Adıyaman University Journal of Science

ADYUSCI

dergipark.ulakbim.gov.tr/adyufbd

6 (2) (2016) 177-186

An Examining from Data Sourced Deviations in Modelling by use of Variance Analysis Method

Ahmet KAYA

Ege University, Tire Kutsan Vocational Training School, 35900 Tire-İzmir, Türkiye
ahmet.kaya@ege.edu.tr

Abstract

In this study, outliers which can cause bias on models in a survey stage using observation values which constitute basement for conducting modelling have been investigated. The processes important for estimation observation, transform to data, modelling from data, and to gain information from models so important by oneself. One of the most important risks we encountered is transformation on data by naturel randomness or derivation by people. If transformation on data is derived by naturel randomness, solution is easy and has to compulsory ultimately. Causes for bias are appeared by time series (ARIMA-Auto Regressive Integrated Moving Average) models, detection, and detection causes on them are modelled by multi factored experimental design with $2^3 3^2$ and significance levels have been investigated.

Keywords: Time Series, Estimation, Modelling, Outliers, Variance Analysis.

Modellemede Veri Kaynaklı Sapmaların Varyans Analizi Yöntemi İle İncelenmesi

Özet

Bu çalışmada bir araştırma sürecinde modelleme planlamaya temel oluşturan, gözlem değerleri üzerinde meydana gelen ve modellerin yanlı olmasına neden olan nedenler incelenmiştir. Gözlem yapmak, gözlemleri verilere dönüştürmek, verilerden modellere ulaşmak ve nihayet modelleri kullanarak bilgi elde etmek başlı başına önemli

Received: 04 April 2016 Accepted: 24 December 2016

süreçlerdir. Bu süreçlerde karşılaşılan en önemli risklerden biri veriler üzerinde meydana gelen değişimlerdir. Bu değişimler doğal yollarla meydana geliyorsa, sorun nispeten basittir. Ancak değişim; kişilerden, aletlerden veya aletlerin yanlış kullanımı ile ortaya çıkıyorsa yanlılık mutlaka giderilmelidir. Bu yanlılıkların güncel bir modelleme yöntemi olan zaman serisi (ARIMA-Auto Regressive Integrated Moving Average) modellerinde ortaya çıkış nedenleri, tespiti ve tespit edilme nedenleri, çok etkenli $2^3 3^2$ deney düzeni ile modellenmiş ve önem düzeyleri araştırılmıştır.

Anahtar Kelimeler: Zaman Serileri, Tahminleme, Modelleme, Sapan Değerler, Varyans Analizi.

1. Giriş

Uygulamalı Matematik, İstatistik, İktisat, Biyoloji, Tıp, Genetik, Gıda ve diğer bilim alanlarında değişimleri gözlemek ya da gözlem yapabilmek çok önemli bir eylem ve bilimsel bir araştırmanın başlangıç aşamasıdır. Nicel kararlara ulaşmada gözlemleri sayısal değerlere dönüştürmek çok daha önemlidir. Sayısal bir nicelik haline dönüşmüş veri, istatistiksel metotlarla işlenebilirse bilgi haline gelir. Bilgi, yönetici veya karar verici durumunda olanların doğrudan kullandığı bir referanstır. Veriden bilgiye giden yolda modeller vardır. Modeller; geleceği öngörmek amacıyla tasarlanır, tahmin aracı olarak kullanıldıkları için de veriye dayalı olarak elde edilirler. Tahmin modellerine temel teşkil eden veriler, ne kadar sağlıklı ve adedi olarak yeterli olursa modeller de o derecede yeterli ve etkin olur. Bir modeli meydana getiren veri seti ne kadar güncel ise, model de o kadar günceldir. Görüldüğü üzere, modeller ile veriler arasında çok sıkı bir ilişki bulunmaktadır.

