Fırat Üniversitesi Müh. Bil. Dergisi
35(1), 81-99, 2023

Araştırma Makalesi
https://doi.org/10.35234/fumbd.1161451

Hidrolojik Sapan Değer Tespitinde Komşu İstasyon Seçimi

Ahu DEDE[1*], Necati AĞIRALİOĞLU[2]
[1]Hidrolik ve Su Kaynakları Bölümü, İnşaat Fakültesi, İstanbul Teknik Üniversitesi, İstanbul, Türkiye
[2]İnşaat Mühendisliği Bölümü, Mühendislik ve Doğa Bilimleri Fakültesi, Antalya Bilim Üniversitesi, Antalya, Türkiye
[*1]dedeah@itu.edu.tr, [2]necati.agiralioglu@antalya.edu.tr

(Geliş/Received: 13/08/2022; Kabul/Accepted: 08/11/2022)

Öz: Bu çalışmada, k-en yakın komşular yöntemine göre komşu istasyon seçmenin kriterleri, sapan değer tespiti üzerinden değerlendirilmiştir. Türkiye'de 129 istasyonda, 1926 – 2012 tarihleri arasında aylık olarak ortalama sıcaklık, ortalama bağıl nem ve toplam yağış verileri kullanılmıştır. Yakınlığa göre komşu seçildiğinde karşılaşılan ilk problem veri eksikliğinden dolayı uzak komşulara başvurulması ve komşuların %0,04 - %3'ünün 140 km'den fazla mesafede ve 0,4'ten düşük korelasyonda olmasıdır. İkinci problem incelenen dizilerin %0,1 - 3'ünde daha uzak olan komşuların en yakın 5 komşudan anlamlı olarak daha yüksek korelasyon göstermesidir. Korelasyon katsayısının yüksekliğine göre komşu seçildiğinde istasyonların %2 - %8'inde her ayın sadece kendi 1.komşusu o ay için kullanılabilir. 2. ve 3.komşular için bu oranlar düşmüştür. İstasyonların %25 - %86'sında belli aylardaki 1.komşular tüm ayları temsil edebilir ama bunun hangi aylardaki komşular olduğu belirlenmelidir. Ayrıca istasyonların %2,5'inde her değişkenin sadece kendi 1.komşusu o değişken için kullanılabilir. 2. ve 3.komşular için bu oran düşmüştür. İstasyonların %29'unda belli değişkenlerdeki 1.komşular tüm değişkenleri temsil edebilir ama bunun hangi değişkenlerdeki komşular olduğu belirlenmelidir. Bunlardan başka, tespit edilen sapan değerler çıkarılınca, sapan değerli dizilere göre korelasyon katsayısı, çoğunlukla %1-36 daha fazla, bazen %1-5 daha az çıkmıştır.

Anahtar kelimeler: K-En Yakın Komşular, Sapan Değer, Histogram, Kutu Grafiği, Korelasyon.

The selection of neighboring station in hydrolojical outlier detection

Abstract: In this study, the criteria of choosing a neighboring station by k-nearest neighbor method were evaluated through outlier detection. Monthly average temperature, average relative humidity and total precipitation data were used between 1926 and 2012 at 129 stations in Turkey. The first problem when choosing a neighbor according to the extent of nearness is that 0.04% - 3% of neighbors are more than 140 km away and less than 0.4 correlation because of choosing distant neighbors due to lack of data. The second problem is the more distant neighbors show a significantly higher correlation than the 5 closest neighbors at 0.1 - 3% of monthly series. When the neighbor is selected according to the correlation coefficient, the 1st neighbor of each month can be used for only that month in 2% - 8% of the stations. These percentages decrease for 2nd and 3rd neighbors. In 25% - 86% of stations, the 1st neighbors in certain months can be used for all months, but it should be determined which certain months are they. In addition, in 2.5% of the stations, the 1st neighbor of each variable can be used for only that variable. The percentage decreases for 2nd and 3rd neighbors. In 29% of the stations, the 1st neighbors in certain variables can be used for all the variables, but it should be determined which certain variables are they. Other than these, when the detected outliers are removed, the correlation coefficient becomes mostly 1-36% higher, sometimes 1-5% less than for the series with outliers.

Key words: K-NearestNeighbors, Outlier, Histogram, Boxplot, Correlation

1. Giriş

Bilimselliğin önemli kriterlerinden biri, rakamlaştırılmış ölçümlerin güvenilir olmasıdır. O nedenle güvenilirlik analizleri pek çok bilimsel disiplinin ayrılmaz bir parçasıdır. Hidrometerolojik ölçümlerin; eksik veri tahmini, türdeşlik, kalite kontrol gibi güvenilirlik analizleri yapılırken en önemli husus, hedef istasyonu kıyaslamak için komşu istasyonlar seçmektir. Komşu istasyonlar, hedef istasyonla aralarındaki benzerliğe göre veya rastgele olarak seçilebilir [1]. Benzerliğe göre seçildiğinde hedef istasyonla arasındaki mesafe, istasyonlarda ölçülen değerler ya da her ikisi komşu istasyonları ağırlıklandırmada kullanılır[2].

Komşu istasyonları ağırlıklandırmada mesafe çeşitli şekillerde kullanılır. Thiessen yönteminde en yakındaki istasyonların ağırlıkları bir, diğer istasyonlarınki sıfır alınır [3]. En yakın komşu yönteminde en yakındaki tek istasyonun ağırlığı bir, diğer istasyonlarınki sıfır alınır [4]. İstasyon ortalama tahminleyici yönteminde tüm

[*]Sorumlu yazar: dedeah@itu.edu.tr. Yazarların ORCID Numarası:[1]0000-0002-0534-6823, [2]0000-0002-5336-9202

istasyonlar eşit ağırlıkta kabul edilir [5]. Ters mesafe ağırlıklı enterpolasyon yönteminde hedef istasyonun diğer istasyonlar ile olan mesafesiyle ,d, ters orantılı olarak belirlenen ve $(1/d^{\alpha})/\sum(1/d^{\alpha})$ şeklinde ifade edilen ağırlıklar kullanılır [5]. Coğrafi koordinatlar yönteminde komşu istasyonların enlem ,e, ve boylam ,b, değerleriyle ters orantılı olarak belirlenen ve $(1/(e^2+b^2))/\sum(1/(e^2+b^2))$ şeklinde ifade edilen ağırlıklar kullanılır [4]. Doğal komşuluk yönteminde en yakındaki komşular temsil ettikleri alanlara göre ağırlıklandırılır [6]. Krigleme yönteminde variogramlarla veya yapay sinir ağlarıyla hesaplanan polinomlarla, hedef istasyonun diğer istasyonlar ile olan mesafesiyle artan ölçüm değerlerinin birbirinden olan farklılıkları kullanılarak oluşturulan denklem sistemlerinin çözümüyle belirlenen ağırlıklar kullanılır [7].

Mesafeye göre komşu istasyon seçmenin temeli yakın istasyonlar arası benzerliğin yüksek olduğunun düşünülmesidir [8]. Fakat mesafenin kısa olması iki istasyonun benzer olduğunu her zaman göstermez [7]. Dolayısıyla komşu istasyonları ağırlıklandırmada ölçülen değerler de kullanılır. Bunlar da mesafeler gibi çeşitli şekillerde kullanılır. Hedef istasyondaki değerler ,x, ile diğer istasyonlardaki değerler ,y, arasındaki korelasyon katsayıları ,r, ile ağırlıklar $(1/r^2)/(\sum(1/r^2)$ şeklinde belirlenebilir [9]. Ters mesafe ağırlıklı enterpolasyon yönteminde hedef istasyonun diğer istasyonlar ile olan mesafesi ,d, yerine istatistiksel bir ölçü olan $\sum(x-y)^2/2n$ katsayısı ile ters orantılı olarak belirlenen ağırlıklar kullanılır [10]. Bir diğer durumda, hedef istasyondaki ölçülen değer diğer istasyonlardaki ölçülen değerlerin bir fonksiyonu olacak şekilde yazılan denklemlerdeki katsayılar komşu istasyonların ağırlıkları olur. Ağırlıklar, denklemlerin çözüm yöntemlerine göre değişebilir. Çok değişkenli doğrusal polinomlardan oluşan denklem sistemleri sadece katsayıları belirleyen yöntemlerle çözülebilir veya buna ek olarak bu denklem sonuçlarının gözlemlenen değerlerle olan farklarını en aza indirecek şekilde çözüme giden yollar izlenebilir. Optimum ağırlıklandırma yöntemi ve negatif olmayan en küçük kareler yönteminde bu yollar izlenmiştir [2, 5]. Normal oran yönteminde hedef istasyondaki değerlerin aritmetik ortalamasının ,M_x, komşu istasyondaki değerlerin aritmetik ortalamasına ,M_y, oranı (M_x/M_y) ile belirlenen ağırlıklar kullanılır [4]. En yakın komşu yönteminde komşu istasyon yakınlığa göre seçildiği gibi hedef istasyonla en yüksek korelasyon katsayısına sahip istasyonun ağırlığı bir, diğerlerininki sıfır alınarak da seçilebilir. Bu ağırlık bazen bir yerine hedef istasyondaki değerlerin aritmetik ortalamasının ,M_x, komşu istasyondaki değerlerin aritmetik ortalamasına ,M_y, oranı (M_x/M_y) olarak da alınabilir. Bazen de tek değişkenli doğrusal regresyon denklemleri şeklinde yazılıp topluca çözülerek ağırlıklar bulunur [11].

Komşu istasyonları ağırlıklandırmada mesafe ve ölçülen değerler birlikte de kullanılabilir. Moran I, Getisord G veya Geary C istatistikleriyle hesaplanan uzaysal otokorelasyon katsayısı hem ölçülen değerleri hem de aralarındaki mesafeyi içerdiğinden buna göre belirlenen komşu istasyonlar da her ikisi de kullanılmış olur [8]. Mesafeyi içeren uzaysal otokorelasyon katsayısının tekrar mesafeyle çarpılması veya bölünmesiyle hesaplanan ağırlıklarda da gene her ikisi de kullanılmış olur [5]. Coğrafi koordinatlı normal oran yönteminde komşu istasyonların enlem, e, ve boylam, b, değerleriyle ters orantılı olarak belirlenen ağırlıkların, hedef ve komşu istasyonlardaki aritmetik ortalamaların oranıyla (Mx/My) çarpılması ile $(1/(e^2+b^2)).(Mx/My)/\sum(1/(e^2+b^2)).(Mx/My)$ şeklinde belirlenen ağırlıklar kullanılır [4]. Sistematik hatalar sapma düzeltmesiyle düzeltilir. Sapma düzeltmeleri, birliktelik kural tabanlı, dağılım tabanlı veya ortalama değer tabanlı yöntemler olarak istasyonlardaki değerleri doğrudan veya dolaylı kullanır. Dolayısıyla komşu istasyonlar için ters mesafe ağırlıklı enterpolasyon veya krigleme gibi mesafenin kullanıldığı yöntemlerde belirlenen ağırlıklar sapma düzeltmesiyle değiştirildiğinde hem mesafe hem değerler kullanılmış olur[9]. İki sonuçlu değişkenler istasyonda değerin olup olmama durumunu temsil eder (ör: yağış var/yok). Optimum mesafe tabanlı yöntemlerde olduğu gibi iki sonuçlu değişkenlerle mesafenin birlikte kullanılmasıyla hesaplanan ağırlıklarda hem mesafe hem değerler kullanılmış olur [5]. K-en yakın komşular yönteminde hedef istasyonla komşu istasyonlar arasındaki mesafe ve komşu istasyonların sayısı kullanılır. Komşu istasyonların sayısı az olursa aşırı uyum çok olursa eksik uyum olacağından değerler mesafeler kadar sonucu etkiler. Bu yöntem başlangıcında Pearsonkorelasyon katsayısı yardımıyla bileşenlerin azaltılması yoluyla da kullanılabilir [12] ya da korelasyon katsayısı ve nominal mesafeyi ölçen değer fark metriğiyle belirlenen ilişkiler arası farkların bulanık kümelerle sayısal olarak gösterilen değeri yardımıyla da kullanılabilir [13]. Her iki durumda da ağırlıklandırmaya hem değerlerin hem mesafelerin etkisi vardır. En iyi tek tahminci yönteminde mesafe ile en yakındaki istasyonlar belirlenir, korelasyon katsayısıyla onların içinden en benzer istasyon komşu istasyon olur [5]. Bulanık kural tabanlı yöntemde eğer üyelik fonksiyonları sadece değerler üzerinden belirlenmişse, ağırlıklar değerler

kullanılarak, sadece mesafeler üzerinden belirlenmişse, ağırlıklar mesafeler kullanılarak, her ikisi üzerinden belirlenmişse, ağırlıklar her ikisi kullanılarak bulunmuş olur [11].

