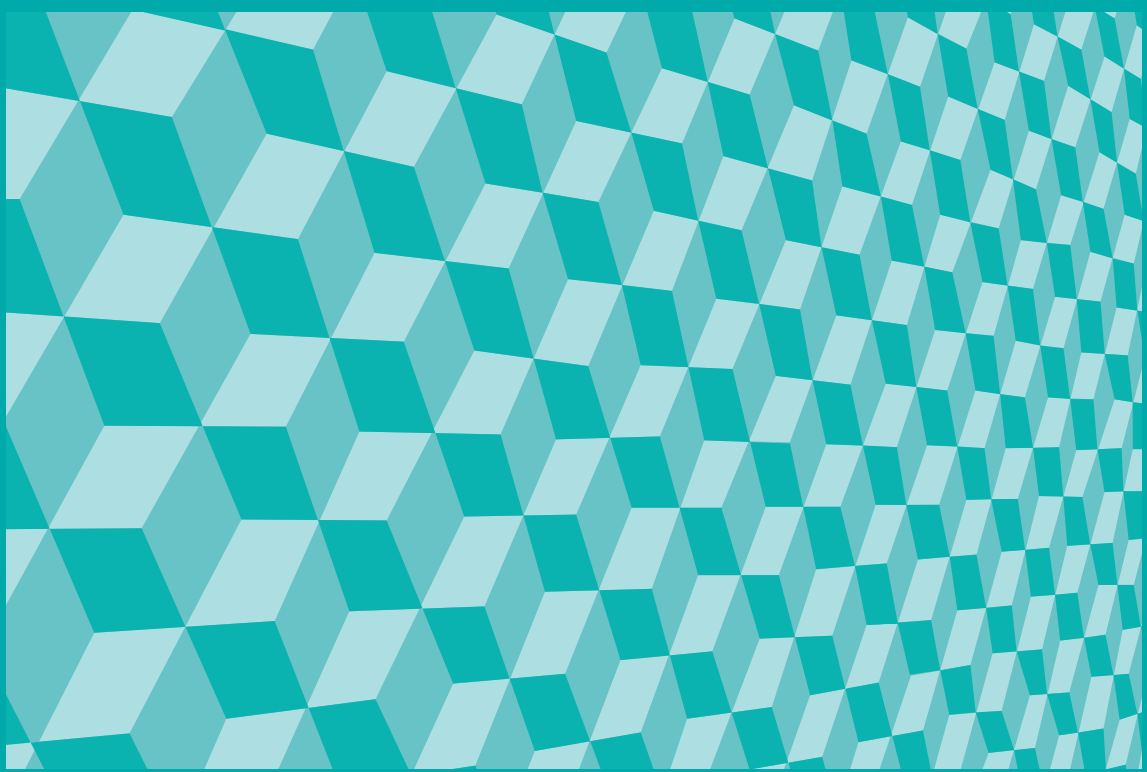
İSTATİSTİK ARAŞTIRMA DERGİSİ
Journal of Statistical Research

Cilt-Volume: 10 Sayı-Number: 03
Aralık-December 2013

ISSN 1303-6319



TÜRKİYE İSTATİSTİK KURUMU
Turkish Statistical Institute

İSTATİSTİK ARAŞTIRMA DERGİSİ
İAD JSR
2001

TÜİK
TÜRKİYE İSTATİSTİK KURUMU

İSTATİSTİK ARAŞTIRMA DERGİSİ
Journal of Statistical Research

Cilt-Volume: 10 Sayı-Number: 03
Aralık-December 2013

TÜRKİYE İSTATİSTİK KURUMU
Turkish Statistical Institute

Yayın istekleri için For publication order

Döner Sermaye İşletmesi Revolving Fund Management

Tel: +90 (312) 425 34 23 - 410 05 96 - 410 02 85
Faks-Fax**:** +90 (312) 417 58 86

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

Dergi Editörlüğü Journal Editorship

Tel: +90 (312) 417 64 45 / 206
Faks-Fax**:** +90 (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
4368

ISSN
1303-6319

Türkiye İstatistik Kurumu Turkish Statistical Institute

Devlet Mah. Necatibey Cad. No: 114 06650 Çankaya-ANKARA / TÜRKİYE

Türkiye İstatistik Kurumu Matbaası, Ankara Turkish Statistical Institute, Printing Division, Ankara
Tel: +90 (312) 387 09 25 * Fax: +90 (312) 418 50 82
Aralık 2014 December 2014
MTB: 2015-106 - 400 **Adet**-Copies

Editör Notu

Değerli Okuyucular ve Meslektaşlarım,

Türkiye İstatistik Kurumu tarafından 2001 yılından bu yana hakemli olarak yürütülmekte olan "İstatistik Araştırma Dergisi" ile istatistiki araştırmaların niteliğinin yükseltilmesi, kuramsal ve uygulama alanındaki araştırmacılar arasında iletişimin ortak çalışma ve yayınlarla güçlenmesi sağlanmaya çalışılmaktadır.

2004 yılı başından bu yana büyük bir özveriyle yürütmeye çalıştığım editörlük görevimden kendi isteğimle; diğer değerli meslektaşlarıma bu ulvi görevde bulunmalarına fırsat tanımak için, ayrılıyorum. Kavramsal, kuramsal ve uygulamalı çalışmalar olmak üzere toplam beş adet çalışmayı siz değerli okuyucularımızla paylaştığımız Dergimizin bu sayısı benim editörlüğünü yaptığım son sayıdır. Bu vesileyle Dergimiz ile ilgili bazı görüşlerimi sizlerle paylaşmak istiyorum.

Öncelikle belirtmeliyim ki; İstatistik Araştırma Dergisi Ülkemizde istatistik alanında yayımlanan, üniversitelerin doğrudan denetiminde olmayan, kurulduğundan bu yana yaşatılabilmiş tek bilimsel dergidir.

Bu derginin sayılarının üretilmesi aşamasında Türkiye İstatistik Kurumu yöneticilerinin herhangi bir şekilde baskıları olmamaktadır. Dergimizde, Türkçe ve İngilizce olmak üzere iki dilde kuramsal ve uygulama içerikli makaleler yayımlanmaktadır. İşte bu özelliklerinden dolayı bu derginin yaşatılması ve desteklenmesi gerektiğine inanmaktayım. Dergimizin istatistik biliminin değişik alanlarında çalışma yapanlarca sürekli olarak talep edilen ve izlenen bir yayın organı olabilmesi için;

1. Dergi için istatistik alnında en az doktora derecesi ve en az 10 yıl deneyimi olan editör dahil beş kişiden oluşan bir Editörler Kurulu'nun oluşturulması,
2. Editörler Kuruluca dergi yayın politikasının yeniden gözden geçirilmesi,
3. Editörler Kurulu üyelerinin, dergiye makale gönderebilecek kurum ve kuruluşlara Dergi hakkında sözlü ve yazılı aydınlatıcı bilgilerin iletilmesi ve makale gönderilmesi için ilgililerin teşvik edilmesi işini kendilerine misyon edinmelerinin sağlanması,
4. Derginin uluslararası kurum veya kuruluşlarca tanınması, tanıtılması ve ilgili endekslerde taranması için yapılması gerekenlerin öncelikle saptanması ve bunun derginin en önemli vizyonu olmasının sağlanması,
5. Dergiye hizmet edenler için, mümkün olduğu takdirde, bir ödül sisteminin oluşturulması,
6. Her yılın sonunda, ilgili alanda birikimi olan bilim insanlarından oluşan bir jüri tarafından o yıl içinde Dergide yayınlanmış en iyi makalenin belirlenmesi ve bu makalenin yazarına veya yazarlarına bir ödül verilmesi,
7. Derginin zamanında yayımlanması için gereken çabanın hem Kurul üyelerince hem de hakemlerce gösterilmesi ve makale sahibi/sahiplerince de zamanında geri dönüşleri yapmalarının sağlanması için önlemlerin alınması,

8. Dergide, kurumlarda ve kuruluşlarda (devlet veya özel sektör kuruluşlarında) istatistiğin belirli bir alanında gerçek verilerin kullanılarak yapılan uygulamalı özgün çalışmaların, belirli bir sayfa sınırlamasıyla, geniş özetinin yayımlanmasının sağlanması (belki bu tür çalışmalar için Dergide ayrı bölümünün ayrılması),

9. Dergi editörünün Ülkemiz dışında daha iyi tanınması için temaslarda bulunmak üzere yurt dışındaki ve yurt içindeki bilimsel toplantılara katılabilmesi için TÜİK tarafından maddi olarak desteklenmesinin sağlanması

gibi düzenlemelerin yapılmasının yerinde olacağı kanaatini taşıyorum.

Uzun süredir Dergimizin editörlüğünü, gücüm ve şartlar elverdiğince yürütmeğe çalıştım. Bu görevimde bana destek olan başta şu an görevde olan TÜİK Başkanı Sayın Birol AYDEMİR'e ve bu süre içinde daha önce görev yapmış başkanlara bana vermiş oldukları destek ve ilgi için, Dergi Sekreteryasında başta sevgili öğrencim Eğitim ve Araştırma Merkezi Müdürü Buket AKGÜN'e ve ekibindeki elemanlarına, Derginin gene benimle birlikte çok emeği geçmiş editor yardımcımız sevgili öğrencim Doç. Dr. Özlem İLK'e ve son olarak, Derginin bu sayısının ve şimdiye kadar yayınlanmış sayılarının gerçekleştirilmesinde emeği geçen, başta Dergimizde yayınlanması amacıyla makalelerini gönderen (yayınlanmasa bile) tüm araştırmacılara, katkılarıyla hiç bir karşılık beklemeksizin bizlere yardımcı olan tüm hakemlerimize ve TÜİK'in tüm değerli çalışanlarına hepimizin adına teşekkür etmeyi bir borç biliyorum.

Yeni görev alacak kişilere şimdiden başarılar dileyerek, içeriği ve kalitesi daha zengin Dergimizin yeni sayılarıyla bizleri buluşturmaları dileğiyle hepinize saygılarımı sunuyorum. Sağlıcakla kalın.

Prof. Dr. Fetih YILDIRIM
Dergi Editörü

TÜRKİYE İSTATİSTİK KURUMU
İSTATİSTİK ARAŞTIRMA DERGİSİ

TURKISH STATISTICAL INSTITUTE
JOURNAL OF STATISTICAL RESEARCH

Sahibi **Owner**
Türkiye İstatistik Kurumu Adına — On Behalf of Turkish Statistical Institute
Birol AYDEMİR — Birol AYDEMİR
Türkiye İstatistik Kurumu Başkanı — President, Turkish Statistical Institute

Editör **Editor**
Prof. Dr. Fetih YILDIRIM — Prof. Dr. Fetih YILDIRIM

Editör Yardımcısı **Assistant Editor**
Doç. Dr. Özlem İLK DAĞ — Assoc. Prof. Özlem İLK DAĞ

Sekreterya **Secretariat**
Buket AKGÜN
Nurdan ELVER

AMAÇ VE KAPSAM

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

"*İstatistik Araştırma Dergisi*"nin kapsamında yer alan tematik konular aşağıda özet olarak verilmiştir:

• Bankacılık, Finans, Sigortacılık, Aktüerya ve Risk Yönetimi; Bayesci İstatistik; Benzetim Teknikleri; Bilgi Sistemleri; Biyoistatistik; Bulanık Teori; Demografi; Deney Tasarımı ve Varyans Analizi; Ekonometri; Genel Sayımlar ve Değerlendirmeleri; İstatistik Eğitimi; İstatistik Etiği; İstatistik Kuramı; İstatistiksel Kalite Kontrolü; Kamuoyu ve Piyasa Araştırmaları; Klinik Denemeler; Mühendislikte İstatistik Uygulamaları; Olasılık ve Stokastik Süreçler; Optimizasyon; Örnekleme ve Araştırma Tasarımları; Parametrik Olmayan İstatistiksel Yöntemler; Resmi İstatistikler; Toplum Bilimlerinde İstatistik; Veri Analizi ve Modelleme; Veri Madenciliği; Veri Yönetimi ve Karar Destek Sistemleri; Verimlilikte İstatistiksel Yaklaşımlar; 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 Türkiye İstatistik Kurumu Başkanlığı'na aittir. Gerçek veya tüzel kişiler tarafından izinsiz çoğaltılamaz ve dağıtılamaz.
- Makale taslakları WORD yazım dilinde, Times New Roman yazı tipinde, 12 punto büyüklükte, satırlar arasında bir satır boşluk bırakılarak yazılmalı, şekil ve grafikler JPG dosyaları olarak hazırlanmalıdır.
- A4 sayfa boyutunda; soldan 3,5 cm, sağdan, yukarıdan ve aşağıdan 2,5 cm boşluk bırakılmalıdır.
- Ana bölüm başlıklarının tümü büyük harf, 12 punto büyüklükte, koyu, ortalı ve Arap rakamları ile numaralandırılarak; alt bölüm başlıklarında ise sadece kelimelerin baş harfleri büyük diğerleri küçük harfle, 12 punto büyüklükte, koyu, sola dayalı ve ana bölüm başlığına endeksli olarak Arap rakamları ile numaralandırılarak yazılmalıdır.
- Makale taslağı yazımında, okuyucunun, çalışmanın her aşamasını anlama ve değerlendirmesine olanak verecek bir anlatım ve plana uyulmalıdır.
- Anlatım olabildiğince sade, anlaşılabilir, öz ve kısa olmalıdır. Gereksiz tekrarlardan, desteklenmemiş ifadelerden ve konu ile doğrudan ilişkisi olmayan açıklamalardan kaçınılmalıdır.
- Yazımda çok genel ifadeler kullanılmamalıdır. Yargı veya kesinlik içeren ifadeler mutlaka verilere/ referanslara dayandırılmalıdır.
- Araştırmacı/araştırmacılar tarafından probleme, hangi kuramsal/kavramsal açıdan yaklaşıldığı, gerekçeleri ile birlikte belirtilmelidir.
- Kullanılan araştırma yönteminin seçilme gerekçesi açıklanmalıdır. Bütün veri toplama araçlarının geçerliliği ve güvenilirliği belirtilmelidir.
- Araştırma sonucunda elde edilen veriler bir bütünlük içinde sunulmalıdır.
- Sadece elde edilen verilere dayanan sonuçlar sunulmalıdır.
- Sonuçların yorumları, varsa, literatürdeki diğer kaynaklarla desteklenerek, değerlendirilmelidir.
- Yararlanılan kaynaklar, çalışmanın kapsamını yansıtacak zenginlik ve yeterlikte olmalıdır.
- Türkçe ve İngilizce özetler; çalışmanın amacı, yöntemi, kapsamı ve temel bulgularını içermelidir.

AIM AND SCOPE

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

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

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

Article Language and General Rules

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

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

Prof. Dr	Öztaş AYHAN	Orta Doğu Teknik Üniversitesi
Prof. Dr	Fetih YILDIRIM	Çankaya Üniversitesi
Prof. Dr	Birdal ŞENOĞLU	Ankara Üniversitesi
Prof. Dr	Olcay ARSLAN	Ankara Üniversitesi
Prof. Dr	M. Qamarul İSLAM	Çankaya Üniversitesi
Prof. Dr	Hasan BAL	Gazi Üniversitesi
Doç. Dr	A. Sinan TÜRKYILMAZ	Hacettepe Üniversitesi
Doç. Dr.	İlknur Yüksel KAPTANOĞLU	Hacettepe Üniversitesi
Doç. Dr	Hasan ÖRKÇÜ	Gazi Üniversitesi
Yrd. Doç. Dr.	Emel ÇANKAYA	Sinop Üniversitesi
Yrd. Doç. Dr.	Hakan Savaş SAZAK	Ege Üniversitesi
Yrd. Doç. Dr.	Haydar DEMİRHAN	Hacettepe Üniversitesi
Yrd. Doç. Dr.	Arzu ALTIN YAVUZ	Eskişehir Osman Gazi Üniversitesi
Yrd. Doç. Dr.	Özge ELMASTAŞ	Ege Üniversitesi

Volume: 10 Number: 03 Category: 01 Page: 1-14 ISSN No: 1303-6319 Cilt: 10 Sayı: 03 Kategori: 01 Sayfa: 1-14

FARKLI ÇALIŞMA ÖLÇEKLERİNDE SUÇ OLUŞUMUNA ETKİ EDEN DEĞİŞKENLERİN MEKANSAL İSTATİSTİK YÖNTEMİYLE KARŞILAŞTIRILMASI

Özlem DALAN* **Vahap TECİM****

ÖZET

20. yüzyıldan bu yana suçların oluşma nedenlerini anlamak ile ilgili birçok çalışma yapılmıştır. İlerleyen yıllarda da coğrafi konumun suçu anlama konusunda çok önemli bir katkısı olduğu tespit edilmiştir. Böylece araştırmacılar çalışma alanlarında suça etki eden faktörleri belirlerken mekansal faktörleri de araştırmalarına dahil etmişlerdir. Farklı bölgelerde gerçekleştirilen çalışmalarda suçu etkileyen faktörlerin konuma göre de farklılık gösterdiği sonucuna ulaşmışlardır. Ancak suçu etkileyen faktörler sadece konuma göre değil aynı konuma ilişkin yapılacak araştırmanın ölçeğine göre de değişiklik göstermektedir. İzmir'in merkez ilçelerini kapsayan alanda gerçekleştirilen bu çalışma ile sadece çalışma ölçeği değiştirilerek suça etki eden değişkenler tespit edilmiştir. Bölgesel ve yerel ölçek karşılaştırılarak gerçekleştirilen çalışmada mekansal istatistik yöntemlerinden yararlanılarak hırsızlık suçunun oluşmasını açıklayan faktörlerin değişiklik gösterdiği saptanmıştır. Elde edilen araştırma sonuçları suç önleme çalışmalarına karar destek oluşturacak geçerli bulgular içermektedir.

Anahtar Kelimeler: Coğrafi Bilgi Sistemleri, Karar destek, Mekansal analiz, Mekansal istatistik, Suç önleme.

1. GİRİŞ

Toplum içerisinde birlikte yaşamanın dezavantajı olarak ortaya çıkan suç oluşumu birçok araştırmacının bu alanda çalışmalar yapmasına neden olmuştur. Suç önleme çalışmaları da bunların bir çeşididir. Suçu önlemenin en başarılı yolu onu ortaya çıkaran nedenleri tespit etmek ve tekrar oluşmasını engellemektir. Bu yüzden suçun oluşmasında rol oynadığı düşünülen sosyal, kültürel, ekonomik ve/veya mekansal birçok faktör günümüze kadar gerçekleştirilen sayısız araştırmaya konu olmuştur (Yavuz ve Tecim, 2011).

Öyle ki suçların konumsal olarak nerede oluştuğunun bilinmesi bile tek başına suçların oluşma nedenini anlamak için önemlidir. Bu kapsamda incelendiğinde suç önleme çalışmalarında "coğrafi konum"un önemi 1930'lu yıllarda Shaw ve McKay tarafından "Chicago School"da oluşturulan *pinmaps* (meydana gelen suçların lokasyonlarının harita üzerine raptiyelerle işaretlenmesi) ile anlaşılmıştır (Chainey ve Ratcliffe, 2005). İlerleyen yıllarda da suçların oluşmasına etki eden değişken sayıları arttırılarak birçok farklı lokasyonda suçların oluşma nedenleri araştırılmaktadır.

Demografik, sosyal, kültürel, ekonomik ve/veya fiziksel çevre faktörleri kullanılarak gerçekleştirilen araştırmalardan elde edilen sonuçlar Tablo 1'de özetlenmiştir.

*Araştırma Görevlisi, Dokuz Eylül Üniversitesi, İktisadi ve İdari Bilimler Fakültesi Yönetim Bilişim Sistemleri Bölümü, e-posta: ozlem.dalan@deu.edu.tr

**Profesör, Dokuz Eylül Üniversitesi, İktisadi ve İdari Bilimler Fakültesi Yönetim Bilişim Sistemleri Bölümü, e-posta: vahap.tecim@deu.edu.tr

Tablo 1. Geçmiş yıllarda gerçekleştirilen araştırma sonuçları

Değişkenler		Araştırmacılar	Strano, 2004	Gillespie vd., 2009	Gorr ve Olligschlaeger, 1994	Olligschlaeger, 1997	Cahill ve Mulligan, 2003	Ceccato vd., 2002	Gruenewald vd., 2006	Schmid, 1960	Ackerman, 1998	Cozens, 2002	Malczewski ve Poetz, 2005	Ergün ve Yirmibeşoğlu, 2005	Newman, 1972	Murray vd., 2001	Ayhan, 2007	Anderson ve Anderson, 1984	Cohn, 1990	Field, 1992	Salleh vd., 2012
Psikolojik	**Karakteristik**	Suçluların karakteristik özellikleri	X	X																	
Sosyal çevre	**Ekonomik**	Fakirlik			X	X	X	X	X												
		İşsizlik						X	X	X	X										
		Düşük gelir								X	X										
		Ev sahipliliği									X										
	Kültürel	Eğitim düzeyi					X		X	X	X	X									
	Demografik	Nüfus yoğunluğu					X	X			X	X	X								
	Sosyal	Yaş				X			X												
		Göç						X			X										
Fiziksel çevre	**Mekansal**	Ticaret alanı varlığı				X						X									
		Şehir merkezinden uzaklık												X							
		Bina formları										X			X						
		Toplu ulaşım durak sayısı														X	X				
		Bina çevresel özellikleri													X		X				
	Zaman-mekan	Ortam sıcaklığı																X	X	X	X
		Rüzgar																			X
		Yağmur																			X
		Yüzey sıcaklığı																			X

İncelenen çalışmaların birçoğunda araştırmacılarınsuç oluşumunu açıklamak için tek bir değişken kümesi içerisindeki farklı değişkenleri kullandıkları görülmektedir.Buna rağmen gerçekleştirilen herbir çalışmanın farklı bulgular içerdiği gözlemlenmektedir. Örneğin; sosyal çevre verileri ile gerçekleştirilen araştırmalarda Schmid (1960) suç oluşumunu etkileyen faktörleri ekonomik ve kültürel değişkenler ile açıklarken Ceccate vd., (2006) ekonomik, demografik ve sosyal değişkenler ile açıklamaktadır. Bunun yanında farklı değişken kümeleri kullanılarak gerçekleştirilen çalışmalar da bulunmaktadır. Örneğin; farklı bulgular elde etmelerine rağmen Olligschlaeger (1997) ve Cozens (2002) suç oluşumunu sosyal çevre faktörlerini fiziksel çevre faktörleri ile birlikte ele alarak açıklamaktadır. Farklı yıllar ve farklı lokasyonlarda elde edilen bu

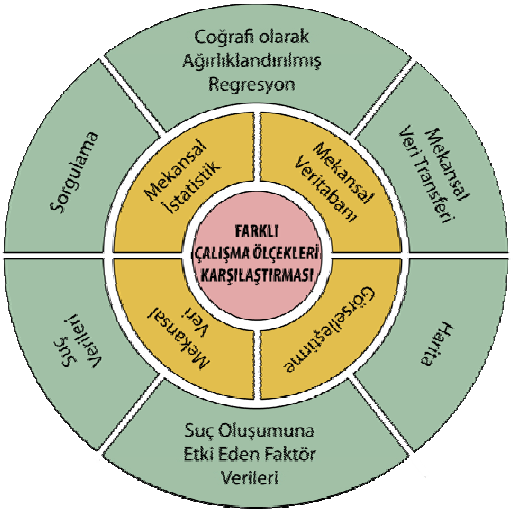
bulgular, bir çalışma alanında suçun ortaya çıkmasında etkili olduğu tespit edilen faktörlerin bir başka çalışma alanı için anlamlı olamayacağı şeklinde yorumlanmaktadır (Yavuz ve Tecim, 2013).

Bu doğrultuda suçların oluşma nedenlerinin ülkeden ülkeye, bölgeden bölgeye, kültürden kültüre, bölgenin sosyal ve ekonomik yapısına göre değişkenlik gösterebildiği kanısına varılmıştır (Aksoy, 2004). Ancak gerçekleştirilen çalışmalar incelendiğinde araştırmalarda kullanılan çalışma alanı büyüklüklerinin ve ölçeklerinin aynı olmadığı görülmektedir. Openshaw (1984) bu durumu suç oluşumunu anlama çalışmalarında farklı alansal büyüklüklerin kullanılmasının araştırma sonuçlarına etki ettiği şeklinde açıklamıştır. Bunun yanında araştırmalarda kullanılan çalışma ölçeklerindeki farklılığın araştırma sonuçlarına etki edip etmediği bilinmemektedir. Bu çalışma ile hiçbir alansal ve konumsal farkın oluşmasına imkan vermeksizin aynı çalışma alanında bölgesel ve yerel olmak üzere iki farklı çalışma ölçeği kullanılarak sadece ölçek etkisi ile suç oluşumuna etki eden değişkenlerin farklılık gösterip göstermediği araştırılmaktadır.

Sonuç olarak İzmir ilinin merkez ilçelerini kapsayan alanda gerçekleştirilen çalışma ile suç oluşumuna etki eden faktörlerin sadece farklı çalışma alanlarına göre değil çalışmada kullanılan ölçeğe göre de değişkenlik gösterdiği tespit edilmiştir. Bu durumun tespit edilmesiyle suç önleme çalışmaları yürüten emniyet teşkilatlarındaki üst düzey yöneticilerin araştırmalarında sadece alansal farklılıkları değil ölçek etkisini de göz önünde bulundurmaları gerektiğine ilişkin kararları destekleyecek geçerli bir sistem oluşturulmuştur.

2. YÖNTEM

Aynı çalışma alanı içerisinde suç oluşumuna etki eden faktörlerin çalışmanın ölçeği değiştirildiğinde farklılık gösterip göstermeyeceğinin belirlenebilmesi için farklı veri kaynaklarından yoğun veriler elde edilmiştir. Verilerin bölgesel ve yerel olarak belirlenen çalışma ölçeklerinin her ikisinde de kullanılabilecek şekilde düzenlenebilmesi için Şekil 1'de görüldüğü gibi bir yöntem izlenmiştir.



Şekil 1. Çalışmanın yöntemi

Ölçek farklılığının suç oluşumunu etkileyen faktörlerin değişiminin incelendiği çalışmada toplanan verilerin harita üzerinde mekansal olarak konumlandırılması gerekmektedir. Bu kapsamda ya veri içerisinde konumsal bilgilerin bulunması ya da konumsal bilgiye sahip olan diğer verilerle ilişkileri kurularak mekansal veriye dönüştürülmeleri gerekmektedir. Böylece farklı formatlara sahip olan veriler çalışma için oluşturulan ortak bir mekansal veri tabanına aktarılabilmektedir. Aynı veri tabanı içerisinde yer alan veriler içerisinden bağımlı ve bağımsız değişkenlerin belirlenmesiyle bölgesel ve yerel ölçekler için ayrı ayrı gerçekleştirilen mekansal istatistik yöntemiyle çalışma alanı içerisinde suçun oluşumunda etkili olan faktörler belirlenmektedir. Bu kapsamda; çalışma yönteminde kullanılan adımlar Tablo 2'de açıklanmıştır.

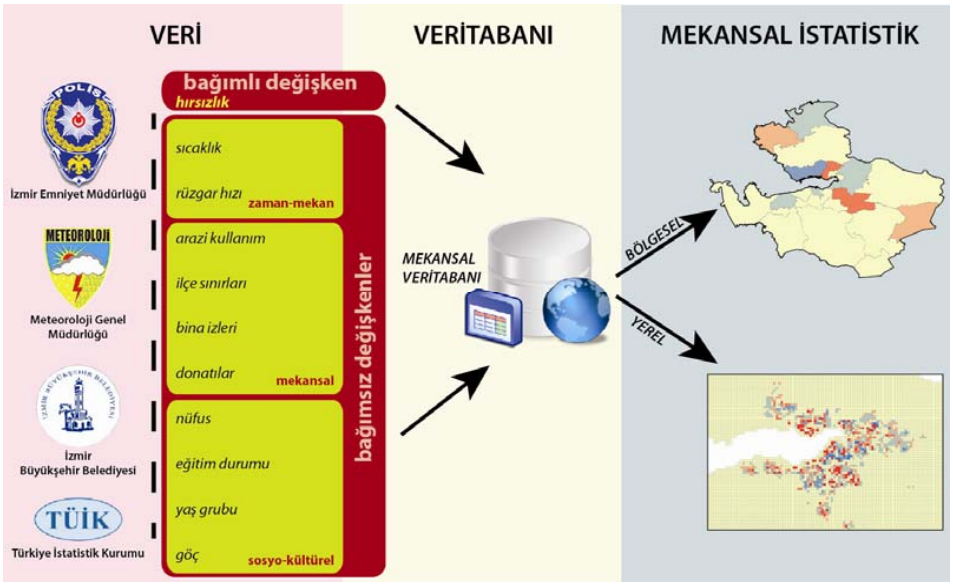
Tablo 2. Çalışma yönteminde kullanılan adımlar

Adımlar	Açıklama
Suç verileri	Suç lokasyonları ve gerçekleşme zamanları ile ilgili bilgilerin mekansal olarak düzenlendiği mekansal veridir.
Suç oluşumuna etki eden faktör verileri	Belirli zaman ve lokasyonda meydana gelen suçun oluşumunu etkileyen faktörlerin belirlenebilmesi için literatür çalışmasından elde edilen sonuçlara göre derlenen ve bağımsız değişkenler kümesini oluşturan mekansal veridir.
Sorgulama	Mekansal veriler (suç verileri ve suç oluşumunu etkileyen faktör verileri) içerisindeki bilgilerin çalışmada doğru bir şekilde kullanılmasını sağlamak amacıyla kullanılan mekansal istatistik yöntemidir.
Coğrafi olarak ağırlıklandırılmış regresyon	Bağımsız değişkenler kümesi içerisinden adımsal değişken seçimi yöntemi kullanılarak değişkenlerin suç oluşumunu etkileme olasılıklarının konumsal olarak değişiminin gösterilmesini sağlayan mekansal istatistik yöntemidir.
Mekansal veri transferi	Çeşitli veri kaynaklarından kağıt ortamda elde edilen ve mekansal veri tabanına aktarılan suç verileri ve suç oluşumunu etkileyen faktör verilerinin harita üzerinde konumsal olarak görselleştirilebilmesi için kullanılan bir yöntemdir.
Harita	Mekansal veri ve mekansal istatistik sonuçlarının görselleştirildiği ortamdır.

