Makale Gönderim Tarihi: 08.03.2021
Yayına Kabul Tarihi: 15.05.2021

GUIDE – Hücresel Küme Ayrıştırma Uygulaması: Yeraltı suyu Arsenik İçeriği Durum Çalışması

GUIDE – Cell Declustering Application: Case Study on Groundwater Arsenic Content

Güneş Ertunç[1]*

[1] *Hacettepe Üniversitesi, Mühendislik Fakültesi, Maden Mühendisliği Bölümü, 06800 Ankara*

** Sorumlu yazar, e-mail: gertunc@hacettepe.edu.tr, ORCID: 0000-0003-0914-2745*

Özet

Bir sahada rastlantı değişkenin çok yüksek veya çok düşük aykırı değerlere sahip alanlarında daha yoğun örnekleme, değişkenin dağılımını daha iyi anlayabilmek adına, genellikle tercihli olarak yapılmaktadır. Ancak bu durum, bu değişkenin gerçek dağılımı ile yanlı bir farkın ortaya çıkmasına neden olmaktadır. Tercihli örneklemeden kaynaklanan yanlı ortalama farkını önlemek için, numuneler düzenli aralıklarla veya rastgele toplanmalıdır. Alternatif olarak, değişkenin saha genelindeki ortalamasını doğru bir şekilde ortaya koymak için hücresel küme ayrıştırma yöntemi (cell declustering) uygulanmalıdır. GUIDE (**G**eostatistical **U**tility **i**n **D**omaining and **E**stimation) mekansal verilerin istatistiksel analizi, krigleme yöntemi ile kestirimi ve jeolojik homojen bölgelerin belirlenmesi için çözüm araçları barındıran modüler uygulamalar bütününden oluşan bir bilgisayar programıdır. Bu çalışmanın amacı, GUIDE ile hücresel küme ayrıştırma uygulaması sunmaktır. Yeraltı suyu arsenik ölçümlerinin analiz değerleri ile programın işleyişi durum çalışması olarak verilmiştir. Durum çalışmasında yüksek arsenik değerlerinin tercihli olarak yoğun şekilde örneklenmesi sonucunda ortaya çıkan istatistikler ile hücresel küme ayrıştırma uygulandıktan sonra oluşan istatistikler ve histogram sonuçları karşılaştırılmıştır.

Anahtar Kelimeler: Hücre küme ayrıştırma, istatistik, yeraltı suyu, arsenik

Abstract

In order to have better understanding on the distribution of the variable, more intensive sampling in the areas of the random variable with very high or very low outliers in a field is usually done preferentially. However, this situation causes a biased difference with the actual distribution. Samples should be collected at regular intervals or randomly to avoid biased mean difference due to preferential sampling. Alternatively, the cell declustering method should be applied to accurately represent the statistics of the variable. GUIDE (Geostatistical Utility in Domaining and Estimation) is a computer program consisting of modular applications that contain tools for exploratory data analysis of spatial data, kriging estimation and domaining. Within the scope of this study, the cell declustering application of the program is presented and the analysis of groundwater arsenic measurements is given as a case study. In the case study, the statistics obtained as a result of preferential sampling of high arsenic values and the statistics and histogram results after cell declustering method is compared.

Keywords: *Cell declustering, statistics, groundwater, arsenic*

1.Giriş

Çoğu zaman bir bölgesel değişkene ilişkin kararlar, mekansal verinin düzenli olmayan konumsal dağılımına göre istatistiksel analizine göre verilmektedir. Özniteliğin değerleri, sahada var olan heterojenlik nedeniyle genellikle farklı şekilde etkilenir. Örneğin, cevherleşme, yeraltındaki jeolojik ortamlara; veya yeraltı suyunda kirletici konsantrasyonları suyun akış yönlerine bağlanabilir. Değişkenin sahanın bir bölümünde anormal derecede yüksek veya düşük değerlere sahip olması çok olağan ve sıklıkla karşılaşılan bir durumdur. Aykırı değerlerin olduğu bölgelerde tercihli şekilde daha sık örneklem alınmasının özniteliğin yapısını daha iyi anlayabilmek için olduğu göz önüne alındığında anlaşılabilir bir durumdur. Ancak, bu değişkenin gerçek dağılımında yanlı bir fark ortaya çıkacağı mutlaka dikkate alınması gereken bir durumdur.

