



**İSTANBUL ÜNİVERSİTESİ  
FEN BİLİMLERİ ENSTİTÜSÜ**

**YÜKSEK LİSANS TEZİ**

**ZAMAN DİZİSİ VERİLERİNDE ANİ  
DEĞİŞİMLERİN TAHMİNİ İÇİN VERİ  
MADENCİLİĞİ YÖNTEMLERİ İLE ÖZNİTELİK  
SEÇİMİ**

**Ahmet DEVELİ**

**Bilgisayar Mühendisliği Anabilim Dalı**

**Tezli Yüksek Lisans Programı**

**Danışman**

**Doç Dr. Olcay KURŞUN**

**İSTANBUL**



**İSTANBUL ÜNİVERSİTESİ  
FEN BİLİMLERİ ENSTİTÜSÜ**

## **YÜKSEK LİSANS TEZİ**

**ZAMAN DİZİSİ VERİLERİNDE ANİ  
DEĞİŞİMLERİN TAHMİNİ İÇİN VERİ  
MADENCİLİĞİ YÖNTEMLERİ İLE ÖZNİTELİK  
SEÇİMİ**

**Ahmet DEVELİ**

**Bilgisayar Mühendisliği Anabilim Dalı**

**Tezli Yüksek Lisans Programı**

**Danışman**

**Doç Dr. Olcay KURŞUN**

**Haziran, 2013**

**İSTANBUL**

Bu çalışma 21/06/2013 tarihinde aşağıdaki jüri tarafından Bilgisayar Mühendisliği Anabilim Dalı Yüksek Lisans programında Yüksek Lisans Tezi olarak kabul edilmiştir.

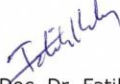
Tez Jürisi



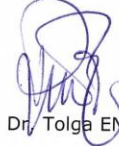
Doç. Dr. Olcay KURŞUN (Danışman)  
İstanbul Üniversitesi  
Mühendislik Fakültesi



Prof. Dr. Ahmet SERTBAŞ  
İstanbul Üniversitesi  
Mühendislik Fakültesi



Yrd. Doç. Dr. Fatih KELEŞ  
İstanbul Üniversitesi  
Mühendislik Fakültesi



Yrd. Doç. Dr. Tolga ENSARİ  
İstanbul Üniversitesi  
Mühendislik Fakültesi



Yrd. Doç. Dr. Niyazi KILIÇ  
İstanbul Üniversitesi  
Mühendislik Fakültesi

## ÖNSÖZ

Bu çalışma, İstanbul Üniversitesi Fen Bilimleri Enstitüsü Bilgisayar Mühendisliği Anabilim Dalı'nda yapılan **Zaman Dizisi Verilerinde Ani Değişimlerin Tespiti İçin Makine Öğrenmesi Yöntemleri ile Öznitelik Seçimi** adlı yüksek lisans tez çalışmasını içermektedir.

Tez çalışmalarım boyunca gösterdiği her türlü destek ve yardımlarından dolayı çok değerli danışmanım Doç. Dr. Olcay KURŞUN'a, çalışmada yer alan veri setini gönderen Çevre ve Şehircilik Bakanlığı'nın Ankara Gölbaşı Laboratuvarı çalışanlarına, çalışma boyunca destek ve tecrübelerini esirgemeyen Araş. Gör. Ergün GÜMÜŞ'e, Araş. Gör. M. Erdem İSENKUL'a, Cihat CENGİZ'e ve tüm çalışma arkadaşlarıma, en içten dileklerle teşekkür ederim.

Son olarak eğitim hayatım boyunca maddi ve manevi desteklerini esirgemeyen sevgili aileme en içten duygularıyla teşekkür eder, sevgilerimi sunarım.

**Haziran, 2013**

**Ahmet DEVELİ**

# İÇİNDEKİLER

Sayfa No

ÖNSÖZ.....	i
İÇİNDEKİLER .....	ii
ŞEKİL LİSTESİ.....	iv
TABLO LİSTESİ .....	v
SEMBOL LİSTESİ.....	vi
ÖZET.....	vii
SUMMARY .....	viii
<b>1. GİRİŞ.....</b>	<b>1</b>
<b>2. GENEL KISIMLAR .....</b>	<b>4</b>
2.1 OZON .....	4
2.1.1. Stratosferik Ozon.....	4
2.1.2. Troposferik Ozon .....	4
2.2 MAKİNE ÖĞRENMESİ METODLARININ OZON TAHMİNİNDE KULLANILMASI.....	5
<b>3. MALZEME VE YÖNTEM .....</b>	<b>13</b>
3.1. OZON VERİKÜMESİ ZAMAN DİZİSİ VERİSİ.....	13
3.2. OZON SEVİYESİ ANİ DEĞİŞİMLER VERİKÜMESİ.....	20
3.3. TEMEL İSTATİSTİKİ KAVRAMLAR.....	21
3.4. KANONİK BAĞINTI ANALİZİ.....	23

3.5.	ISOMAP.....	27
3.6.	DOĞRUSAL AYIRTAÇ ANALİZİ.....	28
3.7.	DESTEK VEKTÖR MAKİNELERİ.....	31
3.8.	NARX YAPAY SİNİR AĞI MODELİ.....	32
3.9.	OZON VERİ KÜMESİ ZAMAN DİZİSİ VERİSİ İLE OZON KESTİRİMİ.....	33
3.10.	ANİ DEĞİŞİMLERİN KESTİRİMİ İÇİN ÖZNİTELİK SEÇİMİ.....	34
<b>4.</b>	<b>BULGULAR.....</b>	<b>36</b>
4.1.	OZON ZAMAN SERİSİ VERİSİNDE KESTİRİM UYGULAMALARI ..	36
4.2.	ACL-OZONE VERİSİNDE ANİ DEĞİŞİM KESTİRİMİ İÇİN ÖZNİTELİK SEÇİMİ.....	43
<b>5.</b>	<b>TARTIŞMA VE SONUÇ.....</b>	<b>57</b>
	<b>KAYNAKLAR.....</b>	<b>1</b>
	<b>ÖZGEÇMİŞ.....</b>	<b>4</b>

## ŞEKİL LİSTESİ

Sayfa No

Şekil 3.1: PM <sub>10</sub> 'un bir yıllık, günlük ortalama gösterimi. ....	14
Şekil 3.2: PM <sub>2,5</sub> 'in bir yıllık, günlük ortalama gösterimi. ....	14
Şekil 3.3:SO <sub>2</sub> 'nin bir yıllık, günlük ortalama gösterimi. ....	15
Şekil 3.4: NO'nun bir yıllık, günlük ortalama gösterimi. ....	15
Şekil 3.5:NO <sub>2</sub> 'nin bir yıllık, günlük ortalama gösterimi. ....	16
Şekil 3.6: NO <sub>x</sub> 'nin bir yıllık, günlük ortalama gösterimi. ....	16
Şekil 3.7: O <sub>3</sub> 'ün bir yıllık, günlük ortalama gösterimi. ....	17
Şekil 3.8: CO'nun bir yıllık, günlük ortalama gösterimi. ....	17
Şekil 3.9: Sıcaklık'ın bir yıllık, günlük ortalama gösterimi. ....	18
Şekil 3.10: Rüzgar hızı'nın bir yıllık, günlük ortalama gösterimi. ....	18
Şekil 3.11: Bağıl nem'in bir yıllık, günlük ortalama gösterimi. ....	19
Şekil 3.12: Hava Basıncı'nın bir yıllık, günlük ortalama gösterimi. ....	19
Şekil 3.13: Yerölçümsel mesafenin gösterimi (Tenenbaum, 2000). ....	28
Şekil 3.14: İki boyutlu, iki kümeli verilerin $w$ üzerine iz düşümü (Alpaydın, 2010). ....	29
Şekil 3.15: Destek vektör makineleri ile iki sınıflı verinin sınıflandırılması. ....	32
Şekil 4.1: (t+1). gün tahmin çalışmasında ozon kullanımı ve kullanılmaması farkı. ....	38
Şekil 4.2: (t+2). gün tahmin çalışmasında ozon kullanımı ve kullanılmaması farkı. ....	40
Şekil 4.3: Seçilen değişkenler ve tüm değişkenler ile yapılan tahmin sonuçları. ....	42
Şekil 4.4:NARX modele ait eğitim ve test regresyon grafikleri. ....	43
Şekil 4.5: En büyük ve en küçük $J$ değerine sahip olan gruplara ait histogramlar. ....	45
Şekil 4.6: DAA'nın ürettiği ani artış ve azalışı en iyi ve en kötü ayıran gruplara ait iz düşümler. ....	46
Şekil 4.7: Dört çekirdek fonksiyonunun en başarıları tahminleri. ....	47
Şekil 4.8: EYKK'nın bulduğu en etkili değişken grubu ile doğru tahmin edilen günler (th=20). ....	49
Şekil 4.9: Ozonun ilişkili olduğu değişken grupları ile bağıntısının görünümü (th=15). ....	52
Şekil 4.10: Ozonun ilişkili olmadığı değişken grupları ile bağıntısının görünümü (th=15). ....	52
Şekil 4.11: Ozonun ilişkili olduğu değişken grupları ile bağıntısının görünümü (th=30). ....	53
Şekil 4.12: Ozonun ilişkili olmadığı değişken grupları ile bağıntısının görünümü (th=30). ....	53
Şekil 4.13:Ozonun ilişkili olduğu değişkenlerden elde edilen katsayıların tüm veri setine uygulanması ile elde edilen ilişkinin görünümü (th=15). ....	54
Şekil 4.14: Ozonun ilişkili olmadığı değişkenlerden elde edilen katsayıların tüm veri setine uygulanması ile elde edilen ilişkinin görünümü (th=15). ....	54
Şekil 4.15: Ozonun ilişkili olduğu değişkenlerden elde edilen katsayıların tüm veri setine uygulanması ile elde edilen ilişkinin görünümü (th=30). ....	55
Şekil 4.16: Ozonun ilişkili olmadığı değişkenlerden elde edilen katsayıların tüm veri setine uygulanması ile elde edilen ilişkinin görünümü (th=30). ....	55

## TABLO LİSTESİ

Sayfa No

<b>Tablo 3.1:</b> Değişkenler ilgili istatistiki değerler.....	20
<b>Tablo 3.2:</b> İki sınıflı durum için karışıklık matrisi.....	22
<b>Tablo 4.1:</b> (t+2). gün tahmin çalışmasında değişkenlerin aldıkları katsayılar.....	39
<b>Tablo 4.2:</b> (t+2). Veri seti için ozon ile diğer değişkenler arası bağıntı katsayıları.....	40
<b>Tablo 4.3:</b> DAA'nın bulduğu ozon seviyesi ani değişimlerinin en fazla bağlı olduğu değişken grupları (th=20).....	44
<b>Tablo 4.4:</b> DAA'nın bulduğu ozon seviyesi ani değişimlerinin en fazla bağlı olduğu değişken grupları (th=30).....	45
<b>Tablo 4.5:</b> DVM'nin dört fonksiyonu kullanılarak seçilen değişkenler (th=20).....	47
<b>Tablo 4.6:</b> DVM'nin 4 fonksiyonu kullanılarak seçilen değişkenler (th=30).....	48
<b>Tablo 4.7:</b> EYKK'nın bulduğu en etkili değişken grubu (th=20).....	48
<b>Tablo 4.8:</b> EYKK'nın bulduğu en etkili değişken grubu (th=30).....	48
<b>Tablo 4.9:</b> Eşik değerlere göre ozon değişimine neden olan etmen grupları.....	50
<b>Tablo 4.10:</b> Bağıntısı en yüksek gruplar için KBA'nın ürettiği katsayılar.....	50
<b>Tablo 4.11:</b> Seçilen eşik değere göre ozon ani değişimi ile bağıntılı olmayan değişken grupları.....	51
<b>Tablo 4.12:</b> Bağıntısı en düşük değişken grupları için KBA'nın ürettiği katsayılar.....	51



## SEMBOL LİSTESİ

<b>C</b>	: Kovaryans Matris
$\rho$	: Bağntı (korelasyon) katsayısı
<b>X</b>	: Çok boyutlu deęişken
$x$	: Tek boyutlu deęişken
$w$	: İzdüşürülecek yön
<b>MAE</b>	: Ortalama mutlak hata
<b>Ppb</b>	:milyar parçada bir
<b>RMSE</b>	: Ortalama karesel hatanın karekökü
<b>BN</b>	: Bağıl Nem
<b>HB</b>	: Hava Basıncı
<b>RH</b>	: Rüzgar Hızı
<b>T</b>	: Sıcaklık

## ÖZET

### ZAMAN DİZİSİ VERİLERİNDE ANİ DEĞİŞİMLERİN TAHMİNİ İÇİN VERİ MADENCİLİĞİ YÖNTEMLERİ İLE ÖZNİTELİK SEÇİMİ

Zaman dizisi analizlerinde genel olarak bir değişken kendisinin ve diğer değişkenlerin önceki değerleri gözönüne alınarak modellenir. Bu çalışmada hem değişkenin zamana bağlı modellenmesi hem de değişken değerlerindeki ani çıkış ve düşüşlerin sınıflandırılması hedeflenmiştir. Uygulama alanı olarak da gittikçe yükselen öneme sahip olan hava kirliliği konusu ele alınmıştır. Hava kirleticilerinden biri olan ozonun tahmininin yanısıra, bu çalışmada ozondaki ani değişimlere sebep olabilecek faktörlerin belirlenmesi de hedeflenmiştir. Yüksek ozon yoğunluğunun stratosfer tabakasındaki pozitif etkisi olan dünyayı güneşten gelen morötesi ışınlarla karşı korumasının yanında troposfer tabakasında insan sağlığı ve çevre üzerinde negatif etkileri vardır. Özellikle astım hastaları ve çocukların yüksek ozon seviyesine iki-üç saat kadar bile maruz kalmaları solunum yollarında ciddi rahatsızlıklar meydana getirebilir. Bu çalışmanın amacı ozon seviyesinin ani değişimine neden olan değişken gruplarını tespit etmektir. Hangi değişken kombinasyonlarının ani değişimle ilgili olduğunun bulunması için Kanonik Bağlantı Analizi (KBA), Destek Vektör Makineleri (DVM), Doğrusal Ayırtaç Analizi (DAA) ve En Yakın K-Komşu (EYKK) teknikleri kullanıldı. Çalışmada kullanılan veri seti Ankara’da bulunan, Çevre ve Şehircilik Bakanlığı tarafından kontrol edilen bir otomatik hava kalitesi izleme istasyonu tarafından toplanmıştır. Veri seti, kirleticiler ve meteorolojik faktörlerin bir yıllık ölçümlerini ihtiva etmektedir. Ozon zaman serisi analizleri sonucunda görülmüştür ki ozonun geleceğe yönelik tahmininde NO, NO<sub>2</sub>, NO<sub>x</sub> ve rüzgar hızı değişkenleri etkili olmaktadır. Diğer yandan ozon seviyesindeki ani artışların ve düşüşlerin sınıflandırılmasında sıcaklık ve bağıl nem daha etkili değişkenlerdir. Ayrıca ozon seviyesindeki çok yüksek artış ve azalış gerçekleştiği günlerin sınıflandırılmasında ise partikül maddeler ve SO<sub>2</sub> ilişkili değişkenler olarak bulunmuştur.

## **SUMMARY**

### **FEATURE SELECTION BY DATA MINING METHODS FOR PREDICTION OF ABRUPT CHANGES IN TIME SERIES**

In time series analysis, generally, a variable is modelled by taking into consideration its previous values and previous/current values of other variables. In this study, we consider not only modelling of a variable as a function of time but also classification of abrupt changes in the values of a variable into two categories: abrupt rise and abrupt fall. As an application of the study, we have picked an air pollution dataset due to its gradually increasing importance. As a result, in this study, besides forecasting of the level of ozone as one of the air pollutants, we have also worked on determining factors that may cause abrupt changes in ozone level. Contrary to the positive effects of high level ozone concentration in stratosphere for protecting the Earth against ultraviolet radiation, in lower troposphere it has negative effects on human health and environment. Exposure to high level ozone concentration even for two-three hours can cause serious damage in respiratory systems of children and asthma patients. The goal of this study is to determine the feature groups that are related to abrupt changes in the level of ozone. Canonical Correlation Analysis (CCA), Support Vector Machines (SVM), Linear Discriminant Analysis (LDA) and k-Nearest Neighbour (KNN) techniques are used to explore which combination of features are predictive of abrupt changes in ozone level. The simulation dataset used in this study is collected in Ankara, Turkey, by an automatic air quality monitoring station operated by the ministry of environment and urban planning. The dataset consisted of one year of measurements of air pollutants and the meteorological factors. The analysis of ozone time series has shown that NO, NO<sub>2</sub>, NO<sub>x</sub> and wind speed are effective variables for forecasting the future ozone levels. On the other hand, temperature and relative humidity are more effective variables for classification of whether an abrupt rise or fall will occur in the level of ozone. Furthermore, particulate matters and SO<sub>2</sub> are found to be the most effective for rise/fall classification when considering even more abrupt changes in ozone levels.

## 1. GİRİŞ

Zaman dizisi verisi, ilgilenilen deęişkenin belirli aralıklarla elde edilmiş ölçümlerinden oluşur. Zaman dizisi analizinde genel olarak hedeflenen deęişkenin kendisinin ve verideki dięer deęişkenlerin önceki deęerleri gözönüne alınarak modellenmesidir. Deęişkenin ilgilenilen zaman diliminde dięer deęişkenler ile ilişkisi zaman dizisi analiz modeli ile ortaya çıkarılır. Ayrıca zaman dizisi analizi ile ilgilenilen deęişkende zamanla meydana gelen deęişimler ve duraęan durumların incelenmesi yapılmaktadır. Bu çalışmada deęişken deęerlerindeki ani çıkış ve düşüşlerin sınıflandırılması da hedeflenmiş olup, uygulama alanı olarak da hava kirleticilerinden biri olan ozonun tahmininin yanısıra, bu çalışmada ozondaki ani deęişimlere sebep olabilecek faktörlerin belirlenmesi de hedeflenmiştir. Ozon seviyesinde ani artış ve azalmaların olduęu günler ele alınarak, hangi deęişkenlerin artma/azalma sınıflandırmasında etkili olduęu da incelenmiştir.

