all (2002) used the projection pursuit method for obtaining robust PCA.

Hubert, Rousseeuw and Vanden Branden (2005) proposed ROBPCA method which is a combination of the projection pursuit method and the MCD estimator.

6. EVALUATION OF THE ESTIMATORS' PERFORMANCE

The assessment of breakdown point and computational time of CPCA, ROBPCA, PCA with OGK and R-OGK were carried out on real data sets and with simulations.

Before illustrating the methods on real data sets, it is necessary to mention the kind of outliers that can occur and their diagnostic plot. Here the definitions are given briefly. A satisfactory explanation with a visual plot can be found in ROBPCA (Hubert et al, 2005).

- Good leverage points: These points lie close to the PCA space but far from the major homogenous data group.
- Orthogonal outliers: These observations have large orthogonal distances to the PCA space; only their projections can be seen on the PCA space.

- Bad leverage points: This type of observations has a large orthogonal distance and its projections on the PCA space are far from typical projections.

The classification of observations can be identified from a diagnostic plot. The horizontal axis of the diagnostic plot consists of the score distance and the vertical axis of the diagnostic plot consists of the orthogonal distance.

- Score distance is calculated for each observation with

$$SD_i = \sqrt{\sum_{j=1}^{k} (t_{ij}^2 / l_j)}$$

where the t_{ij} pca scores are obtained from $T_{n,k} = (X_{n,p} - 1_n \hat{\mu}^{'}) P_{p,k}$. Here, $l_1,...l_k$ stands for the eigenvalues and $P_{p,k}$ represents the matrix which consists of eigenvectors.

- Orthogonal distance is defined for each observation as

$$OD_i = \left\| x_i - \hat{\mu} - P_{p,k} t_i^{'} \right\|$$

For classifying observations two cut-off lines are drawn. The cut off value on the horizontal axis is $\sqrt{\chi^2_{k,0.975}}$. There are several approaches for the distribution of the cut-off value on the vertical axis (Hubert et al, 2005). According to the Wilson-Hilferty approach orthogonal distances to the power 2/3 are normally distributed. Estimations of the mean and the variance of this distribution were found by means of univariate MCD in ROBPCA paper, in a similar way the $\hat{\mu}$ and $\hat{\sigma}^2$ for OGK and R-OGK were found by univariate OGK and univariate R-OGK. Then, the cut-off value on the vertical axis is defined as $(\hat{\mu} + \hat{\sigma} z_{0.975})^{3/2}$.

6.1 Real Data

CPCA, ROBPCA, OGK and R-OGK methods were applied on two data sets* which are commonly used in robust studies.

6.1.1 Car data

The first example is the low dimensional car data set which contains 111 cars and 11 different characteristics of cars. From the Figure 1 observations 25, 30, 32, 34, 36 are seen as good leverage points and observations 103-108, 110 are seen as orthogonal outliers. However the diagnostic plots of ROBPCA (Figure 2) and R-OGK (Figure 3) identifies this orthogonal outlier group and observations 106, 108 and 110 as bad leverage points. OGK (Figure 4) also converts those cases to bad leverage points but with a difference. As it seen from the Figure 2 they are very close to boundary.

*Data were provided by Karlien Vanden Branden

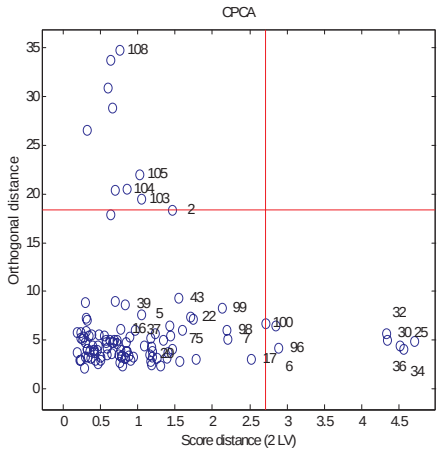


Figure 1. Diagnostic plot of car data set based on two CPCA Principal Components

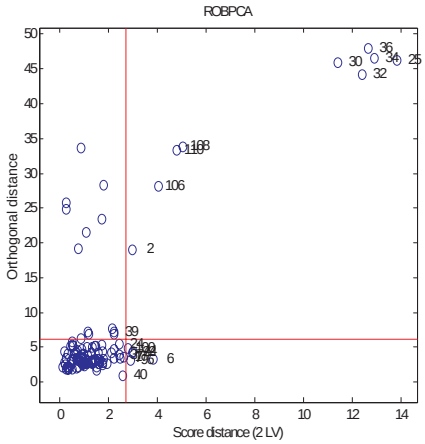


Figure 2. Diagnostic plot of car data set based on two ROBPCA Principal Components

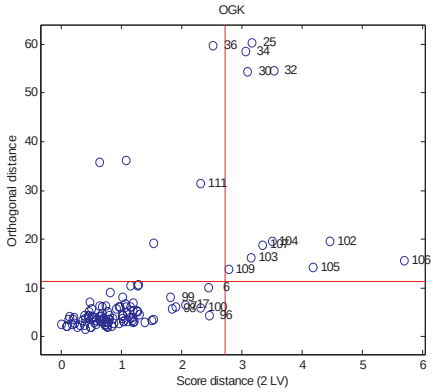


Figure 3. Diagnostic plot of car data set based on two OGK PCA Principal Components

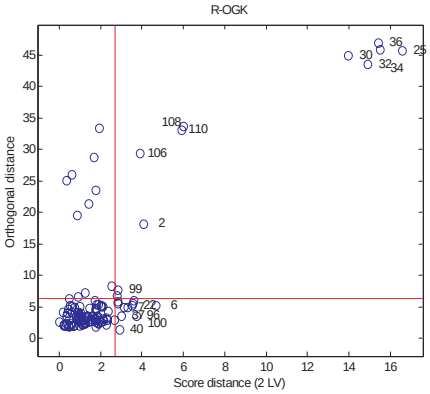


Figure 4. Diagnostic plot of car data set based on two R-OGK PCA Principal Components

6.1.2 Octane data

The second example is the Octane high dimensional data set which consists of 226 variables and 39 gasoline samples. In this data set, six samples contain (25, 26, 36-39) added alcohol.

It is obvious from Figure 5 that CPCA can detect only outlying 26 as a bad leverage point. In contrast, ROBPCA (Figure 6), OGK (Figure 7) and ROBPCA (Figure 8) find all outlying points. This shows that ROBPCA, OGK and R-OGK methods do not contain outliers in their estimated subspaces.

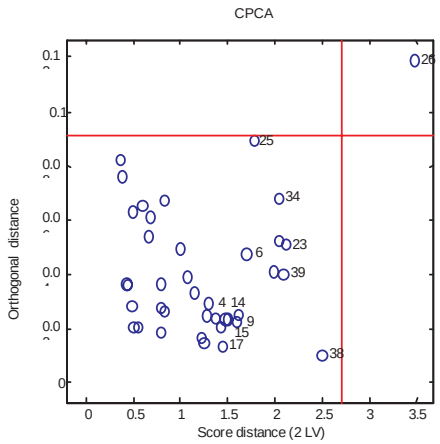


Figure 5. Diagnostic plot of octane data set based on two CPCA Principal Components

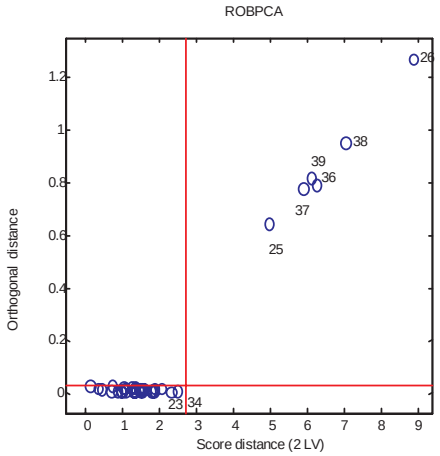


Figure 6. Diagnostic plot of octane data set based on two ROBPCA Principal Components

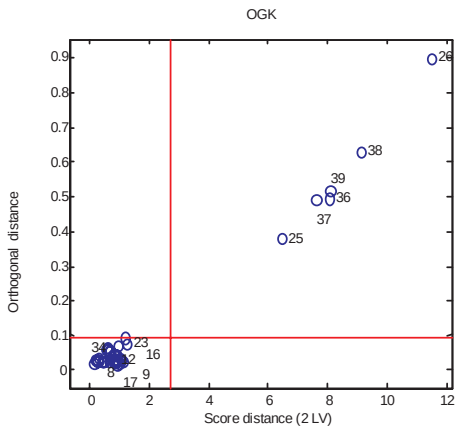


Figure 7. Diagnostic plot of octane data set based on two OGK PCA Principal Components

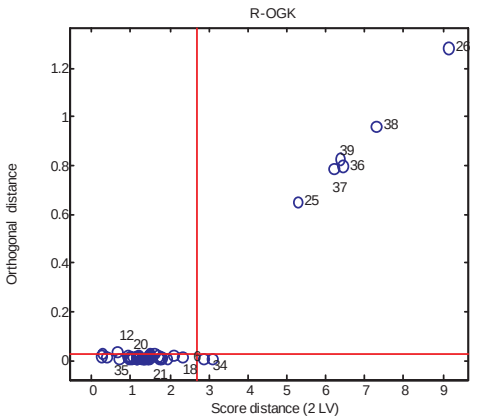


Figure 8. Diagnostic plot of octane data set based on two R-OGK PCA Principal Components

Contrary to the car data, this time OGK on high dimensional data showed similarity to ROBPCA and R-OGK.

6.2 Simulation

In this section a simulation study is performed to compare the performances of CPCA, ROBPCA, OGK and R-OGK on low and high dimensional data sets.

While generating the data, the following contaminated model construction is used

$$(1-\varepsilon)N_p(0,\Sigma)+\varepsilon N_p(\tilde{\mu},\tilde{\Sigma})$$

with different values for epsilon and different sizes of the data matrix. $\tilde{\mu}$ represents the center of outliers and is adjusted to obtain bad leverage points as will become apparent in the following.

For each setting 100 data sets were constructed and two different assessment criteria, MAXSUB and MSE, were used to gain insights about their performance. MAXSUB is the maximal angle between the space spanned by the estimated principal components and E_k, where E_k is the subspace spanned by the *k* dominant eigenvectors of Σ.

The MAXSUB measure is defined as (Hubert et al, 2005) $I'_{k,p}P_{p,k}P'_{k,p}I_{p,k}$ $\text{MAXSUB}=\arccos(\sqrt{\lambda_k})$ where λ_k is the smallest eigenvalue of $I'_{k,p}P_{p,k}P'_{k,p}I_{p,k}$. This gives the largest angle between a vector in E_k and the vector most parallel to it in the estimated PCA subspace. MAXSUB provides the best values when it is close to 0.

The second criterion, MSE, is the mean squared error of *k* largest eigenvalues and defined as:

$$\text{MSE}(\hat{\lambda}_j)=\frac{1}{100}\sum_{l=1}^{100}(\hat{\lambda}_j^{(l)}-\lambda_j)^2$$

Due to their lacking the orthogonal equivariance property, the performance of OGK and R-OGK estimators has also been evaluated on a rotated data matrix which has been obtained by multiplying the original data matrix with an orthogonal matrix.

6.2.1 Simulation study when $\varepsilon=0.20$ and $\varepsilon=0.10$ in low dimension

These are the assigned values of parameters that used for generating low dimensional settings:
n=150, p=5, $\Sigma=\text{diag}(12,8,6,0.20,0.05)$, k=3. It has been decided to assign a value of 3 to k, because three components explain 99% of the data ($(\sum_{i=1}^{3}\lambda_i)/(\sum_{i=1}^{5}\lambda_i)=0.9905$).

As can be seen from Figure 9 and Figure 10 the worst MAXSUB value pertains to CPCA; it is close to 1 when 20% contamination is added to the data. ROBPCA gives the best result and R-OGK pursuits ROBPCA. The most striking result here is that R-OGK is much more equivariant than OGK after rotation, the estimations of R-OGK on rotated data matrix are approximately equivariant. In case of a 10% contamination of the data, OGK is very much in line with the other estimators. Tables 3, 4 give exact values of MAXSUB.

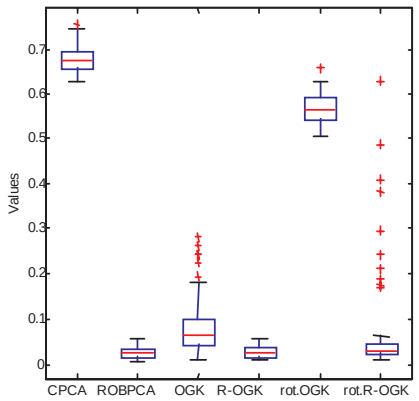


Figure 9. Boxplots of 20% contaminated low dimensional dataset based on MAXSUB values

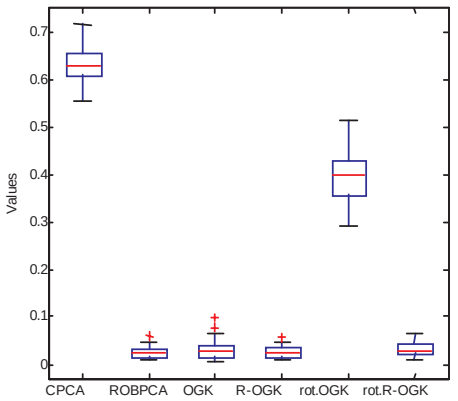


Figure 10. Boxplots of 10% contaminated low dimensional dataset based on MAXSUB values

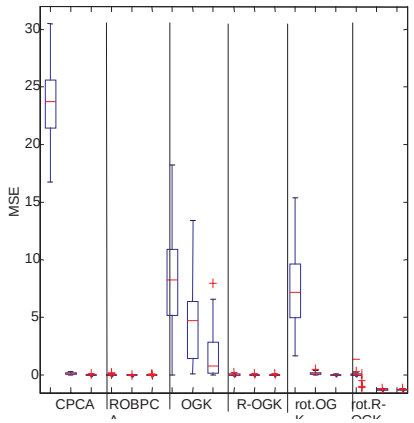


Figure 11. Boxplot of 20% contaminated low dimensional data set based on MSE of eigenvalues

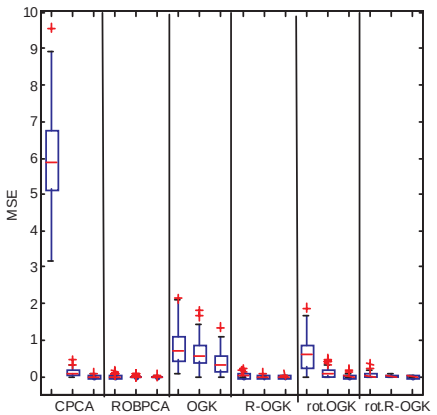


Figure 12. Boxplot of 10% contaminated low dimensional data set based on MSE of eigenvalues

Figures 11 and 12 just enable to evaluate the first eigenvalues but Tables 7, 8 provide detailed information for the MSE of three eigenvalues from which it becomes evident that ROBCPA gives the best results and R-OGK is next in ranking.

6.2.2 Simulation study when $\varepsilon = 0.20$ and $\varepsilon = 0.10$ in high dimension

In high dimensional simulation studies the following parameter values were used: p=100, n=50, $\tilde{\mu} = (6,8,10,12,14,16,0,0....0)$, $\Sigma = (12,8,6,5,3,0.1,0.099,0.098.....0.006)$ and k=5. The first five eigenvalues explain 87% of the data $((\sum_{i=1}^{5} \lambda_i)/(\sum_{i=1}^{100} \lambda_i) = 0.8710)$.

For high dimensional data, MAXSUB values of OGK give surprising results which are evident from Figure 13 and Figure 14. Contrary to what is deduced from the MAXSUB values, the MSE of the eigenvalues indicates that OGK fails like CPCA. R-OGK and ROBPCA, however, give similar and best results for both criteria (Figure15, 16).

Based on the MAXSUB values, OGK on rotated data matrix breaks down. This is in contrast to the MSE of the eigenvalues which tells that the worst outcome is pertained with OGK (See, Appendix Table 5, 6 and Table 9, 10). So there is a serious contradiction between the MAXSUB and MSE values for OGK. When the results of two criteria (MAXSUB and MSE of eigenvalues) are compared, it has to be noticed that except OGK and OGK on rotated X, all the other estimators give coherent results with each other.

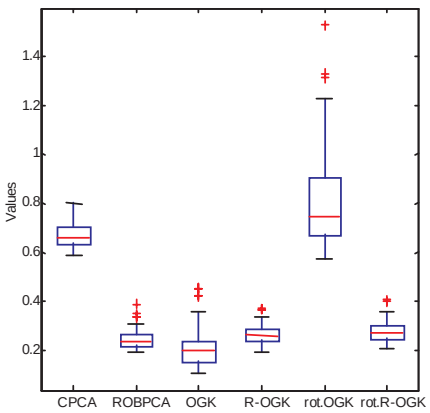


Figure 13. Boxplots of 20% contaminated high dimensional dataset based on MAXSUB values

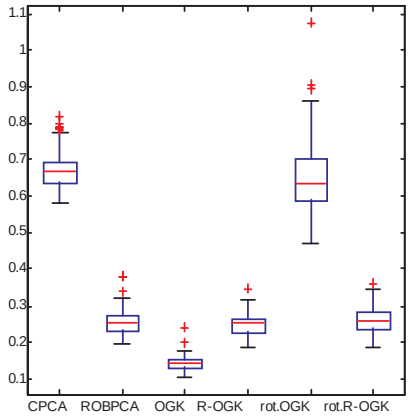


Figure 14. Boxplots of 10% contaminated high dimensional dataset based on MAXSUB values

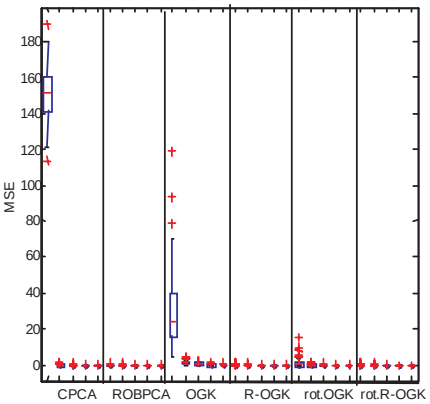


Figure 15. Boxplot of 20% contaminated high dimensional data set based on MSE of eigenvalues

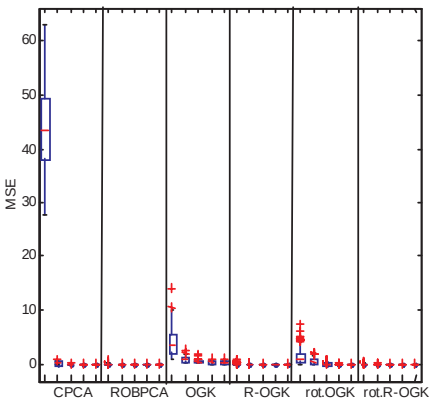


Figure 16. Boxplot of 10% contaminated high dimensional data set based on MSE of eigenvalues

6.2.3 Simulation study when $\varepsilon = 0$ in high and low dimension

For uncontaminated data in the high and low dimensional case CPCA and OGK give the best results, with OGK even performing slightly better than CPCA. Although the MAXSUB results show very similar performances with respect to those of the ROBPCA and the R-OGK estimators, the MSE results indicate that ROBPCA is better. ROBPCA and R-OGK yield higher MAXSUB and MSE values in comparison with lower dimension. OGK and R-OGK estimates on rotated data matrix do not perform extremely different from the original data matrix. The visual and numerical illustrations are provided in the tables (See Appendix-1, Table 1, 2) and below figure 17, 18, and 19, 20.

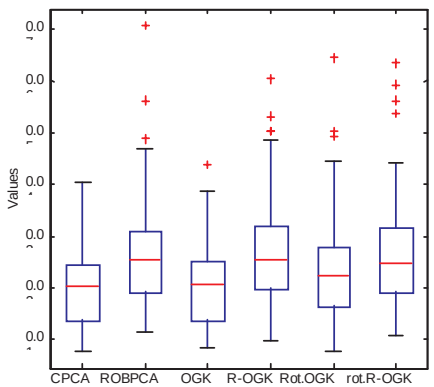


Figure 17. Boxplots of uncontaminated low dimensional dataset based on MAXSUB values

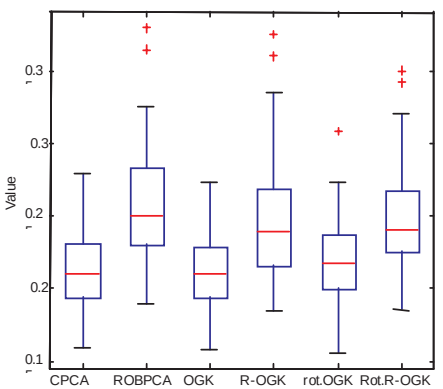


Figure 18. Boxplots of uncontaminated high dimensional dataset based on MAXSUB values

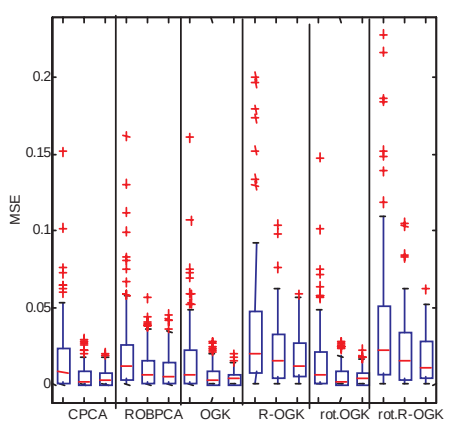


Figure 19. Boxplot of uncontaminated low dimensional data set based on MSE of eigenvalues

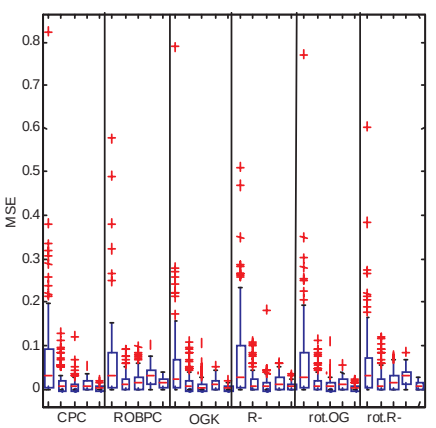


Figure 20. Boxplot of uncontaminated high dimensional data set based on MSE of eigenvalues

7. CONCLUSION

As a general result of simulation it can be said that, when there is contamination in the data, ROBPCA and R-OGK give very similar results, they are both superior to CPCA and OGK but in low dimension ROBPCA slightly comes into prominence whereas in high dimension R-OGK comes into prominence. So, when high dimension is the subject, it can be preferred to use R-OGK since it's computationally easier than ROBPCA. Furthermore, compared to OGK, R-OGK is more equivariant.

Nevertheless when there is no contamination in the data, CPCA and OGK yield best results. In this case inequivariance of OGK does not seem to be an important issue.

Another point, which should be stressed here, is that OGK shows the worst performance of robust estimators in contaminated data sets according to MSE criteria. But in contrast to MSE values, MAXSUB values specify the OGK estimator surprisingly as the best estimator especially in high dimensional data sets. The striking but inevitably incoherent differences between MSE and MAXSUB values of OGK can be seen in appendix–1.

8. REFERENCES

Alqallaf F.A, Konis K.P., Martin R.D. and Zamar R.H., 2002. Scalable robust covariance and correlation estimates for data mining. Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining, 14-23.

Campbell N.A., 1980. Robust procedures in multivariate analysis I: Robust covariance estimation, Applied Statistics, 29, 3, 231-237

Croux, C. and Haesbroeck, G., 2000. Principal component analysis based on robust estimators of the covariance or correlation matrix: influence functions and efficiencies. Biometrika, 87, 603-618.

Gnanadesikan, R., and Kettenring, J.R., 1972. Robust estimates, residuals, and outlier detection with multiresponse data. Biometrics, 28, 81-124.

Huber,P.J.,1981. Robust statistics, John Wiley&Sons, New York.

Hubert, M., Rousseeuw, P.J., and Verboven, S., 2002. A fast method for robust principal components with applications to chemometrics, Chemometrics and Intelligent Laboratory Systems, 60, 101–111.

Hubert M., Rousseeuw P. J., and Vanden Branden K., 2005. ROBPCA: A new approach to robust principal components analysis. Technometrics, 47:64–79.

Li, G., Chen, Z., 1985. Projection-pursuit approach to robust dispersion matrices and principal components: Primary theory and Monte Carlo. J. Amer. Statist. Ass. 80, 759-766.

Maronna, R.A,.1976. Robust M-estimators of multivariate location and scatter, Ann. Stat., 4, 51-67.

Maronna, R.A. and Yohai, V. J., 1995. The Behavior of the Stahel-Donoho robust multivariate estimator. J. Amer. Statist. Assoc. 90, 330-341.

Maronna R.A. and Zamar R.H., 2002. Robust estimates of location and dispersion for high-dimensional data sets. Technometrics, 44, 307-314.

Rousseeuw P.J. and Leroy A. M., 1987. Robust regression and outlier detection. Wiley-Interscience, New York.

Rousseeuw P.J. and Van Driessen K.,1999. A fast algorithm for the minimum covariance determinant estimator. Technometrics, 41, 212–223.

Stahel W., 1981. Breakdown of covariance estimators, Research Report 31, ETH Zurich,fachgruppe fuer Statistik.

GÜÇLÜ İKİLİ KOVARYANS TAHMİNCİSİNİN PERFORMANS DEĞERLENDİRMESİ

ÖZET

Yüksek boyutlu veri kümelerinde aykırı gözlemlerin varlığı halinde, çok değişkenli analiz ve çok boyutlu aykırı gözlem teşhis teknikleri, zamanı etkin kullanan, kırılma noktası yüksek güçlü kovaryans tahmincilerin kullanımını zorunlu kılar. Klasik tahmincilerin aykırı gözlemler karşısında bozulması, güçlü tahmincilerin kullanımnı gerektirir. FMCD kırılma noktası yüksek, yüksek boyutlu verilerde kullanımı uygun olan bir tahmincidir, fakat Maronna ve Zamar (2002), gözlem sayısının artmasıyla FMCD'nin önemli zaman aldığını ve yüksek kırılma noktasına sahip olma özelliğini yitirdiğini vurgular. OGK tahmincisi, yüksek kırılma noktasına sahip güçlü tahmincilerin işlem süresinin uzunluğu problemine yanıt vermek için (Maronna, Zamar, 2002) önerilmiştir. Bu çalışmada OGK tahmincisi ile çeşitli kovaryans tahmincilerinin performansı Temel Bileşenler Analizi (TBA) ile değerlendirilmiştir.

Anahtar Kelimeler: Aykırı gözlemler, Güçlü temel bileşenler analizi, Minimum kovaryans Determinat tahmincisi, Ortogonal Gnanadesikan-Kettering tahmincisi, Temel bileşenler analizi.

Appendix-1

Table 1. Simulation results of MAXSUB when there is no contamination in low dimension

	Mean	Median	Error
CPCA	0.021	0.0204	8.04E-04
ROBPCA	0.0266	0.0251	0.001
OGK	0.0209	0.0206	7.78E-04
R-OGK	0.0264	0.0253	9.42E-04
rot.OGK	0.0232	0.0222	9.55E-04
rot.R-OGK	0.0264	0.0246	9.85E-04

Table 2. Simulation results of MAXSUB when there is no contamination in high dimension

	Mean	Median	Error
CPCA	0.2136	0.2103	0.0024
ROBPCA	0.2569	0.2503	0.0035
OGK	0.2111	0.2095	0.0024
ROGK	0.2457	0.2390	0.0037
Rot.OGK	0.2190	0.2171	0.0026
rot.R-OGK	0.2480	0.2410	0.0034

Table 3. Simulation results of MAXSUB when there is 20% contamination in low dimension

	Mean	Median	Error
CPCA	0.6760	0.6732	0.0025
ROBPCA	0.0263	0.0241	0.0011
OGK	0.0809	0.0641	0.0057
R-OGK	0.0282	0.0257	0.0012
Rot.OGK	0.5669	0.5635	0.0031
Rot.R-OGK	0.0598	0.0310	0.0099

Table 4. Simulation results of MAXSUB when there is 10% contamination in low dimension

	Mean	Median	Error
CPCA	0.6348	0.6297	0.0033
ROBPCA	0.0259	0.0244	9.4153e-004
OGK	0.0300	0.0283	0.0016
R-OGK	0.0266	0.0255	0.0010
Rot.OGK	0.3977	0.4015	0.0051
Rot.R-OGK	0.0319	0.0297	0.0011

Table 5. Simulation results of MAXSUB when there is 20% contamination in high dimension

	Mean	Median	Error
CPCA	0.6736	0.6604	0.0049
ROBPCA	0.2471	0.2413	0.0034
OGK	0.2131	0.1983	0.0074
ROGK	0.2677	0.2638	0.0036
rot.OGK	0.8036	0.7473	0.0178
rot.R-OGK	0.2797	0.2701	0.0046

Table 6. Simulation results of MAXSUB when there is 10% contamination in high dimension

	Mean	Median	Error
CPCA	0.6719	0.6675	0.0050
ROBPCA	0.2567	0.2517	0.0035
OGK	0.1441	0.1437	0.0019
rOGK	0.2508	0.2514	0.0030
rot.OGK	0.6530	0.6356	0.0098
rot.R-OGK	0.2629	0.2575	0.0036

Table 7. Simulation results for MSE of eigenvalues when there is 20% contamination in low dimensional data set

	Mean			Median			Error		
CPCA	23.6913	0.1231	0.0208	23.7911	0.1120	0.0134	0.2771	0.0078	0.0022
ROBPCA	0.0219	0.0081	0.0059	0.0080	0.0046	0.0028	0.0033	0.0010	0.0008
OGK	8.1342	4.5081	1.593	8.2647	4.75	0.7506	0.397	0.311	0.1845
R-OGK	0.0416	0.0169	0.0139	0.0302	0.0096	0.0093	0.0042	0.0024	0.0014
rot.OGK	7.3213	0.1334	0.0181	7.1625	0.1200	0.0106	0.2825	0.0102	0.0018
rot.R-OGK	0.0708	0.0301	0.0163	0.0326	0.0210	0.0127	0.0160	0.0027	0.0014

Table 8. Simulation results for MSE of eigenvalues when there is 10% contamination in low dimensional data set

	Mean			Median			Error		
CPCA	5.9608	0.1181	0.0191	5.8622	0.0985	0.0120	0.1304	0.0082	0.0022
ROBPCA	0.0184	0.0102	0.0078	0.0089	0.0049	0.0034	0.0026	0.0013	0.0009
OGK	0.8031	0.6472	0.3513	0.6913	0.5975	0.3104	0.0482	0.0357	0.0268
R-OGK	0.0453	0.0198	0.0147	0.0279	0.0125	0.0083	0.0049	0.0021	0.0014
Rot.OGK	0.6555	0.1342	0.0400	0.6208	0.1129	0.0199	0.0429	0.0099	0.0042
Rot.R-OGK	0.0619	0.0399	0.0260	0.0383	0.0313	0.0229	0.0065	0.0033	0.0019

Table 9. Simulation results for MSE of eigenvalues when there is 20% contamination in high dimensional data set

	Mean					Median					Error				
CPCA	151,618	0,249	0,052	0,007	0,006	150,766	0,169	0,03	0,003	0,003	1,498	0,023	0,006	0,001	0,0007
ROBPCA	0,0948	0,029	0,008	0,013	0,004	0,0208	0,012	0,004	0,009	0,003	0,015	0,004	0,001	0,001	0,0006
OGK	29,8717	1,388	0,769	0,394	0,213	23,4209	1,194	0,716	0,35	0,171	2,086	0,072	0,04	0,022	0,0147
ROGK	0,0801	0,028	0,01	0,018	0,006	0,0267	0,011	0,005	0,014	0,004	0,013	0,005	0,001	0,002	0,0007
rot,OGK	1,3674	0,306	0,103	0,031	0,029	0,4098	0,245	0,071	0,021	0,021	0,231	0,025	0,009	0,003	0,0027
rot,rOGK	0,0515	0,031	0,016	0,025	0,009	0,0262	0,016	0,011	0,021	0,006	0,007	0,006	0,002	0,002	0,0009

Table 10. Simulation results for MSE of eigenvalues when there is 10% contamination in high dimensional data set

	Mean					Median					Error				
CPCA	43,677	0,243	0,045	0,008	0,006	43,426	0,183	0,029	0,003	0,003	0,768	0,02	0,006	0,001	0,0008
ROBPCA	0,0652	0,02	0,012	0,023	0,009	0,0235	0,012	0,006	0,018	0,006	0,009	0,002	0,001	0,002	0,0009
OGK	4,0256	0,906	0,526	0,318	0,256	3,4681	0,848	0,491	0,305	0,23	0,242	0,046	0,026	0,015	0,0166
ROGK	0,0807	0,02	0,01	0,017	0,006	0,0284	0,013	0,006	0,011	0,005	0,012	0,002	0,001	0,002	0,0006
rot,OGK	1,4375	0,493	0,115	0,023	0,008	0,9977	0,348	0,068	0,01	0,003	0,147	0,045	0,015	0,003	0,0014
rot,rOGK	0,0615	0,024	0,017	0,028	0,008	0,0282	0,014	0,012	0,023	0,006	0,008	0,003	0,002	0,002	0,0007

Appendix-2

MATLAB CODE NOTES

```
function [var,mu]=deviation(x)
med=median(x);
md=mad(x);      # Here, it is also possible to use 'madc' function instead of 'mad'
s=size(x);
Median=(ones(s(1),1))*med;
Mad=(ones(s(1),1))*md;
W=(x-Median)./(Mad);
W=(1 - (W./4.5).^2).^2.*(abs(W)<=4.5);
mu=sum(x.*W)./sum(W);
Mu=(ones(s(1),1))*mu;
rho=((x-Mu)./Mad).^2;                    # First and second steps of the  algorithm
var=((md.^2).*(sum(min(rho,9))))/s(1);
function result =ogk(x)
s=size(x);
[var1,mu1]=deviation(x);
D=diag(sqrt(var1));
y=(inv(D)*x')';                          #Third step of the algorithm
vv=combntns(1:s(2),2);
ss=size(vv);
for i=1:ss(1)
bb{i}=y(:,vv(i,:));
end
for i=1:ss(1)
t(:,i)=bb{i}(:,1)+bb{i}(:,2);
tt(:,i)=bb{i}(:,1)-bb{i}(:,2);
end
[var2,mu2]=deviation(t);
[var3,mu3]=deviation(tt);
U=(var2-var3)/4;                 #Fourth step of the algorithm
UU=zeros(s(2));
for i=1:ss(1)
UU(vv(i,1,:),vv(i,2,:))=U(i);
end
UU=eye(s(2))+UU+UU';
[E,T]=eig(UU);
A=D*E;
z=E'*y';
Z=z';                                    # Fifth step of the algorithm
[var,mu]=deviation(Z);
RO=diag(var);
 v=A*RO*A';
 m=A*mu';
d=sum((((Z-(ones(s(1),1)*mu))./sqrt(ones(s(1),1)*var)).^2)');    # Sixth step,        OGK
estimators
do=(chi2inv(0.9,s(2))*median(d))/chi2inv(0.50,s(2));
 w=((d<=do)*1);
 rm=(x'*w')/sum(w);
 dif=x-(ones(s(1),1)*rm');
 rv=(dif'*(diag(w))*dif)/sum(w);
 result=struct('m',{m},'v',{v},'rm',{rm},'rv',{rv});        #Seventh step, R-OGK estimators
```

Cilt: 06 Sayı: 01 Kategori: 02 Sayfa: 38-51 ISSN No: 1303-6319 Volume: 06 Number: 01 Category: 02 Page: 38-51

OUTLIER DETECTION IN MULTIPLE REGRESSION MODELS USING GENETIC ALGORITHMS AND BAYESIAN INFORMATION CRITERIA

Özlem GÜRÜNLÜ ALMA* **Serdar KURT****
Aybars UĞUR***

ABSTRACT

Statistical models, particularly regression models, are most useful devices for extracting and understanding the essential features of datasets. However, most of the databases in real-world include a particular amount of abnormal values, generally termed as outliers. An accurate identification of outliers plays a significant role in statistical analysis especially regression models. Nevertheless, many classical statistical models are blindly applied to data sets containing outliers, the results can be misleading at best. The appearance of outliers can exert negative influences on the fit of the multiple regression models. The aim of this study is to define outlier detection method using Genetic Algorithms (GA) with Bayesian Information Criterion (BIC) and to illustrate the algorithm with real and simulation data. We use a fitness function which is based on BIC in this algorithm. The criteria's value indicates a better model to fit data, the presence of one or more outliers will negatively impact the regression model and result in larger BIC values.

Keywords: Bayesian information criterion, Genetic algorithms, Multiple regression models, Outlier detection.

1. INTRODUCTION

According to Barnett and Lewis (1994), an outlier is one that appears to deviate so much from other observations of the sample. There are several statistical methods for outlier detection in different conditions. However, it may be difficult to decide which methods can be used in practical work. And if outliers are detected in the data, there are different ways of taking them into account in the analysis. For example, one can either remove the outlying observations from the data or incorporate the detected outliers into the statistical model.

A typical approach of detecting outliers is to characterize what normal observations look like, and then to single out samples that deviate from these normal properties. Existing methods for outlier detection include methods that classify a data point based on a distance from the expected value approaches that use information theoretical principles, such as selecting the subset of data points that minimize the prediction error. Outlier classification based on Mahalanobis distance can work quite well, but tends to

* Research Assistant, Dokuz Eylül University, Department of Statistics, Tınaztepe Campus, İzmir. e-mail: ozlem.gurunlu@deu.edu.tr
** Professor Dr., Dokuz Eylül University, Department of Statistics, Tınaztepe Campus, İzmir. e-mail: serdar.kurt@deu.edu.tr
*** Assistant Prof. Dr., Ege University, Faculty of Engineering, Department of Computer Engineering, İzmir. e-mail: aybars.ugur@ege.edu.tr

require the setting of some threshold that defines whether a point is an outlier or not. This threshold value typically needs to be tuned manually beforehand in order to determine its empirically optimal value for the system. Information theoretical approaches, outlier may be detected active learning (Abe et al., 2006), clustering (Barnett and Lewis, 1994; Breitenbach and Grudic, 2005; MacOueen, 1967) or mixture models (Scott, 2005). These methods may require sampling, the setting of certain parameters such as the optimal k in k-means, and may not all lend themselves to a real time implementation. There exist also a large number of outlier detection methods in literature (Ben-Gal, 2005). Traditionally, these can be categorized into three approaches: the statistical approach, the distance-based approach and the density based approach. But many of them are limited by assumptions of a distribution or limited in being able to detect only single outlier. If there is a known distribution for the data, then using that distribution can aid in finding outliers. Often, a distribution is not known, or the experimenter does not want to make an assumption about a certain distribution (Amidan et al., 2005). In addition to the basic problem of outlier detection mentioned above, there are additional problems in outlier detection for practical work. Data sets with multiple outliers are subject to masking and swamping effects. Although not mathematically rigorous, the following definition gives an intuitive understanding for these effects (Ben-Gal, 2005; Davies and Gather 1993).

According to Acuna and Rodriguez (2005), masking effect is said that one outlier masks a second outlier, if the second outlier can be considered as an outlier only by itself, but not in the presence of the first outlier. Thus, after the deletion of the first outlier the second instance is emerged as an outlier. Swamping effect is said that one outlier swamps a second observation, if the latter can be considered as an outlier only under the presence of the first one. In other words, after the deletion of the first outlier the second observation becomes a non-outlying observation (Acuna and Rodriguez, 2005). Sequential detection of outliers may therefore be misleading, if the detection of one outlier causes the subsequent detection of other outliers to be defective, because of either swamping or masking, or even both. Identification of outliers in Multiple Linear Regression (MLR) models is not trivial, especially when exist several outliers in data. The classical identification method based on the sample mean or sample covariance matrix cannot always find them, because the classical mean and covariance matrix are themselves affected by outliers due to masking effects. Therefore, simultaneous outlier detection method is important issue and in this work it is considered in MLR models.

GA has been used for outlier detection and model selection of linear regression models or times series. Jann (2000) describes a GA for the detection of level shifts in a time series, the problems caused by change points are similar to those caused by outliers. Ishibuchi et al., (2001) were used GA for the feature selection in data mining and they give a lot of references about this literature. Additionally, the use of GA for outlier detection and variable selection can be found in (Tolvi, 2004).

In this work, we are interested with the problem of identifying outliers and detection of outliers in the dependent variable of MLR models using GA. A robust simultaneous procedure is investigated for identification of outliers using Bayesian information criteria (Kullback, 1996). The scalability of information criteria is considered with a real data and also by generating experimental data. We have shown the behavior of our approach for different sample sizes and different percentages of contaminated outliers by simulation. That is, the outliers were produced by adding a given amount to each

dependent variable. We also studied on the affects of Kappa coefficient which is a penalized value of Bayesian information criteria and obtained results for different values of it.

2. METHOD

As mentioned in the first section, outliers can be described as; given a set of n data points and k the expected number of outliers, find the top k outliers that are considerably different, inconsistent with the respect to the remaining data. The outlier detection problem can be viewed as two sub problems:

- which observations data can be considered as inconsistent or exceptional in a given data set,
- finding an efficient method to detection of outliers.

Based on the above sub problems, the purpose of this work is to investigate detection of outliers in MLR models based on GA and Bayesian information criteria, which are described in the next subsections.

2.1 Outlier Detection in Multiple Linear Regression

The purpose of regression analysis is to fit equations to observed variables. The MLR equation takes the following type:

$$\mathbf{Y} = \beta_0 + \beta_1 \mathbf{X}_1 + \beta_2 \mathbf{X}_2 + ... + \beta_k \mathbf{X}_k + \boldsymbol{\varepsilon} \quad (1)$$

$$\hat{\mathbf{Y}} = \hat{\beta}_0 + \hat{\beta}_1 \mathbf{X}_1 + \hat{\beta}_2 \mathbf{X}_2 + ... + \hat{\beta}_k \mathbf{X}_k \quad (2)$$

where:
$Y \in \Re^n$ is a response variable,
$\hat{Y}$ is the predicted value of the dependent variable,
$X_1, \ldots, X_k \in \Re^n$ are different explanatory variables,
β_0 is the intercept on the Y axis, and
$\beta_1, ..., \beta_k$ are the regression coefficients for each of the independent variables.

Ordinary Least Squares (OLS) remains the most often utilized regression coefficient estimation method. This method optimizes the fit of the model by minimizing the sum of the squared deviations between the actual and predicted Y values $\sum e^2 = \sum (Y - \hat{Y})^2$. Computing an intercept term and estimating a set of β coefficient is calculated by $\hat{\beta} = (X^T X)^{-1} X^T Y$. However, some researchers began to realize that real data usually do not completely satisfy the classical assumptions. These are for errors:

- normally distributed,
- have equal variance at all levels of the independent variables and
- uncorrelated with both the independent variables and with each other.

If outliers occur in the data, the errors can be thought to have a different distribution from normal. There are several possibilities, but perhaps the most intuitive one is the mixture model. We assume that the ε's in distinct cases are independent where,

$$\varepsilon \sim \begin{cases} N(0,\sigma^2), & (1-\pi) \\ N(0,K^2\sigma^2), & \pi \end{cases}$$

Here π is the probability of an outlier and K^2 is the variance inflation parameter. In practical works the data sets may have outliers. One outlying observation can destroy least squares estimation, resulting in parameter estimates that do not provide useful information for the majority of the data. Because of these reasons, the detection of outliers is important for multiple regression analysis.

