ZÜMRELERE GÖRE ÖRNEKLEMEDE GENETİK ALGORİTMANIN ETKİNLİĞİNİN ARTTIRILMASINA YÖNELİK BİR ÇALIŞMA: BAŞLANGIÇ POPULASYONUNUN GEOMETRİK YÖNTEMLE BELİRLENMESİ

Şebnem ER

İstanbul Üniversitesi, İşletme Fakültesi, Sayısal Yöntemler Anabilim Dalı, İSTANBUL
sebnemer@istanbul.edu.tr

Timur KESKİNTÜRK

İstanbul Üniversitesi, İşletme Fakültesi, Sayısal Yöntemler Anabilim Dalı, İSTANBUL
tkturk@istanbul.edu.tr

ÖZET

Bu çalışmada, zümrelere göre örneklemede zümre sınırlarının belirlenmesi ve örneklemin zümrelere dağıtımı problemlerinin genetik algoritma ile çözümünde etkinliğin arttırılması amaçlanmaktadır. Zümre sınırları ve örneklem büyüklükleri, amaç fonksiyonu olan tahmin varyansını minimum yapacak şekilde genetik algoritma ile belirlenmiştir. Daha önce bu konuda yapılan çalışmada (Keskintürk, Er, 2007) başlangıç populasyonu rassal olarak belirlenmiştir. Geometrik yöntemle rassal arama daha iyi bir noktadan başlatılarak daha kısa bir zamanda aynı ya da daha iyi sonuçlara ulaşmak hedeflenmiştir. Birçok test problemi üzerinde rassal ve geometrik başlangıç populasyonları kullanılmış ve sonuçlar karşılaştırılmıştır.

Anahtar Sözcükler: *Zümrelere göre örnekleme; genetik algoritma; geometrik yöntem*

1.GİRİŞ

Zümrelere göre örnekleme özellikle farklı değerlere sahip anakütle elemanlarının (N) bir ya da birkaç özelliğe dayanarak, daha homojen alt gruplara (zümre) ayrıldığı bir yöntemdir (Cyert ve Davidson, 1962; Cochran, 1977; Hess, ve diğerleri, 1966; Bretthauer, ve diğerleri, 1999; Rao, 2000). Zümrelere göre örneklemede daha sonra her bir zümreden iadesiz olarak çekilen örneklemler birleştirilerek tek bir örneklem gibi incelenmektedir (Hedlin, 1997). Bu yolla tahmin varyansı minimize edilerek, basit rassal örneklemeye kıyasla istatistiki doğruluk arttırılmaktır (Cochran, 1977). Ancak istatistik doğruluğun arttırılması zümre sınırlarının tahmin varyansını minimize edecek şekilde belirlenmesine dayanmaktadır ve zümrelere göre örneklemenin uygulanması aşamasında karşılaşılan en önemli problemlerden biridir.

Literatürde zümre sınırlarının belirlenmesi konusunda Dalenius-Hodges'in (1959) frekansların kümülatif karekökleri yöntemi, Nicolini'nin (2001) NCM'si, Gunning ve Horgan'ın (2004) geometrik yöntemi, Kozak'ın (2004) rassal arama yöntemi, Ekman'nın kuralı, Sethi'nin kuralı, Singh'in yöntemi, L&H algoritması (Hess, ve diğerleri, 1966) gibi birçok farklı yaklaşım bulunmaktadır. Zümrelere göre örneklemenin ikinci önemli problemi olan örnek büyüklüğünün zümrelere dağıtılması konusunda ise literatürde eşit, orantılı, Neyman (optimal) ve orantısız olmak üzere farklı dağıtım yöntemlerine yer verilmektedir (Hess, ve diğerleri, 1966).

Bu çalışmanın temel amacı, tahmin varyansını minimize edecek zümre sınırlarını genetik algoritma ile belirlemek ve GA'nın performansını arttırmak amacıyla Gunning ve Horgan'ın (2004) geometrik yönteminden elde edilen zümre sınırlarını GA'da başlangıç populasyonu olarak belirlemektedir. Önceden belirli olan toplam örnek büyüklüğünün (n) belirli sayıda zümre arasında dağıtımı ise Keskintürk ve Er (2007)'in çalışmasında yer alan GA ile örnek büyüklüğü dağıtımı yöntemine göre yapılmıştır. İkinci bölümde, zümre sınırlarının nasıl

oluşturulacağı ve bu çalışmaya temel oluşturan geometrik yöntem ile GA'nın işleyiş biçimi açıklanmıştır. Üçüncü bölümde ise zümre sınırlarının belirlenmesi probleminde başlangıç populasyonunun rassal ve Gunning ve Horgan'ın (2004) geometrik yöntemiyle belirlendiği genetik algoritmanın uygulanmasıyla ilgili elde edilen test sonuçlarına yer verilmiştir.

