PARÇALI DAİRESEL VERİ GÖRSELLEŞTİRME TEKNİĞİNİN R "GGPLOT2" PAKETİ İLE NOKTASAL TABANLI OLUŞTURULMASI

Sergen CANSIZ
sergencansiz@gmail.com

ÖZET

Bu makalede çok değişkenli ve çok fazla gözleme sahip olan veri setlerinin görselleştirilmesinde kullanılan parçalı dairesel veri görselleştirme yönteminin (Circle Segments) iki boyutlu düzlemlerde noktasal olarak nasıl uygulanabileceğine değinilmiştir. Bu uygulama sırasında oluşturulan görselde noktaların hangi oranla, nasıl bir algoritma izlenerek dağılacağı ve noktaların renklendirilmesi görünür ışık görüngesine göre hangi değerlerle eşleştirileceği tartışılmıştır.

Anahtar Kelimeler: Çok değişkenli veri görselleştirmesi; Noktasal tabanlı görselleştirme, Parçalı Dairesel veri görselleştirmesi

ABSTRACT

In this article, it is mentioned how to apply fragmental circle data visualisation method used in the visualisation of data sets with many variables and observations in two-dimensional planes in a point-based way. The visual created during the application discusses the ratio and algorithm to use to distribute points, and the values to match point colouring based on the visible light spectrum.

Key Words: Multivariate data visualization; Point-based visualization; Circle Segments data visualization

1. GİRİŞ

Günümüzde birçok deneyde ve araştırmada toplanan veriler çok fazla değişken ve çok fazla gözlem içermektedir. Toplanan veriler nicel veya nitel olmak üzere kabaca iki grupta incelenebilmektedir. Birçok satır ve sütundan oluşan bu veri setlerini görselleştirmek verilerin algılanmasını daha kolay bir hale getirmektedir. Bu bağlamda oluşturulan veri görsellinin tasarımı ve seçilen görselleştirme tekniği algılanma açısından çok önemlidir. Veri türlerine ve veri boyutlarına göre seçilebilecek çeşitli görselleştirme teknikleri bulunmaktadır (Bilgin & Çamurcu 2008).

Bir araştırma sonucunda toplanan verilerden, doğru bilginin eksiksiz şekilde algılanabilmesi için görselde veri manipülasyonunu engellemek büyük önem taşımaktadır. Gözlem sayısı çok olan veri setlerinde kullanılan iki boyutlu veya üç boyutlu çubuk, çizgi, pasta gibi görselleştirme teknikleri, ortalama gibi basit istatistikleri temel alarak görselleştirme uygulandığından dolayı veri manipülasyonuna çok müsaittir. Bu tarz istatistikler veri setlerinde bulunan aykırı değerlerden (Çok büyük ve çok küçük veriler) etkilenmektedirler. Bu durumda alternatif metotların değerlendirilmesi gerekmektedir (Keim 1997).

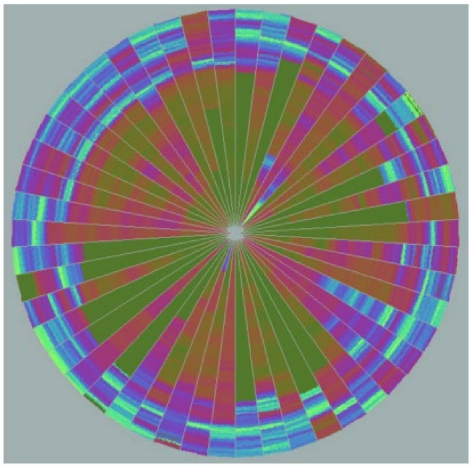
İstatistiksel birçok analizde bir veri setlerinin görselleştirmesinde, veri kaybını korumak ve bütün veri setlerini tek bir görselde görselleştirebilmek için, çeşitli doğrusal veya doğrusal olamayan metotlar uygulanmaktadır. Bunların en yaygınları Regresyon Analizi ve Temel Bileşen Analizi (PCA) olarak bilinmektedir (Dzemyda ve diğ., 2010). Fakat bu tür veri görselleştirmeleri çok hassastır ve bu durum görselin çok dikkatli bir şekilde incelenmesini gerektirmektedir (Liu ve diğ., 2006).

Literatürde "Circle Segments" olarak bilinen Dairesel Parçalı veri görselleştirme tekniği herhangi bir doğrusal ve doğrusal olamayan metot veya ortalama, medyan, yüzdelik gibi temel istatistiksel hesaplar uygulanmadan gerçekleştirilen bir veri görselleştirme tekniğidir (Ankerst ve diğ., 1996). Görselleştirme sayısal veriler görselleştirmesinde kullanılmaktadır ve temel olarak bütün veri nesnelerin hiç işleme tabi tutulmadan tek bir görselde gösterilmesini hedeflemektedir.