Çalışmada uygulama alanı olarak kullanılan ozon verisinin seçilmesi temelde iki sebebe baęlıdır. Bunlardan birincisi bu konuda gerçek veri olarak kullanabileceğimiz kaynakların olması ve dięeri de hava kirlilięi konusunun gittikçe yükselen bir öneme sahip olmasıdır. Büyük ve gelişmiş ülkelerde başta olmak üzere yer yüzünde ve atmosferde, sanayi devriminin gelişmesinden günümüze insan saęlığını ve bitki canlılığını ters yönde etkileyen unsurlar hızla çoęalmaktadır. Çeşitli yanma reaksiyonları sonucu oluşun zararlı gazlar, fabrikalardan çevreye bırakılan katı ve sıvı atıklar bu unsurların bileşenlerini oluşturmaktadır. Katı ve sıvı atıklar topraęa karışmak suretiyle yiyeceklere, yiyeceklerden de insan vücuduna geçmekte ve insan saęlığını kötü yönde etkilemektedir. Atmosfere karışun zararlı gazlar ise solunumla vücuda alınmakta ve özellikle solunum yollarında ciddi zararlara neden olmaktadır. Özellikle büyük ve sanayi şehirlerinde oluşun sis-duman (smog) karışımı birçok zararlı küçük parçacık ve gazlardan oluşmaktadır. Bu parçacıklar ve gazlar katı yakıt kullanımı ile evlerden, fabrikalardan ve motorlu taşıt egzozlarından bulunduęumuz ortama salınmaktadır. Zamanla parçacıkların ve gazların ortamda birikip aęırlıęından dolayı

çökmesiyle şehirlerdeki sis-duman görünümünü oluşturmaktadır. Bu oluşumun temel bileşenlerinden biri de ozon gazıdır. Ozon diğer kirleticiler gibi direk olarak bir kaynaktan doğaya bırakılmamaktadır. Birincil, yani direk olarak bir kaynaktan çıkan kirleticiler güneş ışınlarından yeterli enerjiyi alıp tepkime sonucu bir tane oksijen atomu açığa çıkarır. Ortamda bulunan oksijen molekülleri yine güneş ışınlarının etkisi ile tek kalan oksijen atomu ile birleşerek ozon molekülünü oluşturur. Bu nedenle ozon ikincil gaz olarak adlandırılır. Stratosfer tabakasında bulunan ozon gazı yer yüzündeki canlıları güneşin zararlı etkilerinden korumasına karşın bulunduğumuz troposfer tabakasındaki ozon canlılar üzerinde zararlı etkiler bırakmaktadır. Özellikle astım hastaları ve solunum yolu rahatsızlıkları bulunan kimseler ile çocukların yüksek ozon seviyesine 1-2 saat maruz kalmaları solunum yollarında ciddi zararlara neden olmaktadır. Bu nedenle ozon ve birincil kirleticilerin günlük değişimleri hava kalitesi izleme istasyonları ile gözlenmektedir. Troposferik ozonun yaşadığımız ortamdaki bu zararlarından dolayı ileriye yönelik tahmini ve ortamdaki değişkenlerden ne derece etkilendiği önem arz eden bir konudur. Bu nedenle bu alanda çalışmalar sürdürülmektedir. Türkiye'nin bir çok büyük şehri için troposferik ozon tahmini çalışması yapılmıştır fakat Ankara ili için hali hazırda herhangi bir çalışma bulunmamaktadır.

Yapılan çalışmalarda yapay sinir ağları, doğrusal ve doğrusal olmayan regresyon, temel bileşen analizi, destek vektör makineleri gibi teknikler kullanılmaktadır. Kullanılan teknikler ile genel olarak veri setindeki tüm değişkenler kullanılıp geleceğe yönelik tahminler yapılmaktadır. Tahmin sonucunun kesinliği hatanın mutlak değeri veya karesel hata değerine göre tespit edilmektedir. Bu tez kapsamında da doğrusal regresyon yapıp, veri setindeki değişkenler değiştirilerek tahmin başarısında ilgili değişkenin rolü incelendi. Hata değerinin artmasına neden olan değişken ozon tahmininde kullanışsız değişken, aksi durumda ise değişken ozon tahmini için kullanışlı bir değişken olduğuna karar verildi.

Örüntü tanıma alanında değişken seçmek için kullanılan başka bir yöntem ise bağıntı katsayısıdır. Bu katsayı  $[-1,+1]$  aralığında değer alır. İlgilenilen değişken ile ozon birbiri ile pozitif yönde kuvvetli bir ilişkiliye sahip ise katsayı  $+1$ 'e yakın değer alır. Tam tersi

olarak eğer ilgilenilen deęişken ile ozon arasında negatif bir iliřki var ise baęintı katsayısı -1'e yakın deęer alır. Üçüncü bir durum olarak, katsayının 0'a yakın deęer alması ise ozon ile ilgilenilen deęişken arasında bir iliřki bulunmadığını gösterir.

Ozonun etkilenme derecesini belirlemek için ardışık iki gün arasında ki ozon seviyesi farkının belli bir eşik deęerden yüksek olduęu günler hesaba katıldı. Bu günler için tüm deęişkenlerin üçerli tüm alt kümeleri kullanılan algoritma ile iz düşürölüp yeni koordinatlar elde edildi. Daha sonra elde edilen izdüşümlerin deęişimin olduęu günlerdeki ozon miktarı ile baęıntıları hesaplandı. Baęıntısı yüksek grupların içerdiği deęişkenler ozon deęişime neden olan deęişken gruplarıdır. İncelenen çalışmalarda deęişkenlerin teker teker ele alındığı görüldü. Bu tez kapsamında kullanılan yöntem ile ozonun bir deęişken grubunun tümünden nasıl etkilendięi incelendi. Ayrıca gruptaki deęişkenlerin etkinlięi elde edilen katsayılar ile belirlenebilmektedir. Bu yönü ile bu tez çalışması bu alanda farklılık göstermektedir. Ozon ile ilgili çalışmalarda kanonik baęintı analizi ve isomap algoritmalarına rastlanılmaması ise bir dięer çeşitliliştir.

Tez içerięi řu şekildedir: Bölüm 2'de ozon gazının insan ve çevre üzerindeki zararlı etkileri ve geleceęe yönelik ozon ile ilgili yapılan çalışmalar yer almaktadır. Bölüm 3'te ise kullanılan veri setleri tanıtılmakta, tezde kullanılan makine öğrenmesi yöntemlerinin teorik bilgileri ve izlenen yöntemler bulunmaktadır. Bölüm 4'te de elde edilen sonuçlar tablo ve şekiller ile gösterilmektedir. Bölüm 5'te ise sonuçlara yönelik tartışma ve çıkarım yapılmaktadır.

## **2. GENEL KISIMLAR**

Ozonun keşfinden günümüze kadar ozon ile ilgili çeşitli çalışmalar yapılmıştır. Bu çalışmalar kapsamında ozonun insan sağlığı ve çevre üzerinde etkileri incelenmiştir. Bu amaçla dünyanın değişik yerlerinde ozon ve ozonun etkilendiği düşünülen etmenler düzenli olarak ölçülüp kayıt altına alınmaktadır. Bilgisayar teknolojisinin gelişmesiyle bu işlem için otomatik veri istasyonları kurulup veriler belirli merkezlerde toplanmaya başlandı. Elde edilen bu verilerden çeşitli makine öğrenmesi teknikleri yardımıyla gelecekteki zaman dilimlerinde ozon seviyesinin tahminleri yapılmaktadır. Ayrıca ozon seviyesinin değişimine neden olan etmenlerin etkileri nicellenebilmektedir. Bu bölümde ozonun tanımı ve daha önceki çalışmalara yer verilmiştir.

### **2.1 OZON**

Ozon, güneşten gelen morötesi ışınların oksijen atomları ile girdikleri fotokimyasal tepkime sonucu oluşmaktadır. Açık mavi bir renge sahiptir. 1840 yılında C.F.Schönbein tarafından keşfedilerek ozon olarak adlandırılmıştır. Daha sonra bilim insanları tarafından yapılan çalışmalarla formülü ve yapısı açıklanmıştır (Tozsın, 2003). Ozon atmosferin stratosfer ve troposfer tabakalarında bulunmaktadır.

#### **2.1.1.Stratosferik Ozon**

Stratosfer tabakası yaklaşık olarak ozonun % 90'ını ihtiva etmektedir. Atmosferin bu katmanında ozon, güneşten gelen morötesi ışınların büyük bir bölümünü keserek yeryüzündeki canlı yaşamını etkileyecek zararlarını önlemektedir. Bundan dolayı stratosfer tabakasındaki ozon, iyi huylu ozon olarak adlandırılmaktadır (Meteoroloji).

#### **2.1.2.Troposferik Ozon**

Troposfer tabakası toplam ozonun % 10'unu ihtiva etmektedir. Bu katmanda ki ozon birikimi ya stratosferden rüzgarlar yardımıyla aşağı inmekte ya da kirleticilerin stratosferden sızan morötesi ışınlarla tepkimesi sonucu oluşmaktadır. Yaşadığımız çevredeki bu ozon, solunum yolları rahatsızlıklarına ve ürün kayıplarına yol açmaktadır. Bu nedenle troposfer ozonu kötü huylu ya da zararlı ozon olarak adlandırılmaktadır (Meteoroloji).

## 2.2 MAKİNE ÖĞRENMESİ METODLARININ OZON TAHMİNİNDE KULLANILMASI

Ozon ile ilgili yapılan çalışmalar yukarıda da bahsedildiği gibi genel olarak iki bölüme ayrılabilir. Kolaylık olması açısından tezin bu kısmından itibaren troposferik ya da kirletici ozon için sadece ozon tabiri kullanılacaktır. Hava kalitesi verilerinin analizi için çeşitli makine öğrenmesi teknikleri kullanılmaktadır. Bu bölümde bu alanda yapılmış çeşitli çalışmaların incelemeleri anlatılacaktır.

Ozonla ilgili çalışmalar ölçümlerin yapıldığı yerlere çeşitlilik arz etmektedir. Ölçüm istasyonlarının sanayi bölgelerine yakın yerlere, kalabalık nüfusun yaşadığı çok şehirleşmiş bölgelere ve nüfusun seyrek olduğu kırsal bölgelere kurulması bu çeşitliliği sağlamaktadır. Böylece ozon konsantrasyonunun haritaları çıkarılmakta ve nelerin ozon seviyesine etki ettiği belirlenmeye çalışılmaktadır. Bu amaçla dünya genelinde hem ulusal hemde uluslararası örgütlerin çalışmalarıyla ozon değişimleri gözlem altına alınmaktadır. Türkiyede de bu doğrultuda hem belediyeler aracılığıyla hem de meteoroloji genel müdürlüğünün çalışmalarıyla çok sayıda hava kalitesi gözlem istasyonu kurulmuştur. Yapılan ölçümler belediyelerin ve çevre bakanlığının internet sitelerinde halkla paylaşılmaktadır (Hava kalitesi izleme, 2013). Bu verilerin çeşitli solunum rahatsızlıklarına sahip kişilere daha faydalı olması için ozonun gelecek gün veya saatlerde ki yoğunluğu bu veriler üzerinden tahmin edilmelidir. Bu tahminler için çeşitli algoritmalar geliştirilmiştir ve geliştirilmeye devam edilmektedir. Bu algoritmalarından bazıları şunlardır: yapay sinir ağları, hücresel sinir ağları, doğrusal ve doğrusal olmayan regresyon, temel bileşen çözümlemesi, destek vektör makineleri, karar ağaçları.

Ozon seviyesinde meydana gelen artışların özellikle sanayi devrimi ile birlikte meydana gelen hava kirliliğine bağlı olarak dünya genelinde bir artış olduğu görülmektedir. Bu alanda son 30 yılı kapsayan bir araştırmada ozonun bu süreçte ki davranışı ve uzun vadede ozon kirliliğini azaltmaya yönelik yapılması gerekenler üzerinde durulmuştur (Cape, 2008). Araştırmada ozonun tahmini üzerinde durulmuştur. Ozonun birincillerinden olan NO<sub>x</sub>'in ve taşınabilir organik bileşiklerin (TOB) kaynakları çok



çeşitli olabilir. Özellikle NO<sub>x</sub> yanma olan neredeyse tüm reaksiyonların sonucunda üretilmektedir. Bu nedenle ozon seviyesi tahmininin karmaşıklığı artmaktadır. Ozonun seviyesini yüksek olduğu ilkbahar ve yaz mevsimleri bitkilerin büyüme dönemlerini de kapsamaktadır. Bu nedenle bitkilerin gelişmeleri ozon seviyesi ile ters ilişkilendirilmektedir. Ozona maruz kalmak iki farklı şekilde olmaktadır: ilki aşırı yüksek ozon seviyesine maruz kalmak. İkincisi ise belirli bir eşik değer seviyesindeki ozona maruz kalmaktır. Bitkiler için bu eşik değer her milyarda bir bölüm için 40 olarak ifade edilmektedir. İki şekilde de ozona maruz kalmak bitkiler için zararlıdır fakat hangisinin daha zararlı olduğu tespit edilememiştir. Araştırmada uzun vadede ozona neden olan kirleticilerin atmosferdeki yoğunluğunun artacağı ve iklim değişikliklerinin olacağı öngörülmüş. Sıcaklığın ozon konsantrasyonunu ve aynı zamanda havadaki su buharını da arttırmaktadır. Artan su buharı ozon seviyesinin azalmasını sağlamaktadır. Sonuç olarak uzun vadede özellikle kuzey yarım kürede kıtalar arası parçacıklarında taşınmasının etkisi ile ozon seviyesinde artış beklenmekte olduğu belirtilmiştir.

Yapay sinir ağları tahmin(regresyon) yapmak için en çok kullanılan yöntemlerin başında gelmektedir. İstanbul verileri kullanılarak yapılan bir çalışmada da (İnal, 2010) yapay sinir ağları kullanılarak ozon tahmini yapılmıştır. Giriş katmanında 9 tane hava kirletici ve 9 tane meteorolojik faktör, gizli katmanda 24 tane nöron kullanıp çıkış olarak bir sonraki günün en yüksek ve ortalama ozon seviyelerini tahmin etmiştir. Verisinde bulunan toplam 578 örnek öğrenme, çarpaz-geçerleme ve test (train, cross-validation, test) olarak üçe bölünmüş ve en iyi sonuçlar tüm değişkenlerin kullanılması ile elde edilmiştir. Sonuçların değerlendirilmesi için bağıntı katsayısı, ortalama mutlak hata (MAE), ortalama karesel hata (RMSE), ve Willmott indeksi kullanılmıştır ve değerleri sırasıyla şöyledir: 0.90, 8.78 µg/m<sup>3</sup>, 11.15 µg/m<sup>3</sup> ve 0.95.NO, ortalama sıcaklık, PM10, en yüksek sıcaklık, güneşlenme süresi, rüzgar yönü, solar radyasyon değişkenleri bahsedilen değerleri çok değiştirmedığı için en önemli değişkenler olarak seçilmiştir. Yapay sinir ağı modeli ile doğrusal ve doğrusal olmayan regresyon karşılaştırılmış ve çok büyük farklılık olmadığı tespit edilmiştir.

İstanbul ili için, yapay sinir ağları, sarmalı (wrapper) metod kullanılarak yapılan bir diğer çalışmada öznitelik seçimi yapılmış (Şakar, 2011). Tüm değişkenlerden teker teker değişken azaltarak (backward selection) her değişkenin tek tek yapay sinir ağları ile oluşturulan tahmin modeli üzerinde ki etkisi incelenmiştir. Açık hava sıcaklığı ve solar ışıltama en etkili meteorolojik faktörler olurken, toplam hidrokarbonlar (THC), NO<sub>2</sub> ve NO en etkili kirleticiler olmuştur. Bu şekilde bir değişken seçimi ile hem geçerlilik hem de test kümelerinde en başarılı alt kümedeki değişkenler şunlar seçilmiştir: sıcaklık, solar ışıltama, NO<sub>2</sub>, THC, NO ve SO<sub>2</sub>.

Zaman parametresi kullanılarak yapılan bir başka çalışmada yine yapay sinir ağları ile modellenmiş (Demir, 2008). İstanbul ili için yapılan bu çalışmada 1, 3 ve 7 günlük zaman periyotlarından hangisinin tahmin performansının daha iyi olduğu araştırılmış. Yapay sinir ağları 6 tane kirletici ve 8 tane meteorolojik faktör ile modellenmiş. En iyi sonuçlar 7 günlük zaman periyotları kullanıldığında elde edilmiş, 3 günlük periyotlar kullanılan model ise 1 günlük periyot kullanılan modelden daha iyi sonuçlar üretmiş. En yüksek bağıntı katsayısı 0.8903 olarak 7 günlük periyot için 2 gizli katman kullanılarak elde edilmiş.

İspanyanın Sevilya kenti yüzey seviyesi ozonu için yapılan bir çalışmada yapay sinir ağlarının yanlılık hassaslığı incelenmiştir (Pastor-Bárcenas, 2005). Bu çalışmada kirletici olarak NO ve NO<sub>2</sub>, meteorolojik faktör olarakta rüzgar hızı, sıcaklık, atmosferik basınç, güneş radyasyonu miktarı ve bağıl nem kullanılmıştır. Oluşturulan modelde kırsal bir hava izleme monitöründen elde edilen 30 günlük bir veri setinin 20 günü eğitim için 10 günü ise geçerleme için kullanılmıştır. Bulguların testi için çarpaz-geçerleme kullanılmıştır. Bu çalışmada iki model üzerinde çalışılmıştır: hassaslık analizi ve budama. Birincisi model olan hassaslık analizi için kullanılan iki teknik şunlardır: delta kuralı ve ortalama mutlak eğim (average absolute gradient). Bu modelde en ilgisiz girişler tekrar hesaba katılmadan eğitim yapılmıştır. Şu üç sonuç elde edilmiş: (i) girişlerin önemlerine göre sıralanması, (ii) metodun kolaylığından emin olundu, (iii) tarafsız bir sıralamanın yapılması. İkinci modelde ilişkisiz değişkenler ya buluşsal(heuristic) olarak ya da belirlenen bir eşik değerle elemine edildi. Performans

değerlendirmesi için pearson korelasyon katsayısı (r), deteminasyon katsayısı (R<sup>2</sup>), ortalama mutlak hata ve karesel hatanın ortalamasının karekökü kullanılmıştır. Sonuç olarak ozon ile en ilişkisiz değişkenler NO ve NO<sub>2</sub> bulunmuş öte yandan en çok ilişkili değişkenler ise sıcaklık ve güneş radyasyonu miktarı olarak bulunmuştur. Bu bulgular doğrultusunda bu kırsal kesimde ozon üretilmediği ve rüzgarlar ile buraya taşındığı sonucuna varılmıştır.

Doğrusal olmayan modelleme kabiliyetine sahip yapay sinir ağlarının en çok kullanılan biçimlerinden çok katmanlı algılayıcı (multilayer perceptron) Valensiya şehrinin bir kasabası için ozon tahmininde kullanılmıştır (Gómez-Sanchis, 2006). Bu çalışmada kasabada ki gözlem istasyonundan elde edilen 1997, 1999 ve 2000 yıllarının nisan ayına ait verileri kullanılmıştır. Her yıl için ayrı bir model geliştirilmiştir ve her yıla ait verinin ilk 20 günü öğrenme kümesi olarak son 10 günü ise geçerleme kümesi olarak kullanılmıştır. Veri seti 3 kirletici ve 5 meteorolojik faktörden müteşekkildir ve bunlar sırasıyla: (O<sub>3</sub>, NO, NO<sub>2</sub>) ve (sıcaklık, rüzgar hızı, bağıl nem, güneş radyasyonu, basınç) tir. Modellenen yapay sinir ağının ürettiği tahmin sonuçları değişik hata katsayıları ile değerlendirilmiştir. Bunlarda ortalama karesel hata katsayısı (öğrenme, geçerleme) formatında her yıl için sırasıyla (11.29, 9.04), (8.01, 9.03), (8.14, 7.31) olarak bulunmuştur. Değişkenlerin ilişkileri bakımından 1997 yılında NO ve NO<sub>2</sub> etkisizken 1999 ve 2000 yıllarında NO<sub>2</sub>'nin ozon üzerinde ki etkileri fazladır. Bundan dolayı 1997 yılındaki ozon oluşumunun daha çok taşınım ile gerçekleştiği görülmektedir.

İspanyanın şehirsal ve kırsal yerleşim yerlerinde ozonun tahmininin stokastik bir yaklaşımla modellenmesi için hareketli-ortalama entegreli oto-regresyon (ARIMA; AutoRegressive Integrated Moving Average) kullanılmıştır (Dueñas, 2005). Şehirsal bölgelerde (1,0,0)\*(0,1,1)<sub>24</sub> kırsal bölgelerde ise (2,1,1)\*(0,1,1)<sub>24</sub> ARIMA model yapıları kullanılmıştır. Araştırmada ilgilenilen bölgelere ait verilerin sadece yaz aylarına ait ozon verileri kullanılmıştır. Oluşturulan tahmin modelinde sadece tek bir değişken kullanılmıştır. Yapılan tahminlerde kırsal bölge verileri için alınan sonuçlar şehirsal bölge verilerinden daha iyi sonuçlar vermiştir.

