D.E.Ü.İ.İ.B.F.Dergisi
Cilt:14, Sayı:2, Yıl:1999, ss:123-136

KÜMELEME PROBLEMLERİNE KÜME ÖRTÜLEME MODELİ YAKLAŞIMI VE BİR UYGULAMA

İbrahim GÜNGÖR(*)

ÖZET

İstatistiksel verilerin kümelenmesi sorununa genellikle kümeleme analizi teknikleri ile çözüm aranmakta ancak bu teknikler anlamlılık testlerini yeterince dikkate almamaktadır. Bu çalışma ile, istatistiksel yöntemlere göre alınmış çok sayıdaki örneğin kaç farklı ana kitleden geldiği, başka bir ifade ile bu örneklerin kaç farklı kümede toplanabileceği sorununa çözüm getiren bir algoritma önerilmiştir. Algoritma, Türkiye'deki illerin güneşlenme süresi verilerine göre kümelenmesi problemine uygulanmıştır. Bu algoritma kümeleme işlemlerini anlamlılık testlerine uygun olarak yapmaktadır.

1.Giriş

İstatistiksel verilerin kümelenmesi sorununa genellikle kümeleme analizi teknikleri ile çözüm aranmaktadır. Çok değişkenli istatistiksel analizlerden biri olan kümeleme analizinin asıl amacı, eldeki bir veri grubunu belli bir benzerlik ölçüsünü dikkate alarak iki veya daha fazla kümeye bölmektir. Bu kümelerin kendi içinde yüksek seviyede homojen, kümeler arasında ise heterojen olması aranır (Hair, 1998: 473).

Kümeleme analizinde en önemli sorun küme sayısının belirlenmesidir. Bu konuda son yıllarda yoğun çalışmalar yapılıyor olmakla birlikte hala 1970'li yıllarda geliştirilmiş olan ve çok da güvenilir sonuçlar vermeyen bazı testlerden yararlanılmaktadır (Tadlıdil, 1996: 341). Ayrıca kümeleme analizinde anlamlılık testleri gibi objektif istatistiksel testler kullanılamamaktadır. Bu yüzden araştırmacılar kendilerine göre bir çok sübjektif kriterler kullanmak zorunda kalmışlardır (Hair, 1998: 499).

Bu çalışma ile, istatistiksel yöntemlere göre alınmış çok sayıdaki örneğin kaç farklı ana kitleden geldiği, veya bu örneklerin kaç farklı kümede toplanabileceği şeklindeki kümeleme sorununa çözüm getiren bir algoritma önerilmiştir. Algoritmaya göre, örneklerin geldiği ana kitlelerin normal dağılım gösterdiği ve aynı varyansa sahip oldukları varsayımı (Kartal, 1998: 66) temel alınarak yapılan çok sayıdaki ANOVA testleri sonucunda olası tüm alt kümeler bulunmakta ve bu alt kümeler dikkate alınarak oluşturulan küme örtüleme modelinin çözümü ile, optimum küme sayısına uygun bir kümeleme yapılmaktadır.

(*)Yrd.Doç.Dr., S.D.Ü., İ.İ.B.F. İşletme Bölümü.

Bu çalışmada önerilen algoritmanın kullanılması ile yapılan kümelemenin, kümeleme analizi teknikleri ile yapılan kümelemeye göre en önemli farkı; yeterince anlamlılık testlerinin yapılmış olması nedeniyle çözüm sonunda elde edilen örnek kümelerinden her birinin farklı bir ana kitleden gelmiş olduğunun kanıtlanması ve küme sayısının (farklı ana kitle sayısının) kullanılan verilerin yapısı dikkate alınarak optimum bir sayı olarak belirlenmesidir.

Yapılan literatür taraması sonucunda, kümeleme analizi ile ilgili pek çok çalışmanın yapıldığı gözlenmiş fakat, küme örtüleme modeli kullanılarak yapılmış bir kümeleme çalışmasına rastlamak mümkün olmamıştır. Küme örtüleme modeline yakın bir özellik gösteren küme bölme modeli kullanılarak Gopal ve Ramesh tarafından bir kümeleme çalışması yapılmıştır (Gopal ve Ramesh, 1995: 885-899).

2.Önerilen Algoritma

İstatistiksel yöntemlere göre alınmış çok sayıdaki örneğin kaç farklı ana kitleden geldiği, veya bu örneklerin kaç farklı kümede toplanabileceği şeklindeki kümeleme sorununun çözümü için geliştirilen algoritmanın adımları aşağıdaki gibi tanımlanmaktadır:

Adım 1: Örneklerin alındığı ana kitlelerin normal dağılım göstermesi gerekir. Normal dağılım göstermiyorsa gerekli dönüştürmeler yapılarak (Ramsey ve Schafer, 1997: 65-70) bu şart sağlanır.