2.1 Kitle ve Örneklem

Çalışma İzmir ilinin yüzölçümü olarak %52'sini oluşturan ve birçok şehirsel faaliyeti birarada barındıran 21 merkez ilçesinde gerçekleştirilmiştir. Emniyet teşkilatları tarafından bilişim, ekonomik, narkotik, cinsel, şiddet, trafik, şahıs aleyhine işlenen, mal aleyhine işlenen ve diğer suçlar olmak üzere dokuz grupta sınıflandırılan suçlardan günlük faaliyetlerden en çok etkilenen suç türü olan mal aleyhine işlenen suçlar sınıfında yer alan farklı türdeki hırsızlık suçları çalışmanın araştırma konusunu oluşturmaktadır.

Çalışmanın bağımlı değişkenini oluşturan hırsızlık verileri İzmir Emniyet Müdürlüğü'nden Salleh vd., (2012) tarafından hırsızlık suçlarının en çok görüldüğü ay olarak belirlenen Ağustos ayını kapsayacak şekilde elde edilmiştir. 2010 yılı Ağustos ayında meydana gelen toplam 1.344 adet hırsızlık verisi suçun işlendiği tarih, saat ve gerçekleştiği konum bilgileri ile elde edilmiştir. Çalışmanın bağımsız değişkenlerini oluşturan zaman-mekan, mekansal ve sosyo-kültürel veriler Şekil 2'de gösterildiği gibi farklı veri kaynaklarından belirlenen her iki çalışma ölçeğine uyarlanabilecek formatta elde edilmiştir.



Şekil 2. Çalışma verileri

Bu kapsamda hem zaman hem mekana göre farklılık gösteren meteoroloji verileri, bölgenin sosyal yapısını ortaya koyan göç ve yaş grubu, kültürel yapısını ortaya koyan eğitim durumu, demografik yapısını ortaya koyan nüfus ile çevresel faktörleri ortaya koyan donatı alanlarına ilişikin konum bilgileri elde edilmiştir.

Zaman-mekan verileri çalışma alanı sınırı içerisinde yer alan toplam 6 adet klimatik meteoroloji istasyonu tarafından kaydedilen ve MGM (2012)'den elde edilen bilgilere göre sıcaklık ve rüzgar hızı verilerinde kritik değişimlerin meydana geldiği zaman dilimlerinde (sabah, akşam, gece) elde edilen toplam 558 adet veriden oluşmaktadır.

Sosyo-kültürel veriler Türkiye İstatistik Kurumu (TÜİK)'ndan ilçe bazında elde edilen göç, eğitim durumu ve yaş grupları verileri ile bu ilçelere ait 749 mahalleye ilişkin nüfus verilerinden oluşmaktadır.

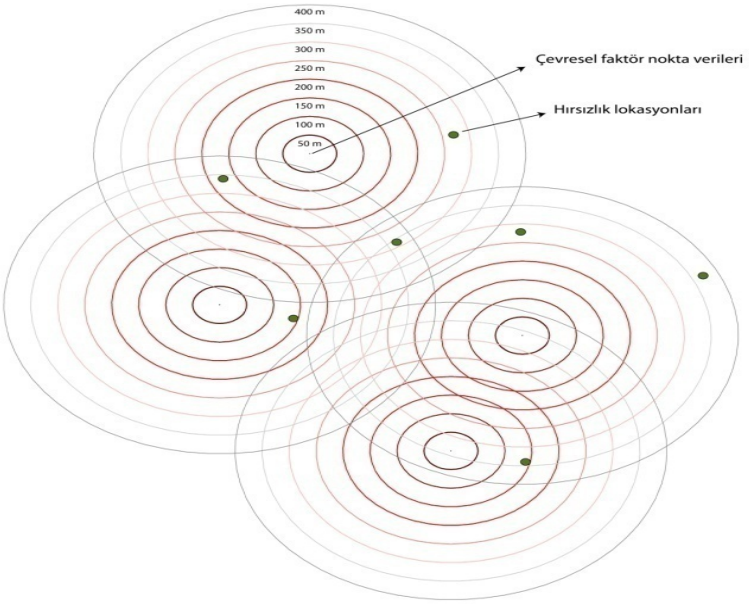
Çevresel faktör verileri İzmir Büyükşehir Belediyesi (İBB)'nden elde edilen koordinatlı toplam 10.536 adet donatı verisinden oluşmaktadır. Bu verilerden çalışma ile ilgili 641'i eğitim tesisi, 470'i dini tesis, 5.904'ü otobüs durakları, 59'u güvenlik birimleri, 377'si bankalar ve 526'sı ticaret alanlarından oluşan toplam 7.508 adet veri kullanılmıştır.

Elde edilen veriler mekansal veri tabanı içerisine aktarılmış ve mekansal istatistik yöntemi kullanılarak bağımsız değişkenler kümesi içerisinden her adımda denklemde olmayan bir başka bağımsız değişkenin seçilerek bölgesel ve yerel ölçeklerde hırsızlık suçunu etkileme olasılığı ölçülmüştür.

2.2 Veri Toplama Yöntemi

Çalışmada hırsızlık suçunu etkileyen faktörlerin belirlenen iki farklı çalışma ölçeğinde yaratacağı değişimi ölçebilmek için bölgesel ölçek olarak ilçe sınırları kullanılırken yerel ölçek olarak aynı çalışma alanının 400x400 metrelik hücrelere bölünmesiyle elde edilen gridler kullanılmıştır. Grid büyüklüğü belirlenirken Şekil 3'te gösterildiği gibi çevresel faktör nokta verilerine çizilen tampon bölge (buffer) analizinden yararlanılmıştır. Buna göre; başlangıç için 50 metre çapında tampon bölgeler

oluşturulmuştur. Oluşturulan tampon bölgenin tüm suç lokasyonlarını kapsayabilmesi için tampon bölge çapı aynı aralıkta arttırılarak denemelere devam edilmiştir. Sonuç olarak, hırsızlık suçlarının meydana geldiği lokasyonun çevresel faktör verilerine yürüme mesafesi olarak da tabir edilen 400 metre çapındaki alanda gerçekleşmiş olduğu tespit edilmiştir. Bu nedenle çalışma alanına uygulanan grid büyüklükleri 400 m^2 olarak seçilmiş olup çalışma alanını kapsayan toplam 35.785 adet grid hücresi elde edilmiştir.



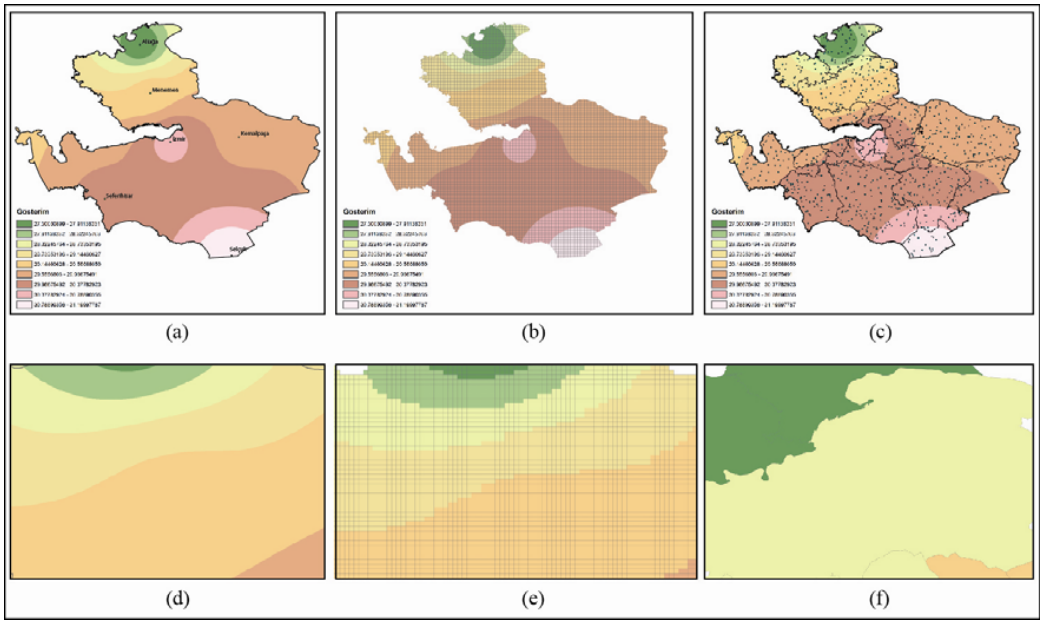
Şekil 3. Grid büyüklüğü seçimi

Kurumlardan elde edilen verilerin her iki ölçekte de kullanılabilmesi için mekansal veri dönüşümleri gerçekleştirilmiştir. Bu kapsamda hırsızlık verileri, tablo bilgisi içerisinde yer alan koordinat bilgileri kullanılarak Coğrafi Bilgi Sistemleri (CBS) teknolojileri vasıtasıyla sayısallaştırılmıştır. Aynı zamanda meteoroloji verileri ile örtüşecek şekilde suçların gerçekleşme saati 8'er saatlik zaman dilimlerine göre sabah (04:00-12:00), akşam (12:00-20:00) ve gece (20:00-04:00) olarak gruplandırılmıştır. Bu sınıflandırmaya göre gerçekleşen suçlar bölgesel ve yerel ölçeklerde kullanılmak üzere mekansal veri tabanına aktarılmıştır.

Birbirlerinden uzak konumlarda yer alan meteoroloji istasyonlarından elde edilen verilerin belirlenen her iki çalışma ölçeğinde kullanılabilmesi için meteoroloji istasyonları arasında sürekliliği bulunan bir yüzey oluşturulması gerekmektedir. Brunsdon vd. (2001)'ne göre birbirinden ayrık konumlanmış olan nokta verileri için sürekli bir yüzey oluşturabilmenin en iyi yolu noktalar arasında enterpolasyon uygulamaktır. Bu kapsamda klimatik istasyonlardan elde edilen sıcaklık ve rüzgar hızı verilerinin enterpolasyonu Ağustos ayı içerisindeki zaman dilimlerine (sabah, akşam, gece) göre ayrı ayrı

$$Z_i = \frac{\sum_{j=1}^{n} \frac{z_j}{d_{ij}^k}}{\sum_{j=1}^{n} \frac{1}{d_{ij}^k}} \qquad (1)$$

eşitliğiyle hesaplanmıştır (Shepard, 1968). Eşitlikte yer alan Z_i i lokasyonundaki sıcaklık/rüzgar hızı, Z_j beklenen j lokasyonundaki sıcaklık/rüzgar hızı, d_{ij} i den j ye olan uzaklık, k enterpolasyon kuvveti (k=2)'dir. Enterpolasyon ile hesaplanan yüzeyin çıktısı Şekil 4(a)'da görüldüğü gibi raster haritadan oluşmaktadır. Raster haritada yer alan herbir piksel değerinin bölgesel ve yerel ölçeklerde kullanılabilmesi için yerel ölçekte grid hücrelerinin merkezleri kullanılarak toplam 35.785 adet (Şekil 4(b)), bölgesel ölçekte ise ilçe sınırları içerisinde kalacak şekilde rastgele yerleştirme yöntemi kullanılarak 1.000 adet nokta üretilmiştir (Şekil 4(c)). Böylelikle herbir pikselin içerisinde kalan noktalara ilgili pikseldeki sıcaklık ve rüzgar hızı değerleri atanmıştır (Şekil 4(d)). Yerel ölçekte üretilen noktalara atanan değerler grid hücre değeri olarak aynen kullanılırken (Şekil 4(e)), bölgesel ölçekte üretilen noktalara atanan değerlerin ortalamaları hesaplanarak bu değerler, içerisinde yer aldığı ilçe sınırına atanmıştır (Şekil 4(f)).



Şekil 4. Meteoroloji verilerine uygulanan enterpolasyon yöntemi

Nüfus yoğunluğunu hesaplayabilmek için bölgesel ölçekte ilçe nüfusları ile ilçelerin yüzölçümü bilgilerinden yararlanılırken yerel ölçekte ise bina başına düşen nüfus ile 0.16 km^2 lik sabit değere sahip olan grid yüzölçümü kullanılmıştır. Nüfus bilgileri ilçe ve mahalle bazında TÜİK'den elde edilmiştir. Bölgesel ölçekte ilçe nüfusları yoğunluk hesabına olduğu gibi dahil edilmiştir. Yerel ölçekte ise bina başına düşen nüfusun hesaplanmasında kullanılmıştır.

Bina başına düşen nüfus; daire sayısı, kat yüksekliği ve hanehalkı sayılarının birbirleriyle çarpılması sonucu elde edilmektedir. Bu kapsamda 2010 yılına ait arazi kullanım haritaları kullanılarak konut alanları vektörel haritalar üzerinde tespit edilmiştir. Farklı kullanım türlerini (garaj, depo, vb.) konut grubundan ayırmak ve yüzeyde kaç m^2 alan işgal ettiklerine göre konutların her katta sahip oldukları daire sayılarını hesaplamak amacıyla Tablo 3'te görüldüğü gibi bir sınıflama yapılmıştır.

Tablo 3. Konutların işgal ettikleri yüzeylere göre daire sayıları

Apartmandaki daire sayısı	İşgal yüzeyleri (m^2)	Bina sayısı
Müştemilat (depo)	< 60	117,340
Tek daire	60 – 200	360,950
İki daire	200 – 300	28,249
Üç daire	300 – 400	8,681
Dört daire	400 – 550	5,712

Daire sayılarına göre sınıflandırılan konut verilerine vektörel haritada yer alan kat yüksekliği bilgileri eklenmiştir. Sayısal olarak bilinmeyen hanehalkı sayısı ise mahalle ve bina başına düşen ortalama nüfusun ortalaması üzerinden hesaplanmıştır (Ural vd., 2011). Buna göre mahallelerin ortalama nüfus yoğunluğu 100 m^2'de 0,26 kişi, konutların ise 3,49 olarak hesaplanmıştır. Hesaplanan bu iki değerin ortalamasını aldığımızda ortaya çıkan "2" değeri hanehalkı sayısı olarak kullanılmıştır. Böylece bina başına düşen nüfusun hesaplanmasında, toplam mahalle nüfusu ile karşılaştırıldığında %96,7 oranında doğru sonuç elde edilmiştir (mevcut mahalle nüfusu 3.479.507, hesaplanan bina nüfusu 3.364.476).

Elde edilen bina nüfusu üzerinden bina bazında göç, yaş grubu ve eğitim durumu çıkarsamaları gerçekleştirilmiştir. TÜİK (2008)'den elde edilen bilgilere göre hırsızlık suçlarından hüküm giyen kişilerin 18-35 yaş arası ilkokul mezunu erkek nüfustan oluşması dolayısıyla yaş grubu ve eğitim durumu verileri bu kapsamda incelenmiştir.

Çevresel faktörler bölgesel ölçekte suçun işlendiği lokasyona olan uzaklıklarıyla değerlendirilmiştir. Yerel ölçekte ise herbir grid hücresine olan mesafeleriyle değerlendirilmiştir. Aynı zamanda bölgesel ölçekte ilçe sınırı içerisinde kalan toplam donatılar ile yerel ölçekte herbir grid hücresi içerisinde kalan toplam donatı değerleri hesaplanmıştır.

2.3 Veri Analizi

Düzenlenerek mekansal veri tabanına aktarılan tüm bağımsız değişkenlerin bölgesel ve yerel ölçeklerde çalışma alanında meydana gelen hırsızlık suçlarının oluşmasını etkileme düzeyleri geleneksel istatistik yöntemlerine kıyasla mekansal ilişkileri modelleyebilme yeteneğine sahip olan mekansal istatistik yöntemleriyle hesaplanmıştır (Fotheringham vd., 2002) .

Bağımlı değişken ile karşılaştırıldığında çok büyük değerlere sahip olan sosyo-kültürel ve mekansal değişkenlere normalizasyon işlemi uygulanmıştır. Bölge hakkında fikir vermesi açısından sosyo-kültürel değişkenler (göç, yaş, eğitim);

$$\frac{x\ bölgesindeki\ miktar}{x\ bölgesindeki\ nüfus} \tag{2}$$

oranına göre normalize edilirken nüfus yoğunluğu ve çevresel faktörlere (donatı sayıları) ilişkin değişkenler;

$$\frac{x\ bölgesindeki\ miktar}{çalışma\ alanı\ toplam\ değeri} \tag{3}$$

oranı kullanılarak normalize edilmiştir.

Bölgesel ve yerel ölçeklerde regresyona dahil edilen nüfus yoğunluğu, eğitim durumu (ilkokul mezunu erkek sayısı), yaş grubu (18-35 yaş arası erkek nüfusu), göç, sıcaklık, rüzgar hızı, donatı sayıları ve herbir donatıya olan mesafelere ilişkin açıklayıcı istatistikler Tablo 4'de verilmiştir.

Tablo 4. Bağımsız değişkenlere ilişikin açıklayıcı istatistikler (bölgesel ölçekte n= 21, yerel ölçekte n=35.785)

Değişken Adı	Bölgesel Ölçekte				Yerel Ölçekte			
	min	max	varyans	Standart sapma	min	max	varyans	Standart sapma
Nüfus yoğunluğu	0,002	0,306	0,006	0,08	0	2,681	0,016	0,13
Eğitim durumu	0,007	0,027	$2{,}765e^{-005}$	0,01	0	2,571	0,016	0,13
Yaş grubu	0,101	0,522	0,160	0,09	0	4,252	0,017	0,13
Göç	0,138	7,29	2,08	1,44	0	1,151	0,001	0,02
Ortalama sıcaklık	2,80	3,10	0,004	0,06	2,750	3,120	0,004	0,06
Ort. rüzgar hızı (km)	1,180	3,323	0,192	0,44	0,963	3,460	0,251	0,50
Donatı sayısı	0,001	0,147	0,002	0,04	0	1,247	0,001	0,03
Durak sayısı	0,002	0,111	0,001	0,03	0	0,675	0,000	0,02
Dini tesis sayısı	0,001	0,291	0,005	0,07	0	1,702	0,001	0,03
Ticaret sayısı	0,001	0,350	0,007	0,09	0	1,711	0,002	0,04
Eğitim tesisi sayısı	0,001	0,222	0,004	0,06	0	1,404	0,001	0,03
Güvenlik birimi sayısı	0,002	0,397	0,008	0,09	0	3,390	0,005	0,07
Banka sayısı	0,001	0,416	0,009	0,09	0	4,775	0,004	0,07
Donatıya olan mesafe	0	0,774	0,027	0,16	0,001	0,199	0,001	0,03
Durağa olan mesafe	0,004	0,221	0,005	0,07	0,001	0,199	0,001	0,03
Dini tesise olan mesafe	0,001	0,160	0,003	0,06	0,001	0,458	0,011	0,11
Ticarete olan mesafe	0,002	0,184	0,004	0,06	0,001	0,447	0,012	0,11
Eğitim tesisine olan mesafe	0	0,725	0,023	0,15	0,001	0,448	0,010	0,10
Güvenlik birimine olan mesafe	0,002	0,149	0,003	0,05	0,001	0,500	0,013	0,11
Bankaya olan mesafe	0,002	0,163	0,003	0,05	0	0,477	0,013	0,11

Nüfus yoğunluğu İzmir'in merkez ilçelerinde körfezin yer aldığı iç kesimlerde maksimum seviyelere sahip iken körfezden uzaklaştıkça ve çeperlere doğru gidildikçe azalmaktadır. Nüfus yoğunluğunun minumum olduğu alanlar sayfiye yerleri ile tarım faaliyetlerinin yürütüldüğü ilin çeperlerini oluşturan kuzey, güney ve doğu bölgelerinde görülmektedir. Turizm faaliyetlerinin de yürütüldüğü Foça, Urla, Seferihisar gibi denize kıyısı bulunan ilçelerde ilkokul mezunu erkek sayısı az iken şehirsel faaliyetlerin daha yoğun bir şekilde gerçekleştirildiği Konak, Bayraklı, Karabağlar, Bornova ve Buca ilçelerinde bu sayı artmaktadır. 18-35 yaş arasındaki genç erkek nüfus ve göç durumu için de günlük ve şehirsel faaliyetlerin yoğun gerçekleştirildiği ilçeler ve bu ilçelerin çeperlerinde yer alan diğer ilçeler maksimum ve maksimuma yakın değerlere sahiptir. Çalışma alanı içerisinde yer alan donatı sayıları iç kesimlerden çeperle doğru azalırken donatılara olan mesafeler çeperlere doğru artış göstermektedir.

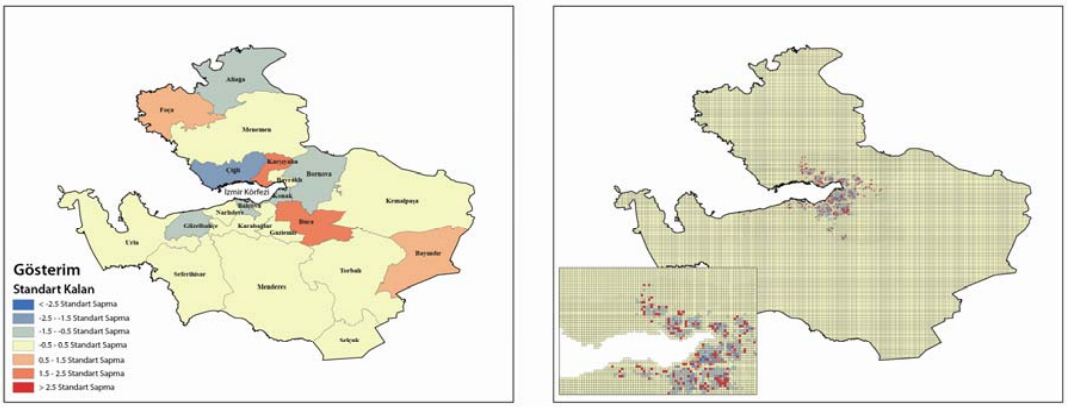
Bu değerlendirmelerin hırsızlık suçlarının oluşmasını etkileme düzeylerini mekansal olarak inceleyebilmek amacıyla coğrafi olarak ağırlıklandırılmış regresyon yöntemi çalışma alanında en küçük kareler yöntemiyle birlikte uygulanmıştır. Bölgesel ve yerel ölçeklerde bağımlı ve bağımsız değişkenler arasındaki mekansal ilişkiler

$$y_i = \beta_0(u_i, v_i) + \sum_k \beta_k(u_i, v_i) x_{ik} + \varepsilon_i \qquad (4)$$

eşitliğiyle hesaplanmıştır. Buna göre; y_i bağımlı değişken olan hırsızlık sayısının her ilçe/grid hücresi içerisindeki i. gözlem değeri, β_0 sabit, β_k k değişkeninin tahminleme parametresi, x_{ik} i. gözlem için k. değişkenin kovaryasyon değeri, (u_i, v_i) i lokasyonunun koordinatı ve ε_i hata terimidir.

Meteoroloji ve hırsızlık verileri için belirlenen zaman dilimleri kullanılarak sabah, akşam ve gece gerçekleşen suçların oluşumunda etkili olan faktörler bölgesel ve yerel

ölçekler için ayrı ayrı hesaplanmıştır. Her iki ölçekte sabah gerçekleşen suçların bağımsız değişkenlerle olan mekansal ilişkisinin tespit edildiği harita Şekil 5'te gösterilmiştir.



Şekil 5. Bölgesel ve yerel ölçeklerde sabah gerçekleşen hırsızlık suçlarına uygulanan coğrafi olarak ağırlıklandırılmış regresyon yönteminin çıktı haritası

Buna göre; bölgesel ölçekte sabah saatlerinde gerçekleşen suçların demografik, sosyo-kültürel çevre özelliklerinin tümü ve mekansal faktörlerden toplu ulaşım ve eğitim tesisi varlığı ile %94 oranında açıklanabileceği saptanmıştır (Şekil 5(a)). Körfezin çevresinde yer alan ilçelerde diğer bölgelere kıyasla daha büyük eğitim tesislerinin bulunması (üniversite kampüsleri) ve alansal büyüklük dolayısıyla toplu ulaşım duraklarının birbirlerinden uzak konumları standart sapmanın bu alanlarda ± 2,5 arasında değişiklik göstermesine yol açmaktadır. Aynı zaman diliminde ve aynı konumsal çerçeve ile sınırlandırılmış çalışma alanında yerel ölçek kullanılarak gerçekleştirilen mekansal analiz sonucuna göre; bölgesel ölçekte belirlenen faktörler kullanılarak suçların açıklanamayacağı tespit edilmiştir. Bunun yanında analiz içerisine nüfus yoğunluğu, eğitim tesislerinden olan mesafe ile dini ve eğitim tesisi sayıları dahil edilerek yerel ölçekte suçların %50 oranında açıklanabileceği saptanmıştır. Yerel ölçekte hesaplanan analiz sonuçlarının düşük olması uygulanan mekansal analizlerde hem yerel hem bölgesel ölçeklerde kullanılabilecek verilerin yer almasından kaynaklanmaktadır.

Akşam ve gece gerçekleşen suçlara neden olan faktörlere ilişkin araştırma sonuçları Tablo 5'te özetlenmiştir.

Buna göre; bölgesel ölçekte akşam gerçekleşen suçların %80 oranında mekansal faktörlerden ticaret, banka, dini tesis sayıları ile yaş ve nüfus yoğunluğundan kaynaklandığı belirlenmiştir. Gece gerçekleşen suçların ise %91 oranında nüfus yoğunluğu, yaş grubu, toplam donatı sayısı, güvenlik sayısı ve rüzgar hızı bilgileri kullanılarak açıklanabileceği saptanmıştır. Yerel ölçekte akşam gerçekleşen suçların % 65 oranında nüfus yoğunluğu, eğitim tesislerine ve toplu ulaşım duraklarına olan mesafe ile eğitim tesisi ve toplu ulaşım durak sayıları faktörleri ile açıklanabileceği tespit edilmiştir. Gece gerçekleşen suçların ise %55 oranında 18-35 yaş arası erkek nüfusu, ticaret ve güvenlik birimlerine olan mesafe ile toplu ulaşım durak sayıları faktörleri ile açıklanabileceği tespit edilmiştir.

Tablo 5. Bağımlı-Bağımsız değişken ilişkisi

Değişken Grubu	Değişken Adı	Bölgesel Ölçek			Yerel Ölçek		
		sabah	akşam	gece	sabah	akşam	gece
Demografik	Nüfus yoğunluğu	X	X	X	X	X	
Kültürel	Eğitim durumu	X					
Sosyal	Yaş grubu	X	X	X			X
	Göç	X					
Zaman-mekan	Ortalama sıcaklık						
	Ort. rüzgar hızı (km)			X			
Mekansal	Donatı sayısı			X			
	Durak sayısı	X				X	X
	Dini tesis sayısı		X		X		
	Ticaret sayısı		X				
	Eğitim tesisi sayısı	X			X	X	
	Güvenlik birimi sayısı			X			
	Banka sayısı		X				
	Donatıya olan mesafe						
	Durağa olan mesafe	X				X	
	Dini tesise olan mesafe						
	Ticarete olan mesafe						X
	Eğitim tesisine olan mesafe				X	X	
	Güvenlik birimine olan mesafe						X
	Bankaya olan mesafe						

3. BULGULAR

Aynı lokasyona ilişkin iki farklı çalışma ölçeğinde gerçekleştirilen mekansal istatistiksel analiz sonuçlarını karşılaştırabilmek için her iki çalışma ölçeğinde de kullanılabilecek özelliğe sahip verilerin farklı veri kaynaklarından elde edilmesi gerekmektedir. Bu kapsamda elde edilen sosyal, kültürel, demografik, zaman-mekan ve fiziksel çevreye ilişkin verilerin büyük çoğunluğu kurumlardan tablo formatında temin edilmiştir. İzmir Büyükşehir Belediyesi'nden elde edilen koordinatlı ilçe sınırları verileri ile öznitelik verileri içerisinde koordinat bilgileri bulunan hırsızlık verilerinden yola çıkılarak tüm veriler CBS teknolojileri vasıtasıyla koordinatlandırılmıştır. Koordinatlandırma işlemi tamamlanan veriler arasındaki mekansal ilişkileri ölçebilmek için tüm veriler mekansal bir veri tabanına aktarılmıştır.

Mekansal olarak birbirlerine yakın lokasyonda bulunan nesnelerin birbirinden uzak olanlara göre daha çok benzerlik gösterdiği bilindiğinden veriler arasındaki etkileşim incelenirken mekansal istatistik yöntemlerinden yararlanılmıştır. Doğrusal bir regresyon denklemi ile birlikte tüm değişkenler hakkında genel bir model sunan en küçük kareler yöntemi uygulanarak coğrafi olarak ağırlıklandırılmış regresyon analizinde pozitif ya da negatif etki yaratan değişkenler tespit edilmiştir. Buradan elde edilen bilgilerle kurulan coğrafi olarak ağırlıklandırılmış regresyon modeli ile bölgesel ve yerel ölçeklerde hırsızlıkların oluşmasına etki eden faktörler belirlenmiştir. Buna göre bölgesel ölçekte hırsızlık suçlarının oluşmasında ağırlıklı olarak demografik ve sosyo-kültürel özelliklerin etkisi olduğu ancak bunların mekansal faktörlerle desteklendiği görülmektedir. Yerel ölçekte ise bölgesel ölçeğe kıyasla mekansal ve zaman-mekan

faaliyetlerinin suçun oluşmasında sosyo-kültürel faaliyetlere göre daha etkin olduğu sonucuna ulaşılmıştır.