Bir başka çalışmada ise trafiğin yoğun olduğu bir bölge ile kırsal bir bölgeden alınan veri setleri kullanılarak zaman serileri, çoklu doğrusal regresyon ve ileri beslemeli yapay sinir ağlarının performansları karşılaştırılmıştır (Sousa, 2006). Yapılan çalışmalar Oporto şehrinin hava kalitesi gözlem istasyonlarının 2002 yılına ait olan veri dört parçaya bölünerek gerçekleştirilmiştir. Veri setinde şu değişkenlerin günlük ortalamaları bulunmaktadır: O<sub>3</sub>, NO, NO<sub>2</sub>, PM<sub>10</sub>, SO<sub>2</sub>, CO, sıcaklık, bağıl nem, rüzgar hızı. Trafiğin yoğun olduğu bölgeye ait veri setinde daha başarılı sonuçlar elde edilmiştir. Zaman serilerine göre yapılan çalışmada ozondan hariç değişkenler kullanılmamıştır. Bu çalışmada en kötü tahmin sonucunu zaman serileri vermiştir. En iyi tahmin sonucunu ise ileri beslemeli yapay sinir ağları vermiştir.

Kuveyt'te yapılan bir çalışmada trafiğin yoğun olduğu bir bölgeden seyyar bir otomatik ölçüm istasyonu ile ölçümler gerçekleştirilmiştir (Abdul-Wahab, 2002). Ölçümü yapılan parametreler şunlardır: CH<sub>4</sub>, NMHC, CO, CO<sub>2</sub>, NO, NO<sub>2</sub>, SO<sub>2</sub>, O<sub>3</sub>, rüzgar hızı, sıcaklık, bağıl nem, güneş enerjisi, şüpheli toz. Bu çalışmada da yapay sinir ağları kullanılmıştır ve üç model geliştirilmiştir. Birincisi 24 satlik ölçüm sonuçları ile tahmin yapıyor ve 4797 örneği var. İkinci model sadece güneşin olduğu saatlerdeki ölçümleri kullanıyor ve 1630 örneğe sahiptir. Üçüncü modelde ise yalnızca ozonun 80 ppb değerinden büyük olduğu ölçümler kullanılmıştır ve bunların sayısı da 19'dur. Bu modellerin test kümelerinde sırasıyla 719, 244, 3 'er tane örnek bulunmaktadır. Modellerin başarısını değerlendirmek için ortalama karesel hataya bakılırsa sırasıyla 6.619, 14.030, 20.998 hata değerleri elde edilmiştir. Ortalama mutlak hata ise yine sırasıyla 1.819, 2.66, 2.049 olmuştur.

Kuveypte yapılan başka bir çalışmada yapay sinir ağlarını temel bileşen analizi ile beraber modelleyerek sadece yapay sinir ağları ile yapılan regresyonun uygunluğu arttırmak istenmiştir (Al-Alawi, 2008). Bu çalışmada da ozonun tahmininde yukarıda ki kirletici etkenler kullanılmıştır. Çalışmada kullanılan meteorolojik faktörler ise şunlardır: rüzgar hızı, rüzgar yönü, sıcaklık, bağıl nem, güneş radyasyonu. Veri seti bir hafta boyunca her gün ver her saat yapılan ölçümlerden oluşmaktadır. Çalışmada temel bileşen ile yapılan doğrusal regresyon sonucunda kalan doğrusal olmayan kısım YSA

ile tahmin edilerek birleştirilmiştir. Temel bileşen regresyon, yapay sinir ağları ve birleşik model ile yapılan yapılan ayrı ayrı regresyonlar sonucunda elde edilen  $R^2$  değerleri sırasıyla şunlardır: 0.965, 0.986, 0.995. Yapılan tahminler sonucu oluşan karesel hatalar ise sırasıyla 9.74, 3.51, 3.81 olarak bulunmuştur. Çalışmada ayrıca Wilcoxon sıra testi uygulanmış ve YSA ile birleştirilmiş model arasında istatistik bir farklılık olmadığı görülmüş fakat temel bileşen regresyon bunlardan farklılık göstermiştir.

Atina'daki dört farklı rüzgar özellikleri olan bölgelerde kurulmuş olan hava kalitesi gözlem istasyonlarından alınan veriler üzerinde yapılan çalışmada sonraki günün en yüksek ozon değerinin tahmini için yapay sinir ağları ile çoklu doğrusal regresyon modellerinin performansları karşılaştırılmıştır (Chaloulakou, 2003). Her model için 1992-1999 yılları arasında nisan ayından ekim ayına kadar ölçümleri yapılmış 8'i meteorolojik 3'ü kirletici olmak üzere toplam 11 değişken kullanılmıştır. Yapay sinir ağları modeli diğer doğrusal regresyon modellerine göre en yüksek ozon seviyesini tahmin etmede daha başarılı olmuştur. Yapay sinir ağları ile oluşturulan model ile Avrupa genelinde eşik değer kabul edilen  $180 \mu\text{g}/\text{m}^3$ 'ü aşan değerler % 72'lik bir başarı ile tahmin edilmiştir. Yapay sinir ağları ile test sette yapılan en küçük karesel hata 18.4 olurken doğrusal regresyon modelinde bu değer 20.0 olmuştur. Ayrıca eğitim kümesi ile test kümesi arasında da yapılan hata değerinde belirli bir fark vardır. Bu çalışmada yapay sinir ağlarının birbirinden farklı bölgelerden alınmış verilerde birbirine yakın hata değerleri vermesi modelin genellenebilir olduğunu göstermektedir.

Ozona maruz kalmaktan en çok zarar gören çocuklara yönelik yapılan Leipzig'deki bir çalışmada, çocukların solunum rahatsızlıkları oldukları günler ile yüksek ozon konsantrasyonu olan günler ilişkisi incelenmiştir (Schlink, 2006). Yaklaşık iki yıl süren çalışmaya yaşları 5-7 arasında 279 çocuk katılmıştır. Çalışma boyunca hergün solunum rahatsızlıkları ile ilgili 25 semptomun bulunduğu bir anket doldurulmuştur. Semptomu görüldüğü her çocuk '+' işareti ile göstermeyenler ise '.' işareti ile ankette işaretlenmiştir. Veri setinde kayıp değerde bulunmaktadır. Meteorolojik etmenler ve kirleticiler ise şunlardır: bağıl nem, buhar basıncı, günlük toplam radyasyon, günlük en yüksek sıcaklık ve günlük ortalama  $\text{NO}_2$ ,  $\text{SO}_2$ ,  $\text{O}_3$ . Çalışma sonucunda  $\text{SO}_2$ 'nin

solunum rahatsızlıkları ile doğrusal olarak ilişkili olduğu görülmüştür. Ozonun sadece 60  $\mu\text{g}/\text{m}^3$ 'ten daha yüksek olduğu seviyelerde çocuklar üzerinde solunum rahatsızlıklarına neden olduğu sonucuna varılmıştır. Çalışmada ayrıca Avrupanın 4 farklı ülkesinin 10 şehrinden toplanan veri setleri kullanılarak geleceğe yönelik ozon tahminleri yapılmıştır. Şehir merkezlerinde, kenar mahallelerde, kırsal bölgelerde ve endüstriyel bölgelerde kurulmuş istasyonlardan toplanan bu ikinci veri seti şu değişkenlerden oluşmaktadır: O<sub>3</sub>, NO, NO<sub>2</sub>, sıcaklık, bağıl nem, rüzgar hızı, rüzgar yönü ve güneş radyasyonu.

Ozon seviyesi tahmini kendi değerleri üzerinden yapıldığı gibi gerçek değerleri farklı uzaylara taşıyıp yapılan tahmin çalışmaları da vardır. Wavelet dönüşümü bu amaçla kullanılan dönüşümlerden biridir. Teksas ta bulunan 14 ozon gözlem istasyonundan elde edilen 2003-2006 yıllarını kapsayan ozon ve şu altı meteorolojik faktörün (en yüksek sıcaklık, en yüksek güneş radyasyonu, ortalama ani rüzgarlar, ortalama bileşke rüzgar yönü, ortalama bileşke rüzgar hızı, ortalama rüzgar hızı) günlük değerlerinin oluşturduğu veri seti üzerinde wavelet dönüşümü uygulanmıştır (Temiyaşatit, 2009). Çalışmada öncelikli olarak adım adım regresyon (stepwise regression) kullanılarak değişken seçimi yapılmıştır. Yapılan seçim işleminde kriter olarak Akaike bilgi kriteri kullanılarak şu dört değişken seçilmiştir: en yüksek sıcaklık, en yüksek güneş radyasyonu, ortalama bileşke rüzgar yönü, ortalama rüzgar hızı. Daha sonra seçilen değişkenler wavelet dönüşüm ile wavelet uzaya taşınmıştır. Bu uzayda tahmin için kriging metodu kullanılmıştır. Tahmin edilen değerler tekrar gerçek değerlerine yani zaman domenine ters wavelet dönüşüm uygulanarak taşınmıştır. Modelin değerlendirilmesinde ozon ölçüm değerlerinin 75'i aştığı günler ile 75'in altında kaldığı günler ayrı ayrı ele alınmış ve çarpaz-geçerleme ortalama mutlak hata (ÇG-MAE) kullanılmıştır. Bu eşik değerinin altında ki günler için hata 9.17 olarak eşik değerinin üstü için ise 21.60 olarak bulunmuştur. Wavelet dönüşümlerine inceltme (refine) uygulanarak bu hatalar sırasıyla 3.95, 8.48 değerlerine küçültülmüştür.

Ozonun karmaşık oluşum ve yok edilmesini modelleyebilen bir diğer algoritma ise genetik algoritmadır (Pires, 2011). Bu çalışmada Pires ve arkadaşları Portekiz'in büyük şehirlerinden Oporto şehrine ait 2004 yılının mayıs-temmuz aylarının ölçümlerini içeren

bir veri seti kullanmıştır. Bu veri seti CO, NO, NO<sub>2</sub>, O<sub>3</sub> kirleticilerinin ve sıcaklık, güneş radyasyonu, bağıl nem, rüzgar hızı meteorolojik değişkenlerinin saatlik ortalamalarını ihtiva etmektedir. Hem veri setinin gerçek değerlerini hem de temel bileşenlerin (principal component) kullanıldığı çalışmada veri seti eğitim kümesi ve test kümesi olarak ikiye bölünerek kullanılmıştır. Genetik algoritma ile önce değişken seçimi yapılmış ve sıcaklık, bağıl nem, O<sub>3</sub> değişkenleri seçilmiştir. Temel bileşenler kullanılırken bunlara ek olarak NO<sub>2</sub> de seçilmiştir. Bu çalışmada da bu değişkenlerin kullanılarak bir gün sonranın ortalama ozon seviyesi tahmini yapılmıştır. Gerçek değerler kullanılarak yapılan tahmin temel bileşenler kullanılarak yapılan tahminden eğitim kümesinde daha iyi sonuç verirken test kümesinde daha kötü sonuç vermiştir.

Ozonun doğrusal olmayan karmaşık yapısını modellemeye uygun bir diğer makine öğrenmesi aracı da destek vektör makineleridir. İspanyanın Aviles kasabasında yapılan bir çalışmada ozonun birincil kirleticiler ile olan ilişkisi destek vektör makineleri (DVM) ile regresyon yapılarak gösterilmiştir (Suárez-Sánchez, 2011). Bu çalışmada birinci kirleticiler ve ozonun 2006-2008 yıllarına ait aylık ortalamaları kullanılmıştır. DVM'nin değişik çekirdek fonksiyonları uygulanarak ozona en iyi uyacak fonksiyon aranmıştır. Ayrıca çok katmanlı algılayıcı kullanılarak bu iki metodun performansları karşılaştırılmıştır. DVM'nin doğrusal çekirdekli yapısı kullanıldığında ozonun en bağımlı olduğu değişkenlerin NO ve NO<sub>2</sub> olduğu görülmüştür. Çekirdek fonksiyonu olarak karesel çekirdek (quadratic kernel) seçildiğinde DVM'nin tüm değişkenler için doğrusal çekirdekte daha yüksek katsayılar üretmiştir, yani ozona daha iyi uyum sonuç üretmiştir. Radyal temelli çekirdek (radial basis kernel) fonksiyonu doğrusal çekirdeğe benzer sonuçlar üretmiştir. Pearson evrensel çekirdeği (Pearson Universal Kernel) kullanıldığında ozona en iyi uyum regresyon sonucu elde edilmiştir. İkinci olarak çok katmanlı algılayıcı en iyi uygunluğu göstermiştir.

### 3. MALZEME VE YÖNTEM

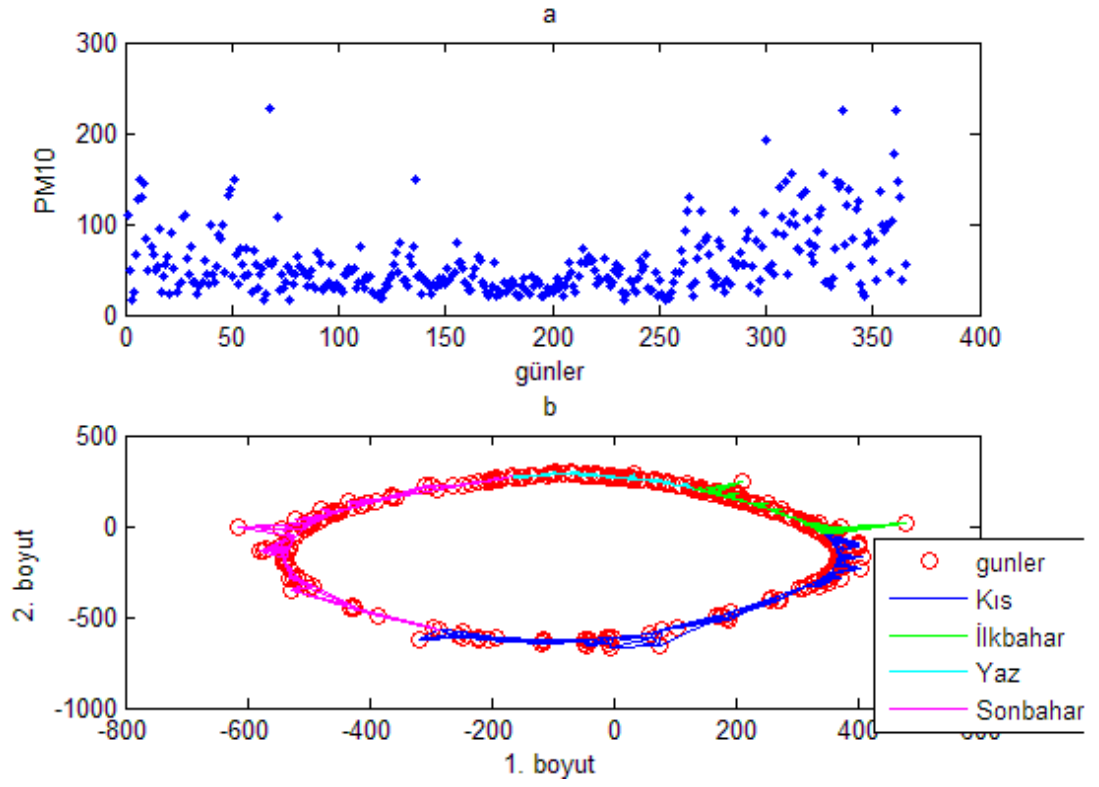
Bu bölümde öncelikle kullanılan zaman dizisi verisi tanıtılacak ve makine öğrenmesi yöntemlerinde kullanılan temel istatistik kavramlara yer verilecektir. Daha sonra tezde kullanılan makine öğrenmesi tekniklerinden kanonik bağıntı analizi'nin, isomap'in ve doğrusal ayrıştırma analizi'nin yapıları teorik olarak anlatılacaktır. Son kısımda ise izlenen yöntem gösterilecektir.

#### 3.1. OZON VERİ KÜMESİ ZAMAN DİZİSİ VERİSİ

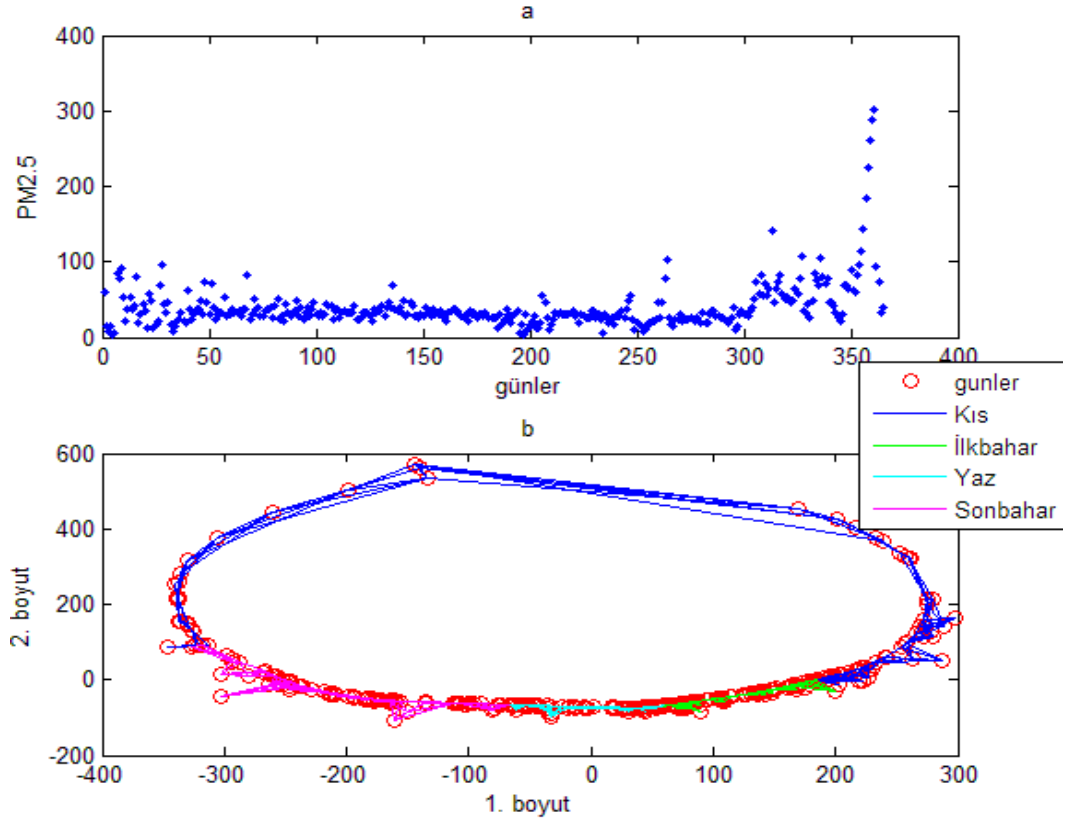
Bu tez çalışmasında Çevre ve Şehircilik Bakanlığının Ankara ili Cebeci semtinde kurmuş olduğu hava kalitesi izleme istasyonundan elde edilen 2010 yılına ait veri seti kullanılmıştır. Bu istasyonda ölçülen parametreler iki kısımdır. Birinci kısım kirleticilerden oluşmaktadır ve şunlardır: PM10, PM2,5, SO2, NO, NO2, NOx, CO, O3. İkinci kısımda meteorolojik faktörler bulunmaktadır ve şöyle sıralanmaktadır: sıcaklık, rüzgar hızı, bağıl nem, hava basıncı. Hava kalitesini etkileyen bu faktörlerin hepsi otomatize istasyon tarafından her saat ölçülmekte ve kaydedilmektedir. Böylece hava kalitesini etkileyen unsurların bir arada olduğu bir zaman dizisi verisi oluşmaktadır. Bu zaman dizisi verisi analiz edilerek hava kalitesi ölçeklenebilir ve bu parametrelerin birbirleri ile etkileşimleri ve ilişkileri tespit edilebilir. Bu tez kapsamında doğada kendiliğinden var olmayan yani başka kirleticilerin güneş ışınları ile etkileşimleri sonucu oluşan ozon (O3) gazının diğer etkenlerle ilişkisi üzerine çalışmalar yapılmıştır.

