Uludağ Üniversitesi
İktisadi ve İdari Bilimler Fakültesi Dergisi
Cilt XXII, Sayı 1, 2003, s. 271-279

ZAMAN SERİLERİNDE SAPAN DEĞER TARAMA SÜREÇLERİNİ MODELLEME ve PERFORMANS ANALİZİ

Ahmet KAYA*

Özet

*Sapan değerler, Parametre tahminlerinin yanlı olmasına neden olan, kişilerin, aletlerin hatalarından veya hatalı kullanımları ile oluşabilen ya da doğal rasgelelik sonucunda ortaya çıkabilen az sayıda gözlem değeridir. Geçmişte tanıma dayalı olarak tespit edilebilen sapan değerler, günümüzde özellikle ARIMA modellerde otokorelasyona dayalı yöntemler ile tespit edilmektedir. Bu çalışma zaman serilerinde, aynı anda test yöntemleri olarak bilinen **LS** (**L**east **S**quare), **M-H** (**M**ethod of **H**uber-Type), **M-B** (**M**ethod of **B**isquare-Type), **GM-H** (**M**ethod of **G**eneralized **H**uber Type), **GM-B** (**M**ethod of **G**eneralized **B**isquare-Type) ve **ITERATIVE** (Ardışık) yöntemlerin performansları, sapan değer sayıları (1, 2 veya 3), ve sapan değer türleri (AO ve IO) özel seçimli, çok-etkenli **(2x3x6)** deney düzeni için tasarlanmış ve varyans analizi tekniği ile test edilmiştir. Araştırma sonucunda etkenler önemli bulunmuştur.*

Summary

Investigated in this study which is called outliers have been regarded as the observations which are extremely different from the others in a sample one, two or more than two observations which effect the estimation and forecasting process in a great extent.

The data used in this study were obtained from the simulation experiments performed on several time series model by Chang, Tiao and Chen The original simulation results were reorganized and remodeled under a suitable experimental design. Two different time series outliers types (AO, IO) in three different size (1, 2

* *Yard. Doç. Dr.; Ege Üniversitesi Tire Kutsan Meslek Yüksekokulu, İzmir.*

or 3) and six different detection methods (LS, M-H, M-B, GM-H, GM-B and iterative) were arranged. The type of factorial design was that is 3 factors each has 1 levels and 1 factor at 3 levels each.

The result of analysis of variance performed for this design indicated that the main effect except that of sensitivity coefficient were statistically significant.

Keywords: *ARIMA, AO, IO, Outlier, Detection Methods.*

1. GİRİŞ

Zaman serilerinde sapan değerler kavramı ise ilk defa Fox (1972) tarafından çalışılmıştır. Fox, otoregresif modellerde sapan değerlerin tespiti amacıyla olabilirlik oran kriteri adını verdiği bir ölçüt geliştirmiş ve ortaya çıkan sapan değerleri birinci ve ikinci tip sapan değerler biçiminde tanımlamıştır. Ayrıca Fox, güç fonksiyonları üzerine çalışmıştır. Bundan sonra bir çok araştırmacı, Fox'un yaptığı çalışmaları daha da geliştirerek, Bütün ARIMA modelleri kapsayacak şekilde, çoklu sapan değerleri taramaya yönelik yöntemler geliştirmişlerdir. Tarama yöntemlerinden etkin olanı belirlemek amacıyla simülasyon çalışmaları yapılmıştır. Bu tür çalışmalarla birlikte; Chang (1982), Chang ve Tiao (1983), ayrıca Hillmer (1983) Tsay (1986), Pena (1987), Abraham ve Yatawara (1988), Bruce ve Martin (1989) tarama yöntemlerinin teorik yapısı üzerinde çalışmalar yapmıştır.

Bununla birlikte Abraham ve Yatawara (1988) lagrange çarpanlar metodu veya skor tabanlı sapan değer testleri üzerine çalışmalar yapmışlardır. Pena (1987), Abraham ve Chuang (1989) ve Bruce ve Martin (1989) sapan değer taramada, sapan olan gözlemin silinmesi esasına dayalı testler ile birlikte zaman serilerinde etkili gözlemler üzerine çalışmalar yapmışlardır (Ljung, 1993: 1-2).

Diğer yandan Harrison ve Stevens (1976), Smith ve West (1983), West, Harrigon ve Migon (1985) ve West (1986) zaman serilerinde sapan değerler elde etmek için sıralı tarama (sequential detecting) yöntemleri üzerinde çalışmışlardır. Sıralı tarama çalışmaları, doğrusal regresyon problemleri içinde de çalışılmıştır. Bu amaçla, Mararsinghe (1985), Kianiard ve Swallow (1990) tarafından sıralı strateji testi adıyla bir yöntem geliştirilmiştir (Pena ve Yohai, 1995: 1-10).