Adım 2: Örneklerin alındığı ana kitlelerin aynı varyansa sahip olmaları gerekir. Gerekli testlerin (Ergün, 1995: 64) yapılması sonucunda farklı varyans durumu var ise verilere uygun dönüştürmeler yapılarak bu şart sağlanmaya çalışılır. Sağlanamıyorsa, aynı varyansa sahip olmayan ana kitlelerden geldikleri saptanan örnekler daha sonra işleme alınmak üzere ayrılır.

Adım3: Ana kitlelerinin normal dağılım gösterdiği ve aynı varyansa sahip olduğu belirlenen örnekler, ortalamaları küçükten büyüğe olacak şekilde sıralanır ve sıra numarası (1,2,...,n) verilir.

Adım 4: Ortalaması en küçük olan örnekten başlayarak sırayla bir örnek (t numaralı örnek) başlangıç olarak alınır ve ANOVA testi ile t ve t+1 numaralı örneklerin aynı ana kitleden gelip gelmedikleri araştırılır. Aynı ana kitleden geliyor ise bir sonraki örnek de dahil edilerek aynı test uygulanır. Bu ardışık test işlemlerine, ardışık ikiden çok H_A hipotezi kabul edilinceye kadar devam edilir. Aynı ana kitleden geldikleri saptanan örneklere t-1 numaralı örnek dahil edilerek aynı testler yapılır. Aynı ana kitleden geliyorlarsa bir önceki örnek de dahil edilerek aynı test tekrarlanır. Bu ardışık test işlemlerine de ardışık ikiden

çok H_A hipotezi kabul edilinceye kadar devam edilir ve bütün bu işlemler sonucunda n tane örnek kümesi belirlenmiş olur.

Adım 5: Ortalaması en büyük olan örnekten başlayarak Adım 4'deki işlemler tersinden (büyükten küçüğe doğru) uygulanarak n tane örnek kümesi daha belirlenir.

Adım 6: Adım 4 ve Adım 5'de belirlenen (2n) tane örnek kümesinden, bir birinden farklı olanlar belirlenir ve küme içindeki örneklerin ortalamasının ortalaması (küme ortalaması) küçükten büyüğe olacak şekilde örnek kümelerine sıra numarası (1,2,...,m) verilir.

Adım 7: Adım 6 ile elde edilen kümeler dikkate alınarak bir küme örtüleme tablosu oluşturulur. n satır m sütundan oluşacak olan bu tablodan gerekli indirimler yapılarak (Garfinkel ve Nemhauser, 1969: 850) elde edilecek yeni tabloyu dikkate alan bir küme örtüleme modeli, küme sayısını minimize edecek şekilde kurulur.

Adım 8: Modelin optimum çözümü ile bulunan kümelerin birden fazlasında yer alan örnekler varsa bu örneklerin ortalaması hangi küme ortalamasına en yakın ise o kümeye dahil edilip diğer kümelerden çıkarılarak, her örneğin sadece bir kümede yer aldığı çözüm elde edilir.

Adım 9: Adım 2 ile ayrılmış olan örneklerden her biri kendi ortalamasına en yakın küme ortalamasına sahip olan kümeye dahil edilerek, dahil edilen örneğin bu küme için varyans eşitliğini bozup bozmadığı araştırılır. Varyans eşitliğini bozan örnekler varsa, bu örneklerin dahil edildikleri ana kitle ile ilgili bilgiler yorumlanırken bu duruma da işaret edilir.

Adım 1 ve 2'deki işlemlerin örnek büyüklüğüne göre uygun bir test tekniğinin seçilerek yapılması gerekir. Bu çalışmada, Adım 4, 5 ve 6'daki işlemleri yapan bir bilgisayar programı da hazırlanmıştır. Algoritmanın tüm adımlarını dikkate alan bir program yapılması halinde, çok büyük boyutlu problemlerin dahi kısa sürede çözülebileceği tahmin edilmektedir.

2.1. Küme Örtüleme Problemi

Yöneylem araştırması literatüründe genellikle örtüleme problemi (covering problem) ismiyle yer alan bu problem aşağıdaki gibi formüle edilmektedir (ROSEAUX, 1991: 311):

$$\text{Min } Z = c.x \tag{1}$$

$$\text{Kısıtlar:} \quad A.x \geq e \tag{2}$$

$$X_j \in \{0,1\} \tag{3}$$

Bu modelde; $A=a_{ij}$ matrisi, mxn boyutunda olan 0 ve 1 değerlerinden oluşan bir matristir. i elemanı j alt kümesi içinde yer alıyorsa $a_{ij}=1$, diğer durumda $a_{ij}=0$ değer alır.

e, mx1 boyutunda ve bütün elemanları 1 olan bir vektördür.

c, 1xn boyutunda olup pozitif katsayılardan oluşan bir vektördür. Bu katsayılar, küme örtüleme probleminin uygulanacağı ana küme içinden önceden belirlenen olası bütün alt kümelerin (n tane) oluşum maliyetleridir.

x, x_j değişkenlerinden oluşan nx1 boyutunda bir vektördür. j alt kümesi optimum çözüm içinde yer alacak alt kümelerden biri ise $x_j=1$, diğer durumda $x_j=0$ değerini alır.

