Volume: 08 Number: 02 Category: 02 Page: 55-75 ISSN No: 1303-6319 Cilt: 08 Sayı: 02 Kategori: 02 Sayfa: 55-75

METEOROLOJİK ZAMAN SERİLERİNDE KAYIP VERİ TAHMİN YÖNTEMLERİNİN BAŞARIMLARININ KORELASYON BOYUTU ANALİZİYLE KARŞILAŞTIRILMASI

Sipan ASLAN[*] **Ceylan YOZGATLIGİL**[**]
Cem İYİGÜN[***] **İnci BATMAZ**[****] **Hasan TATLI**[******]

ÖZET

Bu çalışmada, meteorolojik zaman serilerinde kayıp veri tahmin yöntemlerinin başarımları doğrusal olmayan dinamik zaman serileri analizinde sıklıkla kullanılan korelasyon boyutu belirleme yöntemiyle karşılaştırılmıştır. Bu amaçla, 1965-2006 periyodunda, eksik veri içermeyen aylık meteorolojik zaman serilerinde farklı oranlarda yapay kayıp veriler oluşturulmuş ve oluşturulan yapay kayıp veriler tahmin yöntemleriyle tamamlanarak bu serilerin korelasyon boyutları orijinal serilerden elde edilen korelasyon boyutlarıyla karşılaştırılmıştır. Elde edilen bulgular doğrultusunda, karekök hata kareler ortalaması gibi sadece merkezi eğilimleri dikkate alan doğruluk ölçüm yöntemleri yanında serilerin alansal ve zamansal özelliklerine hassas bağımlı olan korelasyon boyutu belirleme yönteminin de kullanılmasının zaman serilerinde kayıp veri tahmin yöntemlerinin başarımlarının karşılaştırılmasını daha güvenilir düzeye taşıyacağı gözlenmiştir.

Anahtar Kelimeler: Doğrusal olmayan zaman serileri, Kayıp veri, Korelasyon boyutu, Meteorolojik zaman serileri.

1. GİRİŞ

Meteoroloji istasyonlarının gözlemlerinde kayıp veri problemiyle karşılaşmamak neredeyse imkansızıdır. Meteorolojide kayıp verilerin oluşumu, mevsimsel etkilere, meteoroloji istasyonlarının yer değiştirmesine, kapanmasına, istasyonların kullandıkları ekipmanların özelliklerine, personelden kaynaklı hatalara ve hatta ülkenin içinde bulunduğu siyasal konjonktür gibi nedenlere bağlıdır.

Türkiye genelinde 19 meteorolojik değişken üzerine faal olarak gözlem yapabilen yaklaşık 270 klima istasyonu vardır. 1950 – 1960 yılları arasında bu istasyonların toplamında aylık gözlemlerin içerdiği kayıp veri oranı %50 civarındadır. Kayıp veri oranı 1960 sonrasında %50'nin altına düşmekte ve 1980 yılından sonra da %10 kayıplılık düzeyine ulaşmaktadır (Asar *vd*. 2010).

[*]Ar. Gör., ODTÜ, Fen ve Edebiyat Fakültesi, İstatistik Bölümü, Ankara, e-posta: sipan@metu.edu.tr
[**]Yrd. Doç. Dr., ODTÜ, Fen ve Edebiyat Fakültesi, İstatistik Bölümü, Ankara, e-posta: ceylan@metu.edu.tr
[***]Yrd. Doç. Dr., ODTÜ, Mühendislik Fakültesi, Endüstri Mühendisliği, Ankara, e-posta: iyigun@metu.edu.tr
[****]Doç. Dr., ODTÜ, Fen ve Edebiyat Fakültesi, İstatistik Bölümü, Ankara, e-posta: ibatmaz@metu.edu.tr
[*****]Doç. Dr., ÇOMÜ, Fen ve Edebiyat Fakültesi, Coğrafya Bölümü, Çanakkale, e-posta: tatli@comu.edu.tr

İklim belirleme, kümeleme ve modelleme gibi meteoroloji gözlemlerinin uzun dönem ortalamalarına ihtiyaç duyan istatistiksel analizlerde eksiksiz veri setinin elde edilebilir olması oldukça önemlidir. Bu tür veri kümelerinde kayıpların uygun yöntemlerle tamamlanması, uzun dönem verilere ihtiyaç duyan analizlerin güvenilirliğini arttıracaktır. Uygun yöntemlerin belirlenmesi ilgilenilen değişkenin alansal ve zamansal özellikleriyle de yakından ilgilidir. Bu çalışmada biz, kayıp veri atama yöntemlerinin başarımlarının karşılaştırılmasında sadece merkezi eğilim ölçülerini dikkate alan doğruluk ölçümleri yanında Korelasyon Boyutu (KB) belirleme yönteminin de kullanılmasını önermekteyiz.

Bu çalışmada kayıp veri tahmin yöntemlerini KB'ler açısından karşılaştırmak amacıyla 1965-2006 periyodunda eksik veri içermeyen ve Türkiye'nin sahip olduğu farklı iklim rejimlerini yansıtacak şekilde 7 farklı iklim bölgesinden seçilen meteoroloji istasyonlarının yağış gözlemleri kullanıldı. Kayıp veri içermeyen gözlemlerde %10, %20 ve %50 olmak üzere üç farklı düzeyde oluşturulan yapay kayıp veriler Çok Katmanlı Yapay Sinir Ağları (ÇKYSA) yöntemi ve Beklenti Maksimizasyonu tabanlı Monte Karlo Markov Zinciri (BM-MKMZ) çoklu veri atama yöntemleri kullanılarak tahmin edildi. Tahmin edilen değerlerle gerçek değerler Değişim Katsayısı Hata Kök Kareler Ortalaması (DKHKKO) doğruluk ölçümü kullanılarak karşılaştırıldı. Merkezi eğilim ölçülerine dayanan bu karşılaştırmayı etkinleştirmek için tahmin değerleriyle oluşturulan serilerle orijinal serilerden elde edilen KB değerleri karşılaştırılarak tahminlerin başarımları değerlendirildi.