In this work potential outliers can be incorporated into MLR model of equation (1) by the use of dummy variables. A dummy variable is N x 1 vector (N is the number of observations) that has a value of one for the outlier observation, and zero for all other observations. For example, we assume that the last observation is an outlier, then one dummy variable to be added to the model (2), and the independent variable could be below.

$$X_{N\times(k+1)} = \begin{bmatrix} x_{11} & x_{1k} & 0 \\ \dots & \dots & \dots \\ x_{N1} & x_{Nk} & 1 \end{bmatrix}$$

A dummy variable in this experimental study is equivalent to a detected outlier. The problem for outlier detection in MLR is to select of the best model. For this reason, the candidate MLR models have different combination of all possible dummy variables.

The BIC criteria will be used here for outlier detection. For MLR model with dummy variables the criterion can be calculated as,

$$BIC = \log(\hat{\sigma}^2) + m\log(N)/N \qquad (3)$$

where $\hat{\sigma}^2 = (e'e)/(N-k-1)$ is the estimated variance of regression model, and $m = 1+k+m_d$, the total number of parameters in the estimated model, consists of parameters for the constant, the k independent variables and the number of outlier dummies m_d. Generally a good model has small residuals, and few parameters, then it is chosen with the smallest value of BIC is preferred for outlier detection in multiple regression (Tolvi, 2004).

A problem in using the BIC for outlier detection is that by itself tends to include unnecessary outlier dummies. To circumvent this problem, a correction to the criterion is used. The corrected BIC takes into account the different nature of outlier dummies and other variables, and has a different penalty term for different variables. This takes the form of an extra penalty (κ) for the dummies. The corrected BIC, denoted BIC′ (Tolvi, 2004), is given by

$$BIC' = \log(\hat{\sigma}^2) + (1+k)\log(N)/N + \kappa m_d \log(N)/N, \quad (4)$$

where the Kappa $(\kappa > 1)$ is the extra penalty value given to outlier dummies. Simulation experiments are conducted to determine relevant different values of κ and true outlier detection.

2.2 A Genetic Algorithm for Outlier Detection

GA is a stochastic search technique that guides a population of solution towards an optimum using the principles of evolution and natural genetics. The algorithm starts with a randomly generated initial population consisting of sets of chromosomes that represent the solution of the problem. These are evaluated for the fitness function or one of the objective functions, and then selected according to their fitness (Bozdogan, 2004; Goldberg, 1989; Rothlauf, 2006). To perform its optimization like process, the GA employs three operators to propagate its population from one generation to another. The first operator is the selection operator, which mimics the principal of the survival of the fittest. The second operator is the crossover operator, which mimics mating in biological populations. It propagates features of good surviving designs from the current population into the future population, which will have better fitness value on average. The last operator is the mutation operator, which promotes diversity in population characteristics.

In this paper, for the given set of objects located in the space, GA was used to detect the outliers. There are five primary elements in the GA, and the parameter setting of GA was shown as following in details.

- **Parameter Encoding:** The coding of the candidate models for outlier detection is straightforward. Each model also called an individual or chromosome, is fully described by a binary vector "d", $d = (d_1,...,d_N)$, where $d_i = 0$ indicates no outlier dummy and $d_i = 1$ indicates an outlier dummy for observation i, for each $i = 1,...,N$. For example, a model with a dummy variable for the last observation is described by the vector $d = (0\ 0...1)$. These dummy variables for outlier observations must be created before the GA is run on a data set.

In this study, the structure of a chromosome or an individual is shown in Figure1. It has N genes which is the number of observations in data set. Each chromosome consists of p genes, where p is the number of outliers given in a model. For instance, if the second and N-1th observations are outliers in data, the chromosome structure will be as seen in Figure 1.

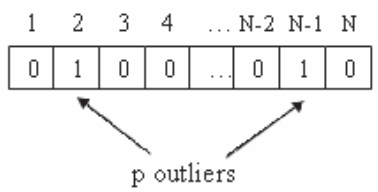


Figure 1. The structure of a chromosome in GA

• **Fitness Function:** The genes, which represent the serial number of outliers, are updated with each new population created. The random population is sorted based on the least fitness is considered to be the elite chromosome within population. The fitness of an individual is computed as the BIC' which is given equation 4 for MLR model with the corresponding dummy variables.

• **The Population and Generations:** The population size in each generation is 40 individuals. The initial population for the algorithm to start with is generated randomly. MLR models corresponding to these individuals are then estimated using the observed data, and BIC' values for them computed. The individuals with smallest values of the fitness function are more likely to pass their genes onto the next generation.

• **Selection Operator:** Stochastic uniform selection function is used in our GA. This function lays out a line in which each parent corresponds to a section of the line of length proportional to its scaled value. The algorithm moves along the line in steps of equal size. At each step, the algorithm allocates a parent from the section it lands on. The first step is a uniform random number less than the step size. It is noted that the results can be improved if a small number of the best individuals. These are kept the same from one generation to the next. In our GA the best two individuals are kept as elite population.

• **Crossover Operator:** The next generation of individuals from the previous one, is based on the BIC' values of the individuals. The best individuals has the smallest value of the fitness function BIC', are more likely to pass their genes onto the next generation. This procedure is repeated to create the same number of individuals as existed in the previous generation. Scattered crossover model is used in our approach and the crossover probability is defined as one. A crossover probability $p_c = 1$ indicates that crossover always occurs between any two parent models chosen from the mating pool; thus the next generation will consist only of offspring models, not of any model from the previous generation.

• **Mutation Operator:** Mating of the individuals from the previous one generation will not be enough for diversity of population. In evolutionary terms, more genetic variation in the population is needed. To this end, the individuals of each generation are also mutated before model estimation. Each gene of each individual is flipped, from zero to one or vice versa, with probability 0.01.

In addition to crossover and mutation, a condition for the maximum number of dummies is used to alter the population. This condition is used in order to keep the candidate models from having too many variables, because only a few dummies will be allowed in the final model. The rule states that if a candidate model has more than N/2 dummy variables, or outliers is more than 50% of the number of observations, it is dropped from consideration. Depending on the particular crossover and mutation rates, the second generation will be composed entirely of offspring models or of a mixture of offspring and parent models. In summary, the outline of the GA is shown in Figure 2.

1. **[Start]** Generate random population of N_C chromosomes. These are suitable solutions for the problem.
2. **[Fitness]** Evaluate the fitness of each chromosome in the population using BIC'.
3. **[New population]** Create a new population by repeating following steps until the new population is complete.
   (a) **[Selection]** Select two parent chromosomes from a population according to their fitness value BIC'. The better fitness, the bigger chance to be selected.
   (b) **[Crossover]** With a crossover probability cross over the parents to form a new offspring. If no crossover was performed, offspring is an exact copy of parents.
   (c) **[Mutation]** With a mutation probability mutate new offspring at each locus
   (d) **[Accepting]** Place new offspring in a new population.
4. **[Replace]** Use new generated population for a further run of algorithm and look for the minimum of the BIC'.
5. **[Test]** If the final condition is satisfied based on the BIC' stop, and return the best solution in current population
6. **[Loop]**Go to step 2.

Figure 2. The outline of the GA

In the approach, the number of outliers was specified firstly in dependent variable of MLR model, and a random population of chromosomes was created representing the solution space. Each chromosome of this random population represents a N observations in data set and each locus in the chromosome is a binary code indicating the outlier observation (1) or non-outlying observation (0) in data set. The GA proceeded to find the optimal solution as fitness function value (BIC') of each chromosome. The process continues one generation after another for a specified number of generations controlled by the researcher.

3. FINDING

A comprehensive performance study has been conducted to evaluate our algorithm. This algorithm was implemented in Matlab. We ran this algorithm on some real life data sets: Scottish Hill Racing and Stack Loss. These data sets demonstrated the effectiveness of our method against other algorithms. Data is generated for N=20, 30, 40, 50 and 100 observations and different number of outliers are inserted for each data set by taking into account of percentage of outliers in the dependent variable.

3.1 Experiments: Simultaneous Outlier Detection

In this paper, two experimental data sets have been used to illustrate outlier detection in MLR modeling. References to these, and other information, including where to obtain the data can be found in (Hoeting et al., 1996)[§]. In this subsection, it is investigated that detect outliers from these data sets with GA. Some information on the data sets and results are following;

[§] These data sets are available from one of the authors' website. This web address is http://www.stat.colostate.edu/~jah/index.html, access date: 30.04.2009

i. Scottish Hill Racing: The first example involves data supplied by Scottish Hill Runners Association (Atkinson, 1986). The purpose of the study is to investigate the relationship between record time of 35 hill races and two explanatory variables: distance is the total length of the race, measured in feet. One would expect that longer races and larger climbs would be associated with longer record times. Several authors have examined these data sets using both predictors in their analysis (Atkinson, 1986; Hadi, 1986; Hoeting et al., 1996). They concluded that races 7^{th} and 18^{th} observations are outliers. After they removed observations 7 and 18, their methods indicated that observation 33 is also an outlier. Thus observations 7 and 18 mask observation 33. After race numbers 7, 18, and 33 are removed from the data, standard diagnostic checking does not reveal any gross violations of the assumptions underlying MLR models (Fox, 1997; Hoaglin and Tukey, 1983; Hoeting et al.,1996). The scatter plot of this data set is shown in Figure 3.

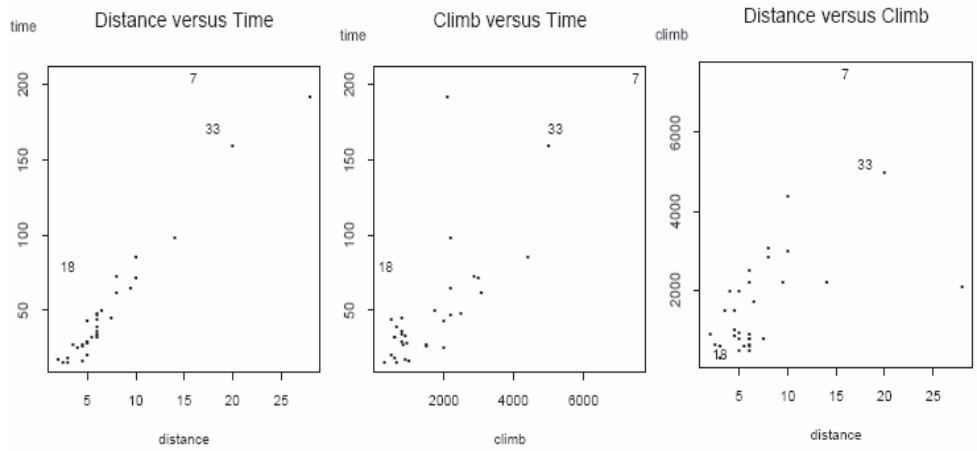


Figure 3. Scatter plot of Scottish hill racing data[**]

The GA described earlier was run many times with this data; all runs result in the same outliers being detected, at observations 7, 18, and 33. The solution was always found quickly by the GA. The estimated model and estimated variance with the three outlier dummies are

$$y = -8{,}45 + 6{,}63x_1 + 0{,}00661x_2 + 57{,}1d_1 + 64{,}6d_2 + 24{,}8d_3 \text{ and } \hat{\sigma}^2 = 22\,.$$

Then, the optimal fitness function value of GA has a BIC' value 4.30.

ii. The Stack Loss Data: The stack loss data consist of 21 days of operation form a plant for the oxidation of ammonia as a stage in the production of nitric acid. The response is called stack loss which is the percent of unconverted ammonia that escapes from the plant. There are three explanatory variables. The air flow is first independent variable which measures the rate of operation of the plant. The second independent variable measures the inlet temperature of cooling water circulating through coils in this tower and the last independent variable is proportional to the concentration of acid in

[**] Numbers correspond to race numbers 7, 18, 33. Distance is given in miles, time is given in minutes, and climb is given in feet.

the tower. Small values of the respond correspond to efficient absorption of the nitric oxides. In earlier research (Atkinson, 1986; Hoeting et al., 1996) been identified as outliers four observations. These are 1, 3, 4, and 21 observations. This data set provides an interesting extreme example of masking (Atkinson, 1986). The detection of any of these outliers is very difficult if only one observation at a time is examined. But the simultaneous methods are able to detect all of four outliers at a time.

The GA was run a lot of times with this data. The entire run gives to result in the same outliers being detected, at observations 1, 3, 4, and 21. The best outlier combination was always found quickly by the GA. The estimated regression model and variance with the four outlier dummies are

$$y = -37{,}7 + 0{,}798x_1 + 0{,}577x_2 - 0{,}0671x_3 + 6{,}22d_1 + 6{,}43d_2 + 8{,}17d_3 - 8{,}63d_4$$

and $\hat{\sigma}^2 = 1{,}57$. The optimal fitness function value of GA has a BIC' value 2.69.

3.2 Data Generation and Outlier Detection in MLR Models

In order to study the performance of the BIC' criterion and also the role of κ values for outlier detection, we conduct a simulation study. The conditions under which the simulation is performed are;

- the linear regression model is selected as $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i$,
- the first explanatory variable X_1 is generated from Normal (3,1), and the second explanatory variable X_2 is generated from Normal (2,1),
- the elements of $\beta_0 = 0$, and β_1, β_2 are generated from Uniform (1,2),
- the error terms are independent and identically distributed according to standard normal distribution N(0,1),
- the sample size N is determined as different sizes N=20, 30, 40, 50, and 100,
- percentages of outliers (P_O) in the dependent variable each of sample size are between %5-%10,
- the outliers are generated from the uniform distribution which lie at least 3σ from the mean of y_i and,
- the Kappa values are selected as $\kappa = 2, 3, 4, \text{and } 5$.

Under these conditions, firstly we simulate the explanatory variables and the error terms for i= 1,…,N observations and N=20, 30, 40, 50, 100. Then, we generate the response variable from y_i each of different sample size. After we generated y_i from normal distribution, we generated outlier observations from uniform distribution take into account of percentage of outliers. For example, for the sample size N=20 and percentage of outlier for the %5, it can be generated 1 outlier observation. However, two outliers must be added for the sample size 30 and 50 for the percentage of outlier 5, because of rounding problems.

Then, the percentage of outliers must be 6% for the sample sizes 30 and 50. The number of outliers for different sample sizes and the percentages are given in Table 1.

Table 1. Number of outliers for different sample size and percentage of outliers in dependent variable

	N				
P_O	20	30	40	50	100
5	1	2	2	3	5
10	2	3	4	5	10

Outliers are then added to the dependent variables. The iteration number for each combination of experiments is 100. Table 2 shows that the parameters of GA with BIC′ as the fitness function for the simulated models. The best models chosen most of the generations of GA can detect the outliers.

Table 2. The parameters of the GA for the simulated model

Sample Size of Simulation Data	**N=20, 30, 40, 50, 100**
Number of Generations	**250**
Population Size	**40**
Fitness Value	BIC′
Crossover Probability	**1**
Mutation Probability	**0.01**
Elitism	**For two parents**

The computational capacity in terms of the number of generations needed to find the true model is increased by an increase in the sample size. GA can simultaneously search in the solution space and find the outliers. The simulation results are shown in Table 3, where the value $T_{Outliers}$ is defined as total numbers of outliers in all iterations and $P_{Outliers}$ is defined as percentage of outliers in dependent variable finding with GA.

Table 3. Generating descriptions of data sets and total number of outliers found

Generating Data Sets Descriptions			Results Finding with GA κ							
			2		3		4		5	
N	$T_{Outliers}$	$P_{Outliers}$	$T_{Outliers}$	$P_{Outliers}$	$T_{Outliers}$	$P_{Outliers}$	$T_{Outliers}$	$P_{Outliers}$	$T_{Outliers}$	$P_{Outliers}$
20	100	5	283	14	105	5	105	5	105	5
30	200	6	296	10	202	6	202	6	202	6
40	200	5	383	10	215	5	215	5	215	5
50	300	6	323	7	300	6	300	6	300	6
100	500	5	683	7	502	5	502	5	502	5
20	200	10	285	14	210	10	210	10	210	10
30	300	10	410	14	302	10	302	10	302	10
40	400	10	660	17	414	10	414	10	414	10
50	500	10	536	11	505	10	505	10	505	10
100	1000	10	1207	12	1002	10	1002	10	1002	10

As seen in Table 3 the true results for experiments are obtained for values of $\kappa = 3, 4,$ and 5 for sample size is N=20, 30, 40, 50, 100, and percentage of outliers %5-10. A simulation study is carried out to support the good behavior of the BIC′ when different percentage of outlier and different sample size. It is clear that from simulation results for high values of Kappa coefficient ($\kappa > 2$) gives true information about how many observations are found as outlier. Therefore, we concluded that the best performing for outlier detection using BIC′ in MLR models is taken by the Kappa

coefficient is bigger than two. The important issue is that the BIC′criteria can not be affected masking or swamping effects finding outliers so we also said that this criteria is robust than other outlier detection methods.

In Figure 4, it is seen that the Kappa coefficient good results when the dependent variable Y containing of %5 outlier observation for all of sample size.

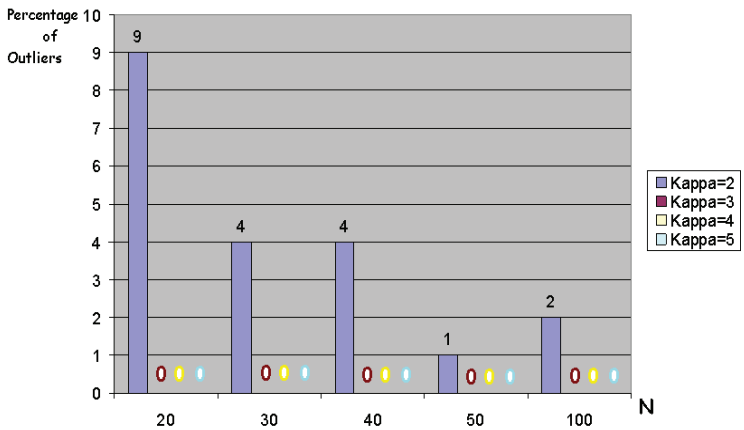


Figure 4. Results for percentage of outliers= 5%

Additionally, the same results for the dependent variable Y containing of %10 outlier observation for all of sample size can be observed from the simulation study. These results are shown in Figure 5.

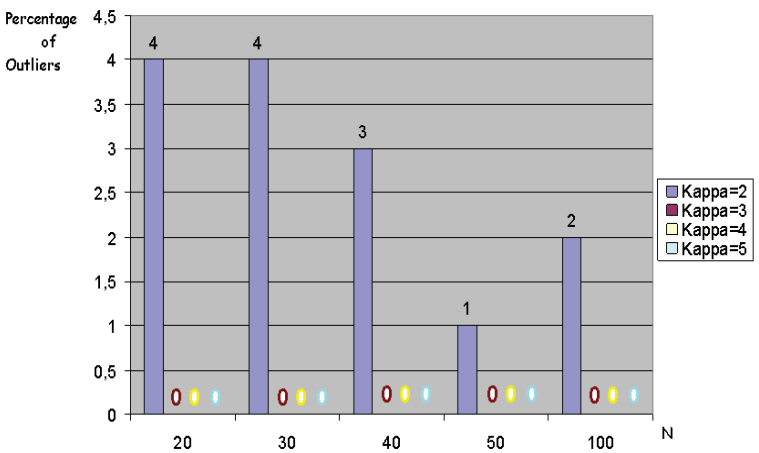


Figure 5. Results for percentage of outliers= 10%

Experiments with real and synthetic data sets show that the information criteria based on outlier detection method using GA in MLR models find the outlier automatically. We tested two types of scalability of the GA for outlier detection on data sets. The first one is the scalability of the GA against the given number of outliers and the second is the scalability against the power of different Kappa coefficients for a given sample size and number of outliers. Figure 4 and 5 show the results of using GA to find diversity number of outliers on data set. One important observation from these figure was that the

GA based on information criteria can be found accurately outliers especially the kappa coefficient bigger than two. The GA was also run with Kappa value bigger than 5 and the same result are obtained for $\kappa = 2,\ 3, 4$ and 5. Therefore the results have been the same with a wide range of Kappa values.

4. DISCUSSION AND RESULT

In this paper, it is demonstrated that Bayesian information criteria and developed a GA for outlier detection in MLR models. The value of BIC' is calculated for each observation as a measure of the fitness of dependent variable in MLR models using GA. GA can simultaneously search in the solution space and find the outliers. The main advantage of this method is that on does not have to bother the distribution of the observed residuals, which has proved to be complicated for the simple reason that the estimated residuals do not have a constant variance. Nevertheless, exact distributions for appropriate test statistics based on these adjusted residuals become intractable (Barnett and Lewis, 1994). The simulation results are shown in Table 3, especially Kappa coefficient ($\kappa > 2$) gives true information about how many observations are found as outlier. Hence, it is confident to claim that the GA based on BIC' criteria is suitable for MLR models.

We are working on comparing other applications of the GA for detection of outliers in MLR models as the future work.

5. REFERENCES

Abe, N., Zadronzy, B., and Langford, J., 2006. Outlier detection by active learning. ACM. Proceedings of the 12th ACM SIGKDD International conference on Knowledge Discovery and Data Mining, 767-772, New York, USA.

Acuna, E., and Rodriguez, C., 2005. On detection of outliers and their effect in supervised classification, http://academic.uprm.edu/~eacuna/vene31.pdf, 30 April 2008

Amidan, B., Ferryman, and T., Cooley S., 2005. Data outlier detection using the Chebyshew theorem. IEEE Aerospace Conference Proceedings, IEEE, Piscataway NJ USA, 3814-3819.

Atkinson, A.C., 1986. Influential observations, high leverage points, and outliers in linear regression. Statistical Science, 1, 397-402.

Barnett, V., and Lewis, T., 1994. Outliers in statistical data. John Wiley and Sons, USA.

Ben-Gal I., 2005. Outlier detection.,131-146. In: Maimon O. and Rokach L., Data mining and knowledge discovery handbook. Springer, USA.

Bozdogan, H., 2004. Statistical data mining and knowledge discovery. Chapman and Hall/CRC, USA.

Breitenbach, M., and Grudic, G.Z., 2005. Clustering through ranking on manifolds. Proceedings of the 22nd International Conference on Machine Learning, 73-80, New York, USA.

Davies L., and Gather U., 1993. The identification of multiple outliers. Journal of the American Statistical Association, 88, (423), 797-801.

Fox, J., 1997. Applied regression analysis, linear models and related methods. Sage Publication, USA.

Goldberg, D.E., 1989. Genetic algorithms in search, optimization, and machine learning. Addison-Wesley, USA.

Hadi, A., 1986. Influential observations, high leverage points, and outliers in linear regression. Journal of the American Statistical Association, Statistical Science, 1 (3), 379-393

Hoaglin, D., and Tukey, J., 1983. Understanding robust and exploratory data analysis. John Wiley and Sons, Canada

Hoeting, J., Raftery, A.E., and Madigan, D., 1996. A method for simultaneous variable selection and outlier identification in linear regression. Computational Statistics and Data Analysis, 22, 251-270.

Ishibuchi, H., Nakashima, T., and Nii, M., 2001. Genetic algorithm based instance and feature selection. In: Liu, H., and Motoda, H., Instance selection and construction for data mining, Kluwer Academic.

Jann, A., 2000. Multiple change point detection with a genetic algorithm. Soft Computing , 4, 68-75.

Kullback, S., 1996. Information theory and statistics. Dover Publications, USA.

MacQueen, J.B., 1967. Some methods for classification and analysis of multivariate observations. IProceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability, 1, 281-297.

Rothlauf, F., 2006. Representations for genetic and evolutionary algorithms. Springer, Netherlands.

Scott, D.W., 2005. Outlier detection and clustering by partial mixture modeling. Phsica-Verlag. In COMPSTAT 2004 Symposium, 453-465, Heidelberg.

Tolvi, J., 2004. Genetic algorithms for outlier detection and variable selection in linear regression models. Soft Computing, Springer, 527-533.

ÇOKLU REGRESYON MODELLERİNDE GENETİK ALGORİTMA VE BAYES BİLGİ KRİTERİ KULLANARAK SAPAN DEĞERLERİN BELİRLENMESİ

ÖZET

İstatistiksel modeller; özellikle regresyon modelleri, veri setlerinin önemli özelliklerinin anlaşılması ve ortaya çıkarılmasında en çok kullanılan araçlardandır. Bununla birlikte, gerçek hayatta birçok veri seti genellikle sapan değer olarak adlandırılan belirli miktardaki anormal değerler içerebilmektedir. Sapan değerlerin doğru bir şekilde tespit edilmesi, istatistiksel çözümlemelerde özelliklede regresyon modellerinde önemli bir rol oynar. Buna rağmen, birçok klasik istatistiksel modeller sapan değer içeren veri setlerine de uygulanmakta, nihayetinde de sonuçlar yanıltıcı olmaktadır. Sapan değerler, uygun olan çoklu regresyon modelinin belirlenmesini de güçleştirir.

Bu çalışmanın amacı, Genetik Algoritma (GA) ve Bayes Bilgi Kriteri (BIC) kullanarak sapan değer belirleme yöntemini tanımlamak ve algoritmayı gerçek ve benzetim verisi ile göstermektir. Genetik algoritmada BIC tabanlı uygunluk fonksiyonu kullanılmıştır. BIC değeri, veri için en uygun modeli göstermekte olup, bir veya daha çok sapan değerin varlığında regresyon modeli bu gözlemlerden olumsuz yönde etkilenecek ve daha büyük BIC değerli sonuçlar verecektir.

Anahtar Kelimeler: Bayes bilgi kriteri, Genetik algoritmalar, Çoklu regresyon modelleri, Sapan değer belirleme.

Volume: 06 Number: 01 Category: 02 Page: 52-64 ISSN No: 1303-6319 Cilt: 06 Sayı: 01 Kategori: 02 Sayfa: 52-64

HİZMET SEKTÖRÜNDE MALİ BAŞARISIZLIĞIN MODELLENMESİ

Özlem İLK* **Murat ÇİNKO****
Deniz AKINÇ*** **Didem PEKKURNAZ******

ÖZET

Ekonomik faliyetlerdeki değişimlerin takip edilmesi ve mali başarısızlığı tetikleyen faktörlerin saptanması, hem ülke ekonomisini hem de firmaların şahsi durumlarını değerlendirmesi açısından önemlidir. Bu çalışmada, Türkiye'de hizmetler sektöründe bulunan firmaların mali başarı olasılıklarının hesaplanması ve bu başarının yıllar içinde değişiminin gözlenmesi amaçlanmıştır. Bu amaçla, İstanbul Menkul Kıymetler Borsası'ndan alınan bilançolar incelenmiş, zaman içinde tekrarlı ölçümlerden oluşan bu karmaşık yapıdaki verilerin analizi için çok seviyeli 'Marjinalleştirilmiş Otoregresif Rastgele Etki Modelleri' (MTREM) kullanılmıştır. Bu modellerle, her şirketin mali başarı olasılıklarını hesaplamak, farklı alt gruplardaki şirketlerin başarılarını karşılaştırmak ve zaman içindeki değişimleri gözlemlemek mümkündür. Karşılaştırma amacıyla, sık kullanılan tek seviyeli lojistik regresyon modelleri de aynı veriye uygulanmıştır. Doğru sınıflandırma oranlarına bakıldığında, MTREM'in lojistik modellere üstünlüğü gözlenmiştir.

Anahtar Kelimeler: Hiyerarşik istatistiki modeller, Panel veri, Şirket rasyoları.

1. GİRİŞ

Panel veri, aynı bireyden birden fazla zamanda alınan ölçümlerden oluşur. Korelasyon yapısı, kayıp verilerin sıklığı gibi nedenlerle karmaşık yapıya sahip olan bu tip verilerin istatistiksel analizi zordur. Panel verilere, diğer bir çok bilim alanın yanında, ekonomi alanında da çok sık rastlanır. Örneğin, şirketlerin mali başarısızlığının tahmini bu şirketlerden zaman içinde alınan tekrarlı verilerle yapılabilir.

Yakın zamanda ekonomi alanındaki panel veri çalışmaları hız kazanmıştır. Mittal vd. (2005) Amerika'daki 77 firmadan topladıkları panel verinin analizi sonucunda müşteri memnuniyeti ile uzun vadeli finansal performansın arasında olumlu ilişki bulmuşlardır. Liao ve Gartner (2006) Amerikan Girişimcileri Panel Çalışması'nı ('U.S. Panel Study of Entrepreneurial Dynamics') kullanarak bu verilere lojistik regresyon metodları uygulamış ve iş planının varlığının, varsa zamanlamasının (erken veya geç plan yapılmasının) ve çevresel belirsizliklerin, yeni iş kuranların işlerini sürdürebilme olasılıklarına etkisini araştırmıştır. Yao vd. (2007) Çin'deki 22 bankadan toplanan 1995-2001 yılları arasındaki panel verilere regresyon metodları uygulayarak, bu ülkenin Dünya Ticaret Örgütü'ne girişinin bankalara etkisini incelemiştir.

* Öğr. Gör. Dr., Orta Doğu Teknik Üniversitesi, İstatistik Bölümü, 06531, Ankara, e-mail: oilk@metu.edu.tr
** Öğr. Gör. Dr., Marmara Üniversitesi, İngilizce İşletme Bölümü, Göztepe Kampusu, 34722, İstanbul, e-mail: mcinko@marmara.edu.tr
*** Arş. Gör., Orta Doğu Teknik Üniversitesi, İstatistik Bölümü, 06531, Ankara, e-mail: denizakinc@yahoo.com
**** Orta Doğu Teknik Üniversitesi, Ekonomi Bölümü, 06531, Ankara, e-mail: didempek@yahoo.com.tr

Mali başarısızlığın modellenmesi konusunda yayın taraması yapıldığı zaman, çoklu regresyon modeli, çoklu diskriminant analizi ve lojistik modellerle karşılaşılır (Altman, 1968; Ohlson, 1980; Aktaş, 1997). Aktaş (1997), modellerde bağımsız değişken olarak kullanılan mali oranların normal dağılım varsayımına genelde uymadığını belirtmiş ve lojistik regresyonun bu durumlarda diğer önerilen metodlara üstünlük sağladığını vurgulamıştır (sayfa 55 ve 77).

Türkiye'deki sektör bilançolarının istatistiksel modellenmesine günümüzde de ihtiyaç vardır. Bu modeller aracılığıyla, firmaların durumları hakkında hem incelenen zamanlarda hem de ileriye yönelik yorumlar yapılabilir. Bu amaç için şu ana kadar kullanılan istatistiki modellerin hepsi tek seviyeden oluşmaktadır. Halbuki, modellerin çok seviyeli olması; verilerin karmaşık yapısını dikkate alacak ve hesaplanan mali başarı olasılığı çok daha güvenilir olacaktır. Ayrıca, Türkiye'de şu ana kadar bu amaçla yapılan çalışmalar, panel veri yerine kesitsel veri bakış açısıyla yürütülmüştür.

Panel veri hem toplanması, hem de istatistiksel analizi açısından zor bir veri tipidir. Çalışmanın başlangıç tarihinde henüz kurulmamış firmalar veya çalışma henüz tamamlanmadan kapanan şirketler nedeniyle dengeli olmayan veri setleri oluşabilir. Kesitsel veriye kıyasla kayıp veriler daha sık karşımıza çıkar. Ayrıca, tekrarlı ölçümler nedeniyle temel istatistiksel metodlar için geçerli olan bağımsızlık varsayımı geçersizdir. Bu tip veri setinin barındırdığı zorluklar çalışmanın yöntem bölümünde daha detaylı olarak tartışılacaktır. Tüm bu zorluklara rağmen, kesitsel veriden farklı olarak, zaman içindeki değişimi ölçebilmesi nedeniyle sık tercih edilir.

Bu çalışmada, Türkiye'de hizmet sektöründe bulunan firmaların mali başarı olasılıklarının hesaplanması ve bu başarının yıllar içinde değişiminin gözlenmesi amaçlanmıştır. Bu amaçla, İstanbul Menkul Kıymetler Borsası (İMKB)'nin internet sayfasından toplanan panel verilerin istatistiksel analizleri, lojistik regresyon ve çok seviyeli modeller uygulanarak şirketlerin başarı tahminleri yapılmıştır. Çalışmanın ikinci amacı da, bahsedilen bu iki modelin sonuçlarının karşılaştırılmasıdır. Makalenin ikinci bölümünde kullanılan veri ve modeller hakkında ayrıntılı bilgi verilmektedir. Üçüncü bölümde modellerden elde edilen parametre tahminleri ve modellerin başarısı, dördüncü bölümde ise sonuçlar tartışılmıştır.

2. YÖNTEM

2.1 Veri Seti

Bu çalışmada, hizmetler sektöründe halka açık verisi bulunan 20 şirketten 1999-2002 yılları arasında toplanan veriler incelenmiştir. Bu 20 şirketin, 4 tanesi elektrik, 3 tanesi ulaşım, 8 tanesi ticaret ve diğer 5 tanesi de turizm alt sektörlerine bağlıdır. Gizlilik ilkesi çerçevesinde şirketlerin ismi bu makalede kullanılmamıştır. Türkiye'de iflas kavramı olmadığı için bu çalışmada başarısızlık tanımı De Andres vd. (2005)'lerinin çalışmasını takiben: Bir şirketin yıllık kâr oranı, o alt sektördeki tüm şirketlerin medyan kârı ile karşılaştırılmış, şirketin oranı bu medyandan büyük ise şirket başarılı, değil ise

başarısız olarak düşünülmüştür. Eldeki verilerden 5 kâr oranı (Özsermaye Kârlılığı Oranı, Aktif Kârlılık Oranı, Brüt Kâr Marjı Oranı, Net Kâr Marjı Oranı, Pay Başına Kâr) hesaplanmıştır. Bu oranlardan bazıları, diğer kâr oranları ile çok yüksek korelasyon katsayısına sahiptir. Kullanılan istatistiki modeller yüksek korelasyonla başa çıkabilecek yapıda olsalar da, 0,96 seviyelerine varan korelasyon katsayıları, bu oranların diğer oranlar tarafından açıklanabileceğini gösterdiği için modelden çıkarılmıştır. Özsermaye Kârlılığı Oranı ve Net Kâr Marjı Oranı'ndan elde edilen başarı göstergeleri bağımlı değişkenler olarak kullanılmıştır. Bağımsız değişkenler olarak 10 rasyo değeri, yıl ve bağımlı değişken göstergesi mevcuttur. Bağımsız değişkenlerin listesi ve açıklamaları Tablo 1'de verilmiştir. Bağımsız değişkenlerden likidite oranı, cari oran ve nakit oran arasında yüksek korelasyon problemi gözlendiği için cari oran ve nakit oran çalışmadan çıkarılmıştır. Çoklu korelasyon problemi gözlendiğinde, probleme neden olan bir veya bir kaç değişkenin modelden çıkarılması, problemin çözümlerinden birisi olarak tavsiye edilir (Neter vd. 1996, sayfa 410).

Tablo 1. Mevcut bağımsız değişkenler ve açıklamaları

Bağımsız Değişken	Açıklamalar
Likidite	Likidite oranı
Kaldıraç	Kaldıraç oranı
KVBTBO	Kısa Vadeli Borçların Toplam Borca Oranı
FKG	Faiz Karşılama Gücü
SDH	Stok Devir Hızı
Aktif BH	Aktif Büyüme Hızı
SBH	Satışların Büyüme Hızı
ÖBH	Özsermaye Büyüme Hızı
Net Kâr BH	Net Kâr Büyüme Hızı
Alt sektör	Alt sektör kodu
Değişken	Bağımlı değişken göstergesi (1= özsermaye kârı, 0= net kâr marjı)
Yıl	Zaman göstergesi (0= 2001 yılı, 1= 2002 yılı)

2.2 Veri Setinin Barındırdığı Zorluklar ve Olası Çözümler

Çalışmada kullanılan ve benzeri türde panel veri setlerinin barındırdığı başlıca zorluklar, karmaşık korelasyon yapısı, kayıp veriler ve veri setinin boyutudur. Bu zorluklara aşağıda kısaca değinilmiştir.

Birey içi korelasyon ve birden fazla bağımlı değişkenle ilgilenildiğinden bunlar arasında gözlemlenen korelasyon sorunları mevcuttur. Birey içi bağımlılığa örnek olarak, bir şirketin 2000 yılındaki başarısının, 1999 yılındaki başarısına bağımlı olması gösterilebilir. Bağımlı değişkenler arasındaki korelasyona ise, herhangi bir yıl içinde gözlemlenen özsermaye kârı ile net kâr marjının bağımlı olması örnek gösterilebilir. Bu çalışmada kullanılan çok seviyeli modeller bu iki tip korelasyon yapısını dikkate almaktadır.

Uzun zaman sürecinde toplanması nedeniyle çok sık karşılaşılan eksik verileri ele almak panel verilerin genel bir zorluğudur. Buna ek olarak, bu çalışmada, kanuni değişiklikler nedeniyle raporlamanın ve hesaplamaların sık sık değişmesi, bazı yıllarda yüksek oranda kayıp veri olmasına yol açmıştır (Tablo 2). Bu son sıkıntıyı aşmak için, çalışmada sadece 4 yıllık (1999-2002) veri kullanılmıştır. Bu 4 yıl içindeki kayıp veriler ise, uygun değerlerle değiştirilmiştir. Bu uygun değerlerin bulunması amacıyla, kayıp veri içeren değişkenlerin birleşik dağılımı, şartlı dağılımların çarpımı olarak yazılıp

uygun regresyon metodlarıyla tahmin edilmiştir. Bu sayede, sürekli ve kesikli dağılımdan gelen değişkenlerin birleşik ve çoklu dağılımlarını dikkate almak mümkündür (İbrahim vd., 2002). Bir başka deyişle, önce 1999 yılındaki kayıp veriler tahmin edilip bu yılda gözlenen ve tahmin edilen kayıp veriler kullanılarak 2000 yılı tahmin edilir ve bu şekilde tahmin işlemine devam edilir.

$$f(\underline{X}_{1999}, \underline{X}_{2000}, \underline{X}_{2001}, \underline{X}_{2002}) = f(\underline{X}_{1999}) f(\underline{X}_{2000} \mid \underline{X}_{1999}) f(\underline{X}_{2001} \mid \underline{X}_{2000}) f(\underline{X}_{2002} \mid \underline{X}_{2001})$$

Yukarıdaki denklemde, $\underline{X}_{1999}$ 1999 yılında analize katılan tüm bağımsız değişkenleri içeren bir vektördür. Bu vektörün elemanları da, yukarıda bahsedilen metotla parçalara ayrılabilir. Bir başka ifade ile, yıl bazında bağımsız değişkenlerin birleşik dağılımları, şartlı olasılıkların çarpımı olarak yazılabilir. Örneğin, 1999 yılında analizde *k* bağımsız değişken varsa, bunların birleşik dağılımı aşağıdaki gibi yazılabilir: $f(\underline{X}_{1999}) = f(X_{1999,1}) f(X_{1999,2} \mid X_{1999,1}) \cdots f(X_{1999,k} \mid X_{1999,1}, \cdots X_{1999,k-1})$

Bu denklemin sağ tarafındaki, her bir fonksiyon, ağırlıklı regresyonlarla modellenir. Ağırlıklı regresyonun amacı, bir değişkende gözlemlenen ekstrem değerlerin etkisini azaltmaktır.

Tablo 2. Yıllara ve değişkenlere göre kayıp veri yüzdesi

	Kârlılık Oranı	Cari Oran	Likidite Oran	Kaldıraç Oranı	KVBTBO	Faiz Karşılama Gücü	Nakit Oran	Stok Devir Hızı	Büyüme Hızları
1991	60	60	60	60	60	70	60	95	95
1992	60	60	60	60	60	70	60	60	60
1993	50	50	50	50	50	60	50	60	60
1994	50	50	50	50	50	65	50	50	50
1995	55	55	55	55	55	65	55	60	60
1996	40	40	40	40	40	50	40	50	50
1997	40	40	40	40	40	40	40	40	40
1998	40	40	40	40	40	40	40	40	40
1999	5	5	5	5	5	10	5	40	35
2000	5	5	5	5	5	10	5	10	5
2001	5	5	5	5	5	10	5	10	5
2002	5	5	5	5	5	10	5	10	5
2003	25	25	25	25	25	45	40	35	25

Kâr oranlarındaki kayıp yüzdesi birbiriyle aynı olduğu için sadece bir başlık altında (kârlılık oranı) verilmiştir. Aynı şekilde, kayıp yüzdeleri eşit olduğu için, büyüme hızları bir başlık altında rapor edilmiştir. Örneğin, 1991 yılında kârlılık değişkenlerinde, toplam 20 gözlemden 12'si kayıp olduğu için, (12/20)*100 %= 60%'lık kayıp rapor edilmiştir. KVBTBO: Kısa Vadeli Borçların Toplam Borca Oranı

Veri setinin küçüklüğü nedeniyle modele bazı bağımsız değişkenler eklenememiştir. Mevcut bağımsız değişkenler arasından Faiz Karşılama Gücü, Stok Devir Hızı, Net Kâr Büyüme Hızı, alt sektör ve yıl göstergeleri modellerde kullanılmamıştır. Genel olarak, kullanılan lojistik regresyon modellerindeki parametrelerin tahminleri sırasında yakınsamama problemleri yaşanabilir. Lojistik regresyon ve benzeri modellerde kullanılan doğrusal olmayan denklemlerin açık çözümleri olmadığı için, denklemler tekrarlı yöntemler aracılığıyla çözümlenir. Olabilirlik fonksiyonu yatay olan verilerde, bu tekrarlı yöntemler çözüm bulmakta zorlanır veya bulamaz. Bu tür problemler, küçük örneklem kümelerinde daha sık yaşanır.

2.3 Modeller

Şirketlerin başarı tahminleri için basit ve pratikte sık kullanılan lojistik modeller ve daha gelişmiş olan Marjinalleştirilmiş Otoregresif Rastgele Etki Modelleri (Marginalized Transition Random Effects Models, MTREM) adı verilen modeller (İlk, 2004; İlk ve Daniels, 2007) kullanılmıştır. MTREM üç seviyeden oluşan ve üç ana konuda yorum yapılmasını sağlayan bir modeldir. Bu üç ana konu; alt grupların başarı olasılıklarının karşılaştırılması, zaman içindeki değişimin gözlenmesi ve bireysel farklılıkları dikkate alarak başarı olasılıklarının hesaplanmasıdır.

MTREM şu üç seviyeden oluşmaktadır:

$$\text{logit } P(Y_{itj}=1) = X_{itj}\,\beta \quad (1)$$

$$\text{logit}P(Y_{itj}=1 \mid y_{i,t-1,j},...,y_{i,t-p,j}, X_{itj}) = \Delta_{itj} + \sum_{m=1}^{p} \gamma_{itj,m}\, y_{i,t-m,j} \quad (2)$$

$$\text{logit}P(Y_{itj}=1 \mid y_{i,t-1,j},...,y_{i,t-p,j}, X_{itj}, b_{it}) = \Delta^{*}_{itj} + \lambda_j b_{it} \quad (3)$$

Burada Y_{itj}, i bireyi için (i=1, ... ,n) t zamanında alınan (t=1, ... ,T) j. (j=1, ... ,J) bağımlı değişkendir ve X_{itj} aynı ölçüme karşılık gelen bağımsız değişken vektörüdür.

Görüldüğü üzere, modelin birinci seviyesi lojistik regresyondan ibarettir ve alt grupların başarı olasılıklarını karşılaştırma amacına yöneliktir. İkinci seviye, zaman içindeki değişimi ölçen AR (autoregressive) modelidir. Modelin üçüncü seviyesi ise, aynı zaman içinde, aynı bireyden alınan birden fazla bağımlı değişkeni birbirine bağlar. Bu seviyedeki, b_{it} değişkenlerinin normal dağılımdan geldikleri varsayılır; $b_{it} \sim N(0, \sigma_t^2)$. Aynı terim, $b_{it} = \sigma_t z_i$, $z_i \sim N(0,1)$ şeklinde de yazılabilir. Bu terim ölçülememiş veya gözlenemeyen faktörleri açıklamak için kullanılır. Saptanılabilirlik için $\lambda_1 = 1$ olarak tanımlanır.