4. TARTIŞMA ve SONUÇ

Suç oluşumuna neden olan faktörlerin araştırıldığı çalışmalarda araştırma bulgularının birbirinden farklı sonuçlar üretmesi bu tür çalışmaların gerçekleştirildiği alana özgü sonuçlar verdiği şeklinde değerlendirilmektedir. Araştırmacılar kendi çalışma alanlarında suçun oluşmasında etkili olduğunu tespit ettikleri faktörlerin bir başka çalışma alanı için anlamlı sonuç vermemesini de bu duruma bağlamaktadırlar. Ancak bu değerlendirme yapılırken gerçekleştirilen çalışmalarda kullanılan alansal büyüklük ve ölçek farklılıklarının gözardı edildiği görülmektedir. Gerçekleştirilen araştırma ile aynı çalışma alanında iki farklı çalışma ölçeği kullanılarak hırsızlık suçuna etki eden faktörler karşılaştırılmıştır. Bölgesel ve yerel ölçek olarak tespit edilen çalışma ölçeklerinde gerçekleştirilen araştırma sonucuna göre aynı çalışma alanı sınırı kullanılmasına karşın hırsızlık suçunun oluşmasını etkileyen faktörler farklılık göstermiştir. Dolayısıyla suç oluşumunu anlama çalışmalarında farklı lokasyonların ve farklı alansal büyüklüklerin kullanılmasının araştırma sonuçlarına etki ettiği görüşüne ek olarak çalışmada kullanılan ölçeğe göre de değişkenlik gösterdiği tespit edilmiştir.

Aynı zamanda araştırmacılar suç oluşumunu etkileyen faktörleri belirlerken genellikle tek bir faktör grubunu (sosyal çevre, fiziksel çevre, vb.) kullanarak çalışmalarını gerçekleştirmektedir. Oysa ki gerçekleştirilen çalışma ile görülmektedir ki bölgesel ölçekte meydana gelen suçlarda sosyal çevrenin yanında fiziksel çevre faktörlerinin de katkısı bulunmatadır. Aynı şekilde yerel ölçekte de fiziksel çevrenin yanında sosyal çevrenin katkısı tespit edilmiştir. Burada önemli olan araştırmanın gerçekleştirileceği çalışma ölçeğine uygun olarak bu verilerin elde edilmesidir.

Sonuç olarak; suç önleme çalışmalarının yürütülmesinde önemli rol oynayan araştırma sonuçlarının emniyet teşkilatlarındaki üst düzey yöneticiler tarafından alınacak mekansal kararları destekleyecek geçerli bir sistem oluşturulmuştur.

NOT: Çalışma Dokuz Eylül Üniversitesi Fen Bilimleri Enstitüsü Coğrafi Bilgi Sistemleri Anabilim Dalı doktora tez çalışmasından üretilmiştir. Çalışma esnasında yürütülen tüm işlemler ArcGIS 9.3 yazılımı ile gerçekleştirilmiştir. Grid işlemi için ArcGIS 9.x versiyonları üzerinde çalışabilen Hawth's Tools (v3.27) kullanılmıştır.

5. KAYNAKLAR

Ackerman, V. W., 1998. Socioeconomic correlates of increasing crime rates in smaller communities. Professional geographer, 50, 372-387.

Aksoy, H., 2004. Coğrafi profilleme, 3. Coğrafi bilgi sistemleri bilişim günleri. Fatih Üniversitesi, İstanbul.

Anderson, C. A., Anderson, D. C., 1984. Ambient temperature and violent crime: Tests of the linear and curvilinear hypotheses. Journal of personality and social psychology, 46(1), 91-97.

Ayhan, İ., 2007. Kentte suç oranlarının ekonomik sosyal ve mekansal değişkenlerle modellenmesi. Yüksek Lisans tezi, Dokuz Eylül Üniversitesi, İzmir.

Brunsdon, C., McClatchey, J., Unwin, D. J., 2001. Spatial variations in the average rainfall-altitude relationship in Great Britain: An approach using geographically weighted regression. International journal of climatology, 21, 455-466.

Cahill, F. M., Mulligan, F. G., 2003. The determinants of crime in Tucson, Arizona. Urban geography, 24, 582-610.

Ceccato, V., Haining, R., Signoretta, P., 2002. Exploring offence statistics in Stockholm city using spatial analysis tools. Annals of the association of American geographers, 92, 29-51.

Cohn, E. G., 1990. Weather and crime. British journal of criminology, 30(1), 51-64.

Cozens, P. M., 2002. Sustainable urban development and crime prevention through environmental design for the British city. Towards an effective urban environmentalism for the 21st century-cities, 19, 129-137.

Chainey, S., Ratcliffe, J., 2005. GIS and Crime Mapping. Wiley Press, Chichester.

Ergün, N., Yirmibeşoğlu, F., 2005. İstanbul'da 2000-2004 yılları arasında suçun mekansal dağılımı. Planlamada yeni politika ve stratejiler riskler ve fırsatlar: 8 Kasım dünya şehircilik günü 29. kolokyumu, İstanbul.

Field, S., 1992. The effect of temperature on crime. British journal of criminology, 32(3), 340-351.

Fotheringham, A. S., Brunsdon, C., Charlton, M., 2002. Geographically Weighted Regression: The Analysis of Spatially Varying Relationships. Wiley Press, Chichester.

Gillespie, W. T., Agnew, A. J., Mariano, E., Mossler, S., Jones, N., Braughton, M., Gonzalez, J., 2009. Finding Osama bin Laden. MIT international review, 1-17.

Gorr, W. L., Olligschlaeger, A. M., 1994. Weighted spatial adaptive filtering: Monte Carlo studies and application to illicit drug market modeling. Geographical analysis, 26, 67-87.

Gruenewald, P. J., Freisthler, B., Remer, L., LaScala, E. A., Treno, A., 2006. Ecological models of alcohol outlets and violent assaults: Crime potentials and geospatial analysis. Addiction, 101, 666-677.

Malczewski, J., Poetz, A., 2005. Residential burglaries and neighborhood socioeconomic context in Londo, Ontario: Global and local regression analysis. The professional geographer, 57(4), 516-529.

MGM (Türkiye Meteoroloji Genel Müdürlüğü), 2012. Telefonla görüşme, 16 Nisan 2012.

Murray, T. A., McGuffog, I., Western, S. J., Mullins, P., 2001. Exploratory spatial data analysis techniques for examining urban crime. British journal of criminology, 41, 309-329.

Newman, O., 1972. Defensible Space: Crime Prevention Through Urban Design. Macmillan, New York.

Olligschlaeger, A. M., 1997. Spatial analysis of crime using GIS-based data: Weighted spatial adaptive filtering and chaotic cellular forecasting with applications to street level drug markets. Doktora tezi, Carnegie Mellon Üniversitesi, Pittsburg.

Openshaw, S., 1984. The Modifiable Areal Unit Problem. Geo Books, Norwich.

Salleh, S. A., Mansor, N. S., Yusoff, Z., Nasir, R.A., 2012. The crime ecology: Ambient temperature vs. spatial setting of crime (burglary). Social and behavioral sciences, 42, 212-222.

Schmid, C. F., 1960. Urban crime areas: Part II. American Sociological Review, 25, 655-78.

Shepard, D., 1968. A two-dimensional interpolation function for irregularly-spaced data. Proceedings of the 1968 ACM National Conference, 517-524.

Strano, M., 2004. A neural network applied to criminal psychological profiling: an Italian initiative. International journal of offender therapy and comparative criminology, 48(4), 1-7 .

TÜİK (Türkiye İstatistik Kurumu), 2008. Ceza İnfaz Kurumuna Giren Hükümlü İstatistikleri. http://tuikapp.tuik.gov.tr/girenhukumluapp/girenhukumlu.zul, 18 Haziran 2013.

Ural, S., Hussain, E., Shan, J., 2011. Building population mapping with aerial imagery and GIS data. International journal of applied earth observation and geoinformation, 13, 841-852.

Yavuz, Ö., Tecim, V., 2011. CBS tabanlı suç önleme çalışmalarında yapay sinir ağları kullanılarak mekansal karar sistemi oluşturulması. Suç önleme sempozyumu, Merinos Atatürk Kültür ve Kongre Merkezi, Bursa.

Yavuz, Ö., Tecim, V., 2013. Exploring scale effect using geographically weighted regression on mass dataset of urban robbery. International archieves of the photogrammetry, remote sensing and spatial information sciences, XL-4/W1, 147-154.

COMPARISON OF VARIABLES AFFECTING ON CRIME OCCURANCES AT DIFFERENT SCALES BY USING A SPATIAL STATISTICS METHOD

ABSTRACT

Many studies have been focused on understanding the occurance of the crimes since 20th century. Following years, the geographical location have been identified as having a very important contribution in understanding the crime. Thus, researchers have also included the spatial factors in their working field while determining the influencing factors on crime. Studies have concluded that these factors also vary according to the location of the study in different regions. However, the factors affecting the crime are changed not only by the location but they also changed according to the scale of the research of the same study area. In this study, the factors influencing crime are identified by changing the scale of the study area that covers the central district of Izmir. The study established the differences of the identified variables which affected on robbery events at the comparison of regional and local scales of the same study area by using spatial statistical methods. The results of the research has current findings to support decisions on crime prevention studies.

Keywords: Crime prevention, Decision support, Geographical Information Systems, Spatial analysis, Spatial statistics.

Volume: 10 Number: 03 Category: 01 Page: 15-34 ISSN No: 1303-6319 Cilt: 10 Sayı: 03 Kategori: 01 Sayfa: 15-34

PERFORMANCE COMPARISONS OF MODEL SELECTION CRITERIA: AIC, BIC, ICOMP AND WOLD'S FOR PLSR

Özlem GÜRÜNLÜ ALMA*

ABSTRACT

Partial least squares regression (PLSR) is a statistical method of modeling relationships between $\mathbf{Y}_{N\times M}$ *response variable and* $\mathbf{X}_{N\times K}$ *explanatory variables which is particularly well suited to analyzing when explanatory variables are highly correlated. In partial least square part, some model selection criteria are used to obtain the latent variables which are the most relevant variables describing the response variables. In typical approach to select the numbers of latent variables are Akaike information criterion (AIC) and Wold's R criterion.*

In this study, we are interested in the performance of Bayesian Information Criterion (BIC) and Information Complexity Criterion (ICOMP) criteria besides the traditional methods AIC and Wold's R criteria as the model selection criteria for partial least squares regression when the number of observations are higher than predictor variables. Performances of AIC, BIC, ICOMP and Wold's R criteria were compared by real life data and simulation study. Simulation results were obtained from different sample sizes, different number of predictor variables and different number of response variables. The simulation results demonstrate that the BIC and ICOMP model selection methods are more effective than AIC and Wold's R criteria selecting of latent variables for known PLSR models.

Keywords: AIC, BIC and ICOMP information criteria, K-fold cross-validation, Model selection, Partial least squares regression, Wold's R criterion.

1. INTRODUCTION

The partial least squares regression is a generalization of multiple linear regression analysis. It was developed by Herman Wold (1966) as an econometric technique but became popular as a tool to analyze data from chemical applications. PLSR is also used in multivariate statistical data analysis (Geladi and Kowalski, 1986; Wold, 1982). It is useful when the predictor variables are highly correlated and/or the number of dependent variables is greater than or equal to the number of observations (Wold, 1982). This success has led to the development of extensions methods of PLSR with objectives other than simple multivariate linear regression. A statistical overview of PLSR can be found in Geladi and Kowalski (1986), Wold et al. (2001), and Abdi and Salkind (2007).

The PLSR's goal is to predict or analyze a set of response variables from a set of independent variables or predictors. This prediction is achieved by extracting from the predictors a set of orthogonal factors called latentvariables which have the best predictive power. Associations are established with latent factors extracted from predictor variables that maximize the explained variance in the response variables. These latent factors are defined as linear combinations constructed between predictor

*Assistant Professor Dr., Department of Statistics, Faculty of Sciences, Muğla Sıtkı Koçman University, 48000 Muğla, Turkey, e-mail: ozlem.gurunlu@gmail.com

and response variables, such that the original multidimensionality is reduced to a lower number of orthogonal factors to detect the structure in the relationships between predictor variables and between these latent factors and the response variables (Abdi and Salkind, 2007; Helland, 1990; Wold, 1982). Although as many latent variables as min(N,K) can be calculated, where N is the sample size and K is number of explanatory variables, it is conjectured that the lower order latent variables are associated with process noise and should be excluded from the model. Therefore to remove the noise, a criterion is required for selecting the number of latent variables to include in the PLSR model (Li et al., 2002).

Various approacheshave been proposed in the literature for model order selection methods, including Final Prediction Error criterion (FPE), Multiple Correlation Coefficient (R^2), Adjusted Multiple Correlation Coefficient (R_a^2), Normalized Residuals Sum of Squares (NRSS), Mallow's Statistics (C_p), Predicted Error of Sum of Squares (PRESS), Wold's R criterion, Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC) and Information Complexity Criterion (ICOMP). A review of these model order selection criteria can be found in Haber and Unbenhauen (1990); Bozdoğan (2000); Li et al. (2002); Clark and Troskie (2006). To evaluate the performance of the different criteria, simulated models allow the underlying structures of the models to be known (Bedrick and Tsai, 1994; Eastment and Krzanowski, 1982). Practical case studies as described in Bozdogan (2000), Myung (2000), Li et al. (2002), and Clark and Troskie (2006).

The PLSR creates latent variables for both explanatory and response variables using different algorithms. As well as the standard NIPALS, SIMPLS and Kernel algorithms, many different algorithms have been proposed to compute PLSR parameters such as IVS-PLS, PoLiSh, UVE-PLS, GA-PLS, and etc (Jouan-Rimbaud Bouveresse and Rutledge, 2009).

In this paper, the subset of latent variables that best fit the data is sequentially determined. Firstly, the latent variables are extracted using partial least squares algorithm, secondly, the number of latent variables can be consistently estimated using information criterion. The performance of information criterion is considered with the generation of experimental data. We have shown the behaviour of AIC, BIC, ICOMP and Wold's R criteria for different sample sizes and different dimension of PLSR models by simulation study.

The article is organized as follows. Section 2 includes the PLSR algorithm and describes how to obtain the latent variables. Section 3 gives summary information about the model selection criteria which are AIC, BIC, ICOMP and Wold's R criteria. In Section 4, real life data and simulation models are described, and the simulation results are given. This section focuses on the empirical results which showthe performance of information criteria for various configurations of data sets. Finally a summary of simulation results and conclusions are given in section 5.

2. PARTIAL LEAST SQUARES REGRESSION MODEL

The objective of all linear PLSR algorithm is to project the data down onto a number of latent variables $(t_a \text{ and } u_a)$, and then to develop a regression model between latent

variables. It uses both the variation of **X** and **Y** to construct latent variables. The intension of PLSR is to form components that capture most of the information in the **X** variables, which is useful for predicting response variables, while reducing the dimensionality of the regression problem by using fewer components than the number of **X** variables (Garthwaite, 1994).

$\mathbf{X}_{N\times K}$ represents the data matrix of N observation units on K explanatory variables and $\mathbf{Y}_{N\times M}$ the data matrix of N observation units on M response variables. t_a and u_a $(a = 1,...,A)$ are latent variables, where A is the number of the latent variables, and then a regression model between latent variables is written as follows:

$$\mathbf{u}_a = b_a \mathbf{t}_a + e_a, \qquad a= 1,\ldots, A \tag{1}$$

where e_a is vector of errors and b_a is an unknown parameter estimated by $\hat{\mathbf{b}}_a=(\mathbf{t}_a'\mathbf{t}_a)^{-1}\mathbf{t}_a'\mathbf{u}_a$. The latent variables are computed by $t_a = X_a w_a$ and $u_a = Y_a q_a$, where both w_a and q_a have unit length and are determined by maximizing the covariance between t_a and u_a.

$X_{a+1} = X_a - t_a p_a'$ where $X_1 = X$ and $p_a = X_a' t_a /(t_a' t_a)$ and $Y_{a+1} = Y_a - b_a t_a q_a'$ where $Y_1 = Y$. Letting $\hat{u}_a = \hat{b}_a t_a$ be prediction of u_a, the matrices **X** and **Y** can be decomposed as the following (Li et al., 2002):

$$\mathbf{X}=\sum_{a=1}^{A} \mathbf{t}_a \mathbf{p}_a' + \mathbf{E}, \quad \text{and} \quad \mathbf{Y}=\sum_{a=1}^{A} \hat{\mathbf{u}}_a \mathbf{q}_a' + \mathbf{F}, \tag{2}$$

where E and F are the residuals of **X** and **Y** after extracting the first "a" pairs of latent variables.

3. ESTIMATING NUMBER OF LATENT VARIABLES USING INFORMATION CRITERION

The problem of estimating the true error of hypothesis using different adjustable parameters in order to choose the best one is known as model selection (Hastie et al., 2001). The necessity of introducing the concept of model evaluation has been recognized as one of the important technical areas, and the problem is posed on the choice of the best approximating model among a class of competing models by a suitable model evaluation criterion given a data set. Model evaluation criteria are defined as figures of merit, or performance measures, for competing models (Bozdoğan, 2000). In this section a number of criteria for PLSR model selection can be briefly summarized for multivariate regression models. In PLSR model, the information criteria used to find the number of latent variables and $\mathbf{T}_a, \mathbf{Y}, M, V(a, T_a), a$ was used instead of $\mathbf{X}, \mathbf{Y}, M, \Sigma, K$. Let T_a be a matrix of latent variables, and $V(a, T_a)$ is the sum of squared residuals.

$$V(a, T_a) = \min_{T_a} \frac{1}{MN} \sum_{k=1}^{M} \sum_{i=1}^{N} (Y_{ik} - \hat{Y}_{ik})^2 \tag{3}$$

$$V(a, T_a) = \min_{T_a} \frac{1}{MN} \sum_{k=1}^{M} \sum_{i=1}^{N} (Y_{ik} - T_{ia} b_{ak})^2 \quad (4)$$

Selection of the number of latent variables to build a representative model is an important issue in PLSR. The main goal of model selection is to approximate the true model using candidate models and then retain the model that entails a minimum loss of information. A metric frequently used by chemometricians for the determination of the number of latent variables is that of Wold's R criterion, whilst more recently a number of statisticians have advocated the use of AIC (Li et al., 2002). Generally a good model has small residuals and few parameters, and then it is preferred, chosen with the smallest value of information criterion. However, it is well known that different information theoretic criteria with proper choice of penalty function can be used to choose the correct model (Kundu and Murali, 1996). Bedrick and Tsai (1994) modified the AIC criterion which is corrected version of the multivariate AIC for the small sample case (Bedrick and Tsai, 1994).

AIC and BIC are the two penalized criteria that are based on two different model selection approaches. AIC is aimed at finding the best approximating model to the unknown data generating process whilst BIC is designed to identify the true model. AIC does not depend directly on sample size. Bozdoğan (1987) noted that because of this, AIC lacks certain properties of asymptotic consistency. Although BIC takes a similar form like AIC, it is derived within a Bayesian framework, reflects sample size and have properties of asymptotic consistency. For reasonable sample sizes, BIC apply a larger penalty than AIC, thus other factors being equal it tend to select simpler models than does AIC. From a Bayesian view point this motivates the adoption of the Bayesian information criteria. AIC and BIC have been compared theoretically and empirically (Kuha, 2004; Weakliem, 2004) and examined empirically with respect to theselection of stock-recruitment relationships (Wang and Liu, 2006; Henry de-Graf, 2010). Although, AIC, BIC, and Bozdogan information criteria compared theoretically and empirically in many areas, there has been no empirical comparison for their relative performance in PLSR modeling context.

3.1 The Akaike Information Criterion

The Akaike information criterion was developed by Akaike (Akaike, 1974). AIC has played a significant role in solving problems in a wide variety of fields for analyzing actual data. The AIC is defined as,

$$AIC = -2\log L(\hat{\theta}) + 2K, \quad (5)$$

where $\hat{\theta}$ is the maximum likelihood estimator of the parameter θ for an approximating statistical model Y, $L(\hat{\theta})$ is the maximized likelihood function, and K is the number of free parameters in $\mathbf{Y}$. The multivariate version of AIC was given by Bedrick and Tsai (1994),

$$MAIC = N(\log|\hat{\Sigma}| + M) + 2d[MK + M(M+1)/2], \quad (6)$$

where $d = N/[N-(K+M+1)]$ and $\hat{\Sigma}$ is the maximum likelihood estimator of Σ. This is the corrected version of the multivariate AIC for the small sample case. When the sample size is large, d value can be equal to one, thus equation (6) may be further simplified. Also, Bozdoğan (2000) derived a score information theoretic criteria under the multivariate normal assumption for the multivariate regression model which are given as follows,

$$AIC = NM\log(2\pi) + N\log\left|\hat{\Sigma}\right| + NM + 2\left[MK + M(M+1)/2\right] \tag{7}$$

Since 1974, AIC has been modified in many ways. For example, many model selection criteria including CAIC, CAICF (Bozdogan, 1987), GAC (Torr, 1998), GAIC (Kanatani, 2002) and MAIC (Boyer et al., 1994) are derived from AIC.

3.2 The Bayesian Information Criterion

The Bayesian Information Criterion is an information criterion based on Bayesian method proposed by Schwarz (1978), has recently been applied to the selection of models. BIC is shown as,

$$BIC = -2\log L(\hat{\theta}) + K\log(N), \tag{8}$$

where $\hat{\theta}$ is the maximum likelihood estimator of the parameter θ for an approximating model M, $L(\hat{\theta})$ is the maximized likelihood function, and K is the number of the estimated parameters. The multivariate version of BIC was given by Bedrick and Tsai (1994). It is shown as follows,

$$BIC = N\log\left|\hat{\Sigma}\right| + \left[\frac{MK + M(M+1)}{2}\right]\log(N). \tag{9}$$

BIC favourizes more parsimonious models than AIC due to its penalization. AIC, but not BIC, is biased in the following sense: if the true model belongs to the family M_i, the probability that AIC chooses the true model does not tend to one when the number of observations goes to infinity. AIC and BIC have similar formulas but originates from different theories and there is no rationale to use simultaneously AIC and BIC: AIC is an approximation of the Kullback-Leibler divergence between the true model and the estimated one, while BIC comes from a bayesian choice based on the maximisation of the posterior probability of the model, given the data (Saporta, 2008).

3.3 The Information Complexity Criterion

The development of ICOMP has been motivated partly by AIC, and partly by information complexity concepts and indices. In contrast to AIC, the new procedure ICOMP is based on the structural complexity of an element or set of random vectors via a generalization of the information based covariance complexity index. ICOMP inverse Fisher information matrix (ICOMP(IFIM)) is shown as for multiple regression (Bozdoğan, 2000; Bozdoğan, 2004),

$$\text{ICOMP(IFIM)} = N\log(2\pi) + N\log(\hat{\sigma}^2) + N + C_1(\hat{F}^{-1}(\hat{\theta})), \tag{10}$$

where $\hat{\sigma}^2$ is the estimated variance of regression model, and K explanatory variables in regression model. Bozdoğan (2000) introduced ICOMP (IFIM) information theoretic criterion for the multivariate regression model, and it is also used when there is multicollinearity in regression model. It is shown as follows,

$$\text{ICOMP} = NM\log(2\pi) + N\log\left|\hat{\Sigma}\right| + NM + 2C_1(\hat{F}^{-1}(\hat{\theta})). \tag{11}$$

The complexity measure $C_1(\hat{F}^{-1}(\hat{\theta}))$ is given by,

$$C_1(\hat{F}^{-1}(\hat{\theta})) = \frac{M(M+K)}{2}\log\left[\frac{\text{tr}(\hat{\Sigma})\text{tr}(X'X)^{-1} + \frac{1}{2N}\left[\text{tr}(\hat{\Sigma}^2) + (\text{tr}\hat{\Sigma})^2 + 2\sum_j \hat{\sigma}_{jj}\right]}{M(M+K)}\right] - \frac{1}{2}(M+K+1)\log\left|\hat{\Sigma}\right| - \frac{M}{2}\log\left|(X'X)^{-1}\right| - \frac{M}{2}\log(2). \tag{12}$$

3.4 Wold's R Criterion

Wold's R criterion is based on cross validation which can be calculated from the Predicted Residual Sum of Squares (PRESS) values, and it can be explained as follows:

$$R = \frac{\text{PRESS}(a+1)}{\text{PRESS}(a)}, \tag{13}$$

where PRESS(a) denotes the PRESS after including the first a latent variables. Wold's R criterion terminates when R is greater than unity or a given threshold and hence A=a (Li et al., 2002). PRESS is a measure of how well the use of the fitted values for a subset model can predict the observed responses of a dependent variable, and its value for the i^{th} observation is calculated as follows:

$$\text{PRESS} = \sum_{i=1}^{n}\left(y_i - \hat{y}_{i(i)}\right)^2, \tag{14}$$

where the notation $\hat{y}_{i(i)}$ is used for the fitted value. By the first subscript i, it is shown that it is a predicted value for the i^{th} case and by the second subscript (i), it is shown that i^{th} case is omitted when the regression function was fitted. The smaller PRESS value shows that it is the best model to predict. In some situations PRESS should reach a minimum and start to rise again. To avoid building a model that is either overfit or underfit, the number of components where the PRESS value reaches a minimum would be the obvious choice for the best model. While the minimum of the PRESS may be the best choice for predicting the particular set of samples, most likely it is not the optimum choice for predicting all unknown samples in the future. That is, the optimum number of factors was determined rather than the selection of the model, which yields a minimum in PRESS; the model selected is the one with the fewest number of factors such that

PRESS for that model is not significantly greater than the minimum PRESS (Niazi and Azizi, 2008). A solution to this problem has been suggested in which the PRESS values for all previous factors are compared to the PRESS value at the minimum.

4. REAL LIFE DATA EXAMPLE, DESIGN OF SIMULATION STUDY AND RESULTS

In this paper, real life data are used and a simulation study is conducted to gain a better understanding of AIC, BIC, ICOMP, and Wold's R criteria performances for PLSR model selection; in fact it is a designed experimental simulation study for choosing thetrue latent variables. The experiment has various characteristics of the simulation models, in order to quantify the expected performance of information criteria. In the next subsection, the steps of data generation and performance results of criteria to PLSR model selection are shown by means of a simulation study. Additionaly, in order to select model number of components to be retained in the final model, k-fold cross validation in kernel PLSR algorithm is used (Kohavi, 1995).

4.1 Real Life Data Example

Performances of AIC, BIC, ICOMP and Wold's R criteria have been tested considering a real life dataset: the Body Fat Measurement. This data set has been used by Bozdoğan (2004) for subset selection of best predictors using Genetic Algorithms. In this data set, it is determined that the best subset predictors of y=Percent body fat from Siria (1956) equation, using k=13 predictors are x_1=Age (years), x_2=Weight(lbs), x_3=Height (inches), x_4=Neck circumference (cm), x_5=Chest circumference(cm), x_6=Abdomen 2 circumference (cm), x_7=Hip circumference (cm), x_8=Thigh circumference (cm), x_9=Knee circumference (cm), x_{10}=Ankle circumference (cm), x_{11}=Biceps (extended) circumference (cm), x_{12} = Forearm circumference (cm), x_{13} =Wrist circumference (cm). The data contain the estimates of the percentage of body fat determined by underwater weighing and various body circumference measurements for n = 252 men. This is a good example to illustrate the versatility and utility of our approach using multiple regression analysis with GA. This data set is maintained by Dr. Roger W. Johnson of the Department of Mathematics & Computer Science at South Dakota School of Mines and Technology[1]. A variety of popular health books suggest that the readers assess their health, at least in part, by estimating their percentage of body fat. In Bailey (1994), for instance, the reader can estimate body fat from tables using their age and various skin-fold measurements obtained by using a caliper. Other texts give predictive equations for body fat using body circumference measurements (e.g. abdominal circumference) and/or skin-fold measurements. See, for instance, Behnke and Wilmore (1974); Wilmore (1976); or Katch and Mc Ardle (1977).Percentage of body fat for an individual can be estimated once body density has been determined. Siria (1956) assumes that the body consists of two components-lean body tissue and fat tissue.

Letting,

D = Body Density (gm/cm^3) $\quad D = 1/[(A/a) + (B/b)]$
A = proportion of lean body tissue
B = proportion of fat tissue (A+B=1) $\quad B = (1/D)*[ab/(a-b)] - [b/(a-b)]$.

[1]E-mail: rwjohnso@silver.sdsmt.edu, and web address: http://silver.sdsmt.edu/~rwjohnso

a = density of lean body tissue (gm/cm^3)
b = density of fat tissue (gm/cm^3)

Using the estimates a=1.10gm/cm^3 and b=0.90 gm/cm^3 (Katch and McArdle, 1977) or Wilmore (1976), we come up with Siri's equation:

Percentage of Body Fat (i.e. 100*B) = 495/D - 450.

Volume, and hence body density, can be accurately measured by a variety of ways. The technique of underwater weighing computes body volume as the difference between body weight measured in air and weight measured during water submersion. In other words, body volume is equal to the loss of weight inwater with the appropriate temperature correction for the water's density (Katch and McArdle, 1977). Using this technique,

Body Density = WA/[(WA-WW)/c.f. - LV]

where, WA=Weight in air (kg), WW=Weight in water (kg), c.f.=Water correction factor (=1 at 39.2 deg F as one-gram of water occupies exactly one cm^3 at this temperature, =.997 at 76-78 deg F), LV=Residual Lung Volume (liters) (Katch and McArdle, 1977). Other methods of determining body volume are given in Behnke and Wilmore (1974).

For this data set, PLSR model is established using Minitab package program tool, and validation technique is selected as k-fold cross validation, k=5. Then, the results of model selection and validation are as follows:

Table 1. Model selection and validation for percent body fat

Number of latentvariables	Relativecumulativevariance of components	Sum of squareerror	R-Square	PRESS	R-Sq (pred)
1	0.59	0.039	0.56	0.041	0.54
2	0.70	0.012	0.86	0.015	0.83
3	0.75	0.004	0.95	0.005	0.93
4	0.81	0.002	0.96	0.003	0.96
5	0.84	0.002	0.97	0.002	0.97
6	0.87	0.002	0.97	0.002	0.97
7		0.002	0.97	0.002	0.97
8		0.002	0.97	0.002	0.97
9		0.002	0.97	0.002	0.97
10		0.002	0.97	0.002	0.97

As seen from the results in Table 1, the number of latent variable is 6, and Figure 1 shows the model selection plot.