Çalışmanın 2. bölümünde yağış gözlemleri kullanılan meteoroloji istasyonları, 3.bölümde kayıp tahmini için kullanılan yöntemler ve KB belirleme yöntemi tanıtılmıştır. Son olarak 4. bölümde çalışmanın sonuç ve tartışma bölümleri yer almıştır.

2. VERİLER

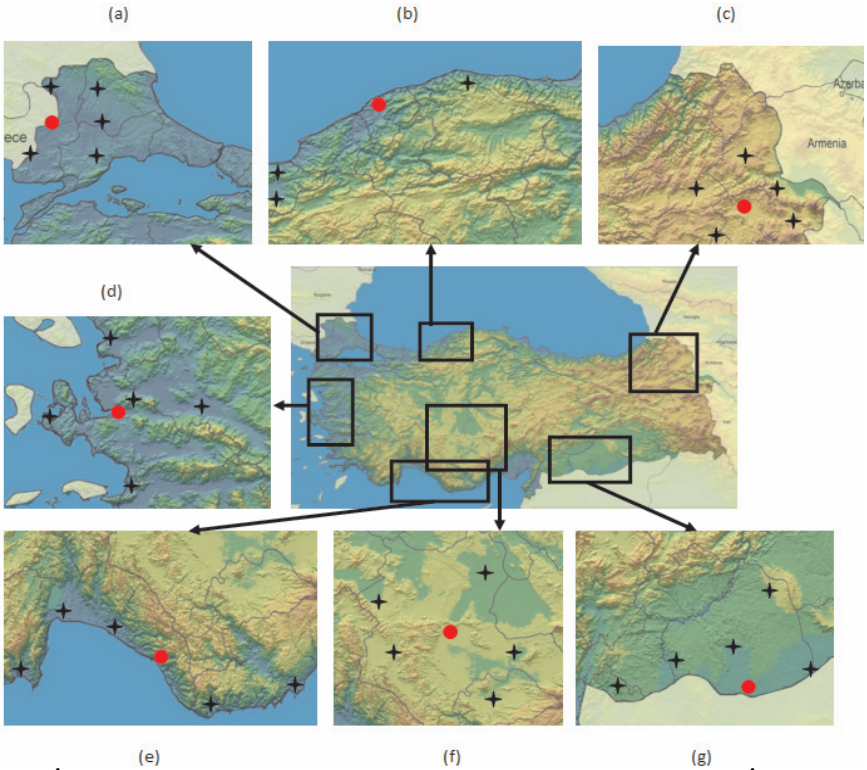
Çalışmada kullanılan yağış verileri Devlet Meteoroloji İşleri Genel Müdürlüğünden (DMİGM) temin edilmiştir. Yağış gözlemleri kullanılan meteoroloji istasyonları 1965-2006 periyodunda kayıp veri içermemektedir. Yağış değişkeninin alansal özelliklerini ve farklı iklim rejimlerinin tahminlerdeki etkilerini de analizlerde değerlendirmek amacıyla Güney Doğu Anadolu Bölgesi (GAB), Doğu Anadolu Bölgesi (DAB), Akdeniz Bölgesi (AB), İç Anadolu Bölgesi (İAB), Ege Bölgesi (EB), Batı Trakya Bölgesi (BTB) ve Kuzey Batı Karadeniz Bölgesi (KBKB) olmak üzere Türkiye'nin 7 farklı iklim bölgesinden seçilen meteoroloji istasyonları kullanılmıştır.

Meteorolojik verilerde karşılaşılan kayıpların tamamlanmasında kullanılan yöntemler genelde yakın istasyonların varolan verilerini kullanmaktadır. Bu çalışmada da kayıp veriye sahip olduğu düşünülen istasyon hedef istasyon olarak, hedef istasyonla korelasyonları yüksek komşu istasyonlar da referans istasyonlar olarak nitelendirilmişlerdir. Gerçekte kayıp veri içermeyen hedef istasyonlarda %10, %20 ve %50 olmak üzere üç farklı kayıplılık düzeyinde kayıplar oluşturulmuş ve bu kayıp veriler komşu istasyonların varolan gözlemleri kullanılarak tahmin edilmiştir. Kullanılan istasyonların lokasyon bilgileri Tablo 1'de verilmiş ve bölgelerde belirlenen hedef istasyonlar *italik* yazıyla gösterilmiştir.

Tablo 1. Çalışmada kullanılan istasyonlar ve lokasyon bilgileri

Istasyon adı	Istasyon numarası	Enlem - Boylam	Istasyon adı	Istasyon numarası	Enlem - Boylam
	GAB			DA	
Akcakale	**17980**	36°43' - 38°56'	***Ağrı***	**17099**	39°43' - 43°03'
Birecik	**17966**	37°01' - 37°57'	**Doğubeyazıt**	**17720**	39°33' - 44°05'
Ceylanpınar	**17968**	36°50' - 40°01'	**Horasan**	**17690**	40°03' - 42°10'
Urfa	**17270**	37°09' - 38°47'	**Igdır**	**17100**	39°55' - 44°03'
Kilis	**17262**	36°42' - 37°06'	**Mus**	**17204**	38°41' - 41°29'
Siverek	**17912**	37°45' - 39°19'			
	AB			İAB	
Alanya	**17310**	36°33' - 32°00'	***Konya***	**17244**	37°59' - 32°33'
Antalya	**17300**	36°52' - 30°42'	**Karaman**	**17246**	37°12' - 33°13'
Finike	**17375**	36°18' - 30°09'	**Akşehir**	**17239**	38°21' - 31°25'
Manavgat	**17954**	36°47' - 31°26'	**Beyşehir**	**17896**	37°41' - 31°44'
Anamur	**17320**	36°05' - 32°50'	**Cihanbeyli**	**17191**	38°39' - 32°57'
Silifke	**17330**	36°23' - 33°56'	**Karapınar**	**17902**	37°43' - 33°32'
	EB			BTB	
İzmir	**17220**	38°23' - 27°04'	***Uzunkopru***	**17608**	41°15' - 26°41'
Kuşadası	**17232**	37°52' - 27°15'	**İpsala**	**17632**	40°55' - 26°22'
Çeşme	**17221**	38°18' - 26°18'	**Kırklareli**	**17052**	41°44' - 27°13'
Dikili	**17180**	39°04' - 26°53'	**Lüleburgaz**	**17631**	41°24' - 27°21'
Bornova	**17790**	38°28' - 27°13'	**Tekirdağ**	**17056**	40°59' - 27°30'
Salihli	**17792**	38°29' - 28°08'	**Edirne**	**17050**	41°41' - 26°33'
				KBKB	
			Bartın	**17020**	41°38' - 32°22'
			Akçakoca	**17612**	41°05' - 31°10'
			Düzce	**17072**	40°50' - 31°10'
			İnebolu	**17024**	41°59' - 33°47'

