KONUMSAL-ZAMANSAL DIŞARIDAKİ ELEMANLARIN DA DÜŞÜNÜLMESİYLE ORMAN YANGINLARININ MODELLENMESİ

Berk Anbaroğlu

Hacettepe Üniversitesi, Jeodezi ve Fotogrametri Mühendisliği Bölümü, banbar@hacettepe.edu.tr

ÖZET

Genel dağılıma uymayan elemanlar popülâsyonun modellenmesindeki en büyük hata kaynağıdır. Bazen ölçümlerdeki hatalardan kaynaklanan bu elemanlar bazen gerçekten de popülâsyonun bir bireyi olabilir. Önceden yapılmış birçok çalışmada bu elemanlar hata olarak kabul edilmiş ve popülâsyonun modellenmesinde düşünülmemişlerdir. Bir başka deyişle dışarıda kalan elemanlar bulunduktan sonra popülâsyondan çıkartılırlar ki; popülâsyonun büyük çoğunluğu daha doğru bir şekilde modellenebilsin. Ancak bu durum konuma ve zamana dayalı bilgi için tekrar düşünülmelidir. Şayet, konuma ve zamana dayalı bilgide dışarıda kalan elemanlar, popülâsyonun genel dağılımdan çok daha önemli bilgi verebilir. Bu çalışma bu düşünceden hareketle yapılmıştır ve araştırmada orman yangınlarına dayalı veri kullanılmıştır. Yanan alan miktarına göre konumsal-zamansal dışarıda kalan elemanlar bulunduktan sonra bu bilgi popülâsyonun destek vektör makineleri ile modellenmesinde bir özellik olarak kullanılmıştır. Oluşturulan model ile konumsal-zamansal dışarıdakilerin de modele dahil edilmesiyle yanan alan miktarının tahmin edilmesi hedeflenmiştir.

Anahtar Sözcükler: Konumsal-Zamansal dışarıdaki elemanlar, Chebyshev Eşitsizliği, Destek Vektör Makineleri

ABSTRACT

FOREST FIRE MODELLING BY USING SPATIO-TEMPORAL OUTLIERS

Outliers are the main error source in population modeling. Outliers may occur from either erroneous observations or they may be a real individual of the population. Most of the previous researches treated outliers as noise and outliers are not used in modeling the population. In other words after detecting outliers, they are removed from the population in order to model the majority of the population with higher accuracy. However this common judgment should be thought again when data has spatial and temporal dimensions. Spatio-temporal outliers may give much more important information than the majority of the distribution. This research is based on this idea and data on forest fire occurrence is used throughout the research. After detecting spatio-temporal outliers based on the burned area of the instance, this information is used as an additional attribute to model the population using support vector machines. It is aimed to predict the burned area of an instance by considering spatio-temporal outliers.

Keywords: Spatio-Temporal Outliers, Chebyshev Inequality, Support Vector Machines

1. GİRİŞ

Konuma ve zamana dayalı verinin artışı ile bu karmaşık veri üstünde veri madenciliğinin önemi de arttı. Böylece büyük miktarda bulunan konuma ve zamana dayalı veriden önemli bilgiler elde edilmesi amaçlanmaktadır. Konumsal-zamansal dışarıda kalan elemanların (bundan sonra konumsal-zamansal dışarıdaki eleman için, "KZ-D" kullanılacaktır) bulunması da bu konuda yapılan çalışmalardan biridir. Bu konu araştırmacıların ilgisini çekmeye başlamıştır; çünkü doğadaki birçok önemli olay konumda ve zamanda dışarıdaki elemandır. Örneğin büyük depremler, volkanik dağların faaliyetleri, heyelanlar ve orman yangınları gibi can ve mal kaybına neden olan doğal afetlerin hepsi de konumda ve zamanda dışarıdaki eleman gibi düşünülebilir.

Şu ana kadar yapılmış birçok çalışma da [Colliez, Fabien, Motulsky, Podgorolec] dışarıdaki elemanların bulunmasının temel amacı, bulundukları popülâsyondan çıkartılarak diğer elemanların daha iyi bir şekilde modellenmesini sağlamaktır. Bir başka deyişle dışarıda kalan elemanlar "hata" olarak düşünülmüştür. Ancak, konuma ve zamana dayalı bilgide esas olan dışarıdaki elemanın nerde ve ne zaman olacağının tahmin edilebilmesidir. Bir başka deyişle konumsal-zamansal dışarıdakiler popülâsyonun bireyleridir ve bulunmaları için özel yöntemler düşünülmelidir.