2. PARÇALI DAİRESEL VERİ GÖRSELLEŞTİRMESİ

Parçalı Dairesel veri görselleştirmesi piksel odaklı oluşturulan veri görselleştirme tekniklerinin arasında yer almaktadır. Piksel odaklı oluşturulan bu görselleştirme yönteminde, her bir piksel aldığı renk değerine göre veri setinde bulunan bir değeri temsil etmektedir. Böylelikle çok boyutlu ve gözlemli veri

setleri için oluşturulan görselde veri değerlerinin üst üste binmesi engellenmiş olmaktadır (Keim & Sips 2008) (Şekil.1). Bir diğer yandan bu görselleştirme yöntemi piksel odaklı çalıştığı için çoğu istatistiksel paket programlar tarafından desteklenmemektedir. Fakat bu görselleştirme tekniğini R istatistiksel programlama dilinde, koordinat düzlemi üzerinde nokta, doğru ve poligon gibi vektör tabanlı oluşturmak mümkündür. Ayrıca bu şekilde daha fazla verinin görselleştirilmesine ve görsel çıktısının istenilen çözünürlükte alınmasına imkan sağlamaktadır.



Şekil.1 50 Değişkenden 265.000 veriden oluşan veri setinin piksel odaklı Parçalı Dairesel veri görselleştirme yöntemi ile görselleştirilmesi (Keim 1996).

2.1 Parçalı Dairesel Veri Görselleştirme Yönteminin Noktasal Tabanlı Uygulanışı

Parçalı Dairesel veri görselleştirme yöntemini, R istatistiksel programlama dilinin içinde yer alan veri görselleştirme paketleriyle, vektör tabanlı oluşturmak mümkündür. Bu görselleştirmenin oluşturulmasında en önemli nokta, algoritmasının doğru bir şekilde kurulabilmesidir. Algoritma, değişkenlik gösteren değişken ve gözlem sayısına uygun hale getirildiğinde, bütün sayısal olarak toplanmış değişkenler için bu görselleştirme tekniği uygulanabilir.

Görsel, temel olarak polar koordinat düzlemi üzerinde, yarıçap ve açı parametrelerini kullanarak, veri sayısı kadar koyulan noktanın, veri değerlerine göre renklendirilmesiyle oluşturulmaktadır. Daire içinde yer alan her bir nokta, bir piksel gibi sabit hale getirilir ve temsil ettiği değişken için, daire dilimi içerisinde orantılı bir şekilde dağılım gösterir. Noktaların orantılı bir şekilde dağılabilmesi, oluşturulacak olan görsel açısından çok önemlidir. Noktaların üst üste binmesi sonucu, noktaların aldığı renk değerleri fark edilemez hale gelebilir, bu da veri görselleştirmesinin yanlış sonuç vermesine neden olmaktadır. Bu görselleştirmeye başlamadan önce veri seti üzerinden belirlenmesi gereken iki nicelik vardır. Bunlardan biri olan değişken sayısı; dairenin kaç dilime bölüneceğini ve dilimlerin açı değerlerinin ne kadar olacağını göstermektedir. Bu da dairenin bir diliminin veri setindeki bir değişkenini görselleştirdiğini göstermektedir. Diğeri ise veri setindeki gözlem sayısıdır. Gözlem sayısının belirlenmesi ile dairede yer alan her bir dilimin içinde kaç tane noktanın bulunması gerektiği tespit edilir. Bu iki değerin tespitinden sonra yapılması gereken, noktaların daire içinde orantılı bir şekilde dağıtacak fonksiyonun bulunmasıdır.

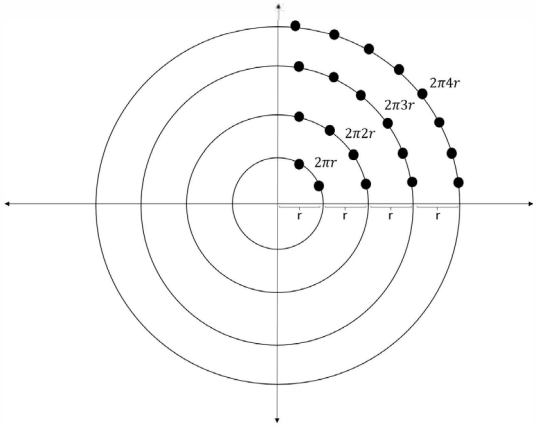
2.1.1 Dilimlerde Bulunan Noktaların Dağılımı

Bir dairenin içinde yer alan noktanın, koordinat değerlerini bulmak için gereken eşitlik, formül (2.1)'de gösterilmiştir. Bu formül temel alınarak, polar koordinat düzlemi üzerinde yer alan bir noktanın bulunması için yarıçap ve açı değerlerinin bulunması gerekmektedir.