2. ZÜMRE SINIRLARININ BELİRLENMESİ VE ÖRNEK BÜYÜKLÜĞÜNÜN DAĞITIMI

Bu bölümde zümre sınırlarının belirlenmesi aşamasında kullanılacak, Keskintürk ve Er (2007)'in çalışmasında yer alan GA ile GA'da başlangıç populasyonunun bir kısmını oluşturacak Gunning ve Horgan'ın (2004) geometrik yöntemi anlatılacaktır. Tüm çalışmada aşağıdaki notasyonlar kullanılmaktadır:

Y	Zümrelere ayrılacak anakütle	σ_{yh}^2	h. zümrenin varyansı
N	Anakütle büyüklüğü	$\bar{Y}_h$	h. zümrenin ortalaması
n	Örnek büyüklüğü	$\bar{y}_{strat}$	Ortalamanın tahmini
H	Zümre sayısı	$b = k_H$	Maksimum değer
N_h	h. zümredeki eleman sayısı	$a = k_0$	Minimum değer
n_h	h. zümreden çekilecek örnek büyüklüğü	k_h	h. zümrenin üst sınırı

Ortalamanın tahmini ve tahmin ortalamasının $\bar{y}_{strat}$ varyansı Cochran (1977)'da aşağıdaki gibi verilmektedir:

$$\bar{y}_{strat} = \frac{\sum_{h=1}^{H} N_h \bar{y}_h}{N}, \qquad S^2_{\bar{y}_{strat}} = \frac{1}{N^2}\sum_{h=1}^{H} N_h^2 \frac{\sigma_{yh}^2}{n_h}\left(1-\frac{n_h}{N_h}\right) \tag{1}$$

Bu formülde her bir zümrenin varyansının bilindiği ve aşağıdaki gibi hesaplandığı varsayılmaktadır:

$$\sigma_{yh}^2 = \sum_{i=1}^{N_h} (Y_{hi} - \bar{Y}_h)^2 / (N_h - 1), \tag{2}$$

Burada Y_{hi} h'ninci zümredeki i'ninci elemanın değerini temsil etmektedir. 3 no'lu denklemde Nh>1 olduğu varsayılmaktadır, dolayısıyla N_h=1 olduğu durumda sapma sıfır olmaktadır. Bu çalışmada örnek büyüklüğü dağıtımı problemi için GA (Keskintürk, Er, 2007) yaklaşımı benimsenmiştir.

2.1 Geometrik Yöntem

Gunning ve Horgan'ın (2004) geometrik yöntemi zümre sınırlarını zümreler arası değişkenlik katsayılarını birbirine eşit yapacak şekilde belirlemektedir. Bu aşamada zümre içi dağılımın tekdüze dağılıma uygunluk gösterdiği varsayılmakta ve zümre içi standart sapma ve ortalama tekdüze dağılımın özellikleri kullanılarak aşağıdaki gibi elde edilmektedir:

$$\sigma = (b-a)/\sqrt{12}\,, \quad \bar{Y} = (a+b)/2 \tag{3}$$

Bu durumda h. zümrenin değişkenlik katsayısı

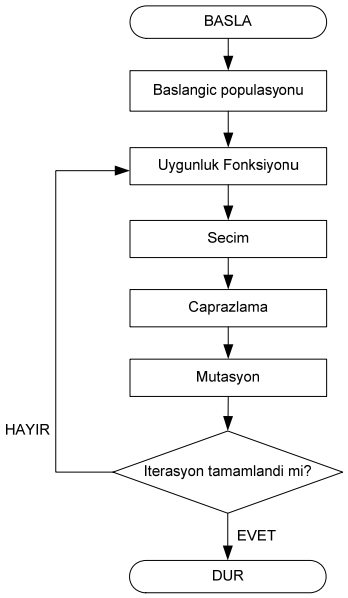
$$CV_h = \frac{\frac{1}{\sqrt{12}}\left(k_h - k_{h-1}\right)}{\frac{\left(k_h + k_{h-1}\right)}{2}} \tag{4}$$