Bu modeldeki parametre tahminleri, Bayesci metodlar kullanılarak gerçekleştirilmiştir. Markov Zinciri Monte Carlo ("Markov Chain Monte Carlo") metodları (Brooks, 1998) karmaşık istatistiki modellerde son zamanlarda çok sık kullanılan metodlardan biridir. Daha teknik bakış açısıyla, bu modelde parametreler Gibbs örneklemesi (Geman ve Geman, 1984) ve Hybrid MC (Neal, 1996) metodları kullanılarak oluşturulan bir algoritma sayesinde tahmin edilmiştir (İlk, 2004).

3. BULGULAR

3.1 Parametre Tahminleri

Rasyoların aynı ölçekte olmaması nedeniyle, veriler standardize edilerek modele dahil edilmiştir. Bir başka ifade ile, her bir gözlemden, o rasyonun ortalaması çıkarılıp, standart sapmasına bölünmüşür.

Yakınsama sağlanarak parametre tahminleri elde edilen modeller arasından birisi seçilmiştir. Bu modelin sonuçları Tablo 3-5 arasında verilmiştir. Bu seçim yapılırken, AIC (Akaike Bilgi Ölçütü) değerleri ve parametre tahminlerinin standart hatalarının mümkün olduğunca küçük olması ve doğru sınıflandırma oranlarının yüksek olması dikkate alınmıştır.

Başlangıç noktasındaki (1999 yılındaki) veriler kullanılarak, tek seviyeli lojistik regresyon modelleri ve üç seviyeli MTREM modelleri oluşturulmuş, bu modellerden elde edilen parametre tahminleri Tablo 3'te verilmiştir. Bu tabloda, MTREM modeli altında her bir parametre için %95 güven aralıkları, Bayes metodu ile elde edilen parametre örnekleminin medyanı, ortalaması ve standart hatası verilmiştir. Lojistik regresyonlar için de parametre tahminleri ve standart hatalar verilmiştir.

Tablo 3. Hizmetler sektörü 1999 yılı verileri için MTREM ve bağımsız lojistik regresyon modellerinin parametre tahminleri

			MTREM			LOJİSTİK		LOJİSTİK	
				posterior		Özsermaye Kârı		Net Kâr Marjı	
Yıl=1999	%2,5	%50	%97,5	Ort.	SH	Katsayı	SH	Katsayı	SH
Sabit	-0,53	0,9	2,38	0,9	0,74	2,36	1,65	0,54	0,83
Likidite	-0,93	1,08	3,19	1,13	1,06	1,59	1,91	2	1,41
Kaldıraç	-1,56	-0,19	1,08	-0,2	0,68	-2,32	1,71	1,56	1,33
KVBTBO	-1,88	-0,73	0,43	-0,73	0,59	-1,35	0,91	-0,23	0,7
Aktif BH	-3,15	-0,64	1,54	-0,7	1,21	-2,76	2,5	-0,65	1,36
SBH	-0,81	0,46	1,71	0,45	0,66	3,14	2,08	-1,01	0,93
ÖBH	-4,47	-1,57	1,05	-1,63	1,39	-1,92	2,09	-0,89	1,77
Değişken	-0,57	0,01	0,58	0,01	0,29				
$\log(\sigma_1^2)$	1,56	1,57	1,61	1,58	0,01				
λ_2^*	0,79	0,8	0,84	0,81	0,01				

Bu tablo 1999 yılındaki yıllık verileri kapsar. Tablodaki kısaltmaların açıklamaları aşağıda verilmiştir.
Ort.: Ortalama,
SH: Standart hata,
Likidite: Likidite oranı,
Kaldıraç: Kaldıraç oranı,
KVBTBO: Kısa Vadeli Borçların Toplam Borca Oranı,
Aktif BH: Aktif Büyüme Hızı,
SBH: Satışların Büyüme Hızı,
ÖBH: Özsermaye Büyüme Hızı,
Değişken: Bağımlı değişken göstergesi (1= özsermaye kârı, 0= net kâr marjı),
σ_1^2 birinci yıldaki varyans parametresidir.

Lojistik modeller tek bağımlı değişken üzerine kurulduğundan, özsermaye kârı ve net kâr marjı için bağımsız ayrı modeller kurulmuştur. MTREM ise çoklu değişkenler için geliştirildiğinden, bu iki değişkeni aynı anda modellemektedir.

Özsermaye kârı için 1999 yılı verileriyle kurulan lojistik regresyon model denklemi;
logit $\hat{P}(Y_{itj}=1)$= 2,36 +1,59 Likidite - 2,32 Kaldıraç - 1,35 KVBTBO -2,76 Aktif BH +3,14 SBH - 1,92 ÖBH
olarak verilebilir.

Net kâr marjı denklemi ise;
logit $\hat{P}(Y_{itj}=1)$= 0,54 + 2,00 Likidite + 1,56 Kaldıraç - 0,23 KVBTBO -0,65 Aktif BH - 1,01 SBH - 0,89 ÖBH
olarak bulunmuştur.

MTREM modelinin 1. seviyesindeki denklem ise;
logit $\hat{P}(Y_{itj}=1)$= 0,9 +1,08 Likidite - 0,19 Kaldıraç - 0,73 KVBTBO -0,64 Aktif BH + 0,46 SBH - 1,57 ÖBH + 0,01 Değişken
olarak verilebilir.

Modeller arasında, parametre tahminlerinde ciddi farklar olduğu gözlenmektedir. Örneğin, MTREM'le, kaldıraç oranının bir standart sapma boyutunda artması durumunda başarı odds'unun exp(-0,19)= 0,83 birim azalacağını söylerken, net kâr marjının lojistik regresyonla modellenmesi, bu odds'un exp(1,56)= 4,76 birim artacağını söylemektedir.

Tablo 4. Hizmetler sektörü 2000 yılı verileri için MTREM ve bağımsız lojistik regresyon modellerinin parametre tahminleri

	MTREM					LOJİSTİK Özsermaye Kârı		LOJİSTİK Net Kâr Marjı	
				posterior					
Yıl=2000	%2,5	%50	%97,5	Ort.	SH	Katsayı	SH	Katsayı	SH
Sabit	-2,42	-0,66	1,08	-0,66	0,99	0,18	0,58	0,22	0,53
Likidite	-0,36	1,29	3,44	1,4	0,95	1,72	1,27	0,71	0,83
Kaldıraç	-2,77	0,61	2,23	0,34	1,29	0,67	1	0,46	0,93
KVBTBO	-1,81	1,53	7,33	1,66	2,27	0,49	0,88	-0,14	0,74
Aktif BH	-1,63	-0,29	3,52	0,09	1,32	-0,91	0,66	-0,35	0,55
SBH	-5,52	1,04	2,91	-0,03	2,42	1,36	0,97	0,78	0,81
ÖBH	-4,9	-0,58	2,25	-0,98	2,08	-1,1	0,82	-1,22	1,07
Değişken	-0,81	0,07	0,8	0,06	0,4				
α_2	-1,39	0,67	2,8	0,66	1,07				
$\log(\sigma_2^2)$	1,27	1,4	1,43	1,38	0,05				
λ_2	1	1,05	1,07	1,04	0,02				

Bu tablo 2000 yılındaki yıllık verileri kapsar.

MTREM modeli denklemi dikkate alındığında, bağımsız lojistik modellerinden farklı olarak bir 'Değişken' teriminin olduğu görülmektedir. Bu terim 0 veya 1 değerlerini alan bir göstergedir. Sıfır değerini aldığı zaman, net kâr marjına karşılık gelen bağımlı değişkenle ilgilenildiğinde, MTREM denklemindeki katsayı 0,9 olarak kalır. Bir değerini aldığı zaman, özsermaye kârındaki başarı tahmin edildiğinde, katsayı 0,9 +

0,01=0,91 değerini alır. Başka bir deyişle, MTREM modelinde her iki bağımlı değişken için, iki farklı kesişim noktası varsayılmıştır. Benzer şekilde, enteraksiyon terimleri aracılığıyla iki farklı eğim varsaymakta mümkündür. Örneğin, finansal kaldıraç oranının özsermaye ve net kâr marjı için oldukça farklı parametre tahminleri verdiği görülmektedir (-2,32 ve 1,56). MTREM modeline Kaldıraç*Değişken terimi eklenirse, kaldıraçın net kâr marjı ve özsermaye kârı için farklı etkisi olduğu dikkate alınır. Ne varki, veri setinin küçüklüğü nedeniyle bu enteraksiyon terimi de eklenmemiştir.

Tablo 5. Hizmetler sektörü 2001 ve 2002 yılı verileri için MTREM ve bağımsız lojistik regresyon modellerinin parametre tahminleri

	MTREM			posterior		LOJISTIK Özsermaye Kârı		LOJISTIK Net Kâr Marjı	
YIL>2000	%2,5	%50	%97,5	Ort.	SH	Katsayı	SH	Katsayı	SH
Sabit	-0,45	0,2	1,04	0,22	0,37	0,54	0,44	-0,41	0,52
Likidite	-1,84	-0,16	0,89	-0,26	0,67	-0,15	0,59	-0,76	0,63
Kaldıraç	-1,38	-0,14	0,57	-0,2	0,48	-0,24	0,47	-0,65	0,6
KVBTBO	-0,61	0,19	1,04	0,19	0,4	0,49	0,4	-0,3	0,39
Aktif BH	-0,64	0,21	1,07	0,22	0,44	0,34	0,64	0,84	0,7
SBH	-0,5	0,16	0,95	0,19	0,38	0,38	0,45	-0,14	0,41
ÖBH	-1,4	-0,2	0,9	-0,23	0,6	0,76	0,83	-3,18	2,03
Değişken	-0,47	-0,06	0,39	-0,05	0,22				
α_{31}	-0,68	0,98	2,71	0,99	0,87				
α_{32}	-0,22	1,11	2,65	1,15	0,74				
α_{41}	0,29	1,85	4,15	1,94	0,99				
α_{42}	-3,23	-0,41	1,58	-0,55	1,24				
$\log(\sigma_3^2)$	1,02	1,05	1,06	1,05	0,01				
$\log(\sigma_4^2)$	0,74	0,78	0,82	0,78	0,02				
λ_2	1,79	1,83	1,88	1,83	0,03				

Bu tablo 2001 ve 2002 yıllarındaki yıllık verileri kapsar.

Modellerde az sayıda veri kullanılması nedeniyle, bu tablolarda, katsayıların standart hatalarının da büyük olduğu görülebilir. Genelde, MTREM ile elde edilen standart hataların lojistik modellerden elde edilenlerden daha küçük olduğu dikkat çekicidir. Çoklu bağımlı değişkeni kullanması nedeniyle verileri birleştiren MTREM daha fazla parametre tahmin etse de, daha fazla veri kullanır.

Tablo 4'teki α_2, çalışmanın ikinci yılındaki bağımlı değişkenin birinci yıldakilere bağımlılığını ölçer. Bağımsız lojistik modeller ve/veya kesitsel veri bakış açısı bu bağımlılığı dikkate alamaz. Tablo 5'teki, α_{31} ve α_{32} üçüncü zamandaki bağımlı değişkenin sırasıyla bir ve iki önceki yıllardaki bağımlı değişkenlerle ilişkisini ölçer. Başka bir deyişle, α_{31}, 2001 yılındaki verinin, 2000 yılındaki veriyle bağlantısını; α_{32} ise 2001 yılındaki verinin, 1999 yılındaki veriyle bağlantısını ölçer. Benzer bir şekilde, α_{41} ve α_{42} 2002 yılındaki bağımlı değişkenlerin sırasıyla 2001 ve 2000 yıllarındaki bağımlı değişkenlerle ilişkilerine ışık tutar. Tablodan α_{41} dışındaki parametrelerin

istatistiksel olarak anlamlı olmadıkları görülse de bu yine küçük veriden kaynaklanan bir sonuç olabilir. Bu parametrenin, α_{41}, tahmininin pozitif olması, önceki yılla bu yıldaki değişken arasındaki ilişkinin pozitif olduğunu gösterir. Bir başka ifade ile, bir şirketin önceki yılda başarılı olması, bu yılda da başarılı olma olasılığının yüksek olduğunu gösterir ki, bu beklenen bir durumdur. Örneğin, 2001 yılında başarılı olan bir şirketin başarısız olana kıyasla, 2002 yılında başarılı olma oddsu exp(1,85)=6,36 kat daha yüksektir.

Tablo 3-5 arasındaki yüksek σ^2 değerleri, şirketler arasında yüksek sapmalar olduğunu gösterir. Zaman içinde azalan sapmalar ($\sigma_1^2 \geq \sigma_2^2 \geq \sigma_3^2 \geq \sigma_4^2$), bu veride de olduğu gibi, panel veride sık görülen bir özelliktir. Bu sapmalar, MTREM modelinin 3. seviyesiyle ölçebildiği için, tek seviyeli lojistik modeller bu bilgiyi veremez.

Modelin üçüncü seviyesi kullanılarak, her bir yılda, her şirket için başarı olasılıkları hesaplanabilir. Örneğin, 2002 yılında *(t=4)* çalışmadaki ilk şirketin *(i=1)* özsermaye kârı *(j=1)* düzeyinde başarılı olma olasılığı (şirketin özsermaye kârının bağlı olduğu alt sektör medyan kârından yüksek olması olasılığı)

$$\exp(\Delta^*_{141} + \lambda_1 b_{14}) / (1 + \exp(\Delta^*_{141} + \lambda_1 b_{14}) =$$
$$\exp(-2.17 + (-0.74)) / (1 + \exp(-2.17 + (-0.74)) = 0.0515$$

olarak bulunmuştur (Bu veri için nxTxJ = 20x4x2 = 160 adet Δ^* ve nxT = 20x4 = 80 adet *b* hesaplandığı için Δ^* ve *b* değerleri raporlanmamıştır). Aynı şirketin, net kâr marjı *(j=2)* için başarı olasılığı ise

$$\exp(\Delta^*_{142} + \lambda_2 b_{14}) / (1 + \exp(\Delta^*_{142} + \lambda_2 b_{14}) =$$
$$\exp(-3{,}04 + 1{,}83 * (-0{,}74)) / (1 + \exp(-3{,}04 + 1{,}83 * (-0{,}74)) = 0{,}0122$$ olarak bulunur.

Şirketin bu yıldaki, 2002 yılındaki, gözlenen ve tahmin edilen değerleri iki başarı ölçümüne göre de başarısızdır. Bu olasılıklar aracılığıyla, her şirketin yıllar içindeki değişimi incelenebileceği gibi, aynı yıl içinde iki farklı şirketin durumları da karşılaştırılabilir.

3.2 Modellerin Başarısı

Tablo 3-5 arasından da görüleceği üzere, hizmetler sektöründeki şirketlerin başarısını açıklamak amaçlı kullanılan bağımsız değişkenlerin hiç biri istatistiksel olarak anlamlı değildir. Yine de, Tablo 6 ve 7'den görüleceği üzere, modelin başarısı ve doğru sınıflandırma başarıları özellikle bu boyuttaki bir veri seti için oldukça iyidir. Posterior tahmin edici kontroller (Gelman vd., 2003) dördüncü zamanda, iki bağımlı değişken arasındaki korelasyon ($Y_{.41}, Y_{.42}$) dışındaki korelasyonların, uygun (doyurucu) boyutta modellendiğini belirtmektedir (Tablo 6). Bu metoddaki ana fikir; eğer model iyi ise, gözlemlenen veri eldeki modelle üretilen suni verilere benzerlik gösterecektir. Bu amaçla, kurulan MTREM modeliyle 1000 adet suni veri seti üretilmiş, her biri gerçek veriyle karşılaştırılmıştır. Bu veri için toplam 16 adet istatistik tanımlanmıştır (bknz. Tablo 6). Bunlar farklı bağımlı değişkenler arasındaki logaritmik odds oranı (LOO) değerlerini hesaplar. Her istatistik için, 1000 adet suni veriden gelen ve bir adet gerçek veriden gelen LOO'lar hesaplanır. Suni olanlardan yüzde kaçının, gerçek verininkinden

daha büyük olduğu hesaplanır ve p-değeri olarak verilir. P-değerleri 0,01'den küçük veya 0,99'dan büyük olan istatistiklerde modelin başarısını arttırmak için çaba harcanabilir. Bu sayede, modelden elde edilen olasılıklar daha başarılı tahmin edileceği için doğru sınıflandırma oranları da artabilir.

Doğru sınıflandırma tabloları ise, (Tablo 7) MTREM'in %70 ile %95 arasında doğru sınıflandırma yapabildiğini göstermektedir. MTREM en az lojistik modeller kadar iyi ve genelde daha iyi sonuçlar vermiştir. Lojistik modellerin bir yılda özsermaye kârı için %40 kadar düşük doğru sınıflandırma vermesi, çok ciddi bir sorunu gösterir. Bir bozuk para atılarak yapılacak deneyin dahi, uzun süre tekrarlanması sonucunda %50 başarı getirmesi beklenir. Rastgele seçimden bile başarısız sonuçlar verecek bir modelin uygulanması tercih edilmez.

Tablo 6. Hizmetler sektörü için model başarısının ölçümü

İstatistik	LOO($Y_{.tj}, Y_{.t'j'}$)	p-değerleri
T_1	$Y_{.11}, Y_{.21}$	0,54
T_2	$Y_{.21}, Y_{.31}$	0,069
T_3	$Y_{.31}, Y_{.41}$	0,398
T_4	$Y_{.12}, Y_{.22}$	0,03
T_5	$Y_{.22}, Y_{.32}$	0,416
T_6	$Y_{.32}, Y_{.42}$	0,071
T_7	$Y_{.11}, Y_{.31}$	0,453
T_8	$Y_{.12}, Y_{.32}$	0,026
T_9	$Y_{.21}, Y_{.41}$	0,86
T_{10}	$Y_{.22}, Y_{.42}$	0,543
T_{11}	$Y_{.11}, Y_{.41}$	0,227
T_{12}	$Y_{.12}, Y_{.42}$	0,016
T_{13}	$Y_{.11}, Y_{.12}$	0,911
T_{14}	$Y_{.21}, Y_{.22}$	0,581
T_{15}	$Y_{.31}, Y_{.32}$	0,804
T_{16}	$Y_{.41}, Y_{.42}$	**1**

LOO: logaritmik odds oranı

Tablo 7. Hizmetler sektöründe özsermaye kârı (j=1) ve net kâr marjının (j=2) lojistik regresyonla ve MTREM ile modellenmesi sonucu elde edilen sınıflandırma başarıları (%)

	MODEL	YIL = 1999	YIL = 2000	YIL = 2001	YIL = 2002
j = 1	MTREM(2)	90	95	75	80
	Lojistik	90	70	40	75
j = 2	MTREM(2)	70	85	90	95
	Lojistik	70	65	60	50

4. TARTIŞMA VE SONUÇ

Bu çalışmada, Türkiye'de kurulmuş hizmetler sektörüne bağlı 20 şirket için 1999-2002 yılları arasında İstanbul Menkul Kıymetler Borsası'nın (İMKB) internet sayfasından toplanan panel verilerin istatistiksel analizleri yapılarak şirketlerin başarı tahminleri yapılmıştır. Bu veriye, biri tek seviyeli diğeri çok seviyeli olmak üzere iki model uygulanmış, bu iki model doğru sınıflandırma oranları üzerinden karşılaştırılmıştır.

Bu çalışma aracılığıyla, MTREM ilk kez Türk literatürüne tanıtılmış ve Türkiye'de toplanan bir veriye uygulanarak lojistik regresyon modeliyle karşılaştırılmıştır. Ayrıca, bilgimiz dahilinde, Türkiye'de ilk kez panel veri bakış açısıyla sektör bilançoları incelenmiştir.

Lojistik modeller sonucunda, farklı yıllarda ve farklı başarısızlık tanımları için değişen, %40 ile %90 arasında doğru sınıflandırma oranları elde edilmiştir. MTREM sonucunda ise bu oranlar %70 ile %95 arasında değişmiştir. Düşük sınıflandırma oranları, veri setinin küçüklüğü nedeniyle bazı bağımsız değişkenlerin modelde kullanılamamasının doğurduğu bir sonuçtur. Daha çok sayıda şirket içeren sanayi sektörü için tekrarlanan analizler net kâr marjı için her yılda %100 doğru sınıflandırma vermiştir. Bu sonuçlar, MTREM'in lojistik modellere üstünlüğünün küçük verilerde dahi korunduğunu göstermektedir.

Kesitsel veri yerine panel veri analizi, geçmiş bilgilerden güç aldığı için daha güçlü sonuçlar doğurur. Birden fazla bağımlı değişkenin kullanılması ve çok seviyeli modeller de sonuçları güçlendirmektedir.

Veri setinin kısıtlı olması nedeniyle, bu çalışmada Türkiye'deki şirketler hakkında genel yorumlar yapmak amaçlanmamıştır. Yine veri setinin küçüklüğü nedeniyle, erken uyarı sistemleri geliştirilememiştir. Bu sistemler, *t-1* zamanındaki bağımsız değişkenin *t* zamanındaki bağımlı değişkeni açıklaması üzerine kurulmuştur. Bu sayede, örneğin geçen seneki finansal rasyolara bakarak bu yıl başarısız olacak şirketler belirlenebilir.

Türkiye'de çalışan istatistikçilerin en büyük sorunlarından birisi, yeterli büyüklükte temiz ve güvenilir veri elde edememektir. Çalışmamızın, verinin akademik camiayla paylaşılması durumunda hem akademisyenler, hem de yöneticiler için yararlı çıktılar vereceği konusunda ikna edici olacağını umuyoruz. Panel veri durumunda, daha fazla bireyden, daha uzun zamanda ve aynı formatta toplanmış olması sonuçları daha da güçlendirecektir. Bu nedenle, sık değişen kanunlar ve alınan resmi kararlar yüzünden sık sık değişen veri formatının karar mercilerini zor durumda bıraktığını vurgulamak isteriz.

Bu çalışmada, istatistiksel analizler, R ve Fortran kullanılarak yapılmıştır. R internetten ücretsiz indirilebilir.

5. TEŞEKKÜR

Bu çalışma, TÜBİTAK (SOBAG-105K048) tarafından desteklenmiştir. Katkılarından dolayı hakemlere ve Dergi Editörlüğüne teşekkür ederiz.

6. KAYNAKLAR

Aktaş, R., 1997. Mali başarısızlık (İşletme Riski) tahmin modelleri, 2. baskı. Türkiye İş Bankası Kültür Yayınları, Ankara.

Altman, E.I., 1968. Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. The Journal of Finance, c.XXIII, 4, 589-609.

Brooks, S.P., 1998. Markov chain Monte Carlo method and its application. The Statistician, 47, 69-100.

De Andres, J., Landajo M., Lorca P., 2005. Forecasting business profitability by using classification techniques: A comparative analysis based on Spanish case. European Journal of Operational Research, 167, 518-542.

Gelman, A. J., Carlin, B., Stern H. S., Rubin D. B., 2003. Bayesian data analysis, 2nd edition. Chapman & Hall, London.

Geman, S., Geman, D., 1984. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. IEEE Transactions on Pattern Analysis and Machine Intelligence, 6, 721-741.

İbrahim, J. G., Chen M.-H., Lipsitz S. R., 2002. Bayesian methods for generalized linear models with covariates missing at random. The Canadian Journal of Statistics- La revue canadienne de statistique, 30, 55–78.

İlk, O., 2004. Exploratory multivariate longitudinal data analysis and models for multivariate longitudinal binary data, (Basılmamış Doktora Tezi). Iowa State University, Ames, United States of America (Ingilizce).

İlk, O., Daniels, M., 2007. Marginalized transition random effects models for multivariate longitudinal binary data. The Canadian Journal of Statistics-La revue canadienne de statistique, 35, 105-123.

Liao, J., Gartner, W.B., 2006. The effects of pre-venture plan timing and perceived environmental uncertainty on the persistence of emerging firms, small business economics, 27, 23–40.

Mittal, V., Anderson, E.W., Sayrak, A., Tadikamalla, P., 2005. Dual emphasis and the long-term financial impact of customer satisfaction. Marketing Science, 24,4, 544–555.

Neal, R.M., 1996. Bayesian learning for neural networks. Springer-Verlag, New York.

Neter, J., Kutner, M.H., Nachtseim, C.J., and Wasserman, W., 1996. Applied linear statistical models, 4th edition. Irwin, Chicago.

Ohlson, J.A., 1980. Financial ratios and the probabilistic prediction of bankruptcy. Journal of Accounting Research, 109-111.

R, 2006. http://www.r-project.org/, Erişim tarihi: 1, Haziran, 2006

Yao, S., Jiang, C., Feng, G., Willenbockel, D., 2007. WTO challenges and efficiency of Chinese banks. Applied Economics, 39, 629–643.

MODELING FINANCIAL FAILURE IN SERVICE SECTOR

ABSTRACT

Observing the economical changes and determining the factors related to financial failure are important for both the economical development of the country and for the self - evaluation of individual firms. In this study, the calculation of the financial success pobabilities for the Turkish firms in service sector and the investigation of temporal change in these probabilities are aimed. With this purpose in mind, financial statements that are collected from İstanbul Stock Exchange are investigated, and multilevel statistical models are used for analysing this complex structured data which consists of repeated measurements in time. Specifically, Marginalized Transition Random Effects Models (MTREM) are fitted. By these models, it is possible to calculate financial success probabilities for each company, to compare success of companies in different subgroups, and to observe the changes in time.With a purpose of comparison, popular single level logistic regression models are fitted as well. In terms of the true classification rates, it is observed that MTREM is superior to logistic regression models.

Key Words: Hierarchical statistical models, Panel data, Financial statements of sectors.

Cilt: 06 Sayı: 01 Kategori: 02 Sayfa: 65-74 ISSN No: 1303-6319 Volume: 06 Sayı: 01 Category: 02 Page: 65-74

GRUPLANDIRILMIŞ VERİLERİN ÜSTEL DAĞILIMA UYUMUNDA AĞIRLIKLANDIRILMIŞ KOLMOGROV-SMIRNOV TESTLERİ İLE OLABİLİRLİK ORANI VE Kİ-KARE TESTLERİNİN KARŞILAŞTIRILMASI

Hamza GAMGAM* **Esra YİĞİT****

ÖZET

Bu çalışmada, gruplandırılmış verilerin üstel dağılıma uyumu için Gulati ve Neus (2003) tarafından önerilen Ağırlıklandırılmış Kolmogrov-Smirnov test istatistikleri tanıtılmıştır. Bu istatistiklerle, olabilirlik oranı ve ki-kare Uyum İyiliği test istatistikleri, farklı alternatif dağılımı, grup sayısı ve örnek çapı için güç bakımından karşılaştırılması yapılmıştır. Karşılaştırmalar sonucunda özellikle sağa çarpık dağılımlarda, Ağırlıklandırılmış Kolmogrov-Smirnov test istatistiklerinin güç bakımından performansının, Pearson χ^2 ve olabilirlik oranı test istatistiğine göre daha iyi olduğu görülmüştür.

Anahtar Kelimeler: Ağırlıklandırılmış Kolmogrov-Smirnov, Bootstrap, Gruplandırılmış veri, Pearson Ki-Kare, Olabilirlik oranı, Uyum iyiliği.

1. GİRİŞ

Üstel dağılım, başarısızlık gerçekleşene kadar geçen sürenin dağılımıdır. Bu sebepten birçok endüstriyel ve biyolojik uygulamalarda sıklıkla kullanılır. Hafızasızlık özelliğinden dolayı bir sonraki başarısızlığın ne zaman gerçekleşebileceğinin tahmini için oldukça faydalıdır. Gruplandırılmış verileri, özellikle bazı uygulama alanında kullanmak gerekli olabilir. Örneğin gözlem birimlerinin tam olarak ölçülemediği durumlarda gruplandırılmış veri kullanmak daha sağlıklı olur. Birçok deneyde birimleri sürekli değişken olarak gözlemlemek çok zor ya da imkânsızdır. Bunun yerine önceden, belirlenmiş belirli aralıklarda ölçüm yapmak hem daha kolay hem de daha ucuzdur. Örneğin belediyenin yeni aldığı bir grup otobüsün bozulma zamanları ile ilgili çalışma yapmak istenmesi durumunda, bunların her birinin bozulma zamanlarının kaydını ayrı ayrı tutmak yerine, belirli zaman aralıklarında kaç tanesinin bozulduğunun ölçülmesi daha kolay ve ucuz bir yoldur.

* Gazi Üniversitesi, Fen-Edebiyat Fakültesi, İstatistik Bölümü Teknikokullar, Ankara, e-posta: gamgam@gazi.edu.tr
** Gazi Üniversitesi, Fen-Edebiyat Fakültesi, İstatistik Bölümü Teknikokullar, Ankara, e-posta: eyigit@gazi.edu.tr

Gruplandırılmış verilerde uyum iyiliğini ölçmek için ilk test istatistiği Pearson (1900) tarafından geliştirilmiştir. Daha sonra Kolmogrov Smirnov ve Cramer Von Mises test istatistikleri geliştirilmiştir. Kolmogrov Smirnov test istatistiğini Schmid (1958) ve Conover (1972) gruplandırılmış veri için düzenlemişlerdir. Kuldorf (1961), gruplandırılmış veriler için üstel dağılımın parametre tahminini en çok olabilirlik yöntemi ile elde etmiştir. Ayrıca gruplandırılmış veriler için Seo ve Yum (1993) tarafından üstel dağılımın parametresinin tahmini için bir yöntem geliştirilmiştir. Gulati ve Neus (2003) gruplandırılmış veriler için Kolmogrov-Smirnov test istatistiğini ağırlıklandırarak, üstel dağılım için iki uyum iyiliği test istatistiği önermiştir. Daha sonra Baklizi (2006), Gulati ve Neus (2003) tarafından üstel dağılım için geliştirilen uyum iyiliği test istatistiğini Rayleigh dağılımına uygulamıştır.

Ayrıca Best ve Rayner (2007) χ^2'nin bileşenlerinden faydalanarak, gruplandırılmış veriler için üstel dağılıma uyum iyiliği test istatistiği geliştirmiştir.

Çalışmanın ikinci bölümünde, Gulati ve Neus (2003) tarafından önerilen SW1 ve SW2 test istatistikleri tanıtılmıştır. Üçüncü bölümde ise bu istatistikler ile olabilirlik oranı ve ki-kare test istatistiğinin güç bakımından karşılaştırılmasında kullanılacak yöntemin algoritması verilmiştir. Ayrıca bu bölümde farklı parametreli bazı dağılımlar için, farklı grup sayıları, örnek çaplarına göre SW1, SW2 olabilirlik oranı ile ki-kare testlerinin güçleri hesaplanmış ve bunlara ait güç grafikleri çizilmiştir.

2. YÖNTEM

2.1 Gruplandırılmış Veriler için Kullanılan SW1 ve SW2 Test İstatistikleri

Olasılık yoğunluk fonksiyonu f(x) olan bir dağılımdan n tane gözlem alınsın ve bunlar, $x_1, x_2, \ldots, x_{k-1}$ kesim noktaları olmak üzere, $(0, x_1)$, $(x_1, x_2), \ldots, (x_{k-1}, \infty)$ k sayıda gruba ayrılsın. Gruplara düşen gözlem sayıları sırasıyla $n_1, n_2, \ldots, n_k$ olsun. Gruplandırılmış verilerin üstel dağılıma uyumu için yokluk hipotezi aşağıdaki gibidir (Gulati ve Neus (2003)).

H_0: $f(x)=\theta\exp(-\theta x)$, ($\theta>0$ bilinmiyor , $x>0$)

θ'nın en çok olabilirlik tahmin edicisi, $n_1<n$ ve $n_k<n$ koşulları altında, Kulldorf (1961) tarafından,

$$\sum_{i=1}^{k-1} \frac{n_i \left(x_i - x_{i-1}\right)}{e^{\theta(x-x_{i-1})} - 1} - \sum_{i=2}^{k} n_i x_{i-1} = 0 \qquad (1)$$

eşitliği her aralığın eşit olduğu varsayımı altında çözülerek;

$$\hat{\theta} = \frac{1}{x_1} \ln\left(1 + \frac{n - n_k}{\sum_{i=2}^{k} (i-1) n_i}\right) \qquad (2)$$

şeklinde elde edilmiştir (Gulati ve Neus, 2003). Eğer bu varsayım kullanılmaz ise $\hat{\theta}$ için tekrarlı çözüm gerekir.

$1 \le i \le k$ için n_i i. gruptaki gözlem sayısı ve x_i de i. grubun üst sınırı olmak üzere, üstel dağılım için x_i değerinden daha küçük olma olasılığı,

$$F(x_i, \hat{\theta}) = 1 - \exp(-x_i \theta)$$

biçiminde ifade edilir. Bu olasılığın deneysel sonucu ise,

$$F_n(x_i) = \sum_{j=1}^{i} n_j / n$$

şeklinde tanımlanır. Bu iki olasılığın farkı,

$$S_i = F_n(x_i) - F(x_i, \hat{\theta})$$

olarak ifade edilir. SW1 ve SW2 test istatistikleri aşağıdaki gibi önerilmiştir.

$$\text{SW1} = \sqrt{n} \sum_{j=1}^{k-1} \frac{|S_j|}{\Psi_1(j)} \tag{3}$$

ve

$$\text{SW2} = \sqrt{n} \sum_{j=1}^{k-1} \Psi_2(j) |S_j| \tag{4}$$

Burada, $\Psi_1(j)$ ve $\Psi_2(j)$ ağırlık fonksiyonları olmak üzere,

$$\Psi_1(j) = \sqrt{F\left(x_i, \hat{\theta}\right)\left(1 - F(x_i, \hat{\theta})\right)},$$

$$\Psi_2(j) = (k/2 - j)^2 +$$

olarak tanımlanır (Gulati ve Neus (2003)). ψ_2'deki (+) ifadesi;

$$x+ = \begin{cases} x & x \neq 0 \\ 1 & x = 0 \end{cases}$$

şeklinde tanımlanır. SW1 test istatistiği dağılımın uçlarına ağırlık verirken, SW2 test istatistiği dağılımın merkezine daha fazla ağırlık verir. SW1 ve SW2 test istatistiklerinin tam dağılımı bilinmediği için Bootstrap yöntemi ile *p* değeri bulunarak, anlamlılık düzeyine göre test sonucu yorumlanır.

3. BULGULAR

3.1 Simülasyon Çalışması ve Güçlerinin Karşılaştırılması

SW1, SW2 olabilirlik oranı ve χ^2 testlerinin güç karşılaştırılması için kullanılan algoritma aşağıda verilmiştir.

1) Çeşitli dağılımlardan ((Ki-kare (1), Ki-kare (4), Weibull (0.8, 1), Lognormal (0, 1), Normal (2, 2), Tekdüze (0, 2.5), Beta (2, 2) ve Lojistik (1.2, 0.35)) n sayıda rassal sayı üretilerek, k gruba bölünür ve (2) eşitliği kullanılarak her biri için $\hat{\theta}$ değeri hesaplanır.
2) Bunların her biri için SW1, SW2, olabilirlik oranı ve χ^2 test istatistiklerinin değerleri elde edilir.

3) 1. adımda elde edilen her bir $\hat{\theta}$ değeri ile üstel dağılımdan, n sayıda rassal sayı üretilerek k gruba bölünür ve bu şekilde 10000 tane Bootstrap örnek oluşturularak, SW1 ve SW2 istatistiklerinin Bootstrap dağılımları elde edilir.
4) Olabilirlik oranı ve χ^2 testi için $\chi^2_{(k-2)}$ dağılımı kullanılarak, *p* değerleri hesaplanır.
5) 2. adımda elde edilen SW1 ve SW2 test istatistiklerinin değerleri için SW1 ve SW2 test istatistiklerinin 3. adımda oluşturulan dağılımları kullanılarak, Bootstrap *p* değerleri hesaplanır.
6) 4. ve 5. adımdaki *p* değerleri seçilen α=0.05 ile karşılaştırılıp, testin sonucu bulunur.
7) Bu işlem 10000 kez tekrar edilerek, 4. ve 5. adımdaki red sayıları bulunur ve bu red sayıları 10000'e bölünerek, her bir test istatistiği için testin gücü hesaplanır.

Simülasyon çalışmasında, yukarıda belirtilen dağılımlardan, *n* çaplı (*n*= 20: 20: 200) örnekler seçilerek farklı grup sayılarına (*k*= 4, 6, 10) ayrılmıştır. Ki-kare (1), Ki-kare (4), Weibull (0.8, 1), Lognormal (0, 1), Normal (2, 2), Tekdüze (0, 2.5) ve Lojistik (1.2, 0.35) için herhangi bir k değeri için i. grubun aralıkları;

$$\left[x_{i-1}, x_i\right) = \begin{cases} \left[\dfrac{2(i-1)}{k-1}, \dfrac{2i}{k-1}\right) & i < k \\ \left[\dfrac{2(i-1)}{k-1}, \infty\right) & i = k \end{cases} \qquad (5)$$

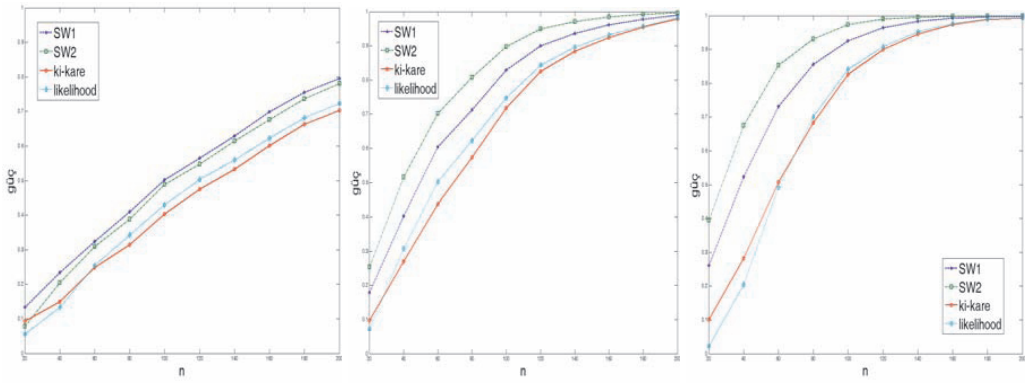
şeklinde alınmıştır. Burada $2i/(k-1)$ kesim noktaları olarak adlandırılır. Beta (2, 2) için;

herhangi bir k değeri için i. grubun aralıkları;

$$\left[x_{i-1}, x_i\right) = \begin{cases} \left[\dfrac{(i-1)}{k-1}, \dfrac{i}{k-1}\right) & i < k \\ \left[\dfrac{(i-1)}{k-1}, \infty\right) & i = k \end{cases} \qquad (6)$$

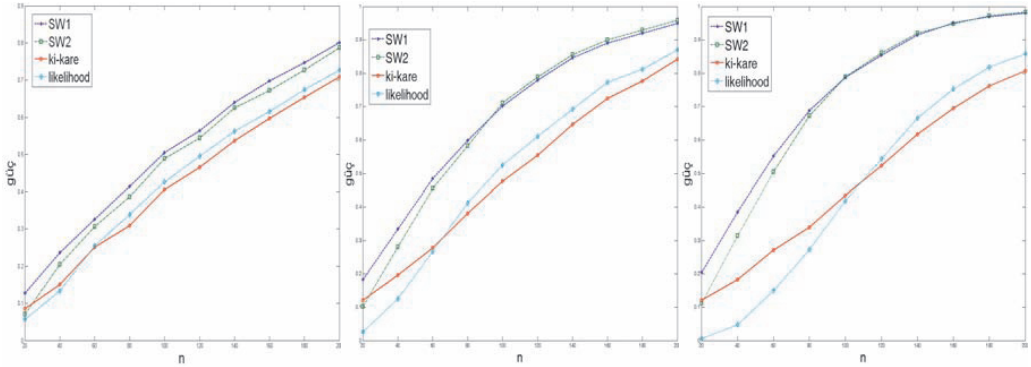
şeklinde alınmıştır. Burada $i/(k-1)$ kesim noktaları olarak adlandırılır.

Her bir durum için SW1, SW2 olabilirlik oranı ve χ^2 test istatistiklerinin değeri hesaplanmıştır. Daha sonra üstel dağılımdan örnek çapı *n* için, 10000 tane Bootstrap örnek üretilerek, SW1 ve SW2 istatistiklerinin Bootstrap dağılımları elde edilmiştir. Oluşturulan bu dağılımlar kullanılarak, SW1 ve SW2 testleri için Bootstrap *p* değeri hesaplanmıştır. Olabilirlik oranı ve χ^2 testi için $\chi^2_{(k-2)}$ dağılımı kullanılarak, *p* değeri elde edilmiştir. Bu *p* değerleri α=0.05 ile karşılaştırılarak, test sonucu bulunmuştur. Bu işlem 10 000 kez tekrarlanarak her bir test istatistiği için red sayıları saptanmış ve bu red sayıları 10 000'e bölünerek, her bir test istatistiği için testin gücü hesaplanmıştır. Bu durum *n*= 20: 20: 200 örnek çapları için yapılarak, Şekil 1-8'deki grafikler elde edilmiştir.