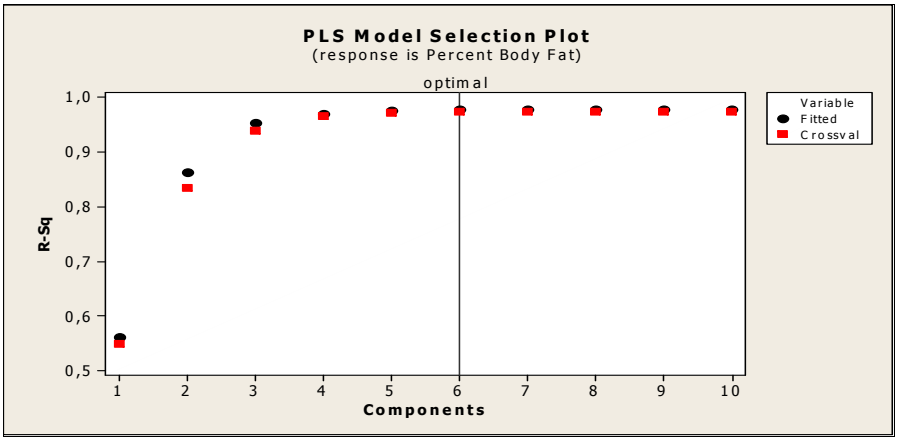


Figure 1. **Partial least squares regression model selection plot**

This data set is used to compare the performance of model selection criteria AIC, BIC, ICOMP and Wold'sR. For this data set, AIC, BIC and ICOMP find the optimal number of components (6) whereas Wold's R finds it as 7.

4.2 Design of Simulation Study

In this subsection, simulation experiments are performed to evaluate the performance comparisons of AIC, BIC, ICOMP, and Wold's R criteria. The framework for the simulation models are based on the study of Naes and Martens (1985), Li and et al. (2002). It is extended in this paper to the situation where there exist multiple response variables and different number of explanatory variables. In the simulation study, the multivariate regression models arefirst developed from which data aregenerated, and then model selection criteria areapplied. The resulting models arethen compared with the true models and finally an evaluation of the different criteria for PLSR model selection ismade through a comparison of the success rate as to which the true model is selected (Bedrick and Tsai, 1994; Li et al., 2002).

The **X** and **Y** block data, with sample size N, aregenerated as:

$$\mathbf{X}=\sum_{i=1}^{A^*}\mathbf{r}_i\boldsymbol{\xi}_i'+\tilde{\mathbf{E}}, \tag{15}$$

$$\mathbf{Y}=\sum_{i=1}^{A^*}\mathbf{z}_i\boldsymbol{\eta}_{A^*i}'+\boldsymbol{\psi}=\sum_{i=1}^{A^*}\mathbf{r}_i\boldsymbol{\eta}_{A^*i}'+\tilde{\mathbf{F}}_{A^*}, \tag{16}$$

where $\tilde{\mathrm{E}}$ and r_i are generated from mutually independent normal variables. It is noted that $\mathrm{var}(\mathrm{r}_i)+\mathrm{var}(\mathrm{e}_j)$ is the largest eigenvalue of $\mathrm{cov}(\mathbf{X})$. $\boldsymbol{\psi}$ is generated from a multivariate normal distribution, $\tilde{\mathbf{F}}$ is a noise matrix, and $\mathbf{Z}$ is constructed as $\mathbf{z}_i=\mathbf{r}_i+\mathbf{f}_i$, $\mathbf{f}_i$ are generated as independent normal variables. $\{\boldsymbol{\xi}_i\}$ and $\{\boldsymbol{\eta}_{A^*i}\}$ are normalized orthogonal vector series, and $\mathbf{r}_i$ are mutually independent random variables.

Comparing equation (2), with equations (15) and (16), it can be calculated that latent variable $\mathbf{t}_i$, loading vectors $\mathbf{p}_i$ and $\mathbf{q}_i$ obtained from PLSR algorithm are approximately

equal to $\mathbf{r}_i$, $\{\xi_i\}$ and $\{\eta_{A^*i}\}$ $i=(1,...,A^*)$, respectively. The Y-block data, $\mathbf{Y}$ of the response variables then essentially depends on $\mathbf{r}_i$, $i=(1,...,A^*)$, plus noise. This means that the theoretical value of the number of latent variable is equal to A^* (Li et. al. 2002).

To carry out simulations run, it is proceeded on different simulation models and a fixed number of blocks for k-fold cross-validation in kernel algorithm; k is selected as 5. The dimensions of explanatory variables are extended as $N\times5$, $N\times8$, $N\times10$, $N\times15$, and $N\times20$. The dimensions of response variables matrix, $\mathbf{Y}$, are chosen as $N\times3$, $N\times4$, $N\times5$, and sample sizes are selected as N=50, 100, 250, 500, 1000. For each of the combinations of parameters in Table 2, 10000 data sets are generated taking into account the dimension of partial least squares regression models and sample sizes, so that 25×10000 data sets are generated.

Table 2. The elements of experimental data sets

Number of latent variables	The dimension of response variables matrix	The dimensions of explanatory variables	Sample sizes
5*3[1]	$Y_{N\times3}$	$X_{N\times5}$	N={50, 100, 250, 500, 1000}
8*4	$Y_{N\times4}$	$X_{N\times8}$	N={50, 100, 250, 500, 1000}
10*4	$Y_{N\times4}$	$X_{N\times10}$	N={50, 100, 250, 500, 1000}
15*5	$Y_{N\times5}$	$X_{N\times15}$	N={50, 100, 250, 500, 1000}
20*5	$Y_{N\times5}$	$X_{N\times20}$	N={50, 100, 250, 500, 1000}

[1] 5*3 shows that the number of predictor variables is 5, and these variables are reduced to number 3 for the number of latent variables.

Then these data sets are applied to AIC, BIC, ICOMP, and Wold's R criteria. Explanatory data matrix, **X**, isgenerated from equation (15), and **Y**, isgenerated from equation (16). Generation of **X** and **Y** data matrices are just explained for 5*3 which is shownin Table 3, and the other data matrices are given in Appendix. The components of **X** and **Y** datamatrices are given in Table 3 ($i=[1,...,A^*]$, $A^*=3$).

Table 3. The components values of X and Y matrices for 5*3

$\tilde{\mathbf{E}}$	$\tilde{\mathbf{E}}=[e_1,...,e_5]\sim N(0,0.02)$
$\mathbf{r}_i$	$r_1\sim N(0,15)$, $r_2\sim N(0,7.5)$, $r_3\sim N(0,3)$
ξ_i	$\xi_1=[0.6247 \quad 0.5635 \quad 0.4472 \quad 0.2871 \quad 0.0989]'$
	$\xi_2=[0.5635 \quad 0.0989 \quad -0.4472 \quad -0.6247 \quad -0.2871]'$
	$\xi_3=[0.4472 \quad -0.4472 \quad -0.4472 \quad 0.4472 \quad 0.4472]'$
$\mathbf{f}_i$	$f_1\sim N(0,0.5)$, $f_2\sim N(0,0.25)$, $f_3\sim N(0,0.1)$
η_{A^*i}	$\eta_{31}=[0.7887 \quad 0.5774 \quad 0.2113]'$
	$\eta_{32}=[0.5774 \quad -0.5774 \quad -0.5774]'$
	$\eta_{33}=[0.2113 \quad -0.5774 \quad 0.7887]'$
ψ	$\psi=\begin{pmatrix}0.00010 & 0.00006 & 0.00006\\ 0.00006 & 0.00010 & 0.00006\\ 0.00006 & 0.00006 & 0.00010\end{pmatrix}$

The variance inflation factor (VIF) values for 5*3 design matrix shows that there ismulticollinearity (Table 4).The VIF values are calculated by Minitab package program. Table 5 shows the relative cumulative variances by the five latent variables for the X and Y blocks, averaged over 10000 simulation experiments. It can be seen from the first two rows of Table 5 that on average, for $A^* = 3$, first three latent variables capture 100% and 98% of the variances in the **X** and **Y** data sets, respectively. This verifies the theoretical value of the number of latent variables $A^* = 3$.

Table 4. The VIF values for 5*3. (N=100, k=5)

	X_1	X_2	X_3	X_4	X_5
VIF	67.6	238.0	142.0	72.3	42.8

Table 5. Relative cumulative variances of X and Y for $5*3$. (N=100, k=5)

True Model	Blocks	Number of Latent Variables				
		1	2	3	4	5
A*=3	X-block	0.55	0.89	1.00	1.00	1.00
	Y-block	0.71	0.84	0.98	0.98	0.98

4.3 Results of Simulation Study and Performance Comparison of Model Selection Criteria

In this study, we compare the performance of model selection criteria by using the percentage of success which shows the precision in finding the number of latent variables by model selection criteria. We also compute performances of all criteria for Li et. al. (2002) data when dimension to reduction PLSR model is 6*4, and the results are shown in Table 6. As seen from the results, AIC, BIC, ICOMP methods provide the best selection of the number of latent variables.

Table 6. Comparison of the percentages of the selected number of latent variables for 6*4

N	AIC	BIC	ICOMP	MAIC	Wold's R
100	100	100	100	84.0*	47.6*
1000	100	100	100	75.8*	49.0*

* MAIC and Wold's R results are taken from Li et al. study (2002).

All results of the simulations for various sample sizes and dimensions are given in Table 7 and these are obtained by 10000 replications. It illustrates the ability of AIC, BIC, ICOMP, and Wold's R criteria in selecting latent variables for all situations.

Table 7. Percentages of performance comparison for each criterion

	AIC					BIC					ICOMP					Wold's R				
N	50	100	250	500	1000	50	100	250	500	1000	50	100	250	500	1000	50	100	250	500	1000
5*3	100	100	100	99	100	100	100	100	99	100	100	100	100	99	100	60	58	53	50	49
8*4	100	100	100	100	100	100	100	100	100	100	99	100	100	100	100	29	20	11	23	13
10*4	59	88	87	73	51	57	88	91	81	64	61	88	91	81	67	47	41	54	43	32
15*5	45	35	24	24	23	37	35	27	34	27	50	35	26	30	26	47	45	47	43	41
20*5	63	59	58	55	43	59	59	64	63	58	67	62	64	63	58	46	47	45	43	42

As it can be seen from Table 7, for all experimental data sets, almost AIC, BIC, and ICOMP criteria have similar performances except Wold's R criterion. It has the lowest success rate compared to other criteria. For $(p*a) \leq (8*4)$, AIC, BIC, and ICOMP criteria perform better than the Wold's R criterion. When $(p*a) > (8*4)$ the success rates of AIC, BIC, ICOMP and Wold's R criteria decrease as the sample sizes and the dimensions of models increase. Overall ICOMP criterion provides the best selection ofthe number of latent variables among AIC, BIC, and Wold's R criteria.

These comparisons of performances are graphically presented in Figure 2(a)-(e) and Figure 3(a)-(f).

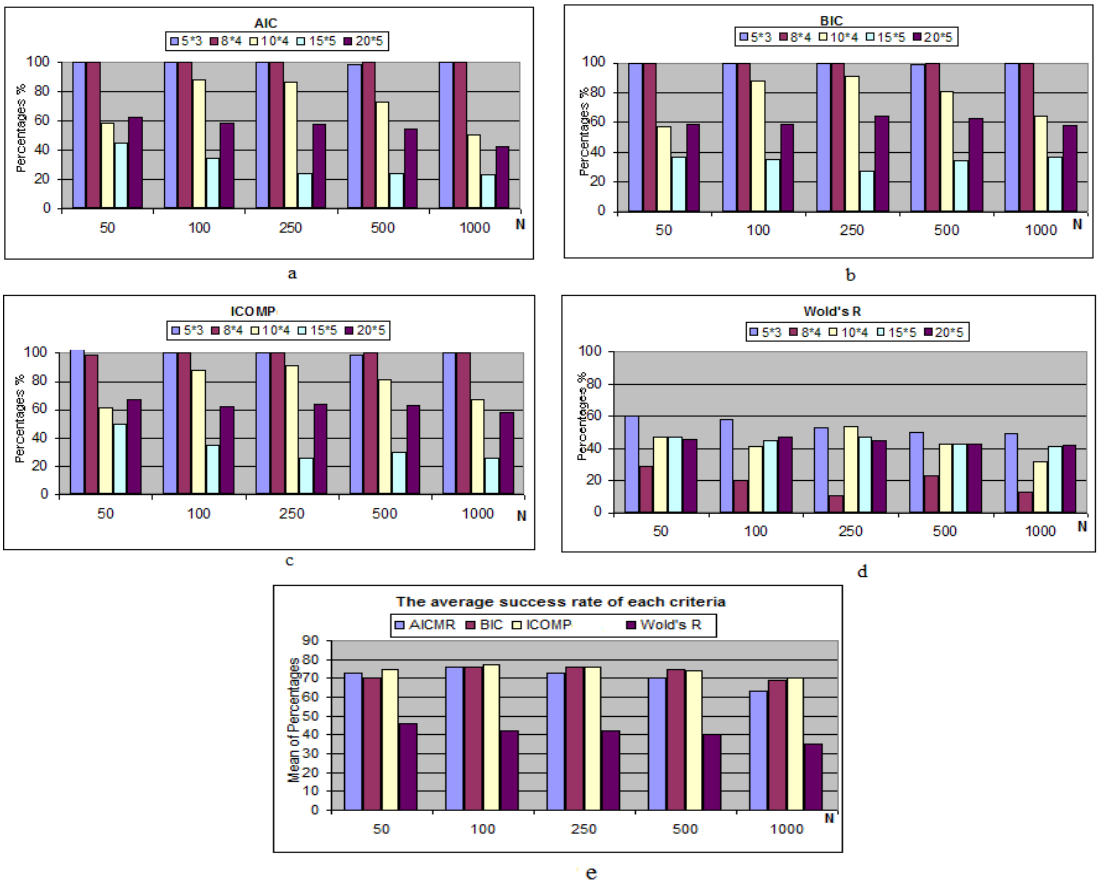


Figure 2. (a)-(d) Performance comparison of each model selection criterion for various dimension of models and sample sizes. (e) The average success rate of model selection criterion for various sample sizes

As can be seen from Figure 2 (a)-(f), there is dependency among dimension of the PLSR models, sample sizes and model selection criteria. AIC, BIC, and ICOMP truly estimate the latent variables of the underlying known PLSR models for the dimension 5*3 and the dimension 8*4. When the sample sizes increase and the dimension of PLSR models isconstant, these criteria have a slight tendency to over-fit their PLSR models. The simulation results show that BIC and ICOMP criteria achieved selecting the true number of latent variables with such a rate of approximately eighty percent for all design matrices. Generally it can be said that, when N and the dimension of PLSR

models increases, PLS creates a model with a high number of latent variables, which is statistically significant.

While the dimensions of the PLSR model change and the sample size is constant, variation in the criteria of performances is shown in Figure 3 (a)-(f).

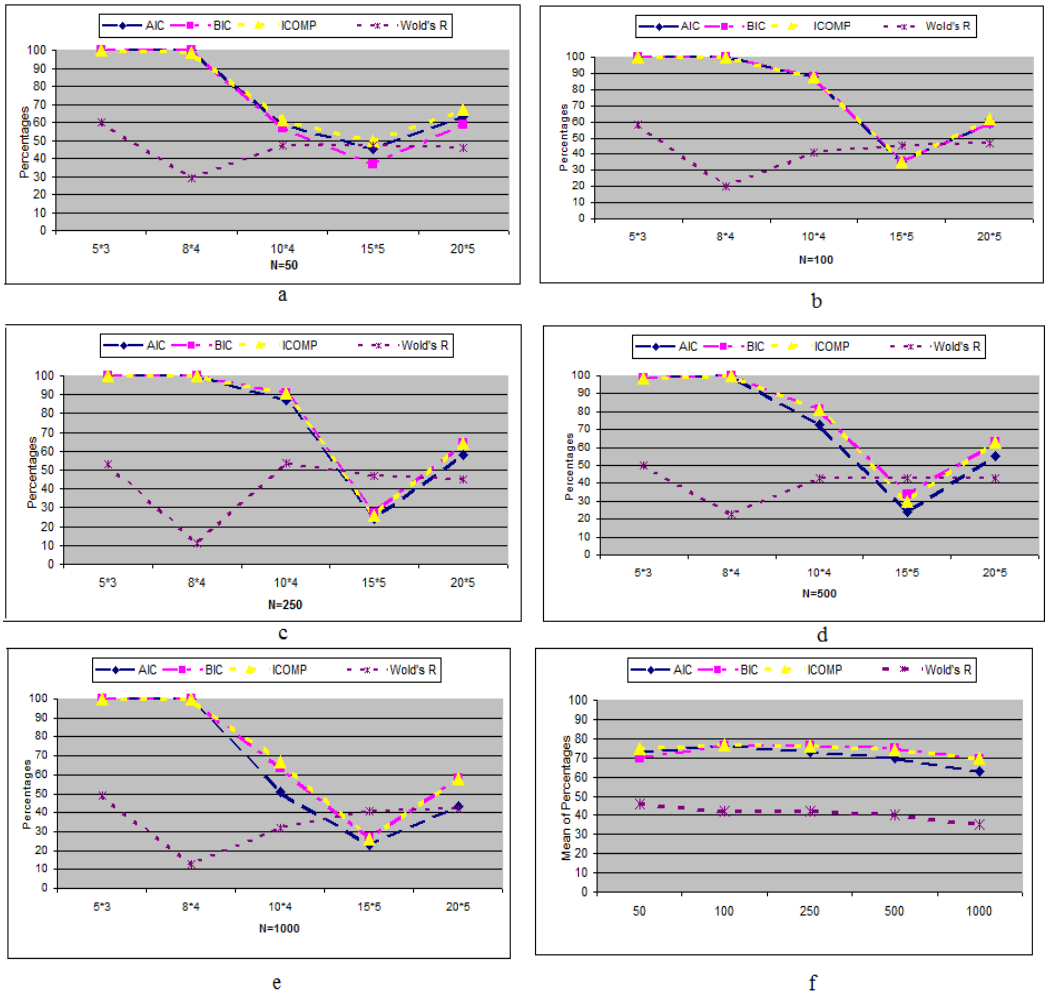


Figure 3. (a)-(e) Performance comparison of each model selection criterion for various dimensions of models for each sample size. (f) The average success rate of model selection criterion for various sample sizes

In each experiment all model selection criteria is applied to test how well they can identify the true known PLSR model. Figure 3(a)-(f) show the success rate of each criterion in identifying the true model. Since the performance of every criterion can be affected by the sample sizes and the dimension of PLSR models, the performances of AIC, BIC, and ICOMP criteria in general show a similar characteristic. Especially, when the dimension of models is smaller than $(p*a) \leq (10*4)$, AIC, BIC, and ICOMP criteria have acceptable performance and almost more accurately select the latent variables than the Wold's R criterion for all dimensions. As shown in Figure 3 (a)-(e), Wold's R criterion does not work well for any sample size and dimension. BIC and ICOMP criteria perform quite well, and in general select the true number of latent variables for known PLSR models.

Results depicted in Figure 3(a)-(e) clearly show that there is a significant reduction in the performances of AIC, BIC, and ICOMP criteria as the dimensions of the PLSR model increase. However, the performance of the Wold's R criterion almost stays the same but it is never satisfactory. In order to provide anoverall measure of success, the average success rate is calculated and shown in Figure 2(f) and Figure 3(f) for various sample sizes. It can be seen from this figure that on the average AIC, BIC, and ICOMP criteria out perform Wold's R criterion, and BIC and ICOMP criteria success rates are higher than the AIC and Wold's R criteria.

5. CONCLUSION

The major contribution of this paper is that this study evaluating the performances of AIC, BIC, ICOMP, and Wold's R criteria for model selection in PLSR, where the number of observations is typically much larger than the number of predictor variables. The aim of the analysis is to extract latent variables with respect to their partial contribution to total variance to build representative models of PLSR. Given this ranking to the latent variables, AIC, BIC, ICOMP, and Wold's R criteria are used to determine a consistent estimate of the dimension of the model. In a real life data set AIC, BIC, ICOMP criteria truly find the number of latent variables. It seems that AIC, BIC and ICOMP criteria are considerably better in choosing the right model when these are applied to our data set.

A simulation study is undertaken to compare the performances of AIC, BIC, ICOMP, and Wold's R criteria. Synthetic data are generated for different number of sample sizes and different dimensions of PLSR models. The simulation studies results clearly show much improved performances of BIC and ICOMP criteria in comparison to AIC and Wold's R criteria methods. The AIC, BIC, and ICOMP criteria properly finds latent variables for (p*a) < (10*4), for all sample sizes except the Wold's R criterion. It is seen from Table 7 that there is a dependency between dimension and sample size of the PLSR models and the success rates of the model selection criteria except for the Wold's R criterion.

In conclusion, these experiments showed that BIC and ICOMP criteria are considerably better than the traditional methods (AIC and Wold's R criteria) in choosing the right model when it is applied to our experimental set of synthetic data. An important point to make is that there is a big difference between the performances of AIC, BIC, ICOMP for different number of observations and the dimensions of PLSR models. Thus, one should select the sample size and the dimensions of the experiment in advance depending on the success rates of the criteria given in this study.

6. REFERENCES

Abdi, H., Salkind N. 2007. Encyclopedia of Measurement and Statistics. Thousand Oaks (CA): Sage.

Akaike, H., 1974. A new look at the statistical model identification, IEEE Transaction on Automatic Control 19, 716-723.

Bailey, C., 1994. Smart Exercise: Burning Fat, Getting Fit. Houghton-Mifflin Co., Boston, pp: 179-186.

Bedrick, E. J., Tsai, C. L., 1994. Model selection for multivariate regression in small samples. Biometrics 50, 226-231.

Behnke, A. R., J.H. Wilmore, 1974. Evaluation and Regulation of Body Build and Composition. Prentice-Hall, Englewood Cliffs, N. J., Pages: 236.

Boyer, K. L., Mirza, M. J., Ganguly, G., 1994. The Robust Sequential Estimator: A General Approach and its Application to Surface Organization in Range Data, IEEE PAMI 16, 987-1001.

Bozdoğan, H., 1987. Model Selection and Akaike's Information Criterion (AIC): The General Theory and Its Analytical Extensions, Psychometrica 52, 345-370.

Bozdoğan, H., 2000. Akaike's Information Criterion and Recent Developments in Information Complexity. Journal of Mathematical Psychology 44, 62-91.

Bozdoğan, H., 2004. Statistical Data Mining and Knowledge Discovery.Chapman and Hall/CRC, USA.

Bozdoğan, H., 2004. Intelligent statistical data mining with Information Complexity and Genetic Algorithms in Statistical Data Mining and Knowledge Discovery.Chapman and Hall/CRC, USA.

Clark, A. E., Troskie, C. G., 2006. Regression and ICOMP: A Simulation Study. Communications in Statistics Simulation and Computation 35, 591-603.

Eastment H. T., Krzanowski W. J. 1982. Cross-validatory choice of the number of components from a principal component analysis.Technometrics 24, 73-77.

Garthwaite, P. H., 1994. An interpretation of partial least squares, Journal of the American Statistical Association 89, 122-127.

Geladi, P., Kowalski, B. R., 1986. Partial least-squares regression a tutorial. Anal. Chim. Acta. 185, 1-17.

Haber, R., Unbenhauen, H., 1990. Structure identification of nonlinear dynamic systems-a survey on input/output approaches. Automatica 26 (4), 651-677.

Hastie, T., Tibshirani, R., Friedman J., 2001. The elements of statistical learning: data mining, inference, and prediction. New York, Springer.

Helland, I. S. 1990. Partial Least Squares Regression and Statistical Models, Scandinavian Journal of Statistics 17(2), 97-114.

Henry de-Graft, A. 2010. Comparison of Akaike information criterion (AIC) and Bayesian information criterion (BIC) in selection of an asymmetric price relationship. Journal of Development and Agricultural Economics 2(1): 001-006.

Jouan-Rimbaud Bouveresse, D., Rutledge, D. N., 2009. Two new extensions of principal component transform to compute a PLS2 model between two wide matrices: PCT-PLS2 and segmented PCT-PLS2. Analytica Chimica Acta. 642 (1-2), 37-44.

Katch, F., W. McArdle, 1977. Nutrition, Weight Control and Exercise. Houghton-MifflinCo., Boston.

Kanatani, K., 2002. Model Selection for Geometric Inference, The 5th Asian Conference on Computer Vision, Melbourne, Australia, pp. xxi-xxxii, January.

Kohavi, R., 1995. A study of cross-validation and boots trap for accuracy estimation and model selection. In Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 2, San Francisco, CA, USA, pp. 1137–1143. Morgan Kaufmann Publishers Inc.

Kuha, J. 2004. AIC and BIC: Comparisons of assumptions and performance. Sociological Methods and Research. (33)2: 188-229.

Kundu, D., Murali G., 1996. Model selection in linear regression, Computational Statistics and Data Analysis 22 (5), 461-469(9).

Li, B., Morris, J., Martin E. B., 2002. Model Selection for Partial Least Squares Regression. Chemometrics and Intelligent Laboratory Systems 64, 79-89.

Myung, I. J., 2000. The Importance of Complexity in Model Selection. Journal of Mathematical Psychology 44, 190-204.

Naes, T., Martens, H., 1985. Comparison of prediction methods for collinear data. Communication in Statistics Simulation and Computation 14, 545-576.

Niazi, A., Azizi, A., 2008. Orthogonal Signal Correction-Partial Least Squares Method for Simultaneous Spectrophotometric Determination of Nickel, Cobalt and Zinc. Turkish Journal of Chemistry 32, 217-228.

Saporta, G., 2008. Models for Understanding versus Models for Prediction. In Compstat 2008, Part IX, 315-322.

Schwarz, G., 1978. Estimating the Dimension of a Model. The Annals of Statistics6 (2), 461-464.

Siria, W. E., 1956. Gross Composition of the Body. InAdvance in Biological and Medical Physics, Lawrence J.H. and C.A. Tobias (Eds.). Academic Press, New York.

Torr, P. H. S., 1998. Model Selection for Two View Geometry: A Review, Model Selection for Two View Geometry: A Review, Microsoft Research, USA, Microsoft Research, USA.

Wang Y, Liu, Q., 2006. Comparison of Akaike information criteria (AIC) and Bayesian information criteria (BIC) in selection of stock recruitment relationships. Fisheries Research 77(2): 220-225.

Weakliem, L. D., 2004. Introduction to the Special Issue on Model Selection. Sociological Methods and Research. 33(2): 167-186.

Wilmore, J., 1976. Athletic Training and Physical Fitness: Physiological Principles of the Conditioning Process. Allynand Bacon, Inc., Boston.

Wold, H., 1966.Estimation of principal components and related models by iterative least squares.In P.R. Krishnaiaah (Ed.). Multivariate Analysis. (pp.391-420) New York: Academic Press.

Wold, H., 1982. Soft Modelling, The basic design and some extensions, in: K.- G. Jöreskog, H. Wold (Eds.), Systems Under Indirect Observation. Vols.I and II, North-Holland, Amsterdam.

Wold, S., Sjöström, M., Eriksson, L., 2001. PLS regression: a basic tool of chemometrics. Chemometrics and Intelligent Laboratory 58, 109-130.

KEKKR İÇİN MODEL SEÇME KRİTERLERİNİN PERFORMANS KARŞILAŞTIRMALARI: AIC, BIC, ICOMP ve WOLD'S R

ÖZET

Kısmi en küçük kareler regresyonu (KEKKR), çoklu bağlantının olduğu durumlarda, yanıt değişkeni $Y_{N\times M}$ *ile açıklayıcı değişkenler* $X_{N\times K}$ *arasında modelleme yapabilen istatistiksel bir yöntemdir. Kısmi en küçük kareler bölümünde, yanıt değişkenini en iyi açıklayabilecek gizli (latent) değişkenlerin elde edilmesi için bazı model seçme kriterleri uygulanır. Gizli değişkenlerin seçiminde kullanılan genel yaklaşımlar Akaike bilgi kriteri (AIC) ve Wold's R kriteridir.*

Bu çalışmada, gözlem sayısının açıklayıcı değişken sayısından fazla olduğu durumlarda, geleneksel yöntemler AIC ve Wold's R'a ek olarak Bayes bilgi kriteri (BIC) ve Bilgi karmaşıklık kriteri de (ICOMP) KEKKR için model seçme kriterleri olarak incelenmiştir. AIC, BIC, ICOMP ve Wold's R model seçme kriterlerinin performansları gerçek veri örneği ve benzetim çalışması yoluyla karşılaştırılmıştır. Benzetim çalışması sonuçları, farklı örneklem büyüklükleri, farklı sayıda açıklayıcı değişken ve yanıt değişkeninin olduğu durumlarda elde edilmiştir. Yapılan benzetim çalışması sonuçlarıBIC ve ICOMP model seçme kriterlerinin KEKKR modelleri için, gizli değişkenin seçiminde diğer model seçme kriterlerinden (AICveWold's R) çok daha etkili olduklarını ve daha doğru sayıda gizli değişken seçimi yaptıklarını göstermiştir.

Anahtar Kelimeler: AIC, BIC ve ICOMP bilgi kriterleri, K çapraz doğrulama, Kısmi en küçük kareler regresyonu, Model seçimi, Wold's R kriteri.