Yukarıdaki ifadelerden de anlaşıldığı gibi, küme örtüleme problemi 0-1 tamsayılı doğrusal model ile ifade edilmektedir. Bu modelin kurulmasında aşağıdaki işlemler takip edilir (Güngör ve Eroğlu, 1997: 378):

1- Küme örtüleme probleminin uygulanacağı S ana kümesinin bütün elemanları (m tane) belirlenir ve numaralandırılır.

2- Ele alınan sorunun yapısına göre, optimum çözümde yer alması olasılığı olan bütün M_j altkümeleri (n tane) eleman numaralarıyla belirlenerek bir alt kümeler seti $F=\{M_1,M_2,...,M_n\}$ oluşturulur.

3- m tane satır n tane sütundan oluşan ve mxn tane hücresi olan küme örtüleme tablosu hazırlanır. M_j alt kümesinde i elemanı yer alıyorsa, tablonun i satırı ve j sütununda bulunan gözün değeri 1 yani $a_{ij}=1$, diğer durumda $a_{ij}=0$ değeri yazılır. Bu tablodaki katsayılar, (3) eşitsizliğinde yer alan A matrisini oluşturur.

4- M_j alt kümelerinin oluşum maliyetleri (bu araştırmada yapılan uygulama çalışmasında bu katsayılar j kümesinde yer alan eleman sayısı olarak alınmıştır) hesaplanır. Bu maliyet katsayılarından oluşan 1xn boyutundaki vektör amaç fonksiyonunun katsayılarını oluşturur.

5- Uygun bir çözümde her bir elamanın en az bir alt kümede yer alması zorunluluğu olduğu için (2) eşitsizliklerinin hepsinin de sağ tarafına 1 yazılır. Bu şekilde elde edilen ve bütün elamanları 1 olan mx1 boyutundaki vektör modeldeki e vektörünü oluşturur.

Küme örtüleme probleminin herhangi bir uygun çözümüyle F 'nin bir alt kümesi elde edilir ki bu alt kümede yer alan M_j alt kümelerinin bileşimleri S ana kümesini vermek zorundadır.

Literatürde genellikle ikisi bir arada ele alınan küme bölme problemi ve küme örtüleme problemini birbirinden ayıran temel özellik; küme bölme probleminde kısıtlar A..x=e şeklinde iken, küme örtüleme probleminde A..x≥e şeklinde olmasıdır (Fisher ve Kedia, 1990: 676). Bu farklı kısıtlayıcı denklemlerden dolayı; küme bölme probleminin herhangi bir uygun çözümünde yer alan M_j alt kümelerinin bileşimleri S ana kümesini ve kesişimleri ise boş kümeyi vermek zorunda olmalarına karşın, küme örtüleme probleminde sadece bileşimlerinin S ana kümesini vermesi zorunluluğu vardır.

Küme bölme modeli ile küme örtüleme modelinin bir birine çok yakın özellik göstermesine rağmen, bu çalışmada küme örtüleme modelinin dikkate alınmasının nedeni, kurulacak küme bölme modelinin çözümsüz olması olasılığıdır. Örneğin, uygulama çalışması ile elde edilen verilerle kurulan küme bölme modeli çözümsüz çıkmıştır.

3.Algoritmanın Uygulaması

Önerilen algoritmanın uygulaması, güneşlenme süreleri açısından Türkiye'deki illerin kümelenmesi konusunda yapılmıştır. Son 15 yıl için günlük ortalama güneşlenme süreleri dikkate alındığında, il merkezlerinin en az kaç ayrı kümede toplanabileceği araştırılmıştır.

Devlet Meteoroloji İşleri Genel Müdürlüğünden 15.04.1999 tarih 1160.4.747 sayılı yazısı ekinde alınan "1984-1998 yılları için il merkezlerinin yıllara göre güneşlenme sürelerinin bir güne düşen ortalama değerleri" verilerinden, SPSS istatistik paket programı ile hesaplanan yıllık ortalama ve standart sapma değerleri (ortalama değerler küçükten büyüğe sıralandırılmış olarak) Tablo 1.de verilmiştir.

Her il için alınan 15 yıllık (15 bireylik) güneşlenme verilerinden oluşan 73 ayrı örnek bulunmaktadır. Tamamı n=15 büyüklüğündeki bu örneklerin tamamı için SPSS paket programı ile Levene Varyansların Homojenliği Testi uygulandığında α=0.01 önem seviyesine göre varyansların homojenliği reddedilmiştir. En büyük varyans değerlerine sahip olan Erzincan ve Van örnekleri hariç tutulduğunda geriye kalan 71 örnek için varyansların homojenliği kabul edilmiştir. Bu nedenle Tablo 1'de Erzincan ve Van örnekleri için numara verilmemiştir. Bu iki örnek, kümeleme analizinin dışında tutulmuş ve işlemler tamamlandıktan sonra ortalama değerlerine uygun olan bir kümeye dahil edilmiştir.