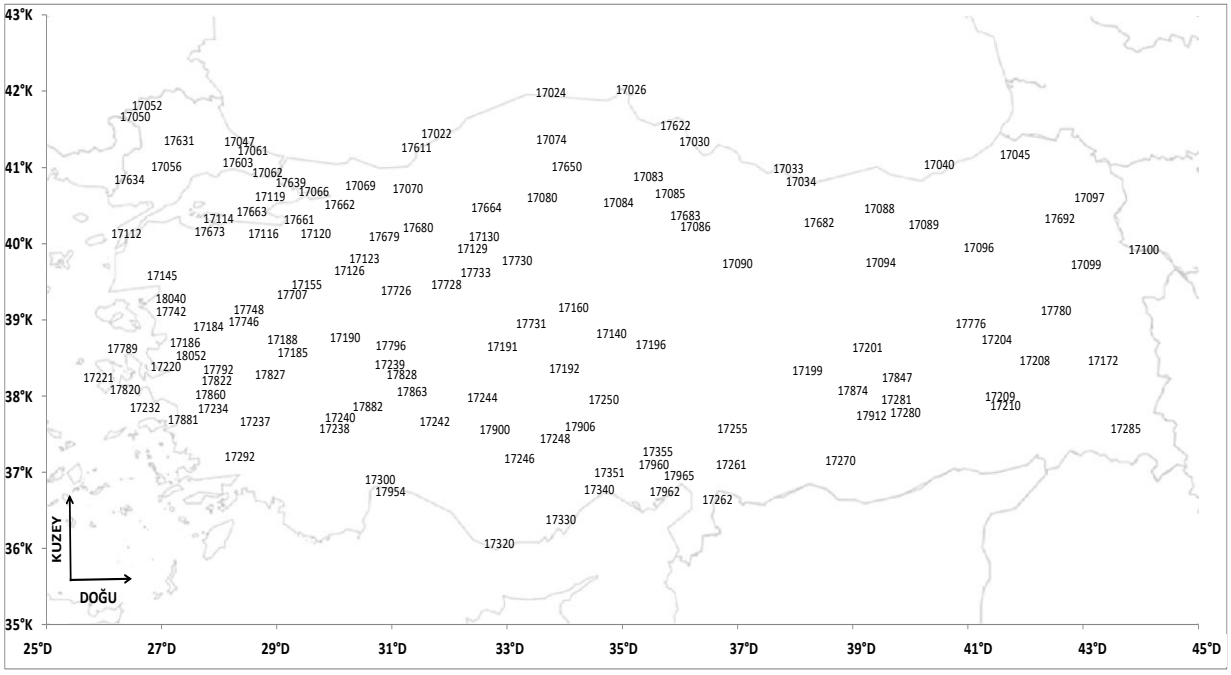
Bu çalışmada hidrolojik sapan değer tespitinde komşu istasyon seçimi için k-en yakın komşular yöntemi kullanılmış ve Türkiye'deki pek çok sıcaklık, bağıl nem ve yağış verileri kullanılarak bulunan sonuçlar birbirleriyle karşılaştırılmıştır. Literatürde en yakın k komşunun seçildiği çalışmalar [14, 15, 16, 17, 18, 19] ve en yüksek korelasyonlu k komşunun seçildiği çalışmalar [16, 20, 21, 22, 23] olmak üzere k komşu sayısını seçme üzerine bir çok çalışma vardır. Mesafe ve korelasyonun sabit alınarak değişen sayıda komşuların belirlendiği çalışmalar da mevcuttur [24]. Çalışmada, en yakın 3 komşu seçildiğinde veri eksikliğinin ve korelasyon katsayısının ortaya çıkardığı problemler ile en yüksek korelasyonlu 5 komşu seçildiğinde hidrolojik verilerdeki çeşitliliğin ortaya çıkardığı problemler değerlendirilmiştir. Bu değerlendirme istasyonlarda sapan değer tespiti üzerinden yapılmıştır. Hidrolojik sapan değerleri tespit etmek için literatürde birçok yöntem mevcuttur [15, 25, 26, 27]. Burada literatürde sıklıkla kullanılan histogram ve kutu grafiği yöntemi kullanılmıştır. Tüm işlemler excel ve excel makro ile yapılmıştır.

2. Materyal

2.1. Genel Değerlendirme

Bu çalışmadaki 1926 – 2012 arası Türkiye genelinde 129 istasyonda aylık ortalama sıcaklık (°C), 128 istasyonda aylık ortalama bağıl nem (%) ve 128 istasyonda aylık toplam yağış (mm) olmak üzere 4620 adet dizi Devlet Meteoroloji İşlerine (DMİ) bağlı Türkiye Meteorolojik Veri Arşiv ve Yönetim Sisteminden temin edilmiştir. University of East Anglia (İngiltere) kurumunun genel ağ açık veri tabanına göre verilerin eksik kısımları tamamlanmıştır. İstasyonların 82 tanesi 2007'ye kadar klima istasyonuyken, sonrasında otomatik istasyon olmuştur [28].

Klima istasyonlarında sıcaklık kuru termometreyle, bağıl nem kuru ve ıslak termometreyle ölçülür. Bunlar otomatik istasyonlarda sensörlerle ölçülür. Yağış ise klima istasyonunda plüviyometreyle, otomatik istasyonlarda elektronik plüviyometreyle ölçülür [29].

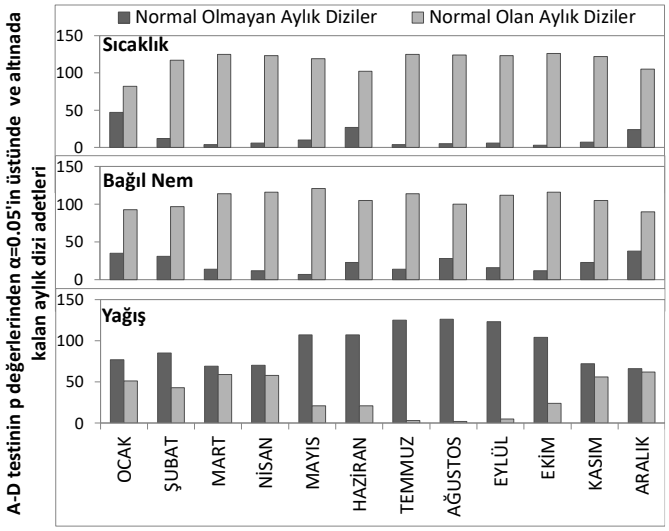


Şekil 1. Uluslararası kodlarıyla Türkiye'deki istasyonların coğrafi dağılımı.

Şekil 1'de 5 haneli istasyon kodlarının kullanıldığı haritada ortadaki hane istasyonun kuzey paralelleri ve doğu meridyenlerindeki coğrafi koordinatlarını göstermektedir. Veriler kesintisiz değildir. Eksik veri en çok 1960 öncesi zaman zarfında görülmektedir. Elimizde bulunan istasyonların Akdeniz ve Güneydoğu bölgelerindekilerde eksik veri daha fazladır. Ölçümler sıcaklıkta [+34.8°C – (-21.3°C)], bağıl nemde [6.6%–98.5%], yağışta [0mm –907.2mm] aralığındadır. Aylık dizilerin eleman sayıları genellikle 45 – 87 arasına toplanmıştır. Bir dizide sıcaklık için en az 11, bağıl nem için en az 10 ve yağış için en az 7 eleman bulunur.

Anderson-Darling testine göre sıcaklık ve bağıl nem için normallik kabulü yapılabilirken yağış için tüm aylarda özellikle yaz aylarında sağa çarpık aylık dizilerin büyük oranda 0.05 anlamlılık seviyesinin altında kaldığı görülmüştür (Şekil 2).

Bu teste normal dağılımlı çıkmayan dizilerin daha esnek değerlendirmeye tabi tutmak amacıyla çarpıklık ve basıklık değerlerinin Çizelge 1'deki formüllerle hangi sınırlarda kaldığına bakılmıştır. Sıcaklık ve bağıl nem dizileri, çarpık veya basık görülmezken, yağış dizileri çarpık veya basık çıkmaya devam etmiştir.



Şekil 2. A-D sınamasının p değerine göre α=0.05'in altında ve üstünde kalan dizi adetleri.

Bağıl nem ve yağışta çarpıklığın basıklığa göre daha çok normalden uzaklaştığı görülmüştür. Sıcaklık için genel olarak çarpık, basık veya sivri denecek istasyon sayısı yok denecek kadar azdır, bağıl nem için kış aylarında sola çarpık ve sivri istasyonlar olsa da bu sayıca azdır, yağış için kış aylarında daha az olmak üzere sağa çarpık ve sivrilik durumu gösteren istasyon sayısı oldukça fazladır.

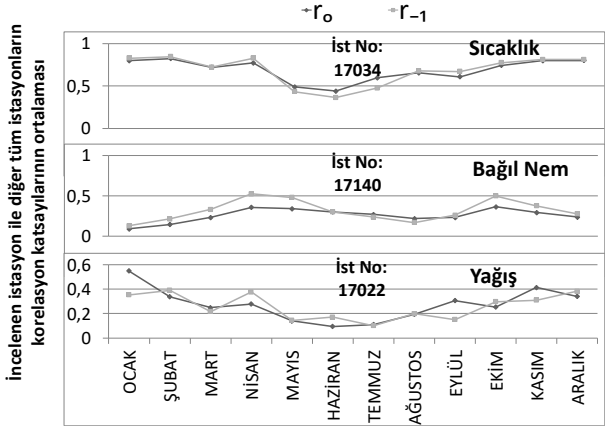
Çizelge 1. Çarpıklık ve basıklık formülleri ve standart hatası.

	Formül	Std Hata
Çarpıklık	$\frac{n}{(n-1)(n-2)}\sum\left(\frac{x_j-\bar{x}}{s}\right)^3$	$\frac{\sqrt{6}}{\sqrt{N}}$
Basıklık	$\left\{\frac{n(n+1)}{(n-1)(n-2)(n-3)}\sum\left(\frac{x_j-\bar{x}}{s}\right)^4\right\}-\frac{3(n-1)^2}{(n-2)(n-3)}$	$\frac{\sqrt{24}}{\sqrt{N}}$

n: veri adeti; s: standart sapma
xj: istasyon değeri; x̄: istasyon ortalaması

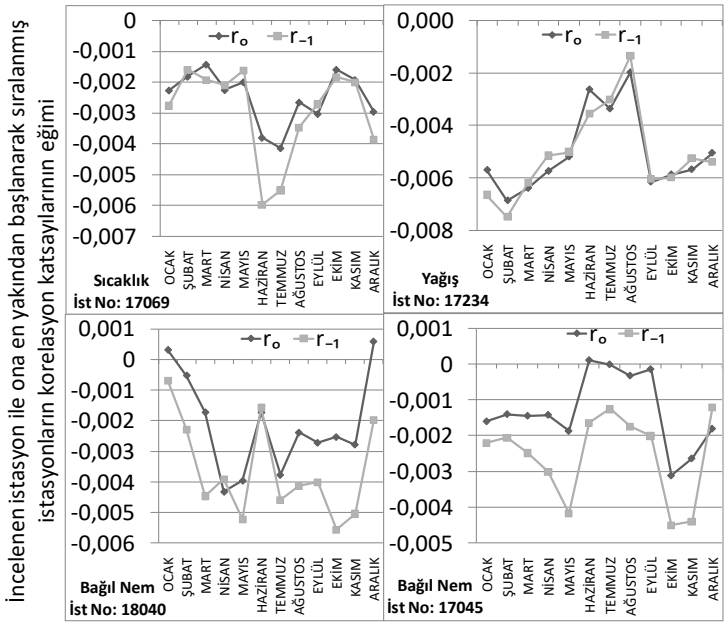
2.2 Korelasyon Katsayısı Değerlendirmesi

Sıcaklık ve yağış değişkeninde incelenen istasyon ile geriye kalan tüm istasyonların korelasyon katsayılarının ortalaması yazın kışa göre daha düşüktür. Bağıl nem değişkeninde korelasyon katsayılarının ortalaması bahar aylarında daha fazladır. Özetle miktar olarak düşük sıcaklık, yüksek yağış ve orta seviye bağıl nemde istasyonlar birbirleriyle daha çok benzerlik göstermiş, birbirleri arasındaki bağlar daha kuvvetli olmuştur. Şekil 3'deki örnek istasyonlarda gösterilen bu davranış sıcaklık için istasyonların %90'ında, bağıl nem için istasyonların %57'sinde ve yağış için istasyonların %75'inde görülmüştür.