Kullanılan meteoroloji istasyonlarının yağış değişkeni için birbirleriyle olan ve istatistiksel olarak anlamlı doğrusal korelasyonları 0.65 ve 0.95 arasında değişmektedir. İstasyonların Türkiye fiziki haritasındaki konumları Şekil 1'de verilmiştir. Hedef istasyonlar kırmızı renkli, referans istasyonlar siyah renkli belirteçlerle gösterilmiştir.



Şekil 1. İstasyonların konumları (a) BTB (b) KBKB (c) EB (d) DAB (e) AB (f) İAB (g) GAB

3. YÖNTEMLER

Literatürde, yağış verilerinde kayıpların tahminiyle ilgili çalışmalar 1950'li yıllara kadar uzanmaktadır. İlk çalışmalarda, yağış verilerinde karşılaşılan kayıp verilerin, aritmetik ortalamalar yöntemi gibi basit yöntemlerle tahmin edilmesi yönünde çalışmalar yapılmıştır. Paulhus ve Koehler 1952'de hedef ve referans istasyonların yağış toplam oranlarından elde edilen ağırlıkları kullanarak aritmetik ortalama yönteminin geliştirilmiş hali olan ve Normal Oran (NO) yöntemi olarak adlandırdıkları yöntemi kullanmışlardır. Bilgisayar teknolojisindeki gelişmlerin yardımıyla, kayıp verilerin tahmininde hesap yoğun yöntemlerin kullanımı yaygınlık kazanmıştır.

Young (1992), kayıp verilerin tahmini için Çoklu Diskriminant Analiz yöntemini kullanmıştır. Makhuvha *vd.* (1997), Çoklu Doğrusal Regresyon (ÇDR) yöntemi ve NO yöntemiyle beraber altı ayrı yöntemi karşılaştırmış ve BM algoritmasının daha doğru sonuçlar verdiğine işaret etmişlerdir. Xia *vd.* (1999a,b), günlük meteorolojik zaman serileri üzerinde ki çalışmaları sonucunda ÇDR yönteminin daha başarılı sonuçlar verdiğini rapor etmişlerdir. Schneider (2001), iklim verilerinde karşılaşılan kayıp verilerin ele alınmasında BM algoritmasını geniş çapta değerlendiren ilk çalışmayı yayınlamıştır. Cano ve Andreu (2010), MKMZ tabanlı çoklu veri atama yönteminin kayıp verilerin tahmininde daha başarılı sonuçlar verdiği yönünde bulgular elde etmişlerdir. Coulibaly ve Evora (2007); Lucio *vd.* (2007); Lo Presti *vd.* (2008); Aly *vd.* (2009); Kalteh ve Berndtsson (2006) ve Kalteh ve Hjorth (2009), çalışmalarında kayıp verilerin tahmini içi Yapay Sinir Ağları (YSA) modellerinin kullanıldığı ve dikkate değer sonuçların elde edildiği karşılaştırmalı analiz sonuçlarını yayımlamışlardır.

Yukarıda adı geçen yayınlar ışığında hedef istasyonlarda oluşturulan yapay kayıp verilerin ÇKYSA ve BM-MKMZ çoklu veri atama yöntemleri kullanılarak tahmin edilmesi uygun görülmüştür. Çalışmada kullanılan yöntemlerin KB belirleme yöntemi açısından karşılaştırılması çalışmanın ana amacını oluşturmaktadır.

3.1 Çok Katmanlı Yapay Sinir Ağları Modeli

Yapay sinir ağları, girdi verileriyle ona karşılık gelen çıktı değerlerini yinelemeli olarak eşleyen yarı parametrik doğrusal olmayan regresyon modeli olarak düşünülebilir. Literatürde pek çok farklı YSA modeli geliştirilmiştir. Detaylı bilgi için Bishop (1995); Haykin (1999) ve Patterson (1996) görülebilir. Zaman serilerinin ve iklim gözlemlerinin modellenmesinde ÇKYSA modeli sıklıkla kullanılan bir yöntemdir (örneğin, Gardner ve Dorling, 1998; ASCE, 2000; Toth *vd*., 2000; Junninen *vd*., 2004; Coulibaly ve Evora, 2007; Kalteh ve Berndtsson, 2006; Kalteh ve Hjorth, 2009; Sorjaama, 2009; Nelwamondo ve Marwala, 2007). ÇKYSA ağ mimarisi girdi, gizli ve çıktı olmak üzere genelde üç katmandan oluşmaktadır. Bu modelde YSA, hataların geri yayılım algoritması kullanılarak eğitilmektedir. Kulanılacak gizli katman sayısı ve gizli katmandaki nöron sayısı deneme yanılma sonucunda bulunabilen değişkenlerdir. Bu çalışmada bir gizli katman ve bu katmanlarda 2-8 sayıları arasında değişen nöronlar kullanılmıştır. Gizli katmanda bulunan nöronlarda hiperbolik tanjant fonksiyonu seçilmiştir ve YSA eğitim algoritması için eşlenik ölçekli gradyan iniş algoritması kullanılmıştır. ÇKYSA modellerinin literatürde yaygın kullanımı nedeniyle model hakkında detaylı bilgiler verilmeyecektir onun yerine ilgilenen okuyucular yukarıda belirtilen referanslara başvurabilirler.