Şekiller 3.1-3.12 ilgili parametrelerin (a) günlük ortalamalarının bir yıllık gösterimi ve (b) ise bir yılın ilk günü ile son günü birbirine komşu günler olduğu göz önünde bulundurularak, komşu günlerin komşuluğunun korunarak elde edilen iki boyutlu ISOMAP iz düşümlerinin gösterimidir.

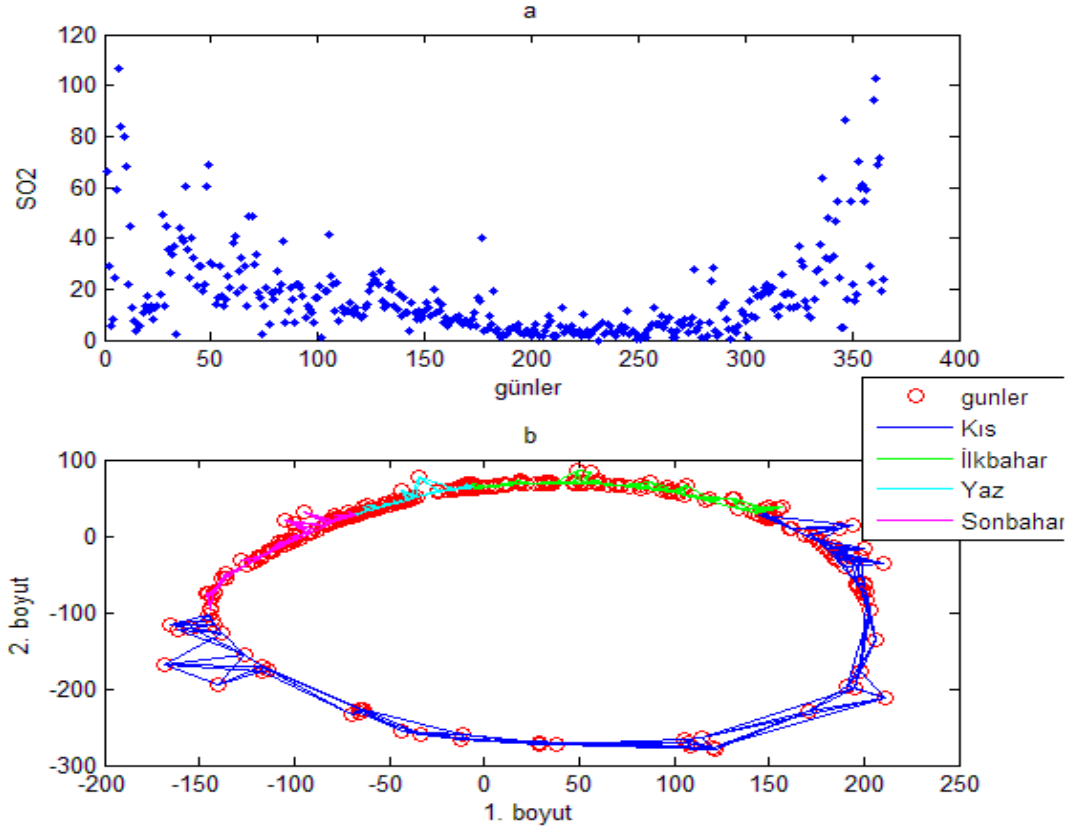




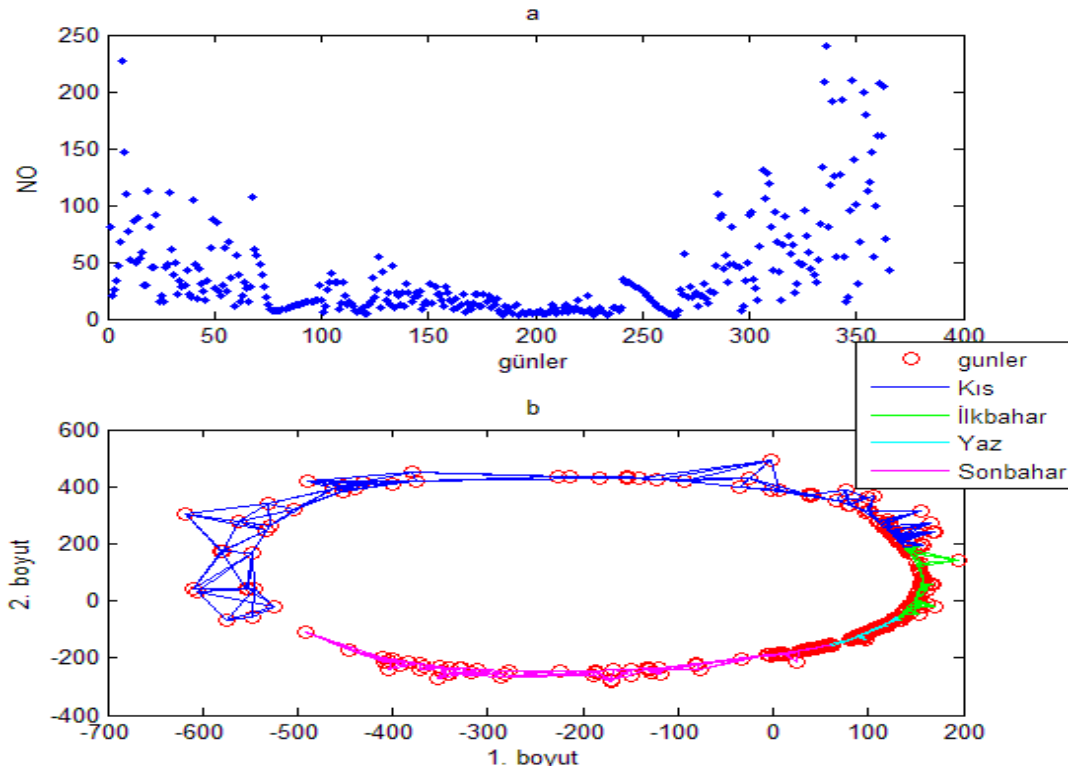
Şekil 3.1: PM<sub>10</sub>'un bir yıllık, günlük ortalama gösterimi.



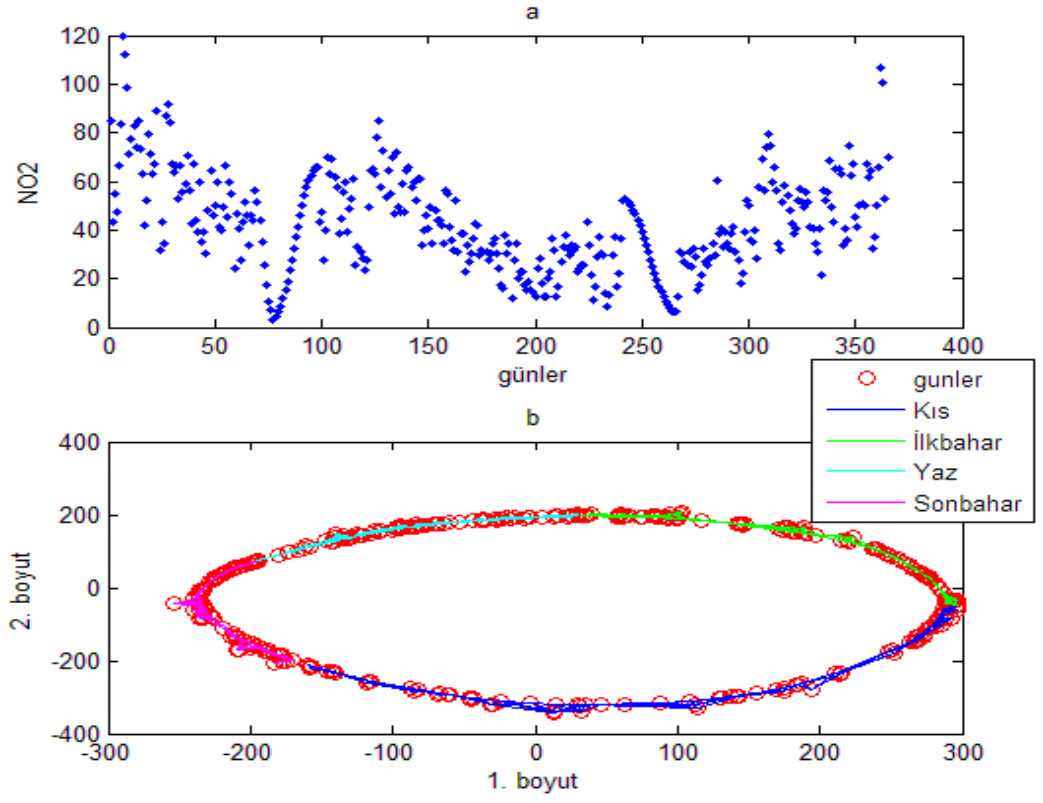
Şekil 3.2: PM<sub>2.5</sub>'in bir yıllık, günlük ortalama gösterimi.



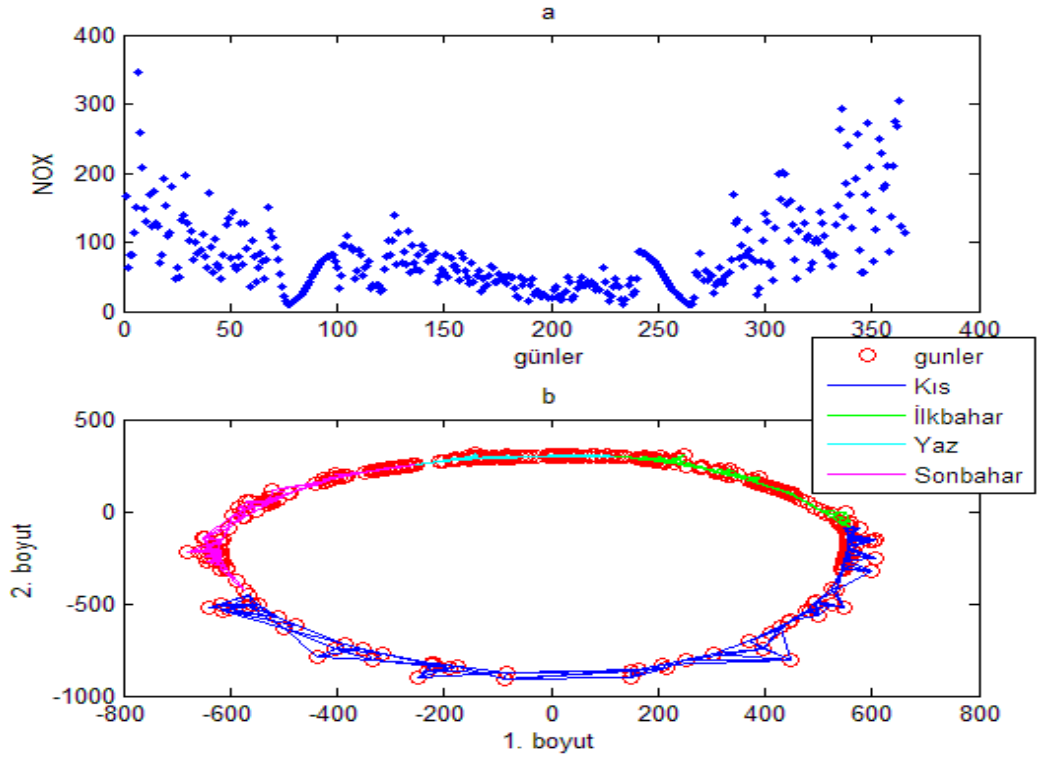
Şekil 3.3:SO<sub>2</sub>'nin bir yıllık, günlük ortalama gösterimi.



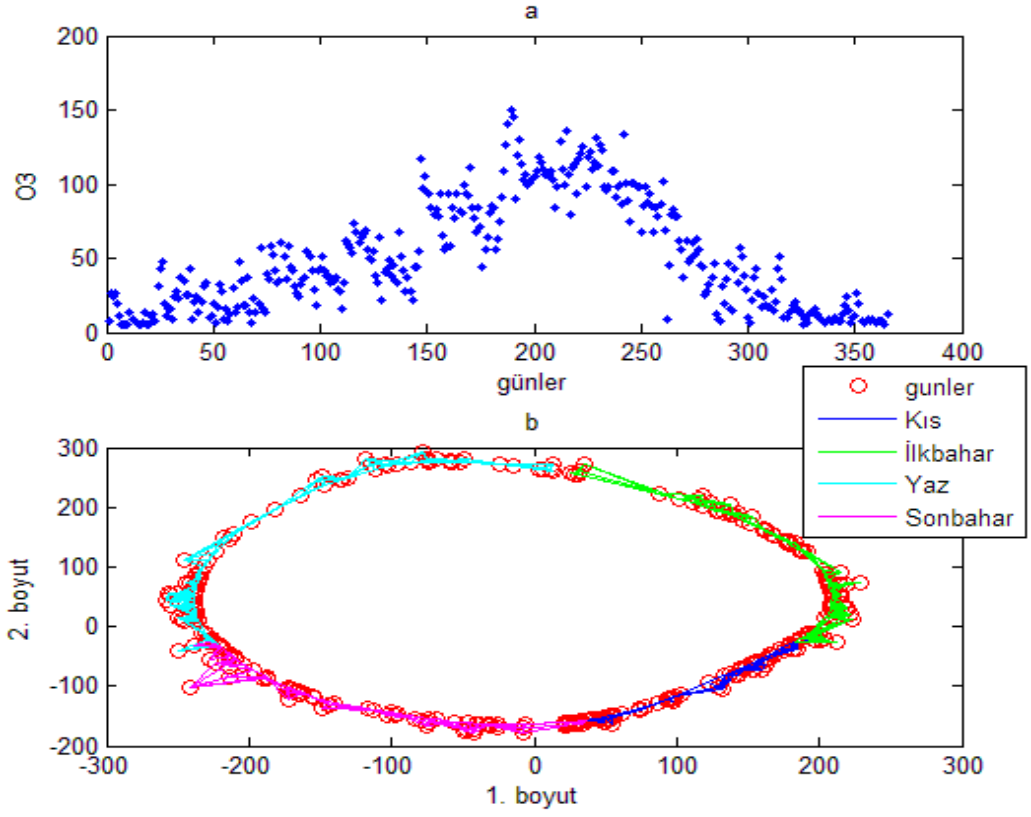
Şekil 3.4: NO'nun bir yıllık, günlük ortalama gösterimi.



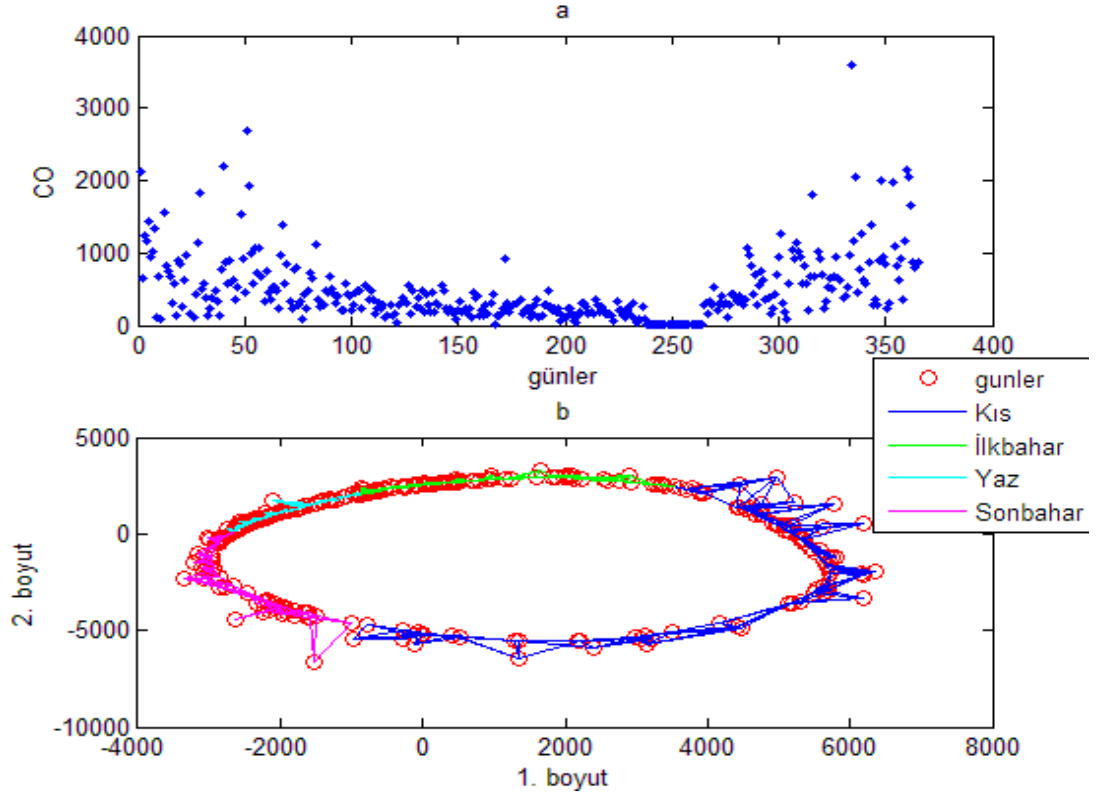
Şekil 3.5: NO<sub>2</sub>'nin bir yıllık, günlük ortalama gösterimi.



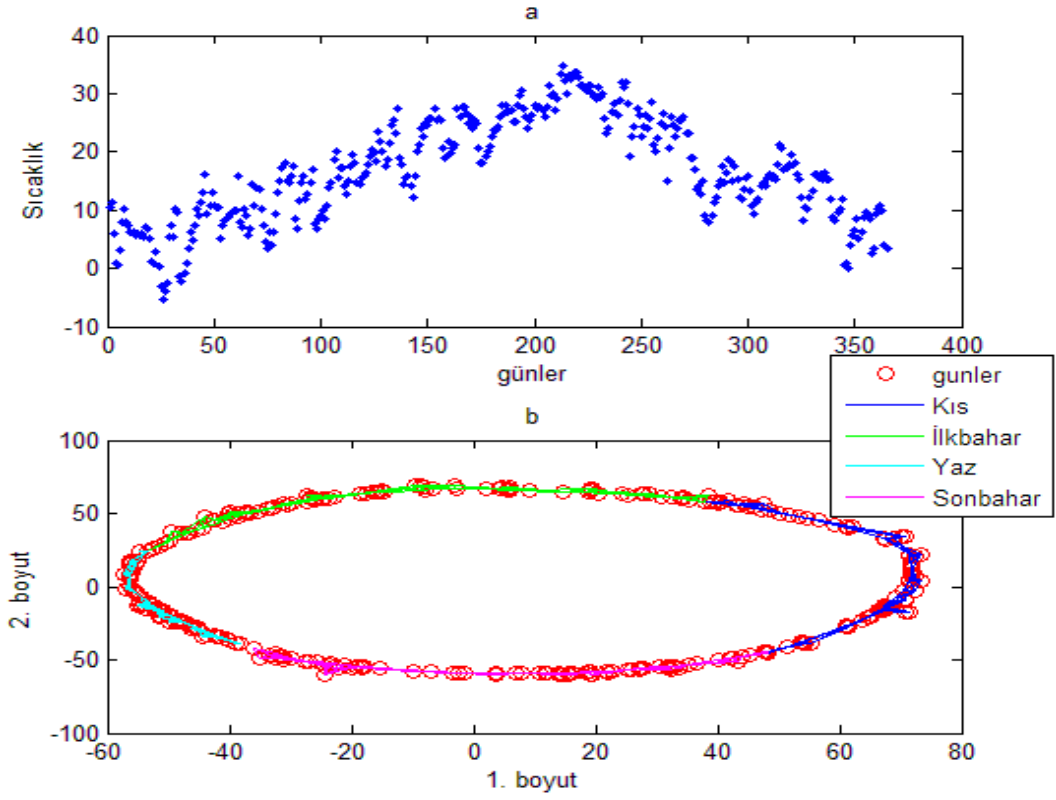
Şekil 3.6: NO<sub>x</sub>'nin bir yıllık, günlük ortalama gösterimi.