Veriler üzerindeki heterojen sahaya özgü etkilerin neden olduğu düzensiz örnekleme ve bu yanlılığı (biasness) hesaba katmak için, mekansal verileri analiz etmek için önemli bir jeoistatistiksel araç olarak hücre küme ayrıştırma yöntemi literatürde bir çok araştırmacı tarafından sunulmuştur (Deutsch ve Journal, 1998; Deutsch ve ark., 1999; Pyrcz ve Deutsch, 2002; Renard ve ark., 2020).

Belirsizlik değerlendirme ve jeoistatistiksel benzetim gibi analizlerde rastlantı değişkenini temsil eden dağılımlar veya kategorik özniteliğe ilişkin oranlar temel girdi parametreleridir. Bu dağılımları veya oranları belirlemede kümeleme teknikleri oldukça sıklıkla kullanılmaktadır. Bu teknik ile her bir veriye (z_i=1,...,n) çevresindeki verilere yakınlığına bağlı olarak bir ağırlık (w_i=1,...,n) atanır. Birbirlerine yakın olan verilere daha az ağırlık atanırken, birbirinden çok uzak olan veriler daha fazla ağırlık kazanır. Buradaki temel varsayım, daha yakın verilerin tercihen düşük veya yüksek değerli alanlardan toplandığıdır. Kümeleme tekniği ile parametrik olmayan dağılım ve özet istatistikler atanan ağırlıklar kullanılarak oluşturulur.

GUIDE (Geostatistical Utility in Domaining and Estimation) mekansal verilerin istatistiksel analizi, krigleme yöntemi ile kestirimi ve jeolojik homojen bölgelerin belirlenmesi için çözüm araçları barındıran modüler uygulamalar bütününden oluşan ve yazar tarafından kodlanan bir bilgisayar programıdır. MATLAB CompilerTM ile derlenmiştir ve çalışmaya konu olan Hücre Küme Ayrıştırma uygulama modülü http://yunus.hacettepe.edu.tr/~gertunc/HKA.rar adresinden ücretsiz bir şekilde indirilebilir.

2.Yöntem

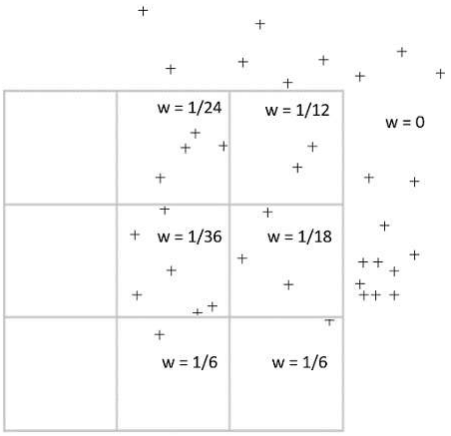
İlk olarak Journel (1983) tarafından sunulan hücresel küme ayrıştırma yöntemi, veri setinin yayıldığı alana uyarlanan iki boyutlu bir ızgara ve bu ızgarayı oluşturan hücrelere düşen örnek noktalarının bulunma yüzdelerine göre, örneklere ağırlık atanması esasına dayanır. Bu ızgarayı oluşturan hücreler, jeoistatistiksel kestirim veya benzetimde kullanılan blok kavramından tamamen farklıdır. Hücre boyutu, bu yöntemde seyrek örneklenen bölgelerdeki veriler arasındaki uzaklığı ile ilişkilidir. Bu yüzden, yöntem ismini Türkçeleştirirken blok yerine hücre kelimesi daha uygun bulunmuştur. Hücrede yalnızca bir tane örnek bulunması durumunda ağırlık aşağıdaki formül ile bulunur.

$$w_i = \frac{1}{\frac{n_{dh}}{n_{vs}}}$$

Burada, w_i: *i* lokasyonundaki örneğe ait ağırlığı, n_{dh}: ızgara yapısındaki toplam dolu hücre sayısını ve n_{vs}: hücre içindeki veri sayısını ifade etmektedir.

Tüm örneklere atanan ağırlıklar ile ağırlıklandırılmış istatistikler (Q1, ortalama, Q3, ve histogram dağılımı) hesaplanır. Bu sayede ağırlıklandırılmış veriler ile hesaplanan istatistikler, tercihli örneklem durumunda yanlı ve yanıltıcı istatistikleri düzelmiş olur.