(2.1)

$$(x, y) = \left(r.\cos\left(\frac{\alpha.\pi}{180}\right), r.\sin\left(\frac{\alpha.\pi}{180}\right) \right)$$

Bu görselleştirme yönteminde, dairenin içinde noktaların eşit aralıklarda orantılı olarak dağılabilmesi için, her yarıçapta oluşturulacak olan üzerine noktaların yerleştiği çemberin çevresinin eşit oranla artması gerekmektedir (Şekil.2).



Şekil.2 Noktaların çember çevresi formülüne göre orantılı şekilde sıralanışı.

Şekil.2 de yer alan görsel incelendiğinde, yarıçaptaki bir birimlik artışta, çemberde yer alan nokta sayısının 2 fazlası kadar bir artışın olduğu görülmektedir. Bu bilgi doğrultusunda gözlem sayısı (bir dilim içerisinde yer alacak nokta sayısı) belirli olan bir veri seti için oluşturulacak olan dairesel parçalı veri görselleştirmesinde, eşit aralıkla yarıçapı artan kaç adet çemberin bulunacağı, ardışık olarak artan çift sayıların toplam formülü ile bulunabilmektedir.

(2.2)

$$\sum_{i=1}^{n} 2k_i = Toplam\ Gözlem\ , \qquad k = (1,2,\dots,n)$$

(2.3)

$$n(n+1) = Toplam\ Gözlem\ (nokta)$$

Formül (2.2)'e göre, n verilen dizide bulunan değerlerin sayısını belirtmektedir. Bu formülün sonucu ise bu değerlerin toplam değerini göstermektedir. Noktaların yerleşim düzeni göz önünde bulundurularak bu formül yorumlandığında, bir dilimin içinde bulunan toplam nokta sayısı (bir değişkenin gözlem sayısı) dizinin toplam değerine eşit olurken, n, dizide bulunan değerlerin sayısı, yani dairenin içinde bulunan çember sayısını (dilim içinde bulunan yay sayısı) belirtmektedir. Ardışık çift sayı olarak artan bu dizide, ilk çemberden son çemberin üzerine yerleştirilecek olan nokta sayısı, bu dizi kümesinin elamanlarının sayısal değeri kadar olacaktır. Dairenin içinde bulunan çember sayısını formül (2.4)'de yer alan eşitliğin iki dereceli denklem çözümü ile tespit edilebilir.

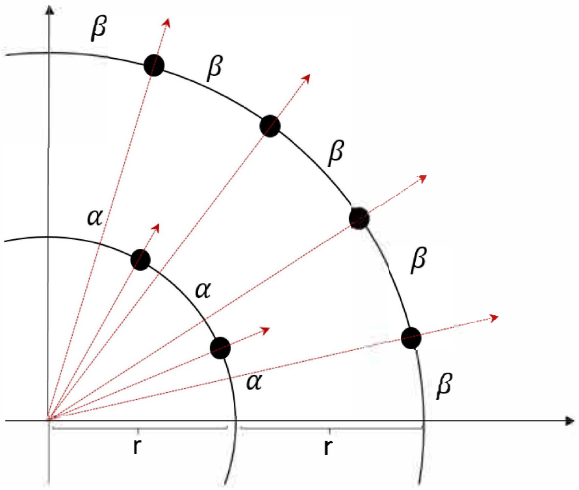
(2.4)