Şekil 3.İncelenen istasyon ile geriye kalan tüm istasyonların korelasyon katsayılarının ortalaması.

Şekil 3'te r_o doğrudan diziye uygulanan korelasyon katsayısıdır. r_{-1} ise 1. yıldaki gözlem 2.yıldan, 2. yıldaki gözlem 3. yıldan çıkarılarak sadece peş peşe yıllar arasındaki değişimleri gösteren ilk fark dizilerine uygulanan korelasyon katsayısıdır. Bu dizilerdeki ortalama veya varyans değişimi, değişim anındaki değerden itibaren



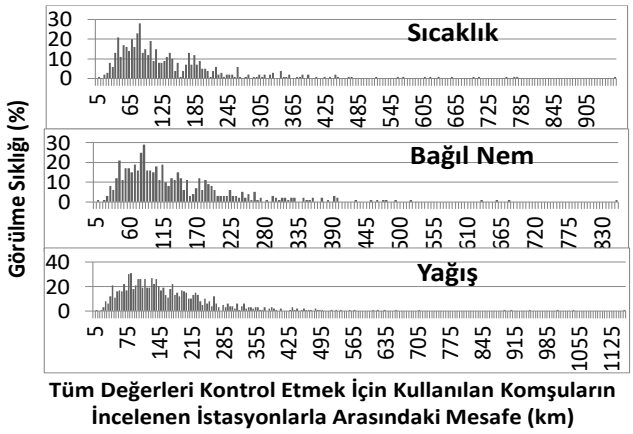
Şekil 4.İncelenen istasyon ile ona en yakından başlanarak sıralanmış geriye kalan tüm istasyonların korelasyon katsayılarının eğimi.

komşusuyla farklılaşmaya başlamaz, sadece yıllar arasındaki değişimler kıyaslanır, değişim anı ve sonrasındakifarklılaşma benzerliği etkilemez ve iki dizi arasındaki gerçek benzerlik ortaya çıkar diye düşünülmektedir [30].

Tüm değişkenler için mesafe arttıkça korelasyon katsayısı azalmıştır. Fakat bu azalış bazen daha geniş bir aralıkta bazen daha dar bir aralıkta olmuştur. İncelenen istasyon ile ona en yakından başlanarak sıralanmış geriye kalan tüm istasyonların korelasyon katsayılarının eğimi sıcaklık için istasyonların %60'ında kış aylarında yaz aylarına göre, yağış için istasyonların %45'inde yaz aylarında kış aylarına göre daha düşüktür. Bağıl nem için istasyonların %15'inde yaz aylarında kış aylarına göre, %15'inde kış aylarında yaz aylarına göre daha düşüktür. Şekil 4'te bu durumdaki istasyonlara benzer istasyonlar gösterilmiştir. Diğer istasyonlarda belirgin bir eğilim gözlenmemiştir.

2.3 İstasyon Dağılımının Değerlendirmesi

Çalışmada kullanılan istasyonların birbirlerine uzaklıkları çoğunlukla 30 – 200 km arasında olduğundan tüm değerlerin komşularıyla mesafelerinin %51'i 0 – 70 km arasında, %29'u 70 – 100 km arasında, %20'si 100 – (960 km - sıcaklık, 854 km - bağıl nem, 1154 km - yağış) arasındadır (Şekil 5).



Şekil 5. İncelenen istasyonla komşusu arasındaki mesafenin dağılımı.

3. Yöntem

3.1 Histogram ve Kutu Grafiği Yönteminin Kullanılma Biçimi

Sapan değerleri tespit etmek için sıcaklık ve bağıl nemde $\mu \pm 3\sigma$ (μ:ortalama, σ:standart sapma) sınırlarıyla histogram, yağışta $Q_{3,1} \pm 2IQR$ ve $Q_{3,1} \pm 4IQR$ ($Q_{3,1}$:üçüncü ve birinci çeyrek, IQR:$Q_3 - Q_1$ çeyrekler açıklığı) sınırlarıyla kutu grafiği yöntemi kullanılmıştır.

İncelenen toplam değer sayısı sıcaklıkta 92787, bağıl nemde 90506 ve yağışta 90618'dir. Literatüre göre komşu istasyonların sayısı 1 tane [31], 2 tane [32], 3 tane [33], 4 tane [32] veya 5 tane [34] olabilir. Uç değerlerle kıyas yapmamak için komşu sayısı ortalama bir değer olarak 3 seçilmiştir. Bu çalışmadaki komşu seçimi mekanizması şu şekilde kurulmuştur: İncelenen istasyondaki değer en yakın 3 komşusundaki karşı gelen değerle kıyaslanmış ve 3 komşuda sapan değer göstermişse bu değer sapan değer kabul edilmiş, herhangi biriyle sapan değer göstermemişse sapan olmadığı kabulü yapılmıştır. Ama herhangi bir (iki) komşusu eksik ama diğer iki (bir) komşusuyla da sapan değer göstermişse o zaman başka bir (iki) komşu ile kontrol edilmiştir. İncelemeler bazen komşuları münferit şekilde kullanarak, bazen komşulardan oluşturulan tek bir referans dizi kullanarak yapılır [24, 30]. 3 dizinin birleşiminden oluşan bir referans dizi seçilseydi, sapan değer kararı için sadece bu tek referans dizinin, incelenen değeri sapan değer göstermesi gerekecekti. Komşu olarak 3 münferit dizi seçildiğinden, üçünün de sapan değer göstermesinin gerekeceği düşünülmüştür. İncelenen istasyonun incelenen yılı için ilk üç komşunun hepsi eksik ise o zaman en yakın 4.komşuya ve sırayla devam eden şekilde diğer

komşulara başvurulmuştur. Dolayısıyla her değer en yakından başlamak üzere farklı komşu kombinasyonlarına sahip olabilir. Eğer değer, 129 istasyon içinde sadece bir/iki komşuyla çakışıyorsa bir/iki komşu ile değerlendirme yapılmıştır. Tüm değerler için kullanılan toplam komşu sayısı sıcaklıkta 233475, bağıl nemde 227831 ve yağışta 223401 olmuştur.

Komşuları seçerken yakınlık temel alındığında aynı zamanda homojen iklim bölgeleri de dikkate alınmış olmaz. Çünkü birbirine yakın iki istasyon farklı iklim bölgelerine düşebilir. İncelenen istasyonun iklim bölgesinin kıyaslanan komşuların iklim bölgeleriyle uyumluluğu için 6 ana iklim bölgesine sahip Aydeniz, De Martonne, Erinç ve Thornthwaite iklim sınıflandırmaları kullanılmıştır [35]. En yakın 1. ve 2. komşular tüm iklim sınıflandırmalarına göre çoğunlukla aynı iklim bölgesinde az bir miktarda (≈%30) %80 benzer iklim bölgesinde, 3. komşular De Martonne sınıflandırması hariç diğerlerinde eşit miktarlarda aynı iklim bölgesinde ve %80 benzer iklim bölgesindedir. İlk üç komşu dışında kullanılan komşular Aydeniz iklim sınıflandırması hariç diğerlerinde eşit miktarlarda aynı iklim bölgesinde ve %80 benzer iklim bölgesindedir.

3.2 Komşu Seçiminde Korelasyon Katsayısının Değişimini İnceleyen Yöntemler

Komşu seçiminde Denklem 1'deki Pearson korelasyon katsayısının (r) etkisini gözlemlemek için onun mesafeyle, zamanla ve sıcaklık, bağıl nem, yağış değişkenleriyle değişimleri incelenmiş, t-değerine göre kontrol edilmiştir. $\bar{x}$, $\bar{y}$: istasyon ortalamaları, x, y: istasyon değerleri, n: veri adetidir. Şekil 2'de gösterildiği gibi normal dağılımlı olan sıcaklık ve bağıl nem Denklem 1'deki Pearson korelasyon katsayısı ile incelenmiştir. Yağış önce excel'de rank.avg fonksiyonuyla sıralanmış sonra bu sıralamalar Denklem 1'deki aynı Pearson korelasyon katsayısı ile incelenmiştir. Spearman katsayısı olarak bilinen bu hesap yağış değişkeni Şekil 2'de gösterildiği gibi normal dağılımlı olmadığı için kullanılmıştır. Ölçülen değerler ve konumları ayrı ayrı değerlendirildiği için uzaysal otokorelasyon katsayısı değil, Pearson ve Spearman korelasyon katsayıları kullanılmıştır. Hem r_o hem r_{-1} için hesaplanan sonuçlar benzer çıkmıştır. Korelasyon katsayısının güvenilirliğini denetlerken, otokorelasyon fonksiyonu grafiğine göre $\pm 2/\sqrt{n}$ sınırını aşmayan istasyon sayıları aşanlardan yüksek olduğundan dizilerde iç bağımlılık olmadığı kabul edilmiş ve etkin eleman yerine veri adeti kullanılmıştır.

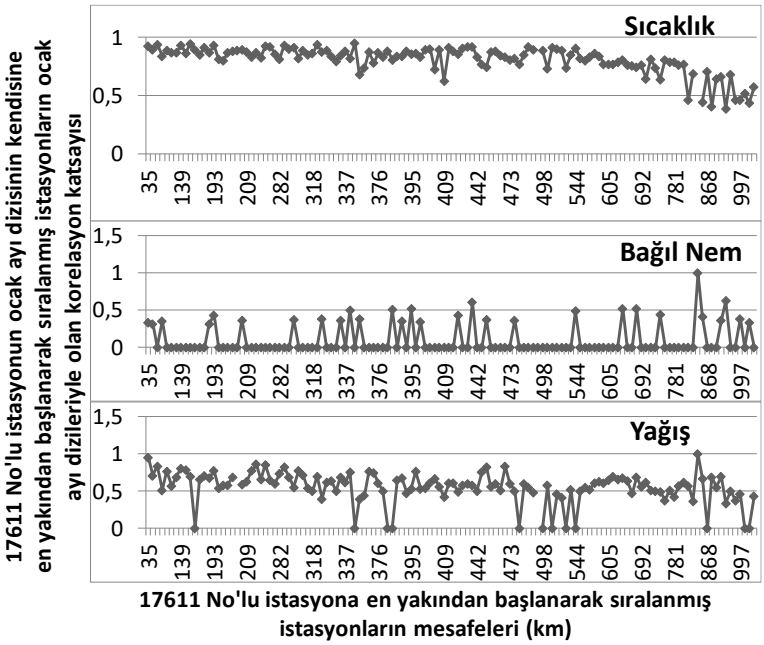
$$r = \frac{\sum (x-\bar{x})(y-\bar{y})}{\sqrt{\sum (x-\bar{x})^2 \sum (y-\bar{y})^2}} \qquad t-\text{deger}i = \frac{abs(r)\sqrt{n-2}}{\sqrt{(1-r^2)}} \tag{1}$$