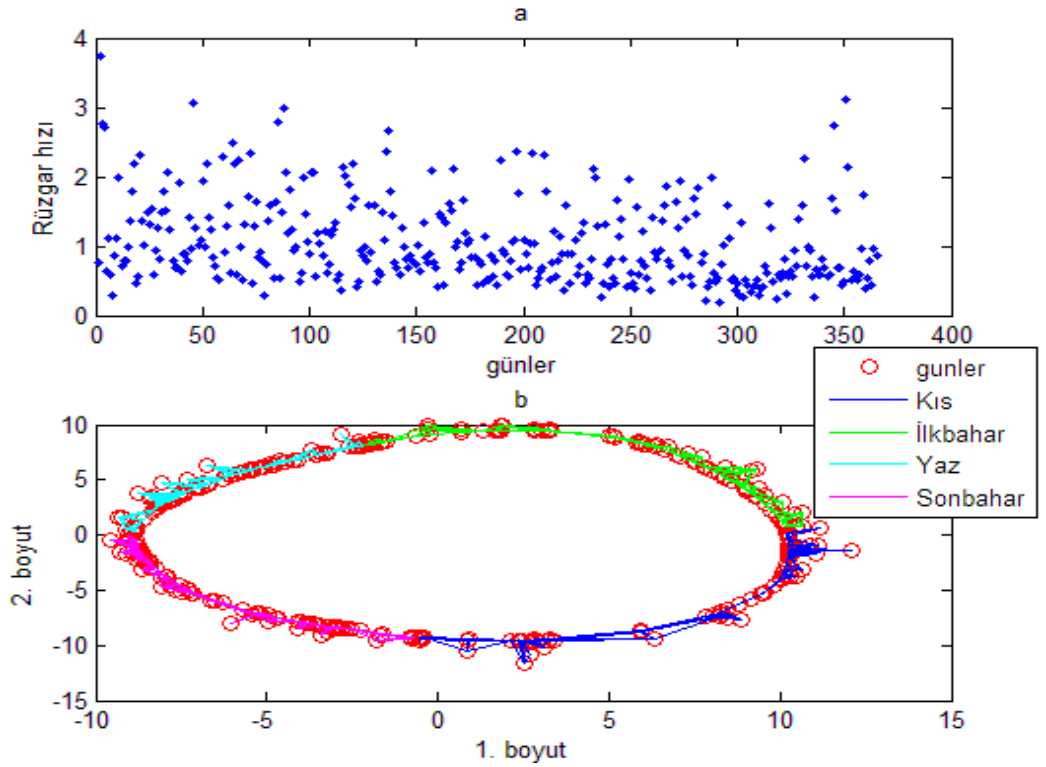
Şekil 3.7:  $O_3$ 'ün bir yıllık, günlük ortalama gösterimi.



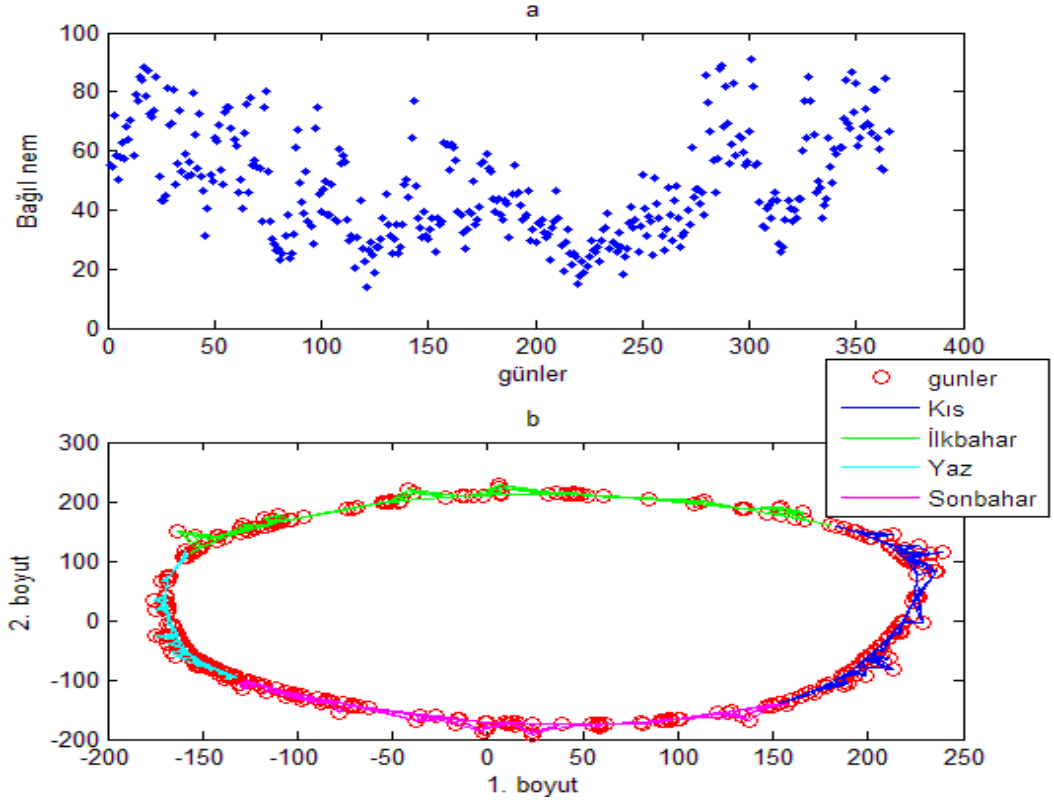
Şekil 3.8: CO'nun bir yıllık, günlük ortalama gösterimi.



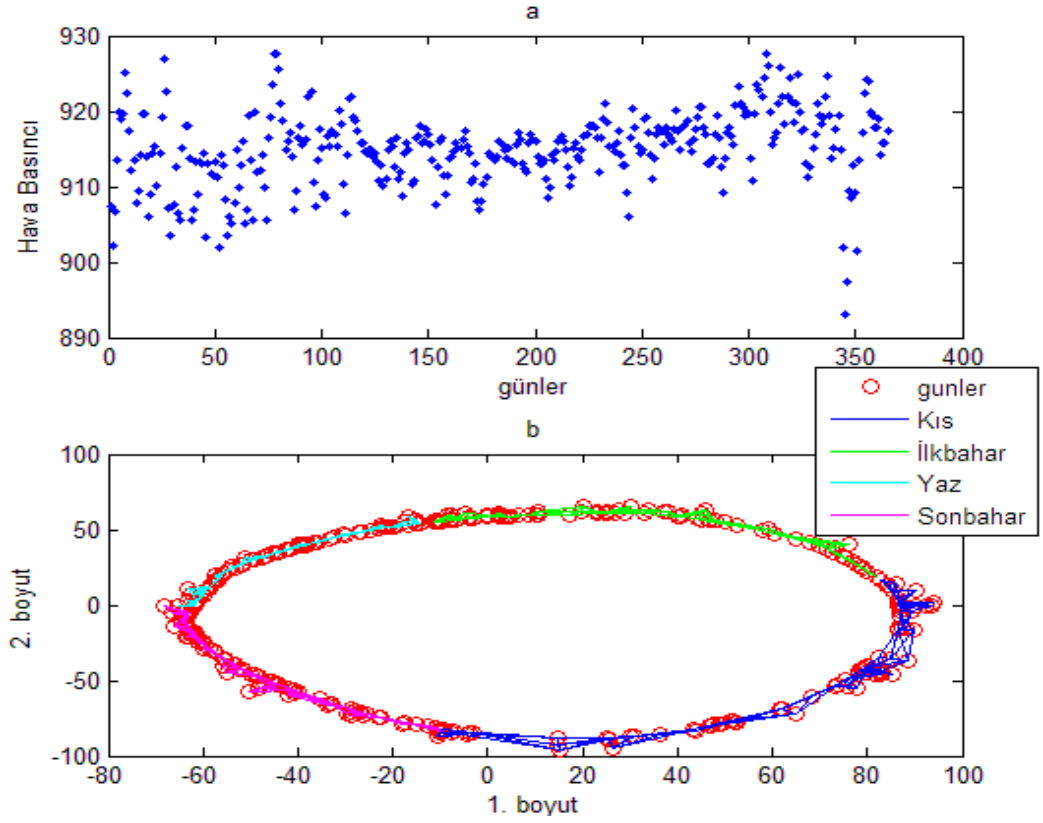
Şekil 3.9:Sıcaklık'ın bir yıllık, günlük ortalama gösterimi.



Şekil 3.10:Rüzgar hızı'nın bir yıllık, günlük ortalama gösterimi.



Şekil 3.11: Bağıl nem'in bir yıllık, günlük ortalama gösterimi.



Şekil 3.12: Hava Basıncı'nın bir yıllık, günlük ortalama gösterimi.

**Tablo 3.1:** Değişkenler ilgili istatistiki değerler.

Değişkenler	Birim	Minimum	Maksimum	Ortalama	StandartSapma	Varyans
PM10	$\mu\text{g}/\text{m}^3$	15.33333	226.4444	57.13546	36.80083	1354.301
PM2.5	$\mu\text{g}/\text{m}^3$	2.745206	300.8394	39.8825	32.60388	1063.013
SO <sub>2</sub>	$\mu\text{g}/\text{m}^3$	0	106.8889	17.14404	17.59755	309.6737
NO	$\mu\text{g}/\text{m}^3$	2.833333	240.2032	37.76197	42.86456	1837.371
NO <sub>2</sub>	$\mu\text{g}/\text{m}^3$	3.018701	119.6111	42.86118	20.03457	401.3841
NO <sub>x</sub>	$\mu\text{g}/\text{m}^3$	9.89203	346.8889	80.60457	56.87081	3234.289
O <sub>3</sub>	$\mu\text{g}/\text{m}^3$	5	150.1609	50.07688	37.23457	1386.413
CO	$\mu\text{g}/\text{m}^3$	0.889885	3587.889	476.2452	463.4958	214828.4
T	°C	-5.24444	34.66667	16.62197	8.731026	76.23082
RH	m/h	0.2	3.744444	1.052053	0.608728	0.370549
BN	%	14.12222	90.85556	47.63121	17.72027	314.0079
HB	mbar	893	927.6667	914.9963	4.940037	24.40396

### 3.2. OZON SEVİYESİ ANİ DEĞİŞİMLER VERİKÜMESİ

Ozonun t. günde aldığı değer , ile gösterilsin. İki ardışık gün arasındaki ozon farkının th eşik değerinden yüksek olduğu gün çifti aşağıda ki gibi tanımlanır.

Farkın th değerinden yüksek olduğu gün çiftleri ve  $[(t+1)]$  \_thile ifade edilsin. Veri seti ozon haricinde ki d değişkenin k adetli kombinasyonlarından  $C(d|k)=c$  tane değişken grubu oluşturulsun. Herbir çok boyutlu değişken , şeklinde ifade edilsin. Böylece veri setinde bulunan 11 tek boyutlu değişkenden 165 çok boyutlu değişken elde edildi. Veri seti bu değişkenlerin t. ve (t+1). günlerinden oluşmaktadır.

Bu veri kümesi bir yıl içinde ozon seviyesinde meydana gelen ani değişimlerin olduğu günler ile bir önceki günlerden oluşmaktadır. Böylece toplamda 330 tane 3 boyutlu

değişken içeren veri setinin örnek sayısı seçilen eşik değere göre değişmektedir. Ani artan ve ani azalan günler  $th=15$  seçildiğinde sırasıyla 57 ve 41,  $th=20$  seçildiğinde sırasıyla 30 ve 24,  $th=25$  seçildiğinde sırasıyla 19 ve 20,  $th=30$  seçildiğinde sırasıyla 12 ve 12 değerlerini almaktadır.

Bu veri seti ozon seviyesi ani değişimleri veri kümesi olarak adlandırıldı. Tez kapsamında İngilizce karşılığı olan “Abrupt Changes in Level of Ozone (ACL-OZONE)” kısaltmasıyla adlandırıldı.

### 3.3. TEMEL İSTATİSTİKİ KAVRAMLAR

*Standart Sapma*: Bir olasılık dağılımından gelen verinin veya bir rassal değişkenin örneklerinin beklenen değerden sapma miktarının beklenen değeridir.  $\sigma$  veya  $s$  ile sembolize edilir. Standart sapmanın matematiksel ifadesi denlem 3.1’de verilmiştir.

$$\sigma = \sqrt{E\left((x - E(x))^2\right)} \quad (3.1)$$

Bir değişkenin beklenen değeri, değişkenin tüm örneklerinin toplanıp örnek sayısına bölümüne eşittir. Beklenen değer aşağıda ki gibi ifade edilir.

$$E(x) = \frac{1}{n} \sum_{1}^n x_n \quad (3.2)$$

Anlatılacak olan denklem dönüşümlerinde rassal değişkenlerin veya vektörlerin sıfır merkezli olduğu varsayılacaktır. Bundan dolayı vektörün standart sapması aynı zamanda vektörün normuna eşit olacaktır. Standart sapmanın karesi ise *varyans*’a eşittir.

*Kovaryans*: İki rassal değişkenin beraber ne kadar değiştiklerinin ölçüsüdür. Diğer bir ifadeyle, birinci değişkenin örneklerinin beklenen değerinden farkları ile ikinci değişkenin örneklerinin beklenen değerinden farklarının birlikteliğinin beklenen değerine kovaryans denir.

$$Cov(x, y) = E\left[(x - E[x])(y - E[y])\right] \quad (3.3)$$

$$Cov(x, y) = E[xy] - E[x]E[y]$$



Değişkenlerin sıfır merkezli olduğu düşünülürse kovaryans denklemi aşağıdaki gibi sadeleşir.

$$\text{Cov}(x, y) = E[xy] \quad (3.4)$$

*Kovaryans matrisi:* X ve Y çok boyutlu değişkenler iseler bu değişkenlerin kovaryansları şöyle ifade edilir.

$$\text{Cov}(X, Y) = E[(X - E[X])(Y - E[Y])^T] = E[XY^T] \quad (3.5)$$

*Pearson Bağlantı katsayısı:* İki rassal değişken arasındaki ilişkinin gücünü ve yönünü belirtir. Bağlantı değeri -1 ve 1 arası değerler alır. Eğer bağlantı -1'e yakın bir değere sahipse, bu negatif yönde güçlü bir ilişkiyi ifade eder. Eğer bağlantı 1'e yakın bir değere sahipse, bu durumda pozitif yönlü güçlü bir ilişki var demektir. İki rassal değişken arasında doğrusal bir ilişki bulunmuyorsa bağlantı 0 değerini alır. İki rassal değişken arasındaki bağlantı, bu iki rassal değişkenlerin kovaryanslarının standart sapmalarının çarpımlarına bölümü olarak ifade edilir.

$$\rho_{x,y} = \frac{\text{Cov}(x, y)}{\sigma_x \sigma_y} = \frac{E((x - E(x))(y - E(y)))}{\|x\| \|y\|} \quad (3.6)$$

*Matthews Bağlantı Katsayısı (MCC):* Sınıf eleman sayılarının birbirinden farklı olduğu problemlerde sınıflandırıcının kesinliğini göstermek için kullanılmaktadır. MCC -1 ve +1 arası değerler alır. MCC'nin -1'e veya +1'e yakın olması iyi bir sınıflandırma yapıldığını 0'a yakın olması ise iyi bir sınıflandırma yapılamadığını göstermektedir. MCC karışıklık matrisinde bulunan sınıflara ait istatistikleri kullanır. Karışıklık matrisi iki sınıflı durum için Tablo 3.2'de gösterilmektedir. Tabloda ki TP ilk sınıfa ait doğru etiketlenen örnek sayısını, FP ilk sınıfa ait yanlış etiketlenen örnek sayısını, FN ikinci sınıfa ait yanlış etiketlenen örnek sayısını, TN ise ikinci sınıfa ait doğru etiketlenen örnek sayısını göstermektedir.

**Tablo 3.2:** İki sınıflı durum için karışıklık matrisi.

Gerçek Değer	Tahmin Edilen değer	
	TP	FP
	FN	TN

Bu durumda MCC şöyle tanımlıdır:

$$MCC = \frac{TP*TN-FP*FN}{\sqrt{(TP+FP)*(TP+FN)*(TN+FP)*(TN+FN)}} \quad (3.7)$$

*F Skor*: Sınıflandırma başarısının değerlendirilmesinde kullanılır. Hassaslık (Precision) ve anma (recall) katsayılarını kullanmaktadır.

$$fskor = 2 * \frac{precision*recall}{precision+recall}$$

$$precision = \frac{TP}{TP+FP} \quad (3.8)$$

$$recall = \frac{TP}{TP+FN}$$

*Dairesel-dairesel bağıntı katsayısı*: iki dairesel değişken arasındaki ilişkinin derecesini belirlemek için kullanılır.  $\alpha$  ve  $\beta$   $n$  örnekle iki dairesel değişken olmak üzere şöyle tanımlanmıştır (Jammalamadaka, 2006):

$$r_c = \frac{\sum_{i=1}^n \sin(\alpha_i - \bar{\alpha}) \sin(\beta_i - \bar{\beta})}{\sqrt{\sum_{i=1}^n \sin^2(\alpha_i - \bar{\alpha}) \sin^2(\beta_i - \bar{\beta})}} \quad (3.9)$$

### 3.4. KANONİK BAĞINTI ANALİZİ

Kanonik bağıntı analizi, çok boyutlu verilerin analizinde kullanılan yöntemlerin en eskilerinden, aynı zamanda da en güçlülerinden biridir (Kurşun, 2011). Birbiri ile alakalı çok değişkenli iki veri kümesi elde edilebilen, çok değişik alanlarda kullanılmaktadır, örneğin: yüz ifadesi tahmini (Şakar, 2012), bir genomda ki virüs yeri tahmini (Gümü, 2012), bölgesel sel taşkınları tahmini (Ouarda, 2001), çok katmanlı sosyal ağların birbiri ile ilişkilerinin değerlendirilmesi (Carroll, 2006).

Kanonik bağıntı analizini ilk defa Hotelling tarafından 1936 yılında bilim dünyasına tanıtılmıştır. Teorik temelleri şöyle anlatılmaktadır (Hardoon, 2004) : kanonik bağıntı analizi, iki çok boyutlu değişken kümeleri arasında öyle baz vektörleri üretir ki bu iki küme bu vektörlere iz düşürüldüğünde aralarında ki doğrusal bağıntı en büyük olur. Bağıntı analizi değişkenlerin tanımlandığı koordinat sistemine bağlıdır. Bundan dolayı çok boyutlu değişkenler arasında kuvvetli bir ilişki olsa dahi kullanılan koordinat sisteminden ötürü bu ilişki görünmeyebilir. Yani değişkenler arasında doğrusal olmayan

bir ilişki olup, doğrusal bir ilişki yoksa kanonik bağıntı analizi bunu gösteremeyebilir. Kanonik bağıntı analizi iki çok değişkenli setin aralarındaki bağıntının en büyük olması için gerekli koordinatları bir dizi doğrusal dönüşümler sonucu üretilir.