$$Toplam\ gözlem = TGS\ ,$$

$$n^2 + n - TGS = 0\ ,$$

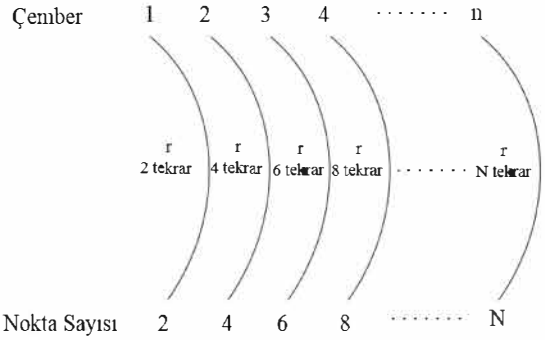
$$n_{1,2} = \pm n$$

Bu eşitlik sonunda bulunan n, çember sayısını belirtmesinden dolayı, ondalık sayı çıkması durumda bir üst tam sayısına tamamlanmalıdır. Dairenin içine veri setinin gözlem sayısına göre yerleştirilecek olan çember sayısı bulunduktan sonra, her çemberin üzerine konumlandırılacak olan noktalardan geçen doğruların, *x eksenine* yaptıkları açı değerleri bulunmalıdır. Bu açı değerleri, oluşturulacak olan görselde noktaların üst üste binmesini engellemek için, hiçbir zaman dairenin içinde bulunan dilimin başlangıç açısına ve bitiş açısına eşit olmamalıdır. Bu durumu sağlamak için, dilimin açısı çemberin içinde bulunan nokta sayısının bir fazlasına bölünmektedir. Bu uygulamanın ardında ise noktaların üzerinden geçen doğruların açı değerleri, birikimli toplam yapılarak bulunmaktadır (Şekil.3).



Şekil.3 Çemberler üzerine konumlandırılacak noktaların üzerinden geçen doğruların açı değerleri.

Bütün bu işlemler tamamlandıktan sonra, bir çemberde bulunan nokta sayısı kadar, o çemberin yarıçapı tekrarlanmalıdır. Sonuçta her bir nokta için sırasıyla dairenin içindeki bir noktayı bulmak için gereken formülün (1.1) yarıçap ve açı parametreleri elde edilmiş olur.



Şekil.4 Nokta sayılarına göre yarıçap frekansları

Elde edilen yarıçap ve açı değerleri sayesinde, bir noktayı koordinat düzleminde tanımlamak için gereken *x* ve *y koordinatları* bulunabilmektedir (1.1).

2.1.2 Parçalı Dairesel Veri Görselleştirmesinin Renklendirilmesi

Birçok veri seti, farklı birimlerde toplanmış verilere sahip olan değişkenlerden oluşmaktadır. Bu bağlamda farklı değişkenlerin sayısal olarak, minimum ve maksimum değerleri bir biriyle uyuşmayabilmektedir. Bu şekilde toplanmış verilere sahip olan değişkenleri parçalı dairesel görselleştirme yöntemi kullanarak veri değerlerinin büyüklük ve küçüklüklerine göre görselleştirdiğimizde, değişkenler arasındaki ilişkiyi renk gösterge çizelgesine göre yorumlamak yanlış sonuçlara varılmasına neden olabilmektedir. Bu durumu önlemek için ilk yapılması gereken bütün değişkenleri 0 ile 1 arasında standartlaştırmak olacaktır. Bu standartlaştırma işlemi için gereken işlem formül (4.1) de gösterilmiştir.

(3.1)

$$X_{std} = \frac{x - \min(x)}{\max(x) - \min(x)}$$

Standartlaştırma işlemi gerçekleştirildikten sonra her bir değere denk gelen RGB değeri hesaplanmalıdır. İnsan gözü yaklaşık olarak 400nm ve 700nm frekansları arasında gelen ışık değerlerini algılayabilmektedir. Bu ekranda görülen yaklaşık 300 farklı rengin algılanabilir olduğunu göstermektedir. Bu durumda 300'den dazla birbirinden farklı değerler barındıran değişkenler için RGB renk değerlerini hesaplarken RGB karılaştırma fonksiyonu kullanılmalıdır (3.1) (Liu ve diğ., 2006).

(3.2)

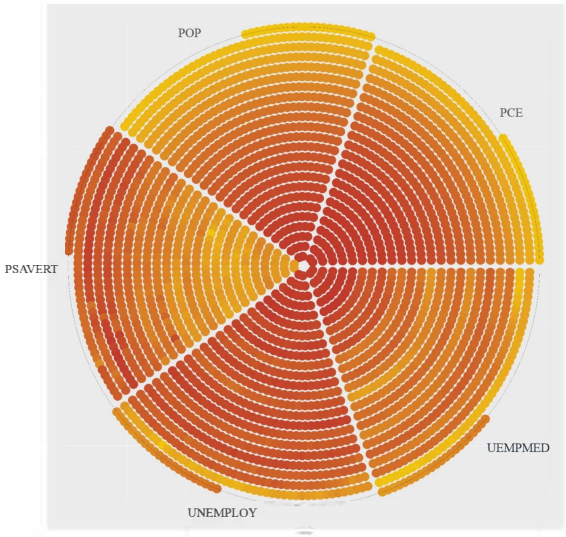
$$R = k * \sum_{\delta} \Phi(\delta) * r(\delta) * \Delta(\delta)$$

$$G = k * \sum_{\delta} \Phi(\delta) * g(\delta) * \Delta(\delta)$$

$$B = k * \sum_{\delta} \Phi(\delta) * b(\delta) * \Delta(\delta)$$

R istatistiksel programlama dilinde yer alan görselleştirme paketleriyle, renk gösterge çizelgesine göre renklendirme RGB fonksiyonu tarafından sağlanmaktadır. Bu fonksiyona göre veri setinde bir birinden farklı yaklaşık 300'den fazla değer varsa, birbirine en yakın değerler aynı RGB değerini almaktadırlar.