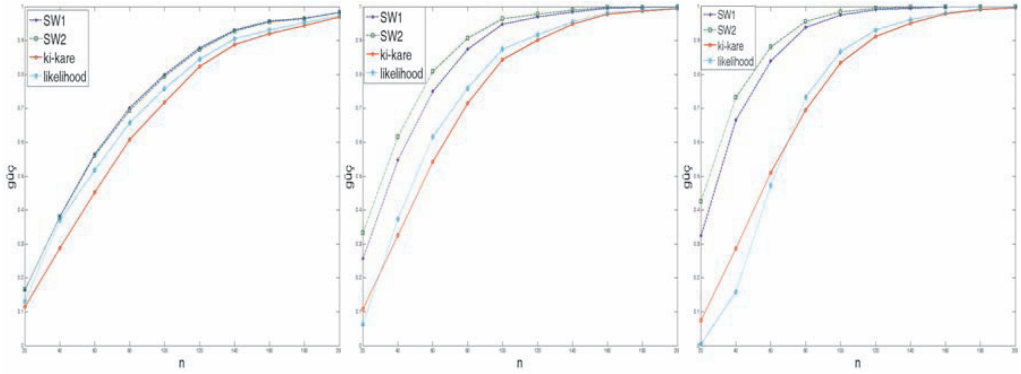
(a) (b) (c)

Şekil 1. χ_1^2 **dağılımından üretilen verilerin, grup sayısı a) k=4; b) k=6; c) k=10 iken, üstel dağılıma uygunluk testinin güç**



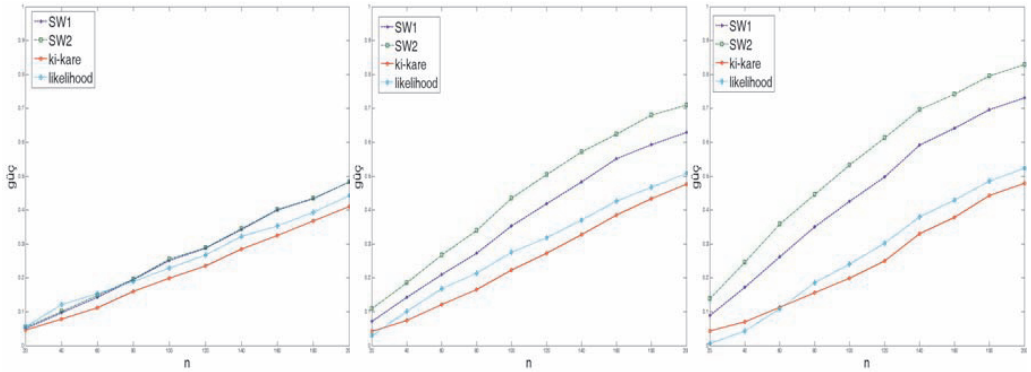
(a) (b) (c)

Şekil 2. χ_4^2 **dağılımından üretilen verilerin, grup sayısı a) k=4; b) k=6; c) k=10 iken, üstel dağılıma uygunluk testinin güç grafikleri**



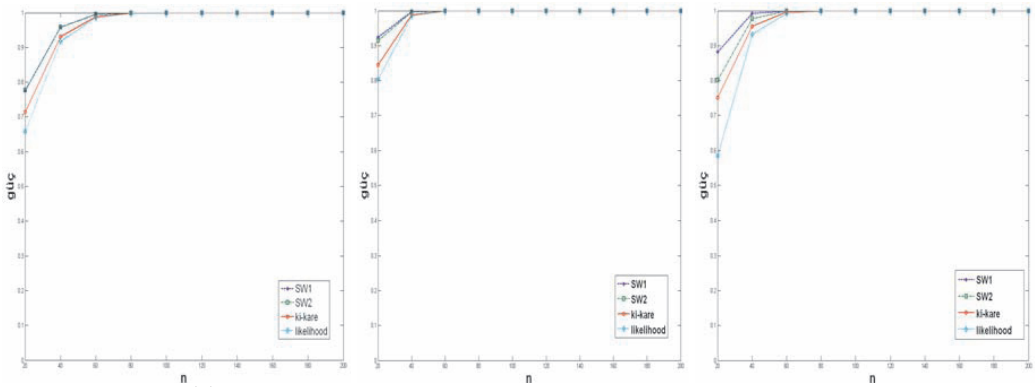
(a) (b) (c)

Şekil 3. Log-normal (0, 2) dağılımından üretilen verilerin, grup sayısı a) k=4; b) k=6; c) k=10 iken, üstel dağılıma uygunluk testinin güç grafikleri



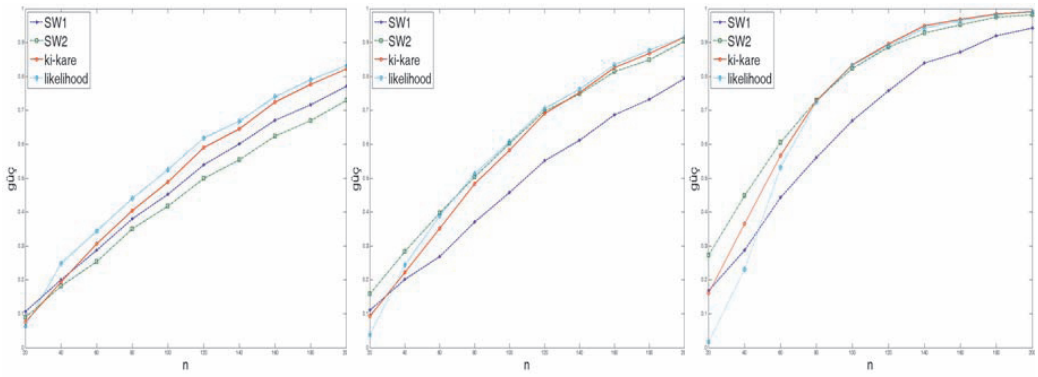
(a) (b) (c)

Şekil 4. Weibull (1, 0.8) dağılımından üretilen verilerin, grup sayısı a) k=4; b) k=6; c) k=10 iken, üstel dağılıma uygunluk testinin güç grafikleri



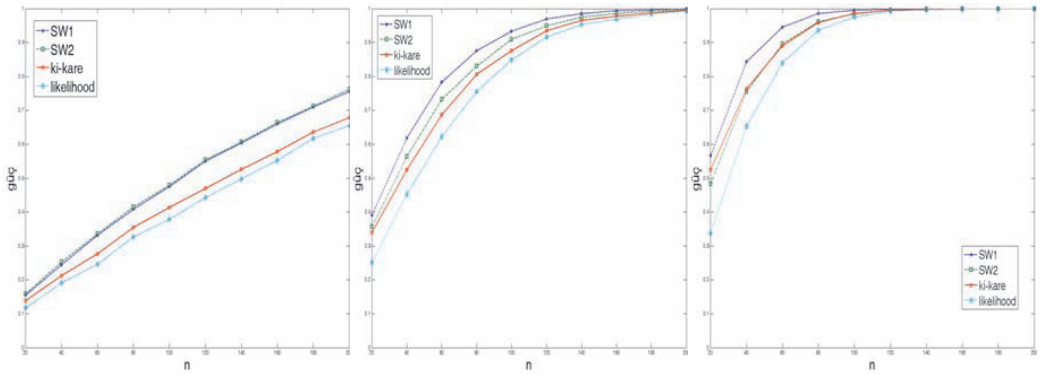
(a) (b) (c)

Şekil 5. Lojistik (1.2, 0.35) dağılımından üretilen verilerin, grup sayısı a) k=4; b) k=6; c) k=10 iken, üstel dağılıma uygunluk testinin güç grafikleri



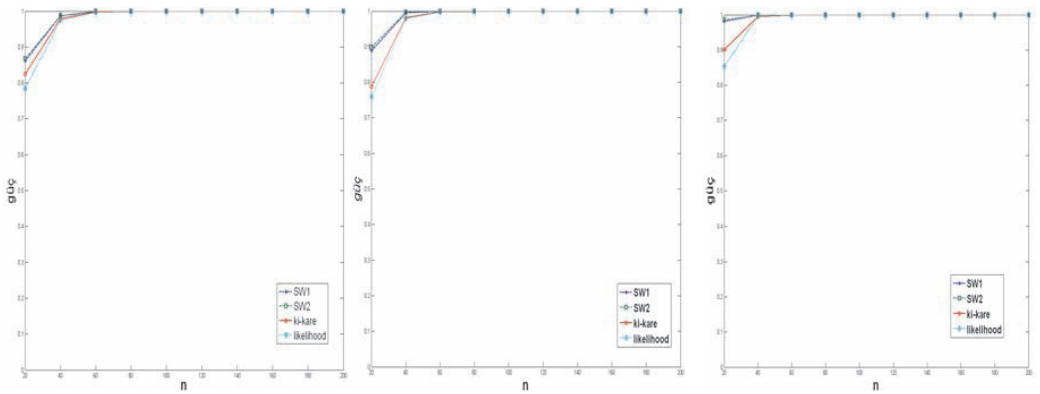
(a) (b) (c)

Şekil 6. Normal (2, 2) dağılımından üretilen verilerin, grup sayısı a) k=4; b) k=6; c) k=10 iken, üstel dağılıma uygunluk testinin güç grafikleri



(a) (b) (c)

Şekil 7. Tekdüze (0, 2.5) dağılımından üretilen verilerin, grup sayısı a) k=4; b) k=6; c) k=10 iken, üstel dağılıma uygunluk testinin güç grafikleri



(a) (b) (c)

Şekil 8. Beta (2, 2) dağılımından üretilen verilerin, grup sayısı a) k=4; b) k=6; c) k=10 iken, üstel dağılıma uygunluk testinin güç grafikleri

Şekil 1, 2, 3, 4, 5'te SW1, SW2, Ki-kare ve olabilirlik oran test istatistikleri, testlerin gücü bakımından alternatif dağılım biçiminin sağa çarpık olduğu durum incelenmiştir. Bu yüzden sağa çarpık dağılımlardan; χ_1^2, χ_4^2, Log-normal (0, 2), Weibull (1, 0.8) ve Lojistik (1.2, 0.35) dağılımları ele alınmıştır. Genel olarak, SW2 test istatistiğinin güç bakımından SW1, χ^2 ve olabilirlik oran test istatistiklerine göre daha iyi olduğu görülmektedir. Ayrıca SW1, SW2 üstel dağılıma benzeyen diğer dağılımları ayırt etmekte daha iyi sonuç vermektedir. Aynı zamanda tüm test istatistikleri artmasına rağmen, bu artış miktarı SW1 ve SW2'de ki-kare ve olabilirlik oranına göre daha fazladır. Grup sayısı küçükken (k=4) tüm test istatistikleri birbirine yakın sonuçlar vermekte, grup sayısı arttıkça (k= 6, 10) SW1 ve SW2 ile χ^2 ve olabilirlik oran test istatistikleri arasındaki fark artmakta ve χ^2 en kötü sonucu vermektedir. Bu durum hücrelerin beklenen değerinin 5'ten küçük olmasından kaynaklanmaktadır. Örneğin, n=20, k=10 durumunda gözlenen grup frekansları 2 olacaktır. Bu durumda ki-kare ve olabilirlik oranı, yapıları gereği iyi sonuç vermeyecektir.

Şekil 6, 7, 8'de kullanılan test istatistikleri, testlerin gücü bakımından dağılım biçiminin simetrik olduğu durumlarda; Normal (2, 2), Tekdüze (0, 2.5), Beta (2, 2) incelenmiştir. Test istatistikleri güç bakımından Normal (2, 2) ve Tekdüze (0, 2.5) dağılımlarında birbirine yakın sonuçlar vermiştir. Normal dağılımda χ^2, olabilirlik oranı ve SW2 en iyi sonucu verirken, SW1 en kötü sonucu vermektedir. Tekdüze dağılımında ise, bunun tersi olarak SW1 en iyi sonucu verirken, olabilirlik oranı en kötü sonucu vermektedir. Beta dağılımında ise, test istatistikleri bütün grup düzeyleri için yüksek güç değerlerine sahip olup, SW1 ve SW2 diğer test istatistiklerine göre daha iyi sonuç vermektedir.

Sola çarpık dağılımlardan; Weibull (1.5, 8) ve Gumbel (2) dağılımları incelendiğinde biçim olarak üstel dağılıma benzemediğinden en düşük örnek çapında (n=20) bile tüm test istatistiklerinin, gücü 1'e yakın çıkmaktadır.

4. TARTIŞMA VE SONUÇ

Grup sayısı ve örneklem çapı arttıkça, tüm test istatistiklerinin özellikle de SW1 ve SW2 test istatistiklerin güç değerleri artmaktadır. Sağa çarpık alternatif dağılımlardan üretilen verilerin üstel dağılıma uygunluğunun testinde, SW1 ve SW2 test istatistikleri daha iyi sonuç verirken, simetrik alternatif dağılımlardan üretilen verilerin üstel dağılıma uygunluğunun testinde SW1 ve SW2'nin, Ki-kare ve olabilirlik oranı test istatistiklerinden üstün olmadığı görülmüştür. Sola çarpık dağılımlardan üretilen veriler için ise, tüm test istatistikleri çok iyi sonuçlar vermiştir. Bu çalışma, farklı dağılımlar için test istatistiği geliştirilerek, genişletilebilir.

5. KAYNAKLAR

Baklizi, A., (2006). Weighted Kolmogrov-Smirnov type tests for grouped Rayleigh data. Applied Mathematical Modelling 30:437-445.

Best, D. J., Rayner, J.C.W. (2007). Chi-Squared components for tests of fit and improved models fort he grouped exponential distribution. Computational Statistics and Data Analysis 51:3946-3954.

Conover, W. J. (1972). A Kolmogrov goodness-of-fit test for discontinuous distributions. Journal of the American Statistical Association 67:591-596.

Gulati, S., Neus, J., (2003). Goodness of fit statistics for the exponential distribution when the data grouped. Comm. Statist. Theory Methods 32, 681-700.

Kulldorf, G., (1961). Estimation from grouped and partially grouped samples. New York: John Wiley & Sons.

Pearson, K., (1900). On a criterion that a given system of deviations from the probable in the case of correlated system of variable is such that it can be reasonably supposed to have arisen from random sampling, Philos. Mag., 5th series 50 (1900) 157-175.

Schmid, P., (1958). On the Kolmogrov and Smirnov limit theorems for discontinuous distribution functions. Annals of Mathematical Statistics 29:1011-1027.

Seo, S. K., Yum, B. J. (1993). Estimation methods for the mean of the exponential distribution based on grouped and censored data. IEEE Transactions on Reliability, 42 (1):87-96.

COMPARISONS OF WEIGHTED KOLMOGROV-SMIRNOV, LIKELIHOOD RATIO AND CHI-SQUARE GOODNESS OF FIT TESTS FOR THE EXPONENTIAL DISTRIBUTION BASED ON THE GROUPED DATA

ABSTRACT

In this paper, Weighted Kolmogrov-Simirnov test statistics which are used to test whether the grouped data fits to exponential distrubution proposed by Gulati and Neus (2003) are defined. These test statistics are compared with Likelihood ratio and Chi-square goodness of test statistic in terms of power under diffrent alternative distribution, group size and sample size. The simulation results showed that specially for right skewed distributuions Weighted Kolmogrov-Smirnov test statistics outperformed Pearson Chi-Square test statistics in terms of power.

Keywords: Bootstrap, Goodness of fit, Grouped data, Likelihood ratio, Pearson Chi-Square, Weighted Kolmogrov-Smirnov.

Cilt: 06 Sayı: 01 Kategori: 02 Sayfa: 75-95 ISSN No: 1303-6319 Volume: 06 Sayı: 01 Category: 02 Page: 75-95

MEVSİMSEL DÜZELTME İÇİN ARIMA MODEL TABANLI YAKLAŞIM

Kemal ÇALIK* **Seçil ÇALIK***

ÖZET

Bu makale, bir zaman serisini karşılıklı olarak birbirinden bağımsız mevsimsel, eğilim ve düzensiz bileşenlerine ayrıştırmak için ARIMA Model Tabanlı yaklaşım üzerinde durmaktadır. Bileşenlerin tahmin edicileri Wiener-Kolmogrov (WK) filtresi aracılığıyla hesaplanmaktadır. Serinin Gaussian ARIMA modeline sahip olduğu kabul edilmektedir. Yöntemin özellikleri açıklanmakta ve gerçek bir örnek verilmektedir. Uygulamada Demetra paket programı kullanılmıştır.

Anahtar Kelimeler: ARIMA model, Kanonikal ayrıştırma, Sinyal çıkarımı, Spektrum, Wiener-Kolmogrov filtresi.

1. GİRİŞ

Konjonktürel dalgalanmaların temelde daha kolay incelenmesi ve güncel ekonomik koşulların değerlendirilebilmesi için zaman serisinden mevsimselliğin kaldırılması gerektiği önemli bir tartışma konusudur (Nerlove vd., 1979, s. 147). Finansal ve ekonomik zaman serileri genellikle her yıl yaklaşık olarak aynı büyüklüklerde tekrarlayan, mevsimsel-periyodik, iniş-çıkışlar gösterir. Mevsimsel hareketlerin arındırıldığı bir zaman serisi, mevsimsel yapının farklı olduğu ikişer aylık dönemler veya mevsimler arasındaki verilerin karşılaştırılmasını sağlamaktadır. Mevsimsel düzeltilmiş veriler sıklıkla, ekonomik modellemede ve dönemsel (cyclical) analizde kullanılır. Mevsimsel düzeltilmiş verilerin sunumu, farklı mevsimsel yapısı olan farklı serilerin karşılaştırılmasını sağlar ve bu seriler farklı ülkelerin aynı yılın aynı ayında farklı mevsimsel bir harekete sahip olacağından, uluslararası karşılaştırmalar bağlamında tutarlıdır. Mevsimsel düzeltme, kısa dönem mevsimsel dalgalanmaları kaldırarak yönetimin fikirlerini dayandırabileceği orta ve uzun dönem tahminleri de belirlemektedir. Mevsimsel düzeltme teorisindeki gelişmeler ekonomik faaliyetlerle ilgili olarak daha güvenilir yorumlar yapmayı mümkün kılmaktadır. Günümüzde konjonktür analistleri, ekonomistler ve politikacıların kullandığı ana bilgi kaynaklarından biri de mevsimsel düzeltilmiş verilerdir.

Seksenlerin ilk yıllarında, zaman serilerinin mevsimsel düzeltmesi için alternatif bir yaklaşım–"ARIMA Model Tabanlı" (AMB) olarak adlandırılan yaklaşım-oluşturulmuştur (Burman, 1980; Hillmer ve Tiao, 1982). Yöntem, bir ARIMA modelinin belirlendiği gözlemlenen zaman serisinde gizli gözlemlenemeyen bileşenlerin Minimum Hata Kare Ortalama (MMSE) tahmininden (veya "sinyal çıkarımı") oluşur (Nerlove vd., 1979). Tipik olarak, bileşenler (veya sinyaller) mevsimsel, eğilim ve düzensiz bileşenlerdir ve mevsimsel etkiden arınmış olan seri, eğilim ve düzensiz bileşeni içermektedir. Üç bileşen karşılıklı olarak bağımsız varsayılır. Eğilim-konjonktür ve mevsimsel bileşen durumlarında genellikle durağan olmayan ARIMA tipi ifadeli doğrusal stokastik süreç izler. Bileşenler için belirlenen modeller, gözlemlenen seriler için belirlenen ARIMA modeli içerisinde bütünleştirilir (Maravall, 1995).

* Başbakanlık Türkiye İstatistik Kurumu, Kars Bölge Müdürlüğü, Kars, e-posta: kemalcalik@tuik.gov.tr

Bileşenlerin tahmin edicileri durağan olmayan serilere uygulandığı gibi Wiener-Kolmogrov (WK) filtresi kullanılarak da hesaplanır (Bell, 1984).

ARIMA Model Tabanlı (AMB) yöntem bazı üstün özelliklere sahiptir. Bir yandan, gözlemlenmiş serinin ARIMA modeli ile uygunluğu, yanıltıcı sonuçlara veya modelin yanlış belirlenmesine karşı iyi bir koruma olarak görülebilir. Diğer yandan, parametrik model tabanlı yöntem analiz ve yorumu kolaylaştırabilir (Pierce (1979, 1980; Bell ve Hillmer, 1984; Hillmer, 1985; Maravall, 1987; Maravall ve Planas, 1999).

Çoğu zaman serisi için ARIMA model belirlenmeden önce öndüzeltme (preadjustment) işlemlerine ihtiyaç duyulur. Önemli düzeltmeler; aykırı değer düzeltmesi, takvim etkilerinin, müdahale değişkenlerinin ve diğer olası regresyon etkilerinin çıkarılması ve eksik gözlemlerin ara değerlerinin eklenmesidir. Zaman serisi literatüründe aykırı değer belirlenmesi; modellemede, yorumlamada ve hatta veri sürecinde önemli bir rol oynar. Aykırı değerler, modelin yanlış belirlenmesine, yanlı parametre tahminine ve başarısız öngörülere yol açabilir. Aykırı değerlerin varlığı otoregresif (AR) ve hareketli ortalama (MA) parametrelerinin tahmininde ciddi sapmalara neden olabilir. Örnek çalışma olarak; Chang vd. (1988), Box ve Tiao (1975), Chen ve Liu (1993), Hillmer vd. (1983), Gómez ve Maravall (2001a), Gómez vd. (1999) verilebilir. Öndüzeltmeye olan ihtiyacın önemi gittikçe artmaktadır ve model tabanlı sinyal çıkarım yöntemlerinin dışına da genişlemektedir (Findley vd., 1998).

2. YÖNTEM

2.1 ARIMA Model Tabanlı Yaklaşım ve Özet Tanımları

B geriye doğru öteleme işlemcisini ifade ettiğinde; $Bx_{(t)} = x_{(t-1)}$ ve ***m*** yıl için gözlem sayısı, $0 < t_1 < ... < t_m$ ve $y = \left(y_{(t1)}, y_{(t2)}, ..., y_{(tm)}\right)$ gözlemleri verildiğinde;

$$y_{(t)} = \sum_{i=1}^{n_{out}} \omega_i \lambda_i(B) d_{i(t)} + \sum_{i=1}^{n_c} \alpha_i cal_{i(t)} + \sum_{i=1}^{n_{reg}} \beta_i reg_{i(t)} + x_{(t)} \qquad (1)$$

genel modeline uyar. Burada:

$d_{i(t)}$: i. aykırı değerin yerini gösteren bir kukla değişkeni,

$\lambda_i(B)$: B'de aykırı değerlerin dinamik yapısını yansıtan bir polinomu,

cal_i : takvim tipi değişkenini,

reg_i : bir müdahale veya regresyon değişkenini,

$x_{(t)}$: ARIMA[7] modelindeki hataları ifade eder.

[7] Autoregressive Integrated Moving Average

ω_i parametresi; i. andaki aykırı değer etkisini, α_i takvim ve β_i regresyon-müdahale değişkenlerinin katsayısı, sırasıyla n_{out}, n_c, n_{reg} Eşitlik (1)'de belirtilen değişkenlerin her birinin toplam sayısını belirtir. Eşitlik (1) kısa gösterim olarak Eşitlik (2)'de verilen biçimde tekrar yazılabilir.

$$y_{(t)} = z'_{(t)}b + x_{(t)} \tag{2}$$

Eşitlik 2'de $\boldsymbol{b}$; ω, α ve β katsayılarıyla bir vektördür ve $\mathbf{z}'_{(t)}$ kolon değişkenleri ile $\left[cal_{1(t)},...,cal_{n_c(t)}, \lambda_1(B)d_{1(t)},...,\lambda_{n_{out}}(B)d_{n_{out}(t)}, reg_{1(t)},...,reg_{n(t)}\right]$ bir matris ifade eder.

Bir ARIMA model izlediği kabul edilebilen zaman serisinden çıkarılması gereken etkiler, Eşitlik (2) ifadesinin ilk teriminde gösterilir ve böylece öndüzeltme bileşeni kapsanmış olur. $X_{(t)}$ için ARIMA modelinin kısaltılmış ifadesi (Box ve Jenkins, 1970);

$$\phi(B)\delta(B)X_{(t)} = \theta(B)a_{(t)} \tag{3}$$

biçiminde yeniden yazılabilir. Burada;

$a_{(t)}$: $N(0,\sigma_a^2)$ dağılımlı beyaz gürültü sürecini, (beyaz gürültü süreci terimi; sıfır ortalamalı ve σ_a^2 varyanslı, özdeş, bağımsız ve normal dağılımlı değişkeni ifade eder)

$\phi(B)$, $\delta(B)$, $\theta(B)$: B'de sonlu polinomları,

$\phi(B)$: durağan otoregresif kökünü,

$\delta(B)$: durağan olmayan AR[8] kökünü,

$\theta(B)$: ters çevrilebilir hareketli ortalama (MA[9]) polinomunu ifade etmektedir.
Genellikle çarpımsal form;

$$\delta(B) = \nabla^d \nabla_s^{ds}$$
$$\phi(B) = \left(1 + \phi_1 B + ... + \phi_p B^p\right)\left(1 + \Phi_1 B^s\right)$$
$$\theta(B) = \left(1 + \theta_1 B + ... + \theta_q B^q\right)\left(1 + \Theta_1 B^s\right)$$

olarak ifade edilir ve burada d fark alma sayısı, $\boldsymbol{d_s}$ mevsimsel fark alma sayısı, $\nabla = 1 - B$ ve $\nabla_s = 1 - B^s$ düzenli ve mevsimsel fark alma işlemcisidir. Uygulamada, Eşitlik (2) ve Eşitlik (3)'te ifade edilen Reg-ARIMA modelleri ele alınmaktadır. Daha

[8] AR: Autoregressive
[9] MA: Moving Average

sonraki aşamada, öndüzeltilmiş seri, bir ARIMA-Model-Taban yöntemine uyan, gözlemlenmemiş bileşenlerine ayrıştırılmaktadır.

Öndüzeltme işlemleri olarak, seride herhangi bir logaritmik dönüşüme ihtiyaç olup olmayacağına karar verilir ve olası takvim etkilerinin varlığı için test uygulanır. Daha sonra üç tip aykırı değer belirlenmeye çalışılır. Karşılaşılan farklı türde aykırı değerler; ek uç değer (Additive Outlier (AO): $e_t = a_t + \omega_A d_{t0}(t)$), geçici değişme (temporary change-TC; $e_t = a_t + \omega_T /(1-\eta B) d_{t0}(t)$), düzey kayması (level shift-LS; $e_t = a_t + \omega_L /(1-B) d_{t0}(t)$)'dır. Bir AO, verideki tek bir noktayı yakalar, bir TC yumuşak bir dönüşle izlenen tek bir nokta değişimini gösterir ve bu iki tip aykırı değer serinin düzensiz bileşeninde gerçekleşir. LS, serinin eğilim düzeyinde gerçekleşen kalıcı bir kaymayı gösterir. Bu durumların hepsinde, seriyi etkileyen her olay oluştuğu zamanla birlikte bilinir. AO, TC ya da LS'yi bağlayıcı (regresör) olarak belirlemek analiste, etkinin belirginliğini ölçme olanağı sağlar. Bu tür analiz, müdahale (intervention) analizi olarak bilinir (Box ve Tiao, 1975). Aykırı değerlerin kaldırılması önemlidir çünkü, örneklem ACF[10] ve PACF[11]'sini kirletebilir. Örneğin, dikkate alınmayan AO fazla fark almaya neden olabilir (Phillips ve Perron, 1988). Aykırı değer tespiti, tanımlanması ve tahmini için metodolojiyi Chang vd. (1988) incelemiştir. Bir aykırı değer bulunduğunda, bu aykırı değer için yorum yapılıp yapılamayacağına karar verilmesi analizin önemli bir safhasını oluşturmaktadır. Modelleme ve tahmin araçları veride düzensizlik olduğunu gösterdiğinde bu düzensizlik açıklanmaya çalışılmalıdır.

2.2 ARIMA Zaman Serilerinin Bileşenlerine Ayrıştırılması

ARIMA modeline uygun olan serilerde gözlemlenemeyen bileşenleri tahmin etmek için ARIMA model tabanlı ayrıştırma kullanılır. Gözlemlenemeyen bileşenler; eğilim $T_{(t)}$, mevsimsel $S_{(t)}$ ve düzensiz $N_{(t)}$ bileşendir ve aşağıdaki şekilde ifade edilir (Hillmer ve Tiao, 1982);

$$X_t = S_t + T_t + N_t \tag{4}$$

X_t için S_t, T_t ve N_t'nin çarpılması durumu daha doğru bir gösterim olacağından Eşitlik (4) modeli orijinal serinin logaritmik dönüşümü için uygundur. Bileşenlerin her birinin bir ARIMA model izlediği kabul edilmektedir.

$$\begin{aligned} \phi_s(B) S_t &= \eta_s(B) b_t \\ \phi_T(B) T_t &= \eta_T(B) c_t \\ \phi_N(B) N_t &= \eta_N(B) d_t \end{aligned} \tag{5}$$

Polinom çiftlerinin her biri $\{\phi_S(B), \eta_S(B)\}$, $\{\phi_T(B), \eta_T(B)\}$ ve $\{\phi_N(B), \eta_N(B)\}$'nin birim çember üzerinde veya dışında ortak kökü yoktur. b_t, c_t ve d_t sırasıyla $N(0, \sigma_b^2)$,

[10] ACF: Autocorrelation Function
[11] PACF: Partial Autocorrelation Function

$N(0,\sigma_c^2)$, $N(0,\sigma_d^2)$ dağılımlı üç karşılıklı bağımsız beyaz gürültü sürecidir ve özdeş, bağımsız dağılımlı olarak temsil edilir. Sonra, X_t için genel modeller kolaylıkla ARIMA model olarak gösterilir.

$$\varphi(B)X_t = \theta(B)a_t \tag{6}$$

$\varphi(B)$; $\phi_S(B)$, $\phi_T(B)$ ve $\phi_N(B)$'nin en yüksek ortak çarpanıdır ve $\theta(B)$ ve σ_a^2;

$$\frac{\theta(B)\theta(F)\sigma_a^2}{\varphi(B)\varphi(F)} = \frac{\eta_s(B)\eta_s(F)\sigma_b^2}{\phi_s(B)\phi_s(F)} + \frac{\eta_T(B)\eta_T(F)\sigma_c^2}{\phi_T(B)\phi_T(F)} + \frac{\eta_N(B)\eta_N(F)\sigma_d^2}{\phi_N(B)\phi_N(F)} \tag{7}$$

ilişkisinden elde edilebilir ve burada $F = B^{-1}$'dir. Eşitlik (6)'da belirtilen parametrelerin bilindiği de varsayılmaktadır. Uygulamada X_t serisi için bir model veriden elde edilebilir ve tahmin edilen parametre değerleri doğruymuş gibi kullanılır. Çoğu gerçek durağan olmayan ve mevsimsel zaman serisinin davranışlarını tanımlayabilen yeterince esnek ARIMA kalıbı kurulabilir (Box ve Jenkins, 1970).

X_t için bir eğilim bileşeni T_t ve mevsimsel bileşen S_t'yi kabul etmeden önce $(1-B)^d$ ve $U(B)$ faktörünü içeren $\varphi(B)$'ye ihtiyaç duyulur. ***m*** yıllık gözlem sayısını göstermek üzere, $s_t - s_{t-m}$ için ardışık mevsimsel bileşen toplamı sıfır olacağından, $U(B)s_t = 0$ ve $U(B) = 1 + B + ... + B^{m-1}$'dir. Eşitlik (5)'te N_t'nin otoregresif polinomu $\phi_N(B)$'nin $(1-B)^d$ veya $U(B)$ her ikisinden biriyle ortak kökünün bulunmaması ek gereksinimdir, aksi taktirde, S_t ve T_t içinde yutulabilen ek mevsimsel ve eğilim bileşeninin varlığı anlamına gelecektir. Bundan dolayı Eşitlik (6)'nın

$$\varphi(B) = (1-B)^d U(B)\phi_N(B) \tag{8}$$

olduğu kabul edilmektedir. Burada sağ taraftaki üç faktörün ortak kökü yoktur. Bir başka ifadeyle, X_t için model biliniyor ve kabul edilebilir bir ayrıştırma mümkünse S_t, T_t ve N_t'nin otoregresif polinomları ayrı ayrı belirlenebilir. Aynı zamanda Eşitlik (7) ifadesi aşağıdaki gibi olur.

$$\frac{\theta(B)\theta(F)\sigma_a^2}{\varphi(B)\varphi(F)} = \frac{\eta_s(B)\eta_s(F)\sigma_b^2}{U(B)U(F)} + \frac{\eta_T(B)\eta_T(F)\sigma_c^2}{(1-B)^d(1-F)^d} + \frac{\eta_N(B)\eta_N(F)\sigma_d^2}{\phi_N(B)\phi_N(F)} \tag{9}$$

Hareketli ortalama polinomlarını ve her bir bileşen için hesaplanan hata varyanslarını ifade eden yenilik varyanslarını belirlemek daha da zor olan bir çalışmadır. En çok (*s*-1) ve *d* dereceli $\eta_S(B)$ ve $\eta_T(B)$ sınıfında σ_b^2, σ_c^2 ve σ_d^2 varyanslı $\eta_S(B)$, $\eta_T(B)$ ve $\eta_N(B)$ üç hareketli ortalama polinomunun herbirinin seçiminin Eşitlik (9)'u sağlaması kabul olunabilir bir ayrıştırma olarak adlandırılabilir, çünkü bu gözlemlenmiş veri X_t modeli tarafından sağlanan bilgiyle tutarlıdır.

2.3 Wiener-Kolmogrov Filtresi

Bileşen tahmin edicisi ve öngörüsü, gözlemlenmiş serinin sinyalinin MMSE'si (normallik varsayımı altında koşullu beklenen değerine eşittir) WK[12] filtresi aracılığıyla elde edilir (Whittle, 1963). WK filtresi; iki yönlü, merkezi, simetrik ve yakınsak filtre olarak AMB çerçevesinde basit analitik gösterimle verilebilir. $X_{(t)}$ serisinin ayrıştırması ele alındığında ARIMA model,

$$\phi(B)X_{(t)} = \theta(B)a_{(t)} \qquad (10)$$

burada $a_{(t)} \sim N(0,\sigma_a^2)$ dağılımına sahip beyaz gürültü sürecidir ve $\phi(B)$ polinomu birim kökleri içerir, "sinyal artı sinyal olmayan" bileşenler $X_{(t)} = s_{(t)} + n_{(t)}$ ve $n_{(t)} = T_{(t)} + N_{(t)}$' dir.

$$\phi_s(B)s_{(t)} = \theta_s(B)a_{s(t)}$$

olarak ifade edildiğinde; $a_{s(t)} \sim N(0,\sigma_s^2)$ dağılımına sahip beyaz gürültü sürecidir ve $\phi_s(B)$ polinomu birim kökleri içerir. F ileri doğru öteleme işlemcisidir ve $F = B^{-1}$'dir. WK filtresi, sinyali tahmin etmede aşağıdaki eşitlikleri kullanır:

$$\hat{S}_t = W_s(B)X_t \quad \text{ve} \quad \hat{T}_t = W_T(B)X_t\,. \qquad (11)$$

Burada

$$W_s(B) = \frac{\sigma_b^2}{\sigma_a^2}\frac{\varphi(B)\varphi(F)\eta_s(B)\eta_s(F)}{\theta(B)\theta(F)\phi_s(B)\phi_s(F)}$$

ve

$$W_T(B) = \frac{\sigma_c^2}{\sigma_a^2}\frac{\varphi(B)\varphi(F)\eta_T(B)\eta_T(F)}{\theta(B)\theta(F)\phi_T(B)\phi_T(F)}$$

olarak gösterilir. Uygulamada, $\left[X_{(1)},X_{(2)},...,X_{(T)}\right]$ gibi bir sonlu seri kullanılır. Genel olarak, verilen $\left[X_{(1)},X_{(2)},...,X_{(T)}\right]$ serisinin bileşenlerinin (sinyallerinin) öngörü ve MMSE tahmin edicileri ileriye ve geriye yönelik tahminlerle genişletilmiş serilere WK filtresi uygulanmasıyla elde edilir.

[12] WK: Wiener Kolmogrov

2.4 Kanonikal Ayrıştırma

Kanonikal ayrıştırma, gürültü (noise) dağıtımı problemi için bazı ek faktörleri kullanan yöntemdir. Bileşenlerin bağımsızlığı varsayımı, $g_x(\omega) = g_s(\omega) + g_n(\omega)$ ilişkisini oluşturur. Burman (1980)'de olduğu gibi, $\varepsilon_s = \min_\omega g_s(\omega)$ ve $\varepsilon_n = \min_\omega g_n(\omega)$ olarak ifade edilir. $\varepsilon_s + \varepsilon_n$ nicelikleri, gözlemlenen serilerin spektrumunda somutlaştırılan saf gürültü bileşenlerinin varyansı olarak kabul edilebilir. Bileşenlere ne kadar gürültü ayrılacağı bilinmediğinden belirleme probleminin doğduğu açıktır. ε_s *ve* ε_n'nin çok küçük bir miktarı, her bileşen spektrumundan ayrılabilir ve diğer bileşene tahsis edilebilir. Eğer s_t'den mümkün olduğu kadar gürültü ayrılır ve bu n_t'ye eklenirse; $g_s^0(\omega) = g_s(\omega) - \varepsilon_s$ ve $g_n^0(\omega) = g_n(\omega) + \varepsilon_s$ elde edilir. Bu ayrıştırma kanonikal olarak bilinir.

Kanonikal ayrıştırma yöntemi ilk olarak Box vd. (1978) ve Pierce (1978) tarafından gösterilmiştir. Yaklaşım, mümkün olduğunca gürültüden arındırılmış bileşenin belirlenmesini sağlar. MA polinomundaki birim köke karşılık gelen kanonikal sinyal spektrumunda sıfıra sahiptir. Bir başka ifade ile kanonikal sinyal ters çevrilemez. Kanonikal ayrıştırmanın önemli bir özelliği, sinyal için kabul olunabilir modellerin, kanonikal ile bağımsız beyaz gürültünün toplamı şeklinde yazılabilmesidir. Ayrıca, Hillmer ve Tiao (1982), kanonikal ayrıştırmanın sinyal yenilik varyansını minimize ettiğini göstermiştir. Düzensiz bileşen izole edileceği ve diğer bileşenler kanonikal olduğu zaman, düzensiz bileşenin varyansı maksimize olur. Kanonikal ayrıştırma, ARIMA Model Tabanlı yaklaşımlarda sıklıkla kullanılır.

s_t ve n_t'nin kanonikal şekilde tanımlanan bileşenler olduğu varsayılsın. Böylece, incelenen ayrıştırma $X_t = S_t + T_t + N_t$ olur. Burada N_t, maksimize edilen varyanslı beyaz gürültüyü gösterir. W_{s0} *ve* W_{T0} katsayıları, kanonikal sinyali ve kanonikal sinyal olmayanı tahmin etmek için planlanan WK filtrelerinin katsayılarını gösterir. Tahmin hatasını minimize eden kanonikal ayrıştırma, tahmin ediciler arasındaki kovaryansı da minimize eder. Bununla birlikte, modelin bütün gürültüsü, nispeten daha önemli olan bileşene tahsis edilmelidir. Bu göreli önemlilik, bileşenleri kanonikal şekillerinde tahmin etmek için planlanan WK filtrelerinin ağırlıklarını karşılaştırarak, incelemek biçiminde gerçekleştirilir.

2.5 Frekans Alanı Analizi

Bir zaman serisinin dinamiği hakkında bilgi, X_t ve geçmişi arasındaki süren ilişki, sürecin otokorelasyon (ACF) incelemesi ile elde edilir. Bunun yanısıra, zaman serileri konusunun diğer bir özelliği, serinin gösterdiği hareketlerin düzenliliğidir. Stokastik süreçlerle ilgilenildiği için yorumlar doğrudan yapılamaz. Frekans alanında zaman serileri analizi için uygun bir araç, spektrumlar aracılığıyla verilir. Durağan bir stokastik süreç için, güç (power) spektrumu aşağıdaki gibi ifade edilir.

$$f(\omega) = \frac{1}{2\pi} \sum_{\lambda=-\infty}^{\infty} \gamma\lambda e^{-i\lambda\omega} \qquad (12)$$

$\omega \in [-\pi, \pi]$ radyanla ifade edilen frekans, i kompleks sayı $\sqrt{i} = -1$ ve γ_λ da λ gecikmesinin otokorelasyonudur. Bu çerçevede $\gamma_\lambda = \gamma_{-\lambda}$ olarak verildiğinde Eşitlik (12), diğer bir ifadeyle,

$$f(\omega) = \frac{1}{2\pi}\left[\gamma_0 + 2\sum_{\lambda=1}^{\infty}\gamma_\lambda \cos\lambda\omega\right] \tag{13}$$

yazılabilir. Bu durumda ilgilenilen uygulamalar için, $f(\omega)$ sıfır etrafında simetrik olacak ve böylece $[0, \pi]$ aralığındaki frekansların ele alınması yeterli olacaktır. Bazen kovaryanslardan korelasyonların elde edilmesiyle benzer bir yolla spektral yoğunluk tanımlanarak, güç spektrum γ_0'a bölünür (Priestley, 1981). Serideki diğer bütün hareketler $[-\pi, \pi]$ aralığında tanımlanır. Bir zaman serisinin spektrumu, ω frekanslı hareketlerin $f(\omega)$ serisinin varyansına katkısını tanımlar. Bu katkılar toplanarak,

$$\int_{-\pi}^{\pi} f(\omega)d\omega = \gamma_0 \tag{14}$$

elde edilir. *k* gecikmeli kovaryanslar ayrıca $f(\omega)$ spektrumundan elde edilebilir ve özel olarak,

$$\int_{-\pi}^{\pi} e^{ik\omega} f(\omega)d\omega = \gamma_k \tag{15}$$

yazılır. Spektral üreten fonksiyon (SGF[13]), bu analizde daha faydalı olmaktadır. $\gamma(\lambda)$ kovaryanslarının yerine ACGF[14] ile bulunan $\gamma(B)$ konulmasıyla elde edilir. Böylece SGF kolayca $\gamma(e^{-i\omega})$ olarak tanımlanır. Durağan bir ARMA süreci için elde edilen ACGF'de;

$$\gamma(B) = \sigma_a^2 \frac{\theta(B)\theta(F)}{\phi(B)\phi(F)} \tag{16}$$

olmaktadır ve gecikme işlemcisi B'nin yerine geçen $e^{-i\omega}$ dönüşümüyle,

$$g(\omega) = \gamma(e^{-i\omega}) = \sigma_a^2 \frac{\theta(e^{-i\omega})\theta(e^{i\omega})}{\phi(e^{-i\omega})\phi(e^{i\omega})} \tag{17}$$

elde edilir. Güç spektrumu ve SGF $2\pi f(\omega) = g(\omega)$ olacak şekilde ilişkilidir. Verilen modellerin otokovaryans fonksiyonları çıkarıldığı zaman, güç spektrumun ya da SGF'nin hesaplamaları önemsizdir. Bu, Fourier dönüşümüdür. Güç spektrumun ters Fourier dönüşümü otokovaryansları verir.

[13] SGF: Spectrum Generating Function
[14] ACGF: Autocovariance Generating Function

Frekans alanında filtrenin seride çalışmasına bakmak, son (final) tahmin edicileri ve bileşenleri arasındaki farkın anlaşılmasına yardım eder. $\hat{s}_t$ tahmin edicisinin spektrumu $g_{\hat{s}}(\omega)$ olarak gösterildiğinde, Eşitlik (18) kullanılarak, Eşitlik (19) aşağıdaki biçimde yazılabilir,

$$W_s(B) = \sigma_s^2 \frac{\theta_s(B)\theta_s(F)\phi_n(B)\phi_n(F)}{\theta_x(B)\theta_x(F)} \tag{18}$$

$$\begin{aligned} g_{\hat{s}}(\omega) &= W_s(e^{-i\omega})W_s(e^{i\omega})g_x(\omega) \\ &= \left[\frac{g_s(\omega)}{g_x(\omega)}\right]^2 g_x(\omega) \end{aligned} \tag{19}$$

Frekans tepki fonksiyonu da aşağıdaki gibi yazılabilir,

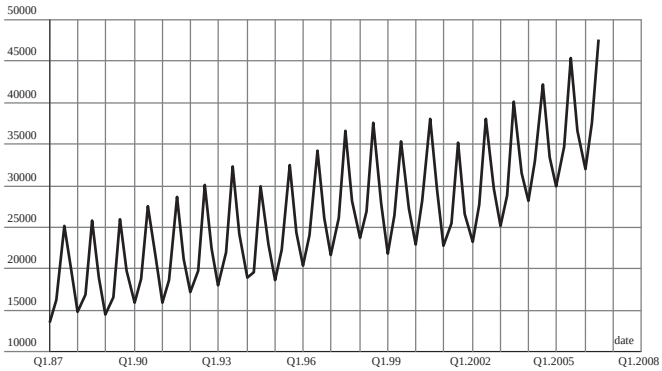
$$\frac{g_s(\omega)}{g_x(\omega)} = \frac{1}{1+\dfrac{g_n(\omega)}{g_s(\omega)}} \tag{20}$$

Bu özellik, WK filtre mekanizması hakkında aşağıdaki yorumların yapılabilmesini sağlar. Sinyalin göreli katkısı ω^* özel frekansında yüksek olduğu zaman, $g_n(\omega^*)/g_s(\omega^*) \approx 0$ olur. Böylece frekans tepki fonksiyonu 1'e yakın olur ve $g_{\hat{s}}(\omega^*) \approx g_s(\omega^*)$ sağlanır. Bileşen ve tahmin edicisi, ω^* frekansı etrafında benzer değişkenliği olan hareketler gösterecektir. Ayrıca, bu frekansla filtrenin artımı (gain) 1'e yakın olur ve böylece $g_{\hat{s}}(\omega^*) \approx g_x(\omega^*)$ elde edilir. Artım hareketlerin genişliği ile ilgilidir. Artımın karesi, ω frekanslarındaki özel bir genişliğe sahip hareketlerin çıktı serisine taşınmasındaki dereceyi kontrol eder. Gözlemlenen serinin spektrumunun çoğu, sinyal tahmininde kullanılır. Aksi durumda, görece katkısı özel bir frekansta düşük olduğu zaman, WK filtresi sinyal tahmini için bu değeri kullanmaz.