Ayrıca, Jones (1980), Ljung (1982, 1989), Harvey ve Pierce (1984), Kohn ve Ansley (1986), Wincek ve Reinsel (1986) ve diğerleri, gözlem değerlerinden bazılarının kayıp olması durumunda, sapan değerlerin teşhis edilmesine yönelik olasılık oran kriterleri geliştirmişlerdir (Ljung, 1993:3).

Zaman serilerinde sapan değerlerin etkisi üzerine çalışmalar, ilk defa Box & Tiao (1965) tarafından çalışılmıştır. Box ve Tiao yaklaşımı, Chen ve

Tiao (1986) tarafından, Chang ve Tiao (1983)'ün iteratif yöntemi kulanılarak genelleştirilmiştir. Son zamanlarda ise Tsay, buna yeni eklemeler yapmıştır. Birinci dereceden AR modellerde sapan değer etkileri ise Whichern, Miller ve Hsu (1976) tarafından yapılmıştır. Bu araştırmacılar, varyans değişimlerinin parametre değerleri üzerindeki etkisini incelemişlerdir (Tsay, 1988: 1-8).

Zaman serilerinde sapan değerlerin teşhis edilmesine yönelik çalışmalardan biri de Robust Procedure (Güçlü Yöntem) adı verilen, Denby ve Martin (1979) tarafından geliştirilen yöntemdir. Bu yöntem Martin ve Yohai (1985) tarafından da çalışılmıştır. Ancak Chang ve Tiao (1983) söz konusu yöntemin ikinci tip sapan değerlerin teşhis edilmesinde etkin olmadığını ortaya koymuştur Bruce ve Martin (1989), birinci tip sapan değerlerin tahminlenmesine yönelik çalışmışlarında, sapan değerlerin sadece bulunduğu pozisyonu değil, yakın pozisyonları da etkilediğini ortaya koymuştur (Tsay, 1988: 1-3).

Chang (1988) ise, yanlış belirlenen sapan değer tiplerinin, test yöntemlerinde etkinlik kaybına neden olduğunu ortaya koymuştur (Muirhead, 1986: 3).

Sapan değerleri tespit etmede kritik değer olarak kullanılan C (ölçüt değer) değerlerini saptamak amacıyla Monte-Carlo simülasyonları kullanılmıştır. Ayrıca, Berman (1964) tarafından bulunan asimptotik sonuçların da kullanılabilir olduğu yine Berman tarafından ispat edilmiştir. (Ljung 1993: 3).

2. ARMA MODEL

Durağan zaman serilerinin modellenmesinde kullanılan model *p* terimli AR ve *q* terimli MA modellerinin bir kombinasyonudur. $p + q + 2$ parametreli model aşağıdaki tanımlı genel gösterimi ile ifade edilir. $\{z_t\}$, ARMA (*p, q*) modeli ile türetilmiş sapan değer içermeyen zaman serisi olsun. ARMA (*p, q*) modeli,

$$\phi(B)z_t = \theta(B)e_t \qquad (2.1)$$

biçiminde ifade edilir. Burada,

$$\phi(B) = 1 - \phi_1 B - ... - \phi_p B^p, \quad \theta(B) = 1 - \theta_1 B - ... - \theta_q B^q, \quad B^k z_t = z_{t-k}$$

$E(z_t) = 0$ ve $\{e_t\}$ $N(0, \sigma^2)$ dağılmaktadır. (2.1) modelinde tanımlanan eşitlikte $\phi(B)$ polinomun kökleri $\phi_1, \ldots, \phi_p$ değerleri birim çemberin dışında kalıyor-

sa model durağanlık, θ(B) polinomun kökleri $\theta_1, \ldots, \theta_p$ değerleri birim çemberin dışında kalıyorsa çevrilebilirlik koşulunu sağladığı ifade edilir (Box-Jenkins, 1976).