3.2.1 Korelasyon Katsayısının Mesafeye Göre Değişimini İnceleyen Yöntem

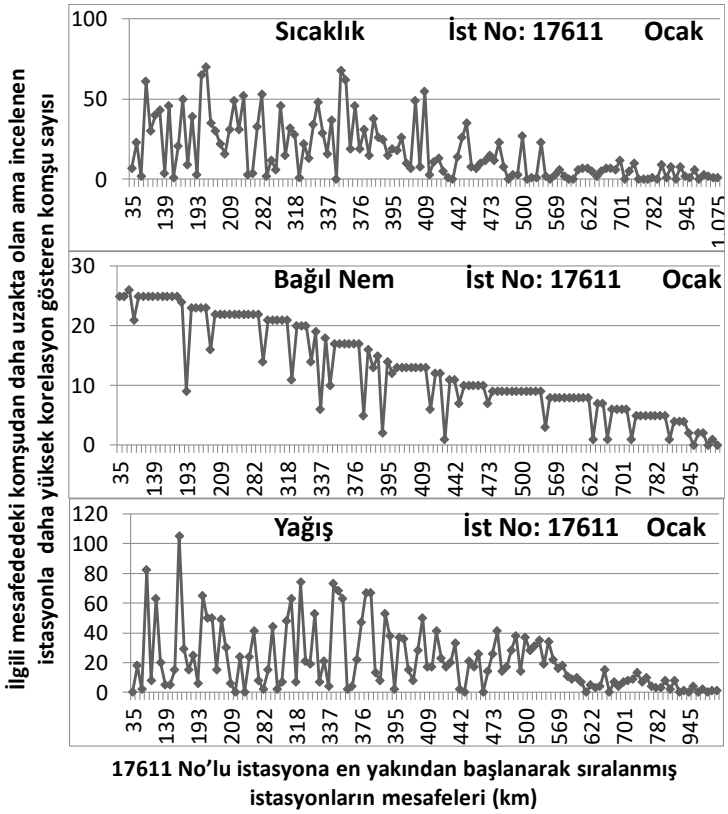
Çalışmada istasyonla komşusu arasındaki enlem boylam ve yükseklik kullanılarak hesaplanan Öklid uzaklığı kullanılmıştır [36]. Uzayda birbirine yakın verilerin, birbirinden uzak verilere göre benzer olma olasılığı daha yüksektir [8]. En yakın komşular yöntemindeki uzaklık algoritmalarından Öklid uzaklığı, uzaklık algoritmaları içinde en kısa mesafeyi hesaplandığı için kullanılmıştır [13]. Bir istasyonun komşusuyla arasındaki mesafe arttıkça aralarındaki korelasyon değişik seyirler takip etmiştir. Belli bir ortalama etrafında yaklaşık olarak sabit kalıp sonra düşmüş, sonra tekrar sabit gidip tekrar düşmüştür. Hiç düşme eğilimi göstermemiş sadece sabit bir seyir takip etmiştir. Hiç sabit kalmadan doğrudan düşme eğilimi göstermiştir. Düşüşler bazen hafif eğimle bazen dik eğimle meydana gelmiştir. Hangi durumda olursa olsun Şekil 6'deki örnek veride görüldüğü gibi mesafe arttıkça korelasyon birebir düşmediği için bir istasyona en yakın istasyon o istasyonla en yüksek korelasyonu gösterir denilemez.

Bir istasyonun komşusuyla arasındaki mesafe arttıkça o mesafedeki komşusundan daha yüksek korelasyon gösteren daha uzaktaki komşuların sayısı değişik seyirler takip etmiştir. Bazılarında önce belli bir ortalama etrafında sabit kalmış sonra yükselmiş ve sabit kalmış en sonunda doğal olarak düşmeye başlamıştır. Çünkü mesafe arttıkça istasyon sayıları da doğal olarak azalmıştır. Bazılarında önce belli bir ortalama etrafında sabit kalıp yükseldikten sonra tekrar sabit kalmadan düşmeye başlamıştır. Bazılarında önce sabit bir seyir takip edip sonra düşmüştür.

Bazılarında sadece düşüş eğilimi vardır herhangi bir kısmında sabit seyir yoktur. Bunlar hem r_o hem de r_{-1} için geçerlidir. Şekil 7'deki örnek veride görüldüğü gibi hangi durumda olursa olsun mesafe arttıkça o mesafedeki komşusundan daha yüksek korelasyon gösteren daha uzaktaki istasyonların sayısı azalsa da başlangıçta dikkate alınır bir sayı değerine sahiptir.



Şekil 6. 17611 No'lu istasyona en yakından başlanarak sıralanmış istasyonların mutlak korelasyon katsayısı.

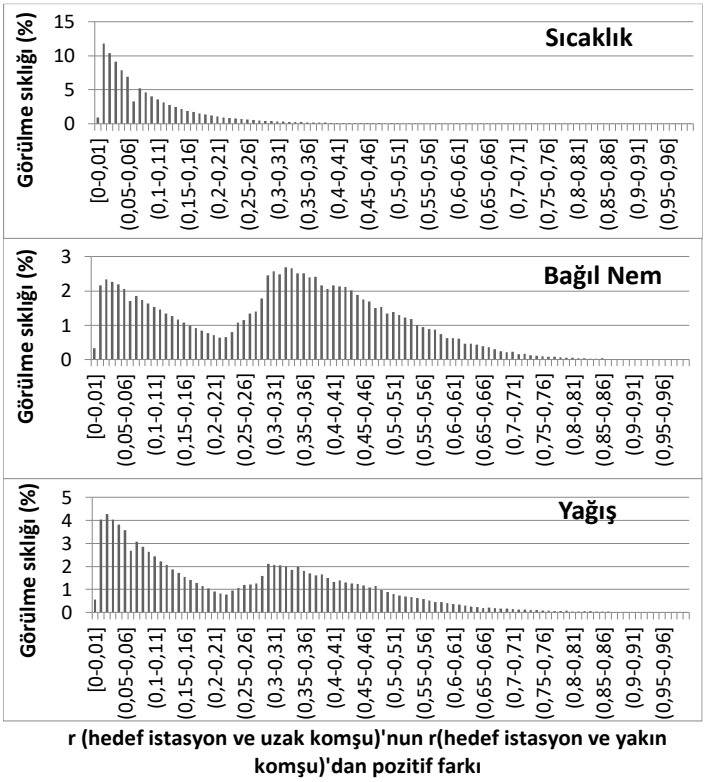


Şekil 7. 17611 No'lu istasyonun ilgili mesafesinden daha uzakta olan ama daha yüksek korelasyon gösteren istasyon sayısı.

Görüldüğü gibi incelenen istasyonun etrafındaki istasyonlarla mesafeleri arttıkça korelasyon katsayısının inişli çıkışlı bir seyir izlemesi nedeniyle zaman zaman uzak korelasyonların yakın korelasyonlardan fazla çıktığı gözlemlenmiştir. Önemli olan bu fazlalığın anlamlı olup olmadığıdır. Bu anlamlılığı belirlemek üzere bu farkların bir dağılımını oluşturmak ve bu dağılımın $Q_{3,1} \pm 2IQR$ içinde kalanlarını anlamsız kabul edip yeterince küçük olduğu için dikkate alınmayacağını, dışında kalanların ise yeterince büyük oldukları için anlamlı farklardan sayılacağını belirtmek amacıyla işlemler yapılmıştır.

Örneğin sıcaklık değişkeni için ocak ayında bir istasyonun 1.komşusundan daha uzaktaki komşuların hepsinin 1.komşusuyla olan korelasyon farkları yazılmıştır. Bu farklar hesaplanırken korelasyonların mutlak değeri alınmıştır. Bu işlem 2., 3. ve diğer tüm komşuları için ve diğer tüm aylar için yapılmıştır. Böylelikle 129 istasyonun her birinin 128 istasyonla teması vardır ve 127 adet fark en yakın 1.komşu için 126 fark en yakın 2.komşu için vs. olacak şeklinde, bir istasyon için 126x127/2 adet fark hesaplanmıştır. Bu farklar 129 istasyon ve 12 ay için bulunmuştur. Ortaya çıkan (126x127/2)x129x12 fark içinde pozitif olanlarla bir dağılım oluşturulmuştur (Şekil 8).

Bu dağılım hem r_o hem r_{-1}'e uygulanmıştır. r_o ve r_{-1} arası farklar Şekil 9'da görüldüğü gibi önce yüksek başlayıp sonra sıfıra yaklaşan ama genel olarak çok küçük olan farklar olmuştur.



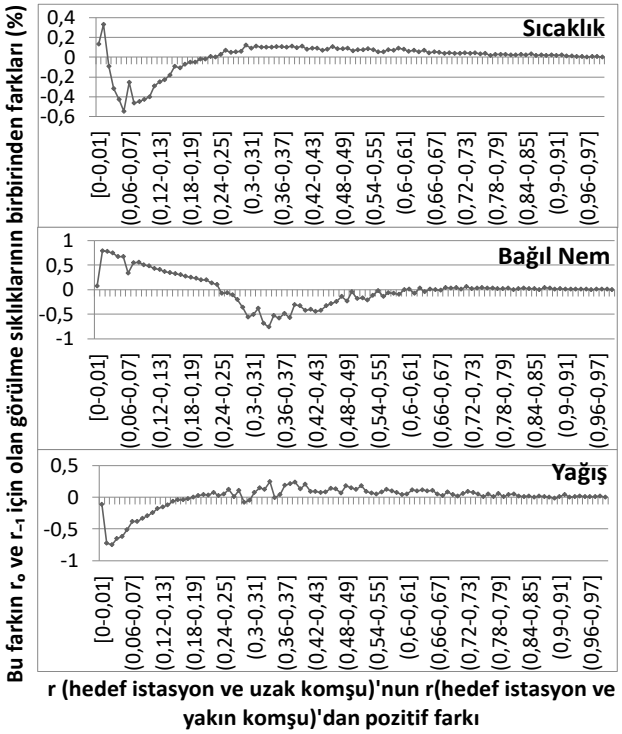
Şekil 8. 129 istasyonun 12 farklı ayda her 128 komşusu için 126x127/2 adet pozitif farkın görülme sıklığı.

Çarpık olan bu dağılımlarda $Q_{3,1} \pm 2IQR$ sınırlarını kullanarak bulunan eşiklere göre r_o için sıcaklıkta 0,33, bağıl nemde 1,04, yağışta 0,97; r_{-1} için sıcaklıkta 0,39, bağıl nemde 1,09, yağışta 1,02 fark eşiğinin dışında kalan ve dolayısıyla anlamlı kabul edilen fark sayısı istasyondan istasyona değişmektedir.

3.2.2 Korelasyon Katsayısının Zamana Göre Değişimini İnceleyen Yöntem

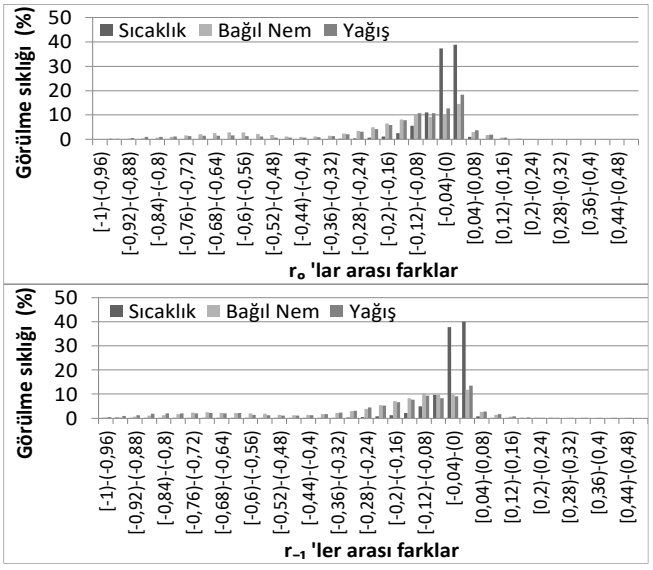
Her istasyonun her ay için benzerliğin yüksekliğine göre farklı bir komşusu olabilir. Bunun olabilirliğini göstermek için, her istasyonun her ay için benzerliğin yüksekliğine göre var olan komşusu diğer aylar için kullanıldığında oluşan farklardan bir dağılım oluşturulmuş ve bu dağılıma göre anlamlı farklar belirlenmiştir.