Bilindiği üzere dışarıda kalan elemanların tanımı; popülasyonun diğer elemanlardan büyük oranda farklı ve sanki o popülasyonun bir bireyi değilmiş gibi düşünülmesine neden olan elemanlar olarak verilir. Bir başka deyişle genel dağılıma uymayan elemanlardır. Konuma ve zamana dayalı veri için aynı tanımı yapmak mümkün değildir; çünkü bu tür veride genel dağılımdan öte yerel dağılım ön plana çıkmaktadır. Bunun nedeni, konuma ve zamana dayalı veride bir bireyin konumsal ve zamansal komşuluğunda (bundan sonra konumsal-zamansal komşuluk için, "KZ-K"

kullanılacaktır) bulunan diğer elemanlardan bağımsız düşünülememesidir ve bireyin özelliklerinde KZ-K'lığındaki bireylerin doğrudan etkisinin olmasındandır. Örneğin bir bölgede deprem olduğunda sırf o bölgeyi incelemek yetersiz olacaktır; çünkü o depremin muhtemelen başka bölgelerdeki fay hatlarına da etkisi olacaktır. Sonuç olarak KZ-D; KZ-K'lığına göre konum ve zamana bağlı olmayan özelliklerinde büyük oranda farklılık gösteren eleman olarak tanımlanmıştır [Barua, Wu]. Burada iki tanımın matematiksel ifadesi verilmesi gerekmektedir: KZ-K nasıl bulunur ve bir bireyin KZ-D olması için KZ-K'lığından ne kadar uzakta olması gereklidir. İlk sorunun cevabı uygulama alanına ve araştırmanın amaçlarına göre değişebilir. Örnek olarak bir ilaç tedavisinde günler veya haftalar zamansal komşuluğu oluşturacakken, bir yıldızın oluşumu gibi evrensel bir durum incelendiğinde yüz yıllar zamansal komşuluk birimi olacaktır. Aynı düşünce, konumsal komşuluk için de geçerlidir. Bu yüzden KZ-K uygulama alanına ve amaca göre şekillenmektedir. İkinci sorunun cevabı için çeşitli yöntemler önerilmiştir [Birant]. Önerilen yöntemler bildirinin kapsamı dışında olduğundan, bu yöntemler üzerinde durulmamıştır.

Bu çalışmanın amacı; orman yangınlarına dayalı bir veri üstünde KZ-D'lerin bulunması ve bu bilginin doğrudan yanan alanın tahmininde bir özellik olarak kullanılarak ortalama hatanın azaltılmasıdır. Böylece KZ-D'lerin modele daha iyi bir şekilde dahil edilmesi hedeflenmiştir.

İkinci bölümde, KZ-D ve orman yangınlarının bulunması ile ilgili yapılmış araştırmalar belirtildikten sonra üçüncü bölümde incelenen veri üstünde KZ-D bulunması için kullanılan yöntem anlatılmaktadır. Dördüncü bölümde, önerilen yöntem kullanılarak tahmin edilen yanan alan miktarları ile orijinal veri kullanılarak elde edilen sonuçlar karşılaştırılmıştır. Beşinci bölümde, elde edilen kazanımlar belirtildikten sonra gelecekte yapılabilecek çalışmalar anlatılmaktadır.

2. YAPILMIŞ ARAŞTIRMALAR

2.1 KZ-D Bulunması

KZ-D bulunması yeni bir konudur. Yapılmış birçok çalışma konumsal dışarıdakileri bulmaya yönelikken [Lu, Shekhar, Zhao] zaman boyutu çoğunlukla ihmal edilmiştir. Bu konudaki sayılı birkaç çalışmanın bazıları aşağıda incelenmiştir:

[Cheng] çalışmasında KZ-D'leri bulmak için dört adımlı bir yöntem önermiştir. Tek özellik üzerinden (yükseklik) bulunan KZ-D'ler için önerilen dört adım: sınıflandırma, 3*3'lük pencereler kullanılarak sınıflandırma, ilk iki adımın farkından olası KZ-D'lerin bulunması ve ardışık seneler içinde olası KZ-D'lerin doğrulanması.

[Yuxiang] ise KZ-D'nin doğru tanımlanması ve doğru tanımın ardından çözümün geleceğini belirtmiştir. Tanımlamanın üç aşamada gerçekleştiği anlatılmıştır. "Basit eleman" tanımı ile veri yapısı tanımlanmakta, "karşılaştırma elemanı" ile basit elemanın ne ile karşılaştırılacağı ve son aşamada ise "karşılaştırma fonksiyonu" ile karşılaştırma ve basit elemanların nasıl karşılaştırılması gerektiği belirtilmektedir.

[Birant] çalışmasında DBSCAN kümeleme algoritmasına üç özellik getirerek ST-DBSCAN, konumsal-zamansal veride kümeleme yaparak, herhangi bir kümeye dahil edilemeyen elemanları KZ-D olarak saptamışlardır. Önerilen algoritma ile konumsal, zamansal ve konumsal olmayan özellikler üzerinden kümeleme yapılabilmektedir. Bunun yanında, her kümeye yoğunluk faktörü ekleyerek, farklı yoğunluklardaki kümelemelerde dahi gürültü bulunabiliyor. Son olarak, kümelerin temsilinde elemanların değerlerinin ortalaması kullanılarak yeni elemanların hangi kümeye atanacağı da bu temsil yöntemiyle sağlanmıştır. Bu yenilikler, 30° - 47.5^o kuzey enlemi ile 17.5^o - 42.5^o doğu boylamı arasındaki bölge olan Ege denizi, Karadeniz, Marmara denizi ve Doğu Akdeniz bölgelerini kapsamaktadır. Deniz yüzeyi sıcaklığı ve yüksekliği ve dalga yüksekliği üzerinden kümeleme başarı ile gerçekleştirilmiştir. Ancak, çalışmada KZ-D'ler konumsal ve zamansal dışarıdakilerin kesişimi olarak bulunmuştur; ancak bu durumun orman yangınları verisinde neden geçersiz olduğu aşağıda anlatılacaktır.