Tablo 1: İl Merkezlerinin 1984-1998 Yılları Günlük Ortalama Güneşlenme Süreleri ve Standart sapmaları (saat/gün)

İLLER	NO	ORT.	S.SAPMA	İLLER	NO	ORT.	S.SAPMA
Rize	1	4,0047	,3577	Yozgat	37	6,7747	,3571
Trabzon	2	4,1220	,3154	Ankara	38	6,7807	,3412
Ordu	3	4,2680	,3220	Kırıkkale	39	6,7900	,4064
Artvin	4	4,9460	,2489	Gaziantep	40	6,9067	,4386
Bolu	5	5,2173	,4054	Manisa	41	7,0133	,3372
Sakarya	6	5,2353	,4368	Nevşehir	42	7,1600	,2673
Sinop	7	5,3567	,3565	İskenderun	43	7,1807	,2419
Samsun	8	5,4287	,4952	Konya	44	7,2187	,3796
Kocaeli	9	5,5107	,5998	K.Maraş	45	7,2220	,4875
Bartın	10	5,6133	,4590	Çanakkale	46	7,2460	,2961
Yalova	11	5,6747	,4492	Kırıkkale	47	7,2727	,3634
Amasya	12	5,6820	,2922	Adana	48	7,3140	,2744
Tekirdağ	13	5,7107	,3829	Muş	49	7,3213	,4901
Kastamonu	14	5,7780	,5464	Batman	50	7,3347	,7176
Gümüşhane	15	5,7920	,4045	Aksaray	51	7,3927	,3051
Zonguldak	16	5,8653	,5004	Muğla	52	7,3960	,2506
İstanbul	17	5,9387	,4855	Burdur	53	7,3993	,3798
Kütahya	18	5,9453	,3080	Antakya	54	7,4140	,4496
Tokat	19	5,9600	,2002	Denizli	55	7,4347	,3638
Çankırı	20	6,0827	,2644	Elazığ	56	7,4607	,4234
Edirne	21	6,1233	,4166	Niğde	57	7,5640	,3340
Çorum	22	6,1380	,3463	Isparta	58	7,5800	,3666
Erzincan	E	6,1853	,7499	Mersin	59	7,6080	,3910
Bursa	23	6,2940	,4801	Aydın	60	7,6527	,2109
Kars	24	6,3573	,4144	Siirt	61	7,6733	,4291
Bilecik	25	6,3693	,3798	Malatya	62	7,7593	,5688
Bandırma	26	6,4420	,4031	Kilis	63	7,8133	,5174
Balıkesir	27	6,4520	,5626	Van	V	7,8847	,7193
Ağrı	28	6,4547	,4796	Karaman	64	7,9000	,2857
Bingöl	29	6,4713	,3834	İzmir	65	7,9120	,4141
Erzurum	30	6,4873	,3827	Diyarbakır	66	7,9407	,3920
Iğdır	31	6,4980	,5099	Hakkari	67	7,9500	,4316
Bitlis	32	6,5253	,4776	Urfa	68	7,9747	,3622
Eskişehir	33	6,5620	,4068	Adıyaman	69	8,1527	,3909
Afyon	34	6,6180	,3629	Mardin	70	8,1873	,4026
Kayseri	35	6,7207	,3922	Antalya	71	8,4180	,2602
Sivas	36	6,7407	,4365	Toplamda		6,6791	1,068