APPENDIX

Table 1. The components of X and Y matrix for 8*4

$\tilde{\mathbf{E}}$	$\tilde{\mathbf{E}} = [e_1,...,e_8] \sim N(0,0.01)$
$\mathbf{r_i}$	$r_1 \sim N(0,10), r_2 \sim N(0,5), r_3 \sim N(0,2), r_4 \sim N(0,0.5)$
$\boldsymbol{\xi_i}$	$\xi_1 = [0.1612 \quad 0.3030 \quad 0.4082 \quad 0.4642 \quad 0.4642 \quad 0.4082 \quad 0.3030 \quad 0.1612]'$ $\xi_2 = [0.3030 \quad 0.4642 \quad 0.4082 \quad 0.1612 \quad -0.1612 \quad -0.4082 \quad -0.4642 \quad -0.3030]'$ $\xi_3 = [0.4082 \quad 0.4082 \quad 0.0000 \quad -0.4082 \quad -0.4082 \quad -0.0000 \quad 0.4082 \quad 0.4082]'$ $\xi_4 = [0.4642 \quad 0.1612 \quad -0.4082 \quad -0.3030 \quad 0.3030 \quad 0.4082 \quad -0.1612 \quad -0.4642]'$
$\mathbf{f_i}$	$f_1 \sim N(0,0.25),\ f_2 \sim N(0,0.125),\ f_3 \sim N(0,0.05),\ f_4 \sim N(0,0.0125)$
	$\eta_{41} = [0.2887 \quad 0.5000 \quad 0.5774 \quad 0.5000]'$ $\eta_{42} = [0.5000 \quad 0.5000 \quad 0.0000 \quad -0.5000]'$ $\eta_{43} = [0.5774 \quad 0.0000 \quad -0.5774 \quad -0.0000]'$ $\eta_{44} = [0.5000 \quad -0.5000 \quad -0.0000 \quad 0.5000]'$
$\boldsymbol{\Psi}$	$\psi = \begin{pmatrix} 0.00010 & 0.00006 & 0.00006 & 0.00006 \\ 0.00006 & 0.00010 & 0.00006 & 0.00006 \\ 0.00006 & 0.00006 & 0.00010 & 0.00006 \\ 0.00006 & 0.00006 & 0.00006 & 0.00010 \end{pmatrix}$

Table 2. The components of X and Y matrix for 10*4

$\tilde{\mathbf{E}}$	$\tilde{\mathbf{E}} = [e_1,...,e_{10}] \sim N(0,0.02)$
$\mathbf{r_i}$	$r_1 \sim N(0,20),\ r_2 \sim N(0,10),\ r_3 \sim N(0,4),\ r_4 \sim N(0,1)$
$\boldsymbol{\xi_i}$	$\xi_1 = [0.4458 \quad 0.4349 \quad 0.4132 \quad 0.3813 \quad 0.3401 \quad 0.2904 \quad 0.2337 \quad 0.1711 \quad 0.1044 \quad 0.0351]'$ $\xi_2 = [0.4349 \quad 0.3401 \quad 0.1711 \quad -0.0351 \quad -0.2337 \quad -0.3813 \quad -0.4458 \quad -0.4132 \quad -0.2904 \quad -0.1044]'$ $\xi_3 = [0.4132 \quad 0.1711 \quad -0.1711 \quad -0.4132 \quad -0.4132 \quad -0.1711 \quad 0.1711 \quad 0.4132 \quad 0.4132 \quad 0.1711]'$ $\xi_4 = [0.3813 \quad -0.0351 \quad -0.4132 \quad -0.3401 \quad 0.1044 \quad 0.4349 \quad 0.2904 \quad -0.1711 \quad -0.4458 \quad -0.2337]'$
$\mathbf{f_i}$	$f_1 \sim N(0,0.5),\ f_2 \sim N(0,0.25),\ f_3 \sim N(0,0.1),\ f_4 \sim N(0,0.025)$
$\boldsymbol{\eta_{A^*i}}$	$\eta_{41} = [0.6935 \quad 0.5879 \quad 0.3928 \quad 0.1379]'$ $\eta_{42} = [0.5879 \quad -0.1379 \quad -0.6935 \quad -0.3928]'$ $\eta_{43} = [0.3928 \quad -0.6935 \quad 0.1379 \quad 0.5879]'$ $\eta_{44} = [0.1379 \quad -0.3928 \quad 0.5879 \quad -0.6935]'$
$\boldsymbol{\Psi}$	$\psi = \begin{pmatrix} 0.00010 & 0.00006 & 0.00006 & 0.00006 \\ 0.00006 & 0.00010 & 0.00006 & 0.00006 \\ 0.00006 & 0.00006 & 0.00010 & 0.00006 \\ 0.00006 & 0.00006 & 0.00006 & 0.00010 \end{pmatrix}$

Table 3. The components of X and Y matrix for 15*5

$\tilde{\mathbf{E}}$	$\tilde{\mathbf{E}}=[e_1,...,e_{15}] \sim N(0,0.02)$
$\mathbf{r}_i$	$r_1 \sim N(0,20),\ r_2 \sim N(0,10),\ r_3 \sim N(0,5),\ r_4 \sim N(0,3.5),\ r_5 \sim N(0,0.8)$
ξ_i	$\xi_{1(1,...,10)}=[0.3646\ \ 0.3607\ \ 0.3527\ \ 0.3409\ \ 0.3253\ \ 0.3062\ \ 0.2838\ \ 0.2582\ \ 0.2298\ \ 0.1989]'$ $\xi_{1(11,...,15)}=[0.1658\ \ 0.1309\ \ 0.0945\ \ 0.0571\ \ 0.0191]'$ $\xi_{2(1,...,10)}=[0.3607\ \ 0.3253\ \ 0.2582\ \ 0.1658\ \ 0.0571\ \ -0.0571\ \ -0.1658\ \ -0.2582\ \ -0.3253\ \ -0.3607]'$ $\xi_{2(11,...,15)}=[-0.3607\ \ -0.3253\ \ -0.2582\ \ -0.1658\ \ -0.0571]'$ $\xi_{3(1,...,10)}=[0.3527\ \ 0.2582\ \ 0.0945\ \ -0.0945\ \ -0.2582\ \ -0.3527\ \ -0.3527\ \ -0.2582\ \ -0.0945\ \ 0.0945]'$ $\xi_{3(11,...,15)}=[0.2582\ \ 0.3527\ \ 0.3527\ \ 0.2582\ \ 0.0945]'$ $\xi_{4(1,...,10)}=[0.3409\ \ 0.1658\ \ -0.0945\ \ -0.3062\ \ -0.3607\ \ -0.2298\ \ 0.0191\ \ 0.2582\ \ 0.3646\ \ 0.2838]'$ $\xi_{4(11,...,15)}=[0.0571\ \ -0.1989\ \ -0.3527\ \ -0.3253\ \ -0.1309]'$ $\xi_{5(1,...,10)}=[0.3253\ \ 0.0571\ \ -0.2582\ \ -0.3607\ \ -0.1658\ \ 0.1658\ \ 0.3607\ \ 0.2582\ \ -0.0571\ \ -0.3253]'$ $\xi_{5(11,...,15)}=[-0.3253\ \ -0.0571\ \ 0.2582\ \ 0.3607\ \ 0.1658]'$
$\mathbf{f}_i$	$f_1 \sim N(0,0.05),\ f_2 \sim N(0,0.025),\ f_3 \sim N(0,0.0125),\ f_4 \sim N(0,0.05),\ f_5 \sim N(0,0.0125)$
η_{A^*i}	$\eta_{51}=[0.6247\ \ 0.5635\ \ 0.4472\ \ 0.2871\ \ 0.0989]'$ $\eta_{52}=[0.5635\ \ 0.0989\ \ -0.4472\ \ -0.6247\ \ -0.2871]'$ $\eta_{53}=[0.4472\ \ -0.4472\ \ -0.4472\ \ 0.4472\ \ 0.4472]'$ $\eta_{54}=[\ 0.2871\ \ -0.6247\ \ 0.4472\ \ 0.0989\ \ -0.5635]'$ $\eta_{55}=[\ 0.0989\ \ -0.2871\ \ 0.4472\ \ -0.5635\ \ 0.6247]'$
ψ	$\psi=\begin{pmatrix}0.00010 & 0.00006 & 0.00006 & 0.00006 & 0.00006\\ 0.00006 & 0.00010 & 0.00006 & 0.00006 & 0.00006\\ 0.00006 & 0.00006 & 0.00010 & 0.00006 & 0.00006\\ 0.00006 & 0.00006 & 0.00006 & 0.00010 & 0.00006\\ 0.00006 & 0.00006 & 0.00006 & 0.00006 & 0.00010\end{pmatrix}$

Table 4. The components of X and Y matrix for 20*5

$\tilde{\mathbf{E}}$	$\tilde{\mathbf{E}}=[e_1,...,e_{20}] \sim N(0,0.02)$
$\mathbf{r}_i$	$r_1 \sim N(0,30),\ r_2 \sim N(0,20),\ r_3 \sim N(0,10),\ r_4 \sim N(0,6),\ r_5 \sim N(0,3)$
ξ_i	$\xi_{1(1,...,10)}$=[0.3160 0.3140 0.3102 0.3044 0.2967 0.2872 0.2759 0.2629 0.2483 0.2322 $\xi_{1(11,...,20)}$=[0.2147 0.1958 0.1757 0.1545 0.1324 0.1095 0.0858 0.0617 0.0372 0.012 $\xi_{2(1,...,10)}$ = [0.3140 0.2967 0.2629 0.2147 0.1545 0.0858 0.0124 -0.0617 -0.1324 -0.19 $\xi_{2(11,...,20)}$ = [-0.2483 -0.2872 -0.3102 -0.3160 -0.3044 -0.2759 -0.2322 -0.1757 -0.1095 -0. $\xi_{3(1,...,10)}$ = [0.3102 0.2629 0.1757 0.0617 -0.0617 -0.1757 -0.2629 -0.3102 -0.3102 -0.26 $\xi_{3(11,...,20)}$ = [-0.1757 -0.0617 0.0617 0.1757 0.2629 0.3102 0.3102 0.2629 0.1757 0.0 $\xi_{4(1,...,10)}$ =[0.3044 0.2147 0.0617 -0.1095 -0.2483 -0.3140 -0.2872 -0.1757 -0.0124 0.15 $\xi_{4(11,...,20)}$ =[0.2759 0.3160 0.2629 0.1324 -0.0372 -0.1958 -0.2967 -0.3102 -0.2322 -0.0 $\xi_{5(1,...,10)}$ = [0.2872 0.0858 -0.1757 -0.3140 -0.2322 0.0124 0.2483 0.3102 0.1545 -0.10 $\xi_{5(11,...,20)}$ = [-0.2967 -0.2759 -0.0617 0.1958 0.3160 0.2147 -0.0372 -0.2629 -0.3044 -0.
$\mathbf{f}_{\cdot 1}$	$f_1 \sim N(0,0.4),\ f_2 \sim N(0,0.1),\ f_3 \sim N(0,0.5),\ f_4 \sim N(0,0.02),\ f_5 \sim N(0,0.00125)$
$\boldsymbol{\eta}_{A^*i}$	η_{51} = [0.4472 0.4472 0.4472 0.4472 0.4472]' η_{52} = [0.4472 0.5635 -0.0989 -0.6247 -0.2871]' η_{53} = [0.4472 -0.0989 -0.2871 0.5635 -0.6247]' η_{54} = [0.4472 -0.6247 0.5635 -0.2871 -0.0989]' η_{55} = [0.4472 -0.2871 -0.6247 -0.0989 0.5635]'
ψ	$\psi = \begin{pmatrix} 0.00010 & 0.00006 & 0.00006 & 0.00006 & 0.00006 \\ 0.00006 & 0.00010 & 0.00006 & 0.00006 & 0.00006 \\ 0.00006 & 0.00006 & 0.00010 & 0.00006 & 0.00006 \\ 0.00006 & 0.00006 & 0.00006 & 0.00010 & 0.00006 \\ 0.00006 & 0.00006 & 0.00006 & 0.00006 & 0.00010 \end{pmatrix}$

Volume: 10 Number: 03 Category: 05 Page: 35-48 ISSN No: 1303-6319 Cilt: 10 Sayı: 03 Kategori: 05 Sayfa: 35-48

İSTATİSTİKTE YENİ EĞİLİMLER VE YÖNTEMLER

Fikri AKDENİZ*

ÖZET

Bu çalışmada istatistik disiplininin önemine değinilerek, istatistiksel düşünme ve veri analizi kavramları açıklanmıştır. Veriyi bilgiye dönüştürmek için veri bilimcilerin yeri vurgulanmıştır. İstatistiğin geleceği hakkında kısaca bilgi verilmiş ve ülkemiz istatistikçilerinin uluslararası yayın yapmak için düzeyli araştırmalara ağırlık vermeleri yönünde yapılması gerekenler verilmiştir.

Anahtar Kelimeler: İstatistik disiplini, İstatistiksel düşünme, Veri, Veri analizi, Veri bilimci.

1. GİRİŞ[1]

İstatistiğin gelişimi: İstatistik uzun bir antikiteye fakat kısa bir tarihe sahiptir. Onun orijini insanlığın başlangıcına dayanmakla birlikte büyük pratik öneminin algılanması çok uzun olmayan bir zamana sahiptir. İstatistiksel metodolojinin (yöntembiliminin) gelişiminde bilgisayarların etkisi büyüktür.

1.1 İstatistik Nedir?

Çağdaş anlamda istatistik; doğal olaylara dayanan ve gözlemle bulunan verilerin, bilimsel yöntemlerle incelenmesi ve doğru sonuç çıkarılması için kullanılan tekniklerin tamamıdır.

Genel olarak istatistik; gözlemle elde edilen bilgileri düzenleme, analiz etme ve bunlardan sonuç çıkarma sanatı ve bilimidir.

İstatistik, İngiltere'de 1834 yılında İstatistik Derneğinin kurulmasından sonra bir bilim dalı olarak kabul edildi ve insanlarla ilgili olguları uygun bir şekilde göstermek için sayılarla ifade edilebilen genel kurallar olarak düşünüldü.

Böylece daha önceleri veri anlamında kullanılan istatistik sözcüğü, veriyi yorumlama ve kaynağı ne olursa olsun veriden sonuç çıkarma anlamını kazanmaya başladı.Kavramlar, tanımlar ve verinin toplanması için ortak yöntemleri tartışmak üzere farklı ülkelerin verilerinin karşılaştırılabilir olması için 1853 te Brüksel'de 26 ülkeden 153 delegenin katılımıyla ilk kez uluslararası istatistik kongresi yapılmış (Osmanlı Devleti de bu toplantıya temsilci göndermişti), bunu sonraki yıllarda diğerleri izlemiştir. (Akdeniz ve Dönmez, 1999) 24 Haziran 1885 yılında da Uluslararası İstatistik Topluluğu kurulmuştur.

*Prof. Dr., Çağ Üniversitesi, Matematik ve Bilgisayar Bölümü, 33800 Yenice-Tarsus/Mersin, e-posta: fikri@akdeniz.edu.tr

[1]Bu çalışma 15. Uluslararası Ekonometri, Yöneylem Araştırması ve İstatistik (EYİ) Sempozyumunda panelist olarak görev alan yazarın 24 Mayıs 2014 günü yaptığı konuşma metnidir. Yazının amacı genel istatistik konusunda yazarın okuyucu kitlesi ile düşündüklerini paylaşmasıdır.

İçinde yaşadığımız yüzyılda bilgi çağı kavramının geliştirilmesinde istatistiğin rolü çok önemlidir. Çünkü her türlü ulusal ve uluslararası, sosyal, ekonomik ve diğer gelişme hedeflerinin belirlenmesi ve bu hedeflerin başarıya ulaşması güncel ve güvenilir istatistiksel çalışmalara dayandırılmasına bağlıdır.

İstatistik, belirsizlik ortamında, araştırma, tahmin yapma ve karar verme mekanizmaları geliştiren bir bilim dalı olup, aynı zamanda diğer bilim dallarının da teknolojisi olarak kabul edilmektedir. Bu bakış açısı dikkate alındığında, özellikle rasgele deney ve gözleme dayalı bilimsel çalışma olup da istatistiksel değerlendirmesinin olmayacağı hiçbir araştırma düşünülemez.

İstatistikçinin kaygısı sadece geçmiş deneyleri analiz etmek değil aynı zamanda yeni deneyler oluşturmak, kaynakların verimli ve doğru bir şekilde kullanıldığını kontrol etmek ve sorulan sorulara uygun deneyler yapılmasını sağlamaktır.

Karar verirken, bazıları uzman görüşü olabilen, farklı kaynaklardan toplananbirçok parçadan oluşan bilginin ve elde bulunan tüm kanıtların göz ününe alınması gerekir. Bu bağlamda aşağıdaki sorularla karşılaşırız.

Kullanılan bilginin her bir parçası ne derece güvenilir olmaktadır? Bu bilginin araştırılan problemle ilgisi ne kadardır? Bilginin farklı parçaları birbiriyle tutarlımıdır? Sonuca varmak için, farklı kaynaklardan gelen ve tümü birbiriyle uyumlu olmayabilen bilgileri nasıl birleştirebiliriz?

Yetersiz bilgilerle karar vererek tümevarımlı sonuç çıkarma yoluna gidilmemelidir. Bu durumda "Kuluçkadan çıkmadan önce civcivleri saymayınız" özdeyişini ya da eski bir Çin atasözü olan "Tahmin etmek ucuzdur, ama yanlış tahmin pahalıya patlar" özdeyişini anımsamak uygun düşüyor. O halde istatistik gerçeğin araştırılmasında kaçınılmaz bir özel araçtır.

1.2 İstatistik; Fizik, Kimya, Biyoloji ya da Matematik gibi Ayrı Bir Temel Bilim midir?

Fizikçi ısı, ışık, elektrik ve devinim kanunları gibi doğal olayları inceler. Kimyacı, kimyasal maddeler arasındaki etkileşimleri ve cevherlerin bileşimlerinin analizini yapar. Biyolog, bitki ve hayvan yaşamlarını inceler. Matematikçi, verilmiş varsayımlara dayanarak çıkardığı önermeler üzerinde çalışır. Bu bilim dallarının her biri, kendi problemlerine ve onların çözümü için kendi yöntemlerine sahiptir. Bu özellik, onlara ayrı bir disiplin olma statüsü vermektedir. Bu anlamda günümüzde uygulanan ve çalışılan istatistik ayrı bir temel bilim alanı mıdır? Çözümünü istatistiğin gerçekleştirdiği tam olarak istatistik problemleri var mıdır?

İstatistik bir temel bilim alanı değilse diğer bilimlerin problemlerinin çözümünde uygulanan bir çeşit sanat, mantık ya da teknoloji midir? Acaba, istatistik üçünün bir kombinasyonu mudur?

1.3 İstatistik Neden Bir Bilimdir?

İstatistik bir bilimdir: Bazı temel ilkelerden çıkarılan, birçok teknikten oluşan bir zenginliğe sahip kimliğinin olması anlamında istatistik, bir bilimdir.

İstatistik bir teknolojidir: istatistiksel yöntem bilimi, endüstriyel üretimdeki kalite kontrol programları gibi bir işletim sistemine uygulanabilir olması bakımından bir teknolojidir.

İstatistiksel yöntemler, bireysel ve kurumsal çabaların etkinliğini maksimuma ulaştırmada ve belirsizliği azaltarak kabul edilebilir düzeye getirmede kullanılır.

İstatistik bir sanattır: Farklı istatistikçiler, aynı veri ile farklı sonuçlara varabilirler. Sunulan veride, çoğu kez var olan istatistiksel araçlarlaelde edilebilecek olandan daha çok bilgi bulunabilir. İstatistikçinin deneyimi burada önem kazanır. Budurum istatistiği sanat yapar. Böylece daha geniş anlamda, istatistik ayrı bir bilim dalıdır, belki de disiplinler arası bir bilim dalıdır (Rao, 1989).

2. İSTATİSTİK MEZUNLARININ SAHİP OLMASI GEREKEN ÖZELLİKLER NE OLMALIDIR?

İstatistik kuramı ve yöntemlerinin çalışmalarda nasıl kullanılacağı ve olaylara hangi açılardan bakılması gerektiği üzerinde özenle durulmalıdır. Birçok bilim insanı neden istatistiksel bir çalışmaya gerek duyulduğunu bilmemektedir.

İstatistikçinin başarılı olabilmesi için özverili, işini bilerek, uygun adımları izleyerek ilerlemesi gerekir. Görevinin bilincinde olan bir istatistikçi başkalarına yardım edeceği düşüncesi ile doğru adımları kullanarak ve analiz ederek araştırmaya devam eder. İstatistikçi olmak isteyen adaylar güçlü bir matematik bilgisine sahip olmanın yanında derinlemesine araştırma yapabilen, sabırlı ve ayrıntılara özen gösteren özelliklere sahip bireyler olmalıdırlar. Bu nedenle istatistik mezunlarının sahip olması gereken özel ve genel becerileri aşağıdaki gibi verebiliriz.

2.1 Özel Beceriler

- İstatistiğin doğasını anlamak (çalışma alanlarını, sınırlamalarını ve istatistiksel araştırmanın rolünü öğrenmek) ve istatistiksel sonuç çıkarmak,
- Bilimsel bir konuyu istatistiksel bir soru haline getirmek,
- Genel olarak kullanılan teknikleri ve bunlarla ilgili modelleri anlamak,
- Verideki varyasyonun doğasını tanımak ve model oluşturmak,
- İstatistikte kullanılan matematiksel yöntemleri bilmek ve bunları uygun durumlarda gereklidönüşümleri yapmak için kullanabilmek,
- İstatistiksel hesaplamaları yapmak için istatistik Paket Programları kullanabilmek,
- İstatistiksel veriyi algılamak, uygun örneklemi seçmek ve deneysel tasarımı oluşturmak için grafiksel teknikleri bilmek,
- Veritabanı teknolojilerini de iyi bilmek gerekiyor.

2.2 Genel Beceriler

- Doğru ve tutarlı düşünmek,
- Bireysel ve ortaklaşa olarak etkili ve üretken çalışma yapabilmek,
- Verilen görevleri programa uygun olarak bitirmek,
- Açık ve düzenli olarak araştırma sonuçları hakkında rapor vermek.

3. ULUSLARARASI İSTATİSTİK ENSTİTÜSÜ (ISI)'NÜN GÖRÜŞÜ

Uluslararası İstatistik Enstitüsü 21 Ağustos 1985'teki kurultayında istatistikçilere güç ve cesaret vermek için meslek etiği ile ilgili bir deklerasyon (duyuru) yayınlamıştı. Bu bildirgede meslek değerleri aşağıdaki gibi belirtilmişti:

1. TANIMAK VE SAYGI DUYMAK (Gelen veriye güvenmek)
2. PROFESYONELLİK (Sorumluluk, uzmanlık bilgisi, yeterlilik, bilgiye dayalı karar),
3 AÇIK SÖZLÜLÜK, DOĞRULUK ve DÜRÜSTLÜK (Bağımsızlık, objektiflik, şeffaflık, gerçekçilik) olarak ayrılmıştı.

3.1 Dünya İstatistik Günü

Birleşmiş Milletler Genel Kurulu 64/267 sayılı kararname ile 20 Ekim 2010 tarihini "Resmi istatistiklerdeki başarıların kutlanması" ana teması altında hizmet, doğruluk ve profesyonellik ana ilkeleri göz önüne alınarak "Dünya istatistik günü" olarak kutlanması kararlaştırılmıştır. Üye ülkelerin yanı sıra Birleşmiş Milletlere bağlı alt kuruluşlar, diğer uluslararası ve bölgesel kuruluşlar ile sivil toplum örgütlerinin de söz konusu kutlamaya katılımı teşvik edilmektedir.

3.2 Neden Dünya İstatistik Günü?

Dünya istatistik gününün kutlanması ulusal ve uluslararası düzeyde küresel istatistiksel sistemle verilen hizmeti onaylamış olacaktır. Resmi istatistiklere kamunun güven duyması ve bilgi sahibi olmaları konusunda düşüncenin güçlendirilmesine yardımcı olması umut edilmektedir. Tüm Dünya istatistikçileri arasında ortak bilinci arttırmayı amaçlar**.**

3.3 Ne Beklenir?

Dünya istatistik gününde, ulusal düzeydeki etkinlikler resmi istatistiklerin rolünü ve ulusal istatistiksel sistemin pek çok başarısını vurgulayacaktır. Uluslararası, bölgesel ve alt-bölgesel organizasyonlar-diğer özel etkinliklerle birlikte- ulusal etkinliklerle birbirini tamamlayacaktır.

4. YENİ BİR MESLEK: VERİ BİLİMCİ

28 Ekim 2012 günü bir gazetede çıkan haberi sizlerle paylaşmak istiyorum.

Yazının başlığı: YENİ BİR MESLEK: VERİ BİLİMCİ

İnternetteki bilgi ve veri bombardımanına akıllı cihazlardan gelenler de eklendikçe iş daha da karmaşık bir hal alıyor. Tahminlere göre 2020 yılında, otomobilden ev aletlerine ve telefonlara kadar yaklaşık 50 milyar cihaz veri üretecek ve birbirileri ile iletişimde olacaktır. İleriye yönelik öngörülerde bulunmak ve karar almak isteyen şirketler için bu "veri tsunamisi" ni doğru analiz edebilmek kritik önem taşıyacak. Bu durum, adına "veri bilimci" denilebilecek yeni bir mesleği doğuruyor. Bunlar, tüm kaynaklardan gelen verileri toplayıp analiz edebilecek. Bunları çalıştıran şirketler de önemli avantaj sağlayacak.

Tüm dünyada büyük şirketlere veri ambarı, kampanya yönetimi ve büyük veri çözümleri sunan bir şirket olan TERADATA'ya göre örneğin, bir perakendeci bu sayede kar marjını yüzde 60 artırabilecektir. Bugün bile ABD sağlık sektörü, veri bilimcileri kullanarak yılda 300 milyar dolarlık tasarruf edebilecek durumdadır. Bu yüzden birkaç yıl içinde veri bilimcilere talep patlayacak ve 2018 de bu alanda 140-190 bin kişilik istihdam açığı doğabilecektir.

4.1 Veri Analizcileri (Data Analysts) ve Veri Bilimciler (Data Scientists)

"Descriptive Analysis-Tanımlayıcı Analiz" yapan veri analizcilerin, organizasyon içindeki rolleri; varolanı raporlama, durumu açıklamakla görevli kişi olarak tanımlandı. Tek bir bakış açısıyla, elindeki verilerle yola çıkan ve her zaman aynı sonuca ulaşabilen veri analizcilerin rolü, şirketin karar destek süreçlerine girdi üretmekle sınırlıydı. Teknolojinin yardımıyla bağlantılı olduğumuz dünyada, verinin hareket hızı inanılmaz düzeydedir. Çok hızlı veri aktarımı, şirketlerin, durumları anında algılayarak hızlı tepki verebilmelerine olanak vermektedir. İnternet döneminin başlamasıyla birlikte artık sadece kurum içinde değil, kurum dışı ile de yoğun bir veri alışverişi var. Veriyi matematiksel işlemler ve istatistiksel yöntemlerle sunuma hazırlayan bu bilimcilerin doğrusal cebir, sayısal analiz ve makina dili gibi alanlarda da çalışmalar yapmaları gerekiyor.

Etkin bir veri analisti olmak için aşağıdaki teknik becerilere hakim ve beşeri niteliklere sahip olmak gerekir:

- Büyük veri kümelerine erişebilme,
- Sorgulama ve seçim ile bilgi kaşifliği (buluculuğu) yapabilme,
- İş ya da uygulama problemlerini çözümleyecek modeller kurma (sınıflandırma, kümelendirme ve anormal durum belirleme),
- Analitik paketlerde SAP, SPSS ve SSIS gibi yazılım programları ile veri transferi yapabilme,
- Veri analizlerini uygulamaya sunmak yani sonuçları (bulguları) görselleştirmek,
- Kullanıcılarınız ile etkin bir iletişim,
- İyi ve güçlü sezgiler,
- Bütün analizlerin ötesinde mantıksal çözümlemeleri derinliğine anlamaktır.

Bilgi çağının en değerli madeni veridir. Bu veri madenini işlemek ve veriyi bilgiye dönüştürmek için veri bilimcilere gereksinim duyulmaktadır. Bunlar analitik düşünebilme yeteneğine sahip kişilerdir. 2000'li yılların başından bu yana etkili olan veri analizi teknolojileri, bu yönde insan kaynaklarının gelişimini cesaretlendirmişti. Değişime ayak uydurabilen organizasyonlar, veri analizi teknolojileriyle bir adım öne geçme konusunda iş çözümleyiciliğinin önemini kavrayarak tüm süreçlerini ve alt yapılarını bu teknolojilerin uygulanmasına elverişli hale getirmeye başladılar.

Google, Facebook, Amazon, Yahoo, Walmart, Facebook, LinkedIn, Twitter gibi öncü kurumlar büyük veriyi yönetmek için geliştirilen teknolojileri kullanabilecek beceri ve zekaya sahip insanların da en az teknoloji kadar önemli olduğuna inanmaları sonucu veri bilimcilerini kullanmaktadırlar.

"**Predictive Analysis-Çıkarımsal Analiz**" yapan veri bilimcilerin organizasyon içindeki rolleri henüz tanımlı değil. Bununla birlikte veri bilimci farklı veri kaynaklarından beslenen büyük veri yönetimi için, bir bilim adamı gibi hipotezler kurup, bu hipotezlerin doğruluğunu ya da yanlışlığını test etmekiçin araştırmalar yapar. **Veri Odaklı Uygulamalar**" geliştirir. Bu uygulamaların birkaç önemli karakteristiği vardır:

- Bu uygulamalar veriden faydalanarak ortaya çıkar.
- Bu uygulamaların kullanımı sonucunda yeni veri ortaya çıkar.
- Yeni çıkan bu veri, uygulamaların iyileştirilmesi için kullanılır.

Farklı kaynaklardan toplanan veri, hiç bir zaman tek ve kesin bir sonuç vermez. Hatta varolan sonuçları yani, bilgiyi bile sorgulama şansı verir. İşin en can alıcı noktası da burasıdır. Farklı sonuçlar üretebilmek. Veri bilimcilerinin yarattığı katma değer, veriyi görsel olarak sunabilmek, verileri ayrıştırabilmek ve organize etmek. Aynı zamanda, bu verileri yorumlayabilmek için gelişmiş algoritmalar hazırlamak ve bu şekilde işlenen veriler ile iş kararları verilmesini sağlamaktır. 21. yüzyılın en cazip mesleklerinden biri olan veri bilimciliği için nasıl bir eğitim ve alt yapı gerekiyor. Henüz veri bilimci yetiştiren bir akademik kurum yok. Ağırlıklı olarak Bilgisayar, Matematik, İstatistik, Yönetim ve bilişim sistemleri eğitimine sahip olan bu kişiler, ekonomi alanında da eğitim görmüş olabiliyor. Donanım, yazılım ve bilişim teknolojisindeki hızlı gelişmenin çok önem kazandığı günümüzde, veri bilimcilerin ne kadar değerli bir insan kaynağı olduğu konusunda her kesim mutabıktır.