Çok değişkenli vektör formu  $(X, Y)$  şeklinde gösterilsin. Örnekler de  $S = ((x_1, y_2), \dots, (x_n, y_n))$  şeklinde verilsin. Bu vektörlerin daha sade gösterimi için  $(x_1, \dots, x_n)$  vektörlerini  $S_x$  ve aynı şekilde  $(y_1, \dots, y_n)$  vektörlerini de  $S_y$  sembolize etsin. Diğer bir ifadeyle birinci ya da sol görü  $S_x$  ve ikinci ya da sağ görü  $S_y$  olarak adlandırılınsın. Sol görü  $(X)$  için yeni koordinatları oluşturmak için  $X$ ,  $w_x$  yönü üzerine izdürülür.

$$X \rightarrow \langle w_x, X \rangle \quad (3.10)$$

Denklem 3.7'de  $\langle \cdot, \cdot \rangle$  vektörlerin içsel çarpımını göstermektedir ve  $w_x^T$  ile  $X$ 'in çarpımına eşittir. Sağ görü  $(Y)$  için de aynı şekilde yeni koordinatları oluşturmak için  $Y$  de  $w_y$  yönü üzerine iz düşürülür.

$$Y \rightarrow \langle w_y, Y \rangle \quad (3.11)$$

Yeni oluşan  $x$  koordinatları şöyle oluşturulur.

$$S_{x, w_x} = (\langle w_x, x_1 \rangle, \dots, \langle w_x, x_n \rangle) \quad (3.12)$$

Aynı şekilde iz düşüm olmuş yeni  $y$  koordinatlarını da şu şekilde gösterilir.

$$S_{y, w_y} = (\langle w_y, y_1 \rangle, \dots, \langle w_y, y_n \rangle) \quad (3.13)$$

Kanonik bağıntı analizi, ilk adım olarak iki iz düşüm vektörü arasında ki doğrusal ilişkiyi en büyük yapacak  $w_x$  ve  $w_y$  yönlerini bulur (Denklem 3.14). Burada  $S_x$  ve  $S_y$  sıfır merkezli matrisler olarak kabul edilmiştir.

$$p = \max_{w_x, w_y} \text{corr}(S_x w_x, S_y w_y) = \max_{w_x, w_y} \frac{\langle S_x w_x, S_y w_y \rangle}{\|S_x w_x\| \|S_y w_y\|}. \quad (3.14)$$

Beklenen değer (Denklem 3.2)  $E$  olmak üzere, en büyük olmasını istediğimiz denklem şu şekilde olur.

$$\begin{aligned}
 p &= \max_{w_x, w_y} \frac{E[\langle w_x, X \rangle \langle w_y, Y \rangle]}{\sqrt{E[\langle w_x, X \rangle^2] E[\langle w_y, Y \rangle^2]}} \\
 &= \max_{w_x, w_y} \frac{E[w_x^T X Y^T w_y]}{\sqrt{E[w_x^T X X^T w_x] E[w_y^T Y Y^T w_y]}}.
 \end{aligned} \tag{3.15}$$

Denklem bir sonra ki aşamada aşağıda ki hale gelir.

$$p = \max_{w_x, w_y} \frac{w_x^T E[XY^T] w_y}{\sqrt{w_x^T E[XX^T] w_x w_y^T E[YY^T] w_y}}. \tag{3.16}$$

Denklem 3.16 incelenildiğinde Denklem 3.5'te ki kovaryans matris ifadesinin denklem içerisinde var olduğu görülmektedir. (X,Y) kovaryans matrisi şu şekilde ifade edilebilir.

$$C = E \left[ \begin{pmatrix} X \\ Y \end{pmatrix} \begin{pmatrix} X \\ Y \end{pmatrix}^T \right] = \begin{bmatrix} C_{XX} & C_{XY} \\ C_{YX} & C_{YY} \end{bmatrix} = C. \tag{3.17}$$

Denklem 3.17'te ki matris bir bütün olarak X ve Y'nin birleşiminin kovaryans matrisini (C) ifade etmektedir.  $C_{XX}$  ve  $C_{YY}$  grup içi kovaryansı,  $C_{YX} = C_{XY}^T$  ise gruplar arası kovaryansı göstermektedir. Denklem aşağıda ki gibi yazılabilir.

$$p = \frac{w_x^T C_{XY} w_y}{\sqrt{w_x^T C_{XX} w_x w_y^T C_{YY} w_y}} \tag{3.18}$$

Üzerine iz düşüm yapılacak olan vektörler  $w_x$  ve  $w_y$  dikkate alınarak bulunacak en büyük  $p$  değeri en büyük kanonik bağıntıyı gösterir. En büyük  $p$  değerini bulabilmek için elde edilen en iyileme problemi standart bir özdeğer problemine dönüştürülür.

Denklem 3.18'te  $w_x$  ve  $w_y$  birlikte veya ayrı ayrı büyütülmesi ya da küçültülmesi denklemin sonucunu etkilemez, örneğin  $w_y$ ,  $\beta$  katsayısı ile çarpılsın

$$\frac{w_x^T C_{XY} w_Y}{\sqrt{w_x^T C_{XX} w_x w_Y^T C_{YY} w_Y}} = \frac{\beta w_x^T C_{XY} w_Y}{\sqrt{\beta^2 w_x^T C_{XX} w_x w_Y^T C_{YY} w_Y}} \quad (3.19)$$

Bundan dolayı seçilecek olan katsayı keyfi bir değer olabilir. Denklem 3.18'te ki en iyileme problemini çözmek için aşağıda ki kısıtları sağlamak kaydıyla payın en büyük tutulması gerekir.

$$\begin{aligned} w_x^T C_{XX} w_x &= 1 \\ w_Y^T C_{YY} w_Y &= 1 \end{aligned} \quad (3.20)$$

Lagrange yöntemine göre bu durum aşağıda ki gibi yazılır.

$$L(\lambda, w_x, w_Y) = w_x^T C_{XY} w_Y - \frac{\lambda_x}{2} (w_x^T C_{XX} w_x - 1) - \frac{\lambda_y}{2} (w_Y^T C_{YY} w_Y - 1) \quad (3.21)$$

$w_x$  ve  $w_Y$ 'ye göre türevler alınır şu iki denklem elde edilir.

$$\frac{\partial L}{\partial w_x} = C_{XY} w_Y - \lambda_x C_{XX} w_x = 0 \quad (3.22)$$

$$\frac{\partial L}{\partial w_Y} = C_{YX} w_x - \lambda_y C_{YY} w_Y = 0 \quad (3.23)$$

Denklem 3.22  $w_x^T$  ile çarpılıp ve Denklem 3.23  $w_Y^T$  ile çarpılıp ikinci denklem birinci denklemden çıkarılıp aşağıda ki denklem elde edilir.

$$\begin{aligned} 0 &= w_x^T C_{XY} w_Y - w_x^T \lambda_x C_{XX} w_x - w_Y^T C_{YX} w_x + w_Y^T \lambda_y C_{YY} w_Y \\ &= \lambda_y w_Y^T C_{YY} w_Y - \lambda_x w_x^T C_{XX} w_x \end{aligned} \quad (3.24)$$

Denklem 3.20 göz önüne alındığında  $\lambda_y - \lambda_x = 0$  eşitliği elde edilir. Buradan bu iki çarpanın birbirinin aynısı  $\lambda = \lambda_x = \lambda_y$  olduğu görülmektedir.  $C_{YY}$  terslenebilir olduğunu varsayılırsa Denklem 3.23'den  $w_Y$  aşağıdaki hale gelir.

$$w_Y = \frac{C_{YY}^{-1} C_{YX} w_x}{\lambda} \quad (3.25)$$

Denklem 3.19'da  $w_Y$  yerine koyulur.

$$C_{XY}C_{YY}^{-1}C_{YX}w_X = \lambda^2 C_{XX}w_X \quad (3.26)$$

Denklem 3.26 genelleşmiş özdeğer problemi  $Ax = \lambda Bx$  haline gelmiştir. Bu problem çözülerek  $w_X$  elde edilir ve bir önceki denklemde yerine konularak  $w_Y$  bulunur. Böylece birbiri ile bağıntısı en yüksek olacak şekilde en iyileme yapılmış yeni koordinatlar elde edilir.

### 3.5. ISOMAP

İsomap, *çok boyutlu ölçekleme* kullanıp çok boyutlu verilerdeki gizli doğrusal yapıları koruyarak boyut indirgeme yapan bir makine öğrenmesi aracıdır (Turhan, 2004). İsomap, boyut indirgeme yaparken komşu olan örnekler arasındaki mesafeyi korur. Diğer bir ifadeyle komşuluğu koruyarak boyut indirgeme yapar. İsomap algoritması aşağıdaki gibi ifade edilmektedir (Tenenbaum, 2000):

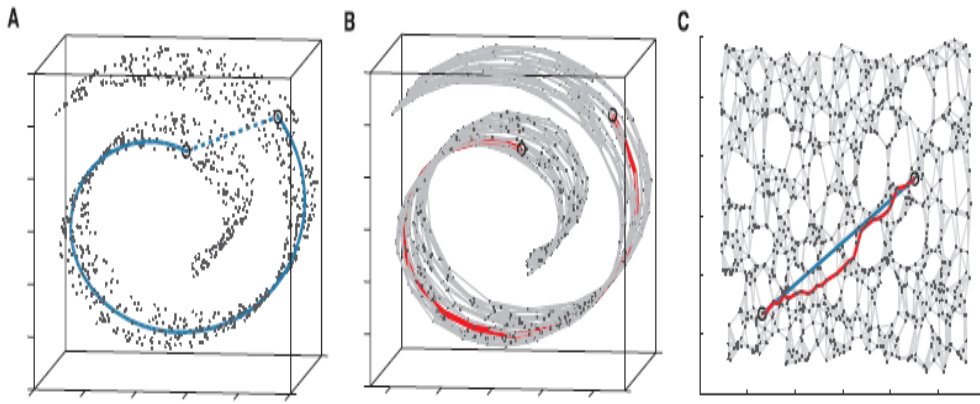
İsomap algoritması giriş olarak çok boyutlu  $X$  verisindeki her nokta çifti  $(i, j)$ , arasındaki mesafeleri  $M_X(i, j)$ , alır. Buradaki uzaklık hesabı için öklit metriği kullanılabilir gibi başka uzaklık metrikleri de kullanılabilir. Çıkış olarak  $X$ 'ten daha az boyuta sahip  $d$  boyutlu  $Y$  koordinat vektörlerini üretir. Algoritmanın adımları şöyle ifade edilmiştir:

1. Komşuluk grafinin oluşturulması.
2. En kısa yolların hesaplanması.
3.  $d$  boyutlu koordinat vektörlerinin oluşturulması.

Algoritmanın birinci adımında komşuluk grafi,  $G$ , oluşturulurken komşuluk ilişkisi, ya en yakın  $k$  nokta ya da belli bir mesafeden,  $\varepsilon$ , daha yakın olan noktalar olarak belirlenmektedir. İkinci adımda eğer nokta çiftleri arasında bir bağ yoksa bu iki nokta arası mesafe sonsuz olarak  $M_G(i, j) = \infty$  değerlendirilir. Eğer bu iki nokta birbirine komşu ise aralarındaki uzaklık grafa aktarılır  $M_G(i, j) = M_X(i, j)$ . Daha sonra birbirine bağlı noktalar üzerinden tüm noktaların birbirlerine olan en yakın yol hesaplanır. Hesap maliyetini düşürmek için seyrek matris (sparse matrix) kullanılır. Üçüncü adımda ise,  $\tau(M_G)$  matrisinin özdeğer ve özvektörleri bulunur. Büyükten küçüğe sıralanarak en

büyük özdeğere karşılık gelen özvektör üzerine iz düşüm yapılır. Burada  $\tau$  fonksiyonu  $\tau = \frac{-H(M_G)H}{2}$  şeklinde ifade edilir. İfade de geçen  $H$  sembolü  $H = \delta_{ij} - \frac{1}{N}$  dır.  $\delta$  sembolüde kronecker delta olarak ifade edilmektedir.

İsomap algoritması yer ölçümsel mesafeyi (geodesic distance) esas almaktadır. Yani graftaki herhangi iki nokta kuş bakışı mesafe ile birbirine çok yakın olsada komşuluk olan düğümler üzerinden birbirlerine bağlı olduklarından aslında birbirine uzak iki noktadrlar. Şekil 3.13 bunu çok güzel görselleştirmektedir.



Şekil 3.13: Yerölçümsel mesafenin gösterimi (Tenenbaum, 2000).

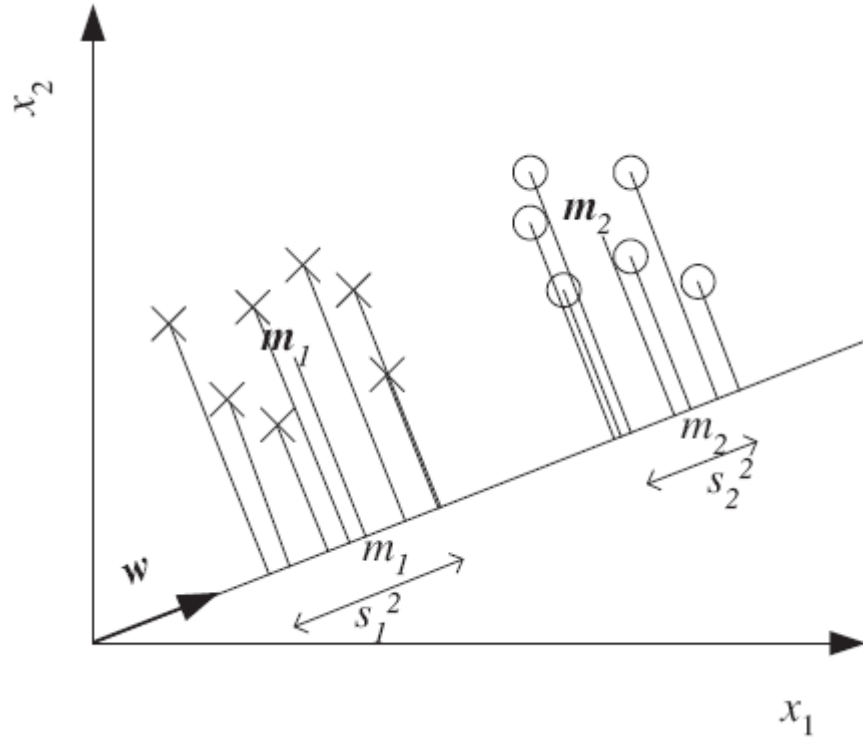
### 3.6. DOĞRUSAL AYIRTAÇ ANALİZİ

Doğrusal ayırtaç analizi (DAA), sınıflandırma için gözetimli bir boyut azaltma yöntemidir. İki sınıflı bir durum teorik olarak aşağıdaki gibi ifade edilmektedir (Alpaydın, 2010):

$C_1$  ve  $C_2$  sınıflarından örnekler verildiğinde,  $w$  yöneyiyle tanımlanan öyle bir yön bulmak isteniyor ki,  $w$  üzerine iz düşümleri alındığında iki sınıfın örnekleri birbirinden olabildiğince iyi ayrılsın. Daha önce görüldüğü gibi

$$z = w^T x \quad (3.27)$$

$x$  örneğinin  $w$  üzerine izdüşümüdür ve sayıl olduğu için boyut sayısını  $d$  değerinden bire indirir.



**Şekil 3.14:** İki boyutlu, iki kümeli verilerin  $w$  üzerine iz düşümü (Alpaydın, 2010).

$m_1 \in \mathfrak{R}^d$  ve  $m_2 \in \mathfrak{R}$ ,  $C_1$  sınıfından örneklerin izdüşümünden önceki ve sonraki ortalamasını gösterir.  $X = \{x^t, y^t\}$  örnekleme ( $x^t \in C_1$  ise  $r^t = 1$  ve  $x^t \in C_2$  ise  $r^t = 0$ ) kullanılarak bunlar kestirilebilir:

$$m_1 = \frac{\sum_t w^T x^t r^t}{\sum_t r^t} = w^T m_1$$

$$m_2 = \frac{\sum_t w^T x^t (1-r^t)}{\sum_t (1-r^t)} = w^T m_2 \quad (3.28)$$

$C_1$  ve  $C_2$  sınıflarında örneklerin izdüşümünden sonraki saçılımı şöyle tanımlanır:

$$s_1^2 = \sum_t (w^T x^t - m_1)^2 r^t$$

$$s_2^2 = \sum_t (w^T x^t - m_2)^2 (1 - r^t) \quad (3.29)$$

İzdüşümünden sonra iki sınıfın iyi ayrılmış olması için ortalamalarının olabildiğince uzak olmasını ve bir sınıfın örneklerinin olabildiğince küçük bir alana saçılması istenir. Yani  $|m_1 - m_2|$  değerini büyük ve  $s_1^2 + s_2^2$  değeri küçük olması istenir (Şekil 3.14). Fisher doğrusal ayırtacı, bunların oranını en büyük yapan  $w$  olarak tanımlıdır:



$$J(w) = \frac{(m_1 - m_2)^2}{s_1^2 + s_2^2} \quad (3.30)$$

Payı açılırsa:

$$\begin{aligned} (m_1 - m_2)^2 &= (w^T m_1 - w^T m_2)^2 \\ &= w^T (m_1 - m_2)(m_1 - m_2)^T w \\ &= w^T S_B w \end{aligned} \quad (3.31)$$

Burada,  $S_B = (m_1 - m_2)(m_1 - m_2)^T$  dizeyine sınıflar arası saçılım düzeyi denir. Payda ise sınıf örneklerinin izdüşüm sonrasında merkez çevresindeki saçılımlarının toplamıdır:

$$\begin{aligned} s_1^2 &= \sum_t (w^T x^t - m_1)^2 r^t \\ &= \sum_t w^T (x^t - m_1)(x^t - m_1)^T w r^t \\ &= w^T S_1 w \end{aligned} \quad (3.32)$$

Burada

$$S_1 = \sum_t r^t (x^t - m_1)(x^t - m_1)^T \quad (3.33)$$

Dizeyine  $C_1$  sınıfı için sınıf içi saçılım düzeyi denir.  $S_1 / \sum_t r^t$ ,  $\sum_1$  dizeyinin kestirimidir. Benzer biçimde,  $s_2^2 = w^T S_2 w$  ve  $S_2 = \sum_t (1 - r^t)(x^t - m_2)(x^t - m_2)^T$  tanımlıdır. Buradan

$$s_1^2 + s_2^2 = w^T S_w w$$

Yazılabilir ve  $S_w = S_1 + S_2$  dizeyine toplam sınıf içi saçılımı denir.  $s_1^2 + s_2^2$ 'in örnek sayısına bölümü birleşmiş verinin değişimsini verir. Denklem 3.30 yeniden yazılırsa:

$$J(w) = \frac{w^T S_B w}{w^T S_w w} = \frac{|w^T (m_1 - m_2)|^2}{w^T S_w w} \quad (3.34)$$

$J$ 'işlevinin  $w$  yönüne göre türevini alıp sifıra eşitlenirse,

$$w = S_w^{-1}(m_1 - m_2) \quad (3.35)$$

olarak yazılabilir.  $p\left(\frac{x}{c_i}\right) \sim N(\mu_i, \Sigma)$  olduğunda  $w = \Sigma^{-1}(\mu_1 - \mu_2)$  biçiminde doğrusal bir sınıflandırıcı oluşturduğunu için Fisher doğrusal ayırtacının sınıflar normal dağıldığında en iyi sınıflandırıcı olduğu söylenebilir. Bu varsayımınla  $w_0$  değeri de hesaplanabilir. Ama Fisher doğrusal ayırtacı sınıflar normal dağılmıyorsa da kullanılabilir; örnekleri  $d$  boyutlu uzaydan bir boyuta izdüşürdükten sonra herhangi bir sınıflandırıcı kullanılabilir. Diğer yandan

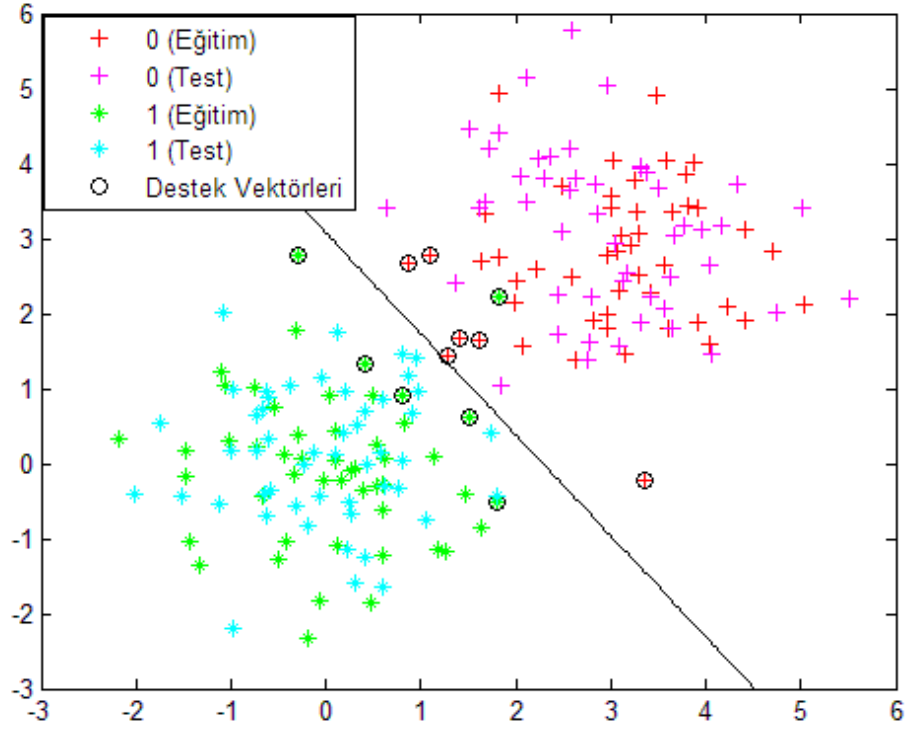
$$J(w) = \frac{w^T S_B w}{w^T S_W w} \quad (3.36)$$

işlevini en büyük yapan  $w$  vektörünü  $S_W^{-1} S_B$  dizeyinin en büyük özvektörleri kullanılarak da bulunabilir. Her ne kadar tek boyuta boyut azaltılırken sınıf ayrılabilirliği kıstas olarak alınmış olsa da, ardından bu boyutta sınıflandırma yapmak ya da threshold belirlemek için herhangi bir sınıflandırıcı kullanılabilir.