2.2 Orman Yangınlarının Bulunması

Orman yangınlarının erken tespiti ve hızlı bulunması üzerine birçok çalışma yapılmıştır. Orman yangınlarının tespiti üç araçla yapılabilir [Cortez]: yer gözlem uyduları, kızılötesi/duman tarayıcıları ve bölgesel sensörler. Yer gözlem uyduları kullanılarak orman yangınlarının tespiti çok uygulanabilecek bir yöntem değildir; çünkü uydulardan elde edilen konumsal çözünürlülük yeterli değildir ve bunun yanında uydu görüntülerin elde edilmesi masraflıdır. Bu yöntem ile imge işleme yöntemleri kullanılarak orman yangınlarının tespiti sağlanır [Dennison, Kalpoma, Li]. İkinci yöntemde Lidar tarayıcılar [Fernandes] kullanılarak olası orman yangınlarının erken tespiti hedeflenmiştir. Yapay sinir ağları

kullanılarak eğitilen komite makineleri ile daha düşük yanlış alarm ile orman yangınlarının tespiti sağlanmıştır. Orman yangınlarında kullanılabilecek son yöntem is bölgesel sensörlerdir. [Cortez] çalışmasında sensörlerden elde ettiği verileri kullanarak yanan orman alanını tahmin etmeye çalışmıştır. Beş farklı teknik kullanılarak yangın tahmini gerçekleştirilmiştir: çoklu regresyon, karar ağaçları, rastgele ormanlar, yapay sinir ağları ve destek vektör makineleri. [Coretz]'in çalışmasının en önemli eksikliği, büyük yangınların (dışarıdaki elemanların) yanan alanının kötü tahmin edilmesidir. Önerilen yöntemin bir diğer yönü de bu dezavantajın giderilmesini sağlamaktır.

3. KONUMSAL-ZAMANSAL DIŞARIDAKİ ELEMANLARIN BULUNMASI

Bu çalışmada KZ-D'ler istatistiksel bir yöntem olan Chebyshev Eşitsizliği [Amidan] ile bulunmuşlardır. Bir örneğin KZ-D olup olmadığı yanan alan miktarına göre belirlenmiştir: Eğer elemanın yanan alan miktarı KZ-K'lığına göre çok büyük veya küçük ise o eleman KZ-D olacaktır.

3.1 Konumsal-Zamansal Komşuluğun Tanımlanması

Daha önceden bahsedilen yöntemlerin eksikliği, konumsal-zamansal dışarıdaki elemanları konumsal ve zamansal dışarıdaki eleman olarak düşünmeleridir. Hâlbuki KZ-D, konumsal ve zamansal dışarıdaki elemanların kesişim kümesi olarak düşünülmemelidir. Bu durumu, aynı tanımların bu araştırma için de geçerli olacağı bir örnekle açıklayalım:

Bir elemanın konumsal komşuluğunda; elemanın bulunduğu konumdan 1 yarıçaplı daire içindeki bütün elemanlar, bulunsun. Bir başka deyişle, {x,y} konumsal değerlerine sahip bir elemanın konumsal komşuluğundaki elemanların konumsal değerlerinde {x, y+1}, {x,y-1}, {x+1,y}, {x-1,y} ve {x,y} değerlerinden biri olmalıdır. Aynı şekilde zamansal komşuluğu da bir çizgi gibi düşünülsün: {t} zamansal değerine sahip bir elemanın zamansal komşuluğunda {t-1}, {t} veya {t+1} değerlerine sahip elemanlar olmalıdır. İki tane 1'den büyük sabit tanımlayalım: c_1 ve c_2. Bir elemanın konumsal ve zamansal değerlerinin {x,y,t} olarak düşünelim. Bu elemanın konumsal komşuluğunda {x,y-1, t-c_1}, {x-1, y, t+c_1}, {x+1,y, t+c_1+c_2}, {x+1, y, t-c_1-c_2},.. değerlerine sahip elemanlar bulunacaktır. Böylece konumsal komşuluğundaki elemanlar bulurken aynı zamanda zamansal komşuluğunu ihlal eden elemanları da incelemiş olacağız. Aynı durum zamansal komşuluk için de geçerlidir. Böylece, konumsal ve zamansal dışarıdakiler ayrı ayrı bulunup bunların kesişiminin konumsal-zamansal dışarıdaki olarak adlandırılması KZ-D'lerin tanımına yanlış düşecektir. Bunun sonucunda konumsal-zamansal dışarıdaki elemanlar da yanlış olarak bulunmuş olacaktır. Özetle, konumsal-zamansal dışarıdaki elemanlar konumsal ve zamansal dışarıdaki elemanların kesişim kümesi değildir.

