ZÜMRELERE GÖRE ÖRNEKLEMEDE ZÜMRE SINIRLARININ VE ÖRNEK BÜYÜKLÜKLERİNİN GENETİK ALGORİTMA KULLANILARAK BELİRLENMESİ

Şebnem Er, Timur Keskintürk
İstanbul Üniversitesi, İşletme Fakültesi, Sayısal Yöntemler Anabilim Dalı
sebnemer@istanbul.edu.tr, tkturk@istanbul.edu.tr

ÖZET

Zümrelere göre örnekleme, heterojen yapıdaki bir anakütlenin homojen alt gruplara ayrılarak incelendiği bir örnekleme türüdür. Bu çalışma, zümrelere göre örneklemede önemli yer tutan zümre sınırlarının belirlenmesi ve örneklemin zümrelere dağıtımı problemlerinin sezgisel bir optimizasyon tekniği olan genetik algoritma ile çözümünü içermektedir. Zümre sayısı ve incelenecek örnek büyüklüğünün belirli olduğu varsayılmıştır. Zümre sınırları, amaç fonksiyonu olan tahmin varyansını minimum yapacak şekilde genetik algoritma ile ve her bir zümreden çekilecek örnek büyüklükleri; eşit, orantılı ve Neyman yöntemleri kullanılarak belirlenmiştir. Bu üç dağıtım yöntemine ek olarak örnek büyüklüğünün dağıtımı hakkında herhangi bir kısıtın olmadığı ve hem örnek büyüklüklerinin hem de zümre sınırlarının genetik algoritma ile belirlendiği dördüncü bir yöntem geliştirilmiştir. Her bir yöntem için örneklere yer verilerek sonuçlar karşılaştırılmış ve orantısız dağıtım yöntemiyle elde edilen tahmin varyansının minimum olduğu gözlenmiştir.

Anahtar Kelimeler:*Zümrelere göre örnekleme, genetik algoritma, zümre sınırları, örnek büyüklüğü.*

1. GİRİŞ

Zümrelere göre örnekleme özellikle farklı değerlere sahip anakütlelerin elemanlarının (N), satışlar, çalışan sayısı gibi önemli bir ya da birkaç özelliğe dayanarak, daha homojen alt gruplara (zümre) ($N_1, \mathbf{L}, N_h$) ayrıldığı bir yöntemdir (Cyert ve Davidson, 1962; Cochran, 1963; Hess, ve diğerleri, 1966; Bretthauer, ve diğerleri, 1999; Rao, 2000). Zümrelere göre örnekleme daha sonra her bir zümreden iadesiz olarak örneklem çekmeye ve çekilen bu örneklerin birleştirilerek tek bir örneklem gibi incelenmesine dayanmaktadır (Hedlin, 1997).

Zümrelere göre örneklemede en önemli hedef tahmin varyansını minimize ederek, basit rassal örneklemeye kıyasla istatistiki doğruluğu arttırmaktır (Cochran, 1963). Bu hedefe ancak her bir zümrenin kendi içindeki değişkenliğinin minimum olması ile ulaşılabilmektedir (Cyert ve Davidson, 1962, p. 116). Sonuç olarak, zümre sınırlarının belirlenmesi zümrelere göre örneklemenin uygulanması aşamasında karşılaşılan en önemli problemlerin başında gelmektedir. Zümrelere göre örnekleme hakkında daha geniş bilgi Cochran (1963)'te bulunabilir.