3.2 Beklenti Maksimizasyonu Monte Karlo Markov Zinciri Tabanlı Çoklu Veri Atama Yöntemi

Kayıp veri kümelerine En Çok Olabilirlik (EÇO) Yöntemi tabanlı yaklaşımlarda, uygun kayıp veri atama yönteminin belirlenmesi kayıp veriyi üreten mekanizmayla yakından ilgilidir (Little ve Rubin, 2002; Schafer, 1997). Rassal Kayıp Mekanizması (RKM), Tümüyle Rassal Kayıp Mekanizması (TRKM) ve Rassal olmayan Kayıp Mekanizması (ROKM) olmak üzere literatürde yagın olarak başvurulan üç kayıplılık mekanizması mevcuttur (Little veRubin, 2002). RKM kayıpların kayıp değişkene bağlı olmadığı ancak gözlenen değişkene bağlı olabileceği durumu, TRKM kayıpların ne kayıp değişkene ne de gözlenen değişkene bağlı olduğu durumu son olarak ROKM kayıpların kayıp değişkene ve gözlenen değişkene bağlı olabileceği durumu ifade etmektedir. RKM ve TRKM varsayımları altında kayıp veri atama modelinde kayıplılık mekanizması göz ardı edilebilir durumlar olarak değerlendirilmektedir ve kayıplılık mekanizmasının modellenmesine ihtiyaç yoktur ancak ROKM varsayımının kabul edilmesi durumunda kayıp veri atama modellerinde kayıplılık mekanizmasının da modellenmesi gerekmektedir (Schafer, 1997).

BM algoritması 1977'de ilk olarak Dempster, Laird ve Rubin tarafından eksik veri içeren veri kümlerinde EÇO tahmincilerinin elde edilmesi için geliştirilmiş yinelemeli bir yöntemdir. Algoritmanın B- adımında eldeki veriler yardımıyla dağılımın EÇO tahmincileri bulunur M- adımında ise bir önceki adımda bulunan tahminciler yardımıyla kayıp verilerin koşullu beklenen değerleri elde edilir. Parametre tahminleri, ardışık parametre tahminleri durağanlaşıncaya kadar iterativ olarak yinelenerek elde edilir. Bu

çalışmada BM algoritması MKMZ çoklu veri atama yöntemi için başlangıç parametre tahminlerini sağlamak için kullanılmıştır.

MKMZ çoklu veri atama yöntemi de BM algoritmasında olduğu gibi yinelemeli bir süreçtir. Çoklu veri atama yöntemide kayıp veri yerine ikiden fazla uygun değerin atanmasıyla kayıplılıktan ötürü oluşan belirsizliğin veri analizlerinde etkisinin indirgenmesi amaçlanmaktadır. MKMZ yönteminde çoklu veri atama, Atama adımı ve Ardıl adım olmak üzere iki adımda gerçekleştirilir. Atama adımında parametre tahminleri kullanılarak kayıp veriler simule edilmektedir. Ardıl adımda, tahmin edilen kayıp verilerle tamamlanmış veri seti üzerinden yeni parametre tahminleri elde edilir. Bu iki adım, simule edilen kayıp değerler ve parametre tahminleri birbirlerinden bağımsız hale gelinceye kadar devam ettirilir ve en son iterasyonda ki kayıp veri tahminleri ve parametre tahminleri Markov Zinciri'nin bir elemanını oluştururlar. Bu çalışmada MKMZ tabanlı çoklu veri atama yöntemi RKM ve çok değişkenli normal dağılım (ÇDND) varsayımları altında çalışıldı. Yağış verileri için ÇDND varsayımı gerçekçi olmasada, kayıp verilerin ÇDND varsayımı altında tahmini yinede başarılı sonuçlar vermektedir (Demirtas *vd*., 2008). Yöntemin ayrıntılı uygulama aşamaları için Schafer (1997) görülebilir.

3.3 Doğrusal Olmayan Zaman Serileri Analizi ve Korelasyon Boyutu Analizi

Son yıllarda meteorolojik zaman serilerinin doğrusal olmayan deterministik özelliklerinin belirlenmesinde dinamik sistem yaklaşımı oldukça yaygınlık kazanmıştır (Sivakumar, 2004). Bu amaçla; faz uzayı geri kurulumu algoritması serinin doğrusal-dışı deterministik özelliklerini incelemek için sıklıkla kullanılan bir yöntem olarak karşımıza çıkmaktadır. Faz uzayı herhangi bir dinamik sistemin tanımlanmasında kullanılan ve sistemin zaman içindeki hareketinin özelliklerini barındıran koordinat düzlemleridir. Diferansiyel denklemlerle ifade edilebilen dinamik sistemler için faz uzayları, ilgili denklemlerin matematiksel çözümleriyle oluşturulabilmektedir. Dinamik sistemi oluşturan diferansiyel denklemlerin bilinmediği ya da belirlenemediği durumlarda ilgili sisteme ait faz uzayını sistemin herhangi bir değişkeninin zamana bağlı gözlemlerini kullanarak faz uzayı geri kurulum algoritmasıyla oluşturmak mümkün olabilmektedir. Zaman dizileri kullanılarak elde edilen faz uzayları bu dizileri üreten sistemin nitel ve nicel özellikleri hakkında bilgiler içermektedir (Kantz ve Schreiber, 2003). Örneğin faz uzayında gelişi güzel saçılan ve uzayın tümüne yayılan koordinat değerleri sistemin tümüyle raslantısal bir süreç olduğunu belirtirken buna karşılık faz uzayında uzayın belli bir bölümünde yoğunlaşan belirgin geometrik yapılar sistemin deterministik özellikte olduğuna ilişkin kuvvetli belirleyicilerdir. Teorik olarak deterministik özellikte dinamik sistemlerin zamana bağlı hareketi faz uzayında belirli geometrik yapılar oluşturacak şekilde bir yörüngeye doğru çekilirler. Bu tür geometrik yapılar çekici(attractor) olarak adlandırılmaktadırlar. Dinamik sistemlerin doğrusal-dışı özellikleri hakkında bilgiler, elde edilen bu çekicilerin boyut analizleri sonucunda ortaya konmaktadır. Bu geometrik yapıların boyutlarının belirlenilmesinde kullanılan en yaygın yöntem korelasyon boyutu analizidir. Sonlu ve kesirsel korelasyon boyutu ilgili zaman dizisini üreten dinamik sistemin periyodik olmadığını doğrusal olmayan özellikte olduğunu belirtirken kesirsel boyutun tam sayı değeri dinamik sistemin serbestlik derecesi olarak değerlendirilmekte ve dinamik sistemi türeten değişken sayısı hakkında bilgi taşımaktadır (Small, 2005).