Bu çalışmada yukarıda belirtilen durum göz önüne alınmıştır ve bir elemanın (I_s) konumsal-zamansal komşuluğundaki elemanlar aşağıdaki formüle göre bulunmuştur. Veri setindeki bütün elemanlar, I_s ile karşılaştırılarak, I(a,b,c) elemanının I_s'nin konumsal-zamansal komşuluğunda olup olmadığı test ediliyor:

$$\text{KZ_K}\ (\ I_S\{x,y,t\}, I\{a,b,c\}) = \begin{cases} 1, \ eger \ \ (|x-a|+|y-b|) \le c_s \wedge |t-c| \le c_t \\ 0, \ aksi \ \ takdirde \end{cases} \tag{1}$$

Her elemanın konumsal-zamansal komşuluğu bulunduktan sonra Chebyshev Eşitsizliği kullanılarak elemanın konumsal-zamansal dışarıda olup olmadığı belirlenir.

Bu çalışmada c_s ve c_t 1 olarak kabul edilmiştir. Daha önceden de belirtildiği gibi bir elemanın konumsal ve zamansal komşuluğu uygulama alanına ve amaca göre değişmektedir. Bu yüzden yukarıda verilen tanımlar bu çalışmaya özeldir.

3.2 Chebyshev Eşitsizliği

Chebyshev Eşitsizliği [Amidan] istatistik temelli bir yöntemdir. Diğer istatistiksel yöntemlerden en önemli farkı ise popülâsyonu, bilinen herhangi bir dağılıma uygun hale getirmeye çalışmamasıdır. Böylece uygulandığı popülasyon için herhangi bir istatistiksel test veya varsayım yapmaya gerek kalmamaktadır. Bu avantajlarından ötürü KZ-D bulunmasında bu yöntem kullanılmıştır.

Cheyshev Eşitsizliği kullanarak dışarıdaki eleman bulmak iki aşamalı bir işlemdir. İlk aşamada olası dışarıdakiler bulunur, ikinci aşamada ise dışarıdaki elemanlar bulunur. İlk aşamada olası dışarıdakiler bulunduktan sonra popülâsyondan çıkartılırlar ki, ortalama ve standart sapma üstündeki olası etkileri kalksın. İkinci aşamada ise, yeni oluşturulan veri setinden birinci aşamada uygulanan işlemler uygulanarak gerçek dışarıdakiler bulunur.

Chebyshev Eşitsizliğine göre; $P(|X-\mu|\geq k\sigma) \leq 1/k^2$ 'dir. μ ortalama, σ standart sapma, k standart sapma sayısı ve X veridir. Bu eşitsizlik, veride ortalamadan k standart sapma dışında kalan elemanların maksimum yüzdesini verir. Örnek olarak, verinin en fazla %25'i ortalamadan 2 standart sapma kadar dışarıdadır. Popülasyonun dağılımının normal dağılıma uyduğu bilinseydi, %25'lik oran %5'e düşecekti. Aradaki fark, Chebyshev Eşitsizliğinin yukarıda belirtilen yönlerinden kaynaklanmaktadır.

Cheyshev Eşitsizliğinde bilinmeyen tek parametre k'dır. k dışarıdaki eleman görme olasılığına göre değişmektedir ve veri tek modlu ise dışarıdaki eleman yakalama olasılığı artmaktadır:

$$k(p) = \begin{cases} \dfrac{1}{\sqrt{p}}, & veri\ birden\ çok\ \text{mod}lu\ ise \\ \dfrac{2}{3\sqrt{p}}, & veritek\ \text{mod}lu\ ise \end{cases} \tag{2}$$

Chebyshev Eşitsizliğinin iki aşamasında da farklı p değerleri verilmelidir. Genelde ikinci aşamada verilen p değeri ilk aşamada verilenden daha küçüktür [Amidan]. Bu çalışmada p değeri olarak ilk aşamada 0.1, ikinci aşamada ise 0.05 kullanılmıştır. Bu değerlerin belirlenmesinde herhangi bir ön bilgi kullanılmamıştır.

Dışarıda kalan eleman limiti (bu değeri aşan elemanlar KZ-D olarak bulunacaktır) alt ve üst limitler için ($\mu \pm k\sigma$) hesaplandıktan sonra ilk aşamada bu limitleri aşan elemanlar olası dışarıdaki olarak bulunur. Bu elemanlar popülâsyondan çıkartıldıktan sonra ikinci aşamada gerçek dışarıdakiler aynı şekilde bulunur.

Her eleman için tek modlu olup olmadığının testi işlem zamanını arttıracağından, bu çalışma her elemanın tek modlu olmadığını kabul edilmiştir.