Örnek olarak 17611 no'lu istasyonun ocak ayındaki en yüksek korelasyonu gösteren 1.komşusu olan 17120 no'lu istasyon, şubat ayında 9. en yüksek korelasyonu gösteren komşudur. Bunun şubat ayındaki en yüksek



Şekil 9. r (hedef istasyon ve uzak komşu)'nun r(hedef istasyon ve yakın komşu)'dan pozitif farkının r_o ve r_{-1} için olan görülme sıklıklarının birbirinden farkları (%).

korelasyonu gösteren 1.komşusu olan 17022 no'lu istasyonla arasındaki korelasyon farkı -0,02'dir. Yani ocak ayındaki en yüksek korelasyonu gösteren 1.komşu, şubat ayındaki en yüksek korelasyonu gösteren 1.komşu yerine kullanılsaydı -0,02 kadar fark oluşacaktı. Mart ayındaki en yüksek korelasyonu gösteren 1.komşu yerine



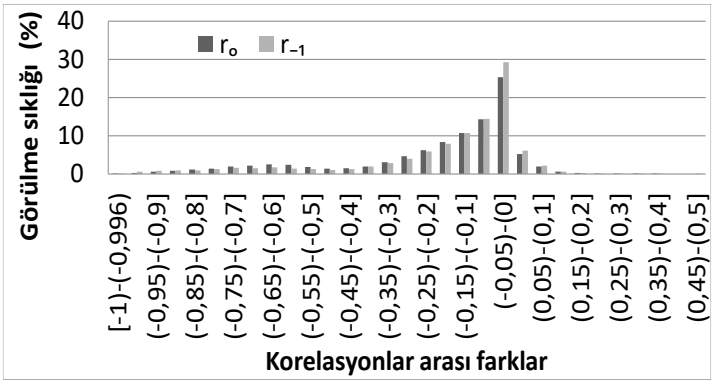
Şekil 10. Farklı aylarda korelasyon farklarının görülme sıklığı (%).

kullanılsaydı -0,06 kadar fark oluşacaktı. Yani ocak ayındaki en yüksek korelasyonu gösteren 1.komşu için 11 adet fark çıkmıştır. Şubat ayında en yüksek korelasyonu gösteren 1.komşu tüm aylar için 1.komşu olarak kullanılırsa hepsinde gene 11 adet fark çıkar. Dolayısıyla bir istasyonun 1.komşusu için 11x12 fark meydana gelir. Dolayısıyla sıcaklık değişkeninde tüm istasyonlar için 129x11x12 fark meydana gelmiştir. Bu işlem en yakın 1., 2., 3., 4. ve 5. komşu için yapıldığında 11x12x129x5 adet fark oluşmuştur. Sıcaklıkta r_o için +0,24 – (-0,996) ve r_{-1} için +0,27 – (-0,996) arası, bağıl nemde r_o için +0,5 – (-0,99) ve r_{-1} için +0,47 – (-0,996) arası, yağışta r_o için +0,4 – (-0,996) ve r_{-1} için +0,46 – (-0,996) arası farklar oluşmuş ve bunlar çarpık bir dağılım meydana getirmiştir (Şekil 10). $Q_{3,1} \pm 2IQR$ sınırlarına göre eşik değerler ($r_o$için: sıcaklık: 0,08 – (-0,12), bağıl nem: 0,54 – (-0,86), yağış: 0,44 – (-0,66), r_{-1} için: sıcaklık: 0,08 – (-0,12), bağıl nem: 0,56 – (-0,94), yağış: 0,61 – (-0,99)) belirlendikten sonra anlamsız farklar çıkarılmıştır. Bu eşiklerin dışında kalan farklar anlamlıdır.

Bu farklardan bazıları pozitiftir çünkü örneğin 17661 no'lu istasyonun ocak ayında en yüksek 2. korelasyonu gösteren 17126 no'lu komşu haziran ayında en yüksek korelasyonu gösteren 1. komşudur ve en yüksek korelasyonu gösteren 17731 no'lu 2. komşudan daha büyük olduğu için fark pozitif çıkmıştır.

3.2.3 Korelasyon Katsayısının Değişkene Göre Değişimini İnceleyen Yöntem

Her istasyonun her değişken için benzerliğin yüksekliğine göre farklı bir komşusu olabilir. Bunun olabilirliğini göstermek için, her istasyonun her değişken için benzerliğin yüksekliğine göre var olan komşusu diğer değişkenler için kullanıldığında oluşan farklardan bir dağılım oluşturulmuş ve bu dağılıma göre anlamlı farklar belirlenmiştir.



Şekil 11. Değişkenler arası korelasyon farklarının görülme sıklığı (%).

Örnek olarak ocak ayında 17611 no'lu istasyonun sıcaklık değişkenindeki en yüksek korelasyonu gösteren 1.komşusu olan 17120 no'lu istasyon, bağıl nem değişkeninde 35. en yüksek korelasyonu gösteren komşudur. Bunun bağıl nem değişkenindeki en yüksek korelasyonu gösteren 1.komşusu olan 17780 no'lu istasyonla arasındaki korelasyon farkı -0,62'dir. Yani sıcaklık değişkeninde en yüksek korelasyonu gösteren 1.komşu bağıl nem değişkenindeki en yüksek korelasyonu gösteren 1.komşu yerine kullanılsaydı -0,62 kadar fark oluşacaktı. Yağış değişkenindeki en yüksek korelasyonu gösteren 1.komşu yerine kullanılsaydı -0,25 kadar fark oluşacaktı. Yani sıcaklık değişkenindeki en yüksek korelasyonu gösteren 1.komşu için 2 adet fark çıkar. Bağıl nem değişkeninde en yüksek korelasyonu gösteren 1.komşu tüm değişkenler için 1.komşu olarak kullanılırsa gene 2 adet fark çıkar. Dolayısıyla bir istasyonun 1.komşusu için 2x3 fark meydana gelir. Bu işlem tüm aylar, tüm istasyonlar ve en yakın 1., 2., 3., 4. ve 5. komşusu için yapıldığında 12x128x5x2x3 adet fark oluşmuştur. Bu farklar r_o için +0,36 – (-0,996), r_{-1} için +0,45 – (-0,996) aralığındadır ve çarpık bir dağılım meydana getirmiştir (Şekil 11).

$Q_{3,1} \pm 2IQR$ sınırlarına göre eşik değerler (r_o için: 0,46 – (-0,74), r_{-1} için: 0,41 – (-0,64)) belirlendikten sonra anlamsız farklar çıkarılır. Bu eşiklerin dışında kalan farklar anlamlıdır. Bu farklardan bazıları pozitiftir çünkü örneğin sıcaklık için 18040 no'lu istasyonun ocak ayında en yüksek 2. korelasyonu gösteren 17742 no'lu komşu yağışta en yüksek korelasyonu gösteren 1. komşudur ve en yüksek korelasyonu gösteren 17220 no'lu 2. komşudan daha büyük olduğu için fark pozitif çıkmıştır.

3.2.4 Korelasyon Katsayısının Dizinin Sapan Değer İçeriğine Göre Değişimini İnceleyen Yöntem

Korelasyonun sapan değerle değişimini göstermek için yukarıda anlatılan yöntemle sapan değer içerdiği tespit edilen istasyonların, kendi komşularıyla olan korelasyonları incelenmiştir. Sapan değer içeren dizinin komşusuyla korelasyonuna göre, sapan değer içermeyen dizinin komşusuyla korelasyonunun % kaçının fazla olduğunu tespit etmek için Denklem 2 kullanılmıştır.

$$r_{fark} = \frac{100 \times (r_{sapan\_değersiz\_dizi} - r_{sapan\_değerli\_dizi})}{r_{sapan\_değersiz\_dizi}} \quad (2)$$

4. Bulgular ve Tartışma

Bulgular iki aşamalı tartışılabilir. Yakınlığa göre komşu seçildiğinde veri eksikliğinden dolayı daha uzak komşulara başvurulduğu için güvenilirlik düşer ve uzaktaki komşunun korelasyonunun daha fazla olduğu durumlar olur. Benzerliğe göre komşu seçildiğinde komşular zamana, değişkene ve sapan değer içeriklerine göre farklılaşır.

4.1 Mesafeye Göre Komşu Seçiminin Değerlendirilmesi

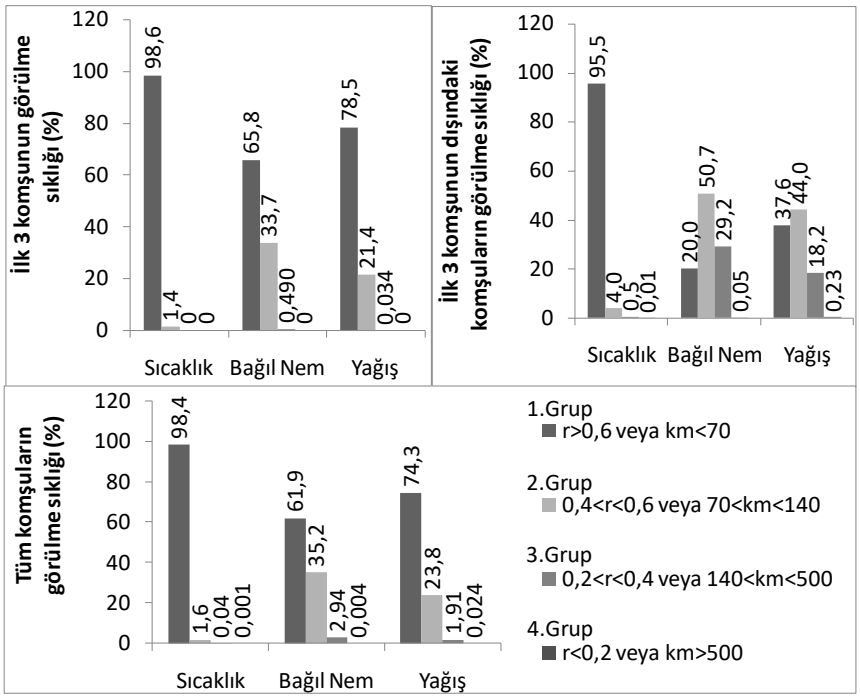
4.1.1 Uzak Komşulara Başvurulmasının ve Güvenilirliğin Düşmesinin Değerlendirilmesi

Belli bir yılın belli bir ayındaki bir dizinin her elemanının aynı yılın aynı ayındaki başka en az bir dizinin karşı gelen elemanıyla çakışması gerekir ki bir dizinin tüm elemanları kontrol edilebilsin. Sıcaklık değişkeninde en az 2 en çok 128 kez çakışmıştır. Fakat yağışta 11, bağıl nemde 16 tane değer diğer istasyonlarda aynı zamandaki başka bir değerle çakışmamış dolayısıyla kontrol edilememiştir. Kontrol edilebilen değerlerin en yakın 3 komşusunun 3'ünde de veri olmaması ve diğerlerinin sapan değer vermesi koşuluyla 1'inde ya da 2'sinde veri olmaması ve diğer komşulara başvurulma durumu sıcaklık için 72, bağıl nem için 85, yağış için 115 istasyonda görülmüştür. Bu, verilerin sıcaklık için %7,4'üne (6840/92787), bağıl nem için %7,2'sine (6493/90506) ve yağış için %8,6'sına (7375/85495) (temmuz, ağustos, eylül için sıfır yağış kayıtları mevcut değilken, sıfır yağış kayıtları mevcutken 7375/90618) denk gelmiştir.

En yakın üç komşu ile incelenen istasyon arasındaki mesafe %56 (sıcaklık - 118569/213000), %55 (bağıl nem - 114464/208403), %54 (yağış - 103323/190879 (temmuz, ağustos, eylülde sıfır yağışlar kayıt yok olarak değiştirilmiş)) oranında 70 km'nin içindedir, ama veri eksikliğinden dolayı kullanılmak zorunda kalınan daha uzaktaki istasyonlarda bu oran %4,4'e (sıcaklık - 897/20475), %4,5'e (bağıl nem - 880/19428), %6,4'e (yağış – 1414/22097 (temmuz, ağustos, eylülde sıfır yağışlar kayıt yok olarak değiştirilmiş)) düşmüştür. Bu güvenilirlik düşüklüğü korelasyonu da hesaba katan bir gruplama metoduyla gösterilmiştir.