Bu çalışmada biz yapay olarak oluşturulan kayıp verileri, yukarıda anlatılan yöntemlerle, tahmin edilen yağış serilerinden elde edilen KB'lerle gerçek yağış serisilerinden elde edilen KB'leri karşılaştırarak yöntemlerin başarımlarını daha etkin bir biçimde karşılaştırmayı hedefledik. KB belirleme yönteminin zaman serisindeki değişimlere hassas bağımlı olması nedeniyle kayıp verisi tamamlanan KB'lerle gerçek KB'lerden olan sapmaların kullanılan yöntemlerin başarımları hakkında bilgi verici olduğunu düşünmekteyiz. KB'yi belirlemek için öncelikle sitemin yaklaşık faz uzayını tek değişkenli zaman serisinden yararlanarak elde edilen geri çatım vektörleriyle geri-kurmak gerekmektedir.

Öncelikle herhangi bir dinamik sistemi belirleyen set üçlüsü $(\mathcal{M}, \Phi, \mathcal{T})$ ile gösterilsin. Burada $\mathcal{M}$, sistemin üzerinde hareket ettiği m boyutlu gerçek faz uzayını; $\mathcal{T}$, zaman belirtecini ve Φ , sistemi hareket ettiren dönüşüm operatörünü göstermektedir. Doğrusal-dışı dinamik sistem analizinde amaç; dinamik sistem hakkında bilgi taşıyan ve matematiksel olarak elde edilemeyen $\mathcal{M}$ 'i yaklaştırımlarla belirlemeye çalışmaktır. Bu yaklaştırım faz uzayı geri-kurulum algoritmasıyla yapılabilmektedir.

3.3.1 Zaman geciktirmeye dayalı yerleştirme yöntemiyle faz uzayı geri-kurulumu

Doğrusal dışı dinamik zaman serisi analizinde, tek değişkenli zaman serisi y_n, $g: \mathcal{M} \to \mathbf{R}$ olmak üzere herhangi bir $g(.)$ fonksiyonula $\mathcal{M}$ 'nin izdüşümü olarak düşünülmektedir. Tek değişkenli zaman serisi y_n kullanılarak d boyutta $\mathcal{M}$ 'nin geri kurulumu, zaman geciktirmeye dayalı yerleştirme yöntemiyle yapılabilmektedir (Takens ,1981; Small ,2005).

Zaman geciktirmeye dayalı yerleştirme yönteminde zaman serisi y_n, d_e boyutta ve τ zaman gecikme değeriyle geri-çatım vektörlerine dönüştürülmektedir.

Zaman serisi y_n ve $n = 1,2,3,\ldots\ldots,N$ olmak üzere oluşturulacak geri-çatım vektörlerinin sayısı $N-(d_e-1)\tau$ olacaktır,

$$x_i = \left[y_i, y_{i-\tau}, y_{i-2\tau}, \ldots\ldots\ldots, y_{i-(d_e-1)\tau}\right], \qquad i = 1,2,3,\ldots\ldots\ldots, N-(d_e-1)\tau. \quad (1)$$

Faz uzayında geri çatım vektörlerinin oluşturduğu geometrik şekiller garip çekerler (strange attractor) olarak adlandırılmaktadır (Takens, 1981). Burada d_e, minimum yerleştirme boyutunu ve τ , zaman gecikme değerini göstermektedir.

Zaman gecikme değerini belirleyebilmek için otokorelasyon fonksiyonunun ilk $\frac{1}{e}$ değerine düştüğü değer gecikme değeri olarak alınabilir ancak bu çalışmada biz uygun zaman gecikme değerinin bulunması için Fraser and Swinney, (1986)'in önerdiği Ortak Bilgi Kriterini (OBK) kullandık. Zaman gecikme değerinin uygun belirlenimi oluşturulacak geri-çatım vektörlerinin birbirlerine istatistiksel olarak bağımlı ancak korelasyonlu olmamasına dayanmaktadır. Küçük zaman gecikme değeri vektörlerin birbirine çok yakın olmasına ve faz uzayında vektörlerin bir noktada yoğunlaşmasına, gecikme değerinin büyük seçilmesiyse oluşturulan vektörlerin birbirleriyle bağımsız olmasına ve faz uzayında vektörlerin gelişi güzel dağılmasına neden olmaktadır. Bu

nedenle OBK'nin ilk minimum değeri uygun zaman gecikme değeri olarak alınabilir (Kantz ve Schreiber, 2003).

Minimum yerleştirme boyutu d_e'nin elde edilmesinde yaygın olarak kullanılan yöntem Kennel *vd.*, (1992)'nin önerdiği Yanlış En Yakın Komşuluklar yöntemidir. Minimum yerleştirme boyutu bulunduğu takdirde zaman serisini üreten dinamik sistem için uygun faz uzayı elde edilmiş olur. Bu yöntemin amacı d_e boyutundan d_e+1 boyutuna geçerken x_i yörüngesindeki doğru ve yanlış komşulukları ayırt etmektir. Bir zaman serisinde yanlış komşuluk, sisteme ait iki noktanın, sistemin olduğundan daha küçük bir boyutta incelenmesinden dolayı komşuluk olarak değerlendirilmesidir.