Literatürde zümre sınırlarının belirlenmesi konusunda Dalenius-Hodges'in (1959) frekansların kümülatif karekökleri yöntemi, Nicolini'nin (2001) NCM'si, Gunning ve Horgan'ın (2004) algoritması, Kozak'ın (2004) rassal arama yöntemi, Ekman'nın kuralı, Sethi'nin kuralı, Singh'in yöntemi, L&H algoritması (Hess, ve diğerleri, 1966; Kish ve Anderson, 1978) 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 belirlemektir. Önceden belirli olan toplam örnek büyüklüğünün (n) belirli sayıda zümre arasında dağıtımı ise eşit, orantılı, Neyman ve orantısız olmak üzere 4 farklı dağıtım yöntemine göre yapılmıştır (Cochran, 1962; Hess, ve diğerleri, 1966). İkinci bölümde, zümre sınırlarının nasıl oluşturulacağı ve örnek büyüklüğünün dağıtımının nasıl yapılacağı özetlenmiştir. Üçüncü bölümde, genetik algoritma ve genetik algoritmanın zümre sınırlarının belirlenmesi problemine nasıl uygulanacağı açıklanmıştır. Dördüncü bölümde ise genetik algoritmanın zümre sınırlarının belirlenmesi ve örnek büyüklüğünün dağıtımı problemlerine uygulanmasıyla ilgili sayısal örneklere yer verilmiştir.

2. ZÜMRE SINIRLARININ BELİRLEMESİ ve ÖRNEK BÜYÜKLÜĞÜNÜN DAĞITIMI

Bu çalışmada, tahmin varyansını ($S^2_{\bar{Y}_{strat}}$) minimize eden zümre sınırları, dört farklı örnek büyüklüğü dağıtım yöntemine göre GA aracılığıyla belirlenmektedir. Zümre sayısının (H) ve toplam örnek büyüklüğünün (n) sabit olduğu varsayılmaktadır. Tüm çalışmada aşağıdaki notasyonlar kullanılmaktadır:

Y	Zümrelere ayrılacak anakütle
N	Anakütle büyüklüğü
n	Örnek büyüklüğü
H	Zümre sayısı
N_h	h. (h=1,…,H) zümredeki eleman sayısı
n_h	h. zümreden çekilecek örnek büyüklüğü
σ^2_{yh}	h. zümrenin varyansı
$\bar{Y}_h$	h. zümrenin ortalaması

Tahmin varyansı Cochran (1963, s. 91)'da aşağıdaki gibi verilmektedir:

$$S^2_{\bar{Y}_{strat}} = \frac{1}{N^2}\sum_{h=1}^{H} N_h^2 \frac{\sigma^2_{yh}}{n_h}\left(1-\frac{n_h}{N_h}\right) \qquad (1)$$

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

$$\sigma^2_{yh} = \sum_{i=1}^{N_h} (Y_i - \bar{Y}_h)^2 / (N_h - 1) \qquad (2)$$

Örnek büyüklüğünün dağıtımı ($n_1, \mathbf{L}, n_h$) ile ilgili yöntemlere ilişkin formüller ise aşağıdaki gibidir:

Eşit Dağıtım Yöntemi **(m1) :** $n_h = \frac{n}{H} \qquad n_1 = \mathbf{L} = n_h \qquad h = 1, 2, \mathbf{L}, H$ (3)

Orantılı Dağıtım Yöntemi **(m2) :** $n_h = n \cdot \frac{N_h}{N} \qquad h = 1, 2, \mathbf{L}, H$ (4)

Neyman Dağıtım Yöntemi **(m3) :** $n_h = n \cdot \frac{N_h \sigma_{yh}}{\sum_{h=1}^{H} N_h \sigma_{yh}} \qquad h = 1, 2, \mathbf{L}, H$ (5)

Orantısız Dağıtım Yöntemi **(m4) :** n_1, …, n_h değerleri genetik algoritma ile belirlenmektedir.

Dolayısıyla GA'daki uygunluk fonksiyonu bu yöntemlerin tümünde (1) no'lu denklem ile verilen tahmin varyansıdır.