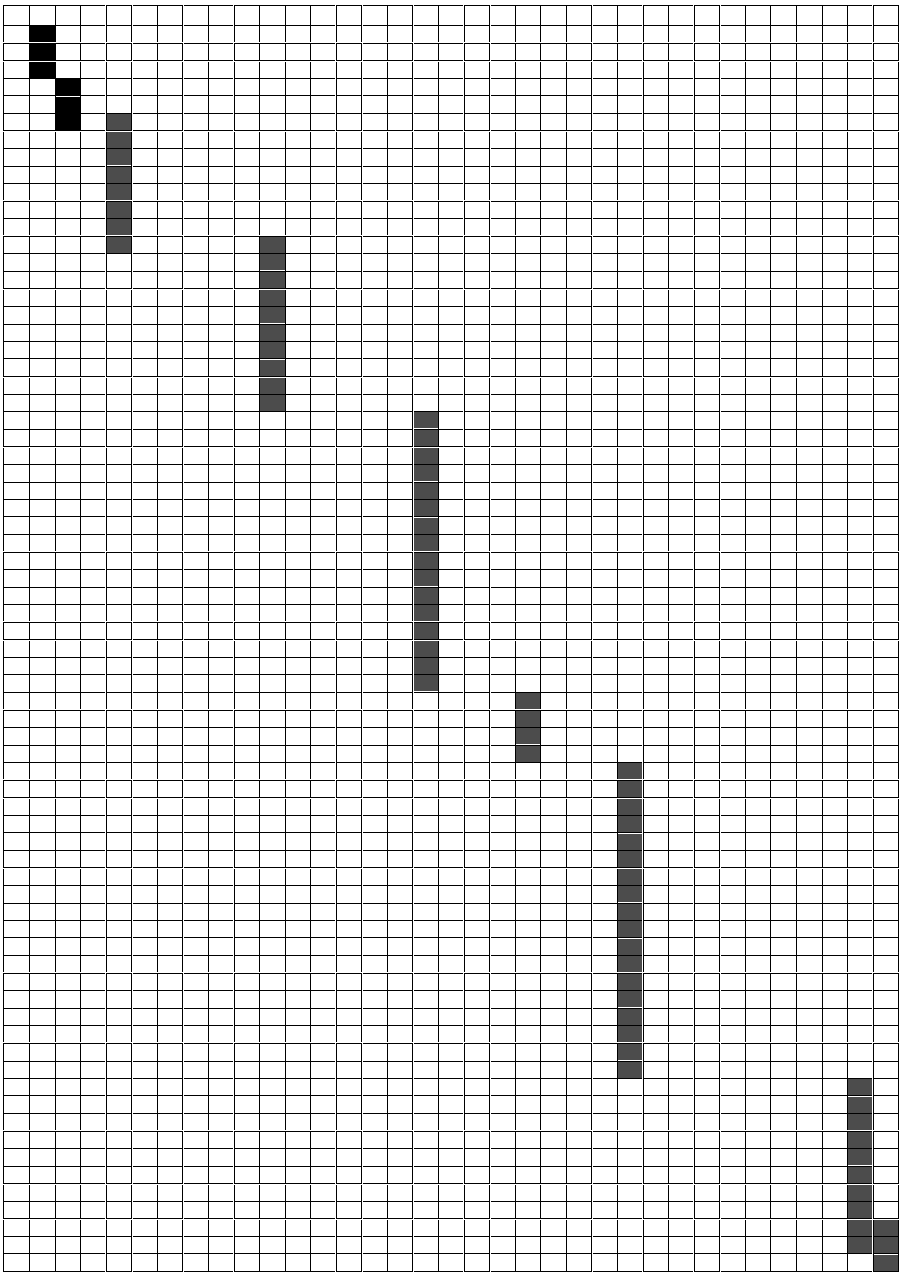
73 ayrı örnek verisinin her biri için, küçük örnekler için daha uygun olan Lilliefor normal dağılımlılık testi (Ergün, 1995: 64) yapılmış ve test işlemleri sonunda tamamının α=0.01 önem seviyesine göre normal dağlım gösterdiği gözlenmiştir.

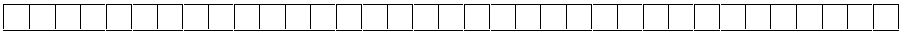
Önerilen algoritmanın 4, 5 ve 6 numaralı adımlarına uygun olarak yapılan ANOVA testleri sonucunda bulunan olası alt kümelerden oluşan "küme örtüleme tablosu" Tablo 2'de verilmiştir. Bu işlemleri yapacak şekilde bir bilgisayar programı yapılmıştır. Programa bu veriler yüklenip çalıştırıldığında işlem süresi 2 saniye olarak gözlenmiştir. Tablo 2'nin birinci satırında olası alt kümelerin kod numaraları (j), birinci sütununda ise örneklerin kod numaraları (i) yer almaktadır. Bu tablodaki a_{ij} matrisi, i örneği j alt kümesinin bir elemanı ise a_{ij}=1, diğer durumda a_{ij}=0 değerlerinden oluşmaktadır. Ancak, tablonun daha rahat bir görüntü vermesi için a_{ij}=0 değerlerinin yeri boş bırakılmıştır.

Tablo 2' nin son satırında yer alan C_j değerleri, j alt kümesinde bulunan örnek sayısını göstermektedir. Küme bölme modelinde bu sayılar amaç fonksiyonunun katsayıları olarak yazılacaktır. Amaç fonksiyonu katsayılarının tamamına C_j=1 değeri verilerek de minimum sayıda kümenin olduğu bir çözüm elde edilebilir. Ancak, bu durumda bulunan çözümde birden çok kümede yer alan örnek sayısının daha çok çıkması olasılığı vardır.