4. ORMAN YANGINLARI ÖRNEĞİ

Bu çalışmada amaç, KZ-D'lerin popülasyonun modellenmesinde kullanılmasıyla popülasyonun daha iyi modellenebilmesini sağlamak ve aynı zamanda KZ-D'lerin de modele dahil edilerek nerde ve ne zaman bir KZ-D olacağını tahmin etmektir. Bunun için aşağıda detayları verilen, orman yangınlarına dayalı bir veri kullanılmıştır. İncelenen veride ve diğer konuma ve zamana dayalı verilerde olacağı gibi, KZ-D popülasyona dahil etmemek popülasyonun modellenmesini iyileştirecektir; ancak bu yukarıda belirtilen nedenlerden dolayı uygulanabilir olmayacaktır.

4.1 Veri Seti

UCI veri deposundaki "Forest Fires" (Orman Yangınları) adlı veri seti araştırmada kullanılmıştır [Cortez]. Bu verinin 13 özelliği ve 517 durumu vardır. Özelliklerin detayları aşağıdaki tabloda belirtilmiştir.

Özellik	**Anlamı ve değer aralığı**
X	x-ekseni. (1-9)
Y	y-ekseni. (2-9)
Month	Ay.
Day	Gün.
FFMC, DMC, DC, ISI	Sıcaklık, nem, rüzgar ve yağmur değerlerine bağlı olarak hesaplanan indisler.
Temp	C° cinsinden sıcaklık. (2 – 33.3)
RH	Bağıl nem. %: 15 – 100
Wind	Rüzgar hızı (km/h): 0.40 to 9.40
Rain	Yağmur (mm/m2) : 0.0 to 6.4
Area	Yanan alan miktarı (hektar): 0.00 to 1090.84

Tablo 1: Verinin özellikleri ve anlamları

X ve Y konumsal özellikleri, Ay ve Gün ise zamansal özellikleri oluşturur. Veri, Ocak 2000 ile Aralık 2004 arasında Portekiz'in kuzey doğusunda bulunan Montesinho doğal parkında gerçekleşen yangınların bilgisini içermektedir. Yıl bilgisi veride olmadığı için, zamansal komşuluk incelendiğinde 4 yıllık bir süre doğrudan incelemeye dahil edilmiş

olacak. Araştırmada "Gün" özelliği zamansal komşuluk bulurken kullanılmamıştır; çünkü yangının hangi gün olduğuyla ilgilenmiyoruz. Yangının hangi aylarda, hangi şartlar altında ve nerde gerçekleştiği bilgisini genellenebilmesi daha önemli olacaktır.

Bu çalışmanın en önemli eksikliği veriyi hatasız kabul etmesidir. Kaydedilen ölçümlerin gerçek ölçümler olduğu varsayılmıştır.

4.1 Sonuçlar

Chebyshev Eşitsizliği kullanılarak 8 KZ-D bulunmuştur. Elemanın KZ-D olup olmadığını belirten yeni bir özellik de veri setine eklenmiştir. Eleman KZ-D ise 1, değil ise 0 olan bu ikil özelliği de içeren yeni veri bundan sonra "türetilmiş veri" olarak adlandırılacaktır.

Bu araştırmada destek vektör regresyonu (DVR) [Gunn] kullanılarak, yanan alan miktarının tahmini hem orijinal veride hem de türetilmiş veride gerçekleştirilmiştir. MATLAB'da yapılan çalışmada, destek vektör makinelerinin bulunduğu libSVM adlı [Chang] kütüphaneden faydalanılmıştır.

Destek vektör yakınlaştırmasında eğitimden önce belirlenmesi gereken 3 parametre vardır: hata fonksiyonu, çekirdek türü ve düzenlileştirme parametresi. Bu parametreler özetle aşağıda anlatılmıştır:

Hata fonksiyonu ile tahmin edilen değer ile gerçek değer arasındaki uzaklığı tanımlar. En çok kullanılan hata fonksiyonu ε-duyarsız hata fonksiyonudur.

$$L_\varepsilon(y) = \begin{cases} 0, \ eger \mid f(x) - y \mid < \varepsilon \\ \mid f(x) - y \mid - \varepsilon, \ aksi\, takdirde \end{cases} \tag{3}$$

Popülâsyon hakkında herhangi bir ön bilgi olmadan ne kadar hata miktarının (ε) göz ardı edilebileceği bilinemez. Bu yüzden bu hata fonksiyonunun; ancak ne kadar risk alınabileceğinin bilinmesi durumunda kullanılması faydalı olacaktır. Bunun için alternatif olarak [Schölkopf] ν-DVR'yi (nu-DVR) önermişlerdir. Bu hata fonksiyonu ε değerini otomatik olarak azaltmaktadır. Bu hata fonksiyonu ile yaratılan destek vektör sayısı ve hata kontrol edilebilir: ν hata için bir üst limit iken, destek vektörler için bir alt limittir. Belirtilen avantajlardan ve popülasyon hakkında bir ön bilgi olmadığından, araştırmada ν-DVR kullanılmıştır.

Çekirdek türü belirlenmesi gereken bir diğer önemli parametredir. Çekirdekler, n-boyutlu bir veriyi daha fazla boyutlu bir özellik uzayına dönüştürerek doğrusal ayrışımın yapılabilir hale gelmesini sağlarlar. En çok bilinen çekirdek türleri arasında Gausssian radyal tabanlı, polinom ve doğrusal çekirdekler vardır. Gaussian ve polinom çekirdeklerinde doğru parametrelerin seçimi kapsamlı ve karmaşık bir araştırma konusu olduğundan [Schölkopf], doğrusal çekirdek tercih edilmiştir.