2. Box-Jenkins Modellemede Sapan Değerler

Box-Jenkins modellerinde iki tip sapan değere rastlanmaktadır. Birinci tip sapan değer (Additive Outlier (AO)) kişilerin veya cihazların hataları ile ortaya çıkmaktadır [1]. Bu hataların doğurduğu etkiler mutlaka gözlemler üzerinden ayrıştırılmalıdır. İkinci tip sapan değer (Innovational Outlier (IO)), doğal rastgelelik sonucu ortaya çıkan ve çıktığı pozisyondan itibaren etkisi azalarak devam eden ancak kendinden sonraki bütün gözlemleri etkileme özelliğine sahip bir türdür [7]. Özetle; seri üzerinde ortaya çıkan sapan veri, bulunduğu pozisyondan önceki ve sonraki gözlemleri etkilememişse, bu hata

kişi kaynaklıdır ve birinci tip sapan değer olarak isimlendirilmektedir. Sapan gözlem; kendinden sonraki gözlemleri etkileme özelliği de göstermişse, bu tür gözlemler doğal rastgelelik sonucunda sapan olmuş denmekte ve ikinci tip sapan değer olarak tanımlanmaktadır. Tarama süreçleri, bu ayırımı seri üzerinde meydana gelen otokorelasyon kayıplarına göre belirlemektedir. Görüldüğü üzere, zaman serilerinde ortaya çıkan sapan değerleri, kaynağı ve nedenleri ile belirlemek mümkün olmaktadır [3]. Ayrıca zaman serilerinde sapan değerlerin tespiti, hata terimleri arasında kurgulu otokorelasyona dayalı olarak yapılmakta olup, birinci tip sapan değerin tespiti, tarama süreçleri tarafından daha kolay bir şekilde yapılabilmektedir [5]. Bununla birlikte birinci tip sapan değer, parametre değerleri üzerinde daha büyük etki meydana getirmekte ve bu etki, şok bir etki olarak tanımlanmaktadır [7].

Box-Jenkins modeli şu şekilde tanımlanır: $\{x_t\}$, $ARMA\ (p,q)$ modeli ile üretilmiş sapan değer içermeyen zaman serisi olsun. $ARMA\ (p,q)$ modeli,

$$\phi(B)x_t = \theta(B)e_t . \qquad (2.1)$$

Burada $\phi(B) = 1 - \phi_1 B - \ldots - \phi_p B^p$, $\theta(B) = 1 - \theta_1 B - \ldots - \theta_q B^q$, $B^k x_t = x_{t-k}$, $E(x_t) = 0$ ve $\{e_t\} \approx (0, \sigma^2)$ özelliklerine sahiptir**.** Eşitlik modelinde tanımlanan $\phi(B)$ fonksiyonunun kökleri $\phi_1, \ldots, \phi_p$ değerleri birim çemberin dışında kalıyorsa model durağanlık, $\theta(B)$ fonksiyonunun kökleri $\theta_1, \ldots, \theta_p$ değerleri birim çemberin dışında kalıyorsa çevrilebilirlik koşulunu, dolaysıyla tersinirlik varsayımlarını sağladığı kabul edilir [6].

Seri durağanlığı kısaca, gözlemler ve hata terimleri arasındaki korelasyonun sınırlar içerisinde olması ve gözlemlerin kısmi otokorelasyon değerlerinin, gecikme değerlerinin artmasına paralel azalması anlamına gelir. Eğer $ARMA\ (p,q)$ modeli, p terimli AR ve q terimli MA modelinin bir kombinasyonu ise $p+q$ terim içerir.

Bu durumda $ARMA\ (p,q)$ modeli,

$$x_t = \underbrace{\phi_1 x_{t-1} + \phi_2 x_{t-2} + \ldots + \phi_p x_{t-p}}_{AR(p)} + \underbrace{e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2} - \ldots - \theta_q e_{t-q}}_{MA(q)} \quad . \qquad (2.2)$$