Şekil 1'de 18 örnek lokasyonu ve 6 dolu hücre için oluşan durum ve ağırlıklar göreselleştirilmiştir.

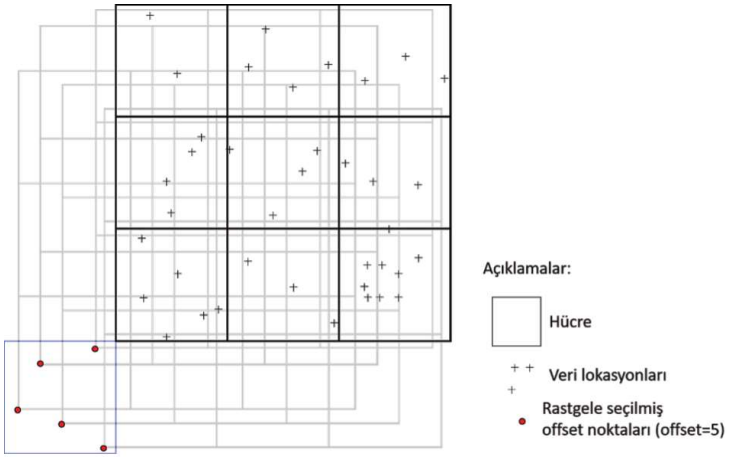


Şekil 1 Dolu hücreler ve örneklerin herbirinin aldıkları ağırlıklar.

Şekil 1'deki duruma göre, ızgaranın en güneyinde orta ve sağ alt hücrelerinde yalnızca birer örnek bulunduğu için her iki örnek ağırlık olarak 1/6 değerini almışlardır. Diğer hücrelerdeki durumlarda artan örneklere göre ters orantı ile ağırlıklar düzenlenmiştir. Bu durumda hücre içine düşen örnek sayısı arttıkça, bu örneklere atanan ağırlıklar azalmaktadır.

2.1 Offset sayısı

Izgaranın hangi lokasyondan başladığının örnek lokasyonlara atanan ağırlıklar üzerinde doğrudan etkisi olduğundan, her numuneye ortalama bir ağırlık sağlamak için çeşitli sayıda başlangıç noktası kullanılır. Rastlantısal seçilen başlangıç noktası sayısı "*offset sayısı*" ile ifade edilmektedir. Endüstri pratiği göz önüne alındığında bu değer 25 ile 100 arasında değişecek şekilde belirlenmiştir. Şekil 2'de, offset sayısının 5 olarak alındığı durum görselleştirilmiştir.



Şekil 2 Beş farklı offset noktasına göre ızgaralar.

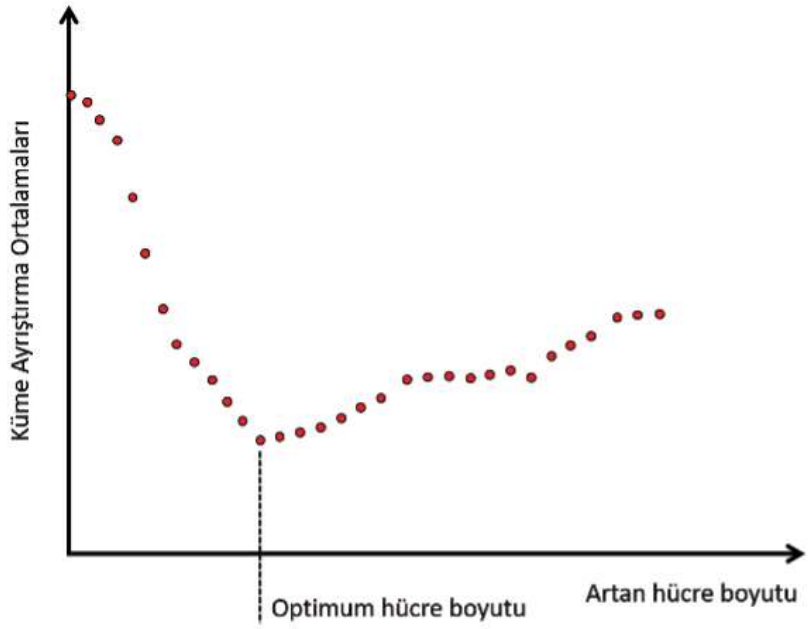
Şekil 2'ye göre 5 farklı durumda örneklere atanan ağırlıkların ortalamaları, seçili hücre boyutuna göre atanan ağırlıklar olarak belirlenir.