3.3.2 Korelasyon boyutu

Korelasyon boyutu, faz uzayı geri-kurulum prosedürünün çıktılarını girdi olarak alır ve minimum yerleştirme boyutu d_e 'nin ve zaman gecikme değeri τ 'nun bir fonksiyonu olarak düşünülebilir. Faz uzayında oluşan yapının boyut analizi korelasyon boyutunu vermektedir. Boyut analizinde amaç dinamik sistemi forse eden değişkenliğin belirlenmesidir. KB çekici üzerindeki bir noktanın başka bir noktayı etkileme derecesinin ölçümüdür ve serideki değişimlere hassas bağımlıdır (Sivakumar vd., 2006).

Bu çalışmada KB tahminleri Grassberger ve Proccacia (1983)'nın önerdiği (GP algoritması) KB belirleme prosedürüyle gerçekleştirilmiştir. GP algoritması KB hesabı için korelasyon toplamını kullanmaktadır. Korelasyon toplamı düşük boyutlu ve yüksek boyutlu dinamik sistemleri ayırt etmek için kullanılan bir yöntemdir. Herhangi bir deterministik sistemin serbestlik derecesinin sonlu sayıda olması gerektiğinden hareketle, artan minimum yerleştirme boyutu için hesaplanan korelasyon toplamlarının belirli bir noktadan (serbestlik derecesi) sonra doyuma ulaşması beklenmektedir. KB, artan yerşetirme boyutu için korelasyon toplamında değişimin stabil hale geldiği nokta olarak düşünülebilir.

Herhangi bir faz uzayında tanımlı x_i noktalar topluluğu düşünüldüğünde, bu noktalar arasındaki komşuluk uzaklığının belirli bir değerden, ε , yakın olan noktaların sayısının, toplam nokta sayısına oranı korelasyon toplamı olarak adlandırılır. $W = N-(d_e-1)\tau$ olmak üzere bu toplam şöyle ifade edilir;

$$C(W, d_e, \varepsilon) = \frac{2}{W(W-1)} \sum_{i=1}^{W} \sum_{j=i+1}^{W} \Theta\left(\varepsilon - (\|x_i - x_j\|)\right) \qquad (2)$$

burada Θ Heavside fonksiyonudur ve aşağıdaki şekilde tanımlanır,

$$\Theta(x) = \begin{cases} 0, & x \leq 0 \\ 1, & x > 0 \end{cases}$$

Başka bir ifadeyle, korelasyon toplamı aralarındaki uzaklık ε ' dan küçük olan (x_i, x_j) noktalarını tanımlar.

Grassberger ve Proccacia (1983)'nın çalışmalarında gösterdiği üzere $\varepsilon \to 0$ ve $N \to \infty$ için korelasyon toplamı $C(W, d_e, \varepsilon)$ ve ε arasındaki ilişki KB , D_c cinsinden şu şekilde ifade edilir;

$$D_c = \lim_{\varepsilon \to 0} \lim_{N \to \infty} \frac{\log C(\varepsilon)}{\log \varepsilon} \qquad (3)$$

ya da başka bir ifadeyle,

$$C(W, d_e, \varepsilon) \propto \alpha \varepsilon^{D_c} \Rightarrow \log C(W, d_e, \varepsilon) \sim \log \alpha + D_c \log \varepsilon \qquad (4)$$

olarak gösterilir. KB, D_c Denklem 4' de gösterilen ifadenin eğimi olarak bulunur.

4. SONUÇ VE TARTIŞMA

Bu çalışmada kayıp veri tahmin yöntemlerinden ÇKYSA ve BM-MKMZ tabanlı çoklu veri atama yöntemleri aylık toplam yağış serileri üzerinde çalışılmıştır. Yapılan kayıp veri tahminlerinin ölçeksiz olarak değelendirilmesi için DKHKKO doğruluk ölçümüyle beraber eksik verileri tamamlanan serilerin KB'lerinde ki değişimler değerlendirilmiştir. Ölçekten bağımsız olarak düşünülen ortalama mutlak yüzde hata ölçümü ve simetrik mutlak yüzde hata ölçümü gibi yüzde hata oranlarına dayanan diğer doğruluk ölçümlerinin kullanımı Hyndman ve Koehler (2006)'in belirttikleri dez avantajlardan dolayı uygun bulunmamıştır.

Çalışmada kullanılan DKHKKO doğruluk ölçümü, a_i gerçek değer ve e_i tahmin edilen değer olmak üzere şu şekilde ifade edilebilir;

$$\text{DKHKKO} = \frac{\text{HKKO}}{\overline{a_{(n)}}} = \frac{\sqrt{\sum_{i=1}^{n} (a_i - e_i)^2 / n}}{\overline{a_{(n)}}}. \qquad (5)$$

Denklem 5'te görüldüğü gibi DKHKKO, $\overline{a_{(n)}}$ tahmin edilen değerlerin ortalaması olmak üzere ölçekten bağımsız hale getirilmiş HKKO doğruluk ölçümüdür. Tablo 1'de belirtilen hedef istasyonlarda %10, %20 ve %50 olmak üzere üç farklı kayıplılık düzeyinde oluşturulan yapay kayıp veriler tahmin edildikten sonra hesaplanan DKHKKO değerleri Tablo 2'de verilmiştir

Tablo 3. DKHKKO değerleri

	17099 Ağrı			17980 Akçakale			17310 Alanya			17020 Bartın		
	%10	%20	%50	%10	%20	%50	%10	%20	%50	%10	%20	%50
ÇKYSA	**0.434**	**0.388**	0.431	**0.667**	0.577	0.551	**0.430**	**0.429**	0.581	0.400	0.462	0.458
BM-MKMZ	0.444	0.395	**0.429**	0.674	**0.573**	**0.537**	0.465	0.466	**0.511**	**0.354**	**0.427**	**0.455**

	17220 İzmir			17244 Konya			17608 Uzunköprü		
	%10	%20	%50	%10	%20	%50	%10	%20	%50
ÇKYSA	0.292	0.276	0.303	0.478	0.391	0.623	0.392	0.401	0.387
BM-MKMZ	**0.288**	**0.272**	**0.296**	**0.477**	**0.389**	**0.560**	**0.382**	**0.401**	**0.375**

Aynı şekilde üç farklı kayıplılık düzeyi için hesaplanan KB'ler ve orjinal seriye ait KB değerleri Tablo 3'te verilmiştir.