Görüldüğü gibi $ARMA\ (p,q)$ modelinde $(\mu, \phi_1, \phi_2, ..., \phi_p, \theta_1, \theta_2, ..., \theta_q, \sigma^2)$ olmak üzere $p+q+2$ adet parametre bulunmaktadır. Geriye öteleme işleci $Bx_t = x_{t-1}, B^2 x_t = x_{t-2}, ..., B^p x_t = x_{t-p}$, $Be_t = e_{t-1}, B^2 e_t = e_{t-2}, ..., B^q e_t = e_{t-q}$ kullanılarak model yeniden yazılırsa, $ARMA\ (p,q)$

$$x_t = \underbrace{\phi_1 Bx_t + \phi_2 B^2 x_t + ... + \phi_p B^p x_t}_{\phi(B)x_t} \underbrace{e_t - \theta_1 Be_t - \theta_2 B^2 e_t - ... - \theta_q B^q e_t}_{\theta(B)e_t} \qquad (2.3)$$

elde edilir. Burada $\phi(B)$ ve $\theta(B)$ sırasıyla p ve q dereceden polinomlar olup, sırasıyla durağanlık ve çevrilebilirlik koşulunu sağlar [6].

2.1. A Tipi Sapan Model (Birinci Tip Sapan Değer) (Additive Outlier (AO))

Bir gözlem değeri, beklenmedik bir şekilde kişi ya da alet kaynaklı veya hatalı ölçümler sonucu beklenenden farklı olarak elde edilmiş ise, bu tip sapan değerler birinci tip sapan değerler olarak tanımlanmaktadır [1]. Parametre tahmin değerleri üzerinde şok değişimlerin ortaya çıkmasına neden olurlar [7]. Bu tip sapan modellerde parametre yanlılığı daha keskin ve daha büyüktür ve bu etki mutlaka giderilmelidir. Bu tip gözlemler A tipi sapan modeli ile tanımlanırlar. A tipi model, ilk kez 1972 yılında Fox tarafından tanıtılmış ve modellenmiştir [4]. Birinci tip sapan değer modeli;

$$y_t = z_t + \delta x_t . \qquad (2.4)$$

Burada y_t, gözlenen değer; z_t, sapan değer etkisi bulunmayan orijinal gözlem; δ, sapan değerin büyüklüğü; x_t, sapan değer anında (T = t) 1, aksi halde 0 değeri alan değişkendir [3].

2.2. B Tipi Sapan Model (İkinci Tip Sapan Değer) (Innovational Outlier (IO))

Bu konuda çok sayıda tanım bulunmasına karşın sapan bir değer; parametre tahminlerinin yanlı olmasına neden olan az sayıda gözlem veya gözlemlerin alt seti olarak tanımlanabilir. AR (Auto Regressif) modellerde sapan model kavramını ilk kez 1972 tarihinde Fox ortaya koymuştur [4]. Ayrıca Ljung, 1993 yılında B tipi sapan

modeli, beklenmeyen ancak etkisi sınırlı bir hata modeli olarak tanımlamıştır [7]. Bununla birlikte, Chang, Tiao ve Chen'in 1988 yılında yaptıkları bir benzetim (simülasyon) çalışması, $ARIMA\ (p,d,q)$ modellerde ortaya çıkması muhtemel iki tip sapan değerin etkilerini ortaya koymayı amaçlamıştır [3].

B tipi model; ortaya çıktığı pozisyondan itibaren kendinden sonraki gözlemleri etkileme özelliğine sahiptir. Bu durum, etkisi uzun süren bir ekonomik krizi, kronik bir hastalık durumunu, bir depremin yol açtığı ekonomik etkiyi ve benzeri durumları betimleyen bir durumdur. Literatürde Innovational Outlier (IO) olarak da isimlendirilir. B tipi model, aşağıdaki biçimde modellenmektedir:

$$y_T = \frac{\theta(B)}{\phi(B)}(e_T - \delta\, x_T)\,. \tag{2.5}$$

Burada $\theta(B)$, MA fonksiyonu; $\phi(B)$, AR fonksiyonu; e_T, sapan değer anında oluşan hata terimi; x_T, $T = t$ anında 1, diğer durumlarda 0 değerini alan değişkendir [3].