2.2 Hücre boyutu

Hücre küme ayrıştırma yönteminin bir diğer önemli parametresi hücre boyutudur. Çok küçük seçilen hücre boyutunda, dolu hücre sayısı örnek sayısına eşit olacağından her bir örneğe atanan ağırlık 1'e eşit olacaktır. Çok büyük boyutlu bir hücre durumunda ise tüm örnek lokasyonları eşit şekilde ağırlandırılacaktır. Her iki durumda ağırlıklı ortalama değerleri birbirine eşittir. Bu yüzden, hücre küme ayrıştırma yönteminde optimum hücre boyutu seçebilmek için küçükten büyüğe doğru değişen her hücre boyutu için ayrı ayrı hesaplama yapılmaktadır.

Optimum hücre boyutu, seyrek örneklenen lokasyondaki verilerin aralığıdır. Bu optimum boyut tam olarak bilinmemektedir ve doğru boyutun seçilmesine yardımcı olmak için bir dizi hücre boyutu girdi parametresi olarak düşünülmelidir. Bu nedenle, en büyük hücre boyutu çok büyük ayarlanmamalıdır.

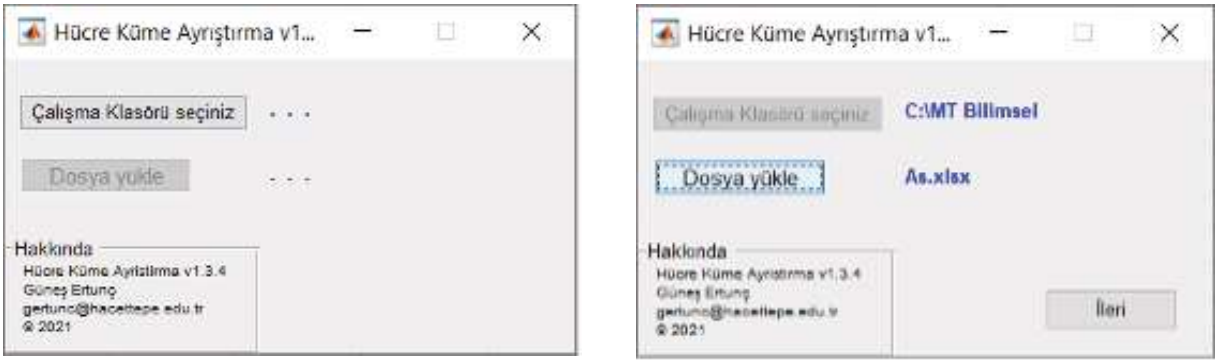
Optimum hücre boyutunu seçmek için en iyi yaklaşım, çeşitli hücre boyutları için küme ayrıştırma ortalamayı değerlendirmek ve bu ortalamayı en aza indirmeye veya en yüksek değere ulaştırmaya yakın bir değer seçmektir. Şekil 3'te, küme ayrıştırılan ortalamanın artan hücre boyutuyla değişimi verilmiştir.