Tablo 4. KB değerleri

	17099 Ağrı			17980 Akçakale			17310 Alanya			17020 Bartın		
	D_c = 3.78			D_c = 2.58			D_c = 3.62			D_c = 3.41		
	%10	%20	%50	%10	%20	%50	%10	%20	%50	%10	%20	%50
ÇKYSA	3.8	3.76	4.22	2.58	2.64	2.59	3.49	3.46	3.49	3.52	3.42	3.35
BM-MKMZ	3.8	3.75	3.93	2.58	2.60	2.63	3.53	3.49	3.95	3.47	3.52	3.25

	17220 İzmir			17244 Konya			17608 Uzunköprü		
	D_c = 2.27			D_c = 3.96			D_c = 5.31		
	%10	%20	%50	%10	%20	%50	%10	%20	%50
ÇKYSA	2.27	2.27	2.27	3.94	3.76	3.93	5.26	5.07	5.29
BM-MKMZ	2.27	2.27	2.19	3.81	3.70	4.23	5.33	5.17	5.52

İklim belirleme ve kümeleme analizleri gibi klimatolojik çalışmalar eksiksiz ve olabildiğince uzun dönemli veriye ihtiyaç duymaktadır. Bu nedenle, meteorolojik zaman serilerinde sıklıkla karşılaşılan kayıp verilerin tahmini için kullanılacak yöntemlerin, serilerin içermiş olduğu alansal ve zamansal özellikleri koruyabilecek derecede etkin olduklarını belirlemek sonraki analizler açısından oldukça önemlidir. Zaman serilerinde tahminlerin başarımlarını belirlemenin en yaygın yöntemi olarak, hata kareler ortalamasına dayanan doğruluk ölçümleri kullanılmaktadır. Sadece merkezi eğilimleri dikkate alan bu tür değerlendirmelerin alansal özelliğe sahip veriler için kullanılan yöntemlerin etkinliğini belirlemede yeterli olamayacağını düşünmekteyiz.

KB'ler zaman serisindeki hassas değişimlere duyarlı olduğundan ve uygun faz uzayı geri-kurulumu altında değişmez (invariant) olduklarından, HKKO'ya dayanan yöntem karşılaştırmalarının KB'lerle birlikte desteklendiği durumlarda, karşılaştırma sonuçlarının daha etkin yapılabileceğini düşünmekteyiz. KB'ler ilgili dinamik sistemin boyutları hakkında da bilgi içerdiklerinden, hesaplanan gerçek KB değerlerinin ilgili bölgedeki yağış sistemi hakkında bilgi taşıdığını söyleyebiliriz. Örneğin, Türkiye'nin Batı Trakya bölgesi istikrarlı iklim koşulları sergilememektedir. Bu bölgede bazı yıllarda Akdeniz iklimi etkisi görülürken bazı yıllarda Karadeniz iklimi etkisi gözlenmektedir. Bu bölge iklim bilimciler tarafından iklim rejimlerinin geçiş bölgesi (transition zone) olarak tanımlanır ve iklimi değişkendir. Buna bağlı olarak yağış rejimindeki değişim de diğer bölgelere göre daha fazla değişkendir. Uzunköprü istasyonu için hesaplanan KB değerinin diğer istasyonlara göre yüksek olması bu olgunun KB analizinde tespit edildiğini göstermektedir. Aynı şekilde, EB'de yer alan İzmir ve GAB'ta yer alan Akçakale gibi yağışların ve yağış değişkenliğinin az olduğu havzalarda KB'lerin diğer bölgelere göre düşük olarak hesaplanması KB analizinin alansal özellikleri yakalayabildiğini göstermektedir.

Aylık toplam yağış verilerinde kayıpların etkin tahmini, seçilen referans istasyonların uygunluğuna bağlıdır. Tablo 3'te görüldüğü gibi her iki yöntem ve üç farklı kayıplılık düzeyi için hesaplanan KB değerleri orijinal serilerden elde edilen KB değerlerine oldukça yakındır. Kullanılan yöntemlerin serilerin alansal ve zamansal özelliklerini dikkate aldığını söyleyebiliriz. Daha ileri klimatolojik araştırmalarda, eksik verisi

tamamlanan meteorolojik serilerin analizler için kullanılabilir olup olmadığını belirlemek, disiplinler arası çalışmaların güvenilir analiz yorumlarına bağlıdır.

Bu çalışmayla birlikte, KB analizini daha fazla yöntem, değişken ve kayıplılık senaryoları üzerinde değerlendirmek, KB'lerden olan sapmaların istatistiksel olarak ifade ettiği anlamı açıklayabilmek ileriki araştırma konularımız olarak ortaya çıkmıştır.

5. KAYNAKLAR

Aly, A., Pathak, C., Teegavarapu, R. S. V., Ahlquist, J., Fuelberg, H., 2009. Evaluation of Improvised Spatial Interpolation Methods For Infilling Missing Precipitation Records. Paper presented at the Proceedings of World Environmental and Water Resources Congress , 342 5914-5923.

Asar, Ö., Kartal, E., Aslan, S., Öztürk, M.Z., Yozgatlıgil, C., Çınar, İ., Batmaz, İ., Köksal, G., Türkeş, M., Tatlı, H., 2010. Descriptive Analysis of Turkish Precipitation data with Data Mining Methods. Presented at the 7th National Symposium of Statistics Days, Ankara, Middle East Technical University.

Bishop, C. M., 1995. Neural Networks for Pattern Recognition. Oxford University Press, New York, USA.

Cano, S., Andreu, J., 2010. Using Multiple Imputation To Simulate Time Series: A Proposal To Solve The Distance Effect. WSEAS Transactions on Computers, 9(7), 768-777.

Coulibaly P., Evora N.D., 2007. Comparison of Neural Network Methods For Infilling Missing Daily Weather Records. Journal of Hydrology, Vol. 341, pp. 27-41.