Genel olarak, eğilim bileşeni serilerin düşük frekansta $(\omega = 0)$ değişimini tutar ve sıfır sıklığında bir spektral uç gösterir. Mevsimsel bileşen sırası ile mevsimsel sıklıklarda (çeyreklik seriler için; $\omega = \pi/2$ ve $\omega = \pi$) spektral uçları tutar ve düzensiz bileşen beyaz gürültü davranışını gösterir ve bu nedenle düz spektruma sahiptir.

3. BULGULAR

Uygulamada, Türkiye İstatistik Kurumu (TÜİK), Gayri Safi Yurtiçi Hasıla (GSYİH), 1987=sabit fiyatlarıyla üç aylık (Bin YTL) serisi kullanılmıştır. Dönem 1987 (I)- 2006 (IV) olup, gözlem sayısı 80'dir. Orijinal seri Şekil 1'de gösterilmektedir.



Şekil 1. Orijinal seri (GSYİH)

3.1 Ön Düzeltme İşlemleri

İlk aşamada seriye logaritmik dönüşüm uygulanıp uygulanmayacağına karar verildikten sonra, takvim etkileri ve aykırı değer etkilerinin varlığının test edilmesi ve düzeltilmesi işlemlerini kapsayan öndüzeltme yöntemleri uygulanmıştır. Ticaret günü (Trading day) etkilerinin belirlenmesinde -Pazar hariç- 6 bağlayıcı ve Artık yıl etkisinin (Leap year effect) belirlenmesi için 1, toplam 7 bağlayıcı kullanılmaktadır. Seride hareketli tatil etkisi ve aykırı değer etkisi araştırılmıştır. Bu süreçte anlamlı bulunan etkiler seriden çıkarılmış ve öndüzeltme işlemi tamamlandıktan sonra model belirleme aşamasına geçilmiştir. Köşeli parantez içinde bulunan değerler %95 güvenilirlik düzeyinde ilgili test istatistiğinin alt ve üst sınır değerini ifade etmektedir.

Seri için model belirlenmeden önce öndüzeltme işlemlerinden takvim etkisi, hareketli tatil etkisi ve aykırı değer etkisi düzeltmesi yapılmaması durumunda ve yapılması durumunda belirlenen modeller sırasıyla; ARIMA modeli $(0,1,0)(0,1,1)_4$, ARIMA modeli $(0,1,0)(0,1,1)_4$, ARIMA modeli $(0,1,1)(0,1,1)_4$, ARIMA modeli $(0,1,0)(0,1,1)_4$'tür. Bu modellere ait model parametre uygunluk kontrol bilgileri ve model uygunluk kontrol bilgilerini içeren sayısal değerler Tablo 1'den Tablo 8'e kadar sıralı olarak verilmiştir.

Model uygunluk kontrol bilgileri tablosu ile ilgili açıklamalar

1) $t(\mu_a)$: H_0 artıklar ortalaması= sıfır ile ilgili t istatistiği değeridir.

2) $Q_a(12)$: artıklarda otokorelasyon için "portmanteau" Ljung-Box testidir (12) otokorelasyon için hesaplanır ve $\chi^2(10)$ asimptotik dağılımlıdır.

3) N_a : artıkların dağılımının normalliği için Jarque-Bera testidir ($\chi^2(2)$).

4) $t_a(skew)$: H_0 artıklar çarpıklık (skewness)= sıfır ile ilgili t istatistiği değeridir.

5) $t_a(kurt)$: H_0 artıklar basıklık (kurtosis)= 3 ile ilgili t istatistiği değeridir.

6) $Q_{as}(2)$: mevsimsel gecikmeli artıklarda otokorelasyon varlığı için Box-Pierce testidir.

7) $t_a(runs)$: H_0 artıkların işaretleri rastgeledir ile ilgili t istatistiği değeridir.

8) Her bir test için kritik değerler %95 güvenilirlik düzeyine göre hesaplanmaktadır.

Ticaret günü etkisi, hareketli tatil etkisi ve aykırı değer düzeltmesi yapılmadığında elde edilen bulgular:

ARIMA modeli $(0,1,0)(0,1,1)_4$

Tablo 1. Model parametre uygunluk kontrol bilgileri

Parametre	Tahmin	Std Hata	T değeri	Periyot	AIC	BIC
Θ	-0.7567	0.0755	-10.02	4	1233.66	13.60

Tablo 2. Model uygunluk kontrol bilgileri

	$t(\mu_a)$	$Q_a(12)$	N_a	t_a(skew)	t_a(kurt)	$Q_{as}(2)$	t_a(runs)
Orijinal Seri	0.45	10.45	0.86	-0.26	2.93	**6.16**	**2.09**
Kritik Değer (%95)	[-1.99, 1.99]	[0, 25.70]	[0, 5.99]	[-0.55, 0.55]	[1.89, 4.11]	[0, 5.99]	[-1.99, 1.99]

Mevsimsel gecikmeli artıklarda otokorelasyon belirlenmiştir ($Q_{as}(2) = 6.16$). Artıkların işaretleri rastgele olduğu hipotezi reddedilmiştir ($t_a(\text{runs}) = 2.09$).

Ticaret günü etkisi, hareketli tatil etkisi düzeltmesi yapıldığında, aykırı değer düzeltmesi yapılmadığında elde edilen bulgular:

ARIMA modeli $(0,1,0)(0,1,1)_4$

Tablo 3. Model parametre uygunluk kontrol bilgileri

Parametre	Tahmin	Std Hata	T değeri	Periyot	AIC	BIC
Θ	-0.9431	0.0384	-24.55	4	1239	13.85

Tablo 4. Model uygunluk kontrol bilgileri

	$t(\mu_a)$	$Q_a(12)$	N_a	t_a(skew)	t_a(kurt)	$Q_{as}(2)$	t_a(runs)
Orijinal Seri	0.64	11.32	2.07	-0.43	3.03	**6.63**	1.23
Kritik Değer (%95)	[-1.99, 1.99]	[0, 25.70]	[0, 5.99]	[-0.59, 0.59]	[1.83, 4.17]	[0, 5.99]	[-1.99, 1.99]

Mevsimsel gecikmeli artıklarda otokorelasyon belirlenmiştir ($Q_{as}(2) = 6.63$). Hareketli tatil etkisi istatistiksel olarak anlamlı bulunmuştur (-3.91 [-1.99, 1.99]). Ticaret günü etkisi istatistiksel olarak anlamlı bulunmamıştır.

Ticaret günü etkisi, hareketli tatil etkisi düzeltmesi yapılmadığında, aykırı değer düzeltmesi yapıldığında elde edilen bulgular:

ARIMA modeli $(0,1,1)(0,1,1)_4$

Tablo 5. Model parametre uygunluk kontrol bilgileri

Parametre	Tahmin	Std Hata	T değeri	Periyot	AIC	BIC
θ_1	-0.7561	0.0800	-9.46	4	1217.77	13.79
Θ	-0.6638	0.0981	-6.76			

Tablo 6. Model uygunluk kontrol bilgileri

	$t(\mu_a)$	$Q_a(12)$	N_a	t_a(skew)	t_a(kurt)	$Q_{as}(2)$	t_a(runs)
Orijinal Seri	0.86	**22.13**	0.70	-0.05	2.52	0.64	0.96
Kritik Değer (%95)	[-1.99, 1.99]	[0, 18.30]	[0, 5.99]	[-0.57, 0.57]	[1.86, 4.14]	[0, 5.99]	[-1.99, 1.99]

ARIMA model ile elde edilen artıklarda otokorelasyon belirlenmiştir $(Q_a(12) = 22.23)$. 1994 yılı 2. çeyrekte (Q2) t-değeri: -5.89 ([-3.075,3.075] %5) geçici değişme (TC), 1999 yılı 1. çeyrekte t-değeri: -6.32 ve 2001 yılı 1. çeyrekte (Q1) t-değeri: -8.29 ([-.075,3.075] %5) düzey kayması (LS) aykırı değerleri belirlenmiştir.

Ticaret günü etkisi, hareketli tatil etkisi düzeltmesi ve aykırı değer düzeltmesi yapıldığında elde edilen bulgular:

ARIMA modeli $(0,1,0)(0,1,1)_4$

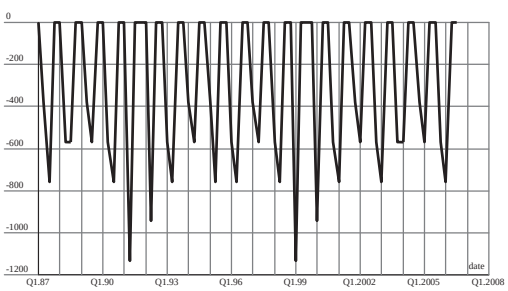
Tablo 7. Model parametre uygunluk kontrol bilgileri

Parametre	Tahmin	Std Hata	T değeri	Periyot	AIC	BIC
Θ	-0.6402	0.0887	-7.22	4	1208.74	13.56

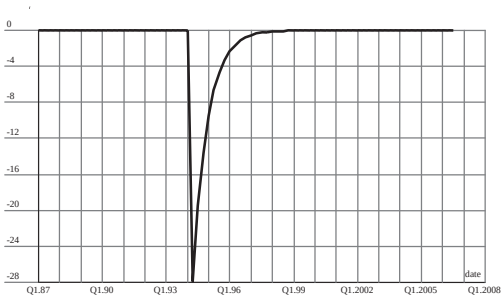
Tablo 8. Model uygunluk kontrol bilgileri

	$t(\mu_a)$	$Q_a(12)$	N_a	t_a(skew)	t_a(kurt)	$Q_{as}(2)$	t_a(runs)
Orijinal Seri	0.61	9.15	2.47	-0.26	2.21	1.28	1.25
Kritik Değer (%95)	[-1.99, 1.99]	[0, 19.70]	[0, 5.99]	[-0.57, 0.57]	[1.87, 4.13]	[0, 5.99]	[-1.99, 1.99]
	$t(\mu_a)$	$Q_a(12)$	N_a	t_a(skew)	t_a(kurt)	$Q_{as}(2)$	t_a(runs)
Mevsimsel Düzeltilmiş Seri	0.30	12.14	2.37	-0.42	3.14	4.33	-1.82
Kritik Değer (%95)	[-1.99, 1.99]	[0, 21.00]	[0, 5.99]	[-0.54, 0.54]	[1.91, 4.09]	[0, 5.99]	[-1.99, 1.99]

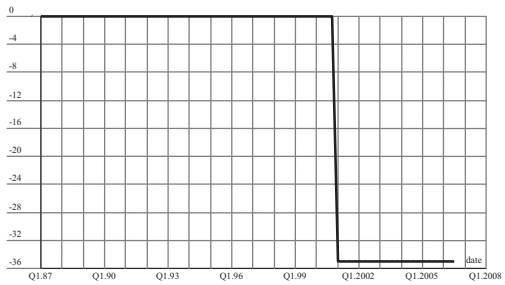
Seriye logaritmik dönüşüm uygulanmamıştır. Ticaret günleri arasında ortalama faaliyetten önemli bir şekilde sapma ve artık yıl etkisi istatistiksel olarak anlamsızdır. Ramazan bayramı ve kurban bayramı tatillerinin değişen gün sayısı ve zamanından kaynaklanan hareketli tatil etkisi, t-değeri: -4.17 ([-1.990,1.990] %5), anlamlı bulunmuş ve modele regresyon etkisi olarak dahil edilmiştir. Hareketli tatil etkisi Şekil 2'de gösterilmiştir. 1994 yılı 2. çeyrekte (Q2) t-değeri: -4.61 ([-3.075,3.075] %5) geçici değişme (TC) ve 2001 yılı 1.çeyrekte (Q1) t-değeri: -4.77 ([-3.075,3.075] %5) düzey kayması (LS) aykırı değerleri belirlenmiştir. Şekil 2'de hareketli tatil etkisi, Şekil 3 ve Şekil 4'te sırasıyla bu aykırı değer etkileri gösterilmektedir. Hareketli tatil etkisi ve aykırı değer etkisi düzeltmesi birlikte uygulanan ve model uygunluk testlerini geçen ARIMA $(0,1,0)(0,1,1)_4$ modeli seri için uygun model olarak kabul edilmiştir. Şekil 5 ve Şekil 6 sırasıyla Model ARIMA $(0,1,0)(0,1,1)_4$ artıklarını ve artıklar ACF'sini göstermektedir.



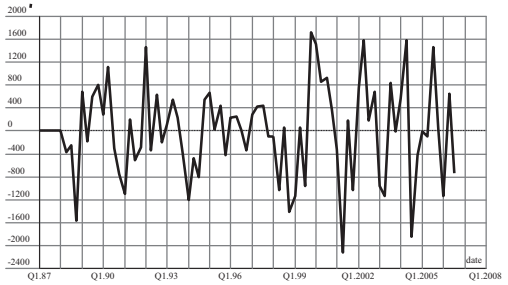
Şekil 2. Hareketli tatil etkisi



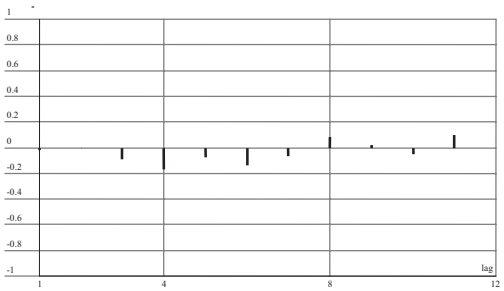
Şekil 3. Aykırı değer etkisi (Geçici Değişme – TC)



Şekil 4. Aykırı değer etkisi (Düzey Kayması – LS)



Şekil 5. Model ARIMA $(0,1,0)(0,1,1)_4$ artıklar



Şekil 6. Model ARIMA $(0,1,0)(0,1,1)_4$ artıklar ACF

3.2 Model Belirleme ve Bileşenlere Ayrıştırma

Öndüzeltilmiş seri $X_{(t)}$ ARIMA Model ve Parametreleri: $(0,1,0)(0,1,1)_4$

$$\nabla\nabla_4 X_{(t)} = (1-\Theta B^4)a_t \qquad (21)$$

$$\nabla\nabla_4 X_{(t)} = (1+0.6402B^4)a_{(t)}$$

Belirlenen ARIMA modeli $(0,1,0)(0,1,1)_4$'ün parametreleri En Çok Olabilirlik yöntemi ile belirlenmiştir. Modelin veriye uygunluğunda; artıklarda otokorelasyon testi Ljung-Box (1978) ve mevsimsel gecikmeli artıklarda Box-Pierce (1970) kullanılmıştır. Bir seri için birden çok modelin karşılaştırılmasında kullanılan seçim ölçütlerinden Akaike bilgi kriteri (AIC) (Hannan, 1980) ve Bayezyen bilgi kriteri (BIC) (Sneek, 1984), en uygun modelin belirlenmesinde dikkate alınmıştır. Bu ölçütlere göre en küçük değere sahip olan model en uygun modeldir.

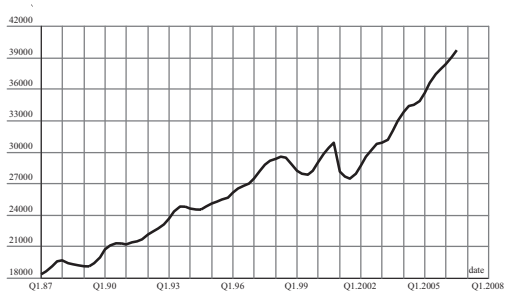
Bileşenlere ayrıştırma

Bileşenlere ayrıştırma için (sinyal çıkarımı) Eşitlik (21)'de ifade edilen öndüzeltilmiş seriyi filtrelemede kullanılan ARIMA modelleri; eğilim bileşeni için IMA(2,2), mevsimsel bileşen için ARMA(3,3)'tür. Uygulamalı ekonometride; öngörü ve mevsimsel düzeltme için IMA(2,2) modeli eğilim bileşeni için yaygın olarak

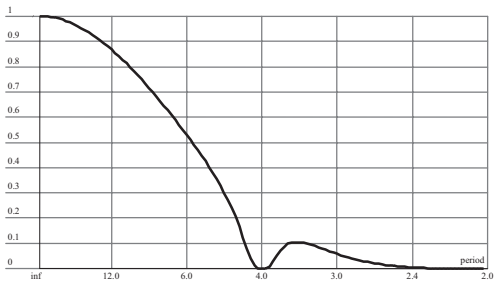
kullanılmaktadır. Eğilim, mevsimsel ve düzensiz bileşenlere ait bileşen ve filtre kareli artım şekilleri sırasıyla; Şekil 7, Şekil 8, Şekil 9, Şekil 10, Şekil 11, Şekil 12'de gösterilmektedir. Bileşenlere ait Wiener-Kolmogrov ağırlıkları da sırasıyla; Tablo 9, Tablo 10, Tablo 11'de verilmektedir.

Eğilim bileşeni;

$$\nabla^2 T_t = (1 + 0.1053B - 0.89471B^2)a_{T(t)}\ ,\ V(a_{T(t)}) = 0.18228,$$



Şekil 7. Eğilim bileşeni



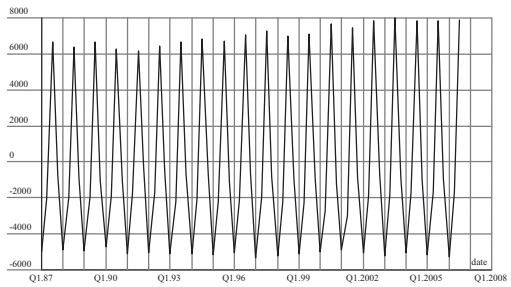
Şekil 8. Eğilim bileşen filtre kareli artım

Tablo 9. Eğilim bileşeni için Wiener-Kolmogrov ağırlıkları

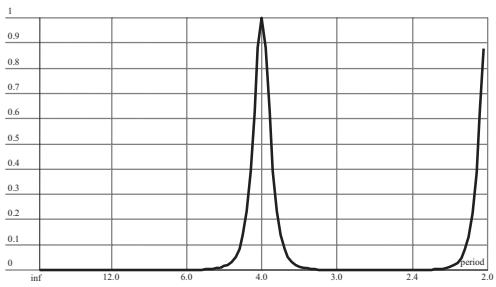
Gecikme j	W_j											
0-11	0.4501	0.2488	0.0450	0.0043	-0.0360	-0.0038	0.0288	0.0027	-0.0230	-0.0024	0.0184	0.0017
12-23	-0.0148	-0.0016	0.0118	0.0011	-0.0094	-0.0010	0.0076	0.0007	-0.0060	-0.0006	0.0048	0.0005
24-35	-0.0039	-0.0004	0.0031	0.0003	-0.0025	-0.0003	0.0020	0.0002	-0.0016	-0.0002	0.0013	0.0001
36-47	-0.0010	-0.0001	0.0008	0.0001	-0.0007	-0.0001	0.0005	0.0000	-0.0004	0.0000	0.0003	0.0000
48-60	-0.0003	0.0000	0.0002	0.0000	-0.0002	0.0000	0.0001	0.0000	-0.0001	0.0000	0.0001	0.0000

Mevsimsel bileşen;

$$Ss_{(t)} = (1 + 0.9961B + 0.3381B^2 - 0.4559B^3)a_{s(t)}\ ,\ V(a_{S(t)}) = 0.01271$$



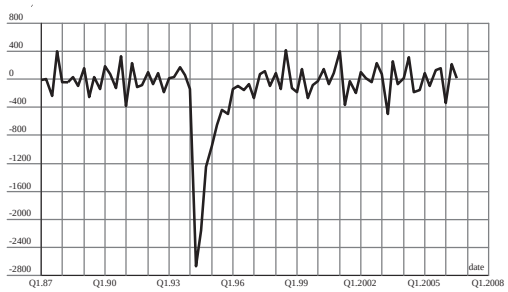
Şekil 9. Orijinal seri mevsimsel bileşen



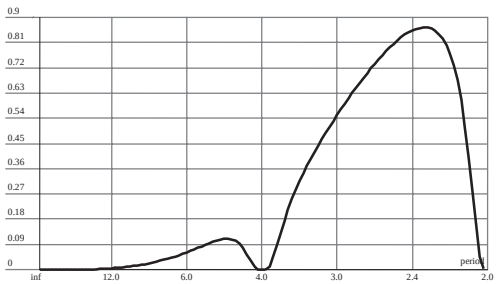
Şekil 10. Mevsimsel bileşen filtre kareli artım

Tablo 10. Mevsimsel bileşen için Wiener-Kolmogrov ağırlıkları

Gecikme j	W_j											
0-11	0.1381	-0.0428	-0.0450	-0.0413	0.1101	-0.0332	-0.0288	-0.0264	0.0705	-0.0213	-0.0184	-0.0169
12-23	0.0451	-0.0136	-0.0118	-0.0108	0.0289	-0.0087	-0.0076	-0.0069	0.0185	-0.0056	-0.0048	-0.0044
24-35	0.0118	-0.0036	-0.0031	-0.0028	0.0076	-0.0023	-0.0020	-0.0018	0.0049	-0.0015	-0.0013	-0.0012
36-47	0.0031	-0.0009	-0.0008	-0.0007	0.0020	-0.0006	-0.0005	-0.0005	0.0013	-0.0004	-0.0003	-0.0003
48-60	0.0008	-0.0002	-0.0002	-0.0002	0.0005	-0.0002	-0.0001	-0.0001	0.0003	-0.0001	-0.0001	-0.0001



Şekil 11. Düzensiz bileşen

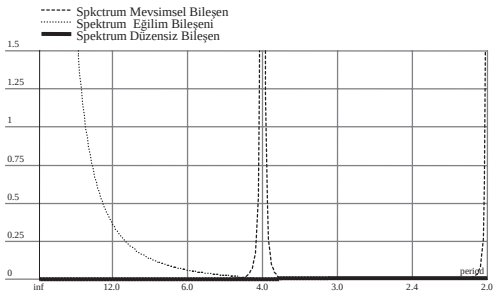


Şekil 12. Düzensiz bileşen filtre kareli artım

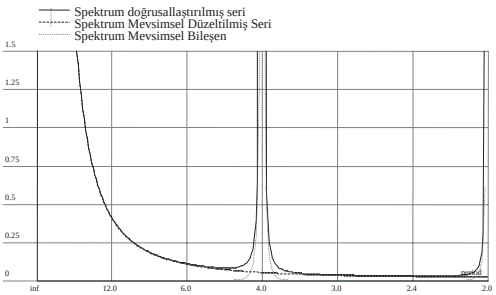
Tablo 11. Düzensiz bileşen için Wiener-Kolmogrov ağırlıkları

Gecikme j	W_j											
0-11	0.4118	-0.2059	0.0000	0.0370	-0.0741	0.0370	0.0000	0.0237	-0.0474	0.0237	0.0000	0.0152
11-23	-0.0304	0.0152	0.0000	0.0097	-0.0194	0.0097	0.0000	0.0062	-0.0124	0.0062	0.0000	0.0040
24-35	-0.0080	0.0040	0.0000	0.0026	-0.0051	0.0026	0.0000	0.0016	-0.0033	0.0016	0.0000	0.0010
36-47	-0.0021	0.0010	0.0000	0.0007	-0.0013	0.0007	0.0000	0.0004	-0.0009	0.0004	0.0000	0.0003
48-60	-0.0005	0.0003	0.0000	0.0002	-0.0004	0.0002	0.0000	0.0001	-0.0002	0.0001	0.0000	0.0001

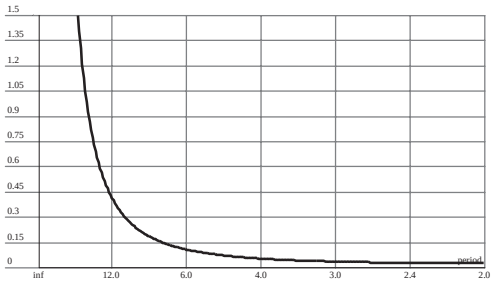
ve düzensiz bileşenin varyansı $V(a_{N(t)}) = 0.16887$ olarak elde edilmiştir. Eğilim IMA (2,2) süreci izler ve π radyan frekansında sıfır spektral ile birleştirilmiş MA polinomunun çarpanları (1+B)'yi gösterir. Bileşen spektrumlarının birlikte gösterildiği **Şekil 13'te** monoton azalan eğilim spektrumu görülmektedir ve sıfıra yakınsadığı noktada mevsimsel bileşen spektrumu uç göstermektedir. Bu durum AMB ayrıştırması içinde eğilim ve mevsimsel bileşenlerin belirlenmesi için kullanılan kanonik özelliği ifade eder. Mevsimsel bileşen, yıllık toplam operatörü ($S = 1 + B + ... + B^{f-1}$) tarafından verilen AR polinomu ile bir ARMA(3,3) süreçtir, bunun spektrumunda spektral sıfır, son iki harmonik arasında yer almaktadır. Düzensiz bileşenin spektrumu düz bir çizgi olarak görülmektedir. Mevsimsel düzeltilmiş serinin spektrumunda, Şekil 14'te, mevsimsellikle ilgili frekanslarda harmoniklere rastlanılmamaktadır. Eğilim, mevsimsel ve düzensiz bileşenin hangi frekanslarda serinin toplam varyansına ne tür bir katkı yaptığı sırasıyla Şekil 8, Şekil 10 ve Şekil 12'de görülebilir. Şekil 15'te serideki mevsimsel etkilerin olduğundan çok (overadjustment) veya olduğundan az (underadjustment) düzeltilmediğini gösteren düzeltilmiş seri, mevsimsel düzeltilmiş seri ve mevsimsel bileşenin spektrumları verilmektedir. Şekil 16'da orijinal seri ve mevsimsel etki arındırılmış seri gösterilmektedir.



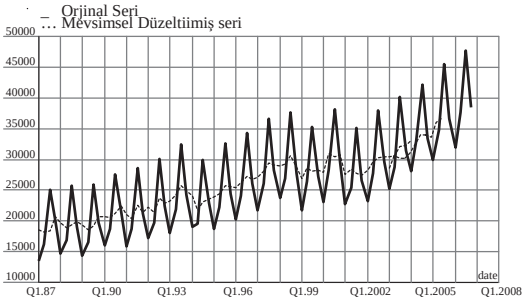
Şekil 13. Mevsimsel-eğilim-düzensiz bileşen spektrumları



Şekil 15. Düzeltilmiş seri, mevsimsel düzeltilmiş seri, mevsimsel bileşen (Spektrumlar)



Şekil 14. Mevsimsel düzeltilmiş seri spektrum



Şekil 16. Orijinal seri–mevsimsel düzeltilmiş seri

Bir zaman serisinde artık mevsimselliği test etmek için kullanılan yaklaşımlardan biri, mevsimsel olarak düzeltilmiş seriye, mevsimsel olmayan bir ARIMA model uydurmak ve artıklar için Ljung-Box testi uygulamaktır (Burman, 1980). Bu çalışmada mevsimsel düzeltilmiş seri ARIMA model (0,1,0) ile modellenebilmektedir. ARIMA model (0,1,0)'e ait model uygunluk test istatistikleri Tablo 8'de verilmektedir. Tablo 12'de bileşenlerine ayrıştırılan orijinal serinin bileşenleri arasındaki korelasyon yapısı gösterilmektedir.

Tablo 12. Bileşenler arasında çapraz korelasyon

	Eğilim	Mevsimsel Bileşen	Düzensiz Bileşen
Eğilim	1		
Mevsimsel Bileşen	0.0340314	1	
Düzensiz Bileşen	0.0343089	0.012605	1

4. TARTIŞMA VE SONUÇ

Bu çalışmada, TÜİK, GSYİH-1987=sabit fiyatlarıyla (üç aylık-Bin YTL) serisinin mevsimsel düzeltmesi için AMB yöntem kullanılmıştır. Orijinal seri için belirlenen model ARIMA $(0,1,0)(0,1,1)_4$, Eğilim bileşeni için IMA (2,2) ve Mevsimsel bileşen için ARMA(3,3)'tür. Düzensiz bileşen beyaz gürültü sürecidir. Mevsimsel düzeltilmiş seri tekrar modellenerek seride mevsimsellik (periyodik yapı) içerilip içerilmediği test edilmiştir. Mevsimsel düzeltilmiş seri için belirlenen model ARIMA(0,1,0)'dir. Bu sonuç, mevsimsel düzeltilmiş serinin eğilim ve düzensiz bileşenden oluştuğunu göstermektedir. Bileşenlerin karşılıklı olarak bağımsız olduğu, bileşenler arasında çapraz korelasyon değerleri Tablo 12'de verilmektedir.

Şekil 13 incelendiğinde; Eğilim, mevsimsel ve düzensiz bileşene ait spektrumların başarılı bir şekilde bileşenlerine ayrıştırılmış seriden beklenilen özellikleri karşıladığı görülmektedir. Bileşenlerin kareli artım (squared gain) filtreleri incelendiğinde ise eğilim ve düzensiz bileşenin mevsimsel sıklıklarda "0" değeri alması ve sadece mevsimsel bileşenin "1" değerini alması kanonikal ayrıştırma özelliğini göstermektedir (Şekil 8, 10, 12). Şekil 15, serideki mevsimsel etkinin olduğundan fazla veya az arındırılmadığını ifade etmektedir.

GSYİH serisinde iki aykırı değer belirlenmiştir. 1994 yılı 2. çeyrekte (Q2) t-değeri: -4.61 ([-3.063,3.063] %5) TC ve 2001 yılı 1. çeyrekte (Q1) t-değeri: -4.77 ([-3.063,3.063] %5) LS'dir. 1994 ve 2001 yıllarında ülkemizde yaşanan siyasi-ekonomik krizlerin etkisi her iki aykırı değerin nedeni olarak değerlendirilebilir. Hareketli tatil etkisi t istatistiği değeri -4.17'dir. Ramazan ve kurban bayramları GSYİH serisi üzerinde daha az çıktı üretilmesine neden olmaktadır.

Çalışmada ulaşılan diğer önemli bulgular ise seri için model belirlemeden önce öndüzeltme işlemlerinin gerekliliğini ortaya koymaktadır. Seri için bir model belirlenmeden önce hareketli tatil etkisi ve aykırı değer etkisi düzeltmesi yapılmaması durumunda; mevsimsel gecikmeli artıklarda otokorelasyon belirlenmiş ve artıkların işaretlerinin rastgeleliği hipotezi red edilmiştir. Hareketli tatil etkisi düzeltmesi yapılması, aykırı değer etkisi düzeltmesi yapılmaması durumunda; hareketli tatil etkisi anlamlı bulunmakla birlikte, mevsimsel gecikmeli artıklarda otokorelasyon belirlenmiştir. Hareketli tatil etkisi düzeltmesi yapılmaması, aykırı değer düzeltmesi yapılması durumunda, artıklarda otokorelasyon belirlenmiştir. Hareketli tatil etkilerinin kaldırılmaması artıklarda mevsimselliğe neden olabilmektedir. Türkiye için örnek çalışmaya Alper ve Bora (2004)'den bakılabilir. Ayrıca; hareketli tatil etkisi düzeltmesi yapılmaması durumdunda; 1994 yılı 2. çeyrekte (Q2) t-değeri: -5.89 ([-3.075,3.075] %5) geçici değişme (TC), 2001 yılı 1. çeyrekte (Q1) t-değeri: -8.29 ([-3.075,3.075] %5) düzey kayması (LS) belirlenmesinin yanısıra, 1999 yılı 1. çeyrekte t-değeri: -6.32 ([-3.075,3.075] %5) (LS) aykırı değeri belirlenmiştir. Bir diğer durum ise serinin diğer modelden farklı olarak ARIMA $(0,1,0)(0,1,1)_4$ ile modellenmesidir.

Bu çalışmada, zaman serilerini karşılıklı olarak bağımsız mevsimsel, eğilim ve düzensiz bileşenlerine ayrıştırmak için ARIMA model tabanlı yöntem üzerinde durulmuştur. Yöntem, farklı modeller içeren birçok zaman serisine uygulanabilir.

5. KAYNAKLAR

Alper C.E., and Bora S., 2004. Moving holidays and seasonal adjustment: The case of Turkey. Review of Middle East Economics and Finance, Volume: 2 , Issue: 3 , Pages: 203-209.

Bell, W.R., 1984. Signal extraction for nonstationary time series. The Annals of Statistics, 12, 2, 646-664.

Bell W.R., and Hillmer, S.C., 1984. Issues involved with the seasonal adjustment of economic time series. Journal of Business and Economic Statistics, 2, 291-320.

Box, G.E.P and Jenkins, G.M., 1970. Time series analysis: Forecasting and control. San Francisco, Holden Day.

Box, G.E.P, Hillmer S.C., and Tiao G.C., 1978. Analysis and modeling of seasonal time series, in seasonal analysis of time series. ed. A. Zellner, Washington, D.C. U.S. Department of Commerce, Bureau of the Census, 309-334.

Box, G.E.P., and Tiao, G.C., 1975. Intervention analysis with applications to economic and environmental problems. Journal of the American Statistical Association, 70, 71-79.

Box, G.E.P. and Pierce, David A., 1970. Distribution of residual autocorrelations in autoregressive-integrated moving average time series models. Journal of American Statistical Association, 65 (December), 1509-1526.

Burman, J.P., 1980. Seasonal adjustment by signal extraction. Journal of the Royal Statistical Society, Ser. A, 143, 321-337.

Chen C. and Liu L. M., 1993. Joint estimation of model parameters and outlier effects in time series. Journal of the American Statistical Association, 88, 284-297.

Chang, I., Tiao, G.C., and Chen, C., 1988. Estimation of time series models in the presence of outliers. Technometrics, 30, 2, 193-204.

Findley, D.F., Monsell, B.C, Bell, W.R., Otto, M.C. and Chen, S., 1998. New capabilities and methods of the X-12 ARIMA seasonal adjustment program (with discussion). Journal of Business and Economics Statistics, 16, 127-177.

Gómez, V., Maravall, A. and Peña, D., 1999. Missing observations in ARIMA models: skipping approach versus additive outlier approach. Journal of Econometrics, 88, 341-364.

Gómez, V. and Maravall, A., 2001a. Automatic modelling methods for univariate series, Ch.7 in Peña D., Tiao G.C. and Tsay, R.S. (eds.) A Course in Time Series Analysis. New York: J. Wiley and Sons.

Hannan, E.J., 1980. The estimation of the order of ARMA processes. Annals of Statistics, 8, 1071-1081.

Hillmer, S.C., 1985. Measures of variability for model-based seasonal adjustment procedures. Journal of Business and Economic Statistics, 3, 1, 60-68.

Hillmer S.C., and Tiao G.C., 1982. An ARIMA-model-based approach to seasonal adjustment. Journal of the American Statistical Association, 77, 63-70.

Hillmer, S.C., Bell, W.R. and Tiao, G.C., 1983. Modeling considerations in the seasonal adjustment of economic time series. in Zellner, A. (ed.), Applied time series analysis of economic data, Washington, D.C.. U.S. Department of Commerce. Bureau of the Census, 74-100.

Ljung, G. and Box G.E.P., 1978. On a measure of lack of fit in time series models. Biometrika, 65, 297-303.

Maravall, A., 1987. On minimum mean squared error estimation of the noise in unobserved component models. Journal of Business and Economic Statistics, 5, 115-120.

Maravall, A., 1995. Unobserved components in economic time series, in handbook of applied econometrics. (eds) Pesaran, M. H., and Wickens, Blackwell, Oxford.

Maravall, A. and Planas, C., 1999. Estimation error and the specification of unobserved component models. Journal of Econometrics, 92, 2, 325-353.

Nerlove, M., Grether, D.M, and Carvalho, J.L., 1979. Analysis of economic time series: A synthesis. New York, Academic Press.

Pierce, D.A., 1978. Seasonal adjustment when both deterministic and stochastic seasonality are present. in seasonal analysis of economic time series. ed. A. Zellner, Washington, D.C.. U.S. Dept. of Commerce, Bureau of the Census, 242-269.

Pierce, D.A., 1979. Signal extraction error in nonstationary time series. Annals of Statistics, 7, 1303-1320.

Pierce, D.A., 1980. Data revisions in moving average seasonal adjustment procedures. Journal of Econometrics, 14, 1, 95-114.

Phillips, P.C.B., Perron, P., 1988. Testing for unit roots in time series regression. Biometrika, 75, 335-346.

Priestley, M.B., 1981. Spectral analysis and time series. New York, Academic Press.

Sneek, M., 1984. Modelling procedures for economic time series. Amsterdam, Free University Press.

Whittle P., 1963. Prediction and regulation using least-square methods. London, English Universities Press.

AN ARIMA-MODEL-BASED APPROACH TO SEASONAL ADJUSTMENT

ABSTRACT

This article presents a model-based procedure to decompose a time series uniquely into mutually independent additive seasonal, trend, and irregular noise components. Estimators of components are calculated by Wiener-Kolmogrow (WK) filter. The series is assumed to follow the Gaussian ARIMA model. Properties of the procedure are discussed and an actual example is given. Demetra package programme was used at implementation.

Key Words: ARIMA model, Canonical decomposition, Signal extraction, Spectrum, Wiener-Kolmogrov filter.

Volume: 06 Number: 01 Category: 02 Page: 96-110 ISSN No: 1303-6319 Cilt: 06 Sayı: 01 Kategori: 02 Sayfa: 96-110

AÇIKLAYICI VE DOĞRULAYICI FAKTÖR ANALİZLERİNİN KARŞILAŞTIRILMASI: BİR UYGULAMA

Bilge ACAR BOLAT*

ÖZET

Çok Değişkenli İstatistik Yöntemlerden biri olan faktör analizi, aralarında ilişki bulunan çok sayıda değişkenin az sayıda faktörler şeklinde tanımlanmasını sağlamaktadır. Yöntem, çok sayıda değişkene ait özet bilgi vermekte ve boyut indirgeme ile sonuçların yorumlanmasını kolaylaştırmaktadır. Yaygın olarak iki faktör analizi yaklaşımı kullanılmaktadır. Bunlardan biri; Açıklayıcı Faktör Analizi, diğeri ise Doğrulayıcı Faktör Analizi'dir. Çalışmada her iki yaklaşım karşılaştırılmakta amaca uygun yaklaşımın seçimiyle ilgili genel bilgi verilmektedir.

Anahtar Kelimeler: Açıklayıcı faktör analizi, Doğrulayıcı faktör analizi.

1. GİRİŞ

Teorik olarak her değişkenin doğrudan gözlenememesi faktör analitik modelinin geliştirilmesine temel oluşturmuştur. Bu gözlenemeyen değişkenler faktör veya gizil değişken (factor-latent variable) olarak adlandırılmaktadır (Long, 1983). Faktör, doğrudan ölçülemeyen ancak bir veya birden fazla gözlenen değişken ile temsil edilebilen değişken olarak ifade edilmektedir. Örneğin, tüketicinin ürün alma davranışı kesin olarak ölçülememekte, ancak tüketiciye sorulacak sorularla değerlendirilebilmektedir (Hair vd., 1998, s.581). Zihinsel işleyiş test (IQ) sonucu, zeka faktörünü temsil etmek için kullanılabilmektedir.

Faktör analizi, faktörler ile bunları temsil eden gözlenen değişkenler arasındaki ilişkiyi ortaya çıkarmaktadır. Yaygın olarak kullanılan temel iki faktör analizi yaklaşımı bulunmaktadır. Önce, genelde *faktör analizi* adıyla ifade edilen, Açıklayıcı Faktör Analizi (AFA), daha sonra Doğrulayıcı Faktör Analizi (DFA) geliştirilmiştir.

Spearman (1904, 1927), hangi değişkenler arasında korelasyon olduğunu ve hangi değişkenlerin birlikte hareket ettiğini belirleyerek faktör modelinin temelini atmıştır. Spearman, faktör analizi tanımını ilk kez zekanın teorisini iki faktör ile incelediği çalışmasında ortaya koymuştur. Lawley (1940) çok sayıdaki gözlenen değişkene ait faktörleri modeline dahil etmiş, Thurstone (1947) faktörler arasında ilişkinin olabileceğini ortaya çıkararak faktor analizi uygulamalarını geliştirmişlerdir. Teorik faktörlerin varlığını test eden DFA, Howe (1955), Anderson ve Rubin (1956), Lawley (1958) tarafından geliştirilmiştir. Jöreskog (1967-1969), Jöreskog ve Lawley (1968) yaptığı çalışmalarla değişken setlerinin faktörleri tanımlayıp tanımlamadığını test ederek analizin şekillenmesini sağlamıştır (Kaplan, 2000; Schumacker, 2004).

* Araştırma Görevlisi, İ.Ü. İşletme Fakültesi, Sayısal Yöntemler Ana Bilim Dalı, e-posta: bacar@istanbul.edu.tr

DFA, AFA'dan farklı olarak her bir faktörü temsil edecek gözlenen değişkenleri önceden belirleme olanağı vermektedir. Genellikle DFA'da bir gözlenen değişkenin bir faktöre ait faktör yükü bulunmaktadır[15]. Oysaki AFA'da tüm değişkenlerin küçük de olsa faktörler üzerinde etkisi bulunmaktadır.

Genelde literatürde yaygın olarak kullanılan faktör analizi yaklaşımı AFA olmakta, DFA'da araştırmacıya teorik modelle ilgili hipotezlerin sınanması ve model geliştirme olanaklarını sunmaktadır.

2. YÖNTEM

AFA ve DFA modellerinin her ikisi de Genel Doğrusal Modeller (General Linear Model-GLM) arasında yer almaktadır. AFA, DFA'ya temel oluşturmakta ancak aralarında önemli farklılıklarda bulunmaktadır. Örneğin DFA'yı uygulamadan önce faktör sayısının, hangi gözlenen değişkenlerin hangi faktörü temsil ettiğini ve hangi faktörler arasında korelasyon olduğuna karar verilmesi gerekmektedir. Bir başka ifadeyle modelin belirlenmesi gerekmektedir. Aynı zamanda AFA'da yapılamayan bazı analizler DFA'da yapılabilmektedir. Örneğin hatalar arasında korelasyon DFA'da incelenebilmektedir. Ancak AFA'da yapılabilen faktör döndürmesi (factor rotation) DFA'da uygulanamamaktadır (Thompson, 2005).

Yöntemler ana hatlarıyla aşağıda açıklanmaktadır.

2.1 Açıklayıcı Faktör Analizi (AFA)

AFA matris formatında aşağıdaki gibi gösterilmektedir.

$$X = FA' + E \qquad (1)$$

Burada:
X : Bileşenler vektörünü,
F : Faktör skorları matrisini,
A' : Faktör yükü matrisini,
E : Hata matrisini göstermektedir (Tacq, 1999).