Şekil 3 Değişen hücre boyutlarına göre hücre küme ayrıştırma ortalamaları

3. GUIDE – Hücresel Küme Ayrıştırma Modülü ve Durum Çalışması

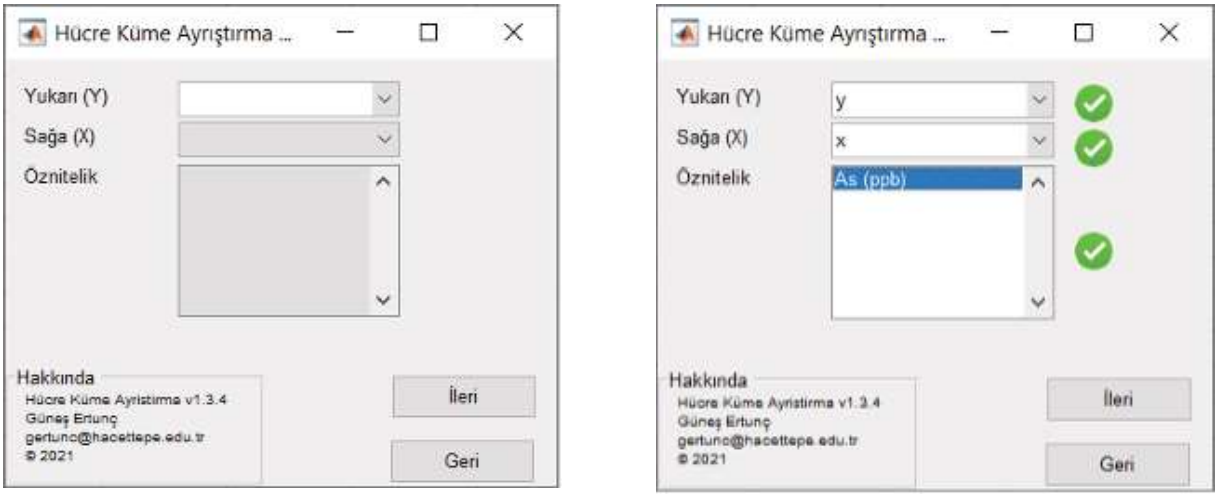
Hücre Küme Ayrıştırma modülü toplam 4 girdi ekranından oluşmaktadır. Öncelikle, kullanıcı MS Office Excel formatındaki dosyasının bulunduğu klasörü "Çalışma Klasörü" olarak belirler.



(a) (b)

Şekil 4 a) Dosya yükleme ekranı, b) Yüklenmiş dosya ekranı

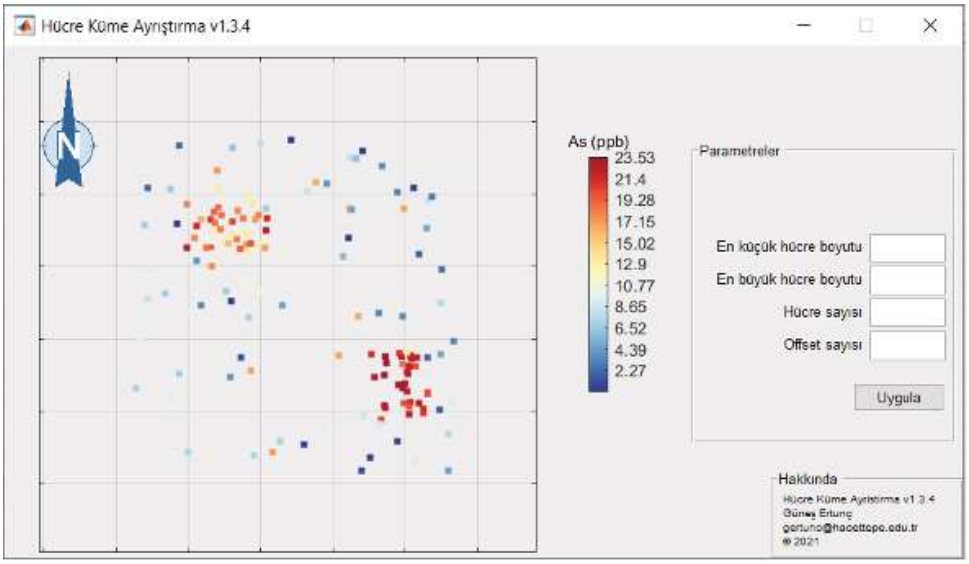
Uygulama iki boyutlu olduğu için kullanıcı bir sonraki öznitelik seçim ekranında Yukarı (Y), Sağa (X) koordinat sütunlarını seçtikten sonra öznitelik seçimi yapar (Şekil 5).



(a) (b)

Şekil 5 a) Öznitelik seçim ekranı, b) Sütunların eşleştirilmesi

İleri butonu ile kullanıcı hücre küme ayrıştırma yöntemi için gerekli parametrelerin girildiği ekrana yönlendirilir (Şekil 6). Seçilen özniteliğin tematik olarak gösterildiği bu ekranda kullanıcı sırasıyla en küçük hücre boyutunu, en büyük hücre boyutunu, bu iki değer arasında kaç hücre boyutu olacağını ve offset sayısını girer.