3. SAPAN DEĞERLER

Sapan değerler kavramı, bir çok araştırmacı tarafından farklı biçimlerde tanımlanmıştır. Mesela sapan değer, Bross (1961) tarafından, "Görünümü geri kalan gözlem değerlerinin görünümüne uymayan bir veya bir kaç gözlem" olarak tanımlanmaktadır. Stefansky (1972), ise "kestirim değerlerinden çok farklı olan gözlem değerleri" şeklinde bir tanım vermiştir. Aynı tanım, "örneklemdeki diğer gözlemlerden farklı olan ve kestirim sonuçlarını büyük ölçüde etkileyen az sayıda gözlem değeri şeklinde de verilmiştir." Bu ve buna benzer bir çok tanım verilebilir. Ancak bir gözlemin sapan olabilmesi değer olarak büyük ya da küçük olmasıyla ilişkilendirilemez. Gerçek model denklemi elde edildiğinde büyük hata kestirimine sahip olan gözlem veya gözlemlerin sapan olmaları olanaklıdır. Bu durumu göz önünde bulundurarak, "Gözlem ve kestirim değerleri arasındaki farkın en büyük olduğu gözlem, sapan gözlem olabilir" denilebilir.

Sapan değerler, model yanlışlığı, gerekli bazı dönüşümlerin yapılmamış olması, ölçüm, tartım, ya da kaydetme hataları gibi çeşitli nedenlerden kaynaklanabilmektedir Örneklemdeki bazı gözlemlerin, diğer gözlemlerden farklı bir etkenin ya da etkenlerin etkisinde kalmış olması da sapan değer oluşturabilmektedir. Bu değerler sadece doğal rasgeleliğin sonucunda da ortaya çıkabilmektedir (Tatlıdil, 1981: 1).

Kalite kontrolde, sürecin kontrol altında olup olmadığını belirlemek amacıyla, süreçten elde edilen gözlemlerin bir zaman serisi oluşturduğu düşünülerek serinin uyum gösterdiği parametre kestirimleri elde edilmekte, daha sonra seride ortaya çıkması olası iki tip sapan değer belirlenmektedir. Sapan değer tiplerine göre problem; girdi, çıktı ya da süreç problemi olarak sınıflandırılmaktadır. (Bayhan, 1992: 17-21).

Tek bir gözlemi etkileyen, bir gözlem veya kayıtta yapılan hata sonucu ortaya çıkan sapan değer tipine birinci tip sapan değer (Additive Outlier), bulunduğu pozisyondan itibaren kendinden sonraki gözlemleri de etkileyen sapan değer tipine ikinci tip sapan değer (Innovational Outlier) denir (Fox. 1972: 1).

Birinci tip sapan değer (Additive Outlier-AO) modeli,

$$y_t = z_t + \delta x_t \qquad (3.1)$$

Birinci tip sapan etki,

$$\delta = y_T - z_T \tag{3.2}$$

Burada y_t, gözlenen değer, δ, sapan değerin büyüklüğü, x_t, sapan değer anında (T = t) 1, aksi halde 0 değeri alan değişkendir.

İkinci tip sapan değer (Innovational Outlier-IO) modeli,

$$y_t = \frac{\theta(B)}{\phi(B)}(e_t + \delta x_t) \tag{3.3}$$

İkinci tip sapan etki,

$$\delta = \frac{\phi(B)}{\theta(B)} y_T - e_T \tag{3.4}$$

Birinci tip sapan değerde sapan değerin bulunduğu pozisyonda büyük bir hata söz konusudur, İkinci tip sapan değer beklenmeyen şok bir durumu ifade eder ve T pozisyonunda meydana gelen şok, z_T, z_{T+1},…, $\varphi(B) = \theta(B)/\phi(B)$'e kadar devam eder. İkinci tip sapan değer (IO), birinci tip sapan değer (AO)'e göre nispeten küçük bir etki yapar (Ljung, 1993 : 1-2).

4. TARAMA SÜREÇLERİ

Seri içerisinde olması muhtemel bütün sapan değerlerin aynı anda testine olanak sağlayan ve hata terimleri arasındaki otokorelasyonu kullanan test yöntemleri, özellikle zaman serileri gözlem değerleri için kullanılan yöntemlerdir. Bu yöntemler aşağıda verilmiştir.

LS (Least Squares), M-H (Method of Huber Type), M-B (Method of Bisquare Type), GM-H (Method of Generalized Huber Type), GM-B (Method of Generalized Bisquare Type), Iterative (Ardışık Yöntem)'den oluşmuştur.

5. VARYANS ANALİZİ

Chang Tiao ve Chen (1988) tarafından yapılmış simülasyon sonuçları kullanılarak; sapan değer taramada; Sapan değer **T**ürü (**T**) (AO, IO) sapan değer sayısı (**A**dedi) (**A**) (1, 2,3) ve sapan değer tarama **S**üreçleri (**S**) (LS, M-H, M-B, GM-H, GM-B, Ardışık) gibi faktörlerin ne derecede etkili olduğunu istatistiksel olarak ortaya koymaya yönelik varyans analizi çalışması yapılmıştır.