4.2 Veri Bilimi

Şekil 1'de Matematik, İstatistik, İleri programlama, Görselleştirme, Bilimsel yöntem, Veri mühendisliği, Alan uzmanlığı, Hacker kafa yapısı etkileşimi görülmektedir.

Şekil 1. Veri bilimi ve diğer alanların ilişkisi

5. VERİ ANALİZİNE BAKIŞ

Şimdi veri analizinin nasıl yapıldığını aşağıdaki izlenen adımlarda gösterilen ilişkilerle inceleyelim:

Gerçek yaşamdaki problemler

1. Veri toplama: Belirli problemlerin formülleştirilmesi:
   a) Deneysel tasarım
   b) Tarihsel kayıtlar
   c) Örneklem surveyleri

2. Gözlemler (Veri):
   a) Kaydedilen ölçümler
   b) Ön bilgi

3. Verinin çapraz sorgulanması:
   a) Sapan değerlerin, hataların (ölçüm hatası ve hatalı kayıt) ayıklanması, verinin uydurma ya da gerçek olup olmadığı,
   b) Ön bilginin (geçerliliğinin)test edilmesi

4. Tanı koyma (teşhis): Model seçimi ve modelin onaylanması

5. Sonuç çıkarma amaçlı veri analizi:
   a) Hipotez testi
   b) Tahmin etme
   c) Karar verme

6. Sonraki araştırma için yol gösterici olma: Yeniden ilk adıma dönme

Böylece **Veri analizi=Belirli soruları yanıtlama+bir araştırmanın yeni aşamaları için bilgi hazırlamak** biçiminde özetlenebilir (Rao, 1989).

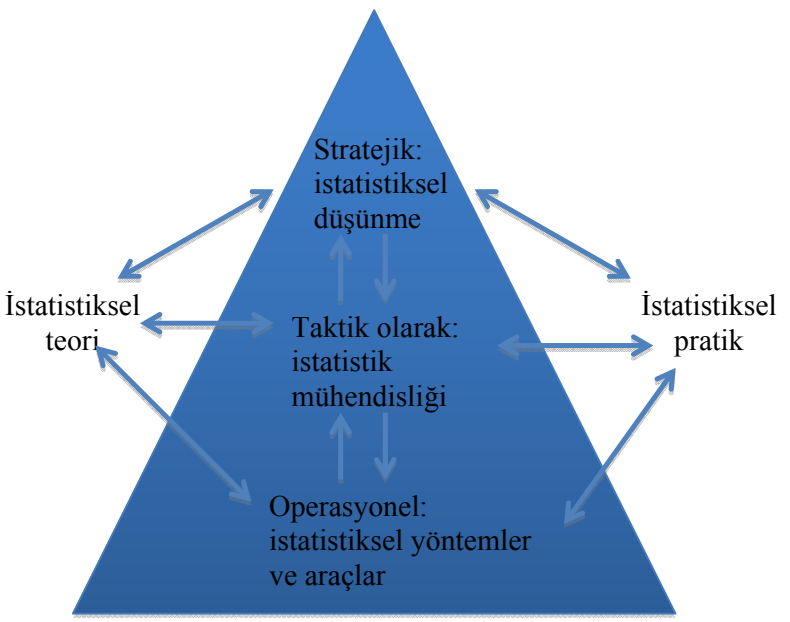
6. İSTATİSTİK MÜHENDİSLİĞİ

İstatistik mühendisliği, var olan bilimi ve istatistik teorisini kullanarak daha büyük bir gelişmenin nasıl çıkarılacağını inceleyen bir disiplindir. İstatistik mühendisliği ile ilgili olarak Snee ve Hoerl (2010) Şekil 2, Şekil 3 ve Şekil 4 deki gösterimleri verdiler. Ayrıca, istatistik mühendisliğini istatistiksel kavramları, yöntemleri ve araçları bilgi teknolojileri ile birleştirerek en iyi şekilde nasıl uygulanacağını ve diğer ilgili bilim dalları için geliştirilmiş sonuçları oluşturmak biçiminde tanımladılar. Daha geniş bilgiye aşağıdaki makalelerden ulaşılabilir. (Hoerl ve Snee (2010a, 2010b)). Snee ve Hoerl, (2011); Steiner ve Mckay, (2014)).

İstatistik mühendisliğini, aşağıdaki üç şekli inceleyerek açıklamaya çalışacağız.

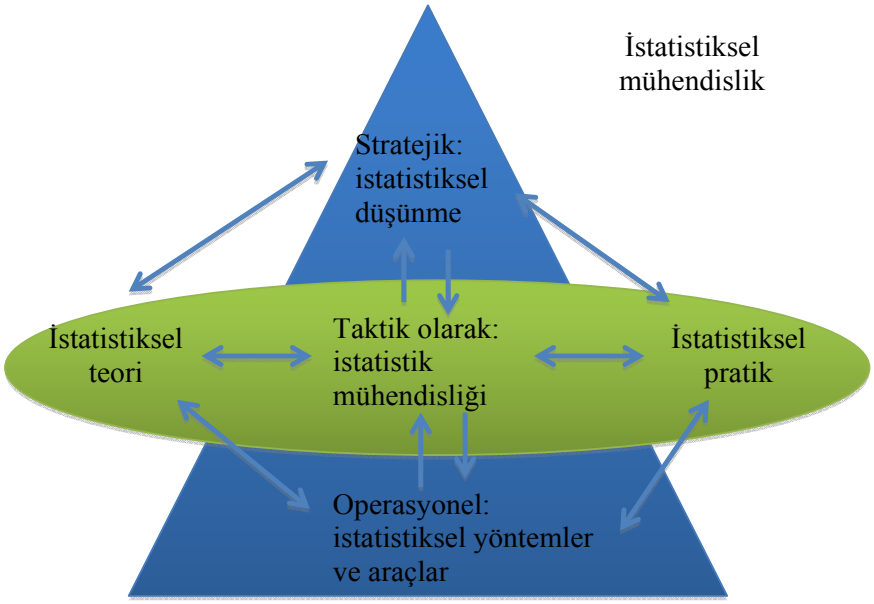
6.1 Bir Sistem Olarak İstatistik Disiplini

STRATEJİK: İstatistiksel düşünme
TAKTİK OLARAK: İstatistik mühendisliği
OPERASYONEL: İstatistiksel yöntemler ve araçlar



Şekil 2. Bir sistem olarak istatistik disiplini

6.2 Bir Sistem Olarak İstatistik Disiplini: İstatistik Mühendisliği

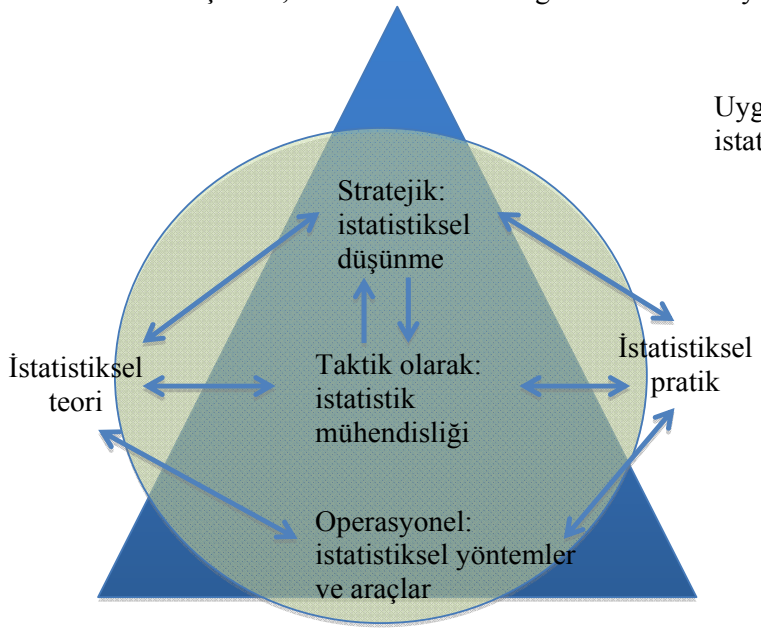


Şekil 3. Bir sistem olarak istatistik disiplini: İstatistik mühendisliği

Şekil 3'teki yatay elips dilimi istatistiksel teori ile istatistiksel uygulama bağlantısını göstermektedir.

6.3 Bir Sistem Olarak İstatistik Disiplini: Uygulamalı İstatistik

Uygulamalı istatistik, gerçek problemlere istatistiğin uygulanmasıdır. Bu nedenle dikey dilim istatistiksel düşünme, istatistik mühendisliği ve istatistiksel yöntemleri içerir.



Şekil 4. Bir sistem olarak istatistik disiplini: Uygulamalı istatistik

7. TÜRKİYEDE İSTATİSTİK BİLİMİNİN GELİŞİMİ

Ülkemizde dönemin hükümetine yardımcı olmak üzere görevli olduğu (1956-1958) sırada Syracusa Üniversitesi öğretim üyesi Professor William Wasserman (1922-2007)'ın 1958 yılında TheAmericanStatisticianVol 12 (2), 1958, 16-18 adlı dergide yayınlanmış «The Teaching and Use of Statistics in Turkey» başlıklı makalesinden alıntı ile konuya girmek istiyorum. Wasserman (1958)'nın görüşüne göre "Türkiye'de Yönetim birimleri, endüstri ve çeşitli bilim dallarında istatistik yöntemlerin potansiyeli henüz bilinmemektedir. Küçük fakat gelişme eğilimi gösteren bir grup insan istatistik yöntemler hakkında bilgi sahibi ve istatistiği uyguluyorlar. 1953-1958 yılları arasında bazı fakültelerde istatistik dersleri verilmektedir, diğerlerinde yetişmiş öğretim üyesi olmadığından istatistik dersi verilememektedir. Türk eğitim sisteminde herhangi bir yerde istatistikte derece almış birini bulmak mümkün değildir. İstanbul Üniversitesi İktisat Fakültesinde öğrencilerin Olasılık ve İstatistiğe giriş dersleri almaktadırlar. Ayrıca üst sınıflarda Örnekleme ve Demografik İstatistik Analizi dersleri de bulunmaktadır. Ankara Üniversitesi Siyasal Bilgiler Fakültesinde tüm öğrencilerin aldıkları "Ekonomi için istatistik" dersi vardır. Ankara Üniversitesi Ziraat Fakültesinde de üçüncü sınıf öğrencileri için bir yarıyıllık derste olasılık ve istatistik bilgisine ek olarak Varyans Analizi ve Deneysel Tasarım dersi verilmektedir. İstanbul Robert Kolej'de de İstatistiğe Giriş dersi bulunmaktadır. Buralarda ki istatistik dersleri ABD'de eğitim görmüş öğretim elemanları tarafından verilmektedir. Bu tarihte İstanbul Teknik Üniversitesi, Ankara Üniversitesi Fen Fakültesi, İstanbul Üniversitesi Tıp Fakültesinde henüz İstatistik dersi yoktur.

Ülkemizde yarım yüzyıl öncesine kadar yönetim amaçları için gereksinim duyulan verinin toplanması ve tablolaştırılması için bazı kamu kuruluşlarında az sayıdaki çalışanın dışında istatistikçi denilen yetişmiş kişiler yoktu.

Günümüzde ise istatistik, alınacak kararların doğruluğunu desteklemek için kullanılan büyülü bir sözcük oldu.

21. yüzyılın ilk çeyreğinde ülkemiz istatistikçileri yönetim kadrolarında, sanayide ve araştırma organizasyonlarında çalışmaya başladılar. Üniversiteler, istatistiği ayrı bir disiplin olarak öğretmeye başladılar. Son 30 yıla damgasını vuran ve çağımızda bilgi çağı olarak adlandırılan gelişmeler istatistiği evrensel bir konuşma dili konumuna getirmiştir.

Ulusal düzeydeki bilimsel toplantılarımızda çok değişik alanlara yayılmış bilimsel çalışmalar bulunmaktadır. SCI, SCI–E ve SSCI kapsamındaki ve alan endekslerince taranan İstatistik dergilerinde özellikle 2000 yılından başlayarak genç istatistikçilerin umut veren katkıları olmasına karşın Türkiye adresli yayınlarımızın yeterli olmadığı açıktır. Birçok disiplin arasında İstatistikte Uluslararası (yalnız İstatistikle ilgili) saygın hakemli dergilerde yayın sayımızın arttırılması için bilimde öncü ülkelerin bilim insanlarının yaptıkları gibi kurumlar arası ve uluslararası işbirliği ile yayın yapma çabaları arttırılmalıdır.

İstatistikçiler fen ve mühendislikte olduğu gibi bilimin tüm alanlarındaki ilginç ve önemli problemlerle ilgilendiklerinden bu durum istatistiğe disiplinler arası bilim olma özelliği kazandırmıştır. "İSTATİSTİK ÜRETMEK KARANLIĞA IŞIK GÖTÜRMEK KADAR KUTSAL BİR GÖREVDİR" deyişinin önemini bir daha hatırlamalıyız..

8. 15. ULUSLARARASI EKONOMETRİ, YÖNEYLEM ARAŞTIRMASI VE İSTATİSTİK (15. EYİ) SEMPOZYUMU'NUN İSTATİSTİK AÇISINDAN DEĞERLENDİRİLİŞİ

15. EYİ Sempozyumu'nda 22 Ekonometri, 17 Yöneylem Araştırması, 19 İstatistik oturumu düzenlenmiştir. İstatistik oturumlarında sunulan bildiriler:

A) OLASILIK TEORİSİ
B) İSTATİSTİK TEORİSİ ve UYGULAMALI İSTATİSTİKSEL ANALİZ
C) İSTATİSTİĞİN UYGULAMASI (Biyoistatistik (tıp alanında), biyometri (ziraat alanında) ve ekonomi alanında)

olarak üç sınıfa ayrılabilir. Kısaca aşağıdaki konular sunulmuştur.

- İnternet bankacılığı,
- Yapısal eşitlik modeli,
- Mekansallık analizi,
- One-way ANOVA analizi,
- Sağlık göstergelerinin analizi,
- Sosyal ağ verilerinin olasılık analizi,
- Üniversite öğrencilerinin davranış ve tutum ölçeği kişilik ve liderlik özellikleri analizi,
- Bağıl değerlendirme sistemi,
- Özgüven ve mezuniyet analizi,
- Süreç yetenek analizi,
- Ölçüm hatalı regresyon modelleri,
- Kur modellemesi,
- Ranktransform metodu,
- Kredi kartlarının tasarrufa etkisi,
- Türkiye de 2023 hedefleri için istatistiğin katkısı,
- Gibbs örneklemesi,
- Meta analizi ve deneysel tasarım,
- Talep tahmininde Bayesçi yaklaşım,
- Sigara içme alışkanlıklarının analizi,
- Kanonik korelasyon analizi ve sağlıkta bir uygulama,
- Müze ziyaretçi profili ve memnuniyet araştırması ölçek tasarımı,
- Faktör analizi uygulaması,
- Temel bileşenler analizi ve uygulama,
- Kümeleme analizi ve uygulama,
- Kasko sigorta yaptırmada belirleyicilerin analizi,
- Rekabet gücü endekslerinin incelenmesi,
- Puan ve gol sayıları için Zaman serileri analizi,
- Ülkemizde enerji tüketim analizi,
- Uzaktan eğitim sisteminde internet tabanlı uygulamanın analizi,
- Path analizi (Gübre değişkenleri üzerinde),
- İstatistiksel süreç kontrolü,
- Poisson prosesi ile veri analizi,
- İki aşamalı Liu tahmin edici,

- Multicollinearity varken aykırı değer sorunu,
- Bulanık regresyon modeli ve uygulama,
- Kantil regresyon yöntemi ve OECD ülkelerinde beklen yaşam süresi analizi,
- Doğrusal olmayan eşitsizlik kısıtlamalı ridge tahmin edicisi,
- RobustBayesian regresyon analizi,
- Su kirliliği ve yoksulluk üzerine alan çalışması,
- Hava yolu ulaşım talebinin tahmini,
- Trafik kazalarının sayısının modellemesi,
- Hibrid sistemler için Bayesçi yaklaşım,
- Araçlarda arıza dağılım parametrelerinin incelenmesi,
- Yapay sinir ağları ile istatistiksel analiz,
- Oransal odds modeli ve performans karşılaştırması,
- Lojistik regresyon ve yapay sinir ağları yöntemleri ile insan gelişme endeksinin sınıflandırılması,
- Çalışma sermayesinin gıda sektöründeki işletmelerin finansal performansı üzerindeki etkisi,
- Liu tipi tahmin edici için test istatistiği,
- Küme örneklemesi,
- Gelen turist sayılarının modellemesi,
- Süt üretiminde modelleme,
- Lojistik regresyon modeli (organik gıda ve yabancı dil başarısı üzerinde) uygulama,
- EYİ ninbibliyometrik analizi,
- Bağıl değerlendirme sistemi,
- Kelime örüntülerinin analizi,
- Örnekleme yöntemlerinin irdelenmesi,
- İşaret levhalarının trafik işleyişindeki etkisi,
- 2011 milletvekili seçim sonuçlarının analizi,
- Yapısal eşitlik modelleme ile ONLINE alışverişlerde müşteri davranışları,
- İş kazalarının gelecek yıllar için tahmini,
- Rüzgar hızı verilerinin istatistiksel analizi.

15. EYİ sempozyum bildiriler kitabı incelendiğinde görülebileceği gibi sunulan bildirilerin %85'den fazlası (C) sınıfındadır. İstatistik bilim dalında uluslararası yayınların artması araştırmaların (A ya da B) sınıfında yer alması ile mümkün olabilmektedir. Ayrıca İstatistik bilim dalında büyük çoğunluğu sayfa başına ücret almayan 120 civarında SCI/SCI-E ya da SSCI kapsamında yer alan dergi vardır.Bu dergilerde yayınlanan makalelere yapılan atıflar Web of Science (WOS)'da yer almaktadır. Bunlara ek olarak alan endekslerinde yer alan hakemli kaliteli dergiler de vardır. Bu dergilerdeki yayınlar da dahil yayınlanan tüm makalelere yapılan atıflar GOOGLE AKADEMİK'te görülmektedir. Yurt içinde ortak çalışmalar yapılabildiği gibi yurt dışından da ortak araştırmayapabilecek bilim insanlarıyla ilişki kurulabilir. Gelişmiş ülkelerin araştırmacılarının yaptıkları yayınlardan kopmamaya özen göstermeliyiz.

9. İSTATİSTİĞİN GELECEĞİ NEDİR?

İstatistikçiler, fen ve mühendislikte olduğu gibi bilimin tüm alanlarındaki ilginç ve önemli problemlerle ilgilendiklerinden bu durum istatistiğe disiplinler arası bilim olma özelliği kazandırmaktadır.

Günümüzde istatistik; durmadan üretilen, araştırılan ve bulunan yeni yöntemlerle gelişen bir bilim dalıdır. İstatistik, diğer bilim dallarındaki karar verme mantığı ve metodolojisine sahiptir. İstatistikçilerin diğer bilim dallarındaki araştırmacılarla ilişkileri sonucunda bu alanlardaki temel problemlerin formülleştirilmesine katkılarıyla istatistik ilgi çekici bir araştırma konusu olmaya devam edecektir.

İstatistikçiler; bilimsel çalışmalarda sonuç almanın önemini bilen, değişik bilim dallarına ve topluma bu alanda yardımcı olabilecek ve gereksinimlerini karşılayabilecek uzmanlaşmış bireyler olacaktır. Böylece, uzmanlaşmış olan bireyler araştırmacı olarak; sosyal ve günlük yaşamın problemlerini çözmede, kaynakların optimum kullanılmasını sağlayarak ekonomik gelişmeye, sanayi üretiminin arttırılmasına, kişisel ve kurumsal düzeylerde optimum kararlar alınmasına önemli katkılarda bulunabilirler.

Bilim ve teknolojinin hakim olduğu 20. ve 21. yüzyılda istatistiğe aşinalığa gereksinim duyulacağını Ünlü İngiliz düşünürü H. G. Wells (1866-1946) önceden görerek şöyle demiştir: **"İstatistiksel düşünme, gün gelecek tıpkı okur yazar olmak gibi, iyi bir yurttaş olmanın en gerekli öğelerinden olacaktır."**

Sonuç olarak, düzeyli araştırmalarla gelecek için bilinçle bilgi üreterek kalıcı izler bırakacak biçimde bilim dünyasında yerimizi almalıyız.

Yazımızı ünlü Çin düşünürü KuanTzu'nun özlü deyişi ile tamamlamak istiyorum. 2600 yıl öncesinden diyor ki: **"Bir yıl sonrasını düşünüyorsan tohum ek, on yıl sonrasını düşünüyorsan ağaç dik; yüz yıl sonrasını düşünüyorsan insan yetiştir. Bir kez ürün verir ektiğin tohum; bir kez diktiğin ağaç on kez ürün verir; eğer insanı eğitirsen yüz kez olur bu ürün."** Bugün, en büyük yatırım insana yapılan yatırımdır. Genç bilim insanlarımızın anlamak ve araştırmak hırslarını uzun süre her şeyin üstünde tutarak, bilimin nabzının bilimsel dergilerde attığı gerçeğini unutmayarak, mutluluklarını orada bulabileceklerine inanıyorum. O halde **"geleceğin anahtarı iyi bir eğitim"** olduğundan duyarlılığı ve cesareti geliştirilen ve başarıyı yakalamak için aklı ve bilimi kullanan insanlar yetiştirmeliyiz.

Teşekkür: Makaleyi dikkatli bir biçimde değerlendirerek daha iyi olmasını sağlayan hakemlere ve şekilleri çizen Yard. Doç. Dr. Esra Akdeniz Duran'a teşekkkürlerimi sunarım.

10. KAYNAKLAR

Akdeniz, F., Dönmez, D., 1999. The History of Statistics in the Ottoman Empire, Chance , Vol 12, No.3, 37-39.

Hoerl, R. W., Snee, R. D., 2010a. "Closing the Gap: Statistical Engineering can Bridge Statistical Thinking with Methods and Tools", QualityProgress, May, 52-53.

Hoerl, R. W., Snee, R. D., 2010b. "Triedand True – Organizations put Statistical Engineering to the Test and See Real Results", QualityProgress, June, 58-60.

Rao, C. R., 1989. Statistics and Truth, International Co-Operative Publishing House, P.O.Box 245, Burtonsville, Maryland, USA.

Snee R. D., Hoerl, R. W., 2010. Further Explanation; Clarifying Points About Statistical Engineering Quality Progress Vol 43 (12), December, 68-72.

Snee, R. D., Hoerl, R. W., 2011. Proper Blending; The Right Mix Between Statistical Engineering and Applied Statistics, Quality Progress, June, 46–49.

Steiner, S. H., MacKay, R. J., 2014. Statistical Engineering and Variation Reduction, Quality Engineering 26(1), 44-60.

Wasserman, W., 1958. The Teaching and Use of Statistics in Turkey, The American Statistician Vol 12 (2) 16-18.

NEW TRENDS AND METHODS IN STATISTICS

ABSTRACT

In this study, statistical thinking and data analysis concepts are explained by emphasizing the importance of statistics. The role of data scientists converting data into knowledge is emphasized. In addition, the future of statistics is discussed and a road map for doing high level research to publish in international prestigious journals is given for the statisticians in our country.

Keywords: Data, Data analysis, Data scientist, Statistics discipline, Statistical thinking.

Volume: 10 Number: 03 Category: 02 Page: 49-63 ISSN No: 1303-6319 Cilt: 10 Sayı: 03 Kategori: 02 Sayfa: 49-63

IMPLEMENTATION OF REGRESSION MODELS FOR LONGITUDINAL COUNT DATA THROUGH SAS

Gül İNAN[*] **Özlem İLK** [**]

ABSTRACT

In this study, we firstly consider the marginal model and generalized linear mixed model classes for longitudinal count data and review the Log-Log-Gamma marginalized multilevel model, which combines the features of marginal models and generalized linear mixed models. Due to the special features of these models, implementation of them requires more special attention. As a consequence, this leads us to use SAS GENMOD procedure for the marginal model, SAS GLIMMIX procedure for the GLMM, and SAS NLMIXED procedure for the Log-Log-Gamma marginalized multilevel model. Since the latter model requires gamma distributed random effects, two different techniques, namely the probability integral transformation technique and likelihood reformulation technique , which are originally used for fitting Gamma Frailty models, are modified and adapted to fit Log-Log-Gamma marginalized multilevel model within the framework of Proc NLMIXED. Finally, we conclude the study with the discussion of the results obtained from the implementation of the models through popular epileptic seizures data.

Keywords: Epileptic seizure count, Gamma random effects, SAS GENMOD, SAS GLIMMIX, SAS NLMIXED.

1. INTRODUCTION

In longitudinal studies, measurements from the same subjects over a sequence of time periods are taken so that changes in measurements over time periods can be observed. In longitudinal count data (LCD), the response variable of the longitudinal dataset represents the counts of a total number of a defined event occurring in a given time interval. Examples from physiological research may include the number of epileptic seizures of each patient per two-weeks over an eight-week treatment period and the number of panic attacks for each patient in a week over a one-month psychological intervention program.

The analysis of longitudinal count data requires more special methods due to the longitudinal feature of measurements and counting process of responses. The most important feature of longitudinal data that motivates the statistical analysis is the association of measurements within a subject since the observations obtained from the same subject over several time periods are expected to be correlated. On the other hand, the statistical distribution of the counts is traditionally assumed to be Poisson distribution (Diggle et al., 2002) and it is well-known that the mean equals to the variance (equi-dispersion) for the Poisson distribution. However, when the variability of counts is greater than its expected value under the Poisson model, the phenomenon is

[*]Dr., Orta Doğu Teknik Üniversitesi, İstatistik Bölümü, Ankara, e-mail: ginan@metu.edu.tr
[**]Doç. Dr., Orta Doğu Teknik Üniversitesi, İstatistik Bölümü, Ankara, e-mail: oilk@metu.edu.tr

called overdispersion. More specifically, extra-Poisson variation occurs (Barron,1992). Although there are additional features that complicate the statistical analysis, these are the two that play a significant role in the estimation of regression parameters for the regression models developed for LCD.

In this sense, this paper aims to summarize the characteristics of the most commonly used general regression model classes, namely marginal models, random-effects models, and marginalized multilevel modelsto analyzelongitudinal count data and to show how these regression models are implemented through SAS, which is not well-documented in the literature, via the popular epileptic seizures example.On the other hand, the main contribution of this paper is that it provides the use of two different techniques to accommodate regression models with gamma distributed random effects for LCD, where non-gaussian random effects are not allowed within SAS framework.

The development of subsequent sections of this paper is organized as follows: Section 2 gives background information on regression model classes for LCD. Section 3 introduces the popular epileptic seizures example. Section 4 is devoted to the implementation of these regression models through the epileptic seizure example within SAS procedures. Section 5 discusses the results and Section 6 concludes the paper.

2. REGRESSION MODELS FOR LONGITUDINAL COUNT DATA

Diggle et al. (2002) classify the models for longitudinal data into three different regression model classes. These are: i) marginal models, ii) random-effects, and iii) transition models. In general, these three regression model classes view the association problem between the repeated measurements of a subject from different perspectives and this leads the models to differ in the interpretation of the regression parameters. In this paper, we restrict ourselves to the marginal and random-effects model classes and reintroduce the Log-Log-Gamma marginalized multilevel models (MMMs).

2.1 Marginal Models

Marginal models directly specify a regression model for the mean response, which depends only on covariates, using a log-link function**.** The mean responses, μ_{ij}^{C}, for the i^{th} subject and j^{th} time related to the covariates as follows:

$$log(\mu_{ij}^{C}) = X_{ij}\beta \qquad (1)$$

The within-subject association, the association between the repeated measurements of a subject, is modeled separately, possibly using additional association parameters**.** The regression parameters, β's, in equation (1) describe the effects of covariates on the population averaged mean response, as in cross-sectional analysis. Their interpretation is independent of specification of within-subject association model (Fitzmaurice and Molenberghs, 2008), which makes them more robust compared to the regression models that will be discussed later.

2.2 Random-Effects Models

The random-effects models assume that there is a natural heterogeneity between the subjects due to unmeasured covariates (Diggle et al., 2002). In this sense, regression parameters randomly varying from one subject to other subject are included into the regression modeling of the mean response. Contrary to the marginal models, GLMMs model the mean response and the within-subject association through a single equation and random effects are viewed as the potential source of within-subject association.

Among the random-effects models, generalized linear mixed models (GLMMs) are the most frequently used one for discrete repeated measurements (Molenberghs and Verbeke, 2005). In GLMMs, the model for the mean response depends both on covariates and random effects, which enter linearly into the linear predictor via a known link function. The simplest case of GLMMs is naturally a model with just a random intercept coefficient.

The formulation of a random-intercept model for LCD can be as follows:

i) Conditional Mean Model: $log(\mu_{ij}^{C}) = X_{ij}\beta + b_{0ij}$
ii) Random Intercept Distribution: $b_{0i} \sim MVN(0, C)$
iii) Conditional Response Distribution: $Y_{ij}^{C} = (Y_{ij}|b_{0ij}) \sim Poisson\ (\mu_{ij}^{C})$

Y_{ij}'s are assumed to be conditionally independent given subject-specific random intercepts,$b_{0i} = (b_{01i}, b_{02i}, \dots., b_{0in_i})'$ and to have Poisson distribution with conditional mean, μ_{ij}^{C}, depending on both fixed and random effects. The subject-specific random intercepts, $b_{0i} = (b_{01i}, b_{02i}, \dots, b_{0in_i})'$ are assumed to have a multivariate normal distribution with zero mean and a common within-subject covariance matrix, *C*.