Kurulacak küme örtüleme modelinde gereksiz olarak yer alacak kısıtları dışlamak amacıyla Tablo 2'de verilen küme örtüleme tablosu sadeleştirilebilir. Bu nedenle Tablo 2'deki a_{ij} matrisinde t satırını kapsayan i satırları iptal edilebilir (Garfinkel ve Nemhauser, 1969: 850). Çünkü, t satırına ilişkin kısıt sağlandıktan sonra bu satırı kapsayan i satırlarına ilişkin kısıtların da sağlanmış olacaktır. Buna göre iptal edilmesi gereken satırlar Tablo 3'de verildiği gibi olacaktır. Tablo 2'nin bu şekilde sadeleştirilmiş hali Tablo 4'de görülmektedir.

Tablo 2: Aynı Ana Kitleden Alındığı Test Edilen Örnek Kümeleri





Tablo 3: Tablo 2'yi Sadeleştirmek İçin İptal Edilmesi gereken Satırlar

Kapsanan Satır Numarası	Kapsayan Satırların Numaraları (İptal Edilebilecek Satır Numaraları)
1	2, 3
4	5, 6
7	8, 9, 10, 11
12	13
15	14, 16
18	17
25	24, 26, 27, 28, 29, 30, 31, 32, 33
37	34, 35, 36, 38
43	44, 45, 46, 47, 48, 49, 50, 51, 52, 53,
55	56, 57
58	59
68	61, 62, 63, 64, 65, 66, 67
71	69, 70

Tablo 4: Tablo 2'in Sadeleştirilmiş Hali

i\j	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34
1	1																																	
4		1																																
7			1	1	1																													
12				1	1	1	1	1	1																									
15					1	1	1	1	1	1																								
18						1	1	1	1	1	1																							
19							1	1	1	1	1	1																						
20								1	1	1	1	1	1																					
21									1	1	1	1	1	1																				
22										1	1	1	1	1	1																			
23											1	1	1	1	1	1																		
25														1	1	1	1																	
37																1	1	1	1															
39																	1	1	1	1														
40																		1	1	1	1													
41																			1	1	1	1												
42																				1	1	1	1											
43																					1	1	1	1										
55																						1	1	1	1	1	1	1						
58																							1	1	1	1	1	1	1	1	1			
60																								1	1	1	1	1	1	1	1	1		
68																															1	1	1	
71																																		1

Tablo 3'deki sadeleştirilmiş küme örtüleme tablosu dikkate alındığında kurulacak küme örtüleme modeli aşağıdaki gibi olacaktır:

EnkZ=$3X_1+3X_2+7X_3+8X_4+10X_5+11X_6+11X_7+11X_8+10X_9+10X_{10}+9X_{11}+7X_{12}+14X_{13}+14X_{14}+15X_{15}+16X_{16}+16X_{17}+15X_{18}+10X_{19}+3X_{20}+15X_{21}+17X_{22}+18X_{23}+18X_{24}+17X_{25}+17X_{26}+15X_{27}+10X_{18}+10X_{29}+11X_{30}+11X_{31}+10X_{32}+10X_{33}+3X_{34}$

Kısıtlar: $X_1 \geq 1$

$X_2 \geq 1$

$X_3 + X_4 + X_5 \geq 1$

$X_4 + X_5 + X_6 + X_7 + X_8 + X_9 \geq 1$

$X_5 + X_6 + X_7 + X_8 + X_9 + X_{10} \geq 1$

$X_6 + X_7 + X_8 + X_9 + X_{10} + X_{11} \geq 1$

$X_7 + X_8 + X_9 + X_{10} + X_{11} + X_{12} \geq 1$

$X_8 + X_9 + X_{10} + X_{11} + X_{12} + X_{13} \geq 1$

$X_9 + X_{10} + X_{11} + X_{12} + X_{13} + X_{14} \geq 1$

$X_{10} + X_{11} + X_{12} + X_{13} + X_{14} + X_{15} \geq 1$

$X_{11} + X_{12} + X_{13} + X_{14} + X_{15} + X_{16} \geq 1$

$X_{13} + X_{14} + X_{15} + X_{16} + X_{17} \geq 1$

$X_{16} + X_{17} + X_{18} + X_{19} \geq 1$

$X_{17} + X_{18} + X_{19} + X_{20} \geq 1$

$X_{18} + X_{19} + X_{20} + X_{21} \geq 1$

$X_{19} + X_{20} + X_{21} + X_{22} \geq 1$

$X_{20} + X_{21} + X_{22} + X_{23} \geq 1$

$X_{21} + X_{22} + X_{23} + X_{24} \geq 1$

$X_{22} + X_{23} + X_{24} + X_{25} + X_{26} + X_{27} + X_{28} \geq 1$

$X_{23} + X_{24} + X_{25} + X_{26} + X_{27} + X_{28} + X_{29} + X_{30} + X_{31} \geq 1$

$X_{24} + X_{25} + X_{26} + X_{27} + X_{28} + X_{29} + X_{30} + X_{31} + X_{32} \geq 1$