Faktör Ana Etkileri ;

Sapan Değer Türü (T),

Sapan Değer Sayısı (Adedi) (A),

Tarama Süreçleri (S).

Faktörler için ikili ve üçlü etkileşimler; T*A, T*S, A*S, T*A*S.

Özel seçimli, çok-etkenli (**2x3x6)** deney düzeni için tasarlanmış model denklemi aşağıdaki gibidir:

$$Y_{ijk} = \mu + T_i + A_j + S_k + (TA)_{ij} + (TS)_{ik} + (AS)_{jk} + \varepsilon_{ijk} \qquad (5.1)$$

$i = 1,2; j = 1,2,3; k = 1,2,3,4,5,6.$

Y_{ijk} : Sapan değer tespiti ile oluşan hata kareler ortalaması.

μ : Genel ortalama.

(5.1) modeli için hata teriminin tahminlenmesine olanak vermek amacıyla faktörlere ilişkin üçlü etkileşim modele dahil edilmemiş ve sıfır kabul edilmiştir. Bir başka deyişle üçlü etkileşim ile açıklanan değişim hata terimi içerisine dahil edilmiştir.

Tablo 1. Tarama Süreçlerine Bağlı Elde Edilen Hata Kareler Ortalamaları

Sapan Değer	ADET	LS	M-H	M-B	GM-H	GM-B	Ardışık
AO	1	0.0306	0.0185	0.0183	0.0119	0.0100	0.0079*
	2	0.0371	0.0237	0.0245	0.0144	0.0127	0.0089*
	3	0.0665	0.0470	0.0502	0.0242	0.0202	0.0120*
IO	1	0.0071	0.0056	0.0055	0.0055	0.0056	0.0049*
	2	0.0058	0.0046	0.0045	0.0047	0.0054	0.0044*
	3	0.0334	0.0203	0.0217	0.0115	0.0099	0.0052*

Tablo 2. (5.1) Modeli Varyans Analizi Sonuçları

D.K.	S.D.	K.T.	K.O.	F	P
T (Tür)	1	0.002070	0.002070	662.02	0.000
A(Adet)	2	0.001837	0.000918	293.65	0.000
S(Süreç)	5	0.002129	0.000426	136.15	0.000
T*A	2	0.000127	0.000063	20.24	0.000
T*S	5	0.000671	0.000134	42.92	0.000
A*S	10	0.000787	0.000079	25.16	0.000
Hata	10	0.000031	0.000003		
Toplam	35	0.007652			

Eldeki veri seti için yapılan varyans analizi işlemi ve elde edilen varyans analizi tablosu ile sapan değerlerin Tür (T), sapan değer sayıları (A) ve tarama süreçleri (S) arasındaki ilişki ele alınmış, etkileri incelenmiştir. Sözü edilen üç faktör için iki'li etkileşimler de incelenerek sonuçlar elde edilmiştir. Tablo 3'te verilen varyans analizi faktör ana etkilerine ilişkin sonuçlar incelendiğinde;

Tablo 3. Faktör Ana Etkilerine İlişkin Varyans Analizi Sonuçları

ETKEN		**Ortalama**	**P**
Tür	**AO**	0.0244	
	IO	0.0092	0.000
Adet	**1**	0.0109	
	2	0.0125	
	3	0.0268	0.000
Süreç	**LS**	0.0300	
	M-H	0.0199	
	M-B	0.0208	
	GM-H	0.0120	
	GM-B	0.0106	
	Ardışık	0.0072	0.000

- Hata kareler ortalaması sonuçlarına göre sapan değer Türü'nün (AO, IO) (T) büyük oranda önemli bir faktör olduğu görülmektedir.
- İkinci faktör olarak belirlenen sapan değer sayısının (1, 2, 3) (A) önemli bulunduğu, dolaysıyla seri içerisinde sapan değer sayısı arttıkça tarama süreçlerinin sapan değer tespit gücü önemli oranda zayıflamaktadır.
- Üçüncü faktör, tarama süreçlerinin performansı ile ilgili olanıdır. Tarama süreçlerine ilişkin performans analiz edildiğinde, süreçler arası farkın istatistik olarak önemli bulunduğu görülmektedir. Bu üç faktör arasında oluşturulan ikili etkileşimler de önemli bulunmuştur. Etkileşimler için ayrıntılı dökümler, hata kareler ortalaması bazında aşağıda Tablo 4, Tablo 5 ve Tablo 6'da ifade edilmiştir.