AFA 4 adımda gerçekleşmektedir:

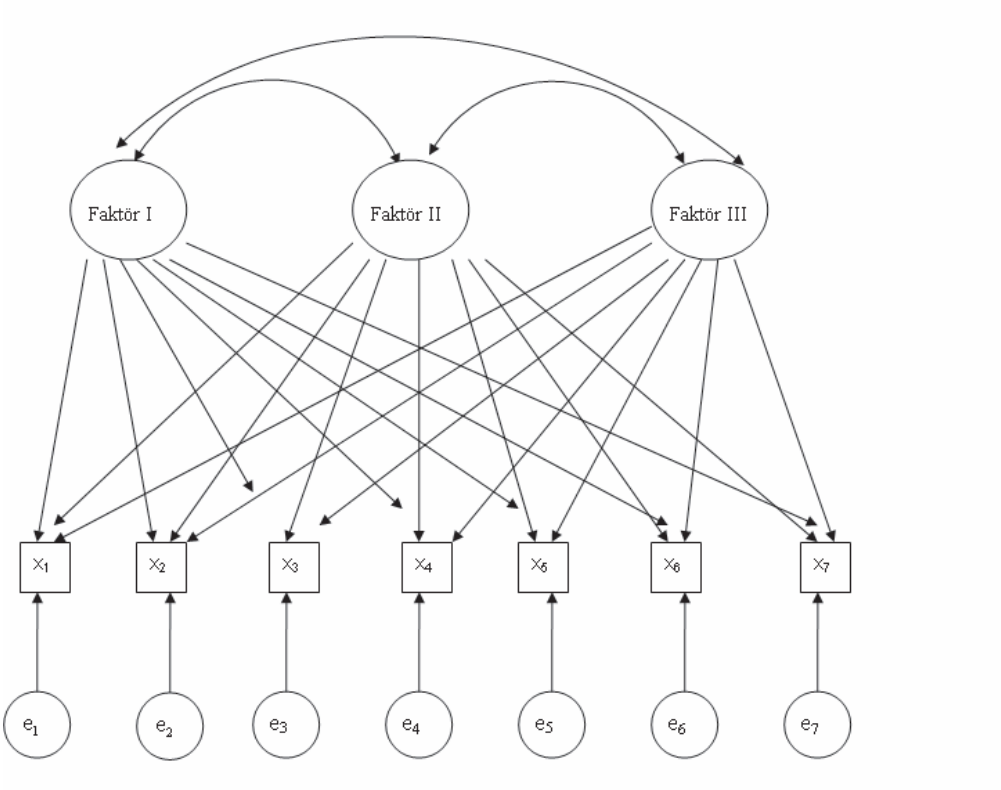
- *Korelasyon Matrisinin seçimi*: Korelasyon matrisi, orijinal değişkenlerden hesaplanıyorsa, gözlenen korelasyon matrisi (observed correlation matrix), faktörlerden hesaplanıyorsa, türetilmiş korelasyon matrisi (reproduced correlation matrix) adını almaktadır. İki korelasyon matrisi arasındaki farklılık hatalara ait korelasyon matrisini (residual correlation matrix) vermektedir. Hatalara ait matrisde korelasyonların küçük olması, her iki matris arasında uygunluğu göstermektedir (Tabachnick, Fidel,1996).

[15] Bir gözlenen değişkenin birden fazla faktörle de ilişkisi bulunabilmektedir. Ancak belirli teorik durumlar dışında uygulanması önerilmemektedir.

- *Faktör Sayısının Belirlenmesi*[Şekil 1]: Birden fazla karar kriteri olsa da yaygın olarak özdeğer (eigenvalues) istatistiği kullanılmakta, birden büyük öz değerler için faktörler anlamlı sayılmaktadır.
- *Faktör Döndürmesi*: Döndürme yapılmasındaki amaç elde edilen faktörlerin yorumlanmasını kolaylaştırmaktır. Faktör döndürmesinde genelde iki sınıflama yapılmaktadır. Bunlardan biri dik döndürme (orthogonal), ikincisi ise eğik döndürme (oblique) uygulanmasıdır. Dik döndürme uygulandığında, faktörler arasında korelasyon olmamakta, yük matrisi (loading matrix) türetilmektedir. Eğik döndürmede ise faktörler arasında korelasyon bulunmaktadır. Eğik döndürme iki farklı matrisi içermektedir. Bunlardan biri faktörler ve değişkenler arasındaki korelasyonları, faktör yüklerini gösteren yapı matrisi (structure mattrix) diğeride her bir faktörle her bir gözlenen değişkene ait tek bir ilişkiyi gösteren model matrisidir (pattern matrix). Dik döndürme yapıldığında her iki matris birbirine eşit olmaktadır. Eğik döndürme ile elde edilen model matrisinde faktörler isimlendirilmekte ve yorum yapılırken genelde bu matris kullanılmaktadır. Her iki döndürme türünde de faktör skor katsayı matrisi (factor-score coefficient matrix) elde edilmektedir (Stevens, 1996; Tabachnick ve Fidell, 1996; Rencher, 1995).

- *Sonuçların Yorumlanması*: Anlamlı faktörlerin türetilip türetilmediğine karar verilmesini ve faktörlerin isimlendirilmesini içermektedir.

Şekil 1'deki AFA modelinde (eğik döndürme yapıldığı varsayılmaktadır.), faktörler daire, gözlenen değişkenler ise kare şekliyle gösterilmiştir. Faktörden gözlenen değişkene doğru giden tek yönlü ok faktörün gözlenen değişken üzerindeki etkisini belirtmektedir. Faktörler arasındaki çift yönlü ok ise faktörler arasında korelasyon olduğunu belirtmektedir. Şekil'de 3 adet faktör bulunmakta, x_1'den x_7'ye kadar olan değişkenlerde gözlenen değişkenleri belirtmektedir. Gözlenen değişkenlere ait hatalar da (e) gösterimi ile şekilde yer almaktadır (Long, 1983; Thompson 2005).



Şekil 1. Eğik döndürme yapıldığı varsayılan açıklayıcı faktör analizi

AFA'da;

- Dik döndürme yöntemi seçilirse faktörler arasında korelasyon olmadığı, eğik döndürme yöntemi seçilirse korelasyon olduğu,
- Tüm gözlenen değişkenler ile faktörler arasında ilişki olduğu,
- Hatalar arasında korelasyon olmadığı,
- Her bir gözlenen değişkene ait hata terimi olduğu,
- Faktörler ile hatalar arasında ilişki olmadığı varsayılmaktadır.

2.2 Doğrulayıcı Faktör Analizi (DFA)

DFA'da, faktörlerle gözlenen değişkenler arasındaki ilişki matematiksel olarak aşağıdaki gibi ifade edilmektedir.

Faktör denklemi:

$$X = \Lambda\xi + \delta \qquad (2)$$

Kovaryans denklemi ise:

$$\Sigma = \Lambda\Phi\Lambda' + \Theta \qquad (3)$$

olarak formüle edilmektedir.

Burada:
X:Gözlenen değişkenlere ait vektörü (qx1),
ξ : Faktörlere ait vektörü (sx1),
Λ :Gözlenen değişkenlerin faktörler üzerindeki faktör yükleri matrisini (qxs),
δ :Ölçüm hatalarına ait vektörü belirtmektedir (qx1).

Gözlenen değişken sayısının faktör sayısından büyük olduğu varsayılmaktadır (q>s). Yukarıda belirtilen ifadeler özet olarak Tablo 1'de verilmektedir.

Tablo 1. Doğrulayıcı faktör analizi modeline ait özet bilgiler

Matris	Boyut	Ortalama	Kovaryans	Boyut	Tanım
ξ	(sx1)	0	$\Phi = E(\xi\xi')$	(sxs)	Faktör
x	(qx1)	0	$\Sigma = (xx')$	(qxq)	Gözlenen Değişken
Λ	(qxs)	-	-	-	Faktör Yükü
δ	(qx1)	0	$\Theta = E(\delta\delta')$	(qxq)	Hata terimi

DFA'daki varsayımlar aşağıdaki biçimde özetlenebilir:
- Değişkenlerin ortalamaları sıfırdır. $E(\xi) = 0, E(x) = E(\delta) = 0$.
- Gözlenen değişken sayısının, faktör sayısından daha fazla olması gerekmektedir (q>s).
- Faktörler ile hatalar arasında korelasyon bulunmamaktadır. $E(\xi\delta') = 0$ veya $E(\delta\xi') = 0$ ve $Kov(\delta\xi') = 0$ 'dır (Long, 1983, s.24-25).

DFA ise beş adımda gerçekleşmektedir.

1) Model belirleme

2) Model tanımlama

3) Model tahmini

4) Model uygunluğu

5) Modelin düzeltilmesi

Model Belirleme:

Model belirleme, araştırmanın yapılacağı konu ile ilgili önceden yapılmış çalışmalardan hareketle teorik modelin geliştirilmesini içermektedir. Teorik modelde, hangi değişkenlerin modelde yer alacağına karar verilmekte ve değişkenler arasındaki ilişkiler belirlenmektedir. Belirleme, modelin tespit edilmesinin ön çalışması olarak ifade edilebilir. Bu aşamada faktörlerin sayısı, faktörler arasındaki varyans-kovaryanslar, gözlenen değişken ile faktör arasındaki ilişkiler, gözlenen değişken ile hatalar arasındaki ilişkiler, faktörler arasındaki varyans-kovaryanslar belirlenmektedir. Bu aşamada parametrelerin büyüklüğü ve işaretleri belirlenirken, model parametreleri serbest parametre (free parameter) ve sabit parametre (fixed parameter) olarak belirlenmektedir.

Model Tanımlama:

DFA'da model tanımlama iki temele dayanmaktadır (Kline, 2005).

1) Varyans-kovaryans matrisindeki eleman sayısının serbest parametre sayısından daha fazla olması, bir başka ifade ile serbestlik derecesinin sıfıra eşit veya büyük olması ($df_m \geq 0$) gerekmektedir.

2) Her bir faktöre ait göstergelerin ve ölçüm hatalarının metrik olması gerekmektedir.

Standart DFA'da her faktörün en az 3 göstergesi varsa model tanımlanmış (identified), iki ve ikiden fazla faktörün olduğu modelde ise her bir faktör için en az 2 gösterge bulunması halinde model yine tanımlanmış olmaktadır. Ancak, küçük örneklem için gerçekleşecek modellerde iki göstergeli faktör tahmin aşamasında sorun yaratacağından genel olarak her bir faktör için en az üç göstergenin olması önerilmektedir.

Modelde q gözlenen değişken sayısını ifade etmek üzere, tahmin edilebilecek varyans-kovaryans sayısı en fazla;

$$[\, q(q+1)/2] \qquad (4)$$

kadardır. t, modelde tahmin edilen serbest parametre sayısını belirtmek üzere;

$t < q(q+1)/2 \Rightarrow$ modelin fazla tanımlanmış olduğu,

$t > q(q+1)/2 \Rightarrow$ tanımlanamadığı,

$t = q(q+1)/2 \Rightarrow$ tam tanımlanmış olduğu kabul edilmektedir (Kaplan, 2000).

İkisi arasındaki fark modelin serbestlik derecesini (df_m) vermektedir. Modelde en azından varyans-kovaryans matrisindeki eleman sayısının parametre sayısından fazla olması beklenmekte, bu durumda serbestlik derecesi de sıfırdan büyük olmaktadır ($df_m \geq 0$) (Kline, 2005).

Model Tahmini:

DFA'nın parametre tahmini için En Çok Olabilirlik Tahmini, Genelleştirilmiş En Küçük Kareler Yöntemi ve Ağırlıklandırılmamış En küçük Kareler Yöntemleri kullanılabilmektedir. DFA'da standardize edilmemiş tahminler aşağıdaki gibi yorumlanmaktadır. Faktör çiftleri veya ölçüm hataları arasındaki analiz edilemeyen önceki ilişkiler kovaryans olarak tanımlanmaktadır. Faktör yükleri, faktörün gösterge üzerindeki doğrudan etkisini gösteren standardize edilmemiş regresyon katsayıları gibi yorumlanmaktadır. Faktöre ait göstergelerin bire sabitlenmesi, ilgili faktörün standardize edilmemiş olduğunu ve standart hataları da olmadığından istatistiki anlamlılığının test edilemediğini belirtmektedir.

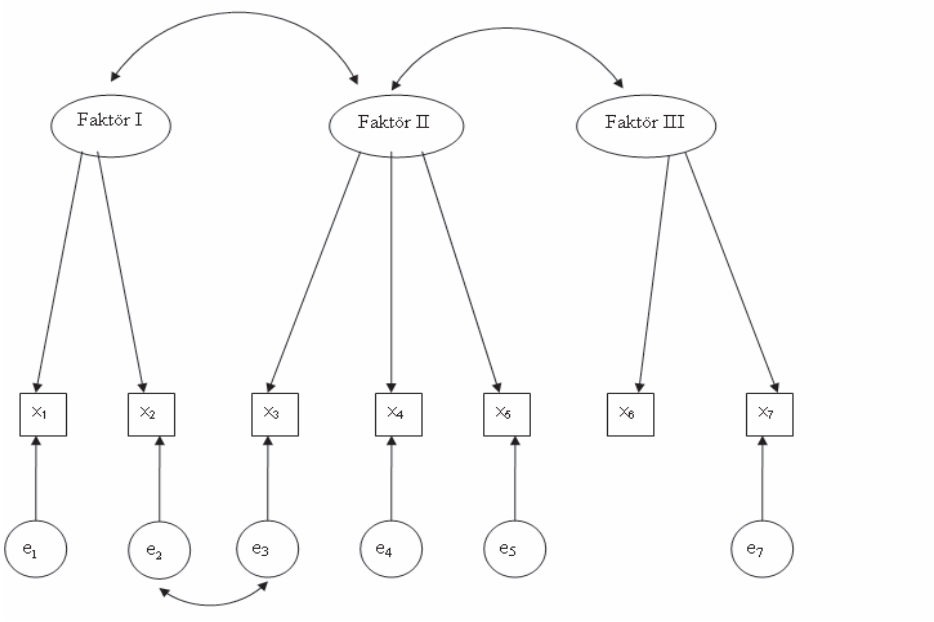
Model Uygunluğu:

DFA'da; Ki-Kare istatistiği, Uygunluk endeksi/Uyum iyiliği endeksi, Hata kareleri ortalamasının karekökü, Hata kareleri ortalaması yaklaşımı, Akaike bilgi kriteri Standartlaştırılmış uygunluk endeksi gibi birden çok uygunluk endeksi kullanılmaktadır. İyi uygun modeller farklı uygunluk endeksleri için benzer sonuçları vermektedir. Birçok endeks benzer sonucu veriyorsa, karar vermek için herhangi bir endeks tercih edilebilmektedir ancak genelde çoklu endeks sonuçlarına da yer verilmektedir. Uygunluk endeksleri sonuçları benzer değilse modelin yeniden incelenmesi gerekmektedir (Tabachnick, Fidell 1996).

Modelin Düzeltilmesi:

Modelin düzeltilmesi için iki neden vardır. Bunlardan biri modelin uygunluğunu arttırmak, diğeri de hipotezlerin test edilmesine olanak vermektedir. Modelin düzeltilmesi için serbest parametreler sabitlenerek veya sabit parametreler serbest parametreye dönüştürülerek uygun olmayan model yeniden belirlenmeye çalışılmaktadır (Hoyle, 1995).

DFA modeline ilişkin bir örnek Şekil 2'de gösterilmektedir.



Şekil 2:Doğrulayıcı faktör analizi

AFA'dan farklı olarak DFA'da:

- Hangi faktörler arasında ilişki (korelasyon) olduğu,
- Hangi faktörlerin hangi gözlenen değişkenden etkilendiği,
- Hangi gözlenen değişkenin hatadan etkilendiği,
- Hangi hatalar arasında ilişki olduğu belirlenebilmekte ve uygulanacak istatistik testler ile veri setinin uygunluğuna karar verilebilmektedir.
- AFA'da faktör yükleri için 0.30'dan büyük değerler anlamlı sayılırken, DFA'da her bir faktör yükünün anlamlılığı test edilebilmektedir.

AFA ve DFA arasındaki farklılıklar Şekil.1 ve Şekil.2 üzerinden rahatlıkla görülebilmektedir. DFA modeline örnek olarak gösterilen Şekil 2'deki modelde Faktör 1 ve Faktör 3 arasında korelasyon olmadığı varsayılmaktadır. Oysaki Şekil 1'deki AFA'da faktörler arasında ilişki olduğu varsayılmaktadır (alternatif olarak faktörler arasında ilişki olmadığı da varsayılabilmektedir). DFA'da bir gözlenen değişken sadece tek bir faktörü (örneğin x_3 değişkeni Faktör 1 ve Faktör 3'ten etkilenmemektedir) temsil etmektedir. DFA'da hatalar arasında ilişki olabilmekte, şekilde görüldüğü gibi x_2 ve x_3 gözlenen değişkenine ait hatalar arasında (e_2 ve e_3) ilişki görülmektedir. DFA'da bazı gözlenen değişkenler hataya sahip olmamaktadır. Örneğin x_6 değişkeni hataya sahip değildir. Oysaki AFA'da hatalar birbirinden bağımsız ve her gözlenen değişkene ait hata terimi bulunmaktadır.

3. BULGULAR

Bilindiği gibi finansal oranlar genel olarak karlılık, likitide, mali yapı ve faaliyet oranları olmak üzere dört ana başlık altında toplanmaktadır. Finansal oranları içeren çok değişkenli istatistik yöntemlerin uygulandığı bir çok çalışmada, yapılacak olan analizden önce genelde AFA yöntemine başvurulmaktadır. Bunun nedeni ise çok sayıda oran arasından veri seti için temsil gücü yüksek olan oranların tespit edilmesi olarak görülmektedir. Çalışmada İstanbul Menkul Kıymetler Borsasında (İMKB) işlem gören 280 firmaya ait Ek.1'de yer alan 11 adet oran[16] (gözlenen değişken) hesaplanmış ve bu veri setine AFA ile DFA ayrı ayrı uygulanarak sonuçlar yorumlanmıştır. Analizler SPSS ve LISREL paket programlarıyla gerçekleştirilmiştir.

3.1 AFA Sonuçları:

AFA'da, faktörler arasındaki korelasyonun hesaplanmasına olanak veren eğik döndürme kullanılarak analiz uygulanmıştır. İyi uygunluk testi sonucuna göre veri setinin AFA için uygun olduğu ($p<0.05$), 4 adet faktörün elde edildiği ve bu faktörlerin toplam varyansı açıklama oranının yaklaşık % 71.76, birinci faktörün açıklama oranına katkısının ise %37.25 olduğu görülmektedir.

Model matrisi incelendiğinde birinci faktör mali yapı, ikincisi faaliyet, diğerleri ise sırasıyla likitide ve karlılık oranları olarak isimlendirilebilir.

Mali yapı faktörüne en çok etki eden oranın BORCOZS (0.987**)** olduğu görülmekte, bunu KALDIRAÇ oranı ve OZSAKT oranı izlemektedir.

[16] Yorumları kısaltmak amacıyla dört ana başlık altında yer alan oranlardan sadece Ek.1'de yer alanlar analize dahil edilmiştir.

Faaliyet faktöründe STOKDH (0.998), AKTDH ve ALCDH oranları dikkat çekmektedir.

Üçüncü faktör olan Likitide ise CARİ (0.988) ve ASIT oranları öne çıkmaktadır.

Dördüncü faktörde ise FVAKT (0.845), VKOZS, BKSAT oranları öne çıkmaktadır.

Tablo 2. Model matrisine ait sonuçlar

	Faktör			
	1	**2**	**3**	**4**
VKOZS				0.658
FVAKT				0.845
BKSAT				0.653
CARI			0.988	
ASIT			0.882	
BORCOZS	0.987			
OZSAKT	0.756			
KALDIRAC	-0.856			
ALCDH		0.362		
AKTDH		0.470		
STOKDH		0.998		

Tablo 3 incelendiğinde mali ve likitide faktörleri (0.458), kar ve likitide faktörleri arasında aynı yönde (0.275) bir ilişki olabileceği görülmektedir. Ancak AFA'da faktör yüklerinin anlamlığı ve faktörler arasındaki korelasyonun anlamlılığı test edilememektedir.

Tablo 3. Faktör korelasyon matrisi

Factor	1	2	3	4
1	1	0.111	0.458	0.094
2	0.111	1	-0.083	-0.165
3	0.458	-0.083	1	0.275
4	0.094	-0.165	0.275	1

3.2 DFA Sonuçları:

DFA'da analize başlamadan önce teoriye dayanarak finansal oranlar 4 faktör şeklinde belirlenmiştir. Analiz sonucunda modele ait şemasal gösterime bakıldığında her bir gözlenen değişkenin tek bir faktörle ilişkisinin olduğu görülmektedir (Bkz. Ek.2).

Tablo 4 incelendiğinde, her bir orana ait faktör yükü ve hangi oranların modelde yer alacağına karar verilmesini sağlayan "t değerleri" görülmektedir. Faktör yükünün karesi her bir oranın ilgili faktörü açıklama oranını göstermektedir. Karlılık faktörü en çok FVAKT oranı (R^2=0.64), likitide faktörü CARİ oran (R^2=0.94), mali yapı faktörü BORCOZS oranı (R^2=0.98), faaliyet faktörü de STOKDH (R^2=0.63) oranı tarafından açıklanmaktadır.

Tablo 4. Finansal oranlara ait faktör yükleri ve t değerleri

Faktör	Faktör Yükü	t değeri
Karlılık		
VKOZS	0.57	9.33 (0.011)
FVAKT	0.80	13.28 (0.0048)
BKSAT	0.79	13.05 (0.0071)
Likitide		
CARI	0.97	20.45 (0.043)
ASIT	0.94	19.44 (0.036)
Mali Yapı		
BORCOZS	0.99	22.74 (0.0055)
KALDIRAC	0.92	19.89 (0.014)
Faaliyet		
ALCDH	0.44	6.07 (0.35)
AKTDH	0.58	7.43 (0.72)
STOKDH	0.79	8.71 (0.044)

Yukarıdaki tablo incelendiğinde ALCDH ve AKTDH oranlarının anlamsız olduğu görülmektedir ($p>0.05$).

Tablo 5. Faktör korelayon matrisi

	Karlılık	Likitide	Mali Yapı	Faaliyet
Karlılık	1			
Likitide	0,32 (0,06)	1		
MaliYapı	-0,11 (0,07)	-0,55 (0,04)	1	
Faaliyet	-0,21 (0,08)	0,00 (0,07)	0,01 (0,07)	1

Tablo 5 incelendiğinde mali yapı ve likitide faktörleri arasında anlamlı ve ters yönlü bir ilişki olduğu görülmektedir ($p<0.05$).

Modelin uygunluğuyla ilgili çok sayıda endeks değerinin hesaplandığı görülmekte bunlardan birkaçı değerlendirildiğinde GFI (0.91), CFI(0.89) değerleri bire yaklaşık olduğu için, RMSEA değeri de sıfıra yaklaşık bir değer (0.081) olduğu için modelin iyi uygun olduğuna karar verilebilmektedir. Ancak anlamsız oranlar modelden çıkarıldığında daha iyi uygun modelin elde edilebileceğinin de dikkate alınması gerekmektedir.

4. TARTIŞMA VE SONUÇ

Her iki faktör analizi yaklaşımı, çok sayıda değişkenin az sayıda faktörler şeklinde tanımlanmasını sağlamaktadır. Bununla birlikte önemli farklılıkları bulunmaktadır. Bunlar kısaca şu şekilde özetlenebilir.

- AFA'da faktörler analizin sonunda isimlendirilebilmekte, DFA'da teoriye dayanarak analize başlamadan önce tanımlanmaktadır.
- AFA'da gözlenen değişkenlerin her faktörle ilişkisi bulunmakta, DFA'da ise gözlenen değişkenin sadece ilgili olduğu faktörle ilişkisi bulunmaktadır (Şekil.1 ve Şekil.2).
- AFA'da dik döndürme uygulanıyorsa faktörler arasında ilişki bulunmamakta, eğik döndürme yapılıyorsa Şekil 1'de görüldüğü gibi faktörler arasında ilişki bulunmaktadır. DFA'da ise faktörler arasında ilişki hesaplanmakta ancak hangi faktörler arasında ilişki olacağına teoriye dayanılarak karar verilmektedir.
- AFA ve DFA analizinin her ikisinde de hatalar hesaplanabilmekte ancak hatalar arasındaki ilişki sadece DFA'da incelenebilmektedir.
- AFA'da 0.30'dan büyük değerler için faktör yüklerinin anlamlı olduğu kabul edilmekte, DFA'da ise faktör yüklerinin anlamlılığı da test edilebilmektedir.
- DFA'da teorik modelle ilgili hipotezler sınanmakta ve model düzeltilerek modelin uygunluğu arttırılabilmektedir.

Uygulamada finansal oranlar için her iki faktör analizi yaklaşımına ait sonuçlar değerlendirilmiştir. AFA'da faktörler, analiz sonuçlarına göre isimlendirilmiş ve aynı zamanda teoriye de uygun olan model DFA ile yorumlanmıştır. AFA'da 4 faktör elde edilmiş ve faktör yükleri 0.30'dan büyük olduğu için model anlamlı kabul edilmiştir. Model, faktör yüklerinin ve faktörler arasındaki korelasyonların anlamlılığının test edilmesine olanak veren DFA kullanılarak yorumlandığında ALCDH ve AKTDH oranlarının anlamsız, mali yapı ve likitide faktörleri arasındaki ilişkinin ise anlamlı olduğu tespit edilmiştir.

Teorik model biliniyorsa DFA uygulaması daha kapsamlı bilgi vermektedir. Teorik modelle ilgili bilgi olmadığında ise genelde AFA kullanılmaktadır. Yapılan bazı çalışmalarda ise AFA ile faktörler isimlendirilmekte daha sonra DFA ile modelin geçerliliği test edilmektedir. Bu nedenle amaca uygun yaklaşımın kullanılması önerilmektedir.

5. KAYNAKLAR

Anderson T.W., Rubin H., 1956. Statistical inference in factor analysis, Proceedings of the third Berkeley Symposium on Mathematical Statistics and Probability, 5, 111-150, Berkeley: University of California Pres.

Hair, J., Anderson R.,Tatham R., Black W., 1998. Multivariate data analysis, Prentice Hall, New Jersey.

Hoyle, R., 1995. Structural equation modeling: Concepts, issues, and applications, Sage Publications,U.S.A.

Howe, W.G., 1955. Some contributions to factor analysis, Oak Ridge, Tennesee.

Joreskog, K.G., 1967. Some contributions to maximum likelihood factor analysis, Psychometrica, 32, 443-482.

Joreskog, K.G., 1969. A general approach to confirmatory maximum likelihood factor analysis, Psychometrica, 34, 183-202.

Joreskog, K.G., Lawley D.N., 1968. New methods in maximum likelihood factor analysis, British Journal of Mathematical and Statistical Psychology, 21, 85-96.

Kaplan, D., 2000. Structural equation modeling foundations and extensions, Sage Publications, U.S.A , 2000.

Kline R. B., 2005. Principles and practice of structural equation modeling, Second Edition (Methodology In The Social Sciences), The Guilford Press.

Lawley, D.N., 1940. The estimation of factor loadings by the method of maximum likelihood, Proceedings of the Royal Society of Edinburgh, 60, 64-82.

Lawley, D.N., 1958. Estimation in factor analysis under various initial assumptions, British Journal of Statistical Psychology,11,1-12.

Long, J.S., 1983. Confirmatory factor analysis, Sage publication, A.B.D.

Rencher, A., 1995. Methods of multivariate analysis, John Wiley & Sons, Kanada.

Schumacker, R.E., 2004. Beginner's guide to structural equation modeling, Lawrence Erlbaum Associates, A.B.D.

Spearman, C., 1904. General intelligence, objectively determined and measured, American Journal of Psychology,15,201-293.

Spearman, C., 1927. The abilities of man. London, Macmillan.

Stevens, J., 1996. Applied multivariate statistics for the social sciences, 3. baskı, Lawrence Erlbaum, New Jersey.

Tabachnick, B., Fidell L., 1996. Using multivariate statistics, Harper Collins, A.B.D.

Tacq, Jacques, 1999. Multivarate technique in social sciences, Sage Publications, Great Britain.

Thompson, Bruce, 2005. Exploratory and confirmatory factor analysis, American Psychological Association, A.B.D.

Thurstone, L.,1947. Multiple factor Analysis, Chicago, University of Chicago Press.

COMPARISON OF EXPLORATORY AND CONFIRMATORY FACTOR ANALYSIS: AN APPLICATION

ABSTRACT

Factor analysis is one of the multivariate statistical methods that can be used to analyze interrelationships among large number of variables and to explain these variables into smaller set of factors. The method summarizes a special information that belongs to a large number of variables and facilitates the interpretation of the results with data reduction. There are two factor analysis approaches that are widely used. One of them is Explanatory Factor Analysis and the other is Confirmatory Factor Analysis. The aim of this study is to compare two approaches and to give general information about the selection process.

Keywords: Exploratory factor analysis, Confirmatory factor analysis.

Ek.1 Finansal Oranlar

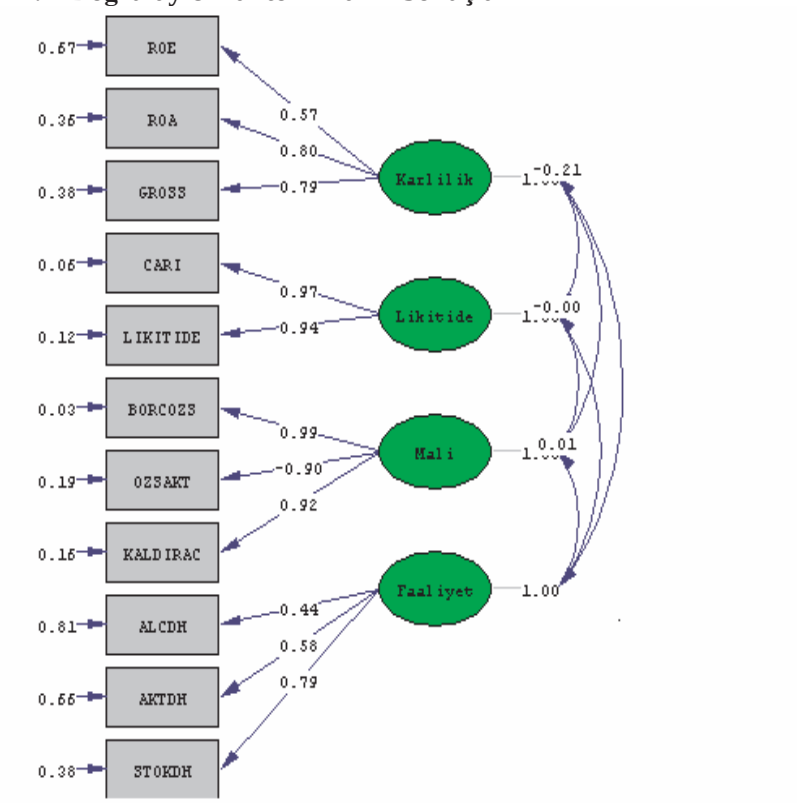
1. Karlılık Oranları
- Faiz ve Vergi Öncesi Kar/Kaynaklar Toplamı (FVAKT)
- Vergi Öncesi Kar/Öz Sermaye (VKOZS)
- Brüt Satış Karı/Net Satışlar (BKSAT)

2. Likitide Oranları
- Dönen Varlık/K.V Borc (CARI)
- (Dönen Varlık–Stoklar)/K.V. Borc (ASIT)

3. Mali Oranlar
- T. Borç/Öz Sermaye (BORCOZS)
- Öz Sermaye/Aktif Toplamı (OZSAKT)
- T. Borç/Aktif Toplamı (KALDIRAC)

4. Faaliyet Oranları
- Net Satışlar/Ortalama Ticari Alacaklar (ALCDH)
- Satılan Malın Maliyeti/Ortalama Stoklar (STOKDH)
- Net Satışlar/Aktif Toplamı (AKTDH)

Ek. 2 Doğrulayıcı Faktör Analizi Sonuçları



Cilt: 06 Sayı: 01 Kategori: 02 Sayfa: 111-122 ISSN No: 1303-6319 Volume: 06 Number: 01 Category: 02 Page: 111-122

A BOOLEAN APPROACH IN FOREST MANAGEMENT

Nurcan TEMİZ* **Vahap TECİM****

ABSTRACT

Multi criteria analyses have been used mostly to deal with spatial decision problems since their emergence. Spatial multi criteria analysis is different from conventional multi criteria decision analysis (MCDA). Because it includes a geographic component. Multi criteria evaluation (MCE) and multi criteria decision making (MCDM) are very important concepts in Geographical Information Systems (GIS). Many spatial decison problems entail GIS and MCDA integration. GIS-MCDM integration can be thought of a process that uses value judgements and then represents results of these judgements spatially on a digital map. Forestry decision problems involve many alternatives and evaluation criteria. Most of the forest management problems are spatial in their nature and usually involve multi criteria. Fire management is an important component of forest management. In this study Boolean approach is used for the fire management. The areas that can cope with fire effectively are examined according to their distances from water resources, streams and settlement areas criteria for İzmir Forest Administration Chief Office by using Boolean approach. IDRISI Software Package is used for all analyses.

Keywords: Boolean analysis, Forest fire management, GIS, Spatial multi criteria decision making.

1. INTRODUCTION

Many problems in life can be thought of as multi criteria decision making problems. As stated by Vassilev et al. (2005), multi criteria decision making problems can be divided into two distinct classes. In the first class, a finite number of alternatives are explicitly given in a tabular form. These problems are called discrete multi criteria decision making problem or multi criteria analysis problems. In the second class, a finite number of explicitly set of constraints in the form of functions define an infinite number of feasible alternatives. These problems are called continuous multi criteria decision making problem or multi criteria optimization problems. The methods used in the different approaches of decision analysis are called Multi Criteria Decision Methods (MCDM).

Multi criteria analyses have been used largely to deal with spatial decision problems since their emergence. The preliminary works including integration of Geographical Information Systems (GIS) and multi criteria analysis were in the late 1980s and the early 1990s (Chakhar and Martel, 2003).

Banai (1993) used Analytic Hierarchy Process (AHP), a multi criteria decision making technique, and GIS in order to find optimally suitable sites for landfill. In another study,

* Research Assistant Dr., Dokuz Eylul University, Faculty of Arts and Sciences, Department of Statistics, e-mail: nurcan.temiz@deu.edu.tr

** Professor Dr., Dokuz Eylul University, Faculty of Economics and Administrative Sciences, Department of Econometrics, e-mail: vahap.tecim@deu.edu.tr

GIS-MCDM was used to improve quality of landscape ecological forest planning (Kangas et al., 2000). Ananda and Herath (2003), examined the use of AHP in regional forest planning. Jumppanen, et al. (2003) applied GIS-MCDM in spatial harvest scheduling approach for areas involving multiple ownership. Evans et al. (2004) used Boolean (suitable/unsuitable) and weighted map overlays in the site search problem for waste management. Mau-Cummins et al. (2005) used AHP in the selection of forest wilderness sites. Our study used Boolean approach in determining the most appropriate areas that can cope with the forest fires subject to a defined set of criteria.

Spatial multi criteria analysis requires information on criterion values and the geographical locations of alternatives, and the results of analysis are represented visually on a digital map (Jankowski, 1995; Malczewski and Ogryczak, 1996). As stated by Carver (1991) and Jankowski (1995), two important components of spatial multi criteria decision analysis are GIS component and multi criteria decision making component.

In this study integration of GIS and MCDM is applied for İzmir Forest Administration Chief Office. The main objective is to do spatial MCDA, which is different from conventional MCDA, for our study area. For this reason the most appropriate areas that can cope with forest fires are represented according to Boolean approach. IDRISI software package, which is a GIS package named after the famous geographer Abu Adb Allah Muhammed al-Idrisi (1100-1166 A.D.) and is dedicated to him, is used for all analyses.

1.1 An Overview of GIS

Environmental management has been a major motivator of developments in GIS. Some authors suggest that the roots of current GIS is Canada Geographic Information System (CGIS), which emerged in the 1960s. It was designed to produce the map of land capability for forestry. Its initial task was to classify and map the land resources of Canada. The second objective of the system was to provide data to the Government of Canada on land resources and their management.

Created maps were classified according to various themes. Some of these themes were soil capability for agriculture, forestry capability, and present land use (DeMers, 1997; Heywood et al., 2002; Goodchild, 2003). When these systems were first developed in the early 1960s, they were no more than a set of innovative computer-based applications for data processing on maps. Today GIS is one of the fastest growing sectors in computer industry and an important component of the information technology (Franklin, 2001; Lo and Yeung, 2002). GIS technology offers combined power of both geography and the information systems and provides ideal solutions for effective natural resource management (Shamsi, 2005). This technology integrates common database operations such as query and statistical analysis with visualization and geographic analysis offered by maps. These abilities distinguish GIS from other information systems and make it valuable for several applications (Lang, 2001).

1.2 Integration of GIS and MCDM

Spatial multi criteria decision problems typically involve a set of geographically defined alternatives from which a choice of one or more alternatives is made with respect to a given set of evaluation criteria.

Multi Criteria Evaluaiton (MCE) and MCDM are very important concepts in GIS. Many spatial decision problems lead to GIS and MCDA integration. These two disciplines can benefit from each other. On the one hand, GIS techniques have an important role in analyzing decision problems and it is a decision support system that integrates spatially referenced data into a problem solving environment. On the other hand, MCDA provides many techniques and procedures for structuring decision problems, and evaluating and prioritizing alternative decisions. GIS-multi criteria decision making integration can be thought of as a process that transforms and combines geographical data and value judgements of the decision maker to obtain information for decision making (Malczewski, 2006).

In the context of GIS, two procedures are common for MCE. The first includes Boolean overlay, the second is known as Weighted Linear Combination (WLC). In Boolean approach, all criteria are assessed by thresholds of suitability to produce Boolean maps, which are then combined by logical operators such as intersection (AND) and union (OR). With WLC, continuous criteria (factors) are standardized to a common numeric range, and then combined by weighted averaging. The result is a continuous mapping of suitability (Jiang and Eastman, 2000).

Boolean analysis is used only when two states are possible (criterion satisfied and not satisfied). This analysis was developed by George Boole, who devised rules and methodologies for combining two-state variables. In Boolean search it is generally concerned with the AND operator. The logical AND operator produces a true result from the phrase “A AND B” only if both A and B are “true”. In GIS, this methodology is used in a multiplication overlay between layers containing only “zeroes” (representing areas where conditions are “false” or “criterion is not satisfied”) and “ones” (representing areas where conditions are “true” or “criterion is satisfied”) (Eastman, 2003).

1.3 Forest Management

Forestry involves the management of a wide range of natural resources. In addition to timber, forests provide various resources such as land for livestock to graze, recreation areas and water supply resources. In this context, forest management includes management of harvesting and recreational areas, protection of endangered species and archaeological sites. Management of forest resources is a complex task due to multi-functional nature of these resources. Therefore, the problems of forest management and planning usually involve decisions which have to take into account multiple objectives (Aronoff, 1995; Kazana et al., 2003; Mohren, 2003).

The amount of data and information involved in the forest management is often overwhelming. Integrated decision support systems help forest managers to make consistently good decisions about forest ecosystem management (Potter et al., 2000). Compared to previous forest management approaches, new forest management strategies require integration of spatial information technologies, such as GIS, remote sensing, and decision support systems (Franklin, 2001).

The designing of a forest database is crucial in a comprehensive forest management plan. Data should be accurate, properly organized, detailed and obtained easily and economically. The gathering of spatial and nonspatial data and their analysis determine the quality of forest management plans. Forest management constists of several subsystems one of which is the fire management system. It is very important to minimize damage caused by a forest fire. This can be achieved by developing an efficient fire management system. Fire fighting planning is an important component of fire management system. Martell (1982) reviewed Operations Research approaches in forest fire management comprehensively.

2. MATERIALS AND METHODS

In this study, integration of GIS and MCDM is applied for İzmir Forest Administration Chief Office. This Institution is subordinate to İzmir Directorate of Forest Administration, which has eleven chief offices. It is aimed to represent the most appropriate areas that can cope with forest fires effectively in the boundary of our study area according to Boolean approach.

Figure 1 shows the forest boundary map of İzmir Forest Administration Chief Office. General area is 39270 ha and 50.88 % of this area is forested land. Total forest area is 19983.5 ha of which 11494.5 ha (57.52 %) is productive forest and 8489 ha is unproductive forest.

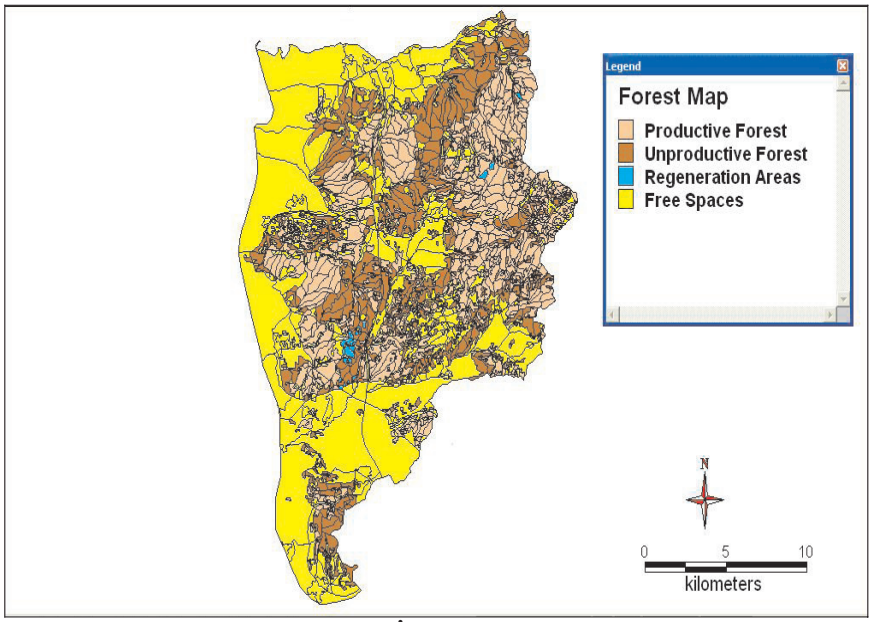


Figure 1. Forest boundary map of İzmir Forest Administration Chief Office

There are several criteria that must be considered in forest fire fighting planning process, such as fuel/vegetation type, soil properties, topographical information, slope, aspect and altitude information, distance from roads, distance from water resources, distance from settlement areas, and distance from streams. However, in this study only the last three criteria were used. This was due to the fact that maps of the other criteria were unavailable to authors, whereas maps for the three criteria above could be constituted by the data obtained from the study area. The most important point that must be taken into in spatial multi criteria decision making is the availability of maps of all criteria. Water resources map, settlement areas map and stream map are used for the analyses. First phase of the application is the conversion of all vector-based maps to the raster-based maps. Then Boolean analysis is done and results are visualized by the maps.

3. RESULTS

3.1 The Boolean Approach

In our study Boolean approach is used to determine the most appropriate areas that can cope with forest fires effectively. Firstly all criteria are standardized to Boolean values (0 and 1). Factors (criteria) of our study are distances from water resources, streams and settlement areas.

3.1.1 Distances from Water Resources, Streams and Settlement Areas

Water resources and streams are very important in fire management. The areas closer to the water resources and streams are considered more appropriate to cope with fire than the areas that are distant from water resources. Settlement areas can be considered as an important factor to intervene and control fire. However, according to different points of

view settlement areas can also be considered as a risky factor. In some cases, the areas closer to the settlement areas are more fire prone because of the human factor.

As interviewed with the directorates of fire combatting department of İzmir Forest Administration Chief Office the areas closer to the water resources, streams and settlement areas were considered as appropriate (1) and the others were considered as not appropriate (0) in this study.

There are four water resources in our study area namely Buca Gölet, Kaynaklar Göleti, Sarnıç Göleti and BP Olduruk. Water resources map was derived by rasterizing and using the module DISTANCE in IDRISI software package. Then the distance image, which shows a simple linear distance from all water resources in our study area, was obtained as shown in Figure 2.