### 3.7. DESTEK VEKTÖR MAKİNELERİ

Destek vektör makineleri özellikle son yıllarda sınıflandırma ve regresyon işlemlerinde ki başarılarından dolayı sıkça tercih edilen bir makine öğrenmesi yöntemidir. Destek vektör makinelerinin kullanıldığı alanlar çok çeşitlilik göstermektedir. Örneğin; ozon seviyesi tahmini (Suárez-Sánchez, 2011), yüz tanıma (Meshgini, 2013), EEG sinyali sınıflandırma (Subasi, 2010), sismik dayanıklılık belirleme (Chen, 2012) destek vektör makinelerinin kullanıldığı bazı alanlardır.

Sınıflandırma problemleri için destek vektör makineleri veri setindeki örnekleri en iyi ayıracak üstün düzlemi bulmaya çalışır. Eğer sınıflar doğrusal olarak ayrılamıyorsa özel çekirdek fonksiyonları ile örnekler daha çok boyutlu bir uzaya taşınarak bu uzayda sınıflandırma yapılır. Sınıflar en iyi ayıran üstün düzlemin bulunmasında kenar payı kullanılır. Kenar payına giren örnek sayısının en az olduğu üstün düzlem sınıfları en iyi ayıran üstün düzlem olarak belirlenir. Şekil 3.15'te 50 eğitim ve 50 test örneğine sahip iki boyutlu bir verinin destek vektör makineleri ile sınıflandırılması görülmektedir.



Şekil 3.15: Destek vektör makineleri ile iki sınıflı verinin sınıflandırılması.

### 3.8. NARX YAPAY SİNİR AĞI MODELİ

Narx (Nonlinear AutoRegressive Exogeneous input) model yapay sinir ağlarının zaman serilerinde en yaygın kullanılan bir geri beslemeli tahmin modelidir. narx modelin değişik tahmin çalışmalarında kullanılmaktadır, örneğin uzun vadede zaman serisi tahmini (Menezes, 2008) ve hava kirlenmelerinin tepe noktaları tahmini (Pisoni, 2009).

Narx model hafızalı bir yapıya sahiptir. Ekstra girişin ve tahmin edilecek hedefin istenilen miktarda zaman adımı ağına içine gömülür. Daha sonra bunlar adım adım kaydırılarak hedefin bir sonraki zaman adımı için tahmini yapılır. Ağa gömülmek istenen zaman adımı  $n$  olmak üzere narx model şöyle tanımlanır.

$$y(t+1) = f(y(t), \dots, y(t-n), x(t), \dots, x(t-n)) + b$$

Modelde  $y$  çıkışı,  $x$  extra girişi ve  $b$  dengeleyiciyi ifade etmektedir. Çıkışın  $(t+1)$ . adım değeri giriş ve çıkışın  $t$ . adımdan  $(t-n)$ . adıma kadar olan değerleri ve dengeleyiciye bağlı olarak üretilir.

### 3.9. OZON VERİ KÜMESİ ZAMAN DİZİSİ VERİSİ İLE OZON KESTİRİMİ

Bu veri kümesinin  $t$ . ölçümleri kullanılarak ozonun  $(t + 1)$ . ve  $(t + 2)$ . zaman basamaklarına ait değerleri tahmin edildi. Tahmin için doğrusal regresyon kullanıldı. Tahmin doğruluğunun değerlendirilmesi için üç değişik katsayı kullanıldı. Bunlar; karesel hata (Root Mean Square Error, RMSE), mutlak hata (Mean Absolute Error, MAE) ve bağıntı ( $\rho$ ) katsayılarıdır.

Ozonun  $(t + 1)$ . güne ait değerinin tahmin çalışmalarında veri setinin ilk yarısı eğitim kümesi ikinci yarısı ise test kümesi olacak şekilde veri seti ikiye bölündü. Ozonun  $(t + 2)$ . güne ait değerinin tahmin çalışmalarında veri setinin  $t$ . günleri eğitim kümesi ve veri setinin  $(t + 1)$ . günleri ise test kümesi olacak şekilde veri seti ikiye bölündü. Bu iki regresyon çalışmasında da veri seti hem ölçüm değerleri ile hem de sıfır merkezine çekilerek ayrı ayrı kullanıldı. Ayrıca veri setinden ozonun  $t$ . ölçüm değerlerinin çıkarılmasının kestirim sonucunu nasıl etkilediği incelendi. Doğrusal regresyon çalışmasında tüm değişkenlerin kullanımının yanında bağıntı katsayısına ( $\rho$ ) göre seçilen en iyi üç değişkenin de kestirim sonuçları incelendi.

Ayrıca bu veri seti üzerinde ozonun  $(t + 2)$ . güne ait değerinin tahmin çalışmalarında örneklerin fiziksel komşuluklarını koruyarak verideki gizli komşulukları ortaya çıkaran *isomap* yöntemi kullanıldı. Bütün bir yıla ait veri dikkate alındığında yılın son günleri ile ilk günlerinin birbiri ile komşu olduğu bir komşuluk ilişkisi vardır. Bu ilişki korunarak veri setindeki her değişken iki boyuta iz düşürüldüğünde dairesel bir geometrik şeklin veri içinde saklı olduğu görüldü. Bu iki iz düşüm boyutu açı değerleri olarak bir boyutta ifade edildi. Ozona ait elde edilen açı değerleri *dairesel-dairesel bağıntı* katsayısı kullanılarak diğer değişkenlerden elde edilen açı değerleri ile ilişkilendirildi. Eklemeli öznitelik seçme (forward selection) kullanılarak seçilen en iyi 4 değişkenin gerçek değerleri kullanılarak doğrusal regresyon yapıldı.

Doğrusal regresyona ek olarak yapay sinir ağlarının zaman serisi tahminlerinde kullanılan dinamik bir modeli olan NARX (Nonlinear AutoRegressive Exogeneous input model) modeli kullanıldı. Zaman serisi tahmin modelleri genel olarak tahmin edilecek değişkeni giriş olarak alır ve buna göre çıkış üretir. Fakat NARX modele ozona ek olarak diğer kirleticiler eklenebilmektedir. Ayrıca çıkışı bir sonraki adımda giriş

olarak almaktadır. NARX model uygulanırken eğitim kümesinde 300 gün test kümesinde 65 gün olacak şekilde veri seti ikiye bölündü. Model hem tüm değişkenler için hem de bağıntı katsayısına ( $\rho$ ) göre en iyi üç değişken kullanılarak regresyon yapıldı.

### 3.10. ANİ DEĞİŞİMLERİN KESTİRİMİ İÇİN ÖZNETELİK SEÇİMİ

Ozon seviyesi ani değişimlerinin kestiriminde ACL-OZONE veri seti kullanıldı. Bu veri seti ani artış ve ani azalış olmak üzere iki sınıf içermektedir. Bu ani değişimlere neden olan değişken üçlülerinin tepiti için doğrusal ayırtaç analizi (DAA), destek vektör makineleri (DVM), en yakın k-komşuluk (EYKK) ve kanonik bağıntı analizi (KBA) makine öğrenmesi yöntemleri kullanıldı.

( $t + 1$ ). gün ani artışın olduğu gün olsun. Ani artış ve ani azalış gerçekleşen günleri en iyi sınıflandıran değişken grupları ozon seviyesinde meydana gelen bu ani değişimlerden sorumlu olan değişken grupları olarak tespit edilebilir.

DAA sınıf içi varyansın en az ve sınıflar arası varyansın en fazla olmasını sağlamaya çalışır. Bu amaçla Denklem 3.38'deki  $J$  katsayısını üretir. Bu katsayının büyük olması örneklerin iyi ayrıldığını gösterir. DAA'nın ürettiği en büyük  $J$  katsayısının ait olduğu kombinasyon bu iki kümeyi birbirinden en iyi ayıran değişken kombinasyonudur. Bu çalışmada ani değişimin olduğu  $t$ . ve ( $t + 1$ ). gündeğerleri beraber kullanıldı.

Destek kuvvet makineleri (DVM) sınıflandırma problemlerinde en sık kullanılan makine öğrenmesi yöntemlerinden biridir. Bu tez kapsamında da ozon seviyesinin ani artış ve ani azalış gösterdiği günlerin tespiti için ACL-OZONE veri setine uygulandı. DVM uygulanırken değişken üçlülerinin hem  $t$ . hem de ( $t + 1$ ). gün değerleri beraber kullanıldı. Sonuçların testi için *birini dışarıda bırak (leave-one-out, LOO)* yöntemi uygulandı. DVM'nin dört değişik çekirdek makinesi fonksiyonu kullanıldı, bunlar: doğrusal, rbf, polinomial, kuadratik tir.

En yakın-k komşuluk algoritması da bu kapsamda uygulanan doğrusal olmayan makine öğrenmesi yöntemlerindedir. Uygulamada komşuluk kriteri olarak  $k=7$  seçildi ve sonuçların testi için LOO kullanıldı. DVM ve EYKK yöntemleri ile oluşturulan

öznitelik seçim modellerinin kesinliği matthews-çarpraz-bağıntı, f-skor ve doğruluk katsayıları ile değerlendirildi.

Kanonik bağıntı analizi kullanılarak ani değişimlere neden olan değişken grupları bulunabilir. Bunun için her değişken kombinasyonunun  $t$ . ve  $(t + 1)$ . gün değerleri sırasıyla sol ve sağ görü olacak şekilde atandı. Bu iki görü arasında ki bağıntıyı en büyük yapacak  $w_X$  ve  $w_Y$  yönleri Denklem 3.39'dan hesaplandı. Daha sonra sol görü  $w_X$  yönü üzerine iz düşürülerek sol izdüşüm  $S_X$  elde edildi. Sonra ki adımda Denklem 3.6'dan sol görü  $S_X$  ile ozonun  $t$ . günde aldığı değerler arasındaki bağıntı katsayısı hesaplandı. En yüksek bağıntı katsayısına sahip olan gruptaki değişken bileşimleri ozonun ani değişimlerine neden olan değişkenlerdir.

## 4. BULGULAR

### 4.1. OZON ZAMAN SERİSİ VERİSİNDE KESTİRİM UYGULAMALARI

Ozonun geleceğe yönelik tahmini için çok değişkenli doğrusal regresyon kullanıldı. Ozon seviyesinin ( $t + 1$ ). gündeki değeri girdi olarak alınan değişkenlerin  $t$ . gündeki değerlerinin ağırlıklı toplamı olarak kestirildi. İlk olarak veri setindeki ozon dahil tüm değişkenler kullanılarak tahmin yapıldı. Oluşturulan modelde değişkenler aşağıdaki gibi ağırlıklandırıldı.

$$\begin{aligned} O_3(t + 1) = & 0.069 * PM_{10}(t) - 0.053 * PM_{2.5}(t) - 0.188 * SO_2(t) + 6.076 * NO(t) \\ & + 5.966 * NO_2(t) - 6.045 * NO_x(t) + 0.627 * O_3(t) - 0.002 * CO(t) \\ & + 0.873 * T(t) - 2.368 * RH(t) + 0.084 * BN(t) + 0.009 * HB(t) \end{aligned}$$

Bu modelde enyüksek katsayıları nitrik oksitlerin ve bağıl nemin aldığı görülmektedir. Bunlardan sonra ise ozon ve sıcaklık modelin en etkin değişkenleridir. Modelde en etkili değişkenlerden  $NO_x$  ve  $RH$  modele negatif katkı sağlarken diğer etkili değişkenler modele pozitif katkı yapmaktadır. Modelin doğruluğunun değerlendirilmesinde RMSE, MAE ve  $\rho$  katsayıları kullanıldı. Bu katsayılar eğitim kümesi için sırasıyla 13.042, 9.766, 0.869 değerlerini alırken test kümesi için sırasıyla 16.952, 13.8272 0.934 değerlerini aldı. Bu modele değişkenlere ek olarak bir dengeleyici (bias) olarak '1' eklendiğinde model aşağıda ki gibi oldu. Modelin doğruluk katsayılarının yukarıdaki sıraya göre eğitim kümesinde aldıkları 13.041, 9.768, 0.869 değerlerine karşın test kümesinde 17.023, 13.889, 0.934 değerlerini aldı. Bu modelin performansını çok azda olsa düşürdü.

$$\begin{aligned} O_3(t + 1) = & 0.069 * PM_{10}(t) - 0.055 * PM_{2.5}(t) - 0.189 * SO_2(t) + 6.080 * NO(t) \\ & + 5.973 * NO_2(t) - 6.051 * NO_x(t) + 0.626 * O_3(t) - 0.002 * CO(t) \\ & + 0.891 * T(t) - 2.229 * RH(t) + 0.092 * BN(t) + 0.063 * HB(t) \\ & - 49.980 \end{aligned}$$

Bu veri setinde ki deęişkenler birbirinden deęişik aralıklarda deęerler almaktadır. Bu handikaptan kurtulmak için deęişkenler sıfır merkezine çekilerek kestirim yapıldığında modelin performansında ciddi oranda bir iyileşme olduęu görülmektedir.

$$\begin{aligned} O_3(t + 1) = & 0.068 * PM_{10}(t) - 0.048 * PM_{2.5}(t) - 0.089 * SO_2(t) + 7.0 * NO(t) \\ & + 3.214 * NO_2(t) - 9.242 * NO_x(t) + 0.626 * O_3(t) - 0.027 * CO(t) \\ & + 0.208 * T(t) - 0.036 * RH(t) + 0.044 * BN(t) + 0.008 * HB(t) \\ & - 0.003 \end{aligned}$$

Üstteki modelde sıcaklık ile baęlı nem deęişkenlerinde önemli oranda bir deęişiklik görüldü. RMSE, MAE ve  $\rho$  katsayıları eğitim kümesi için sırasıyla 0.350, 0.262, 0.869 ve test kümesi için sırasıyla 0.457, 0.373, 0.934 olarak bulundu.

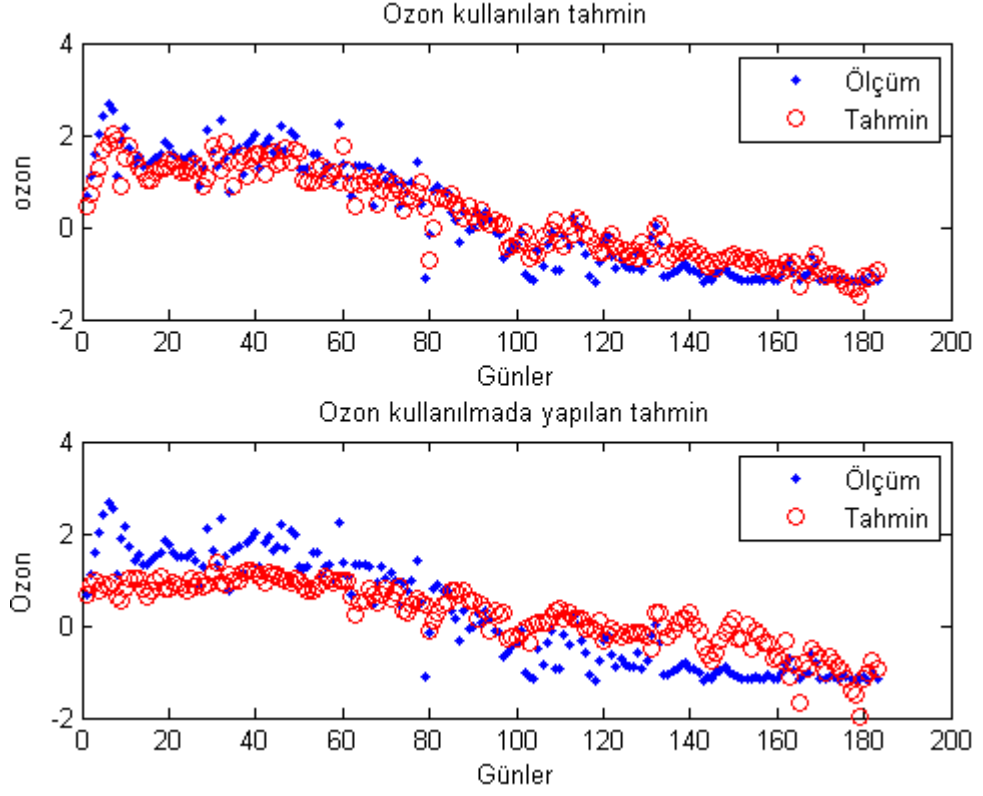
Tahmin modelinde ozonun  $t$ . güne ait deęerinin modelde olmasının gereklilięini incelemek için bu deęişkeni modelden çıkarıp ozonun  $(t + 1)$ . gün deęerinin kestirimi yapıldı. Dięer deęişkenler ile oluşturulan tahmin modeli aşıęıdaki gibidir. Modelin kesinlięinin tespiti için hesaplanan RMSE, MAE ve  $\rho$  katsayıları eğitim kümesi için sırasıyla 15.503, 12.661, 0.809 ve test kümesi için sırasıyla 26.227, 22.148, 0.843 deęerlerini aldı. Hata katsayı deęerlerinin yükselmesi ve beraberinde baęıntı katsayısının düşmesi  $t$ . güne ait ozon deęerinin katsayısı nitrik oksitlere göre düşük olmasına raęmen  $(t + 1)$ . güne ait ozon kestirimi için ozonun önemli bir deęişken olduęunu görüldü.

$$\begin{aligned} O_3(t + 1) = & -0.044 * PM_{10}(t) - 0.081 * PM_{2.5}(t) - 0.309 * SO_2(t) + 11.815 \\ & * NO(t) + 11.471 * NO_2(t) - 11.716 * NO_x(t) - 0.006 * CO(t) \\ & + 1.936 * T(t) - 3.153 * RH(t) - 0.098 * BN(t) + 0.295 * HB(t) \\ & - 222.319 \end{aligned}$$

Şekil 4.1'de üstteki panel ozon deęişkeni olmadan yapılan regresyonun test deęerlerine uygunluęu alttaki panelde ise ozon deęişkeni olduęunda yapılan regresyonun test deęerlerine uygunluęu görülmektedir.