3. Modelleme Planlaması

Chang, Tiao ve Chen tarafından elde edilen simülasyon verileri kullanılarak; sapan değer taramada; Model (M) (AR, MA), sapan değer Büyüklüğü (B) (3σ, 5σ), Genişliği (G) (50, 100, 150), Duyarlılığı (D) (C=3.0, C=3.5, C=4.0), Türü (T) (AO, IO) gibi faktörlerin ne kadar etkili olduğunu istatistiksel olarak ortaya koymaya yönelik Varyans Analizi (VA) çalışması yapılmıştır. Söz konusu problemlerin çözülmesine yönelik faktörler; Sapan değer Modeli (M), sapan değer Büyüklüğü (B), seri Genişliği (G), ölçüt değer Duyarlılığı (D), sapan değer Türü (T) biçiminde belirlenmiştir. Etken düzeyleri özel seçilmiş olup, çok-etkenli $2^3 3^2$ deney düzeni için model denklemi şöyledir:

$$\begin{aligned} Y_{ijklm} = {} & \mu + M_i + B_j + G_k + D_l + T_m + (MB)_{ij} + (MG)_{ik} + (MD)_{il} + (MT)_{im} + \\ & (BG)_{jk} + (BD)_{jl} + (BT)_{jm} + (GD)_{kl} + (GT)_{km} + (DT)_{lm} + \varepsilon_{ijklm} \end{aligned} \tag{3.1}$$

$i = 1,2; \quad j = 1,2; \quad k = 1,2,3; \quad l = 1,2,3; \quad m = 1,2.$

Bu modelde Y_{ijklm} Sapan değer tespit olasılığı, μ genel ortalama, M_i model türü, B_j sapan değer büyüklüğü, G_k seri genişliği, D_l ölçüt değer duyarlılığı, T_m sapan değer türü, ε_{ijklm} hata terimi, diğerler ise, ikili etkileşimler ifade edilmektedir.

Tablo 1. Tek Sapan Değer Tarama Olasılıkları

			C=3.0		C=3.5		C=4.0	
			AO	IO	AO	IO	AO	IO
AR	3σ	50	0.75	0.83	0.75	0.85	0.72	0.88
		100	0.81	0.82	0.80	0.88	0.80	0.87
		150	0.81	0.77	0.82	0.81	0.81	0.87
	5σ	50	0.86	0.93	0.86	0.94	0.91	0.95
		100	0.90	0.93	0.90	0.93	0.90	0.93
		150	0.92	0.90	0.92	0.90	0.92	0.91
MA	3σ	50	0.78	0.91	0.75	0.94	0.70	0.97
		100	0.83	0.87	0.83	0.90	0.82	0.92
		150	0.85	0.88	0.84	0.89	0.84	0.88
	5σ	50	0.84	0.98	0.84	0.98	0.83	0.99
		100	0.91	0.96	0.91	0.96	0.90	0.96
		150	0.93	0.97	0.93	0.96	0.93	0.97

Tablo 2. (3.1) Modeline İlişkin Varyans Analizi Tablosu

Değişim Kaynağı	**Serbestlik Derecesi**	**Kareler Toplamı**	**Kareler Ortalaması**	**F Değeri**	**Olasılık Değeri**
M (Model)	1	0.016501	0.016501	41.78	0.000
B (Büyüklük)	1	0.134335	0.134335	340.09	0.000
G (Genişlik)	2	0.006808	0.003404	8.62	0.001
D (Duyarlılık)	2	0.001225	0.000613	1.55	0.229
T (Tür)	1	0.078013	0.078013	197.50	0.000
M*B	1	0.002335	0.002355	5.91	0.021
M*G	2	0.001353	0.000676	1.71	0.198
M*D	2	0.001203	0.000601	1.52	0.235
M*T	1	0.011001	0.011001	27.85	0.000
B*G	2	0.000586	0.000293	0.74	0.485
B*D	2	0.000486	0.000243	0.62	0.547
B*T	1	0.003335	0.003335	8.44	0.007
G*D	4	0.000192	0.000048	0.12	0.974
G*T	2	0.040908	0.020454	51.78	0.000
D*T	2	0.004408	0.002204	5.58	0.009
Hata	45	0.015698 6	0.0019811		
Toplam	71	0.318388			