One of the most important characteristics of GLMMs is that they have the ability to accommodate complex within-subject association structures for subject-specific random effects. Weiss (2005) lists a large number of covariance structures and detailed information on these covariance structure specifications, but among them, most commonly used ones are unstructured (UN), first order autoregressive (AR(1)), and compound symmetry (CS).

In GLMMs, the aim is to make inference on individual subjects rather than the population average; for that reason the fixed effects regression parameters, β's, in (i) describe the effects of covariates on an individual's mean response by controlling for the random-effects. However, interpretations being dependent on random effects and being sensitive to within-subject association specifications and robustness of estimates being dependent on the distribution of the random effects reflect the disadvantages of GLMMs (Heagerty and Zeger, 2000).

2.3 Log-Log-Gamma Marginalized Multilevel Model

Marginalized multilevel models are proposed by Heagerty and Zeger (2000). These models combine the features of marginal models and GLMMs with an aim to compensate the distinctions of these two models. While marginalized multilevel models take the interpretation and robustness of regression parameters from marginal models,

they take likelihood-based inference capabilities and flexible within-subject association specifications from GLMMs (Griswold and Zeger, 2004). Accordingly, Griswold and Zeger (2004) expand the marginalized multilevel model of Heagerty and Zeger (2000) for LCD and name this model as Log-Log-Gamma marginalized multilevel model (MMM).

The formulation of the Log-Log-Gamma MMM which assumes only a subject-specific intercept coefficient,b_{0i}, in the linear predictor, in addition to fixed effects, is as follows:

i) Marginal Mean Model: $log(\mu_{ij}^{M}) = X_{ij}\beta^{M}$

ii) Association Model: $log(\mu_{ij}^{C}) = \Delta_{ij} + b_{ij}$

iii) Random Effects Distribution:$g_{ij} \sim Gamma(1/\theta_1, \theta_2)$ where$b_{ij} = log(g_{ij})$

iv) Conditional Response Distribution: $Y_{ij}^{C} = (Y_{ij}|g_{ij}) \sim Poisson\ (\mu_{ij}^{C})$.

It defines a general linear model (GLM) for the marginal mean model in i) and a nonlinear mixed model (NLMM) for the within-subject association in ii).

Griswold and Zeger (2004) follow the same logic and assume a gamma distribution for subject-specific random effects and a Poisson distribution for the conditional response distribution, so that the marginal distribution of responses becomes negative-binomial distribution, which accommodates overdispersion well (Greenwood and Yule, 1920; Barron, 1992; Cameron and Trivedi, 1998; Jowaheer and Sutradhar, 2002). Contrary to GLMMs, subject-specific random effects in Log-Log-Gamma MMM follow a non-Gaussian distribution, that's Gamma distribution, and are allowed to enter nonlinearly into the model.

The log-link function and Poisson-gamma mixing distribution, together with the connection between marginal mean and conditional mean model, lead to $\Delta_{ij} = X_{ij}\beta^{M} - log\ (v_{ij})$where$v_{ij} = E(g_{ij}) = 1/\theta_1 \times \theta_2$(Griswold and Zeger, 2004). Hence, the conditional mean, μ_{ij}^{C}, can be written in terms of the marginal regression parameters, β^{M}, such that

$$\mu_{ij}^{C} = exp(\Delta_{ij} + b_{ij}) = exp(X_{ij}\beta^{M} - log(v_{ij}) + b_{ij}). \quad (2)$$

Since equation (2) includes the marginal regression parameters, β^{M}, the estimation of β^{M} can be performed by fitting the conditional model, μ_{ij}^{C}, via standard NLMM techniques. The regression parameters, β^{M}, describe the effects of covariates on the population averaged mean response, over the random effects.

3. EPILEPTIC SEIZURE COUNT DATA

The illustration of model fitting will be through an epileptic seizure count data, which is publicly availablein R package Mass (Venables and Ripley, 2002). We preferred this data set since it is the most commonly used one in the literature. This data comes from a randomized placebo-controlled clinical trial which was conducted by Leppik et al. (1985). 59 patients with simple or complex partial seizures were participated in the study and were randomized to receive either the antiepileptic drug progabide or a

placebo, as an adjuvant to the anti-epileptic standard chemotherapy. Before receiving treatment, the number of epileptic seizures of each patient over an eight-week period was recorded as baseline data. After treatment, the number of epileptic seizures of each patient per two-weeks over an eight-week treatment was also recorded at clinic visits. Apart from these, age information related to each patient was recorded as well. The question of interest is whether progabide has an effect in reducing the epileptic seizure counts or not.

The summary statistics for the epileptic seizure count data are displayed in Table 1. It is obvious that counts show overdispersion across visits within placebo group, progabide group, and complete data. When we do not take visits into account, the same case still continues, and counts exhibit high overdispersion in placebo group, progabide group, and complete data as an overall.

3.1 Covariates for Regression Models

To relate the covariates to the seizure counts, the covariates those listed in Thailand Vail (1990) are used. These are:

X_{1i}= logage=The natural logarithm of age in years, *log (Age)*,
X_{2i}= lgbsl=The natural logarithm of ¼ of the 8-week baseline counts,$log\ (Base/4)$,
X_{3i}= trt=Trt is a binary variable taking a value of 1 if progabide, 0 if placebo,
X_{4i}= v4=$Visit_4$ is a binary variable taking a value of 1 if visit number is 4, 0 otherwise,
X_{5i}= int=Interaction of Trt and $log\ (Base/4)$,

Here, β_3, which corresponds to the Trt variable, represent the parameter of interest for our research question.

Table 1. Summary statistics for epileptic seizure count data

	Placebo		Progabide		Complete	
	Mean	Variance / Mean	Mean	Variance / Mean	Mean	Variance / Mean
Visit 1	9.36	10.98	8.58	38.78	8.95	24.59
Visit 2	8.29	8.04	8.42	16.71	8.36	12.42
Visit 3	8.79	24.50	8.13	23.75	8.44	23.72
Visit 4	7.96	7.31	6.71	18.92	7.31	12.75
Overall	30.79	22.13	31.65	24.76	8.27	18.45

4. FITTING THE REGRESSION MODELS IN SAS

For model fitting of the regression models, SAS (version 9.2) is used.

4.1 Marginal Models

When the responses are discrete, i.e., binary or count, it is hard to estimate regression parameters of the marginal models by likelihood-based methods (Fitzmaurice and Molenberghs, 2008). That is because the complete joint distribution of longitudinal responses requires the specification of two-way associations between the responses and

in turn, building models for these associations that are consistent with the model for the mean response in an interpretable manner is difficult in the framework of marginal models (Lipsitz and Fitzmaurice, 2008).

When distributional assumption on repeated responses is avoided, an estimation method that is called generalized estimating equation (GEE) is considered. It is developed by Liang and Zeger (1986) by including additional parameters in the formulation of within-subject covariance matrix of responses. GEE provides as efficient estimates as maximum likelihood estimation (MLE), as well as consistent and asymptotically normal estimates provided that the mean response model is correctly specified. One disadvantage of GEE is that avoiding defining the complete joint distributions deprive us of using likelihood-based methods.

4.1.1 SAS GENMOD

SAS procedure that gives the opportunity to fit the GEE to repeated measures data is Proc GENMOD.

The marginal model equation for the epileptic seizure example can be given by

$$\mu = exp\left(\beta_0 + \beta_1 \times X_{log\,(Age)} + \beta_2 \times X_{log\,(\frac{Base}{4})} + \beta_3 \times X_{(Trt)} + \beta_4 \times X_{\left(Trt*log\left(\frac{Base}{4}\right)\right)} + \beta_5 \times X_{(Visit_4)}\right),$$

and the code related to our data and to our research question is given as follows:

```
procgenmoddata=seizure;
class id;
model count=logagelgbsltrtint v4 /dist=poissonlink=log scale=deviance;
repeated subject=id / type=UN;
run;
```

The model statement defines the relation between the response variable, count and the covariates logagelgbsltrtint v4, listed in Section 3. While dist option defines the distribution of counts, link option refers to the link function used in the model. On the other hand, scale=deviance enables the scale parameter to be fixed at 1 during estimation. Subjectthrough repeated statement identifies the subjects in the model and the variable identifying subjects should also be listed through the class statement. Finally, type refers to working correlation structure used in the model. SAS GENMOD allows user different working correlation structure types, such as unstructured, exchangeable and autoregressive AR (1).

4.2 Random-Intercept Model

When the interest is on the fixed effects regression parameters, β's, rather than random effects in the random-intercept model; the model fitting and inference on β's, requires the maximization of the likelihood of the data. This maximization is obtained by treating random intercepts, b_i's, as if they were nuisance parameters and by integrating over their distribution (Diggle et al., 2002). In other words, if the i^{th} subject's contribution to the likelihood of the data is defined as

$$L_i(\beta|\boldsymbol{Y_i}, \boldsymbol{b_i}) = \int_{b_i} \left[\left(\prod_{j=1}^{n_i} f_{ij}(y_{ij}|\beta, b_{ij}) \right) f(b_i|\theta) \right] db_i,$$

and then, the expression in equation (3)is expected to be maximized

$$L(\beta|\boldsymbol{Y}, \boldsymbol{b}) = \prod_{i=1}^{N} L_i(\beta|\boldsymbol{Y_i}, \boldsymbol{b_i})$$
$$= \prod_{i=1}^{N} \int_{b_i} \left[\left(\prod_{j=1}^{n_i} f_{ij}(y_{ij}|\beta, b_{ij}) \right) f(b_i|\theta) \right] db_i \quad (3)$$

where θ is the vector of parameters for the distribution of b_i.

In GLMMs, the distribution of random effects and high-dimensional integration of them together with a possibly nonlinear link function may cause computational difficulties in the evaluation of the likelihood and as consequence; closed-form solutions cannot be provided. In random-intercept models, being normal distribution not conjugate to Poisson distribution make the implementation of approximation harder.

Molenberghs and Verbeke (2005)divide the approaches toward the evaluation of the likelihood into three categories according to the frequency of usage and to the availability in statistical software. These are the approaches based on the approximation of i) the integrand, ii) data, and iii) integral itself. While Laplace–type approximations fall in the first category, penalized quasi-likelihood (PQL) and marginal quasi-likelihood (MQL) fall in the second category. The numerical integration methods such as adaptive and nonadaptive Gaussian quadrature fall in the latter category.

In this sense, SAS GLIMMIX procedure has the ability to fit the approximation and methods mentioned above.

4.2.1 SAS GLIMMIX

SAS GLIMMIX procedure is a built-in SAS procedure and is an appropriate choice for generalized linear mixed models, in which random effects are restricted to appear linearly in linear predictor. This procedure is especially recommended for models when the number of random effects per subject is large (Flom et al., 2006).

The random-intercept model equation for seizure data is given by

$$\mu = exp\Big(\beta_0 + \beta_1 \times X_{log\,(Age)} + \beta_2 \times X_{log\,(\frac{Base}{4})} + \beta_3 \times X_{(Trt)} + \beta_4 \times X_{\left(Trt*log\left(\frac{Base}{4}\right)\right)} + \beta_5 \times X_{(Visit_4)} + b\Big),$$

and $b \sim MVN_4(0, C)$ and C is assumed to be an unstructured within-subject covariance matrix.

SAS GLIMMIX code related to our data and to our research question can be given as follows:

```
procglimmix data=seizure MAXOPT=500 method=RSPL;
class id;
model count=logage lgbsl trt int v4 /dist=poisson link=log s;
random intercept /subject=id type=UN;
run;
```

Within the framework of Proc GLIMMIX, the random-intercept model is fitted by using PQL, based on REML for the linear mixed models. The option for PQL in **procglimmix** statement is the "method=RSPL", which is the default method. dist option through the model statement specifies the conditional distribution for the response variable given the random effects to come from any distribution in the exponential family. As in Proc GENMOD, link specifies the link function. interceptthrough random statement specifies a random intercept in the model. This procedure allows random effects to have only normal distribution and offer a straightforward fitting of a wide variety of within-subject covariance structures such as AR (1), CS and UN through type option.

4.3 The Log-Log-Gamma MMM

4.3.1 SAS NLMIXED

SAS NLMIXED procedure is a built-in SAS procedure and is preferred for the Log-Log-Gamma MM as in Griswold and Zeger (2004).

Proc NLMIXED is an appropriate choice for nonlinear mixed models, in which random effects are allowed to enter nonlinearly into the linear predictor of the model. It specifies the conditional distribution for the response variable given the random effects, either by standard distributions such as normal, binomial, and Poisson or by general distributions that can be coded using SAS statements. The only distribution available for random effects is normal distribution. The way of model specification in Proc NLMIXED has a high degree of flexibility, compared to other SAS procedures (Molenberghs and Verbeke, 2005). This advantage enables any non-normal distribution of interest for random effects to be implemented within the numerical integration techniques available in Proc NLMIXED via probability integral transformation(PIT) technique (Nelson et al., 2006) or likelihood reformulation (LR) technique (Liu and Yu, 2008). When the random effects are normally distributed, SAS NLMIXED procedure does not offer a straightforward option for the specification of any within-subject covariance structure. But, by the help of its flexibility, it is possible to allow the within-subject covariance matrix of the random effects to be, for instance, an AR(1), when specifying the mean and covariance components of the normal distribution (Molenberghs and Verbeke, 2005). Apart from these, Proc NLMIXED procedure requires the specification of initial values for all parameters in the model. Initial values for regression parameters can be obtained by the resulting parameter estimates after fitting a GLM in SAS.

In this sense, two different techniques, which Nelson et al. (2006) and Liu and Yu (2008) originally used for fitting Gamma Frailty models, are modified and adapted to fit

Log-Log-Gamma MMM by accommodating gamma distributed random effects within the framework of Proc NLMIXED.

1. PIT Technique by Nelson et al. (2006)

To accommodate gamma distributed random effects in Proc NLMIXED, we firstly use PIT technique proposed by Nelson et al. (2006). Similar to them, a_i is assumed to be a random effect from standard normal distribution, such that $a_i \sim N(0,1)$, and then by the use of PIT, it can be shown that $\Phi(a_i) = u_i \sim Unif(0,1)$ where $\Phi(.)$ is of the standard normal distribution. Again by the help of PIT, it can also be shown that $F_\theta(g_i) = u_i \sim Unif(0,1)$ where $F_\theta(.)$ is cumulative distribution function (CDF) of the gamma distribution of g_i, with $\theta = (1/\theta_1, \theta_2)$. For identifiably, θ_2 will be taken as equal to $\theta_1$on the forthcoming parts of the paper. Then it turns out that $g_i = F_\theta^{-1}(u_i) = F_\theta^{-1}(\Phi(a_i))$ has the gamma distribution of interest, where $F_\theta^{-1}(.)$is the inverse CDF of gamma distribution. Similarly, i^{th} subject's contribution to the likelihood of the data can be defined as in equation (4).

$$L_i(\beta|\boldsymbol{Y_i}, \boldsymbol{b_i}) = \int_{b_i} \left[\left(\prod_{j=1}^{n_i} f_{ij}(y_{ij}|\beta, b_{ij}) \right) f(b_i|\theta) \right] db_i, \tag{4}$$

where $b_i = log(g_i)$.

The expression in equation (5)**,** which is nowwritten in terms of random effects, a_i, is expected to be maximized such that

$$L(\beta|\boldsymbol{Y}, \boldsymbol{a}) = \prod_{i=1}^{N} L_i(\beta|\boldsymbol{Y_i}, \boldsymbol{a_i})$$
$$= \prod_{i=1}^{N} \int_{a_i} \left[\left(\prod_{j=1}^{n_i} f_{ij}\left(y_{ij}|\beta, F_\theta^{-1}(\Phi(a_i))\right) \right) \phi(a_i) \right] da_i \,, \tag{5}$$

where $\phi(.)$ is the standard normal distribution density function. Nelson et al. (2006) suggest that the likelihood in equation (5) can be approximated well by the Gaussian quadrature numerical integration technique. The approximation with Gaussian quadrature to integrals in equation (4) is achieved such that i^{th} subject's likelihood is approximated by a weighted sum

$$L_i(\beta|\boldsymbol{Y_i}, \boldsymbol{a_i}) = \int_{a_i} \left[\left(\prod_{j=1}^{n_i} f_{ij}\left(y_{ij}|\beta, F_\theta^{-1}(\Phi(a_i))\right) \right) \phi(a_i) \right] da_i$$
$$L_i(\beta|\boldsymbol{Y_i}, \boldsymbol{a_i}) \cong \sum_{q=1}^{Q} \left(\prod_{j=1}^{n_i} f_{ij}\left(y_{ij}|\beta, F_\theta^{-1}(\Phi(a_i))\right) \right) \phi(z_q) w_q,$$

and, thus, the likelihood in equation (5), which is expected to be maximized, turns out that

$$L(\beta|\boldsymbol{Y},\boldsymbol{a}) \cong \prod_{i=1}^{N}\sum_{q=1}^{Q}\prod_{j=1}^{n_i} f_{ij}\left(y_{ij}|\beta, F_{\theta}^{-1}(\Phi(a_i))\right)\phi(z_q)w_q,$$

where z_q is quadrature point and indexed by q = 1,...,Q, Q is the order of approximation, w_q is the standard Gauss-Hermite weight. Since the approximations will be more accurate as Q increases, we use Gaussian quadrature with 30 points like Griswold and Zeger (2004) and Nelson et al. (2006). The values of z_q and w_q can be obtained from tables in Abramowitz and Stegun(1972) (Table 25.10).

The Log-Log-Gamma MMM equation for seizure data is given by

$$\mu = exp\left(\beta_0 + \beta_1 \times X_{log\,(Age)} + \beta_2 \times X_{log\,(\frac{Base}{4})} + \beta_3 \times X_{(Trt)} + \beta_4 \times X_{\left(Trt*log\left(\frac{Base}{4}\right)\right)} + \beta_5 \times X_{(Visit_4)} + b\right),$$

and $b_{ij} = log(g_{ij}) \sim log - Gamma(1,1)$ or $e^{b_{ij}} = g_{ij} \sim Gamma(1,1)$.

SAS NLMIXED code by the help of PIT method that is related to our data and to our research question can be given as follows:

```
procnlmixed data=seizure noad fd qpoints=30;
PARMS theta1=1 beta0_m=-2.3492 beta1_m=0.7722 beta2_m=0.9582 beta3_m=-1.3299 beta4_m=-0.1565 beta5_m=0.5397;
eta_m=beta0_m + beta1_m*logage + beta2_m*lgbsl + beta3_m*trt + beta4_m*int + beta5_m*v4;
ui=CDF('Normal',ai);
if (ui >0.9999 ) then ui=0.9999;
gi1=quantile('GAMMA',ui,1/theta1,theta1);
v=1/theta1*theta1;
delta=eta_m-log(v);
eta_c=delta + log(gi1);
mu_c=exp(eta_c);
Model count ~ Poisson(mu_c);
Random ai ~ Normal(0,1) subject=id;
run;
```

noad in **procnlmixed** step refers to nonadaptive Gaussian quadrature. Finite difference approximation with fd is required for the derivative of CDF of normal distribution, that's CDF and the derivative of inverse CDF of gamma distribution, that's quantile. For that reason fd is there to specify that all derivatives to be computed using finite difference approximations. fd is equivalent to 100 as default and high fd values indicates better approximation. qpointsrefers to the number of quadrature points to be used during evaluation of integrals.PARMS statement allows to set the initial values for all unknown parameters in the model. The next eight SAS statements are used for defining Log-Log-Gamma MMM by PIT method. Model statementdefines the response variable and the form of the distribution of the conditional likelihood. Random statement declares the distribution of subject-specific random-intercept terms.

2. LR Technique by Liu and Yu (2008)

Another approach for accommodating gamma distributed random effects within the framework of the Proc NLMIXED is proposed by Liu and Yu (2008). This method aims to transform the formulation of likelihood that is conditional on non-normal random effects to a likelihood that is conditional on normal random effects in the framework of Gaussian quadrature. In this sense, they multiply and divide the likelihood in equation (6) by a standard normal density function, $\phi(.)$ such that

$$L_i(\beta|\boldsymbol{Y_i}, \boldsymbol{b_i}) = \int_{b_i} \left[\left(\prod_{j=1}^{n_i} f_{ij}(y_{ij}|\beta, b_{ij}) \right) f(b_i|\theta) \right] db_i \,, \tag{6}$$

$$L_i(\beta|\boldsymbol{Y_i}, \boldsymbol{b_i}) = \int_{b_i} \left[\left(\prod_{j=1}^{n_i} f_{ij}(y_{ij}|\beta, b_{ij}) \right) \frac{f(b_i|\theta)}{\phi(b_i)} \phi(b_i) \right] db_i \,,$$

$$L(\beta|\boldsymbol{Y_i}, \boldsymbol{b_i}) = \int_{b_i} \left[exp\left(log\left(\prod_{j=1}^{n_i} f_{ij}(y_{ij}|\beta, b_{ij}) \right) + log\big(f(b_i|\theta)\big) - log\big(\phi(b_i)\big) \right) \phi(b_i) \right] db_i \,,$$

$$= \int_{b_i} \left[exp\big(l_i^A + l_i^B - l_i^C\big) \phi(b_i) \right] db_i$$

where l_i^A is the conditional log-likelihood

$$l_i^A = log\left(\prod_{j=1}^{n_i} f_{ij}(y_{ij}|\beta, b_{ij}) \right) = log\left(\prod_{j=1}^{n_i} \left(\frac{e^{-\mu_{ij}^C} \mu_{ij}^{C\,Y_{ij}}}{y_{ij}!} \right) \right)$$

$$= -\sum_{j=1}^{n_i} exp\big(X_{ij}\beta^M - log\big(v_{ij}\big) + b_{ij}\big) + \sum_{j=1}^{n_i} y_{ij}\big(X_{ij}\beta^M - log\big(v_{ij}\big) + b_{ij}\big) - \sum_{j=1}^{n_i} log\big(y_{ij}!\big)$$

l_i^B is the log of the $log - Gamma \sim (1/\theta_1\,, \theta_2)$

$$l_i^B = log\big(f(b_i|\theta)\big) = -\frac{1}{\theta_1} log(\theta_1) - log\,\Gamma\left(\frac{1}{\theta_1}\right) + \frac{b_i}{\theta_1} - \frac{exp(b_i)}{\theta_1},$$

and l_i^C is the log of the standard normal distribution

$$l_i^C = log\big(\phi(b_i)\big) = -0.5 b_{ij}^2 + constant.$$

SAS NLMIXED code by the help of LR technique that is related to our data and to our research question can be given as follows:

```
procnlmixed data=seizure qpoints=30;
PARMS theta1=1 beta0_m=-2.3492 beta1_m=0.7722 beta2_m=0.9582 beta3_m=-
1.3299 beta4_m=-0.1565 beta5_m=0.5397;
eta_m=beta0_m + beta1_m*logage + beta2_m*lgbsl + beta3_m*trt + beta4_m*int +
beta5_m*v4;
v=1/theta1*theta1;
eta_c=eta_m-log(v)+ b;
mu_c=exp(eta_c);
expb=exp(b);
fc=fact(count);
loglik=-mu_c+count*eta_c-log(fc);
if lastid=1 then do;
lB=-((1/theta1)*log(theta1))-lgamma(1/theta1)+((1/theta1)*b)-((1/theta1)*expb);
lC=-1/2*(b**2);
loglik=loglik+lB-lC;
end;
Model count ~ general(loglik);
Random b ~ Normal(0,1) subject=id;
run;
```

Contrary to nonadaptive Gaussian quadrature**,**theadaptive Gaussian quadrature considers the shape of the likelihood when placing quadrature points and this result in better approximations (Liu and Yu,2008). For that reason, this technique prefers adaptive Gaussian quadrature contrary to PIT method which is the default option in **procnlmixed.** The next four SAS statements after the PARMS statement are there to specify the Log-Log-Gamma MMM. Similarly, Model statementshows the response variable and the form of the distribution of the conditional likelihood but this time through a general log-likelihood.

Contrary to PIT technique which requires the inverse CDF to have a closed form or to be available in SAS, LR technique requires that distribution function of the non-normal random effect to have a closed form or to be available in SAS. Further information on the description of the SAS NLMIXED and SAS GLIMMIX procedures and their options can be obtained from SAS (2000).

5. FINDINGS

Table 2 displays the regression parameter estimates and corresponding standard errors produced from the models and estimation methods mentioned above through the epileptic seizure data.

We find that results from four methods are similar except the estimates of regression parameter,β_4**.** Large differences are observed in this parameter between the regression models. It is found that the treatment effect has a statistically significant effect on the number of seizures count. As the negative sign on $\beta_3$indicates, the treatment reduces the seizure numbers.

Table 2. Results from the Marginal Model (ProcGenmod), Random-Intercept Model (Proc GLIMMIX), Log-Log-Gamma MMM by PIT (Proc NLMIXED) and Log-Log-Gamma MMM by LR (Proc NLMIXED)

Parameter	Marginal Model	Random-Intercept Model	Log-Log-Gamma MMM by PIT	Log-Log-Gamma MMM by LR
β_0	-2.5426 (0.9051)	-0.8776 (1.1217)	-1.1020 (0.5663)	-0.7594 (1.0863)
β_1	0.8417 (0.2608)	0.3558 (0.3259)	0.3556 (0.1680)	0.3378 (0.3195)
β_2	0.9455 (0.0931)	0.8780 (0.1369)	1.0774 (0.0823)	0.8926 (0.1273)
β_3	-1.4867 (0.4425)	-0.8671 (0.4139)	-1.4672 (0.2920)	-0.8173 (0.3826)
β_4	0.6019 (0.1789)	0.2984 (0.2096)	0.7044 (0.1013)	0.2971 (0.1914)
β_5	-0.1520 (0.0822)	-0.1565 (0.0544)	-0.1565 (0.0545)	-0.1565 (0.0545)

When we compare, the Log-Log-Gamma MMM by PIT and that by the LR method in terms of computational time, it is observed that LR method with adaptive Gaussian quadrature with 30 points option reduces the computational time considerably compared to PIT method with non-adaptive Gaussian quadrature 30 points option. While the estimation time takes approximately 2 seconds in LR technique, it takes about 22 seconds in PIT technique. This is due to that the LR technique does not need any finite difference approximation; hence it reduces implementation duration considerably.

6. CONCLUSION

This paper summarizes the marginal and random-effects model classes dealing with longitudinal count data in the literature and the implementation of longitudinal count data within SAS. One regression model class that is not mentioned in this paper is the transition models, but interested reader is kindly invited to read the Diggle et al. (2002). Diggle et al. (2002) reviews three different models and discusses the models with their pluses and minuses.

We especially focus on the Log-Log-Gamma marginalized multilevel model, which was developed by Griswold and Zeger (2004). This model is a likelihood-based model and offers a GLM for the mean response model, and a nonlinear mixed model for the within-subject association model. Separation of the model for mean response from that for within-subject association eases the interpretation of regression parameters of interest. Moreover, the Log-Log-Gamma MMM specifies a gamma distribution for the random effects which is conjugate to the Poisson distribution of conditional mean model. This is a great advantage over normally distributed random effects model since the Poisson-gamma mixture is able to remedy the overdispersion problem. As Nelson et al. (2006) stresses, non-normal random effects are taking progressive attention not only from longitudinal data analysis field, but also from different areas in statistics, and are more realistic than normally distributed random effects. However, non-normal random effects within the nonlinear mixed models suffer from the lack of computational implementation in the literature. In this sense, the main contribution of this paper is to show how a regression model with gamma distributed random effects, contrary to normally distributed random effects, can be handled within SAS, where non-Gaussian random effects are not allowed. We hope that the proposed algorithm would be helpful

for statisticians who work on models with non-Gaussian random effects and who would like to implement those models through user-specified algorithms within a standard software, i.e. SAS.

7. REFERENCES

Abramowitz, M., Stegun, I., 1972. Handbook of Mathematical Functions. Dover, New York.

Barron, D. N., 1992. The Analysis of count data: Overdispersion and Autocorrelation. Sociological Methodology, 22:179-220.

Cameron, A. C., Trivedi, P. K., 1998. Regression Analysis of Count Data. Econometric Society Monographs. Cambridge University Press, New York.

Diggle, P. J., Heagerty, P., Liang, K. Y., Zeger, S., 2002. Analysis of Longitudinal Data. Oxford University Press.

Fitzmaurice, G., Molenberghs, G., 2008. Advances in longitudinal data analysis: A historical perspective, in *Longitudinal Data Analysis: A Handbook of Modern Statistical Methods*, eds. G. Fitzmaurice, M. Davidian, G. Molenberghs, G. Verbeke, pp 3-27. Boca Raton, FL: Chapman & Hall/CRC Press, Florida.

Flom, P. L., McMahon, J. M., Pouget, E. R., 2006. Using PROC NLMIXED and PROC GLMMIX to analyze dyadic data with binary outcomes.Northeast SAS Users Group (NESUG) Proceedings, SAS Inc., Cary, NC.

Greenwood, M., Yule, G.U., 1920. An Inquiry into the Nature of Frequency Distributions Representative of Multiple Happenings with Particular Reference to the Occurrence of Multiple Attacks of Disease or of Repeated Accidents. Journal of the Royal Statistical Society, 83:255-279.

Griswold, M. E., Zeger, S. L., 2004. On Marginalized Multilevel Models and their Computation. The Johns Hopkins University, Department of Biostatistics Working Papers.

Heagerty, P. J., Zeger, S. L. 2000. Marginalized multilevel models and likelihood inference. Statistical Science, 15(1):1-26.

Jowaheer, V., Sutradhar, B. C., 2002. Analyzing longitudinal count data with overdispersion. Biometrika, 89(2):389-399.

Leppik, I. E., Dreifuss, F. E., Bowman-Cloyd, T., Santilli, N., Jacobs, M., Crosby, C., Cloyd, J., Stockman, J., Graves, N., Sutula, T., Welty, T., Vickery, J., Brundage, R., Gumnit, R., Gutierres, A., 1985. A double-blind crossover evaluation of progabide in partial seizures. Neurology, 35:285.

Liang, K. Y., Zeger, S. L., 1986. Longitudinal data analysis using generalized linear models. Biometrika, 73:13-22.