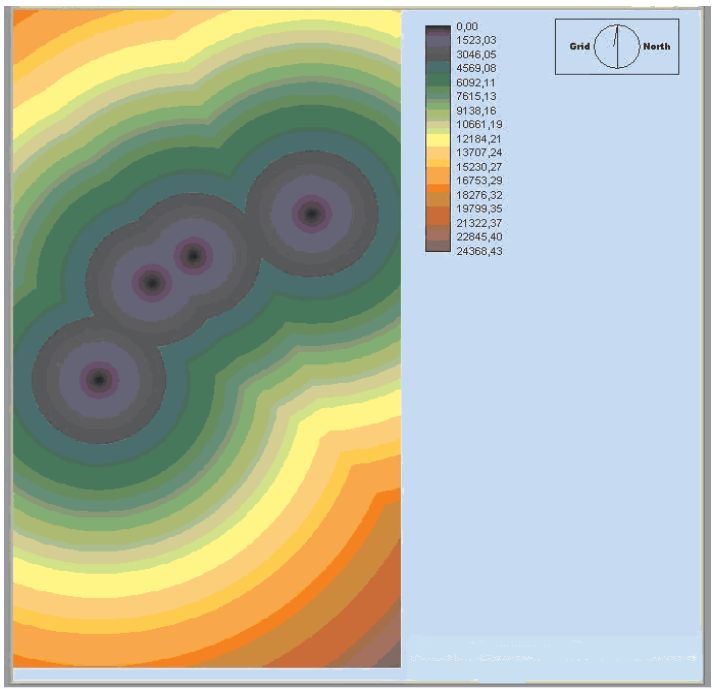


Figure 2. Distance map of the water resources

In this stage it was needed to RECLASSIFY continuos image of distance from water resources to determine the distances that are appropriate and the distances that are not appropriate. As interviewed with the directorates of fire combatting department of İzmir Forest Administration Chief Office, the areas that have a distance less than 5000 meters to the water resources were considered as appropriate (1) and those equal to or larger than 5000 meters were considered as not appropriate (0). Reclassified distance map of the water resources is shown in Figure 3.

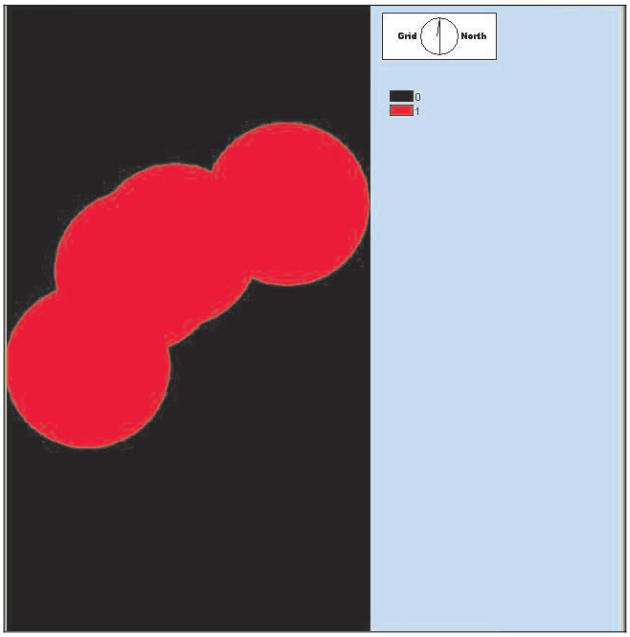


Figure 3. Reclassified distance map of the water resources

The same procedures were followed for the distance from streams and the distance from settlement areas. For reclassification of distance from streams factor, areas that have a distance less than 5000 meters to the streams were considered as appropriate (1) and those equal to or larger than 5000 meters were considered as not appropriate (0). For reclassification of the distance from settlement areas factor, areas that have a distance less than 2000 meters to the settlement areas were considered as appropriate (1) for effectively struggling with the fire and those equal to or larger than 2000 meters were considered as not appropriate (0). Figure 4 and Figure 5 show reclassified distance maps of the streams and the settlement areas, respectively.

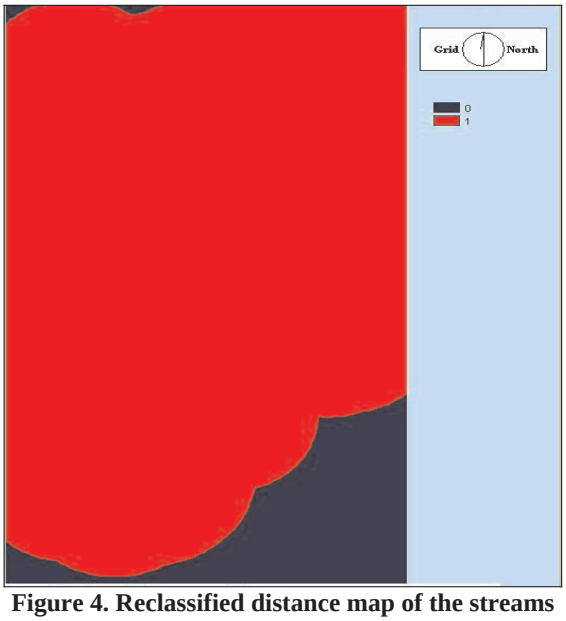


Figure 4. Reclassified distance map of the streams

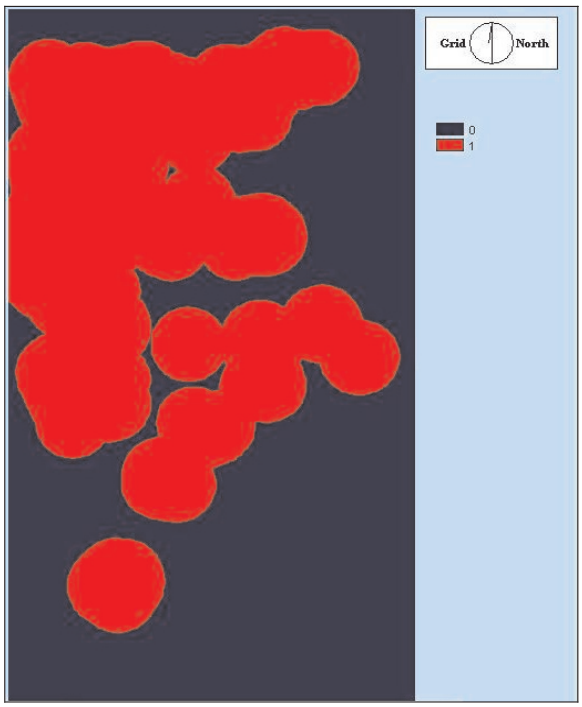


Figure 5. Reclassified distance map of the settlement areas

3.1.2 Boolean Aggregation of Factors

All factors have been transformed into Boolean images and they were ready to be aggregated. All of these three factors were multiplied together to produce a single image of appropriate areas that can effectively cope with the forest fire. This aggregation process was done by using image calculator with the AND operation in IDRISI software package. By using the AND operation it is aimed to represent the intersection of the areas according to the distance from water resources, streams and settlement areas criteria and to visualize the 'most appropriate' areas in terms of meeting all of these criteria simultaneously.

At the end of Boolean approach process, the most appropriate areas that can cope with forest fire were determined as shown in Figure 6.

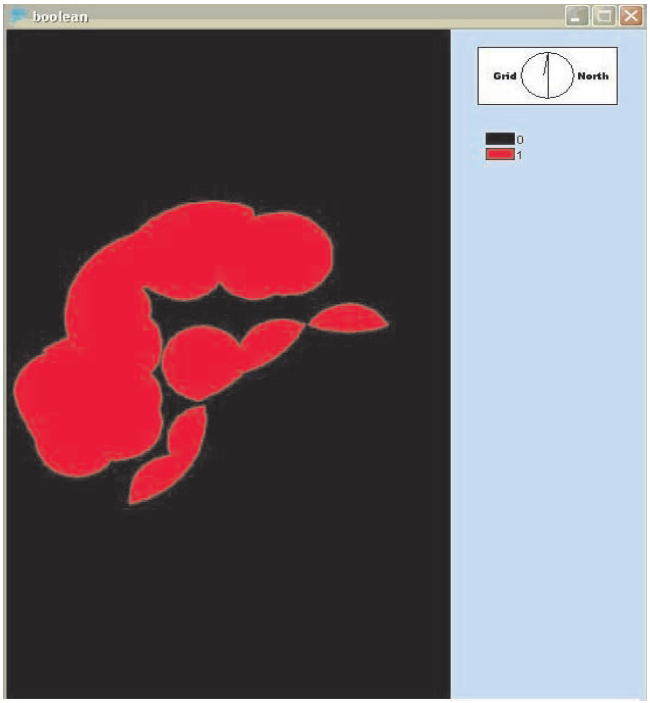


Figure 6. The most appropriate areas that can cope with forest fires effectively according to Boolean approach

5. CONCLUSIONS

This study represents the use of Boolean analysis to determine areas which are the most appropriate in fire fighting according to distances from water resources, streams and settlement areas. Several criteria can be added to this analysis. Hovewer, the most important point that must be taken into account is the availability of maps of these criteria. By looking at the results of this study subject to three criteria, it is proposed that İzmir Forest Administration Chief Office must take proactive measures and pay more attention to the areas shown as (0) in Figure 6. The results of this study can change in the case of adding different and more criteria to the analysis.

This study can be further extended by increasing the numbers of criteria. The next step of this study is to use AHP for determining the most appropriate areas that can cope with forest fires. Then the results of Boolean analysis and AHP can be compared as to the details of information they give.

ACKNOWLEDGEMENTS

The authors would like to thank staff of İzmir Regional Directorate of Forestry for their kindness, help, and the time they spent to give information and for providing the data needed for this study. The authors want to thank Rectorship of Dokuz Eylul University for funding support of this study as a Scientific Research Project with the 2006 KB FEN 4 project number.

6. REFERENCES

Ananda, J., Herath, G., 2003. The use of analytic hierarchy process to incorporate stakeholder preferences into regional forest planning. Forest Policy and Economics, 5, 13–26.

Aronoff, S., 1995. Geographic information systems: A management perspective. WDL Publications, Ottawa.

Banai, R., 1993. Fuzziness in geographical information systems: Contributions from an analytic hierarchy process. International Journal of Geographical Information Systems, 7, 315–329.

Carver, S.J., 1991. Integrating multi criteria evaluation with geographical information systems. International Journal of Geographical Information Systems, 5, 321-339.

Chakhar, S., Martel, J-M., 2003. Enhancing geographical information systems capabilities with multi criteria evaluation functions. Journal of Geographic Information and Decision Analysis, 7 (2), 47-71.

DeMers, M., 1997. Fundamentals of geographic information systems. John Wiley&Sons Inc., New York.

Eastman, J. Ronald., 2003. Idrisi Klimanjaro tutorial. Clark Labs Clark University. Worchester, USA.

Evans, A.J., Kingston, R., Carver, S., 2004. Democratic input into the nuclear waste disposal problem: The influence of geographical data on decision making examined through a Web-based GIS. Journal of Geographical Systems, 6, 117–132.

Franklin, S.E., 2001. Remote sensing for sustainable forest management. CRC Press LLC., USA.

Goodchild, M.F., 2003. Geographic information science and systems for environmental management. Annual Review of Environment and Resources, 28, 493-519.

Heywood, I., Cornelius, S., Carver, S., 2002. An introduction to geographical information systems (2nd Edition). Prentice Hall, Inc., United Kingdom.

Jankowski, P., 1995. Integrating geographical information systems and multiple criteria decision-making methods. International Journal of Geographical Information Science, 9, 251-273.

Jiang, H., Eastman, J.R., 2000. Application of fuzzy measures in multi criteria evaluation in GIS. International Journal of Geographical Information Systems. 14, 173–184.

Jumppanen, J., Kurttila, M., Pukkala, T., Uuttera, J., 2003. Spatial harvest scheduling approach for areas involving multiple ownership. Forest Policy and Economics, 5, 27-38.

Kangas, J., Store, R., Leskinen, P., Mehtätalo, L., 2000. Improving the quality of landscape ecological forest planning by utilising advanced decision support tools. Forest Ecology and Management, 132, 157-171.

Kazana, V., Fawcett, R.H., Mutch, W.E.S., 2003. A decision support modeling framework for multiple use forest management: The Queen Elizabeth forest case study in Scotland. European Journal of Operational Research, 148 (1), 102-115.

Lang, L., 2001. Managing natural resources with GIS. ESRI Press, USA.

Lo, C.P., Yeung, A.K.W., 2002. Concepts and techniques of geographic information systems. Prentice Hall, Inc., New Jersey.

Malczewski, J., Ogryczak, W., 1996. The multiple criteria location problem: 2nd preference-based techniques and interactive decision support. Environment and Planning A, 28, 69-98.

Malczewski, J., 2006. GIS-based multi criteria decision analysis: A survey of literature. International Journal of Geographical Information Science, 20 (7), 703-726.

Martell, D.L., 1982. A review of operational research studies in fores fire management. Canadian Journal of Forest Research, 12, 119-140.

Mau-Cummins, T., de Steiguer, J.E., Dennis, D., 2005. AHP as a means for improving public participation: a pre-post experiment with university students. Forest Policy and Economics, 7, 501–514.

Mohren, G.M.J., 2003. Large-scale scenario analysis in forest ecology and forest management. Forest Policy and Economics, 5 (2), 101-206.

Potter, W.D., Liu, S., Deng, X., Rauscher, H.M., 2000. Using DCOM to support interoperability in forest ecosystem management decision support systems. Computers and Electronics in Agriculture, 27, 335-354.

Shamsi, U.M., 2005. GIS applications for water, wastewater, and stormwater systems. CRC Press, Boca Raton.

Vassilev, V., Genova, K., Vassileva, M., 2005. A brief survey of multi criteria decision making methods and software systems. Cybernetics and Information Technologies, 5 (1), 3-13.

ORMAN YÖNETİMİNDE BOOLEAN YAKLAŞIMI

ÖZET

Çok kriterli karar problemleri ortaya çıkışlarından beri büyük ölçüde konumsal (mekansal) karar problemlerini çözmek için kullanılmıştır. Konumsal çok kriterli karar analizi, klasik Çok Kriterli Karar Analizi (ÇKKA)'dan farklıdır. Çünkü coğrafi bileşen içermektedir. Çok Kriterli Değerlendirme (ÇKD) ve Çok Kriterli Karar Verme (ÇKKV, Coğrafi Bilgi Sistemleri (CBS) oldukça önemli kavramlardır. Birçok konumsal karar problemi CBS ve ÇKKA'nın entegrasyonunu gerektirmektedir. CBS-ÇKKV entegrasyonu, çıkarımları kullanan ve daha sonra bu çıkarımların sonuçlarını konumsal olarak sayısal harita üzerinde gösteren bir süreç olarak düşünülebilir. Ormancılıkla ilgili karar problemleri birçok alternatifi ve değerlendirme kriterini içermektedir. Çoğu orman yönetimi problemi yapısal olarak konumsaldır ve genellikle çoklu kriter içermektedir. Yangın yönetimi, orman yönetiminin önemli bir bileşenidir. Bu çalışmada yangın yönetiminde Boolean yaklaşımı kullanılmıştır. İzmir Orman İşletme Şefliği için yangınla etkin olarak mücadele edebilen alanlar; su kaynaklarından uzaklık, akarsulardan uzaklık, yerleşim birimlerinden uzaklık kriterlerine göre incelenmiştir. Tüm analizler için IDRISI paket programı kullanılmıştır.

Anahtar Kelimeler: Boolean analizi, Orman yangını yönetimi, Coğrafi bilgi sistemleri, Konumsal çok kriterli karar verme.

Cilt: 06 Sayı: 01 Kategori: 02 Sayfa: 123-143 ISSN No: 1303-6319 Volume: 06 Number: 01 Category: 02 Page: 123-143

KARDEŞ CİNSİYET BİLEŞİMİNİN EĞİTİMSEL ERİŞİMLERE ETKİSİ[†]

Ali BERKER[*]

ÖZET

Aile ekonomisi teorileri, çocuklar için yapılan insan sermayesi yatırımlarının belirlenmesinde kardeş cinsiyet bileşiminin önemli bir etken olduğunu belirtmektedir. Bu çalışmada, rassal etkileri tahmin etme yöntemi çerçevesinde aileler arasındaki kız kardeş sayısındaki değişkenlik kullanılarak, kardeş cinsiyet bileşiminin çocukların ilköğretim ve lise mezunu olma olasılıklarına olan etkileri incelenmiştir. Türkiye İstatistik Kurumu (TÜİK) tarafından gerçekleştirilen 2000 Genel Nüfus Sayım sonuçlarından elde edilen bulgular, kız kardeş sayısındaki artışın çocukların eğitimsel erişimlerini arttırdığını göstermektedir. Bu olumlu etkinin, ailesi düşük ve orta sosyo-ekonomik konumda olan çocuklar için daha büyük olduğu bulunmuştur. Ayrıca, kız kardeş sayısının eğitimsel erişimlerdeki erkek-kız farkını etkilemediği gözlenmiştir.

Anahtar Kelimeler: Aile yapısı, Eğitimsel başarılar.

1. GİRİŞ

Bu çalışma Türkiye'deki ailelerde kardeş cinsiyet bileşiminin, aile içinde yapılan insan sermayesi yatırımlarının kız ve erkek çocukları arasındaki bölüşümüne olan nedensel etkilerini analiz etmektedir.

Aile ekonomisi teorileri; ekonomik karar alma ve uygulama birimi olan ailede çocuklara yapılan insan sermayesi yatırımlarının çocukların cinsiyetine göre farklılıklar göstermesinin etkinlik-eşitlik ikilemi ekseninde nasıl gerçekleştiğini, bunun kuşaklar arası insan sermayesi, gelir ve servet aktarımını nasıl belirlediğini ayrıntılarıyla incelemiştir (Becker, 1991; Becker, 1993; Behrman vd., 1982). Çocuğun cinsiyetine göre insan sermayesi yatırımlarının farklılaşmasının nedenlerinden ilki, çocuklara yapılacak yatırımların hem işgücü piyasasında, hem de evlilik piyasasında beklenen getirisinin çocuğun cinsiyetine göre farklılaşmasıdır. İkincisi, insan sermayesi üretim fonksiyonlarında kız ve erkek çocuklarının aynı girdiden farklı miktarda ve/veya farklı girdileri kullanmaları nedeniyle yapılacak yatırımların aile bütçesine olan yükünün cinsiyete göre farklılaşmasıdır. Üçüncüsü, ailelerin kız ve erkek çocukları için beklenen gelirlerin eşitsizliği yönündeki tercihlerinin ve her bir çocuğa verilen göreli önemin cinsiyete göre farklılaşmasıdır. Son olarak, aile içinde çocuklar arasında ve çocuklar ile anne-baba arasındaki etkileşimlerin çocukların cinsiyetinden etkilendiği ölçüde, çocukların gelişiminin, amaçlarının ve ailenin çocukları için amaçladığı insan sermayesi yatırımlarının cinsiyete göre farklılıklar göstermesidir.

[†] Bu çalışma, TÜBİTAK'ın Hızlı Destek Programı (Proje Kodu: 105K-130) tarafından desteklenmiştir. Derginin hakemlerine ve Editör Yardımcısı Sevil UYGUR'a önerilerinden, düzeltmelerinden ve katkılarından dolayı teşekkür ederim. Ayrıca, makalenin hazırlanmasında katkıda bulunan İnsan Tunalı, Derya Erel, İsmail Erol, Nebile Korucu'ya, TÜBİTAK'ın ve Türkiye Ekonomi Kurumu'nun hakemlerine teşekkür ederim. Makaledeki hataların ve noksanlıkların sorumluluğu sadece bana aittir.

[*] Yrd. Doç. Dr., Abant İzzet Baysal Üniversitesi İktisat Bölümü, Bolu. e-posta: berkera@gmail.com

Bu teorik nedenlerden dolayı, literatürde kardeş cinsiyet bileşimleri ile çocukların insan sermayesi çıktıları arasındaki ilişki çok yönlü bir şekilde incelenmiştir. Özellikle, kardeş cinsiyet bileşiminin çocukların hem gençlik hem de yetişkinlik dönemindeki eğitimsel çıktılarına ve başarılarına etkileri (Butcher ve Case, 1994; Kaestner, 1997, Garg ve Morduch, 1998a), çocuk ölümleri, bodur olma, zayıf olma, düşük-kilolu olma gibi çocukların sağlık çıktılarına etkileri (Das Gupta, 1987; Garg ve Morduch, 1998b), çocuk emeğine etkileri (Edmonds, 2006) inceleme konusu olmuştur.

Ayrıca, 2000 TÜİK Genel Nüfus Sayım sonuçları incelendiğinde, kentte yaşayan çocukların eğitimsel erişimlerindeki başarısızlıkları önemli boyutlardadır. 16-18 yaş grubundaki çocukların sadece yarısından biraz fazlası (% 52) ilköğretimi tamamlayabilmiştir. Benzer şekilde, 18-20 yaş grubundaki çocukların sadece % 42'si lise mezunudur[17]. Aynı yaş grupları için bu eğitimsel erişimlerdeki çocukların cinsiyete göre farklılıkları incelendiği zaman, erkek çocukların kız çocuklarına göre daha büyük olasılıkla ilköğretimi (erkek: % 58, kız: % 46) ve liseyi (erkek: % 44, kız: % 40) tamamladıkları gözlenmiştir. Bu eğitimsel erişimlerdeki başarının düşük olması ve cinsiyete özgü farklılıklar göstermesinden dolayı, aile içindeki kardeş cinsiyet bileşiminin erkek ve kız çocuklarına yapılan insan sermayesi yatırımlarını nasıl ve hangi boyutlarda etkilediğini incelemek önem kazanmaktadır.

Çalışmada, aile içindeki çocukların kardeş cinsiyet bileşimleri kız kardeş sayısıyla, çocuklara yapılan insan sermayesi yatırımları ise çocukların ilköğretimi ve liseyi bitirme olasılıkları ile ölçülmüştür. Ekonometrik analizin sonuçları, kız kardeş sayısı ile ilköğretim ve lise mezunu olmaları arasında pozitif bir korelasyon olduğuna işaret etmektedir. Bütün kardeşleri erkek olan bir erkek çocuğuyla karşılaştırıldığında, bütün kardeşleri kız olan bir erkek çocuğun ilköğretimi ve liseyi tamamlama olasılıkları sırası ile % 11.29 ve % 15.69 oranında daha büyüktür. Aynı durum tahminleri kız çocukları için de sırasıyla % 10.17 ve % 12.24'tür. Ayrıca, kardeş cinsiyet bileşiminin etkilerinin ailenin maddi olanaklarına göre farklılaştığı da belirlenmiştir. Kız kardeş sayısının eğitimsel erişimlere olan olumlu etkileri düşük ve orta sosyo-ekonomik statülü ailelerin çocukları için gözlemlenirken, yüksek sosyo-ekonomik statülü aileler için gözlenmemiştir.

Bu sonuçlar gelişmekte olan ülkeler için yapılan çalışmaların sonuçlarıyla tutarlıyken, gelişmiş ülkeler için yapılan çalışmaların sonuçlarından bazı farklılıklar göstermektedir. Örneğin, Butcher ve Case (1994) Amerika Birleşik Devletleri (ABD) için yetişkinlerin oluşturduğu bir örneklem kullanarak yaptığı çalışmada, sadece erkek kardeşleri olan kızların kız kardeşine sahip olan kızlardan daha yüksek eğitim seviyesine sahip olduklarını göstermiştir. Öte yandan, Kaestner (1997) ABD için daha yeni doğum kohortları (yaş grupları) için kardeş cinsiyet bileşimi ile eğitimsel erişimler arasında istatistiksel olarak anlamlı bir ilişki bulamamıştır. Ayrıca, Kaestner (1997) 12-18 yaş grubundaki çocukların akademik başarılarını incelediğinde, kız kardeş sayısının çocukların akademik başarılarını etkilemediğini ve bu sonucun ailenin maddi olanaklarına göre de farklılaşmadığına dair bulgular sunmuştur. Benzer bir şekilde, Bauer ve Gang (2001) Almanya için kardeş cinsiyet bileşimi ile eğitimsel erişimler arasındaki ilişkiyi destekler nitelikte güçlü bulgular elde edememiştir.

[17] Bu belirtilen sonuçlar yazar tarafından hesaplanmıştır.

Gelişmekte olan ülkeler incelendiğinde, kardeş cinsiyet bileşiminin eğitimsel erişimlere olan etkilerinin göreli olarak daha güçlü olduğu gözlenmektedir. Garg ve Morduch (1998b) Gana için yaptıkları çalışmada, kız kardeş sayısının orta öğretime kayıtlı olma olasılığını olumlu bir şekilde etkilediğini göstermişlerdir. Bu olumlu etki Tanzanya'da 13-16 yaş grubundaki çocukların eğitim süreleri için de bulunmuştur, fakat Güney Afrika'da aynı yaş gurubundaki çocuklar için benzer etki bulunamamıştır (Morduch, 2000).

Bu çalışmaların en önemli ortak özelliği, ailedeki kardeş sayısı, doğum sırasına koşullu olan indirgenmiş-modelleri kullanarak kardeş cinsiyet bileşiminin etkilerini incelemeleridir. Bu çalışmalardan farklı olarak, Kırdar vd. (2007) Türkiye için yaptıkları çalışmada ikiz kardeş doğumlarını araçsal değişken kullanarak hem kardeş sayısının, hem doğum sırasının, hem de kardeş cinsiyet bileşiminin okula kayıtlı olma olasılığına olan nedensel etkilerini tahmin etmişlerdir. Elde ettikleri sonuçlar, genel olarak erkek kardeş sayısının kız çocuklarının okula gitme davranışlarını olumsuz bir şekilde etkilediğini göstermektedir. Sadece kız çocukları için gözlemlenen bu etkiler, ailelerin maddi olanakları ve kardeşlerin doğum sırasına göre önemli farklılıklar göstermektedir.

Çalışmanın ikinci bölümünde, kardeş cinsiyet bileşiminin çocukların eğitimsel çıktılarına olası etkilerini bir çerçeveye oturtabilmek için iki önemli teorik modelin – yatırım ve aile tercihi- kısa bir özeti sunulmuştur. Daha sonraki bölümde kardeş cinsiyet bileşiminin etkilerini ölçmek için kullanılan veri ve ekonometrik yöntemler açıklanmaktadır. Dördüncü bölümde sonuçlar sunulmuştur. Sonuç bölümü de çalışmanın son bölümünü oluşturmaktadır.

2. YÖNTEM

2.1 Kuramsal Arka Plan[18]

2.1.1 Yatırım Modeli

Becker (1993) aileyi fayda fonksiyonunu parasal ve zaman bütçe kısıtlarına bağlı olarak azamileştirmeye çalışan karar alıcı ve uygulayıcı bir birim olarak tanımlamaktadır[19]. Bu azamileştirme probleminin çözülmesi sonucunda, ailedeki her bir çocuk için indirgenmiş insan sermayesi talep fonksiyonu elde edilir. Bu yatırım modeli çerçevesinde, Becker (1993) ailenin yapılacak insan sermayesi yatırımlarının marjinal getirisiyle marjinal maliyetlerini -piyasa faiz oranını- karşılaştırarak yatırımların çocuklar arasında etkin bir şekilde bölüştürüleceğini belirtmektedir.

Yatırım modelinde, kardeş cinsiyet bileşiminin insan sermayesi yatırımlarının çocuklar arasındaki bölüşümüne olan etkileri sermaye piyasasının tam, mükemmel olmamasına

[18] Kardeş cinsiyet bileşimlerinin çocuklara yapılan insan sermayesi yatırımlarına olan etkileri antropoloji, sosyal psikoloji, sosyoloji ve iktisat literatüründe de ayrıntılı bir şekilde incelenmiştir. Bu çalışmada, bu konuyu sadece ailenin iktisadi davranışları açısından ele alan yatırım ve aile tercihleri modelleri incelenmiştir. Bu konunun sosyoloji literatüründe nasıl incelendiğini görmek için Dalton (2000)'a bakılabilir. Ayrıca, sosyal psikoloki literatürü için bu konuyu Türkiye bağlamında inceleyen Kağıtçıbaşı (1981)'nın çalışmasına bakılabilir.

[19] Becker (1993)'e göre ailenin fayda fonksiyonu ailenin tüketim harcamaları, çocukların yetişkinlik dönemleri için beklenen gelirleri ve aile içindeki bireylere göre farklılaşmayan tercihlerin toplamından oluşmaktadır.

ve/veya ailenin maddi olanaklarına bağlıdır (Behrman vd., 1986; Butcher ve Case, 1994; Garg ve Morduch, 1998a). Sermaye piyasasının mükemmel olması veya ailenin maddi olanaklarının yüksek olması durumunda, ailenin borçlanması için bir kısıt oluşmamakta ve aile her bir çocuk için insan sermaye yatırımının marjinal getirisi piyasa faiz oranına eşit oluncaya kadar yatırım yapabilmektedir. Eğer işgücü piyasasında kızlar erkeklerden daha az ücret alıyorsa, yapılacak yatırımın marjinal getirisi kızlar için daha düşük olacak ve sonuç olarak aile içinde, erkek çocuklarla karşılaştırıldığı zaman, kız çocuklarına yapılacak yatırımlar daha düşük seviyelerde gerçekleşecektir. Fakat, kardeş cinsiyet bileşimi çocuğa yapılacak yatırımın marjinal getirisini belirlemediği için, cinsiyet bileşiminin insan sermaye yatırımlarının bölüşümünde hiçbir etkisi olmayacaktır (Kaestner, 1997).

Yatırım modelinde, ailedeki çocuklar için ayrılan kaynaklar sınırlı ve ailenin borçlanma olanaklarının kısıtlı olduğu durumlarda, ailenin insan sermayesi yatırımının marjinal getirisi yüksek olan çocuğa daha fazla yatırım yapacağı belirtilmektedir. Bu durumda, insan sermayesi getirisinin erkek çocukları için daha yüksek, kız çocukları için daha düşük olduğu toplumlarda, aile içinde bulunan kız çocukları erkek çocuklarına yapılan insan sermayesi yatırımlarını olumlu bir şekilde etkileyecektir. Benzer şekilde, sadece erkek kardeşlerine sahip bir kız çocuğu ile karşılaştırıldığında, sadece kız kardeşlerinin bulunduğu bir ailedeki kız çocuğuna daha fazla insan sermayesi yatırımı yapılacaktır. Maddi olanakları kısıtlı ailelerde kız kardeşe sahip olmanın olumlu etkisi -erkek kardeşe sahip olmanın olumsuz etkisi- hem kız, hem de erkek çocukları benzer bir şekilde gerçekleşecektir (Kaestner, 1997; Garg ve Morduch, 1998a). Sonuç olarak, yatırım modelinde kardeş cinsiyet bileşiminin etkisi ailenin maddi olanaklarına bağlı olarak değişmektedir. Maddi olanakları yüksek olan ailelerin çocukları ile karşılaştırıldığında, kız kardeş sayısının olumlu etkisi, maddi olanakları kısıtlı olan ailelerin çocukları için daha önemli ve büyük olabilir.

2.2 Aile Tercihi Modelleri

Yatırım modelinde olduğu gibi, aile tercihi modellerinde çocuklar arasında insan sermayesi yatırımının dağılımını belirleyen en önemli etmen çocukların beklenen servetleridir. Fakat bu iki model çocukların beklenen gelecek ekonomik servetlerinin bileşimi konusunda farklılaşmaktadır. Yatırım modelinde, beklenen kazanç çocuğun yetişkinlik dönemindeki ekonomik serveti olarak tanımlanmıştır, aile tercihi modelinde ise beklenen kazanç ile çocuğa bırakılan miras payının toplamı çocuğun yetişkinlik dönemindeki ekonomik serveti olarak tanımlanmıştır.

Ayrıca, yatırım ve aile tercihi modelleri aile tercihlerinin insan sermayesi yatırımlarının çocuklar arasında dağılımının belirlenmesindeki önemi konusunda farklılaşır. Yatırım modelinde, aile tercihlerinin çocuklara yapılan yatırımların bölüşümünde hiçbir etkisi yoktur. Bu anlamda, yatırım modeli ailenin sadece yatırımlarının etkinliğine odaklanarak çocuklar arasında yatırımlarının paylaştırılacağını belirtmektedir.
Aile tercihi modelinde ise, ailede insan sermayesi yatırımlarının bölüşülmesinde önemli rol oynayan ailenin tercih yapısı iki kısımdan oluşmaktadır. İlki, ailenin çocukları arasında eşitsizlikten kaçınma yönündeki tercihleri diğeri ise, her bir çocuğun aile fayda fonksiyonunda sahip olduğu göreli önemi, ağırlığıdır. Bu tercihlerin önemi nedeniyle, yatırım modelinin aksine, aile tercihi modelinde çocuklar arasındaki yatırımların bölüşümünde sadece etkinlik ilkesi değil, etkinlik ilkesi ile birlikte çocuklar arasındaki

eşitliğin belli ölçülerde gerçekleşmesi yönündeki ailenin duyarlılıkları da önemli rol oynamaktadır.

Bu çerçevede, aile tercihi modelinde kardeş cinsiyet bileşiminin insan sermayesi yatırımlarının çocuklar arasındaki bölüşümüne olan etkileri, ailenin çocuklarına miras bırakabilme olanaklarına, kız ve erkek çocukları arasındaki beklenen kazanç farklılıklarına, ailenin kız ve erkek çocuklar arasındaki eşitsizliklerden kaçınma yönündeki tercihlerine ve ailenin kız ve erkek çocuklarına fayda fonksiyonlarında verdikleri göreli öneme bağlıdır (Becker ve Tomes, 1979; Behrman vd., 1982, 1986).

Aile tercihi modelleri, çocukları arasındaki eşitsizlikten tamamen kaçınan ailelerde, insan sermaye yatırımının beklenen getirisi daha düşük olan çocuğa daha fazla yatırım aktarılacağını belirtmektedir (Behrman vd., 1982). Yapılan yatırımın kız çocuğu için daha az olması durumunda, kız çocuğuna daha fazla yatırım yapılacaktır. Bu durumda, yatırım modelinin aksine, ailedeki kız kardeşleri erkeklere yapılacak yatırımları olumsuz etkileyecektir. Ailedeki erkek kardeşleri ise kız çocuklarına daha fazla yatırım yapılmasına sebep olacaktır. Sonuç olarak, ailenin kız ve erkek çocukları arasındaki eşitsizlikten kaçınma yönündeki tercihleri kız kardeşlerinin olumlu etkisini olumsuza dönüştürebilir. Ayrıca, yatırım modelinin aksine, aile tercihi modelinde kız kardeşlerinin yatırımlara olan etkisi çocukların cinsiyetine göre simetrik değildir. Kız kardeşleri sadece erkek çocukları için yapılan yatırımları olumsuz bir şekilde etkilemektedir. Buna benzer bir etki kız çocukları için geçerli değildir.

Aile tercihi modelleri kız kardeşlerin etkilerinin ailelerin maddi olanaklarına göre nasıl farklılaştığı konusunda bir öngörüde bulunmamaktadır. Ancak, bu farklılaşmanın ailenin maddi olanakları ile çocuklar arasındaki eşitsizlikten kaçınma yönündeki tercihlere bağlı olduğu ileri sürülebilir. Eğer çocuklar arasındaki eşitsizlikten kaçınma derecesi ailenin maddi imkanları ile birlikte artıyorsa, kız kardeş sayısının erkek çocuklarına olumsuz etkisi maddi olanakları yüksek aileler için daha güçlü olacaktır. Öte yandan, eğer bu ikisi birbiriyle negatif bir şekilde bağlantılıysa, kız kardeş sayısının etkisi maddi imkanları düşük olan aileler için daha güçlü olacaktır.

Sonuç olarak, yatırım ve aile tercihi modelleri kız kardeş sayısıyla çocukların eğitimsel erişimleri arasında ilişkinin yönü hakkında birbirine karşıt öngörülerde bulunmaktadır. Yatırım modeli kız kardeş sayısının artmasıyla hem kız, hem erkek çocukların eğitimsel erişimlerinin artacağını öngörmektedir. Öte yandan, aile tercihi modelleri kız çocuklarının sayısıyla çocukların -sadece erkek çocukların- eğitimsel erişimleri arasında negatif bir ilişki olduğunu ileri sürmektedir. Dolayısıyla, kardeş cinsiyet bileşimlerinin çocukların eğitimsel erişimlerine olan etkilerini önceden kestirmek mümkün değildir. Bu nedenle, bu çalışma bir sonraki bölümde açıklanan ekonometrik analiz yöntemlerini kullanarak, bu etkilerin yönünü ve büyüklüğünü belirlemeye çalışmaktadır.

2.3 Verinin Yapısı ve Ekonometrik Analiz Yöntemleri

2.3.1 Verinin Yapısı

Bu çalışmada, kardeş cinsiyet bileşiminin aile içindeki çocuklara yapılan insan sermayesi yatırımlarına olan nedensel etkilerini incelemek amacıyla 2000 TÜİK Genel Nüfus Sayımı (GNS)'nin %5'lik rassal örneklemi kullanılmıştır. 2000 GNS'de hanehalkında bulunan bireylerin yaşı, cinsiyeti ve medeni durumları, hanehalkı reisine yakınlık derecesi gibi demografik özellikleri ve eğitimsel erişimleri konusunda detaylı bilgileri içermektedir. Her bir hanehalkı için bireyin hanehalkı reisine yakınlık derecesi ve cinsiyet bilgisi kullanılarak baba, anne ve çocuklardan oluşan bir çekirdek-merkez aile yapısı belirlenmiştir; hanehalkı reisi, eşi ve çocukları bu çekirdek-merkez ailenin temel bileşenlerini oluşturmaktadır. Ekonometrik analizde kullanılan örneklem bu çekirdek-merkez ailelerdeki çocuklarla sınırlı tutulmuş, ailenin ve çocuklarının demografik, eğitim ve sosyo-ekonomik özelliklerini ölçen değişkenler oluşturulmuştur[20]. Bu kısıtlama sonucu birden fazla ailenin bulunduğu hanehalkında yaşayan veya kendi başına hanehalkı kuran çocuklar örneklemden çıkarılmıştır. Bu kısıtlamanın tahminlere olası etkileri bir sonraki bölümde detaylı şekilde tartışılacaktır. Ailelerin çocuklarının eğitimleri için talep fonksiyonlarının ve çocukların okula gitme olanaklarının kırsal ve kentsel yerleşim birimlerine göre farklılık göstermesinin tahmin edilen sonuçları etkilememesi için sadece ilçe ve il merkezinde yaşayan aileler analiz edilen veriye dahil edilmiştir.

Aile içindeki çocukların cinsiyet bileşim yapısını ölçmek için kız kardeş sayısı kullanılmıştır. Çocukların eğitimsel erişimlerini ölçmek için iki eğitimsel çıktıya odaklanılmıştır: İlköğretim ve lise mezunu olma olasılıkları. Çocukların okula geç başlama ve öğretim yılını tekrar etme olasılıkları göz önüne alınarak, 16-18 yaş grubu için ilköğretim mezunu olma olasılığı ve 18-20 yaş grubu için lise mezunu olma olasılığı incelenmiştir. Ayrıca, diğer kardeş cinsiyet bileşimi ölçümlerinin çocukların eğitimsel çıktılarına olan etkileri de incelenmiştir.

2.3.2 Ekonometrik Analiz Yöntemleri

Bu çalışmada, gözlemler aile ekseninde kümelendirilerek panel veri benzeri bir veri oluşturulmuştur[21]. Bu bağlamda, bir ailenin analiz örneklemine dahil olabilmesi için incelenen yaş grubunda en az iki çocuğa sahip olması gerekmektedir. Bu oluşturulan

[20] Bu çalışmada, 2000 GNS kullanılarak ölçülen aile içindeki kardeşlerinin cinsiyet yapısının, çocuklar için insan sermayesi yatırımlarının yapıldığı dönemdeki kardeşlerinin cinsiyet yapısı ile aynı olduğu varsayılmıştır. Bu varsayım, genel olarak ailenin bütün özellikleri için geçerlidir. Çocuğun kardeşlerinin cinsiyet yapısının ve ailesinin temel özelliklerinin çocuğa yapılan insan sermayesi yatırımlarının yapıldığı zamanda ne olduğunu bilmek için, aynı çocukları zaman içinde takip eden ve bu çocukların bilgilerini barındıran panel verinin kullanılması gerekmektedir. Maalesef, böyle kapsamlı bir panel veri bulunmamaktadır.

[21] Kullanılan verinin panel veri olabilmesi için iki özelliğe sahip olması gerekmektedir. İlk olarak, verinin yatay-kesitli olması gerekmektedir. Bu çalışmada kullanılan verideki her bir aile, verinin yatay-kesit boyutunu oluşturmaktadır. İkinci olarak, aileye-özgü etkileri kontrol edebilmek için aynı ailenin birden fazla gözlenmesi gerekir. Veri sadece birden fazla çocukları olan ailelerden oluşturulduğu zaman bu özellik sağlanmış olur. Böylece incelenen verinin gözlem birimi olan aile, hem yatay-kesit, hem de zamansal boyutta gözlenmiş olur.

panel veri kullanılarak, rassal-etki (RE) tahmin yöntemiyle aşağıda belirtilen denklem tahmin edilmiştir[22].

$$Y_{ij} = \alpha + \beta_1 KKS_{ij} + \beta_2 KKS_{ij}^{\ 2} + X_{ij}\delta + Z_j\psi + a_j + \mu_{ij}$$

(1)

Burada:

j : Aile endeksini,

i : Çocuk endeksini,

Y_{ij} : J ailesindeki i çocuğunun incelenen eğitim süresini tamamlayıp tamamlamadığını gösteren iki değerli gösterge değişkenini,

KKS_{ij} : Kız kardeş sayısını,[23]

X_{ij} : Çocuğa özgü değişkenlerin vektörünü, (Çocuğun yaşı, cinsiyeti, doğum sırası)

Z_j : Aileye özgü değişkenlerin vektörünü, (Kardeş sayısı, anne ve babanın eğitimsel erişimleri için kukla değişkenleri, hanehalkında bulunan toplam insan sayısı, ailenin il merkezi veya ilçe merkezinde yaşadığını gösteren kukla değişkeni, ailenin hangi bölgede yaşadığını belirten kukla değişkenleri)

a_j : Ailedeki bütün çocuklar için geçerli olan aile özgü sabit etkileri,

u_{ij} : Rassal hatayı ifade etmektedir.

Doğrusal olasılık modeli kullanılarak elde edilen β'ların rassal-etki tahminleri, hem çocuğun, hem de çocuğun ailesinin gözlemlenebilen özellikleri kontrol edildiği zaman, kardeşlerin cinsiyet bileşiminin çocukların eğitimsel çıktılarına olan etkilerini ölçmektedir. Bu çalışmada bütün regresyon modellerinde, aileler arasındaki sabit olmayan varyans yapıları ve gözlemlerin aynı aileden gelmesi dolayısıyla ortaya çıkacak korelasyon dikkate alınarak tahmin edicilerinin standart hataları düzeltilerek hesaplanmıştır. Bu regresyon modeli metinde ve tablolarda etkileşimsiz model olarak belirtilecektir.

Ekonometrik analizde, ilk önce yukarıdaki denklem bütün veriler kullanılarak tahmin edilmiştir. Daha sonra, kız kardeş sayısının etkilerinin kız ve erkek çocuğa göre nasıl farklılaştığını inceleyebilmek için denklemdeki bütün değişkenlerin kız çocuk kukla değişkeniyle çarpımıyla elde edilen etkileşim terimleri yukarıda belirtilen modele dahil edilerek, aşağıdaki etkileşim modeli tahmin edilmiştir.

$$Y_{ij} = \alpha_0 + \alpha_1 KIZ_{ij} + \beta_1 KKS_{ij} + \beta_2 KKS_{ij}^{\ 2} + \lambda_1 (KIZ_{ij} * KKS_{ij}) + \lambda_2 (KIZ_{ij} * KKS_{ij}^2) + X_{ij}\delta_0 + (KIZ_{ij} * X_{ij})\delta_1 + Z_J\psi_0 + (KIZ_{ij} * Z_j)\psi_1 + a_j + \mu_{ij} \quad (2)$$

Bu etkileşim modelinde, tahmin edilen λ_1 ve λ_2 kız kardeş sayısının eğitimsel erişimlere olan etkilerinin kız ve erkek çocukları arasında nasıl farklılaştığını ölçmektedir.