Her bir zümrenin değişkenlik katsayıları birbirine eşitlendiğinde zümre sınırlarının geometrik bir dizi şeklinde arttığı görülmektedir:

$$k_h = k_0\left(k_H/k_0\right)^{h/H} \tag{5}$$

2.2 Genetik Algoritma

Genetik algoritma populasyon temelli sezgisel bir optimizasyon yöntemidir (Goldberg, 1989). Problem değişkenleri kromozomlarla temsil edilmektedir. GA'ya ait akış diyagramı Şekil 1'deki gibi çalışmaktadır:



Şekil 1. GA akış diyagramı

Başlangıç populasyonu oluşturulduktan sonra, her kromozomun değeri uygunluk fonksiyonu ile belirlenir. Başlangıç populasyonu genel olarak rassal olarak belirlenmektedir. Bu çalışmada rassal başlangıç populasyonu ile birlikte geometrik yöntemin sonuçlarının populasyonunun belirlenen miktardaki kromozomlarının değerini oluşturduğu farklı bir başlangıç populasyonu denenmiştir. Böylelikle arama surecine daha iyi bir noktadan başlayarak aramanın etkinliğinin arttırılması amaçlanmıştır. Daha sonra çeşitli GA operatörleri (seçim, çaprazlama ve mutasyon) ile daha uygun çözümler için değişiklikler yapılır. Bu süreç önceden belirlenmiş olan iterasyon sayısına ulaşılana kadar tekrarlanır.

Bu çalışmada zümre sınırlarının belirlenmesi ve örnek büyüklüğü dağıtımı için ikili ve reel değerli kodlama yöntemleri kullanılmıştır. Bu iki kodlama yöntemi Şekil 2'deki gibidir.

Değer	1.5	1.0	5.2	2.8	4.5	1.1	0.2	
İkili & reel-değerli kodlama	0	0	0	1	0	0	1	3	2

Şekil 2. Zümrelere göre örneklemede ikili ve reel-değerli kodlama örneği

Birinci "0"dan birinci "1"e kadar olan genlerin sayısı birinci zümrenin büyüklüğünü (N_1), birinci "1"den sonra gelen "0"dan ikinci "1"e kadar olan genlerin sayısı ikinci zümrenin büyüklüğünü (N_2) göstermektedir. Dolayısıyla "1" ile kodlanmış genler her bir zümrenin sınırını temsil etmektedir. 2.8 ve 0.2 değerleri zümre sınırlarını göstermekte ve böylelikle zümre büyüklükleri sırasıyla 4 ve 3 olmaktadır. Bu zümrelerden çekilecek örnek büyüklüklerini gösteren son 2 gene bakıldığında da örnek büyüklüklerinin 3 ve 2 olduğu anlaşılır.

Uygunluklar hesaplandıktan sonra seçim operatörü ile uygunluk değeri göz önüne alınarak kromozomların bir sonraki nesle geçip geçmeyeceklerine karar verilir. Bu çalışmada rulet tekerleği seçim yöntemi kullanılmıştır. Çaprazlama kromozomlar arası bilgi değişimini sağlamakta olup bu çalışmada tek-nokta çaprazlama yöntemi kullanılmıştır. Daha sonra her bir kromozom mutasyona tabi tutulur. Mutasyon yerel optimuma takılmayı önlemekte olup rassal değişim mutasyonu kullanılmıştır. Kromozomdaki ikinci kısım için iki genin rassal olarak seçildiği ve bu genlerden birinden bir birimlik örnek büyüklüğü çıkartılarak diğerine eklendiği bir mutasyon türü geliştirilmiş ve kullanılmıştır.

3. UYGULAMA

Bu çalışmada zümre sınırları ve örnek büyüklükleri GA ile belirlenmiştir. GA'nın performansını arttırmaya yönelik olarak başlangıç populasyonunun belli bir orandaki kısmında geometrik yöntemin sonuçları kullanılmıştır. Karşılaştırma için başlangıç populasyonunun tamamen rassal olarak belirlendiği ve geometrik yöntemle belirlendiği algoritmaların sonuçları elde edilmiştir. Çalışmada iki farklı anakütle ile çalışılmış olup bunlardan ilki 2005 yılında İSO tarafından yayımlanan Türkiye'nin 500 büyük imalat sanayi firması içerisinden seçilen 485 firmanın net satış değerleri (iso); diğeri ise Gunning ve Horgan'ın çalışmasından alınan verilerdir (JH4). Anakütleler 2, 3 ve 4 zümreye ayrılmış olup, toplam örnek büyüklüğü 100 olarak belirlenmiştir. GA parametrelerinin belirlenmesi aşamasında çeşitli parametre kombinasyonları her bir problem ve zümre büyüklüğü için 100'er kere çalıştırılmıştır. 2, 3 ve 4 zümre için iterasyon sayısı sırasıyla 1000, 1000 ve 1500 olarak seçilmiştir. GA'da populasyon büyüklüğü 100 ve çaprazlama oranı 0.99 olarak belirlenmiştir. Elde edilen sonuçlardan en iyi değeri veren parametreler Tablo 1'deki gibidir.