Tablo 3. Tablo 1, Ana Etken Ortalamaları

ETKEN		Ortalama	P
Model	AR	0.86278	
	MA	0.89306	0.0000
Büyüklük	3σ	0.83472	
	5σ	0.92111	0.0000
Genişlik	50	0.86417	
	100	0.88500	0.0001
	150	0.88458	
Duyarlılık	3.00	0.87250	
	3.50	0.87875	0.2290
	4.00	0.88250	
Tür	AO	0.84500	0.0000
	IO	0.91083	

VA Sonuçları (Tablo 2) İçin Yorumlar:

Tablo 2 sonuçları, yapılan modelleme sonuçlarını ortaya koymaktadır. Buna göre; Model farklılığının sapan değer tespitinde önemli olduğu belirlenmiştir. Modelin AR veya MA model olması sapan değer tespitinde belirleyici olmaktadır. MA modellemede sapan değerlerin daha yüksek bir olasılıkla tespit edildikleri görülmektedir. Bunun için Tablo 1 verilerine bakmak yeterlidir. Buna neden faktörün, MA modellemenin hata terimlerinin doğrusal fonksiyonu olarak tanımlanmasının etkisi olduğu değerlendirilmektedir. Bunun içim (2.2)'deki eşitliğe bakılabilir.

Beklentiler doğrultusunda; sapan gözlemin 3σ yerine 5σ büyüklüğünde olması, tespitini daha kolay bir hale getirmektedir. Burada F değerinin çok yüksek değerlerde olması dikkate değerdir. Nitekim sapan değer etkileri, ölçüt değerlerle kıyaslanarak sapan olup olmadıkları sonucuna ulaşılmaktadır.

Ulaşılan diğer bir sonuç, seri genişliğinin sapan değer tespitinde etkili bir faktör olduğu biçimindedir. Bu bakımdan sapan değer analizlerinde seri genişliği arttıkça tarama süreçlerinin performanslarını artırdıkları söylenebilir. Başka deyişle, seri genişliklerinin yüksek olması, parametre tahmin değerlerinin daha etkin olması

sonucunu doğurmaktadır. Birçok araştırma makalesi ve tezlerde zaman serileri analizlerinde minimum 50 gözlemle tahmin işlemlerinin yapılması tavsiye edilmektedir.

Sapan değer tespit etmede kullanılan, ölçüt değer duyarlılıklarının, sapan değer tespitinde önemli bir faktör olmadığı görülmektedir. Ölçüt değer duyarlılığının hangi büyüklükte seçileceğine, araştırmacı karar vermektedir. Bu, istatistik araştırmalarda önem düzeyinin ne şekilde belirlenmesi ile tam bir benzerlik göstermektedir.

4. Sonuçlar

Yapılan özel seçimli ($2^3 3^2$) çok etkenli deney tespitinde model farklılığının, sapan değer büyüklüğünün, seri genişliğinin ve sapan değer türlerinin önemli olduğu, ölçüt değer duyarlılığının önemsiz olduğu sonucuna ulaşılmıştır. Bu sonuçlar şunu söylemektedir: Sapan değer büyüklüğünün, seri içerisinde tespit edilmesinde önemlidir. Bununla birlikte modellerin AR ya da MA olması sapan değer tespitinde belirleyicidir. MA modellerde daha yüksek tespit edilme olasılıklarına ulaşılmaktadır. Ayrıca, modelleri meydana getiren seriler genişledikçe, sapan değerleri tespit etme olasılıkları artmaktadır. Sapan değer türlerinin de tespit edilmelerinde etken oldukları ulaşılan diğer bir sonuçtur. Böylece, A tipi veya birinci tip sapan değerlerin seri içerisinde tespit edilme olasılıklarının daha yüksek olduğu doğrulanmıştır [5].