[22] Bu çalışmada, Garg ve Morduch (1998a ve 1998b) kullandığı ekonometrik analiz yöntemleri bazı değişiklikler yapılarak kullanılmıştır.

[23] Literatürde yapılan diğer çalışmalara uygun bir şekilde, kız kardeş sayısının etkisinin doğrusal olmadığı düşünülerek kız kardeş sayısının karesi de regresyon modeline dahil edilmiştir. Ayrıca, kardeş cinsiyet bileşimi kız kareş sayısının toplam kardeş sayısına oranıyla da ölçülmüştür. Bu değişken kullanılarak yapılan ekonometrik analizlerde elde edilen sonuçların yönü ve istatistiksel anlamlılık seviyesi değişmemektedir.

RE tahmin edicilerinin sapmasız ve tutarlı olabilmesi için aileye özgü sabit etkileriyle, a_j, kız kardeş sayısının arasında korelasyon olmaması gerekmektedir. Ancak, bu varsayımın gerçekleşmesi imkansızdır. Çünkü aileler birbirinden farklıdır. Bu farklılık, ailelerin sahip olmak istedikleri çocuk sayısı, cinsiyeti ve çocuklarının sahip olmak istedikleri insan sermayesinin niceliği ve niteliği yönündeki tercihlerinin farklılığından kaynaklanmaktadır. Aileye özgü sabit etkilerin olumsuz etkilerini gidermek için sabit-etki tahmin etme yöntemi kullanılabilir. Fakat sabit-etki tahmin etme yöntemini gerçekleştirmek için aynı ailede bulunan çocuklar için kız kardeş sayısında değişkenlik olması gerekmektedir. Bu çalışmada incelenen çocukların yaş aralıklarının dar olması nedeniyle, aynı ailede bulunan çocuklar için bu değişkenlik yeterli düzeyde gerçekleşmemektedir. Bundan dolayı rassal-etki tahmin etme yöntemi uygulanmıştır.

Ayrıca, sabit-etki tahmin etme yöntemi kız kardeş sayısının içsel bir şekilde belirlemesinin ortaya çıkaracağı sorunları çözen bir yaklaşım değildir. Hem aileye, hem çocuğa özgü özelliklerden dolayı, ailede bulunan çocukların demografik yapısı -çocuk sayısının ve cinsiyetlerinin- çocuklar için yapılacak insan sermayesi yatırımlarıyla birlikte eşanlı olarak belirlenmektedir (Becker ve Tomes, 1976). Dolayısıyla, kız kardeş sayısının içsel bir şekilde belirlenme olasılığı, onun sabit-etki tahmin edicisinin sapmalı ve tutarsız olmasına neden olabilir. Bu içsellik problemini çözebilmek için en ideal çözüm, kız kardeş sayısıyla bağıntılı ama aileye ve çocuğa özgü sabit etkilerle bağıntısız bir araçsal değişken (instrumental variable) bulmaktır. Ancak, sayım verisinin sağladığı sınırlı bilgiler ve aileye özgü sabit etkilerle bağıntısı olmayan bir değişkenin bulunmasının zor olması bu çözümü olanaksız kılmaktadır. Özetlemek gerekirse, kız kardeş sayısının aynı ailedeki çocuklar için değişkenlik göstermemesi ve ideal bir araçsal değişken bulunmaması, araçsal değişken kullanarak sabit-etki tahmin etme yönteminin uygulanmasını imkansız kılmaktadır. Bu nedenle, ilgili literatürde de yapıldığı şekilde, bu çalışmada rassal-etki yöntemiyle tahmin edilen, çocukların eğitimsel çıktılarının fonksiyonlarının ailedeki çocukların demografik yapısına -kardeş cinsiyet bileşimine- koşullu indirgenmiş talep fonksiyonları olduğu unutulmamalı ve elde edilen sonuçlar bu çerçevede değerlendirilmelidir[24].

Diğer önemli bir sorun ise, kız kardeş sayısının çocukların gerçek kardeş cinsiyet bileşimini yansıtıp yansıtmadığıdır. Çalışmada ebeveynlerinden en az biri ile yaşadığı tespit edilen çocuklardan oluşan örneklem kullanılmıştır. Hanehelkı reisine yakınlık değişkeni kullanılarak anne, baba ve çocukları tespit edilmiştir. Verinin bu şekilde işlenmesi hanehalkında sadece tek bir çekirdek-merkez aile olması durumunda ebeveynleri ile birlikte yaşayan çocukları doğru bir şekilde temsil etmektedir. Ama hanehalkında birden fazla aile bulunması veya çocuğun kendi başına hanehalkı kurması durumunda bütün çocukları ve onlarla ilgili bilgileri türetmek imkansızdır. Bu nedenle, çalışmada sadece çekirdek-merkez ailelerde yaşayan çocuklar kullanılmıştır. Bu

[24] Bu indirgenmiş modellerde, kardeş cinsiyet bileşiminin etkisi kardeş sayısıyla birlikte doğum sırasına göre de koşullu olarak tahmin edilmektedir. Ailenin karar alma ve uygulama süreçleri düşünüldüğünde indirgenmiş modeldeki bütün değişkenlerin içsel değişken olma olasığı yüksektir. Daha önce metin içinde de belirtildiği gibi sonuçlar bu kısıtlı çerçevede değerlendirilmelidir. Bu sorunu aşmak amacıyla, Kırdar vd. (2007) yaptıkları çalışmada içsel bir değişken olan kardeş sayısı için ikiz kardeş doğumlarını araçsal değişken olarak kullanmışlardır. Yaptıkları bu çalışmada, sıradan en küçük kareler yöntemi tahminlerinin aksine, araçsal değişkenle tahmin etme yöntemini uygulayarak kardeş sayısıyla çocukların eğitimsel çıktıları arasında negatif bir ilişki olmadığına dair bulgular sağlamaktadır. Ancak, aynı çalışmada, sıradan en küçük kareler ve araçsal değişkenle tahmin etme yöntemlerinin farklılaşması kardeş cinsiyet bileşimi için gözlenmemiştir. Kız çocuklar için, her iki tahmin etme yönteminde erkek çocuk oranının etkisi negatif olarak tahmin edilmiştir.

kısıtlama sonucu, 16-18 yaş grubundaki çocukların % 17.4'ü, 18-20 yaş grubundaki çocukların ise % 23.6'sı kapsam dışında bırakılmıştır.

Böyle bir örneklem seçimi kardeş cinsiyet bileşimlerinin tahmin edilen etkilerinin sapmalı tahmin edilmesine neden olabilir. Çocukların anne ve/veya babalarıyla birlikte aynı hanehalkında çekirdek-merkez aile içinde bulunma eğilimleri, çocukların eğitimsel erişimleri ve kardeş cinsiyet bileşimlerinden bağımsız, rassal bir şekilde belirleniyorsa, tahmin edilen etkiler sıfıra doğru sapmalı olacaktır. Öte yandan, çocukların aileleriyle birlikte yaşama eğilimleri hem eğitimsel erişimleriyle, hem de kardeş cinsiyet bileşimleriye bağıntılı ise tahmin edilen etkiler yine sapmalı olacak, ama bu sapmanın yönünün önceden kestirilmesi olanaksız olacaktır. Benzer bir durum kardeş sayısının tahmin edilen etkileri için de geçerlidir. Bu nedenlerden dolayı, örneklem seçiminin tahmin edilen etkilere olan olumsuz etkilerini bertaraf etmek amacıyla kardeş sayısı ve kız kardeş sayısı için annenin canlı doğurduğu toplam çocuk sayısı ve kız çocuk sayısı araçsal değişkenler olarak kullanılmıştır[25]. Bu uygulanan rassal-etki, iki aşamalı en küçük kareler tahmin (RE-2AEEK) etme yöntemi kardeş ve kız kardeş sayısındaki ölçüm hatalarının olumsuz etkilerini telafi etmeyi amaçlamaktadır, ama bu yöntem bu iki değişkenin içselliğinden ortaya çıkan sorunlara bir çözüm olarak düşünülmemelidir.

3. BULGULAR

3.1 Ekonometrik Analizin Bulguları

3.1.1 Temel Bulgular

Kız kardeş sayısının aile içindeki çocuklar için yapılan insan sermayesi yatırımlarına olan etkilerini incelemek amacıyla 16-18 yaş grubundaki çocuklar için ilköğretim mezunu olma olasılığı ve 18-20 yaş grubundaki çocuklar için lise mezunu olma olasılığı incelenmiştir.

Tablo 1 kız kardeş sayısının çocukların eğitimsel erişimlere olan etkilerini betimlemektedir. İlköğretim mezunu olma olasılığı incelendiğinde, etkileşimsiz model (EM1) kız kardeş sayısının, çocukların ilköğretim mezunu olma olasılığıyla pozitif bir şekilde bağlantılı olduğunu göstermektedir ve bu tahmin edilen etki istatistiksel olarak % 1 anlamlılık düzeyinde anlamlı bulunmuştur[26].

[25] Regresyon modelinde, içsel değişken olarak kız kardeş sayısı ve kız kardeş sayısının karesi, kardeş sayısı ve kardeş sayısının karesi bulunmaktadır. Bu fonksiyonel yapı göz önüne alınarak, annenin canlı doğurduğu toplam çocuk sayısı ve onun karesi, kız çocuk sayısı ve onun karesi araçsal değişkenler olarak kullanılmıştır. Her bir içsel değişken için yapılan analizde, birinci aşamada tahmin edilen regresyon modellerinde kullanılan araçsal değişkenler istatistiksel olarak anlamlı bulunmuştur. Yapılan F-testlerinin en küçük değeri 8.77 olarak bulunmuştur

[26] Bu tahmin edilen etkilerin çocukların eğitimsel erişimleri için ne anlam ifade ettiği sonuç bölümünde hesaplanacak ve tartışılacaktır.

Tablo 1. Kız kardeş sayısının ilköğretim ve lise mezunu olma olasılığına tahmin edilen etkileri: RE ve RE-2AEKK tahminleri

	İlköğretim				Lise			
	RE tahmini		RE-2AEKK tahmini		RE tahmini		RE-2AEKK tahmini	
Bağımsız değişkenler	EM1	EM2	EM1	EM2	EM1	EM2	EM1	EM2
Kız kardeş sayısı	0.028***	0.034***	0.074***	0.044***	0.025***	0.033***	0.078***	0.035**
	(0.006)	(0.008)	(0.010)	(0.014)	(0.007)	(0.009)	(0.012)	(0.015)
Kız kardeş sayısının karesi	-0.001	-0.002	-0.010***	-0.005*	0.000	-0.002	-0.011***	-0.003
	(0.001)	(0.001)	(0.002)	(0.003)	(0.001)	(0.001)	(0.002)	(0.003)
Kardeş sayısı	-0.069***	-0.062***	-0.166***	-0.152***	-0.062***	-0.058***	-0.163***	-0.127***
	(0.005)	(0.007)	(0.013)	(0.017)	(0.006)	(0.007)	(0.014)	(0.017)
Kardeş sayısının karesi	0.003***	0.002***	0.008***	0.008***	0.002***	0.002***	0.008***	0.006***
	(0.0004)	(0.0004)	(0.001)	(0.001)	(0.0004)	(0.001)	(0.001)	(0.001)
Etkileşim terimleri								
Kız* kız kardeş sayısı		-0.003		0.012		-0.002		0.028
		(0.011)		(0.018)		(0.012)		(0.020)
Kız*kız kardeş sayısının karesi		0.001		-0.0004		0.0004		-0.003
		(0.002)		(0.004)		(0.002)		(0.004)
Kız*kardeş sayısı		-0.019**		-0.034*		-0.016		-0.064***
		(0.009)		(0.019)		(0.012)		(0.022)
Kız*kardeş sayısının karesi		0.001		-0.001		0.001		0.002
		(0.001)		(0.001)		(0.001)		(0.001)
R^2	0.186	0.194	0.155	0.538	0.195	0.205	0.463	0.466
Gözlem sayısı	29217	29217	28275	28275	21287	21287	20595	20595
B-P LM testi	1290.15 (0.000)	1303.07 (0.000)			1245.56 (0.000)	1263.58 (0.000)		

EM1: Etkileşimsiz Model EM2: Etkileşimli Model

Eğitimsel çıktılar için doğrusal olasılık modeli kullanılmıştır. RE tahminleri genelleştirilmiş en küçük kareler yöntemiyle elde edilmiştir. RE-2AEKK tahminleri için ise rassal etkiler modeli kullanılarak 2AEKK yöntemi kullanılmıştır. Regresyon modellerinde kullanılan bütün değişkenlerin listesi Ek 1'de verilmiştir. Standart hatalar parantez içinde verilmiştir. Aileler arasındaki değişen varyanslılık ve gözlemlerin aynı aileden gelmesi sonucu ortaya çıkacak korelasyon dikkate alınarak standart hatalar hesaplanmıştır. B-P LM testinin p-değeri parentez içinde verilmiştir. ***: % 1 anlamlılık düzeyi, **: % 5 anlamlılık düzeyi, *: % 10 anlamlılık düzeyidir.

Tablo 1'de sonuçları gösterilen EM2'deki *Kız*kız kardeş sayısı* etkileşim teriminin tahmin edilen katsayısı kız kardeş sayısının ilköğretim mezunu olma olasılığındaki erkek-kız farkını nasıl etkilediğini ölçmektedir. Bu etkileşim terimi sayısal ve istatistiksel olarak anlamlı bulunmamıştır. Öyleyse, ailedeki kız kardeş sayısı hem kız, hem erkek çocukların ilköğretim mezunu olma olasılığını olumlu bir şekilde etkilemekte, ama bu olasılıktaki kız-erkek farkının kapanmasına bir etkisi olmamaktadır.

Benzer bir ekonometrik analiz, kız kardeş sayısının lise mezunu olma olasılığına olan etkilerini incelemek için gerçekleştirilmiştir. Tablo 1'de gösterildiği gibi, kız kardeş sayısı 18-20 yaş grubundaki çocukların lise mezunu olma olasılığını olumlu bir şekilde etkilemektedir, ama erkek-kız arasındaki lise mezunu olma farklılıklarını etkilememektedir.

Burada tahmin sonuçları betimlenen RE modeli, hem kardeş cinsiyet bileşimi hem de eğitimsel çıktıları etkileyen aileye özgü gözlemlenemeyen sabit özelliklerinin, a_j ortalaması sıfır ve varyansı σ_a^2 olan bir dağılıma sahip olduğu varsayılarak, Genelleştirilmiş En Küçük Kareler (GEKK) yöntemiyle tahmin edilmiştir. Sıradan En Küçük Kareler (SEKK) yönteminin aksine, RE modeli eğitimsel erişimlerin belirlenmesindeki aileler arası gözlenmeyen farklılıkların kontrol edilmesine olanak tanır. Tablo 1'de sunulan Breusch-Pagan Lagrange Multiplier (B-P LM) testi aileye özgü özelliklerin ne kadar önemli olduğunun sınanmasına olanak vermektedir. Bu testin sonuçları, aileler arasındaki aileye özgü etkilerin önemli olduğunu ve dolayısıyla RE tahminlerinin daha etkin olduğuna dair bulgular sağlamaktadır. Bu nedenden dolayı, analize RE modeli kullanılarak devam edilmiştir[27].

GNS, diğer yatay-kesitli hanehalkı verilerinde olduğu gibi, anketin yapıldığı zaman ailesiyle birlikte yaşayan çocuklar için bilgiler sunmaktadır. Bu nedenle, nüfus sayımının hanehalkı kütüklerinden türetilen kardeş ve kız kardeş sayısı yanlış ölçülmüş olabilir. Ayrıca, çocuğun ailesiyle birlikte yaşama olasılığı çocuğun eğitimsel erişimleri ve kız kardeş sayısı ile bağlantılı olması durumunda elde edilen tahmin ediciler sapmalı ve tutarsız olabilir. Bu nedenlerden dolayı, kardeş sayısı ve kız kardeşleri sayısı için annenin canlı doğurduğu toplam çocuk sayısı ve kız çocuk sayısı araçsal değişkenler olarak kullanılmıştır. RE-2AEKK yönteminin sonuçları Tablo 1'de verilmiştir. Sonuçlar kız kardeş sayısının ilköğretim ve lise mezunu olma olasılığına olumlu katkısının devam ettiğine işaret etmektedir. İlköğretim mezunu olma olasılığı için kız kardeş sayısının tahmin edilen katsayısı 0.028'ten, 0.074'e, lise mezunu olma olasılığı ise 0.025'ten, 0.078'e yükselmiştir. Bütün bu tahmin edilen etkiler istatistiksel olarak anlamlı bulunmuştur.

3.2 Ailenin Sosyo-Ekonomik Statüsüne göre Kız Kardeş Sayısının Etkileri

Kız kardeş sayısının çocukların ilköğretim ve lise mezunu olma olasılıklarına olan etkilerinin ailenin maddi olanaklarına göre nasıl farklılaştığı Tablo 2'de incelenmiştir. Sayım verisi ailenin maddi olanaklarını ölçebilecek bilgilerden yoksun olduğu için

[27] Karşılaştırma yapabilme amacıyla, SEKK yöntemi kullanılarak elde edilen tahminler Ek.2'de verilmiştir. RE ve SEKK tahminleri, hem niceliksel hem de niteliksel olarak benzer sonuçlar vermektedir.

ebeveynlerin eğitim seviyesi kullanılarak aileler sosyo-ekonomik statülerine göre üç ayrı grupta sınıflandırılmıştır. Ebeveynlerinin en yüksek eğitim seviyesi;

- liseden az olanlar düşük
- lise olanlar orta
- üniversite olanlar ise yüksek

sosyo-ekonomik statülü aileler olarak tanımlanmıştır.

Tablo 2'de görüldüğü gibi, hem düşük hem orta sosyo-ekonomik statülü ailelerdeki kız ve erkek çocuklarının ilköğretimden ve liseden mezun olma olasılıkları kız kardeş sayısından olumlu bir şekilde etkilenmektedir. Düşük sosyo-ekonomik konumundaki ailelere odaklanıldığında, kız kardeş sayısının tahmin edilen katsayısı 0.025 ile 0.046 arasında değişmektedir. EM2'deki RE-2AEKK tahmin edicileri dışında, kız kardeş sayısının bütün tahminleri istatistiksel olarak anlamlıdır.

Düşük sosyo-ekonomik statülü ailelerdeki çocuklar ile karşılaştırıldığında, orta sosyo-ekonomik statülü ailelerden gelen çocuklar için kız kardeş sayısının tahmin edilen etkileri göreceli olarak daha güçlü bir şekilde ortaya çıkmaktadır. Kız kardeş sayısının tahmin edilen katsayısı 0.032 ile 0.110 arasında değerler almaktadır. Ayrıca, bu katsayıların istatistiksel anlamlılık seviyesi daha yüksektir.

Etkileşimli modellerde, etkileşim katsayılarının tahminlerinin küçük ve istatistiksel olarak anlamsız olması, kız kardeş sayısının bu incelenen iki eğitimsel erişimdeki erkek-kız farklılıklarını etkilemediğini işaret etmektedir.

Öte yandan, yüksek sosyo-ekonomik statüde olan aileler incelendiğinde, kız kardeş sayısının erkek çocuklarının eğitimsel erişimlerini etkilemediği bulunmuştur. Fakat, etkileşimli modeldeki etkileşim teriminin tahmin edilen katsayısı göreli olarak büyüktür ve bu katsayılarının RE-2AEKK tahminleri istatistiksel olarak anlamlıdır (ilköğretim için: 0.072, lise için: 0.078).

3.3 Farklı Kardeş Cinsiyet Bileşimlerinin Çocukların Eğitimsel Erişimlerine Etkileri

Çalışmanın bu kısmında farklı kardeş cinsiyet bileşim ölçümlerinin çocukların eğitimsel erişimlerine olan etkileri incelenmiştir. Tablo 3'te farklı modeller için GEKK yöntemiyle elde edilmiş RE tahminleri gösterilmiştir. İlk olarak, birinci modelde büyük ve küçük kız kardeş sayısının etkileri incelenmiştir (Model 1). Toplam kız kardeş sayısının tahmin edilen etkilerine benzer bir şekilde, hem büyük kız kardeş sayısı, hem de küçük kız kardeş sayısı çocukların ilköğretimi ve liseyi bitirme olasılıklarını pozitif bir şekilde etkilemektedir. Büyük kız kardeş sayısının tahmin edilen katsayısı 0.039-0.059 arasında, küçük kız kardeş sayısının katsayısı 0.023-0.045 arasında değişmektedir. GNS sonuçlarında büyük kız kardeş sayısının daha büyük olasılıkla yanlış ve eksik olarak ölçülebileceği dikkate alınırsa, küçük kız kardeş sayısının da eğitimsel erişimlerle pozitif bağlantısının bulunması bu çalışmada bulunan sonuçlara güven arttırmaktadır. Ayrıca, etkileşimli model incelendiğinde büyük kız kardeş sayısının liseyi bitirme olasılığındaki erkek-kız farklılıklarını azalttığına dair bulgular elde edilmiştir.

Tablo 2. Kız kardeş sayısının ailenin sosyo-ekonomik statüsüne göre tahmin edilen etkileri: RE ve RE-2AEKK tahminleri

	İlköğretim				Lise			
	RE tahmini		RE-2AEKK tahmini		RE tahmini		RE-2AEKK tahmini	
Bağımsız değişkenler	EM1	EM2	EM1	EM2	EM1	EM2	EM1	EM2
A. Düşük sosyo-ekonomik statülü aileler								
Kız kardeş sayısı	0.028**	0.031**	0.053**	0.031	0.025**	0.039**	0.046*	0.041
	(0.012)	(0.018)	(0.025)	(0.034)	(0.012)	(0.017)	(0.024)	(0.031)
Kız kardeş sayısının karesi	-0.002	-0.002	-0.005	-0.003	-0.001	-0.003	-0.004	-0.003
	(0.002)	(0.003)	(0.004)	(0.006)	(0.002)	(0.002)	(0.004)	(0.005)
Kız*kız kardeş sayısı		-0.002		0.020		-0.028		-0.002
		(0.022)		(0.045)		(0.022)		(0.044)
Kız*kız kardeş sayısının karesi		-0.0003		-0.001		0.002		-0.001
		(0.003)		(0.007)		(0.003)		(0.007)
R^2	0.076	0.082	0.265	0.256	0.063	0.069	0.177	0.175
Gözlem sayısı	4983	4983	4755	4755	4132	4132	3963	3963
B. Orta sosyo-ekonomik statülü aileler								
Kız kardeş sayısı	0.039***	0.051***	0.107***	0.077***	0.032***	0.037***	0.110***	0.057***
	(0.008)	(0.010)	(0.014)	(0.018)	(0.009)	(0.011)	(0.017)	(0.020)
Kız kardeş sayısının karesi	-0.001	-0.004*	-0.016***	-0.010**	-0.001	-0.003	-0.019***	-0.008**
	(0.002)	(0.002)	(0.003)	(0.004)	(0.002)	(0.002)	(0.004)	(0.004)
Kız*kız kardeş sayısı		-0.008		0.013		0.012		0.038
		(0.015)		(0.025)		(0.015)		(0.027)
Kız*kız kardeş sayısının karesi		0.001		-0.002		-0.001		-0.005
		(0.003)		(0.005)		(0.002)		(0.005)
R^2	0.107	0.114	0.484	0.483	0.075	0.086	0.370	0.372
Gözlem sayısı	19492	19492	18922	18922	13886	13886	13473	13473
C. Yüksek sosyo-ekonomik statülü aileler								
Kız kardeş sayısı	-0.006	-0.018	0.023	-0.034	0.004	-0.011	0.029	-0.025
	(0.013)	(0.017)	(0.021)	(0.029)	(0.016)	(0.023)	(0.025)	(0.032)
Kız kardeş sayısının karesi	0.002	-0.002	-0.005	0.004	0.005*	0.009*	0.002	0.012
	(0.003)	(0.004)	(0.006)	(0.008)	(0.003)	(0.004)	(0.006)	(0.008)
Kız*kız kardeş sayısı		0.032		0.072*		0.054		0.078*
		(0.026)		(0.041)		(0.033)		(0.042)
Kız*kız kardeş sayısının karesi		0.005		-0.005		-0.015**		-0.015
		(0.006)		(0.011)		(0.007)		(0.010)
R^2	0.134	0.144	0.743	0.740	0.070	0.085	0.741	0.743
Gözlem sayısı	4742	4742	4598	4598	3269	3269	3159	3159

Açıklamalar için Tablo 1'in açıklamalarına bakınız.

Tablo 3. Çeşitli kardeş cinsiyet bileşimi ölçümlerinin ilköğretim ve lise mezunu olma olasılıklarına tahmin edilen etkileri: RE tahminleri

		İlköğretim		Lise	
	Bağımsız değişkenler	EM1	EM2	EM1	EM2
Model 1	Büyük kız kardeş sayısı	0.059***	0.047***	0.045***	0.039***
		(0.010)	(0.010)	(0.010)	(0.011)
	Büyük kız kardeş sayısının karesi	-0.007**	-0.007**	0.002	0.001
		(0.003)	(0.003)	(0.003)	(0.003)
	Küçük kız kardeşi sayısı	0.045***	0.024***	0.041***	0.023**
		(0.009)	(0.009)	(0.010)	(0.010)
	Küçük kız kardeşinin sayısı	-0.003	-0.001	-0.003	-0.001
		(0.002)	(0.002)		(0.002)
	Kız*büyük kız kardeş sayısı		-0.012		0.028*
			(0.013)		(0.015)
	Kız*büyük kız kardeş sayısının karesi		0.009**		-0.004
			(0.004)		(0.004)
	Kız*küçük kız kardeşi sayısı		0.002		-0.018
			(0.011)		(0.013)
	Kız*küçük kız kardeşi sayısının karesi		-0.001		0.003
			(0.002)		(0.002)
	R^2	0.189	0.195	0.200	0.207
	Gözlem sayısı	29217	29217	21287	21287
Model 2	Erkek kardeş sayısı	-0.030***	-0.038***	-0.026***	-0.030***
		(0.006)	(0.007)	(0.007)	(0.009)
	Erkek kardeş sayısının karesi	0.001	0.003**	0.0005	0.001
		(0.001)	(0.001)	(0.001)	(0.001)
	Kız*erkek kardeşi sayısı		0.008		-0.006
			(0.011)		(0.012)
	Kız*erkek kardeşi sayısının karesi		-0.002		0.001
			(0.002)		(0.002)
	R^2	0.186	0.195	0.196	0.205
	Gözlem sayısı	29217	29217	21287	21287
Model 3	Büyük erkek kardeş sayısı	-0.039***	-0.044***	-0.053***	-0.038***
		(0.007)	(0.009)	(0.008)	(0.010)
	Büyük erkek kardeş sayısının karesi	0.001	0.004**	0.002	0.001
		(0.002)	(0.002)	(0.002)	(0.002)
	Küçük erkek kardeş sayısı	-0.016**	-0.031***	-0.003	-0.011
		(0.006)	(0.008)	(0.007)	(0.009)

Tablo 3. Çeşitli kardeş cinsiyet bileşimi ölçümlerinin ilköğretim ve lise mezunu olma olasılıklarına tahmin edilen etkileri: RE tahminleri

	İlköğretim		Lise	
Bağımsız değişkenler	EM1	EM2	EM1	EM2
Küçük erkek kardeş sayısının karesi	0.00007	0.003*	-0.0005	-0.0004
	(0.001)	(0.002)	(0.001)	(0.001)
Kız*büyük erkek kardeş sayısı		0.011		-0.024
		(0.013)		(0.016)
Kız*büyük erkek kardeş sayısının karesi		-0.005*		0.003
		(0.003)		(0.004)
Kız*küçük erkek kardeş sayısı		0.023**		-0.003
		(0.011)		(0.012)
Kız*küçük erkek kardeş sayısının karesi		-0.004**		0.002
		(0.002)		(0.002)
R^2	0.186	0.195	0.196	0.206
Gözlem sayısı	29217	29217	21287	21287

Model 2'de erkek kardeş sayısının eğitimsel erişimlere etkileri tahmin edilmiştir. Tablo 1'deki sonuçlarla tutarlı bir şekilde, erkek kardeş sayısının çocukların eğitimsel erişimlerine negatif etkisi saptanmıştır. Model 3'te ise büyük ve küçük erkek kardeş sayısının etkileri ayrı bir şekilde tahmin edilmiştir. Büyük erkek kardeş sayısı ilköğretim ve liseden mezun olma olasılıklarını negatif bir şekilde etkilerken, küçük erkek kardeş sayısının negatif etkisi sadece ilköğretimden mezun olma olasılığı için saptanmıştır. Etkileşimli modeldeki, kız*küçük erkek kardeş sayısı etkileşim teriminin pozitif ve istatistiksel olarak anlamlı olan katsayısı küçük erkek kardeş sayısının kız çocuklarını daha az olumsuz etkilediğini göstermektedir.

4. TARTIŞMA VE SONUÇ

Bu çalışmada, kardeş cinsiyet bileşimiyle çocuklar arasındaki insan sermayesi yatırımlarının dağılımı arasındaki nedensel ilişkiler incelenmiştir. Bu amaçla, ailedeki kardeş cinsiyet bileşimini ölçmek için kız kardeş sayısı, çocuklarına yapılan insan sermayesi yatırımlarının sonucunu ölçmek için ise çocukların ilköğretimden ve liseden mezun olma olasılıkları kullanılmıştır. Uygulanan ekonometrik analizlerinin sonuçları kız kardeş sayısının hem ilköğretimden mezun olma, hem de liseden mezun olma olasılıklarını olumlu bir şekilde etkilediğini göstermektedir. Ayrıca, kız kardeş sayısının etkilerinin çocukların cinsiyetine göre farklılaşmadığı önsavı reddedilememiştir. Başka bir şekilde belirtmek gerekirse, kız kardeş sayısı hem erkek, hem de kız çocuklarının eğitimsel erişimlerini aynı yönde ve şiddette etkilemekte ve bunun sonucunda eğitimsel erişimdeki cinsiyete özgü farklılıkları etkilememektedir. Bu bulgular, kız kardeş sayısının artmasıyla hem kız, hem de erkek çocuklarına yapılan insan sermayesinin artacağını öngören yatırım modelini destekler niteliktedir.

Kardeş cinsiyet bileşimlerinin tahmin edilen etkilerinin somut olarak ne anlama geldiğini gösterebilmek için kız çocuklarının sadece kız kardeşleri olduğu ve sadece erkek kardeşleri olduğu zaman ilköğretimden ve liseden mezun olma olasılıkları Tablo 1'deki etkileşimsiz modelin tahmin edilen katsayıları kullanılarak kestirilmiştir, benzer bir kestirme işlemi erkek çocukları için de yapılmıştır. Tablo 4'te gösterildiği gibi, bütün kardeşlerinin erkek olduğu bir erkek (kız) çocuk ile karşılaştırıldığı zaman, bütün kardeşleri kız olan erkek (kız) çocuğunun ilköğretim mezunu olma olasılığı % 11.29 (10.17), lise mezunu olma olasılığı ise % 15.69 (12.24) daha yüksektir. Bu elde edilen bulgular, aile içindeki kız kardeş sayısındaki artışın bütün çocukların eğitimsel başarılarını olumlu bir şekilde etkilediğini göstermektedir.

Tablo 4. Kardeş cinsiyet bileşimlerinin ilköğretim ve lise mezunu olasılıklarına etkilerinin kestirimi

Eğitimsel erişim	Çocuğun cinsiyeti	Kardeşlerin cinsiyet bileşimi		
		Hepsi erkek	Hepsi kız	Değişim (%)
	A. Bütün aileler			
İlköğretim mezunu olma	Erkek	0.62	0.69	11.29
	Kız	0.53	0.59	10.17
	Erkek/Kız	1.17	1.17	0
Lise mezunu olma	Erkek	0.51	0.59	15.69
	Kız	0.49	0.55	12.24
	Erkek/Kız	1.04	1.07	2.88
	B. Düşük sosyo-ekonomik statülü aileler			
İlköğretim mezunu olma	Erkek	0.37	0.45	21.62
	Kız	0.23	0.30	30.43
	Erkek/Kız	1.60	1.5	6.25
Lise mezunu olma	Erkek	0.26	0.32	23.07
	Kız	0.20	0.25	25.00
	Erkek/Kız	1.30	1.28	-1.53
	C. Orta sosyo-ekonomik statülü aileler			
İlköğretim mezunu olma	Erkek	0.60	0.67	11.66
	Kız	0.50	0.56	12.00
	Erkek/Kız	1.20	1.20	0
Lise mezunu olma	Erkek	0.48	0.55	14.58
	Kız	0.46	0.50	8.69
	Erkek/Kız	1.04	1.10	5.76
	D. Yüksek sosyo-ekonomik statülü aileler			
İlköğretim mezunu olma	Erkek	0.84	0.83	-0.01
	Kız	0.81	0.79	-0.02
	Erkek/Kız	1.03	1.05	0.02
Lise mezunu olma	Erkek	0.80	0.83	0.04
	Kız	0.83	0.84	0.01
	Erkek/Kız	0.96	0.99	0.03

Eğitimsel erişimlerde cinsiyete özgü farklılıkları ölçen Erkek/Kız oranlarının kardeşlerin hepsinin erkek veya kız olmasına göre değişmemesi, kız kardeş sayısının eğitimsel erişimlerdeki erkek-kız farklılıklarını etkilemediğini göstermektedir. Ayrıca, kardeş cinsiyet bileşiminin tahmin edilen etkilerinin ailenin sosyo-ekonomik statüsüne göre farklılaştığı gözlenmiştir. Bu etkiler en güçlü bir şekilde ailesi düşük ve orta sosyo-ekonomik konumda olan çocuklar için gözlenirken, yüksek sosyo-ekonomik statülü ailelerde gözlenmemiştir. Genel olarak, bu bulgular yatırım modelinin öngörülerini desteklemektedir. Kız kardeş sayısının artması çocukların eğitimsel erişimlerini arttırmakta ve bu pozitif etki maddi olanakları kısıtlı olan aileler için daha yüksek bir şekilde gerçekleşmektedir.

Sonuç olarak, bu çalışma aile içindeki çocukların eğitimsel erişimleriyle onların kardeşlerinin cinsiyet bileşimi arasındaki ilişkiler hakkında bulgular sağlamaktadır. Erkek ve kız çocukları için kardeşlerinin cinsiyet yapısına göre insan sermayesi çıktılarında farklılıklar gözlenmesi, en azından aynı cinsiyet grubundaki çocuklar arasında insan sermayesi yatırımlarının eşitsiz bir şekilde dağılacağına işaret etmektedir. Bundan dolayı, çocukların insan sermayesi düzeyini arttırmayı amaçlayan sosyal politikalar tasarlanırken, ailedeki çocukların demografik yapısından kaynaklanan eşitsiz insan sermayesi yatırımları dağılımının dikkate alınması gerekmektedir. Örneğin, aile içindeki çocuklarda okula kimin devam edebileceği kardeşlerinin cinsiyet ve diğer özelliklerine bağlıysa, okullarda uygulanan bedava kitap, bedava süt vb. uygulamaları çocukların insan sermayesi çıktılarındaki eşitsizlikleri bertaraf etmeyecek, daha da güçlendirecektir. Çünkü ancak okula gidebilen çocuklar bu politikalardan faydalanabilir. Öte yandan, örneğin, zorunlu eğitim süresinin uzatılması belli ölçülerde kardeş cinsiyet bileşiminin, çocukların insan sermayesi çıktılarına olan olumsuz etkilerini azaltabilir.

Bu çalışmada, çocukların kardeşlerinin cinsiyet bileşimleri onların insan sermayesi çıktılarına olan etkileri incelenirken, aile yapısı sabit olarak alınmıştır. Kardeşlerin sayıları ve cinsiyet bileşimine koşullu olarak kardeşlerin cinsiyet bileşimlerinin etkileri tahmin edilmiştir. Bu nedenle, bu çalışmada elde edilen sonuçlar değerlendirilirken dikkatli olunmalıdır. Ayrıca, GNS sonuçlarının sınırlı bilgi içermesinden dolayı, bu çalışmada aileleriyle birlikte yaşayan genç ve yaş aralığı dar olan bir nüfus incelenmiştir. Bu nedenle, bu çalışmada elde edilen bulgular kardeş cinsiyet bileşiminin kısa dönem etkileri olarak düşünebilir. Kardeşlerin cinsiyet bileşiminin uzun dönem ve kalıcı etkilerini incelemek amacıyla, yetişkin bireylerin ve onların kardeşlerinin kazanç, mesleki erişim, servet, miras ve ailelerinin demografik yapısı bilgilerini içeren bir veri seti oluşturulmalı ve incelenmelidir. Böyle bir yetişkin-kardeşler verisi sadece kardeş cinsiyet bileşiminin etkilerinin incelenmesinde değil, Türkiye'nin diğer sosyal ve ekonomik sorunlarının incelenmesinde ve nedensellik analizlerinin yapılmasında büyük olanaklar sağlayacaktır.

5. KAYNAKLAR

Bauer, A.T., Gang, I.N., 2001. Sibling rivalry in educational attainment: The German case. Labour, 15, 2, 237-255.

Becker, G.S., 1991. A treatise on the family. Harvard University Press, Cambridge, MA.

Becker, G.S., 1993. Human Capital: A theoretical and empirical analysis with special reference to education, Third Edition. The University of Chicago Press, London.

Becker, G.S., Tomes, N., 1976. Child endowment and the quantity and quality of children. Journal of Political Economy, 84, 4, 2, 143-162.

Becker, G.S., Tomes, N., 1979. An equilibrium theory of the distribution of income and intergenerational mobility. Journal of Political Economy, 87, 6, 1153-1189.

Behrman, J.R., Pollack R., Taubman, P., 1982. Parental preferences and provision for progeny. Journal of Political Economy, 90, 1, 52-73.

Behrman, J.R., Pollack R., Taubman, P., 1986. Do parents favor boys. International Economic Review, 27, 1, 33-54.

Butcher, K., Case A., 1994. The effect of sibling composition on women's education and earnings. Quarterly Journal of Economics, 04, 3, 531-563.

Das Gupta M., 1987. Selective discrimination against female children in rural Punjab, India. Population and Development Review, 13, 1, 77-100.

Dalton, C., 2000. Sibship sex composition: Effects on educational achievement. Social Science Research, 29, 441-457.

Edmonds, E.V., 2006. Understanding sibling differences in child labor. Journal of Population Economics, 19, 4, 795-821.

Garg, A., Morduch, J., 1998a. Sibling rivalry, Development Discussion Paper No. 630. Harvard Institute for International Development, Harvard University.

Garg, A., Morduch, J., 1998b. Sibling rivalry and the gender gap: Evidence from child health outcomes in Ghana. Journal of Population Economics, 11, 471-493.

Morduch, J., 2000. Sibling rivalry in Africa. The American Economic Review, 90, 2, 405-409.

Kaestner, R., 1997. Are brothers really better? Sibling sex composition and education achievement revisited. Journal of Human Resources, 32, 2, 250-283.

Kağıtçıbaşı, Ç., 1981. Çocuğun değeri: Türkiye'de değerler ve doğurganlık. Bogaziçi Üniversitesi İktisadi İdari Bilimler Fakültesi Yayınları, İstanbul.

Kırdar, M.G., Dayıoğlu, M., Tansel A., 2007. Impact of sibship size, birth order, and sex composition in urban Turkey. MPRA Working Paper No: 2755.

THE EFFECTS OF THE GENDER COMPOSITION OF SIBLINGS ON EDUCATIONAL ATTAINMENTS

ABSTRACT

Theories of family economics suggest that the gender composition of siblings could be an important determinant of the distribution of human capital investment among children. This paper uses variations in the number of female siblings across families to identify the effects of siblings' gender composition on the children's likelihood of completing second level primary and high school education. The results indicate that an increase in the number of female siblings boosts the completion rates of second level primary and high school education for both male and female children. Furthermore, the estimated effects appear to be larger for children from families with low and medium socio-economic status. However, there is no evidence that changes in the number of female siblings might alter male-female differences in the educational attainments.

Keywords: Family structure, Educational achievements.

Ek 1. Ekonometrik analizde kullanılan bağımlı değişkenlerin ve bazı bağımsız değişkenlerin ortalamaları: TÜİK, 2000 GNS

Değişkenin adı	16-18 yaş grubu		18-20 yaş grubu	
	Erkek	Kız	Erkek	Kız
Bağımlı değişkenler				
İlköğretim mezunu olma	0.565	0.468		
Lise mezunu olma			0.448	0.421
Bağımsız değişkenler				
Kız kardeş sayısı	1.649	1.673	1.721	1.691
Kız kardeş sayısının karesi	4.847	4.974	5.170	5.070
Kardeş sayısı	3.544	3.587	3.736	3.644
Kardeş sayısının karesi	18.861	18.539	21.018	19.502
Doğum sırası	2.335	2.276	2.350	2.210
Yaş	16.980	17.005	18.885	19.016
Hanehalkı büyüklüğü	7.13	7.10	7.48	7.27
Annenin eğitimi: Eğitimsiz	0.452	0.435	0.465	0.455
Annenin eğitimi: İlkokul mezunu	0.432	0.448	0.429	0.433
Annenin eğitimi: Orta okul mezunu	0.040	0.036	0.033	0.034
Annenin eğitimi: Lise mezunu	0.036	0.038	0.030	0.034
Annenin eğitimi: Üniversite mezunu	0.009	0.010	0.010	0.010
Annenin eğitimi: Bilgisi eksik	0.032	0.032	0.032	0.033
Babanın eğitimi: Eğitimsiz	0.154	0.143	0.168	0.160
Babanın eğitimi: İlkokul mezunu	0.528	0.526	0.510	0.507
Babanın eğitimi: Ortaokul mezunu	0.103	0.110	0.100	0.105
Babanın eğitimi: Lise mezunu	0.101	0.105	0.097	0.096
Babanın eğitimi: Üniversite mezunu	0.041	0.047	0.041	0.048
Babanın eğitimi: Bilgisi eksik	0.073	0.069	0.084	0.084

Ek 2. Kız kardeş sayısının ilköğretim ve lise mezunu olma olasılığının SEKK yöntemiyle tahmin edilen etkileri

	İlköğretim		Lise	
	SEKK tahminleri		SEKK tahminleri	
Bağımsız değişkenler	EM1	EM2	EM1	EM2
Kız kardeş sayısı	0.033***	0.033***	0.029***	0.029***
	(0.00)	(-0.083)	(-0.007)	(-0.094)
Kız kardeş sayısının karesi	-0.001*	-0.002	-0.001	-0.0013
	(0.001)	(0.001)	(0.001)	(0.001)
Kardeş sayısı	-0.069***	-0.060***	-0.061***	-0.054***
	(0.006)	(0.002)	(0.006)	(0.007)
Kardeş sayısının karesi	0.003***	0.002***	0.002***	0.002***
	(0.0004)	(0.0005)	(0.0004)	(0.0005)
Etkileşim terimleri				
Kız* kız kardeş sayısı		-0.001		0.00004
		(0.011)		(0.013)
Kız*kız kardeş sayısının karesi		0.0006		0.0002
		(0.002)		(0.002)
Kız*kardeş sayısı		-0.0175**		-0.017
		(0.01)		(0.012)
Kız*kardeş sayısının karesi		0.001		0.0008
		(0.0007)		(0.001)
R^2	0.186	0.197	0.195	0.205
Gözlem sayısı	29217	29217	21287	21287

DANIŞMA KURULU ÜYELERİ - ADVISORY BOARD MEMBERS