Tablo 1. Genetik algoritma parametreleri

Anakütle	Zümre	2	3	4
iso	Geo oranı	0.05	0.05	0.05
	Mut oranı	0.35	0.50	0.50
JH4	Geo oranı	0.50	0.50	0.50
	Mut oranı	0.50	0.50	0.35

Tablo 2 ve 3'te çalışmada kullanılan verilerin zümre sınırlarının ve örnek büyüklüğünün belirlenmesi için geliştirilen GA'da başlangıç populasyonu rassal ve geometrik yöntemle

belirlenmiş olan ve 100 çalıştırmada elde edilen tahmin varyansı değerlerinin ortalaması, standart hatası ve değişim katsayısı değerlerine yer verilmiştir.

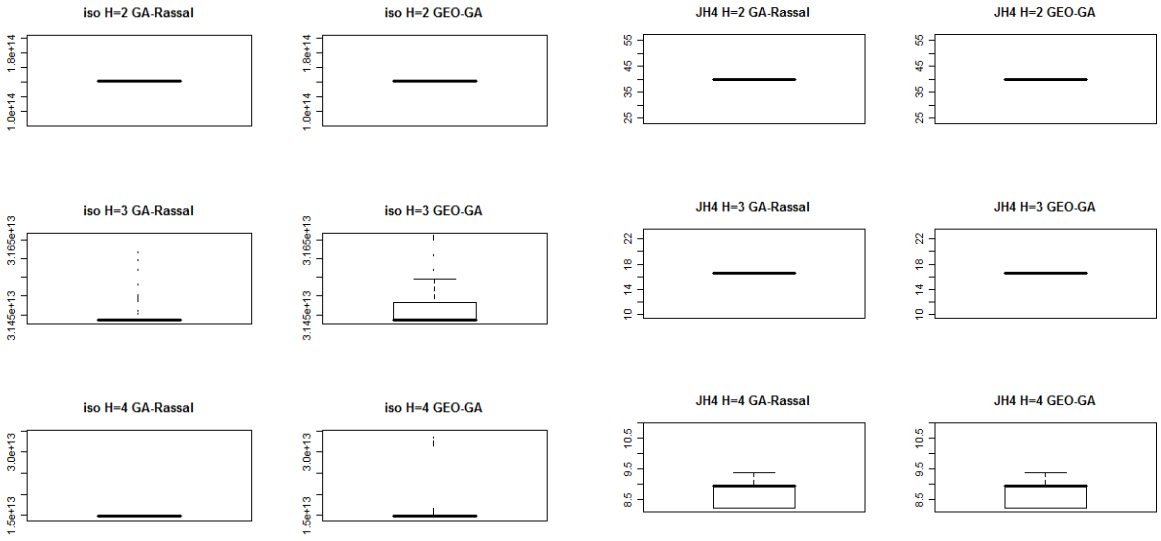
Tablo 2. ISO verisi için GA ile her iki başlangıç populasyonundan elde edilen tahmin varyansının ortalaması, standart hatası ve değişim katsayısı değerleri

	Ortalama (10^{13})		Standart Sapma (10^{12})		Değişim Katsayısı (%)	
H	Başl.Pop. Rassal	Başl.Pop. Geometrik	Başl.Pop. Rassal	Başl.Pop. Geometrik	Başl.Pop. Rassal	Başl.Pop. Geometrik
2	14.18056	14.18056	0.00000	0.00000	0.000	0.000
3	3.14486	3.14768	0.03630	0.07672	0.115	0.244
4	1.46725	1.59795	0.10069	4.47576	0.686	28.009

Tablo 3. JH4 verisi için GA ile her iki başlangıç populasyonundan elde edilen tahmin varyansının ortalaması, standart hatası ve değişim katsayısı değerleri