$X_{31} + X_{32} + X_{33} \geq 1$

$X_{34} \geq 1$

$X_{ij} = 0$ veya 1

Modelin çözümü için Pentium 166 MMX (166 MHz) kişisel bilgisayar ve LINDO paket programı (Demo LINDO/PC Release 6.01 1997) kullanılmıştır. Çözüm süresi 1 saniye olarak gözlenmiştir. Optimum çözüm sonuçları aşağıdaki gibi bulunmuştur:

$X_1 = X_2 = X_4 = X_{10} = X_{16} = X_{20} = X_{24} = X_{33} = X_{34} = 1$

Diğer $X_j = 0$, Enk Z = 75

Bu çözüm sonucuna göre bulunan alt kümeler Tablo 2'de görülen j=1, 2, 4, 10, 16, 20, 24, 33 ve 34 indisli (koyu renkli olarak işaretlenmiş) kümelerdir. Bu kümelere dikkatle bakıldığında i=6,13,69 ve 70 indisli örneklerin iki kümede birden yer aldıkları görülmektedir. Min.Z değerinin 75-71=4 fazla çıkmasının nedeni budur. Bu örneklerin, iki kümeden hangisine daha yakın ise sadece o kümede yer alması uygun olacaktır. Bu amaçla yapılan ANOVA testleri sonucunda daha düşük bir güven katsayısı verdiğinden 6 ve 13 numaralı örneklerin 4 numaralı kümeden, 69 numaralı örneğin 34 numaralı kümeden ve 70 numaralı örneğin 33 numaralı kümeden çıkartılması, alt küme sınırlarının netleşmesi açısından uygun olacaktır. Bu düzenlemelerin yanında Tablo 2 ve Tablo 1 birlikte dikkate alındığında, en uygun kümeleme Tablo 5'de görüldüğü şekilde olmaktadır.

Örneklerin tümü birlikte dikkate alınarak yapılan Levene testine göre varyansların eşitliği sağlanmadığı için daha önce ayrılmış olan Erzincan ve Van iline ait örnekler de, ortalama değerine en yakın olan kümelere dahil edilmişlerdir. Bu ilavelerden sonra yapılan Levene testlerine göre, 8. kümeye dahil edilen Van örneği bu küme içindeki varyans eşitliğini bozmamıştır. 4. kümeye dahil edilen Erzincan örneği ise varyans eşitliğini bozmaktadır. Erzincan örneğinin 5. Kümeye dahil edilmesi durumunda ise, varyans eşitliği bozulmamakta ancak bu kümeyi temsil eden ana kitleden gelmiş olması, α=0.0089 önem seviyesine göre kabul edilebilmektedir. Bu sonuçlara göre, Erzincan örneğinin 5. Kümeye dahil edilmesi uygun olmaktadır.

Yapılan uygulama çalışmasının sonucuna göre, Türkiye'deki iller güneşlenme süresi verilerine göre 9 farklı küme oluşturmaktadır. Her kümede yer alan iller Tablo 5'de açık olarak görülmektedir. Bu kümelerde yer alan örneklerin alındığı ana kitlelerin ortalama ve standart hata değerleri her kümenin altında görülmektedir.

Tablo 5: Türkiye'deki İllerin Güneşlenme Süresi Verilerine Göre Kümelenmesi

Küme Örtüleme Modeli

1. KÜME	**4. KÜME**	**6. KÜME**	**8. KÜME**
1 Rize	13 Tekirdağ	39 Kırıkkale	61 Siirt
2 Trabzon	14 Kastamonu	40 Gaziantep	62 Malatya
3 Ordu	15 Gümüşhane	41 Manisa	63 Kilis
Ort= 4,03 - 4,23	16 Zonguldak	42 Nevşehir	V Van
S.Hata= 0,0510	17 İstanbul	*Ort= 6,87 - 7,07*	64 Karaman
2. KÜME	18 Kütahya	S.Hata= 0,0496	65 İzmir
4 Artvin	19 Tokat	**7. KÜME**	66 Diyarbakır
5 Bolu	20 Çankırı	43 İskenderun	67 Hakkari
6 Sakarya	21 Edirne	44 Konya	68 Urfa
Ort= 5,02 - 5,25	22 Çorum	45 K.Maraş	69 Adıyaman
S.Hata= 0,0579	*Ort*= 5,87 - 6,00	46 Çanakkale	*Ort= 7,82 - 7,97*
3. KÜME	*S.Hata*= 0,0336	47 Kırıkkale	*S.Hata= 0,0382*
7 Sinop	**5. KÜME**	48 Adana	
8 Samsun	E Erzincan	49 Muş	**9. KÜME**
9 Kocaeli	23 Bursa	50 Batman	70 Mardin
10 Bartın	24 Kars	51 Aksaray	71 Antalya
11 Yalova	25 Bilecik	52 Muğla	*Ort= 8,17 - 8,43*
12Amasya	26 Bandırma	53 Burdur	*S.Hata= 0,0645*
Ort= 5,45 - 5,64	27 Balıkesir	54 Antakya	
S.Hata= 0,0482	28 Ağrı	55 Denizli	
	29 Bingöl	56 Elazığ	
	30 Erzurum	57 Niğde	
	31 Iğdır	58 Isparta	
	32 Bitlis	59 Mersin	
	33 Eskişehir	60 Aydın	
	34 Afyon	*Ort= 7,34 - 7,44*	
	35 Kayseri	*S.Hata= 0,0245*	
	36 Sivas		
	37 Yozgat		
	38 Ankara		
	Ort= 6,46 - 6,57		
	S.Hata= 0,0293		