İncelenen istasyonla 70 km'den az bir mesafede veya 0,6'ten daha yüksek bir korelasyona sahip komşular birinci güvenilir grup, incelenen istasyonla 70 km'den fazla ama 140 km'den az bir mesafede veya 0,6'tan daha düşük ama 0,4'ten yüksek bir korelasyona sahip komşular ikinci güvenilir grup, incelenen istasyonla 140 km'den fazla ama 500 km'den az bir mesafede veya 0,4'tan daha düşük ama 0,2'ten yüksek bir korelasyona sahip komşular üçüncü güvenilir grup ve geri kalan incelenen istasyonla 500 km'den fazla bir mesafede veya 0,2'ten yüksek bir korelasyona sahip komşular dördüncü güvenilir grup olarak ayarlanmıştır. İlk grup, istasyonlar arası mesafede 70 km'nin kullanıldığı [37] ve hidrolojide 0,6'dan büyük korelasyon katsayılarının genellikle anlamlı bir bağımlılık ifade ettiği kabulünün yapıldığı [38] komşularda dikkate alınarak güvenilir kabul edilmiştir. İkinci grup ilk gruba yakın olduğu için yaklaşık olarak kabul edilmiştir. Son iki grup ise güvenilirliği düşük kabul edilmiştir.

r_{-1}, r_o'dan en yakın 3 komşunun, sıcaklığın %90'ında %0 – 10 kadar, bağıl nemin %88'inde %10 – 50 kadar, yağışın %45'inde %0 – 50 kadar daha fazla çıkmıştır. Dolayısıyla sıcaklık ve bağıl nem değişkeninde grup belirleme işlemleri r_o ile değil r_{-1} ile yapılınca, daha yüksek güvenilirlikli gruplarda olan istasyon sayıları artmıştır. Yağış değişkeninde ise tam tersi bir durum vardır. Dolayısıyla yağış değişkeni r_{-1} ile yapılınca Şekil 12'de görülen güvenilir olmayan yani 3. ve 4. gruplara giren istasyon miktarı biraz daha artmıştır.



Şekil 12. Komşuların güvenilirliğini belirten grupların görülme sıklığı.

En yakın 3 komşuda düşük güvenilirlikli gruplar sıcaklık değişkeni için yoktur, bağıl nem değişkeni için %0,5 ve yağış değişkeni için %0,03'dir. En yakın 3 komşuda veri olmayan ve diğer yakın komşulara başvurulan durumlarda düşük güvenilirlikli gruplar sıcaklık değişkeni için %0,51, bağıl nem değişkeni için %29,25 ve yağış değişkeni için %18,43'dir. Toplamda kullanılan tüm komşularda düşük güvenilirlikli gruplar sıcaklık değişkeni için %0,041, bağıl nem değişkeni için %2,94 ve yağış değişkeni için %1,93'tür (Şekil 12). Dolayısıyla sapan değerleri tespit etmek için kullanılan komşuların yaklaşık %0,04'ü, (sıcaklık için), %3'ü, (bağıl nem için) ve %2'si (yağış için) güvenilirliği düşük komşulardır.

4.1.2 Uzaktaki Komşuların Daha Yüksek Korelasyon Göstermesinin Değerlendirilmesi

En yakın 1., 2., 3., 4. ve 5 komşudan daha yüksek korelasyon gösteren anlamlı istasyon sayısı sırasıyla sıcaklık için 8, 11, 18, 20 ve 19 (r_{-1} için: 5, 9 ,15 ,22 ve 18)'dur, bağıl nem için tüm farklar anlamsızlık sınırlarının içinde kalmıştır, yağış için 0, 2, 1, 2 ve 3 (r_{-1} için tüm farklar anlamsızlık sınırlarının içinde kalmıştır)'tür. Örneğin sıcaklık için 17120 no'lu istasyonun dışında kalan 128 adet istasyon mayıs, haziran, ağustos ve eylül aylarında toplam 18 defa 17120 no'lu istasyonun 1.komşusundan daha yüksek korelasyon göstermiştir. 17673 no'lu istasyonun dışında kalan 128 adet istasyon mart, mayıs, haziran, temmuz, ağustos ve eylül aylarında toplam 488 defa 17673 no'lu istasyonun 3.komşusundan daha yüksek korelasyon göstermiştir.

Özetle Çizelge 2'de görüldüğü gibi sıcaklık ve yağış değişkeni için ilk 5 komşunun korelasyon katsayısı farklı aylarda farklı istasyonlarda değişik defalar daha uzak istasyonların korelasyon katsayısından düşük kalmıştır. Sıcaklık değişkeninde 1548 dizide (129 istasyon x 12 ay) ilk 5 komşu için sırasıyla 11, 19, 42, 42, 43 (r_{-1} için: 6, 14, 30, 37 ve 35) defa yani %1 ila %3 arasında daha uzak komşular ilk 5 komşunun yerine geçebilecek durumdadır. Yağış değişkeninde 1536 dizide (128 istasyon x 12 ay) ilk 5 komşu için sırasıyla 0, 3, 1, 3, 4 defa yani %0,1 ila %0,3 arasında daha uzak komşular ilk 5 komşunun yerine geçebilecek durumdadır. Özetle yakınlığı kullanarak komşu seçildiğinde ilk komşulardan anlamlı olarak daha yüksek korelasyon gösteren komşuların daha uzak komşular olduğunu görülmektedir.

Çizelge 2. En yakın 5 komşu ile olan r_o'ın daha uzak komşularla olan r_o'dan düşük olduğu durumlar.

	Mesafeye göre en yakın komşular	İncelenen İstasyon No	İncelenen İstasyon Ay												ilk 5 komşu daha uzak komşuların kaçından daha az korelasyon gösterir
			OCAK	ŞUBAT	MART	NİSAN	MAYIS	HAZİRAN	TEMMUZ	AĞUSTOS	EYLÜL	EKİM	KASIM	ARALIK	
SICAKLIK	1.KOMŞU	17827									X				4
		17881					X								2
		17040							X						1
		17119									X				14
		17120					X	X		X	X				18
		17188									X				123
		17682					X								1
		17882									X				3
	2.KOMŞU	17827									X				4
		17040					X								1
		17045						X	X	X	X				48
		17066					X								10
		17237									X				118
		17238									X				1
		17240									X				16
		17300							X	X					2
		17320						X	X						8
		17355						X							1
		17954						X	X	X	X				67
	3.KOMŞU	17663					X	X	X	X	X				252
		17661						X							1
		17673			X		X	X	X	X	X				488
		17827									X				4
		17863									X				1
		17881	X				X								8
		17022					X								6
		17034					X								4
		17045							X						9
		17061					X				X				115
		17069					X	X		X					14
		17119					X	X		X					83
		17300						X	X						3
		17330						X	X	X					8
		17692						X	X						6
		17906						X							1
		17954						X	X	X	X	X	X		215
		17062									X				114

	Mesafeye göre en yakın komşular	İncelenen İstasyon No	İncelenen İstasyon Ay												ilk 5 komşu daha uzak komşuların kaçından daha az korelasyon gösterir
			OCAK	ŞUBAT	MART	NİSAN	MAYIS	HAZİRAN	TEMMUZ	AĞUSTOS	EYLÜL	EKİM	KASIM	ARALIK	
SICAKLIK	4.KOMŞU	17611									X				7
		17663									X				53
		17661					X	X	X		X				275
		17827									X				4
		17024					X	X							6
		17030						X							3
		17034					X	X	X						27
		17040					X	X	X	X	X				212
		17045							X						53
		17088					X								5
		17089							X						1
		17096					X	X	X	X	X				68
		17097							X						1
		17114									X				26
		17116					X	X		X	X				104
		17239									X				89
		17246						X	X	X					67
		17300						X	X	X	X				8
		17748									X				116
		17954									X				6
	5.KOMŞU	17663			X		X	X	X		X				471
		18040						X		X					201
		17827									X				4
		17863									X				2
		17022					X	X							6
		17033					X								2
		17040					X	X	X						3
		17061					X	X		X					34
		17238						X	X	X					78
		17244						X			X				13
		17300						X	X	X					56
		17330							X	X					2
		17340									X				1
		17355						X							1
		17820		X			X								19
		17860									X				111
		17882									X				6
		17954					X	X	X	X	X				309
		17062					X	X		X	X				59
YAĞIŞ	2.KOMŞU	17827												X	1
		17746		X										X	2
	3.KOMŞU	17827												X	1
	4.KOMŞU	17827										X		X	3
		17123						X							1
	5.KOMŞU	17731						X							1
		17827										X		X	3
		17126										X			1

4.2 Benzerliğe Göre Komşu Seçiminin Değerlendirilmesi

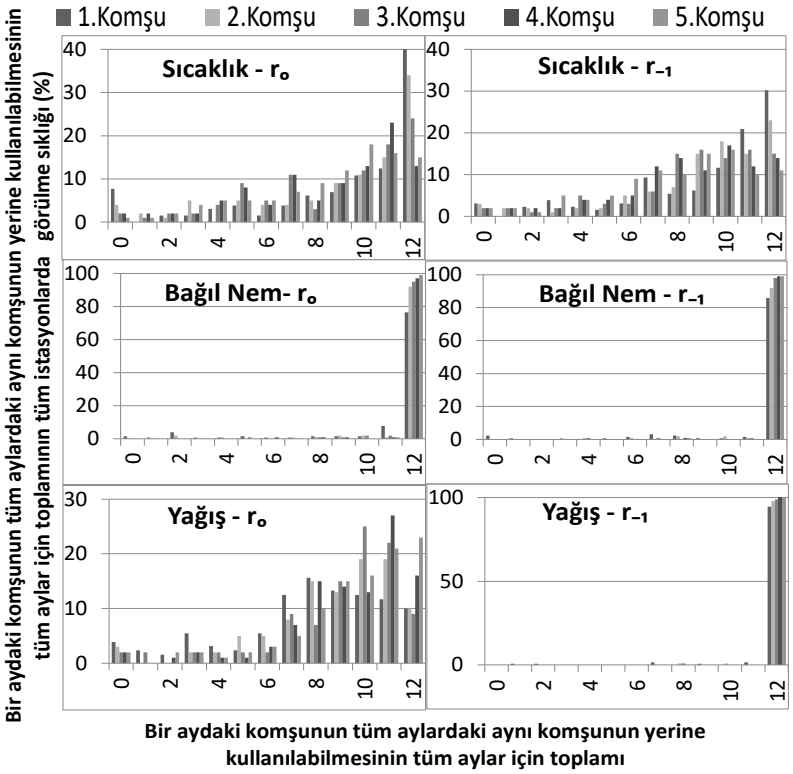
4.2.1 Benzerliğin Zamanla Birlikte Değişiminin Değerlendirilmesi

Korelasyon katsayısının yüksekliğine göre komşu seçildiğinde istasyonların sıcaklık için %40'ında, bağıl nem için %76,5'inde ve yağış için %10'unda herhangi bir aydaki 1.komşu tüm ayları temsil edebilir. 2. ve 3. komşuda bu oranlar düşmüştür.

Ama istasyonların sıcaklık için %8'inde, bağıl nem için %1,6'sında ve yağış için %4'ünde her ayın sadece kendi 1.komşusu o ay için kullanılabilir. 2. ve 3. komşuda sıcaklık için bu oran düşmüş, diğerleri için yükselmiştir. İstasyonların sıcaklık için %52'sinde, bağıl nem için %21,9'unda ve yağış için %86'sında belli

aylardaki 1.komşular tüm ayları temsil edebilir ama bunun hangi aylardaki komşular olduğu belirlenmelidir. 2. ve 3. komşuda sıcaklık için bu oran yükselmiş, bağıl nem için tamamen, yağış için büyük çoğunlukla düşmüştür. Sıcaklık ve bağıl nem değişkeninde r_o ve r_{-1} birbirine yakın çıkar ama yağış değişkeninde belirgin bir fark gözlemlenmiştir (Şekil 13).