Ölçüt değer duyarlılığının sapan değerleri tespit etmede önemsiz olduğu sonucuna ulaşılmıştır.

5. Çıkarsamalar

1. Birinci tip sapan değer; kişi, ölçüm hatası ve alet kaynaklı bir durumu ifade eder iken, ikinci tip sapan değer, süreçsel özellikli hata durumlarını ifade eder [1].
2. Sapan değer, ekonomik özellikli bir veri setinde, bir grev ya da kriz durumunu [3], bir kalite kontrol sürecinde girdi, çıktı ya da süreç problemini [1], veri tabanı üzerinde meydana gelen bir hatayı [9] ve daha birçok farklı nedenin ortaya çıkardığı bir durumu ifade edebilir.
3. Sapan değer analizleri; hisse senedi ve borsa analizleri, kalp düzensizlik analizleri, genetik algoritma çalışmaları, fuzzy mantık çalışmaları, parametre

iyileştirme analizleri ve kalite kontrol süreçlerinde kullanılan analizler için çok önemli bir analiz aracı haline gelmiştir.

4. Sapan değer analiz süreci ile birlikte, veriler üzerinde bazı dönüşümler gerekli olabilir. Dönüşüm işlemleri (Karekök ve logaritma dönüşümleri) veri düzeltmeye katkı yapıyor ise sapan değer analizine gereksinim duyulmayabilir [10].
5. Sapan değer analizleri, veri düzeltme, parametre iyileştirme özellikleri yanında, hata kareler ortalamasını küçülten bir etki oluşturduğundan, model, dolaysıyla varyans iyileştirme özelliği de sağlayan bir analizdir.
6. Sapan değer analizleri homojen ve hassas veri setlerine uygulanmalıdır. Veri setleri yeterince homojen değil ise, tarama testleri ne kadar güçlü olursan olsun, testlerin gücünde bir azalma olacağı kaçınılmazdır.
7. Yanlış belirlenen sapan değer tiplerinin, test yöntemlerinde etkinlik kaybına neden olduğu belirlenmiştir [8].

Ek Bilgi

Bu makale, “The XIII’th. International Conference on Applied Stochastic Models and Data Analysis (ASMDA)” kongresinde Litvanya’da bildiri olarak sunulmuştur.

Kaynaklar

[1] Bayhan, M., *Kalite Kontrolünde Zaman Serisi Analiz*i, Endüstri Mühendisliği Dergisi, **21**, 17-21, 1992.

[2] Box, G. E. P., Jenkins G. M., Time series analysis: Forecasting and control, Sect 6.4.3. San Francisco, Holden-Day, 1976.

[3] Chang, I., Tiao G. C., Chen, C., Estimation of Time Series Parameters in the Presence of Outliers, American Statistical Association and the American Society for Quality Control, 1988.

[4] Fox, A. J., *Outliers in Time Series*, J. Royal Statistical Society B., **34(3)**, 350-363, 1972.

[5] Kaya, A., *AR(1) Modelinde A Tipi Sapan Etki*, İstatistikçiler Dergisi, **3**, 1-7, 2010.

[6] Ljung, G. M., Box, G. E. P., *The Likelihood Function of Stationary Autoregressive-Moving Average Models*, Biometrika, **66**, 265-270, 1979.

[7] Ljung, G. M., *On Outlier Detection in Time Series*, J. Royal Statistical Society B, **55**, 559-567, 1993.

[8] Muirhead, C. R., *Distinguishing Outlier Types in Time Series*, J. Royal Statistical Society B, **48(1)**, 39-47, 1986.

[9] Kaya, A., *Outlier Effects On Databases*, ADVIS 2004-Advances in Information Systems, Dokuz Eylül Üniversitesi, İzmir, 2004.

[10] Kurt, S., *Çok Etkenli Deneylerde Tek Sapan Değer Çözümlemesi*, Seminer Çalışması, Ege Üniversitesi Fen Fakültesi İstatistik Bölümü, İzmir.