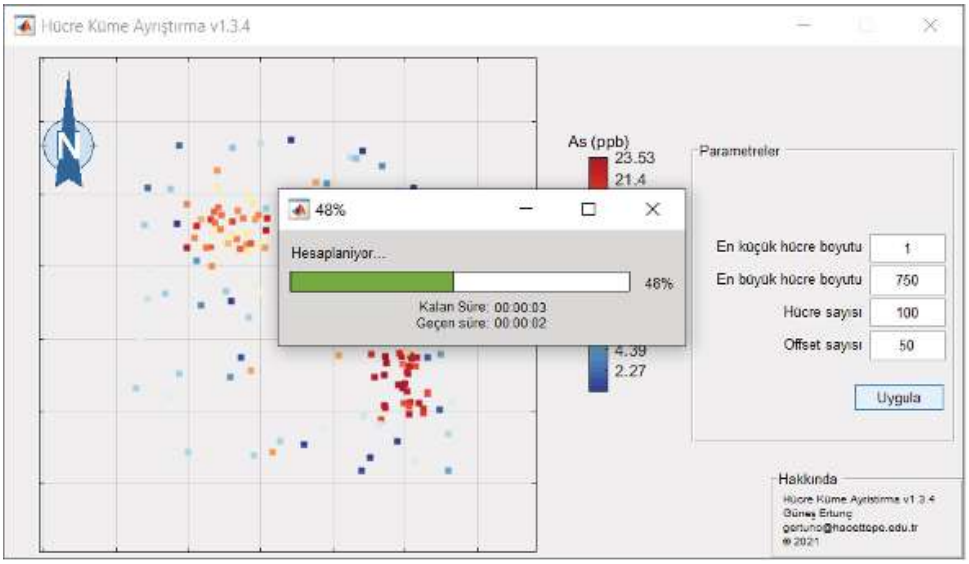
Şekil 6. Hücre küme ayrıştırma parametre ekranı

3.1 Durum Çalışması

Çalışmada kullanılan veri seti, µg/L cinsinden ölçülen arsenik konsantrasyonlarını içeren iki boyutlu uzayda dağılmış 153 yeraltı suyu örneğinden oluşan sentetik bir arsenik değerlerinden oluşur (Şekil 6). Yaklaşık 1 km^2'lik alana yayılan veri setinde örnekler arasındaki mesafeye ilişkin tanımlayıcı istatistikler Tablo 1'de yer almaktadır. Bu veri setine http://yunus.hacettepe.edu.tr/~gertunc/decluster.rar adresinden erişilebilir.

Tablo 1 Örnekler arasındaki mesafelere ilişkin tanımlayıcı istatistikler.

En küçük değer	2.3
En büyük değer	1278
Ortalama	459.5
Q1	249.52
Ortanca (Q2)	481
Q3	648
Varyans	58334



Şekil 7. Değişen hücre boyutlarına göre hücre küme ayrıştırma ortalamaları

Tablo 2 Arsenik verisi tanımlayıcı istatistikleri

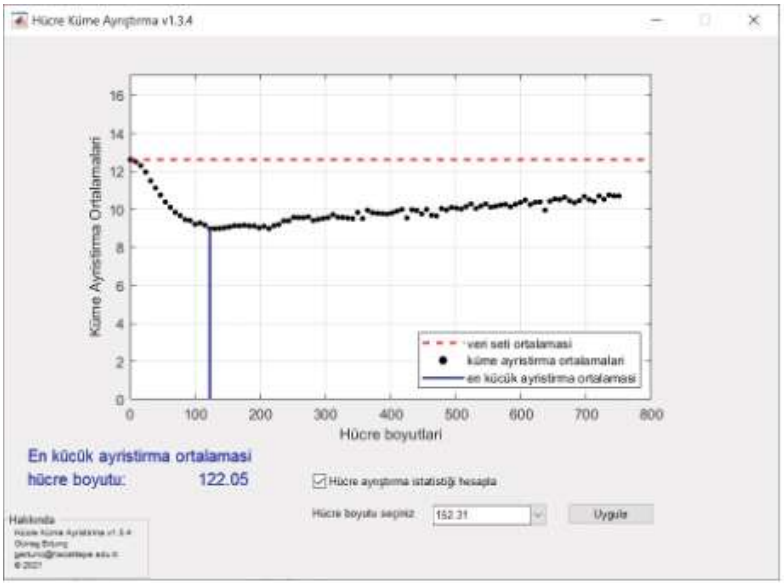
Veri sayısı	153
En küçük değer	0.148
En büyük değer	23.53
Ortalama	12.61
Q1	6.78
Ortanca (Q2)	12.76
Q3	19.36
Varyans	53.87

Bu veri seti için seçilen hücre küme ayrıştırma parametreleri aşağıda sıralanmıştır.