Bu algoritmalar izlenerek R istatistiksel programla dilinde "ggplot2" paketi kullanılarak noktasal tabanlı parçalı dairesel veri görselleştirmesi uygulanmıştır. Görselleştirme uygulanan veri seti, 1967-2015 arası Amerika Birleşik Devletleri'nde, nüfusu (POP), kişisel tüketim harcamalarını (PCE), kişisel tasarruf oranını (PSAVERT), bin kişi arasındaki işsiz sayısını (UNEMPLOY), haftalar bazında işsizlik oranını (UEMPMED) göstermektedir. Veri seti 5 değişken 574 gözlemden oluşmaktadır. Toplamda 2870 veri nesnesi Parçalı Dairesel veri görselleştirme tekniği kullanılarak noktasal tabanlı görselleştirilmiştir (Şekil.5).



Şekil.5 Noktasal Parçalı Dairesel görselleştirme

Şekil.5'de yer alan Parçalı Dairesel görselleştirmesinde, noktalar 1997'den 2015'e doğru sıralanmış (merkezden, dışarıya doğru), "sarı" renkler yüksek değerleri "kırmızı" renkler düşük değerleri göstermektedir. Zaman serisi halinde toplanmış bu veri seti kolaylıkla renkler temel alınarak ilişkisel olarak yorumlanabilmektedir.

3. SONUÇ

Bu makalede çok değişkenli ve çok fazla gözleme sahip olan veri setlerini daha anlaşılır hale getirebilmek için tercih edilebilecek yöntemler arasında bulunan ve piksel odaklı oluşturulan Parçalı dairesel veri görselleştirme yönteminin, koordinat düzlemi üzerinde görselleştirmeler uygulayan programlama dillerinde nasıl bir algoritma izlenerek oluşturulabileceği incelemiştir. Daire üzerinde noktaların dağılımının

parametrik bir şekilde sağlanabilmesi için gerek duyulan formüllere ve adımlara ayrıntılı olarak değinilmiştir. Sonuç olarak, değinilen bütün formüller ve adımlar kullanılarak, zaman serisi şeklinde toplanmış bir veri seti anlaşılır bir şekilde Parçalı Dairesel veri görselleştirme yöntemiyle görselleştirilmiştir.

Birçok veri görselleştirme tekniği, gerek doğrusal veya doğrusal olmayan iz düşüm yöntemleri, gerekse ortalama gibi aykırı değerlerden etkilenen temel istatistikler kullanılarak uygulanmaktadır. Bu durum ise oluşan görselde yanlış kararlar alınmasına sebep olabilmektedir. 2870 veri nesnesinden oluşan veri seti, noktasal tabanlı tasarıma sahip olan Parçalı Dairesel görselleştirme tekniği kullanılarak, hiçbir veri kaybı veya veri manipülasyonu olmadan, verilerin doğrudan aldığı değerler doğrultusunda görselleştirmiştir.

4.KAYNAKLAR

Ankerst, M., Keim, D. & Kriegel, H., 1996. "Circle Segments": A Technique for Visually Exploring Large Multidimensional Data Sets. *Proc. IEEE Visualization '96, Hot Topic Session*, 5–8.

Bilgin, T.T. & Çamurcu, A.Y., 2008. Çok Boyutlu Veri Görselleştirme Teknikleri. In *Akademik Bilişim, Çanakkale Onsekiz Mart Üniversitesi*. Çanakkale, 107–112.

Dzemyda, G., Olga, K. & Julius, Z., *Multidimensional Data Visualization* P. M. Pardalos & D.-Z. Du, eds., Springer Optimization and Its Applications Volume 75.

Keim, D.A., 1997. Visual Techniques for Exploring Databases Daniel Exploration. *Institute for Computer Science, University of Halle-Wittenberg*, T6-104.

Keim, D.A. & Sips, M., 2008. Circle View - A New Approach for Visualizing Time-related Multidimensional Data Sets.

Keim, D. a., 1996. Pixel-oriented database visualizations. *ACM SIGMOD Record*, 25(4), 35–39.

Liu, K., Liu, P. & Jin, D., 2006. Stimulation Spectrum Based High-dimensional Data Visualization. , 1–4.