Belirlenmesi gereken son parametre düzenlileştirme (veya kapasite kontrol, C) parametresidir. Bu parametrenin görevi, oluşan hata ile yaratılan modelin karmaşıklığı arasındaki dengeyi belirlemektir. C arttıkça yaratılan modelin karmaşıklığı ve hesaplama zamanı artacaktır. Bunun yanında oluşan hata azalacaktır. Ancak, C aşırı arttırıldığında bu sefer modelleme sadece eğitimde kullanılan örnekleri çok iyi modelleyip diğer elemanlar için iyi sonuçlar vermeyecektir. Bir başka deyişle ezberleme gerçekleşmiş olacaktır.

Eğitim ve test için ayrı ayrı veri kümeleri oluşturulmuştur. 517 veri elemanına sahip popülasyondan rastgele seçilen 345 veri (popülasyonun 2/3'ü) eğitimde, 172'si ise yaratılan modelin geçerliliğini test etmek için kullanılmıştır. Bu veri seti üstünde değişik ν ve C değerleri için ortalama kare hatası hesaplanmıştır. Aşağıdaki tablolarda sonuçlar gösterilmiştir:

ν	Orijinal - eğitim	Türetilmiş - eğitim	Orijinal-test	Türetilmiş-test
0.1	685.9	685.82	10920.03	10919.2
0.3	566.82	566.73	11175.72	11174.86
0.5	588.08	587.98	11266.29	11265.43
0.7	607.76	607.67	11328.92	11328.05
0.9	615.32	615.23	11348.45	11347.58

Tablo 2: C = 0.01

ν	Orijinal - eğitim	Türetilmiş - eğitim	Orijinal-test	Türetilmiş-test
0.1	685.81	685.09	10919.89	10911.54
0.3	566.72	565.81	11175.37	11166.75
0.5	587.84	586.91	11265.77	11257.08
0.7	604.51	603.56	11320.54	11311.78
0.9	611.66	610.73	11341.46	11332.75

Tablo 3: C = 0.1

ν	Orijinal - eğitim	Türetilmiş - eğitim	Orijinal-test	Türetilmiş-test
0.1	685,03	677,38	10917,86	10834,4
0.3	565,87	556,82	11172,86	11086,79
0.5	585,51	576,29	11260,62	11173,87
0.7	603,95	594,7	11320,48	11233,68
0.9	609,45	600,17	11338,95	11252,06

Tablo 4: C = 1

ν	Orijinal - eğitim	Türetilmiş - eğitim	Orijinal-test	Türetilmiş-test
0.1	682,8	606,86	**10909,57**	10097,07
0.3	563,84	481,09	11184,45	10350,26
0.5	581,76	498,7	11260,25	10421,19
0.7	594,35	508,74	11297,24	10452,96
0.9	598,46	513,39	11313,88	10463,15

Tablo 5: C = 10

ν	Orijinal – eğitim	Türetilmiş - eğitim	Orijinal-test	Türetilmiş-test
0.1	623,06	414,89	10997,68	7735,05
0.3	571,34	315,26	11142,16	**7429,31**
0.5	592,15	324,22	11246,56	7559,03
0.7	573,26	341,4	11207,47	8110,86
0.9	562,92	332,79	11134,64	8100,94

Tablo 6: C = 100

İncelenen bütün durumlarda türetilmiş verideki ortalama kare hata orijinal veridekinden daha azdır. C'nin küçük olması daha basit modelleme demektir. Daha basit modelleme yapıldığında dahi türetilmiş veri ile daha iyi sonuçlar alınmıştır. C arttıkça türetilmiş ile orijinal veri arasındaki hata farkı da artmaktadır. Bir başka deyişle, modellemenin karmaşıklığı arttıkça türetilmiş veride daha iyi bir modelleme yapılabilmiştir. Bu durumun aksine, karmaşıklık arttığı halde orijinal veride hem eğitimde hem de testte oluşan hata nerdeyse aynı kalmıştır.

Yaratılan modelin geçerliğinin yapıldığı test aşamasında ulaşılan en düşük ortalama kare hata; 7429,31'dir ve türetilmiş veriden elde edilmiştir. Aynı şartlarda orijinal verinin ortalama kare hatası 11142,16'dır. Bir başka deyişle, önerilen metot uygulandığında %33'lük bir hata azalması gözlemlenmiştir.

Bu sonuçlar göstermektedir ki, KZ-Dlerin ikil bir özellik ile veriye eklenmesi modellemenin başarısını arttırmıştır.

Gözlemlenen bu durumu doğrulamak ve seçilmiş veri setine olan bağlılığı azaltmak için rastgele seçilen veri elemanlarından oluşan 10 veri seti hazırlanmıştır. Her bir veri setinde 345 eğitim ve 172 test elemanı vardır. Eğitimler ve testler C için 100, ν içinse 0.1 ve 0.3 değerleri için yapılmıştır.