Tablo 3. Hücre Küme Ayrıştırma Parametreleri

En küçük hücre boyutu	1 m
En büyük hücre boyutu	750 m
Hücre sayısı	100
Offset sayısı	50

Seçilen hücre sayısı 100 olduğundan, hücre boyutu 1 m'den başlayarak 7.56 m'lik artışlar ile 750 m'ye kadar ayrıştırılmış ortalamalar hesaplanmıştır ((750-1)/99). Parametrelere göre çözüm sonuçları Şekil 8'de verilmektedir. Şekil 8'de kesikli kırmızı çizgi orijinal verinin ortalama değeridir (12.61).



Şekil 8. Değişen hücre boyutlarına göre hücre küme ayrıştırma ortalamaları

En küçük ayrıştırma ortalaması, hücre boyutu 122.05 m olduğu durumda hesaplanmıştır. Bu durumda, hücre ayrıştırma istatistikleri bu boyut seçilerek yapılmıştır. Seçilen boyuta göre ortaya çıkan histogram ve istatistikler Şekil 9'da verilmektedir. Normal dağılım göstermeyen veri setinin, hücre ayrıştırma sonrasında uniform dağılıma yakın bir davranış gösterdiği görülmektedir.

Şekil 9. Değişen hücre boyutlarına göre hücre küme ayrıştırma ortalamaları

4 Sonuç

DSÖ (Dünya Sağlık Örgütü), içme suyu arsenik içeriği yönergesine (WHO, 2021) göre içme suyunun arsenik içeriğinin 10 *u*g/L geçmemesi gerektiğini belirtmektedir ve orijinal veri setinin arsenik ortalaması 12.60 *u*g/L, bu sınırın üzerindedir. Yüksek değerlerin olduğu bölgelerden fazla örnek toplama, başka bir deyişle tercihli örnek toplama stratejisi gereği çıkan bu sonucun yanlı ve dahası, yanıltıcı bir sonuçtur.

Orijinal veri histogramı ile hücre küme ayrıştırma ile elde edilen histogram karşılaştırıldığında,

orijinal verinin yüksek değerlerinin daha sık olduğu görülmektedir. Veri seti incelendiğinde, yüksek değerlerin daha sıklıkla örneklendiği ve hücre kümeleme ayrıştırma yönteminin bu değerlerin ağırlıklarını düşürerek yanlı istatistikleri düzelten nitelikte sonuçlar ortaya koyduğu anlaşılmaktadır. Düzeltilen istatistikler sonucunda saha genelindeki ortalamanın 9.03 *u*g/L olduğu ve kabul edilebilir sınırlar içerisinde kaldığı ortaya konmuştur.

Dikkat edilmesi gereken diğer bir bulgu ise, örnek değerlerinin değişmemesidir. Hücre küme ayrıştırma yönteminde verinin kendisi değil, onu karakterize eden istatistiklerin -varsa tercihli örnekleme stratejisi yüzünden yanlı istatistiklerin- düzeltilmesidir. Düzeltilen istatistikleri oluşturan temel unsur, örnek değerlerlerinin birbirlerine yakınlığı esas alınarak atanan ağırlıklarıdır. Bu nedenle en küçük ve en büyük değerler hücre küme ayrıştırması yöntemi uygulamasından etkilenmezler (Şekil 9).

Kaynaklar

Deutsch, C.V., Journel, A.G., 1998. GSLIB: Geostatistical Software Library: and User's Guide,2nd Ed. New York: Oxford University Press.

Deutsch, C.V., Frykman, P., Xie, Y.L., 1999. Declustering with Seismic or "soft" Geologic Data, Centre for Computational Geostatistics Report One 1998/1999, University of Alberta.

Pyrcz, M.J., Deutsch, C.V, 2002. Declustering and Debiasing. In Centre for Computational Geostatistics, Paper 62, Annual Report 4.

Renard, D., Bez, N., Desassis, N., Beucher, H., Ors, F., Laporte, F., 2020, RGeostats: The Geostatistical package. MINES ParisTech

WHO, 2021, https://www.who.int/news-room/fact-sheets/detail/arsenic, (Erişim tarihi: 25/03/2021)