Lipsitz, S., Fitzmaurice, G., 2008. Generalized estimating equations for longitudinal data analysis. In *Longitudinal Data Analysis: A Handbook of Modern Statistical Methods*, eds. G. Fitzmaurice, M. Davidian, G. Molenberghs, G. Verbeke, pp 43-78. Boca Raton, FL: Chapman & Hall/CRC Press, Florida.

Liu, L., Yu, Z., 2008. A likelihood reformulation method in non-normal random effects models. Statistics in medicine, 27:3105-3124.

Molenberghs, G.,Verbeke, G., 2005. Models for Discrete Longitudinal Data. Springer, New York.

Nelson, K. P., Lipsitz, S. R., Fitzmaurice, G. M., Ibrahim, J., Parzen, M., Strawderman, R., 2006. Use of the Probability Integral Transformation to Fit Nonlinear Mixed-Effects Models With Nonnormal Random Effects. Journal of Computational & Graphical Statistics, 15(1):39-57.

SAS/STAT User's Guide, Version 8, Chapter 46 (SAS Institute Inc.,Cary, NC, 2000).

Venables, W. N., Ripley, B. D. 2002. Modern Applied Statistics with S. Fourth Edition. Springer, New York.

Weiss, R. E., 2005. Modeling longitudinal data. Springer, New York.

UZUNLAMASINA KESİKLİ VERİLER İÇİN REGRESYON MODELLERİNİN SAS İLE UYGULANMASI

ÖZET

Bu çalışmada, öncelikle, uzunlamasına kesikli veri için marjinal model ve genelleştirilmiş lineer karma model sınıflarını ele alacağız ve sonra marjinal ve genelleştirilmiş lineer karma modellerinin özelliklerini birleştiren, Log-Log-Gamma marjinalleştirilmiş çok seviyeli modellerini yeniden gözden geçireceğiz. Bu modellerin bilgisayar ortamına aktarılması onların sahip olduğu birtakım özelliklerden dolayı, dikkat gerektirmektedir. Bu nedenden dolayı, bu durum marjinal modeller için SAS GENMOD, GLMM için SAS GLIMMIX ve Log-Log-Gamma marjinalleştirilmiş çok seviyeli modelleriiçin de SAS NLMIXED prosedürünü kullanmamıza öncülük etmektedir. Son model, gamma dağılımlı rassal etkiler içerdiğinden, ilk olarak Gamma Frailty modelleri için kullanılmış olan iki farklı yöntem, isim vermek gerekirse, olasılık integral dönüşümü ve olabilirlik yeniden formülasyonu yöntemleri değiştirilerek, Log-Log-Gamma marjinalleştirilmiş çok seviyeli modelleri için PROC NLMIXED prosedürü çerçevesinde uyarlanmıştır. Son olarak, çalışmamızı bu modellerin popüler epilepsi nöbet sayısı verisine uygulanmasından elde edilen sonuçları tartışarak bitirmekteyiz.

Anahtar Kelimeler: Epilepsi nöbet sayısı, Gamma dağılımlı rassal etkiler, SAS GENMOD, SAS GLIMMIX, SAS NLMIXED.

Cilt: 10 Sayı: 03 Kategori: 03 Sayfa: 64-75 ISSN No: 1303-6319 Volume: 10 Number: 03 Category: 03 Page: 64-75

A K-NEAREST NEIGHBOR BASED APPROACH FOR DETERMINING THE WEIGHT RESTRICTIONS IN DATA ENVELOPMENT ANALYSIS

Elvan AKTURK HAYAT* **Olcay ALPAY****

ABSTRACT

Data Envelopment Analysis (DEA), a method commonly used to measure the efficiency is becoming an increasingly popular management tool. On the contrary to classical efficiency approaches, the most important advantage of DEA is that researchers can determine the weight restrictions of input and output variables. Variable selection and determination of weight restrictions are important issues in DEA. This work investigates the use of K-nearest neighbor (KNN) algorithm in the definition of weight restrictions for DEA. With this purpose a new approach based on KNN is proposed. Applications are constructed with empirical and real data sets depending on the specific constraints. Performance scores were calculated for both KNN based restricted and unrestricted DEA models and the results are interpreted.

Keywords: Data envelopment analysis, Efficiency, K-nearest neighbor, Weight restrictions.

1. INTRODUCTION

Data Envelopment Analysis (DEA) is a nonparametric technique for measuring the relative efficiency of a set of similar units, usually called the Decision Making Unit (DMU), which use a variety of identical inputs to produce a variety of identical outputs. DEA based on Frontier Analysis was introduced by Farrell in 1957, but the recent series of discussions started with the article by Charnes et al. (Charnes et al., 1978).

DEA provides efficiency score through linear programming when there are multiple inputs and outputs. One of the most important differences of DEA from the other efficiency measurement models is allowance to use input-output weights. In recent years, weight restrictions and value judgments have become one of the major issues in the DEA literature. The traditional DEA formulation allows for unrestricted model weights, which may result in inadequate weight values (zero, for instance, implying that a variable with relevance to the model would not be used for parameter estimation) (Gonçalves et.al., 2013). To deal with this kind of problem, Thompson et al. (1986) were the first to propose the use of weight restrictions in DEA. Many methods for estimating restrictions for the DEA weights have been developed by several researchers in the area, including Charnes et al. (1979); Charnes et al. (1985); Golany (1988); Thompson et al. (1990); Roll et al. (1991); Thanassoulis et al. (1995); Podinovski and Athanassopoulos (1998); Thanassoulis and Allen (1998); Podinovski (1999, 2001,

*Yrd. Doç. Dr. Sinop University, Faculty of Arts and Sciences, Department of Statistics, 57000, Sinop, Turkey, e-mail: elvanhayat@sinop.edu.tr
**Yrd. Doç. Dr. Sinop University, Faculty of Arts and Sciences, Department of Statistics, 57000, Sinop, Turkey, e-mail: olcayb@sinop.edu.tr

2004); Zhu (2003); Allen and Thanassoulis (2004). For a detailed review of such methods, see Allen et al. (1997) and Sarrico and Dyson (2004).

Jahanshahloo et al. used goal programming and Big M method techniques to obtain feasible weights for DMU's in 2005. Dimitrov and Sutton (2010) proposed symmetric weight assignment technique (SWAT) which does not affect feasibility. Mecit and Alp (2012) used correlation coefficients to determine the weights of inputs and outputs and also they compared this new method with cross efficiency evaluation model.

In our study, we proposed the use of a weight restriction technique based on the K-nearest neighbor (KNN) algorithm in order to define variation limits for the DEA model parameters. The KNN based restricted model is applied to three data sets and the results are compared with the unrestricted model.

The rest of paper is organized as follows. In the next section, the basic DEA model and the weight restrictions in DEA, also the proposed approach are briefly explained. In Section 3, the proposed approach and classical DEA model are applied to three data sets and the application results are reported. The first two data sets are from Roll et al.'s (1991) and Beasley's (1990) studies; the last one is a real data related to Turkey health system in 2013. Some concluding remarks are given in the final section.

2. METHODOLOGY

2.1 Data Envelopment Analysis

DEA does not require any assumptions about the functional form of the production function. In the simplest case of a unit having a single input and output, efficiency is defined as output/input. Charnes, Cooper and Rhodes, who developed Farrell's idea, extended the single-output/input ratio measure of efficiency to the multiple output/input measure of efficiency (Cooper et al., 2000).

The efficiency score in the presence of multiple input and output factors is defined as:

$$Efficiency = \frac{weighted\ sum\ of\ outputs}{weighted\ sum\ of\ inputs}$$

The first DEA model was introduced by Charnes et al. in 1978, known as the CCR model. This model measures the total efficiency under the assumption of constant returns to scale (CRS).

CCR is a linear program measuring the efficiencies of DMUs with respect to weighted inputs and outputs (Charnes et al., 1978). The model did not have any restrictions on the weights of inputs and outputs and found the optimal combination of weights that maximizes the efficiency score (Cooper et al., 1996).

Assume that there are n DMUs, each with m inputs and s outputs. The relative efficiency score of a DMU_p is obtained by solving the following proposed model (Charnes et al., 1994).

$$\max \frac{\sum_{k=1}^{s} v_k y_{kp}}{\sum_{j=1}^{m} u_j x_{jp}}$$

$$s.t. \quad \frac{\sum_{k=1}^{s} v_k y_{ki}}{\sum_{j=1}^{m} u_j x_{ji}} \leq 1 \qquad \forall i \tag{1}$$

$$\mathrm{v_k}, u_j \geq 0 \qquad \forall k,j$$

where,

$k = 1,2,..., s$
$j = 1,2,...,m$
$i = 1,2,...,n$

$y_{ki} =$ amount of output k produced by DMU i,
$x_{ji} =$ amount of input j utilized by DMU i,
$v_k =$ weight given to output k,
$u_j =$ weight given to input j.

The fractional program shown as (1) can be converted to a linear program as given in (2).

$$\max \sum_{k=1}^{s} v_k y_{kp}$$

$$s.t. \sum_{j=1}^{m} u_j x_{jp} = 1$$

$$\sum_{k=1}^{s} v_k y_{ki} - \sum_{j=1}^{m} u_j x_{ji} \leq 0 \qquad \forall i \tag{2}$$

$$\mathrm{v_k}, u_j \geq 0 \qquad \forall k,j.$$

The above problem is run n times in identifying the relative efficiency scores of all the DMUs. Each DMU selects input and output weights that maximize its efficiency score. In general, a DMU is considered to be efficient if it obtains a score of 1 and if it has a score of less than 1, it is implied as inefficient.

The weights given by the DEA model may be inconsistent with prior knowledge or accepted views on the relative values of the outputs and the inputs. DEA model can assign lower or higher weights to some inputs and/or outputs than they actually are. The model can give high weights for some inputs and/or outputs which give the impression that these attributes are over represented. As a result, the relative efficiency of a DMU

may not really reflect its performance on the inputs and outputs taken as a whole (Talaue et al., 2011).

In recent years, many new kinds of methods were proposed for weight restrictions such as analytic hierarchy process (AHP) and Delphi. A common characteristic of these approaches is based on specialists' own experiences and subjective judgment, to determine each of the indices that will be used to evaluate. The main disadvantage of this approach is that it is subjective (Allen et al., 1997).

Wong and Beasley (1990) proposed the use of proportions to introduce restrictions in the virtual inputs and outputs, seeking to make the quantification of value judgments easier for decision makers. Thus, they could set weights as varying, for instance, between 10% and 90% of the total contribution of inputs and outputs.

To constitute the weight restrictions in DEA, some methods such as assurance regions type, cone-ratio and absolute weight restriction were developed. In this research, assurance regions method is used to determine the weight restrictions.

2.2 KNN Based Algorithm

Many data mining techniques are based on similarity measures between objects. Measures of similarity may be obtained indirectly from vectors of measurements or characteristics describing each object (Hand et al., 2001).

The KNN prediction model simply stores the entire data set. As the name implies, to predict for a new observation, the predictor finds the *k* observations in the training data with feature vectors close to the one for which we wish to predict the outcome (Ye, 2003). Many applications of nearest neighbor methods adopt a Euclidean metric. Euclidean distance between i^{th} and j^{th} objects is defined as follows (Hand et al., 2001):

$$d_{ij} = \sqrt{\sum_{k=1}^{p} (x_{ik} - x_{jk})^2}$$

The nearest neighbor method has several attractive properties. It is easy to program and no optimization or training is required. Its classification accuracy can be very good on some problems, comparing favorably with alternative more unfamiliar methods (Hand et al., 2001). Also, nearest neighbor methods are very simple and therefore suitable for extremely large data sets (Felici and Vercellis, 2008). From a theoretical perspective, the nearest neighbor method is a valuable tool: as the design sample size increases, the bias of the estimated probability will decrease for fixed k (Hand et al., 2001).

In this study, we present a new algorithm based on KNN to determine the weight restriction matrices:

Step-0: Each of input/output numbers must be greater than 2.
Step-1: Determine the input and output variables.
Step-2: Construct the distance matrices using KNN for inputs and outputs.
Step-3: Find min and max values of matrices.
Step-4: Compute the decimal scale bandwidths which include all values in distance matrices.

Step-5: Determine the weight matrices according to bandwidths (in step-4) and relative to each other rates of variables.
Step-6: Construct the constraints with assurance regions method and calculate the efficiencies by DEA.

3. APPLICATION

In this section, three applications are illustrated and DEA is performed using LINDO (Linear, Interactive, and Discrete Optimizer) program. In the first application, we use the data of Roll et al. (1991), in the latter we use the data of 52 universities in Beasley (1990). In the last, we use the selected health statistics data for 12 statistical regions in Turkey. CCR model is used to calculate the efficiency scores and assurance region method is used to compose the restrictions obtained from our KNN based approach. Also, we determine the weight restricted models and compare the efficiencies with unrestricted models.

Roll et al.'s (1991) data consists of 10 DMUs with 3 inputs and 2 outputs. Table 1 summarizes the input and output variables for DMUs. The Euclidean distance matrix and the weight restrictions table for inputs are calculated by proposed KNN based algorithm. Finally, efficiency scores are calculated with DEA and given in Table 2. According to Table 2, 4 DMU reached 100% of efficiency in the unrestricted model, but 3 DMU achieved 100% in the proposed model. In unrestricted model, 2 DMU and in the proposed model 3 DMUs efficiency had less than 70%.

Table 1. Input-output variables (from Roll et. al.'s (1991) study)

DMU	I1	I2	I3	O1	O2
1	1.00	0.80	5.40	0.90	7.00
2	1.50	1.00	4.80	1.00	9.50
3	1.20	2.10	5.10	0.80	7.50
4	1.00	0.60	4.20	0.90	9.00
5	1.80	0.50	6.0	0.70	8.00
6	0.70	0.90	5.20	1.00	5.00
7	1.00	0.30	5.00	0.80	7.00
8	1.20	1.50	5.50	0.75	7.50
9	1.40	1.80	5.70	0.65	5.50
10	0.80	0.90	4.50	0.85	9.00

Table 2. Efficiency scores (%) for unrestricted and proposed model

	Efficiency (%)	
DMU	Unrestricted Model	Proposed Model
1	84.7	80.1
2	97.2	93.2
3	73.4	68.3
4	100.0	98.2
5	82.9	78.6
6	100.0	100.0
7	100.0	100.0
8	66.0	64.1
9	53.2	49.6
10	100.0	100.0

In Beasley's 1990 data set, there were 3 inputs and 8 outputs for 52 DMUs. Kocakoç (2003) used the same data set to determine the constraints for weight restrictions with analytic hierarchy process (AHP). We determine the weight restricted model and compare the efficiencies with AHP and unrestricted models. KNN based algorithm is performed on this data set, and the Euclidean distance matrices and the weight restriction tables are constructed for inputs and outputs. In Table 3 efficiency scores computed for 3 models are given. In the unrestricted model, there is no difference between the 52 DMUs in terms of efficiency. The results obtained from AHP and proposed models are quite similar. In the AHP model 39^{th} and 41^{th} DMU reached 100% of efficiency, whereas in the proposed model only 39^{th} achieved 100%. Average efficiency score is 71.52% in the AHP model, 69.65% in the proposed model.

In the last application, the selected health statistics data related to 12 statistical regions in Turkey consists of 3 inputs and 3 outputs. Physician number (per 100.000), beds number (per 10.000) and inpatient number (%) were taken as the inputs. Also, operation number (per 1000), mortality rate and average hospitalization days were taken as the outputs. Table 4, summarizes the input and output variables for DMUs. Calculated efficiency scores by without restrictions and proposed model are given in Table 5. In considering the efficiency scores obtained by each model, the average efficiency score of unrestricted model and our proposed model is 98.91% and 87.58%, respectively.

Table 3. Efficiency scores (%) for unrestricted, AHP and proposed models

	Efficiency (%)		
DMU	Unrestricted Model	AHP	Proposed Model
University 1	100	65.80	63.19
University 2	100	88.23	89.98
University 3	100	69.17	71.37
University 4	100	65.95	63.51
University 5	100	66.47	63.15
University 6	100	86.13	88.49
University 7	100	63.00	65.04
University 8	100	64.37	58.92
University 9	100	77.53	74.91
University 10	100	62.37	59.26
University 11	100	95.41	91.40
University 12	100	73.29	71.79
University 13	100	70.50	59.22
University 14	100	63.09	58.56
University 15	100	75.90	70.79
University 16	100	59.83	61.02
University 17	100	56.58	55.82
University 18	100	81.42	83.60
University 19	100	68.63	69.25
University 20	100	36.05	31.66
University 21	100	67.92	59.24
University 22	100	68.58	64.76
University 23	100	64.70	61.40
University 24	100	56.70	54.36
University 25	100	58.63	57.38
University 26	100	60.24	59.56
University 27	100	62.38	61.12
University 28	100	78.01	77.58
University 29	100	62.18	58.70
University 30	100	82.02	82.01
University 31	100	89.04	85.02
University 32	100	76.68	75.37
University 33	100	56.18	44.85
University 34	100	74.44	75.10
University 35	100	82.23	80.63
University 36	100	79.01	75.63
University 37	100	66.68	61.87
University 38	100	69.33	65.34
University 39	100	100.00	100.00
University 40	100	68.09	69.48
University 41	100	100.00	99.20
University 42	100	82.55	86.06
University 43	100	74.48	73.68
University 44	100	76.40	72.77
University 45	100	69.35	67.82
University 46	100	52.66	55.15
University 47	100	87.84	84.85
University 48	100	71.07	72.24
University 49	100	78.95	83.64
University 50	100	78.37	73.79
University 51	100	60.71	62.62
University 52	100	74.03	69.72

Table 4. Input-output variables for statistical regions in Turkey

	Inputs			Outputs		
Regions	**# Physician (per 100.000)**	**# Beds (per 10.000)**	**# Inpatient (%)**	**# Operation (per 1000)**	**Mortality rate**	**Average hospitalization days**
Akdeniz	161	23.8	54	66.2	15.5	4.2
Ege	191	27.4	60	61.2	18.7	4.4
Batı Anadolu	274	34.4	53	77.3	16.9	5.0
Güneydoğu Anadolu	124	20.2	63	52.9	11.2	3.4
Batı Karadeniz	156	29.8	67	56.5	18.2	4.9
İstanbul	184	23.4	43	59.6	15.3	5.0
Kuzeydoğu Anadolu	148	29.5	68	56.0	10.7	4.2
Batı Marmara	154	27.2	66	48.1	21.3	4.3
Ortadoğu Anadolu	146	27.7	60	53.6	6.4	4.1
Doğu Karadeniz	160	32.6	64	58.7	19.3	4.7
Doğu Marmara	160	25.8	61	62.5	18.0	4.2
Orta Anadolu	164	27.6	54	65.2	13.2	3.9

Resource: T.C. Minister of Health, Health Statistics Year Book, 2013

Table 5. Efficiency scores (%) for unrestricted and proposed model

	Efficiency (%)	
Regions (DMUs)	**Unrestricted Model**	**Proposed Model**
Akdeniz	100	100
Ege	97	80
Batı Anadolu	100	70
Güneydoğu Anadolu	100	100
Batı Karadeniz	100	88
İstanbul	100	81
Kuzeydoğu Anadolu	96	89
Batı Marmara	100	79
Ortadoğu Anadolu	96	86
Doğu Karadeniz	100	89
Doğu Marmara	100	95
Orta Anadolu	98	94

4. CONCLUSION

In this study, a new approach is proposed for determining the weight restrictions in DEA without any information or expert opinion about constraints. This approach is based on using K-nearest neighbor method establishing of constraint conditions and has several advantages. Firstly, this is a new kind of approach to determine the weight restrictions; it's easy to implement as well. Another advantage of this model is that it does not require expert opinions or value judgments.

Applications are performed to demonstrate the use of the proposed model and calculated efficiency scores for unrestricted model and the proposed model with using different data sets. The first data set, which consists of 10 DMUs with 3 inputs and 2 outputs, is obtained by Roll et al. (1991) study. The second data set from Beasley (1990) consists of 3 inputs and 8 outputs for 52 DMUs. Lastly, in real data application we used the selected health statistics for 12 DMUs in Turkey. As it can be seen from the results of application, the efficiency scores obtained from the proposed and restricted model based on AHP are quite similar. Thus, our proposed model can identify these restrictions objectively if there is no pre-information about weight restrictions. Undoubtedly, it cannot be expected to obtain such results in each time and every application. In a future study the real performance of our model can be evaluated by using simulation study.

5. REFERENCES

Allen R., Athanassopoulos A., Dyson R. G., Thanassoulis, E., 1997. Weights Restrictions and Value Judgements in Data Envelopment Analysis: Evolution, Development and Future Directions, Annals of Operations Research, 73:13 – 34.

Allen R., Thanassoulis E., 2004. Improving Envelopment in Data Envelopment Analysis, European Journal of Operational Research, 154, 363–79.

Beasley J. E., 1990. Comparing University Departments, Omega, International Journal of Management Science 18:171 - 183.

Charnes A., Cooper W. W., Rhodes E., 1978. Measuring the Efficiency of Decision Making Units, European Journal of Operational Research, Vol.2 No: 6, 429-444.

Charnes A., Cooper W. W., Rhodes E., 1979. Short Communication: Measuring the Efficiency of Decision Making Units, European Journal of Operational Research, 3, 339.

Charnes A., Cooper W. W., Golany B., Seiford, L., 1985. Foundations of Data Envelopment Analysis for Pareto-Koopmans Efficient Empirical Production Functions, Journal of Econometrics, 30, 91–108.

Charnes A., Cooper W. W., Lewin A. Y., Seiford L. M. (Eds.)., 1994. Data Envelopment Analysis: Theory, Methodology, and Applications, Boston: Kluwer.

Cooper W. W., Thompson R. G., Thrall R. M., 1996. Introduction: Extensions and New Developments in Data Envelopment Analysis, Annals of Operations Research 66(1):1-45.

Cooper, W. W., Seiford, L. M., Tone, K., 2000. Data Envelopment Analysis, Kluwer Academic Publishers, Boston, USA.

Farrell M. J., 1957. The Measurement of Productive Efficiency, Journal of the Royal Statistical Society, Series A, 120:253–81.

Dimitrov S., Sutton W., 2010. Promoting symmetric weight selection in data envelopment analysis: A penalty function approach, European Journal of Operational Research, 200, 281-288.

Felici G., Vercellis C., 2008. Mathematical Methods for Knowledge Discovery and Data Mining, Information Science Reference, Hershey, New York.

Golany B., 1988. A Note on Including Ordinal Relation among Multipliers in DEA, Management Science, 34, 1029–33.

Gonçalves A. C., Almeida R. M. R. V., Lins M. P. E., Samanez C. P., 2013. Canonical Correlation Analysis in the Definition of Weight Restrictions for Data Envelopment Analysis.

Hand D. J., Mannila H., Smyth P., 2001. Principles of Data Mining, MIT Press.

Jahanshahloo G. R., Memariani A., Hosseinzadeh F., Shoja N., 2005. A feasible interval for weights in data envelopment analysis, Applied Mathematics and Computation, 160, 155–168.

Kocakoç İ. D., 2003. Veri Zarflama Analizi'ndeki Ağırlık Kısıtlamalarının Belirlenmesinde Analitik Hiyerarşi Sürecinin Kullanımı, D.E.Ü.İ.İ.B.F.Dergisi 18(2):1-12.

Mecit E. D., Alp İ., 2012. A New Restricted Model Using Correlation Coefficients as an Alternative to Cross-Efficiency Evaluation in Data Envelopment Analysis, Hacettepe Journal of Mathematics and Statistics, 41(2), 321-335.

Podinovski V. V., Athanassopoulos A. D., 1998. Assessing the Relative Efficiency of Decision Making Units using DEA Models with Weight Restrictions, The Journal of the Operational Research Society, 49, 500–08.

Podinovski V. V., 1999. Side Effects of Absolute Weight Bounds in DEA Models, European Journal of Operational Research, 115, 583–95.

Podinovski V. V., 2001. DEA Models for the Explicit Maximisation of Relative Efficiency, European Journal of Operational Research, 131, 572–86.

Podinovski V. V., 2004. Production Trade-offs and Weight Restrictions in Data Envelopment Analysis, The Journal of the Operational Research Society, 55,1311–22.

Roll Y., Cook W. D., Golany B., 1991. Controlling Factor Weights in Data Envelopment Analysis, IIE Transactions, 23:1, 2-9.

Sarrico C. S., Dyson R. G., 2004. Restricting Virtural Weights in Data Envelopment Analysis, European Journal of Operational Research, 159, 17–34.

Thanassoulis E., Boussofiane A., Dyson R. G., 1995. Exploring Output Quality Targets in the Provision of Perinatal Care in England using Data Envelopment Analysis, European Journal of Operational Research, 80, 588.

Thompson R. G., Singleton, F. D., Thrall R. M., Smith B. A., 1986, Comparative site evaluations for locating a high-energy physics lab in Texas, Interfaces 16:35-49.

Thompson R. G., Langemeier L. N., Lee C. T., Lee E., Thrall R. M., 1990. The Role of Multiplier Bounds in Efficiency Analysis with Application to Kansas Farming, Journal of Econometrics, 46, 93–108.

Talaue C. O., Diesta N. A. N., Tapia C. G., 2011. Weights Restriction by Multiple Decision Makers in Data Envelopment Analysis Using Fuzzy Programming, Proceedings of the 11th Philippine Computing Science Congress, Ateneo de Naga University.

Ye N., 2003. The Handbook of Data Mining, Lawrence Erlbaum Associates, Publishers Mahwah, New Jersey, London.

Zhu J., 2003. Imprecise Data Envelopment Analysis (IDEA): A Review and Improvement with an Application, European Journal of Operational Research, 144, 513–29.

VERİ ZARFLAMA ANALİZİNDE AĞIRLIK KISITLARININ BELİRLENMESİNDE K-EN YAKIN KOMŞULUĞA DAYALI BİR YAKLAŞIM

ÖZET

Genellikle etkinlik ölçümünde kullanılan Veri Zarflama Analizi (VZA), popüler bir yönetim aracı olmaya başlamıştır. Klasik etkinlik yaklaşımlarının tersine, VZA'nın en önemli avantajı, girdi ve çıktı değişkenlerinin ağırlık kısıtlarını araştırmacıların belirleyebilmesidir. Değişken seçimi ve ağırlık kısıtlarının belirlenmesi VZA' da önemli konulardır. Bu çalışma VZA için ağırlık kısıtlarının tanımlanmasında K-en yakın komşuluk algoritmasının kullanımını araştırmaktadır. Bu amaçla K-en yakın komşuluk temeline dayanan yeni bir yaklaşım önerilmiştir. Belirlenen kısıtlara bağlı olarak ampirik ve gerçek veri setleri ile uygulamalar yapılmıştır. K-en yakın komşu temelinde kısıtlı model ve ağılık kısıtlamasız VZA modeli için performans skorları hesaplanmıştır ve sonuçlar yorumlanmıştır.

Anahtar Kelimeler: Ağırlık kısıtları, Etkinlik, Veri zarflama analizi, K-en yakın komşuluk.

DANIŞMA KURULU ÜYELERİ - ADVISORY BOARD MEMBERS

Ali YAZICI	Atılım Üniversitesi
Alper GÜVEL	Çukurova Üniversitesi
Asaf Savaş AKAT	Bilgi Üniversitesi
Aşır GENÇ	Selçuk Üniversitesi
Aydın ÖZTÜRK	İzmir Üniversitesi
Ayşe GÜNDÜZ HOŞGÖR	Orta Doğu Teknik Üniversitesi
Bedriye SARAÇOĞLU	Gazi Üniversitesi
Coşkun Can AKTAN	Dokuz Eylül Üniversitesi
Deniz GÖKÇE	Bahçeşehir Üniversitesi
Ekrem ERDEM	Erciyes Üniversitesi
Ercan UYGUR	Türkiye Ekonomi Kurumu
Erdem BAŞCI	T.C. Merkez Bankası
Erinç YELDAN	Bilkent Üniversitesi
Erol TAYMAZ	Orta Doğu Teknik Üniversitesi
Eser KARAKAŞ	Bahçeşehir Üniversitesi
Fatih ÖZATAY	TOBB Ekonomi ve Teknoloji Üniversitesi
Fatin SEZGİN	Bilkent Üniversitesi
Fikri AKDENİZ	Çukurova Üniversitesi
Fikri ÖZTÜRK	Ankara Üniversitesi
Gülay BAŞARIR KIROĞLU	Mimar Sinan Güzel Sanatlar Üniversitesi
Güven SAK	TOBB Ekonomi ve Teknoloji Üniversitesi
Haluk LEVENT	Galatasaray Üniversitesi
Hamza EROL	Abdullah Gül Üniversitesi
İlhan TEKELİ	Orta Doğu Teknik Üniversitesi
İmdat KARA	Başkent Üniversitesi
İnsan TUNALI	Koç Üniversitesi
Levent KANDİLLER	Yaşar Üniversitesi
Mehmet KAYTAZ	Işık Üniversitesi
Meltem DAYIOĞLU TAYFUR	Orta Doğu Teknik Üniversitesi
Metin TOPRAK	İstanbul Üniversitesi
Mustafa ACAR	Aksaray Üniversitesi
Mustafa AYTAÇ	Uludağ Üniversitesi
Nihat BOZDAĞ	Gazi Üniversitesi
Orhan GÜVENEN	Bilkent Üniversitesi
Ömer Faruk ÇOLAK	Gazi Üniversitesi
Ömer L. GEBİZLİOĞLU	Kadir Has Üniversitesi
Özkan ÜNVER	Ufuk Üniversitesi
Öztaş AYHAN	Orta Doğu Teknik Üniversitesi
Reşat KASAP	Gazi Üniversitesi
Savaş ALPAY	SESRTCIC
Seyfettin GÜRSOY	Bahçeşehir Üniversitesi
Süleyman GÜNAY	Hacettepe Üniversitesi
Turan EROL	Ankara Strateji Enstitüsü
Ümit OKTAY FIRAT	Marmara Üniversitesi
Yasin AKTAY	Stratejik Düşünce Enstitüsü
Yılmaz AKDİ	Ankara Üniversitesi