4. Sonuç

Bu çalışmada, istatistiksel yöntemlere göre alınmış çok sayıdaki örneğin kaç farklı ana kitleden geldiği, veya bu örneklerin kaç farklı kümede toplanabileceği sorununa çözüm getiren bir algoritma önerilmiştir.

Önerilen algoritmanın, Türkiye'deki illerin güneşlenme süresi açısından kaç kümede toplanabileceği şeklindeki kümeleme probleminin çözümü için bir uygulaması yapılmıştır. Uygulama çalışması sonunda, Türkiye'deki illerin güneşlenme süresi açısından 9 farklı küme oluşturduğu ortaya çıkmış ve her kümede yer alan illere ilişkin örneklerin aynı ana kitleye ait oldukları kanıtlanmıştır. Bu sonuçlar, Türkiye'de potansiyel güneş enerjisinin hesaplanmasında, güneş enerjisi ile ilgili yatırımlarda öncelikli yerlerin belirlenmesinde, bu yatırımların karlılığı açısından bölgeler arasındaki farklılıkların ortaya konulmasında, tarım ve ormancılık alanında yapılacak yatırımlarda ve bilimsel araştırmalarda veri olarak kullanılabilir.

İstatistiksel verilerin kümelenmesi sorununa genellikle kümeleme analizi teknikleri ile çözüm aranmaktadır. Ancak bu teknikler anlamlılık testlerini yeterince dikkate almamaktadır. Bu çalışma ile önerilen algoritmaya göre yapılan kümeleme işlemlerinde gerekli testler yapılmakta ve çözüm sonucunda elde edilen kümelerdeki örneklerin aynı ana kitleden gelmiş olmaları sağlanmaktadır. Bu yöntemle küme sayısı, kullanılan verilerin yapısına göre optimum bir sayı olarak belirlenebilmektedir.

ABSTRACT

A solution to the problem of clustering statistical data is usually approached by clustering analyzing techniques. However, these techniques do not take the significance tests into consideration enough. The question in this study is to determine from how many different populations the samples are coming, when numerous samples are taken statistically. In other words the problem is to find out in how many different clusters the samples can be formed. In this study an algorithm for this problem is suggested. The algorithm has been applied to cluster the cities in Turkey according to the amount of daylight. The clustering process used in this algorithm is appropriate for significance tests.

KAYNAKÇA

ERGÜN, Mustafa (1995), *Bilimsel Araştırmalarda Bilgisayarla İstatistik Uygulamaları: SPSS For Windows*, Ankara: Ocak Yayınları.

FISHER, M.L. ve KEDIA, P. (1990), "Optimal Solution of Set Covering /Partitioning Problems Using Dual Heuristics", *Management Science*, 36, 675-686.

GARFİNKEL, R.S. ve NEMHAUSER,G.L. (1969), "The Set Partitioning Problem: Set Covering With Equality Constraints", *Operations Research*, 17, 848-856.

GOPAL, R.D. ve RAMESH R. (1995), "The Query Clustering Problem: A SetPartitioning Approach", *IEEE Transactions On Knowledge and Engineering*, 7, 885-899.

GÜNGÖR, İbrahim ve EROĞLU, Abdullah (1997), "Küme Örtüleme Problemi ve Bir Uygulama", *S.Demirel Üniversitesi İktisadi ve İdari bilimler Fakültesi dergisi,* 2, 377-386

HAIR, F.H., ANDERSON, R.E., TATHAM, R.L. ve BLACK, W.C. (1998), *Multivariate Data Analysis*, New Jersey: Prentice-Hall.

KARTAL, Mahmut (1998), *Bilimsel Araştırmalarda Hipotez Testleri*, Erzurum: Şafak Yayınevi.

RESEAUX, (1991), *Exercices et Problèmes Résolus De Recherche Opérationnelle*, Tome 3, Paris Milan Barcelona Bonn: Mason.

TATLIDİL, Hüseyin (1996), *Uygulamalı Çok Değişkenli İstatistiksel Analiz*, Ankara: Akademi matbaası.

RAMSEY F.L. ve SCHAFER D.W. (1997), *The Statistical Sleuth: A Corse in Methods of Data Analysis*, Belmont CA: Duxbury Press.