Tablo 4. Hata Kareler Ortalaması Verileri için Tür x Adet (T*A) Etkileşimi

		Adet			Ortalama
		1	2	3	
Tür	AO	0.0162	0.0202	0.0367	0.0244
	IO	0.0057	0.0049	0.0170	0.0092
	Ortalama	0.0109	0.0125	0.0268	0.0168

(Tür x Adet) etkileşim tablosundan görülebileceği gibi sapan değer türleri (AO, IO) arasında ortalama bazında meydana gelen farklılık, sapan değer sayısının artmasıyla belirginleşmiş Bunun sonucunda ikili etkileşim önemli bulunmuştur.

Tablo 5. Hata Kareler Ortalaması Verileri için Tür x Süreç (T*S) Etkileşimi

		Süreç						Ortalama
		LS	M-H	M-B	GM-H	GM-B	ARD	
Tür	AO	0.0447	0.0297	0.0310	0.0168	0.0143	0.0096	0.0244
	IO	0.0154	0.0101	0.0106	0.0072	0.0069	0.0048	0.0092
	Ortalama	0.0300	0.0199	0.0208	0.0120	0.0106	0.0072	0.0168

(Tür x Süreç) etkileşim tablosu incelendiğinde hata kareler ortalaması bazında minimum değerler, ardışık yöntemin bulunduğu sütunda gözlenmektedir. Sapan değer türlerinden AO'nun ortaya çıkması hata kareler ortalamasında önemli artışlara sebep olmaktadır.

Tablo 6. Hata Kareler Ortalaması Verileri için Adet x Süreç (A*S) Etkileşimi

		Süreç						Ort.
		LS	M-H	M-B	GM-H	GM-B	ARD	
Adet	1	0.0188	0.0120	0.0119	0.0087	0.0078	0.0064	0.0109
	2	0.0214	0.0141	0.0145	0.0095	0.0090	0.0066	0.0125
	3	0.0499	0.0336	0.0359	0.0178	0.0150	0.0086	0.0268
	Ort	0.0300	0.0199	0.0208	0.0120	0.0106	0.0072	0.0168

(Adet x Süreç) etkileşim tablosu, sapan değer sayısının artışı ile hata kareler ortalamasının arttığını ifade etmektedir.

6. SONUÇ

Varyans analizi sonuçlarına göre, sapan değer türleri, sapan değer sayısı, ve tarama süreçlerinin sapan değer tespit performansları arasındaki fark, istatistiksel açıdan önemli bulunmuştur. Söz konusu faktörlerin ikili etkileşimleri de aynı şekilde önemli bulunmuştur. Bu sonuçlardan, doğal rasgelelik sonucunda oluşan sapan değer tespitinin daha kolay olduğu sonucuna varılmıştır. Ayrıca seri içerisinde sapan değer sayısının artması, tarama süreçlerinin güçlerini azaltan bir unsur olmaktadır. Ulaşılan diğer bir önemli sonuç, tarama süreçlerinden Ardışık Yöntemin performansının istatistiksel açıdan diğer süreçlere oranla daha iyi bulunmasıdır.

KAYNAKLAR

Bayhan, M.: Kalite Kontrolünde zaman serisi analizi, *Endüstri Mühendisliği Dergisi*, 21:17-21.

Box, G.E.P ve Jenkins, G.M. (1976), Time series analysis : Forecasting and control, Sect 6.4.3 San Francisco: Holden-Day.

Chang, IH., Tiao, George. C., Chen, Chung. (1988), Estimation of time series parameters in the presence of outliers., American Statistical Association and the American Society for Quality Control.

Corlett, D. ve Lewis, T. (1976), The subjective nature of outliers rejection procedures., *Applied statistics*, 25, no:3 s: 288.

Fox, A.J. (1972), Outliers in time series., J.R. Statist. Soc. B, 34, 350-363.

Kurt, S. (1990), Çok etkenli deneylerde tek sapan değer çözümlemesi, seminer notları, E.Ü Fen Fakültesi İstatistik Bölümü. Bornova-İzmir.

Ljung, Greta .M., Box, G.E.P. (1979), The likelihood function of stationary autoregressive-moving average models., *Biometrika*, 66, 265-270.

Pena, D. (1987), Measuring the importance of outliers in ARIMA models in New perspectives in theoretical and applied statistics., Newyork: Wiley.

Tatlıdil H. (1981), Doğrusal regresyonda ve çok değişkenli verilerde kuşkulu gözlemlerin testi., Basılmamış doktora tezi., Hacettepe Üniversitesi., Ankara.

Tsay, Ruey S. (1988), Outliers, level shifts, and variance changes in time series: *Journal of forecasting.*, 7: 1-20.