Yağış değişkeninde r_{-1} r_o'a göre daha düşük çıkmıştır ve anlamsız korelasyonlar daha fazladır. Dolayısıyla yağış değişkeninin r_{-1}'de anlamsız korelasyonların daha fazla olması korelasyonlar arası farkların daha yüksek çıkmasına neden olmuştur. Daha yüksek farklar, farkların dağılımındaki sınırları daha yukarı çekmiş ve böylece anlamlı olan sınırların dışındaki farklar azalmıştır. Çoğu fark anlamsız kabul edileceğinden bir istasyon için herhangi bir aydaki komşu diğer aylar için kullanılabilir olmuştur.



Şekil 13. Bir aydaki komşunun tüm aylardaki aynı komşunun yerine kullanılabilmesinin tüm aylar için toplamının tüm istasyonlarda görülme sıklığı (%).

Hesaplar r_{-1}'e göre yapıldığında istasyonların sıcaklık için %30'unda, bağıl nem için %86'sında ve yağış için %94,5'inde herhangi bir aydaki 1. komşu tüm ayları temsil edebilir. 2. ve 3. komşuda bu oranlar düşmüştür. Ama istasyonların sıcaklık için %3'ünde, bağıl nem için %2,3'ünde ve yağış için %0'ında her ayın sadece kendi 1. komşusu o ay için kullanılabilir. 2. ve 3. komşuda sıcaklık için bu oran düşmüş, bağıl nem ve yağış için yükselmiştir. İstasyonların sıcaklık için %67'sinde, bağıl nem için %11,7'sinde ve yağış için %5,5'inde belli aylardaki 1.komşular tüm ayları temsil edebilir ama bunun hangi aylardaki komşular olduğu belirlenmelidir. 2. ve 3. komşuda sıcaklık için bu oran yükselmiş, bağıl nem ve yağış için düşmüştür (Şekil 13).

Şekil 3'te açıklandığı gibi tüm değişkenlerde, incelenen istasyon ile geriye kalan tüm istasyonların korelasyon katsayılarının ortalaması aydan aya değişmiştir. Şekil 13'teki aylık değişim de bu sonucu doğrulamaktadır.

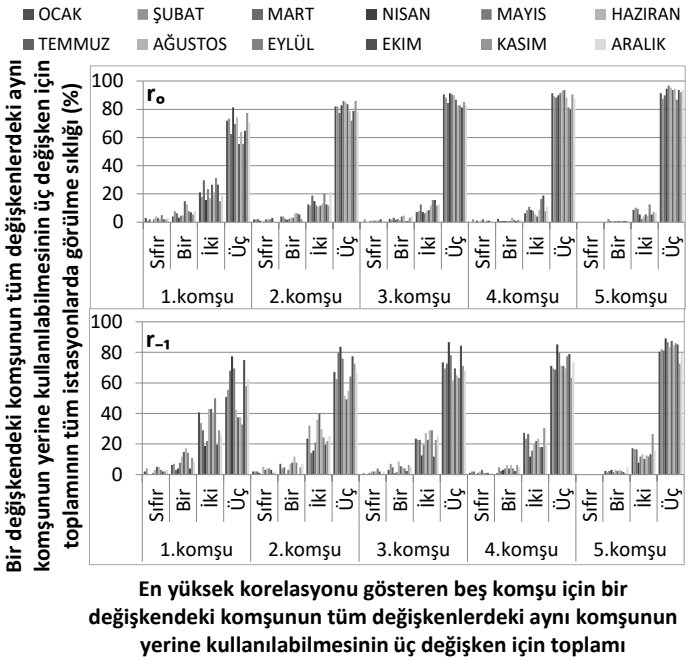
Şekil 4'te incelenen istasyon ile ona en yakından başlanarak sıralanmış istasyonların korelasyon katsayılarının eğimi kışın sıcaklıkta düşük, yağışta yüksektir. Yani uzaktaki komşulara doğru korelasyonun düşüşü kışın yağışta sıcaklığa göre fazladır. Bu durumla uyumlu olarak Şekil 13'te bir aydaki komşunun başka aylar için kullanılamaması uzaklara gidildikçe yağışta sıcaklığa göre daha fazla artmış; tam tersi olarak da birkaç aydaki komşunun tüm aylar için kullanılabilmesi uzaklara gidildikçe yağışta sıcaklığa göre daha fazla azalmıştır.

Bir aydaki komşunun tüm aylar için kullanılması zor denk gelebilecek bir durum olduğu için tüm aylarda uzak komşularda daha az görülmesi Şekil 4 ile çelişmez.

4.2.2 Benzerliğin Değişkenle Birlikte Değişiminin Değerlendirilmesi

Tüm ayların ortalaması olarak düşünüldüğünde, istasyonların 1.komşu için %68'inde, 2.komşu için %81'inde, 3.komşu için %86'inde, 4.komşu için %89'unda ve 5.komşu için %93'sinde [r_{-1}: 1.komşu için %56, 2.komşu için %67, 3.komşu için %72, 4.komşu için %73, 5.komşu için %83] herhangi bir değişkendeki komşu tüm değişkenlerde aynı komşu yerine kullanılabilir. 2. ve 3.komşuda bu oran düşmüştür.

İstasyonların 1.komşu için %2,4'ünde, 2.komşu için %1,4'ünde, 3.komşu için %0,8'inde ve 4.komşu için %0,7'sinde [r_{-1}: 1.komşu için %2,5; 2.komşu için %2,4; 3. ve 4. komşu için %1,3] tek bir değişkendeki komşu sadece kendi değişkeni için kullanılabilir. 2. ve 3. komşuda bu oranlar yükselmiştir. İstasyonların 1.komşu için %29'unda, 2.komşu için %18'inde, 3.komşu için %13'ünde, 4.komşu için %11'inde ve 5.komşu için %8'inde [r_{-1}: 1.komşu için %42, 2.komşu için %31, 3.komşu için %27, 4.komşu için %26 ve 5.komşu için %17] herhangi bir ya da iki değişkendeki komşu tüm değişkenlerde aynı komşu yerine kullanılabilir. Ama bu bir ya da iki değişkenin bu üç değişkenden hangisi olduğunu belirlemek gerekir. 2. ve 3. komşuda bu oranlar düşmüştür (Şekil 14).



Şekil 14. Üç değişkende, kendileri dışındaki değişkenlerdeki en yüksek korelasyonu gösteren ilk beş komşu yerine kullanılabilecek komşu sayısının görülme sıklığı (%).

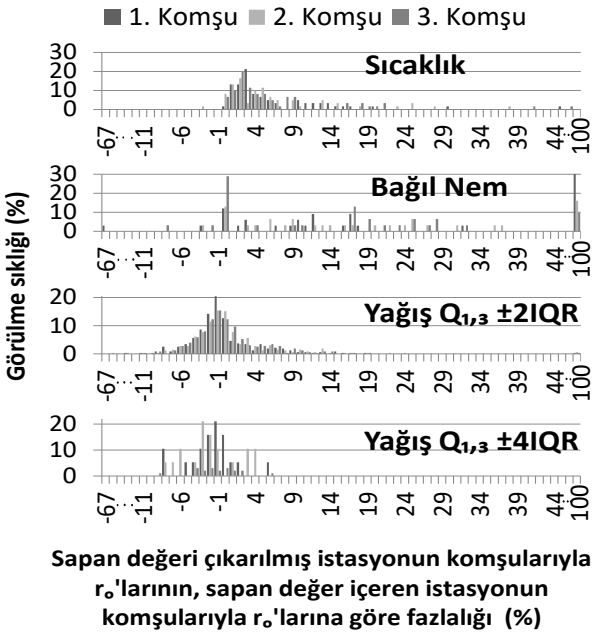
4.2.3 Benzerliğin Sapan Değerle Birlikte Değişiminin Değerlendirilmesi

$\mu \pm 3\sigma$ sınırlarının dışında kalmış sapan değerler sıcaklık değişkeni için 19, bağıl nem değişkeni için 26, yağış değişkeni için $Q_{3,1} \pm 4IQR$ sınırlarına göre 19, $Q_{3,1} \pm 2IQR$ sınırlarına göre 231 adettir.

Tespit edilen sapan değerleri çıkarılmış istasyonların komşularıyla korelasyonları, bunların çıkarılmadığı istasyonların komşularıyla korelasyonundan sıcaklıkta çoğunlukla %1 - 5 daha fazladır, az bir kısmı (%3) değişmez, daha az bir kısmı %1 daha azdır. Bağıl nemde bir kısmında (%18) değişmez, çoğunlukla %1 - 36 daha fazladır, çok az (%5,5) bir kısmında %67'ye kadar yükselen bir azalma görülmüştür. Bunun yanında sapan değer dizideyken komşularıyla korelasyonları anlamsız çıktığı için sıfır kabul edilen istasyonlar, sapan değer

çıkarılınca anlamlı çıkmaya başladıkları için, istasyonların %18'inde sapan değeri çıkarılmış istasyonun komşularıyla korelasyonları %100 daha fazla çıkmıştır.

Yağışta bir kısmında ($Q_{3,1}$ ±2IQR için %13, $Q_{3,1}$ ±4IQR için %11) değişmez $Q_{3,1}$ ±2IQR için üçte birinde $Q_{3,1}$ ±4IQR için dörtte birinde %1- 10 olan bir artış, yarısından çoğunda ($Q_{3,1}$ ±2IQR için %54, $Q_{3,1}$ ±4IQR için %67) %15'e kadar inen ama çoğunlukla %1-5 olan bir azalma vardır (Şekil 15). Az miktar bağıl nemde ve büyük çoğunlukta yağışta gözlemlenen sapan değer içeren istasyonun komşularıyla korelasyonlarının, sapan değeri çıkarılmış istasyonun komşularıyla korelasyonlarından %1 - 10 daha fazla olmasının sebebi çarpıklığa bağlanabilir. Çarpık istasyon sayısı, bağıl nemde az, yağışta çoktur. Sapan değerin çıkarılmasının olumsuz etki etmesi durumu bu durumla paralellik gösterdiğinden birbiriyle ilgili olması beklenebilir.



Şekil 15. Sapan değeri çıkarılmış istasyonun komşularıyla korelasyonlarının, sapan değer içeren istasyonun komşularıyla korelasyonlarına göre fazlalığının görülme sıklığı (%).

5. Sonuç

Bu çalışmada k-en yakın komşular yöntemi ile komşu seçiminin kriterleri bu yöntemin en bilindik zorluğu olan k'nın miktarıyla değil [12], k sabit alındığında ortaya çıkan diğer problemler üzerinden değerlendirilmiştir.

Türkiye geneline yayılmış 129 meteoroloji istasyonundaki 3 farklı değişkenin her biri için 129x12 tane aylık dizinin, 87 yıllık zaman içinde 273,911 adet değerinin her birine 1, 2 veya 3 komşu karşılık gelecek şekilde 684,707 komşu kullanılarak ortaya çıkarılan problemlerden birincisi k-en yakın komşular yönteminde yakınlığa göre komşu seçildiğinde tüm tarihlerde her istasyonda veri bulunmadığından daha uzaktaki komşulara başvurulması dolayısıyla tüm komşuların %0,04 - %3'ünün incelenen istasyon ile mesafesinin 140 km'yi aşarak ve korelasyonunun 0,4'ün altına düşerek güvenilirliğin gerilemesidir.

İkinci problem, k-en yakın komşular yönteminde yakınlığa ve benzerliğe göre iki farklı şekilde seçilebilen komşular, yakınlığa göre seçildiğinde benzerliğe göre çelişki oluşturmasıdır. İncelenen dizilerin %0,1 – 3'ünde en yakın 5 komşudan daha uzak olan komşular daha yüksek korelasyon gösterdiğinden bunların yakınlığa göre mi korelasyona göre mi seçilmesi gerektiği belirsizdir.