3. ZÜMRELERE GÖRE ÖRNEKLEMEDE GA YAKLAŞIMI

İlk olarak J. Holland tarafından geliştirilen GA sezgisel bir optimizasyon yöntemidir (Holland, 1975; Goldberg, 1989; Michalewicz, 1992). Problem değişkenleri kromozom adı verilen yapılarla temsil edilmektedir. Bu çalışmada ikili kodlama (m1,2,3) ile ikili ve reel-değerli kodlama (m4) yöntemleri kullanılmıştır. Başlangıç populasyonu oluşturulduktan sonra, her kromozomun değeri uygunluk fonksiyonu ile belirlenmektedir. Bu çalışmada uygunluk değeri iterasyon süresince minimize edilecek olan ve Eq. (1) ile gösterilen tahmin varyansıdır. Birinci ve dördüncü yöntemlerde zümre örnek büyüklüklerinin zümre büyüklüğünden fazla olmaması koşulundan dolayı ceza fonksiyonları ile bu özellikteki kromozomların elenmesi sağlanmıştır. Problemin yapısına göre daha iyi sonuca sahip bireyler

gelecek nesillere aktarılmakta ve bu bireylerden daha kaliteli yeni bireyler elde edilmektedir (Man ve Kwong, 1996). Seçim operatörü ile kromozomların, uygunluk değerlerine dayanarak, gelecek jenerasyona geçip geçmeyeceğine karar verilmektedir. GA'nın en önemli operatörlerinden çaprazlama ile de kromozomlar arası özelliklerin değişimi sağlanmaktadır. Bu çalışmada problem büyüklüklerine göre tek-nokta, iki-nokta ve çok-nokta çaprazlama teknikleri kullanılmıştır. Çaprazlama sonrası mutasyon için rassal değişim mutasyonu (Nearchou, 2004) kullanılmıştır. GA'daki bu süreç önceden belirlenmiş olan iterasyon sayısına ulaşılana kadar tekrarlanmaktadır.

4. UYGULAMA

Zümre sınırlarının belinlenmesi için önerilen algoritma Matlab programlama dili kullanılarak geliştirilmiş ve farklı özelliklere sahip iki sayısal örnek (iso487 ve p75) üzerinde uygulanmıştır. Birinci uygulama (iso487), İSO'nun 2004 yılı Birinci 500 Büyük İmalat Sanayi Kuruluşu net satışlar verileri (N=487) ile ikinci uygulama ise literatürde Hedlin (1997)'in çalışmasında kullanılmış olan nüfus verileri (N=284) ile yapılmıştır. Her iki örnek için de örneklem büyüklüğü 80 olarak varsayılmış ve GA ile, dört dağıtım yönteminin her biri için iki ve dört zümre oluşturulmuştur.

Tablo 1: 4 dağıtım yöntemi için iso487 örneğine ait zümre sınırları (H=2)

	m1		m2		m3		m4	
Zümre	N_h	n_h	N_h	n_h	N_h	n_h	N_h	n_h
1	447	40	485	79	442	35	442	35
2	40	40	2	1	45	45	45	45
Tahmin varyansı (10^{14})	2.1591		18.129		**2.133**		**2.133**	

Tablo 2: 4 dağıtım yöntemi için iso487 örneğine ait zümre sınırları (H=4)

	m1		m2		m3		m4	
Zümre	N_h	n_h	N_h	n_h	N_h	n_h	N_h	n_h
1	290	20	397	65	261	21	252	12
2	124	20	67	11	119	8	125	11
3	53	20	19	3	72	16	74	21
4	20	20	4	1	35	35	36	36
Tahmin varyansı (10^{13})	3.0519		13.409		2.8604		**2.4948**	

Tablo 3: 4 dağıtım yöntemi için p75 örneğine ait zümre sınırları (H=2)

	m1		m2		m3		m4	
Zümre	N_h	n_h	N_h	n_h	N_h	n_h	N_h	n_h
1	244	40	282	79	236	32	236	32
2	40	40	2	1	48	48	48	48
Tahmin varyansı	1.49500		7.04500		**1.41450**		**1.41450**	