ν	Orijinal - eğitim	Türetilmiş - eğitim	Orijinal-test	Türetilmiş-test
0.1	3895,85	2525,9	4447,91	2941,64
0.3	3917,75	2424,97	4474,3	2842,90

Tablo 7: Modellerin ortalama hatası

Görüldüğü gibi incelenen 10 veri setinin ortalaması da türetilmiş verinin hem eğitimde hem de testte daha iyi sonuç verdiğini doğrulamıştır.

Regresyon analizlerinde bir diğer değerlendirme ölçütü de regresyon hata karakteristiği (RHK) [Bi] eğrisidir. RHK eğrisi hata toleransını x-ekseninde ve hata toleransı içinde doğru tahmin edilen noktaların yüzdesini de y-ekseninde gösterir. Bir başka ifadeyle, RHK eğrileri hatanın kümülatif dağılım fonksiyonunun tahmininde kullanılır. Eğrinin üstündeki alan ile ortalama kare hatanın veya ortalama mutlak sapma tahmininde kullanılır.

Belirtilen 10 model için eğrinin üstündeki alan hesaplaması, ortalama kare hata tahmini için kullanılmıştır. Aşağıdaki tabloda ortalama ortalama kare hata tahmini değerlerini, türetilmiş ve orijinal veri için bulunabilir.

ν	Orijinal - eğitim	Türetilmiş - eğitim	Orijinal-test	Türetilmiş-test
0.1	21.4	19.40	21.5	19.83
0.3	15.06	12.43	15.02	13.05

Tablo 8: Eğrinin Üstündeki Alan – Ortalama Kare Hata

Eğrinin üstündeki alan sonuçları da türetilmiş verinin daha iyi sonuçlar ürettiğini göstermektedir.

Önerilen yöntemin başarısını gösterebilecek son yöntem olarak modellerdeki hatanın standart sapması gösterilecektir. Yaratılan modellerdeki hatanın standart sapması ne kadar düşükse, modellerin o kadar tutarlı olacağı ve görülen sonuçların genel geçer olduğunu belirtecektir. Modellerdeki hataların standart sapması aşağıdaki tabloda görülebilir.

ν	Orijinal - eğitim	Türetilmiş - eğitim	Orijinal-test	Türetilmiş-test
0.1	2342.11	1589.29	4697.83	3187.81
0.3	2411.8	1589.53	4837.62	3188.29

Tablo 9: Modellerin hatalarının standart sapması

Türetilmiş verinin standart sapması, hem eğitim hem de test için orijinal veriye göre daha düşük çıkmıştır. Bu sonuçla görülmektedir ki, önerilen yöntem aynı zamanda modellemenin daha tutarlı olduğunu göstermektedir. Bir başka deyişle, eklenen özellik sadece veriye ek bilgi değil aynı zamanda bütünlük teşkil edecek bir nitelik de katmıştır.

5. KAZANINIMLAR

Bu çalışmada konumsal-zamansal dışarıdakilerin önemi üzerinde durulmuştur. Başka çalışmalarda dışarıda kalan elemanlar kolayca hata (gürültü) gibi değerlendirilip popülâsyondan çıkartılabiliyorken bu durumun konuma ve zamana dayalı veride uygulanamayacağı üzerinde durulmuştur. Konumsal-zamansal dışarıdakilerin tanımı ve Chebyshev Eşitsizliği kullanarak nasıl bulunabilecekleri anlatıldıktan sonra bu bilginin nasıl kullanılabilir hale getirileceği incelenmiştir. Destek vektör makineleri kullanılarak orman yangınlarına dayalı bir veri üstünde yapılan araştırmalar ile bir elemanın konumsal-zamansal dışarıda olup olmadığı bilgisinin veriye eklenmesiyle popülâsyonun modellenmesinde önemli iyileştirmeler gerçekleştirebileceği gösterilmiştir. Destek vektör makinelerinde artan C ve azalan ν değerinin daha iyi sonuçlar verdiği gözlemlenmiştir. Gözlenen sonuçların seçilen veri setine olan bağlılığını azaltmak için rastgele seçilen 10 veri seti için orijinal-türetilmiş veri karşılaştırması yapılmıştır. Bu incelemelerin sonucunda, türetilmiş verinin hem daha az hatalı hem de daha tutarlı olduğu görülmüştür.

Bu araştırmada konumsal-zamansal dışarıdakiler tek bir özellik üstünden (yanan alan) yapılmıştır. Bir elemanın yanan alan miktarı konumsal-zamansal komşuluğundan önemli oranda farklı ise o eleman konumsal-zamansal dışarıdaki olarak bulunmuştur. Ancak, gerçekte çok özellikli verilerde her özelliğin dışarıdaki elemanın bulunmasına katkısının olması gerekmektedir. Bu durumda her bir özellik için uzaklık tanımı yapılabilmeli ve elemanın dışarıdaki eleman olmasına ne

kadar katkıda bulunacağının belirlenmesi gerekecektir. Bunlardan ötürü konumsal-zamansal dışarıdaki elemanların bulunması konusu karmaşık bir problem olma özelliğini koruyacaktır.