Demirtas, H., Freels, S.A., Yucel, R.M., 2008. Plausibility of Multivariate Normality Assumption When Multiply Imputing Non-Gaussian Continuous Outcomes: A Simulation Assessment. Journal of Statistical Computation and Simulation, 78 (1), pp. 69-84.

Dempster A.P., Laird N.M., Rubin D.B., 1977. Maximum Likelihood From Incomplete Data via the EM Algorithm. Journal of the Royal Statistical Society .B., 39, pp. 1-38.

Fraser, A. M. ve Swinney, H. L. 1986. Independent Coordinates for Strange Attractors from Mutual Information. Phys. Rev. A, 33, 1134.

Gardner, M. W., Dorling, S. R., 1998. Artificial Neural Networks (The Multiplayer Perceptron)--A Review of Applications in The Atmospheric Sciences. Atmospheric Environment, 32, 2627-2636.

Grassberger, P., Procaccia, I., 1983. Measuring The Strangeness of Strange Attractors. Physica D, 9,189-208.

Haykin, S., 1999. Neural Networks: A Comprehensive Foundation. 2nd Edition, Prentice-Hall.

Hyndman, R. J., Koehler, A. B., 2006. Another Look at Measures of Forecast Accuracy. International Journal of Forecasting, 22 (4), 679-688.

Junninen, H., Niska, H., Tuppurainen, K., Ruuskanen, J., Kolehmainen, M., 2004. Methods For Imputation of Missing Values in Air Quality Data Sets. Atmospheric Environment, 38(18), 2895-2907.

Kalteh, A. M., Berndtsson, R., 2007. Interpolating Monthly Precipitation By Self-Organizing Map (SOM) And Multilayer Perceptron (MLP). Hydrological Sciences Journal, 52(2), 305-317.

Kalteh, A. M., Hjorth, P., 2009. Imputation of Missing Values in a Precipitation-Runoff Process Database. Hydrology Research, 40(4), 420-432.

Kantz, H., Schriber, T., 2003. Nonlinear Time Series Analysis. Cambridge University Press, Cambridge UK. 2nd Edition.

Kennel, M. B., Brown, R., Abarbanel, H. D. I., 1992. Determining Embedding Dimension For Phase-Space Reconstruction Using A Geometrical Construction. Phys. Rev. A, 45, 3403. Reprinted in Ott et al. (1994).

Little, R. J. A., Rubin, D. B., 2002. Statistical Analysis with Missing Data. 2nd Edition. Chichester: Wiley.

Lo Presti, R., Barca, E., Passarella, G., 2010. A Methodology For Treating Missing Data Applied To Daily Rainfall Data in The Candelaro River Basin (Italy). Environmental Monitoring and Assessment ,160 (1-4), pp. 1-22.

Makhuvha, T., Pegram, G., Sparks, R., Zucchini, W., 1997. Patching Rainfall Data Using Regression Methods. 2. Comparisons of Accuracy, Bias And Efficiency. Journal of Hydrology, 198(1-4), 308-318.

Paulhus, J. L. H., Kohler, M. A., 1952. Interpolation of Missing Precipitation Records. Mon. Weather Rev. 80, pp. 129–133.

Schafer, J. L., 1997. Analysis of Incomplete Multivariate Data. London: Chapman and Hall / CRC Press.

Schneider, T., 2001. Analysis of Incomplete Climate Data: Estimation of Mean Values and Covariance Matrices and Imputation of Missing Values. Journal of Climate, Vol. 14, pp. 853–871.

Sivakumar, B., 2004. Chaos Theory in Geopyhsics: Past, Present and Future. Chaos, Solitons and Fractals. No:19, Sh. 441-462.

Sivakumar, B., Wallender, W. W., Horwath, W. R., Mitchell, J. P., Prentice, S. E., Joyce, B. A., 2006. Nonlinear Analysis of Rainfall Dynamics in California's Sacramento Valley. Hydrological Processes, No:20 (8), Sh. 1723-1736.

Small, M., 2005. Applied Nonlinear Time Series Analysis: Applications in Physics, Physiology and Finance. Nonlinear Science Series A, World Scientific.vol 52.

Takens, F., 1981. Detecting Strange Attractors in Turbulence. Lecture Notes in Math. Vol. 898, Springer, New York.

Toth, E., Brath, A., Montanari, A., 2000. Comparison of Short-Term Rainfall Prediction Models For Real-Time Flood Forecasting. Journal of Hydrology, 239(1-4), 132-147.

Xia Y., Fabian P., Stohl A., Winterhalter M., 1999a. Forest Climatology: Estimation of Missing Values For Bavaria Germany. Agricultural and Forest Meteorology, Vol. 96 (1-3), pp. 131-144.

Xia, Y., Fabian, P., Stohl, A., Winterhalter, M., 1999b. Forest Climatology: Reconstruction of Mean Climatological Data for Bavaria, Germany. Agricultural and Forest Meteorology, 96(1-3), 117-129.

Young, K.C., 1992. A Three-Way Model For Interpolating For Monthly Precipitation Values. Monthly Weather Review,Vol. 120., pp. 2562–2569.

COMPARISON OF MISSING DATA IMPUTATION METHODS FOR METEOROLOGICAL TIME SERIES DATA VIA CORRELATION DIMENSION TECHNIQUE

ABSTRACT

In this study, the performances of missing value imputation methods for meteorological data are compared by Correlation Dimension Technique, which is frequently used in nonlinear dynamic time series analysis. For this purpose, artificial missing data sets are created with different missing data ratios from complete monthly meteorological time series in the spanning period of 1965-2006. Comparisons were made between original Correlation Dimensions which are calculated by using complete time series and with Correlation Dimensions calculated on re-estimated time series. Since the Correlation Dimension is highly dependent on auto-correlation structures of time series and according to our findings using Correlation Dimension, besides central tendency measures, will make comparisons more appropriate and reliable.

Keywords: Nonlinear time series, Missing data, Correlation dimension, Meteorological time series.