Tablo 4: 4 dağıtım yöntemi için p75 örneğine ait zümre sınırları (H=4)

	m1		m2		m3		m4	
Zümre	N_h	n_h	N_h	n_h	N_h	n_h	N_h	n_h
1	150	20	185	52	170	36	111	10
2	80	20	71	20	61	8	73	8
3	34	20	25	7	26	9	58	20
4	20	20	3	1	27	27	42	42
Tahmin varyansı	0.26729		1.0046		0.30968		**0.23957**	

Bu dört tablodan anlaşılan en önemli husus 2 zümreli durumda Neyman (m3) ve orantısız dağıtım (m4) yöntemlerinde GA ile elde edilen sonuçların, her iki örnek için de minimum tahmin varyansına

sahip olmasıdır. Bunun yanında zümre sayısı arttıkça, zümrelere dağıtılacak örnek birim sayılarının da GA ile belirlendiği m4 yöntemi diğer yöntemlere kıyasla en iyi sonucu vermektedir. Ayrıca p75 örneği için, m4 yöntemiyle GA ile elde edilen 0.23957 değerinin, Hedlin'in (1997) çalışmasında yer alan parametrelere (N1-4 = 111, 73, 51, 49; n1-4 =12, 10, 9, 49) dayanarak Eq. (1)'den hesaplanan 0.28668 değerinden daha iyi olduğu görülmektedir.

5. SONUÇ

Zümrelere göre örnekleme heterojen yapıdaki anakütlelerin örneklenmesinde sıklıkla kullanılan bir yöntemdir. Bu yöntemin uygulanması aşamasındaki en önemli problem zümre sınırlarının nasıl belirleneceğidir. Bu çalışmada GA sezgisel yaklaşımı ile 4 farklı örnek büyüklüğü dağıtım yöntemine göre bu probleme çözüm üretilmeye çalışılmıştır. Hem zümre sınırlarının hem de örnek büyüklüklerinin GA ile belirlendiği m4 yönteminin en iyi sonucu verdiği gözlenmiştir.

KAYNAKLAR

Bretthauer, Kurt M., Ross, Anthony, Shetty, Bala, (1999). "Nonlinear integer programming for optimal allocation in stratified sampling", European Journal of Operational Research, 116, 667-680.

Cochran, William G., (1963). Sampling Techniques. 2nd ed., John Wiley-Sons, Inc. USA.

Cyert, R.M. & Davidson, H.Justin, (1962). Statistical Sampling for Accounting Information, Prentice-Hall Inc., Englewood Cliffs, New Jersey, 116-127.

Dalenius, Tore, Hodges, Joseph 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, Dan, (1997). "Minimum Variance Stratification of a Finite Population", SSRC Methodology Working Paper, M03/07.

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

Holland J.H., (1975). Adaptation in natural and artificial systems. University of Michigan Press Ann Arbor.

Kish, Leslie, Anderson, Dallas W., (1978). "Multivariate and Multipurpose Stratification", Journal of the American Statistical Association, 73 (361), 24-34.

Kozak, Marcin, (2004). "Optimal Stratification Using Random Search Method in Agricultural Surveys", Statistics in Transition, 6 (5), 797-806.

Man, T., Kwong, S., (1996). Genetic algorithms: concepts and applications. IEEE Transaction on Industrial Electronics, 43 (5), 519-534.

Michalewicz, Z, (1992). Genetic Algorithms + Data Structure = Evolution Programs. Springer-Verlag, Berlin.

Nearchou, A.C., (2004). "The effect of various operators on the genetic search for large scheduling problems", International Journal of Production Economics, 88, 191-203.

Nicolini, Giovanna, (2001). "A Method to Define Strata Boundaries", Department of Economics University of Milan Italy, Departemental Working Paper, 2001-01.

Rao, Poduri S.R.S., (2000). Sampling Methodologies with Applications. Chapman & Hall/CRC, Washington D.C.