	Ortalama		Standart Sapma		Değişim Katsayısı (%)	
H	Başl.Pop. Rassal	Başl.Pop. Geometrik	Başl.Pop. Rassal	Başl.Pop. Geometrik	Başl.Pop. Rassal	Başl.Pop. Geometrik
2	39.95957	39.95957	0.00000	0.00000	0.000	0.000
3	16.53336	16.53336	0.00000	0.00000	0.000	0.000
4	8.780142	8.95456	0.36411	0.17126	4.147	1.913

Şekil 3'teki kutu-nokta diyagramları ise her iki örnekte başlangıç populasyonu rassal ve geometrik yöntemle belirlendiği durumda 100 deneme sonucunda elde edilen tahmin varyansı değerlerinin dağılımını göstermektedir.



Şekil 3. ISO ve JH4 verilerinde tahmin varyansı değerlerinin kutu-nokta diyagramları

Tablo 2-3 ve Şekil 3'ten de görülebileceği gibi GA yönteminde başlangıç populasyonunun rassal ya da geometrik yöntemle belirlendiği her iki durumda da 2 zümre için aynı tahmin varyansı değerine ulaşılmıştır. Bunun yanında zümre sayısı arttıkça çarpıklık değeri 12.672 olan iso verisinde başlangıç populasyonunun bir kısmının geometrik yöntemle belirlenmesinin tahmin varyansına önemli bir katkıda bulunmadığı; çarpıklık değeri 2.076 olan JH4 verisinde önceden belirlenen iterasyon sayısına ulaşıldığında GA'nın benzer tahmin varyansı değerlerinde durduğu gözlenmiştir.

4. SONUÇ

Zümrelere göre örnekleme oldukça heterojen yapıdaki anakütleler için istatistik kesinliği arttırmada yaygın olarak kullanılan bir örnekleme türüdür. Bu çalışma Keskintürk ve Er (2007)'in çalışmasında zümre sınırlarının ve örnek büyüklüğünün belirlenmesi problemine önerilen GA yaklaşımını başlangıç populasyonunda geometrik yöntemden elde edilen sınırları kullanarak geliştirmeyi amaçlamaktadır. Yapılan analizlerden de görülebildiği gibi başlangıç populasyonun geometrik yöntemle belirlenmesi çok çarpık veri setlerinde önemli bir katkı sağlamazken, nispeten daha az çarpık verilerde GA sonucu elde edilen tahmin varyansındaki değişkenliği azaltmıştır.

KAYNAKÇA

Bretthauer, K. M., Ross, A., Shetty, B., 1999. Nonlinear integer programming for optimal allocation in stratified sampling. *European Journal of Operational Research* 116, 667-680.

Cochran, W. G., 1977. *Sampling Techniques*, 3rd ed., John Wiley & Sons, Inc. USA.

Cyert, R.M., Davidson, H.J., 1962. *Statistical Sampling for Accounting Information*, Prentice-Hall Inc., Englewood Cliffs, New Jersey, 116-127.

Dalenius, T., Hodges, J. L.Jr., 1959. Minimum Variance Stratification. *Journal of the American Statistical Association* 54, 285, 88-101.

Goldberg, D.E., 1989. *Genetic Algorithms in Search Optimization and Machine Learning*, Addison Wesley Publishing Company, New York.

Gunning, P., Horgan, J.M., 2004. A New Algorithm for the Construction of Stratum Boundaries in Skewed Populations. *Survey Methodology* 30, 2.

Hedlin, D., 1997. Minimum Variance Stratification of a Finite Population. *SSRC Methodology Working Paper*, M03/07.

Hess, I., Sethi, V.K., Balakrishnan, T.R., 1966. Stratification: A Practical Investigation. *Journal of the American Statistical Association* 61, 313, 74-90.

Keskintürk, T., Er, Ş., 2007. A Genetic Algorithm Approach to Determine Stratum Boundaries and Sample Sizes of Each Stratum in Stratified Sampling. *Computational Statistics & Data Analysis* 52, 1, 53-67.

Kozak, M., 2004. Optimal Stratification Using Random Search Method in Agricultural Surveys. *Statistics in Transition* 6, 5, 797-806.

Nicolini, G., 2001. A Method to Define Strata Boundaries. *Department of Economics University of Milan Italy, Departemental Working Paper*, 2001-01.

Rao, P. S.R.S., 2000. *Sampling Methodologies with Applications*. Chapman & Hall/CRC, Washington D.C.