Komşular benzerliğe yani korelasyon katsayısının yüksekliğine göre seçildiğinde ise istasyonların %2 - %8'inde her ayın sadece kendi 1.komşusu o ay için kullanılabilir. İstasyonların %25 - %86'sında belli aylardaki 1. komşular tüm ayları temsil edebilir ama bunun hangi aylardaki komşular olduğu belirlenmelidir. Aynı şekilde istasyonların %2,5'inde her değişkenin sadece kendi 1. komşusu o değişken için kullanılabilir. İstasyonların %29'unda belli değişkenlerdeki 1. komşular tüm değişkenleri temsil edebilir ama bunun hangi değişkenlerdeki

komşular olduğu belirlenmelidir. Tüm durumlar için 2. ve 3. komşularda bu oranlar değişmektedir. Görüldüğü gibi komşular k-en yakın komşular yönteminde komşular benzerlik ölçütüne göre seçildiğinde sabit komşular bulmak yerine değişkene ve aylara göre değişen komşular belirlemek gerekir. Ayrıca istasyonlar sapan değer barındırdıkları için korelasyon katsayısı sapan değersiz dizilerde çoğunda %1- 36 daha fazla, bazısında %1-5 daha az çıkmıştır.

Sonuç olarak diğer çalışmalarda korelasyon mesafe ile düşen bir katsayı olarak görülmüş [7, 24], ve bu düşüşün birebir olmayışıyla ilgilenilmemiştir. Fakat bu çalışmada genel düşüş içinde zaman zaman yükselen korelasyonların incelenmesiyle, uzaktaki komşuların yakındaki komşulardan anlamlı olarak yüksek korelasyona sahip olabileceği gösterilmiştir. Ayrıca diğer çalışmalarda komşu istasyon seçimlerinde bir korelasyon katsayısı aylık dizilerin hepsini temsil etmiştir [24]. Fakat bu çalışmada, korelasyon katsayısına bağlı komşu istasyon seçiminde farklı ayların ve farklı değişkenlerin anlamlı olarak farklı komşulara sahip olabileceği gösterilmiştir.

Teşekkür

Bu çalışma, birinci yazarın ikinci yazar danışmanlığında hazırladığı doktora tezi esas alınarak üretilmiştir. Tez çalışmasına katkıda bulunan jüri üyeleri Prof. Dr. Ercan Kahya'ya ve Prof. Dr. C. Melek KazezyılmazAlhan'a teşekkür ederiz.

Kaynaklar

[1] Çubukçu, A., Demir, V., & Sevimli, M. F. (2019). Türkiye'nin uzun vadeli aylık sıcaklıklarının yapay sinir ağlarıyla tahmin edilmesi. 10. Ulusal Hidroloji Kongresi (p. 871). Muğla: Muğla Sıtkı Koçman Üniversitesi.

[2] Teegavarapu, R. S. (2013). Statistical corrections of spatiallyinterpolatedmissingprecipitationdataestimates. HydrologicalProcess, 3789-3808.

[3] Raghunath, H. M. (2006). Hydrologyprinciples, analysis, design. 4835/24, Ansari Road, Daryaganj, New Delhi - 110002: New Age International (P) Ltd.

[4] Rafii, F.,&Kechadi, T. (2019). Collection of HistoricalWeather Data: IssueswithMissingValues. the 4th internationalconference on smartcityapplications, (pp. 1-8).

[5] Teegavarapu, R. S. (2012). Spatialinterpolationusingnonlinearmathematicalprogrammingmodelsforestimation of missingprecipitationrecords. HydrologicalScienceJournal, 383 - 406.

[6] Karagiannidis, A. F.,&Feidas, H. (2014). Comparison of sixspatialinterpolationmethodsfortheestimation of missingdailytemperatureandprecipitationdata. 12th Pan-Hellenicand International Conference on Meteorology, ClimatologyandAtmosphericPhysics.

[7] Teegavarapu, R. S. (2007). Use of universalfunctionapproximation in variance-dependentinterpolationmethod: An application in Hydrology. Journal of Hydrology, 332: 16 - 29.

[8] O'Sullivan, D.,&Unwin, D. (2010). Geographical Information Analysis. New Jersey: John Wiley&Sons, Inc.

[9] Teegavarapu, R. S. (2009). Estimation of missingprecipitationrecordsintegratingsurfaceinterpolationtechniquesandspatio-temporalassociationrules. Journal of Hydroinformatics, 11(2), 133-146.

[10] Ahrens, B. (2006). Distance in spatialinterpolation of dailyraingaugedata. Hydrol. Earth Syst. Sci., 10, 197–208.

[11] Bárdossy, A.,&Pegram, G. (2014). Infillingmissingprecipitationrecords – A comparison of a newcopula-basedmethodwithothertechniques. Journal of Hydrology, 519, 1162–1170.

[12] Rajakumari, D. (2020). Pearsoncorrelationcoefficient k-nearestneighboroutlierclassification on real-time datasets. ICTACT Journal on Soft Computing., 10, 2045-2053.

[13] Rodriguez, Y. D. (2008). A Correlation-BasedDistanceFunctionforNearestNeighborClassification. InRuiz-Shulcloper, J.,Kropatsch, W.G. (eds) Progress in PatternRecognition, Image Analysis and Applications (pp. LectureNotes in ComputerScience, vol 5197). Berlin, Heidelberg: Springer.

[14] Doğan, Y. (2009). Outlierdetectionwith K nearestneighborclustering. Master'sthesis, DEÜ Thegraduateschool of naturalandappliedsciences, İzmir, 80.

[15] Gürünlü Alma, Ö. (2009). Geneticalgorithmbasedoutlierdetectionusinginformationcriterion. Phdthesis, DEU TheGraduate School of Natural andAppliedSciences, İzmir, 152.

[16] Li, X.,&Xiang, C. (2012). Correlation-based K-nearestneighboralgorithm. IEEE International Conference on ComputerScienceandAutomationEngineering, (pp. 185-187). doi:10.1109/ICSESS.2012.6269436

[17] Mehrotra, R.,&Sharma, A. (2006). Conditionalresampling of hydrologic time seriesusingmultiplepredictorvariables: A K-nearestneighbourapproach. Advances in WaterResources, 29, 987-999. doi:10.1016/j.advwatres.2005.08.007.

[18] Meng, Q.,Cieszewski, C. J., & Madden, M. (2007). K NearestNeighborMethodforForest Inventory Using Remote Sensing Data. GIScience& Remote Sensing, 44:2, 149-165. doi:10.2747/1548-1603.44.2.149.

[19] Rajagopalan, B.,&Lall, U. (1999). A k-NearestNeighbour Simulator for Daily PrecipitationandOtherWeatherVariables. WaterResourcesResearch, 35 , 3089-3101. doi:10.1029/1999WR900028
[20] Aieb, A.,Madani, K., Scarpa, M., Bonaccorso, B., &Lefsih, K. (2019). A newapproachforprocessingclimatemissingdatabasesappliedtodailyrainfalldata in Soummamwatershed, Algeria. Heliyon, 5.
[21] Nemes, A.,Rawls, W., &Pachepsky, Y. (2006). Use of theNonparametricNearestNeighborApproachtoEstimateSoilHydraulicProperties .SoilScienceSociety of AmericaJournal, 70, 327-336. doi:10.2136/sssaj2005.0128
[22] Shabani, S.,Samadianfard, S., Sattari, M., Mosavi, A., Band, S., Kmeť, T., &Varkonyi-Koczy, A. (2020). ModelingPanEvaporation Using GaussianProcessRegression K-NearestNeighborsRandomForestandSupportVectorMachines; Comparative Analysis. Atmosphere, 11, 1-17. doi:10.3390/atmos11010066
[23] Shi, J.,&Yang, L. (2019). A ClimateClassification of Chinathrough k -Nearest-NeighborandSparseSubspaceRepresentation. Journal of Climate, 33, 243-262. doi:10.1175/JCLI-D-18-0718.1.
[24] Vicente-Serrano, S.,Beguería, S., López-Moreno, J., García-Vera, M., &Stepanek, P. (2010). A completedailyprecipitationdatabasefornortheastSpain: reconstruction, qualitycontrol, andhomogeneity. Int. J. Climatol, 30, 1146-1163. doi:https://doi.org/10.1002/joc.1850
[25] Erdoğan, G. (2012). Spectralmethodsforoutlierdetection in machinelearning. Master'sthesis, Boğaziçi UniversityGraduate Program in ComputerEngineering, İstanbul,110.
[26] Gupta, M.,Gao, J., Aggarwal, C., & Han, J. (2014). OutlierDetectionforTemporal Data: A Survey. IEEE Transactions on Knowledge and Data Engineering, 26:9, 2250-2267. doi:10.1109/TKDE.2013.184.
[27] Bartolucci, A., Singh, K., &Bae, S. (2015). Introductiontostatisticalanalysis of laboratorydata. WILEY. Retrievedfrom https://0-doi-org.divit.library.itu.edu.tr/10.1002/9781118736890
[28] TÜMAS. (2013). 01.01.2007 tarihinde "Otomatik İstasyon" olan Klima İstasyonları. Retrievedfrom Meteoroloji Genel Müdürlüğü TÜMAS: http://tumas.mgm.gov.tr
[29] Yalçın, G. D. (2005). Klimatoloji – I. Ankara: DMİ Genel Müdürlüğü Matbaası.
[30] Aguilar, E.,Auer, I., Brunet, M., Peterson, T. C., &Wieringa, J. (2003). Guidelines on climatemetadataandhomogenization. Geneva: WMO/TD No. 1186.
[31] Wang, X. L.,& Feng, Y. (2010). RHtestsV3 User Manual. ScienceandTechnologyBranch, Environment Canada, ClimateResearchDivision, AtmosphericScienceandTechnologyDirectorate. Toronto: CanadianCentreforClimateModellingand Analysis.
[32] NCDC. (1993). Report of theinternationalworkshop on qualitycontrol of monthlyclimate data. Theinternationalworkshop on qualitycontrol of monthlyclimate data. Asheville, NC (United States), 5-6 Oct 1993: NationalClimatic Data Center.
[33] Sönmez, İ. (2013). Qualitycontroltestsfor western TurkeyMesonet. Meteorological Applications, 20(3), 1469-8080. doi:10.1002/met.1286
[34] Bayazıt, M. (1999). Hidroloji. İstanbul: İTU İnşaat Fakültesi Matbaası.
[35] Bölük, E. (2016). Aydeniz, De Martonne, Erinç, Thornthwaite İklim Sınıflandırmasına Göre Türkiye İklimi. Araştırma Dairesi Başkanlığı Klimatoloji Şube Müdürlüğü. Ankara: T.C. Orman ve Su İşleri Bakanlığı Meteoroloji Genel Müdürlüğü.
[36] MGM. (2013). İstasyon Bilgileri Veritabanı. Retrievedfrom Meteoroloji Genel Müdürlüğü: http://www.mgm.gov.tr/kurumsal/istasyonlarimiz.aspx
[37] Chiu, C.-A., Lin, P.-H., & Lu, K.-C. (2009). GIS-basedtestsforgualitycontrol of dataandspatialinterpolation of climate data. MountainResearchand Development, 29(4), 339-349. doi:http://dx.doi.org/10.1659/mrd.00030
[38] Bayazıt, M. (1981). Hidrolojide İstatistiksel Yöntemler. İstanbul: İstanbul Teknik Üniversite Matbaası.
[39] Burn, D.,&Boorman, D. (1993). Estimation of hydrologicalparameters at ungauged. Journal of Hydrology, 143, 429-454.
[40] Hohmann, C.,Kirchengast, G., O, S., Rieger, W., &Foelsche, U. (2021). Small CatchmentRunoffSensitivityto Station DensityandSpatialInterpolation: HydrologicalModeling of HeavyRainfall Using a Dense RainGauge Network. Water, 13, 1381. doi:https://doi.org/10.3390/w13101381
[41] Teegavarapu, R. S.,Aly, A., Pathak , C., Ahlquist, J., Fuelbergd, H., &Hoode, J. (2018). Infillingmissingprecipitationrecordsusingvariants of spatialinterpolationanddata-drivenmethods: use of optimal weightingparametersandnearestneighbour-basedcorrections. International journal of climatology, 38, 776-793. doi:10.1002/joc.5209
[42] Lu, Y.,Qina, X., &Mandapakaa, P. (2015). A combinedweathergeneratorand K-nearest-neighbourapproachforassessingclimatechangeimpact on regionalrainfallextremes. Int. J. Climatol, 35, 4493–4508. doi:10.1002/joc.4301
[43] Bartolucci, A., Singh, K. P., &Bae, S. (2015). Introductiontostatisticalanalysis of laboratorydata. Hoboken, New Jersey: John Wiley&Sons.