Bu çalışmada her eleman ya konumsal-zamansal dışarıdaki ya da değil olarak düşünülmüştür. Ne kadar konumsal-zamansal dışarıda ne kadar değil gibi bir yaklaşım da (bulanık mantık yaklaşımlarının kullanılması ile) popülâsyonun modellemesini iyileştirebilecek bir durum olacaktır.

Gelecekte yapılabilecek çalışmalar hakkında önerilecek son başlık ise hata ve konumsal-zamansal dışarıdaki ayrımının yapılmasıdır. Yapılan çalışmada verinin hatasız olduğu kabul edilmiştir, oysa veri de ölçümlerden kaynaklanabilecek hatalar ile gerçekten dışarıdaki elemanların ayrımının yapılması gerekmektedir.

KAYNAKLAR

AMİDAN B. G.; FERRYMAN T. A.; COOLEY S. K. (2005), "Data Outlier Detection Using the Chebyshev Theorem", *IEEE Aerospace Conference*, (3814-3819)

BARUA S.; ALHAJJ R. (2007), "Parallel Wavelet Transform for Spatio-temporal Outlier Detection in Large Meteorological Data", *IDEAL 2007*, vol.4881 (684-694)

BI J.; BENNETT K.P.,(2003), "Regression Error Characteristic Curves", *ICML*

BİRANT D.; KUT A. (2007), "ST-DBSCAN: An algorithm for clustering spatial-temporal data", *Data and Knowledge Engineering*, Vol.60:1, (208-221)

CHANG C-C.; LIN C-J. (2001), "LIBSVM: A library for support vector machines", Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm

CHENG T.; LI Z. (2004), "A Multiscale Approach to Detect Spatial- Temporal Outliers", *20th ISPRS Congress*, (1008-1013)

COLLİEZ J.; DUFRONOİS F.; HAMAD D. (2006), "Robust Regression and Outlier Detection with SVR: Application to Optic Flow Estimation", *British Machine Vision Conference*

CORTEZ P.; MORAIS A. (2007), "A Data Mining Approach to Predict Forest Fires using Meteorological Data", *13th Portuguese Conference in Artificial Intelligence*, (512-523)

DENNISON P.E. (2006), "Fire Detecttion in Imaging Spectrometer Data Using Atmospheric Carbon Dioxide Absorption", *International Journal of Remote Sensing*, (3049-3055)

FABİEN L.; ANATOLE L.; FABRİCE L.; BRUNO A. (2007), "Studying the Use of Fuzzy Inference Systems for Motor Imagery Classification", *IEEE Transactions on Neural Systems and Rehabilitation Engineering*

FERNANDES A.M.; UTKIN A.B.; LAVROV A.V.; VILARA R.M. (2004), "Development of Neural Network Committee Machines for Automatic Forest Fire Detection Using Lidar", *Pattern Recognition Society*, (2039-2047)

GUNN S. (1998), "Support Vector Machines for Classification and Regression", *ISIS Technical Report*

KALPOMA K.A.; KUDOH J.I. (2006), "A New Algorithm for Forest Fire Detection Method with Statistical Analysis Using NOAA AVHRR Images", *International Journal of Remote Sensing*, (3867-3880)

LI Y.; VODACEK A.; KREMENS R. L.; ONONYE A.; TANG C. (2005), "A Hybrid Contextual Approach to Wildland Fire Detection Using Multispectral Imagery", *IEEE Transactions on Geoscience and Remote Sensing*, (2215-2126)

LU C. T.; CHEN D.; KOU Y. (2003), "Detecting Spatial Outliers with Multiple Attributes", *15th IEEE International Conference on Tools with Artificial Intelligence*

MOTULSKY H. J.; BROWN R. E. (2005), "Detecting Outliers When Fitting data with nonlinear regression- a new method based on robust nonlinear regression and the false discovery rate", *BMC Bioinformatics*, Vol.7:123

PODGOROLEC V.; HERİCKO M.; ROZMAN I. (2005), "Improving Mining on Medical Data by Outliers Prediction", *18th IEEE Symposium on Computer Based Medical Systems*, (91-97).

SHEKHAR S.; LU C. T.; ZHANG P. (2003), "A Unified Approach to Detecting Spatial Outliers", *GeoInformatica*, Vol:7:2, (139-166)

SCHÖLKOPF B.; SMOLA A.; WILLIAMSON R.; BARTLETT P. L. (2000), "New support vector algorithms", *Neural Computation*, (1207-1245)

WU E.; LIU W.; CHAWLA S. (2008), "Spatio-Temporal Outlier Detection in Precipitation Data", *Sensor-KDD*, (6-14)

YUXİANG S.; KUNQİNG X.; XİUJUN M. (2005), "Detecting Spatio-Temporal Outliers in Climate Dataset: A Method Study", *IEEE*, (760-763)

ZHAO J.; LU C. T.; KOU Y. (2003), "Detecting Region Outliers in Meteorological Data", **GIS'03**, (49-55)