Yukarıdaki çalışmada bir yılın ilk yarısı eğitim kümesi ikinci yarısı ise test kümesi olarak alındı. Aşağıdaki kısımda ise bir gün eğitim sonraki gün ise test kümesinde olacak şekilde veri seti iki kümeye bölündü.



**Şekil 4.1:**  $(t+1)$ . gün tahmin çalışmasında ozon kullanımı ve kullanılmaması farkı.

Ozon tahmin çalışmasında değişik eğitim ve test kümesi seçimi de yapıldı. Bu kısımda  $(t + 2)$ . günündeki ozon seviyesi kestirimi için veri setindeki değişkenlerin  $t$ . gün değerleri kullanıldı. Veri seti  $t$ . gün eğitim ve  $(t + 1)$ . gün test kümesi olacak şekilde ikiye bölündü. Bu kısımda yapılan  $t$ . gün değerlerinin kestirimi iki gün öncenin verileri ile yapıldığından dolayı yukarıda anlatılan tahmin problemine göre daha zor bir tahmin problemidir. Modele ait katsayılar Tablo 4.1’de görünmektedir.

**Tablo 4.1:** (t+2). gün tahmin çalışmasında değişkenlerin aldıkları katsayılar.

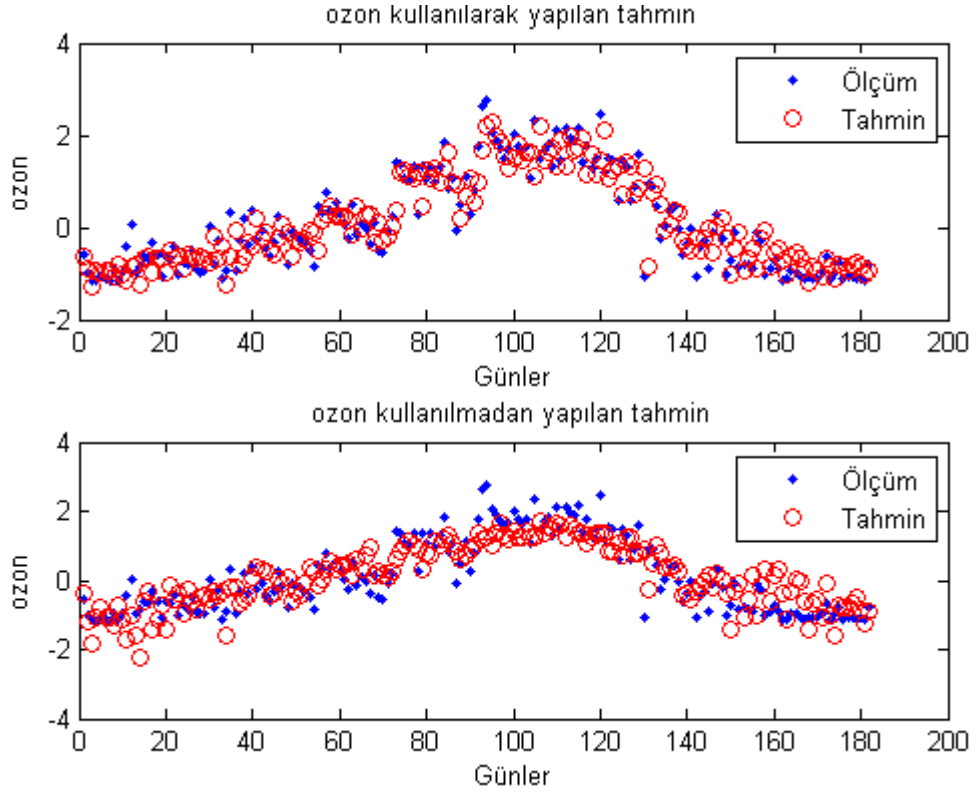
Değişkenler	Katsayılar (Ozon dahil)	Katsayılar (ozon hariç)
PM <sub>10</sub>	-0.120	-0.338
PM <sub>2.5</sub>	0.047	0.088
SO <sub>2</sub>	0.071	0.201
NO	-2.216	1.663
NO <sub>2</sub>	-2.176	1.507
NO <sub>x</sub>	2.181	-1.701
O <sub>3</sub>	0.661	0
CO	0.001	-0.002
T	1.368	2.977
RH	-3.225	-3.981
BN	0.211	-0.028
HB	-0.104	-0.285
Dengeleyici(bias)	86.714	291.086

Ozon tahmin edici olarak modelde yer aldığı anda oluşan RMSE, MAE ve  $\rho$  katsayıları eğitim kümesinde sırasıyla 15.369, 11.682, 0.909 değerlerini alırken test kümesinde ise sırasıyla 18.223, 13.968, 0.874 değerlerini aldı. Tahmin edici olarak ozon kullanılmadığında ise bu katsayılar eğitim kümesinde sırasıyla 18.546, 14.475, 0.865 ve test kümesinde ise sırasıyla 21.141, 16.976, 0.825 oldu. Tahmin modelinde ozon kullanmanın modelin kesinliğini pozitif yönde etkilediğini bu katsayılarıdaki değişim ortaya koydu. Değişkenler sıfır merkezine çekildiğinde üstteki katsayılar eğitim değerleri sırasıyla 0.523, 0.408, 0.865 ve test değerleri sırasıyla 0.596, 0.478, 0.825 oldu. Bağıntı katsayısının değişmemesi modelin uygunluğunu gösterdi. Hata katsayılarının düşmesi ise veri setindeki değişkenlerin farklı aralıklarda değerler almalarından kaynaklandı. Şekil 4.2’de ozon kullanıldığında regresyon sonucunun test kümesine daha iyi uyduğu görüldü.

Sadece ozon tahmin edici olarak kullanılarak (t + 1). ve (t + 2). zaman adımlarında ozon seviyesinin tahmini için oluşturulan doğrusal regresyon modelleri şöyledir:

$$O_3(t + 1) = 0.985 * O_3(t)$$

$$O_3(t + 2) = 0.961 * O_3(t)$$



**Şekil 4.2:**  $(t+2)$ . gün tahmin çalışmasında ozon kullanımı ve kullanılmaması farkı.

Bu modellerden  $(t + 1)$ 'e ait RMSE, MAE ve  $\rho$  katsayıları eğitim kümesinde sırasıyla 14.655, 10.555, 0.843 değerlerini alırken test kümesinde ise sırasıyla 15.117, 10.806, 0.939 değerlerini aldı. Diğer model  $(t + 2)$  için bu katsayılar eğitim kümesinde sırasıyla 17.058, 12.515, 0.891 değerlerini test kümesinde ise sırasıyla 19.880, 14.770, 0.855 değerlerini aldı. Tahmin modelinde yalnızca ozonu kullanmak diğer değişkenleri kullanmaktan daha düşük bir hata üretmediği görüldü.

Regresyon analizi yapılırken ilgili değişkenleri seçip ilgisizleri dışarıda bırakma zaman açısından ve modelin sadeliği bakımından olumlu etkiler yapar. Bu işlem *öznitelik seçme* olarak adlandırılmaktadır. Öznitelik seçimi doğrusal bağıntıya göre yapıldı. Tablo 4.2'de ozonun  $(t + 2)$ . günü ile diğer değişkenlerin  $t$ . gün değerleri arasındaki bağıntı katsayıları gösterilmektedir.

**Tablo 4.2:**  $(t+2)$ . Veri seti için ozon ile diğer değişkenler arası bağıntı katsayıları.

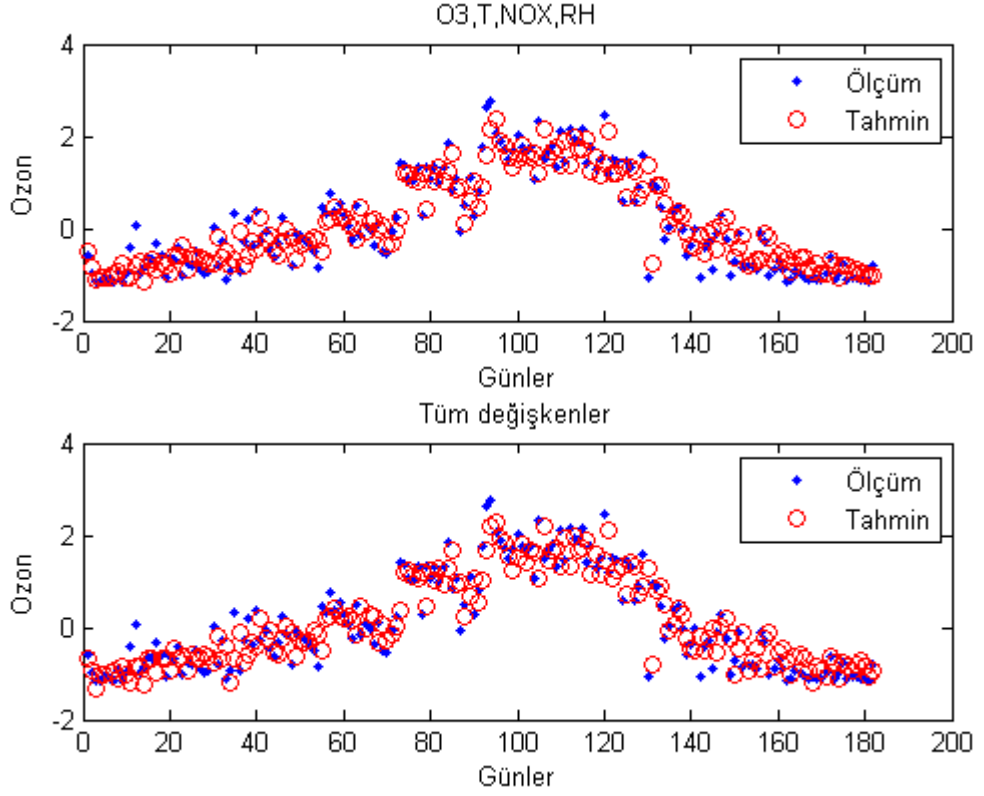
PM <sub>10</sub>	PM <sub>2.5</sub>	SO <sub>2</sub>	NO	NO <sub>2</sub>	NO <sub>x</sub>	O <sub>3</sub>	CO	T	RH	BN	HB
-0.46	-0.34	-0.52	-0.54	-0.53	-0.59	0.89	-0.56	0.80	-0.03	-0.57	-0.01

Tablo 4.2'ye göre en yüksek katsayıya sahip değişkenler  $O_3$ ,  $T$ ,  $NO_x$ ,  $BN$  değişkenleridir. Sadece bu değişkenler kullanılarak oluşturulan tahmin modeli şöyledir:

$$O_3(t + 2) = 0.705 * O_3(t) + 1.231 * T(t) - 0.023 * NO_x(t) + 0.238 * BN - 15.160$$

Modelin doğruluğunun değerlendirilmesinde kullanılan hata katsayıları RMSE, MAE ile bağıntı katsayısı  $\rho$  eğitim verisinde sırasıyla 15.638, 11.765, 0.906 ve test verisinde sırasıyla 17.937, 13.339, 0.878 değerlerini aldılar. Değerlendirme sonuçlarından görüldüğü üzere ilgisiz değişkenlerin modelden çıkarılması karmaşıklığı ve çalışma zamanını düşürmesinin yanında hata katsayılarında azalma ve bağıntı katsayısında artma sağladı. Şekil 4.3'ten üstteki panelde ki öznetelik seçimi yapılması sonucu elde edilen regresyon sonucunun, alttaki panelde ki tüm değişkenler kullanılarak yapılan regresyon sonucundan daha uygun olduğu görüldü.

Yapay sinir ağlarının geri beslemeli bir modeli olan *narx ağları (ekstra girişli doğrusal olmayan otoregresyon model)* zaman serisi tahminlerinde ilgilenilen değişkenden farklı değişkenlerde kullanılmak istendiğinde tercih edilen bir zaman serisi tahmin modelidir. Narx ağlarının bir özelliği de hafızalı yapıya sahip olmalarıdır. Bu çalışmada veri setinin 182 günü eğitim için kalan 183 günü ise test için ayrıldı. Modelin eğitiminde ilk 5 zaman dilimi yapay sinir ağının hafızasına gömüldü. Daha sonra eğitim kümesinin diğer günleri kullanılarak model eğitildi. Test setindeki gün değerleri kullanılarak model test edildi. Modelin eğitimi 18 iterasyon sürdü ve geri yayılım algoritması ile eğitildi. Modelin eğitim kümesine ait RMSE, MAE ve  $\rho$  katsayıları sırasıyla 11.275, 8.515, 0.907 olarak elde edilirken test kümesine ait aynı katsayılar sırasıyla 15.672, 11.464, 0.933 olarak elde edildi. Performans katsayılarının eğitim ve test kümelerinde birbirlerine yakın olması modelin ozon tahmininde kullanılabileceğini göstermektedir. Şekil 4.4'te üst panelde eğitim kümesine ait tahmin grafiği alt panelde ise test kümesine ait tahmin grafiği görülmektedir.



**Şekil 4.3:** Seçilen değişkenler ve tüm değişkenler ile yapılan tahmin sonuçları.

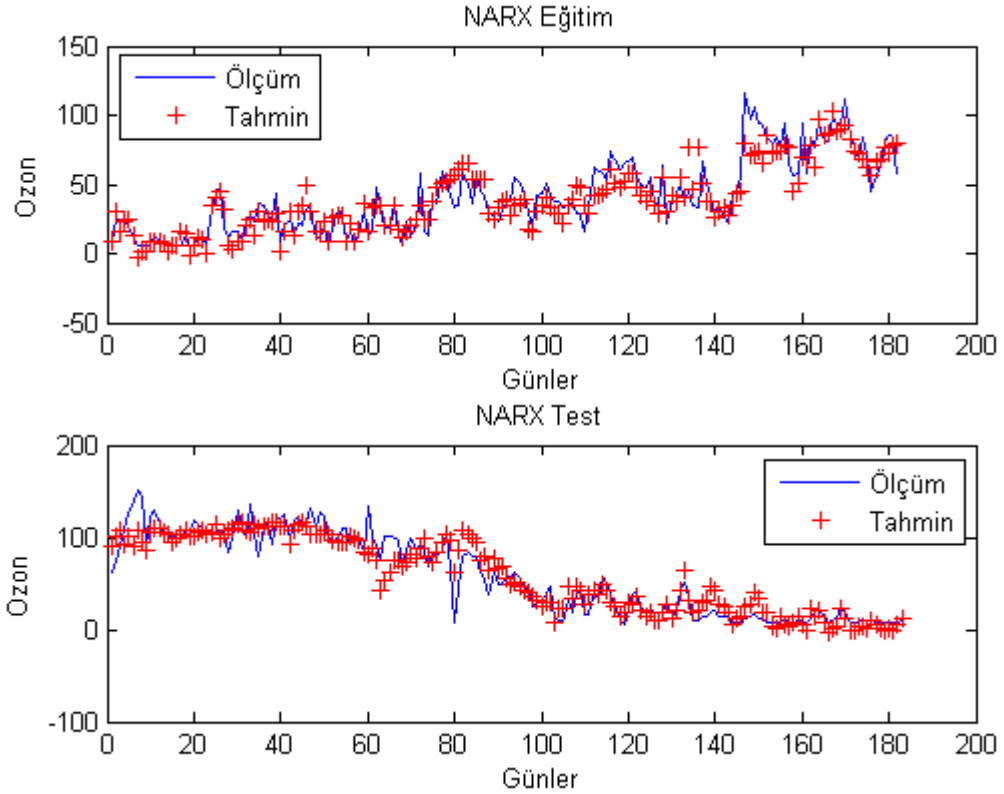
Isomap ile eklemeli öznitelik seçimi (forward feature selection) çalışmasında veri seti  $t$ . gün eğitim ve  $(t + 1)$ . gün ise test kümesi olacak şekilde ikiye bölündü ve  $(t + 2)$ . gün ozon seviyesi tahmini yapıldı. Eğitim kümesi hem öznitelik seçiminde hem de modelin eğitiminde kullanıldı. Öznitelik seçimi *isomap* ile *dairesel bağıntı* kullanılarak *eklemeli* olarak yapıldı. Eğitim kümesindeki tüm değişkenlerin  $t$ . gününün  $(n - 2)$ . ve  $(n + 2)$ . günleri ile komşuluğu korunacak şekilde değişkenler *isomap* ile iki boyuta iz düşürüldü. Her değişkene ait bu iki boyutlu izdüşümler dairesel bir geometrik yapıya sahip olduğu görüldü. Bu nedenle *isomap*'in ürettiği bu iki boyut aslında yeni uzaya taşınan değişkenin bu uzaydaki sinüs ve kosinüsü temsil etmektedir. Bu iki boyut (sinüs ve kosinüs) ters tanjant trigonometrik fonksiyonu kullanılarak bir açı değerine dönüştürüldü. Ayrıca ozonun  $(t + 2)$ . gününün  $t$ . ve  $(t + 4)$ . günleri ile komşuluğu korunacak şekilde *isomap* ile iz düşürülüp yeni uzaydaki hedef açı değeri bulundu. Yeni uzayda elde edilen bu iki değişken arasındaki ilişki dairesel bağıntı katsayısı ile derecelendirildi. Tahmin yapmak için 4 değişken seçildi. İlk değişken hedef açı ile en yüksek dairesel bağıntı katsayısına sahip değişken oldu. Seçilen diğer değişkenler

sırayla bu deęişkenin yanına eklenerek izdüřürüldüęünde en yüksek dairesel baęıntı katsayısını veren deęişkenlerdir.

Seçilen deęişkenler ozon, rüzgar hızı, sıcaklık ve hava basıncıdır ve dairesel korelasyon katsayıları sırasıyla 0.999, 0.999, 0.998, 0.992 dur. Oluřturulan doğrusal tahmin modelinde deęişkenlerin katsayıları ařaęıdaki gibidir.

$$O_3(t + 2) = 0.704 * O_3(t) - 1.035 * RH(t) + 0.971 * T(t) + 0.00 * HB$$

Oluřturulan modelin RMSE ve MAE katsayıları sırasıyla 18.536 ve 14.133 olarak elde edildi.



Şekil 4.4: NARX modele ait eğitim ve test regresyon grafikleri.

## 4.2. ACL-OZONE VERİSİNDE ANİ DEĞİŐİM KESTİRİMİ İÇİN ÖZİNTELİK SEÇİMİ

Veri setinin ozon seviyesinde belirli bir eşik deęerin üstünde deęişim olduęu günleri kullanılarak bu deęişime neden olan deęişken kombinasyonlarının tespiti çalışmasında *doęrusal ayırtaç analizi, destek vektör makineleri, en yakın k-komşuluk, kanonik baęıntı*

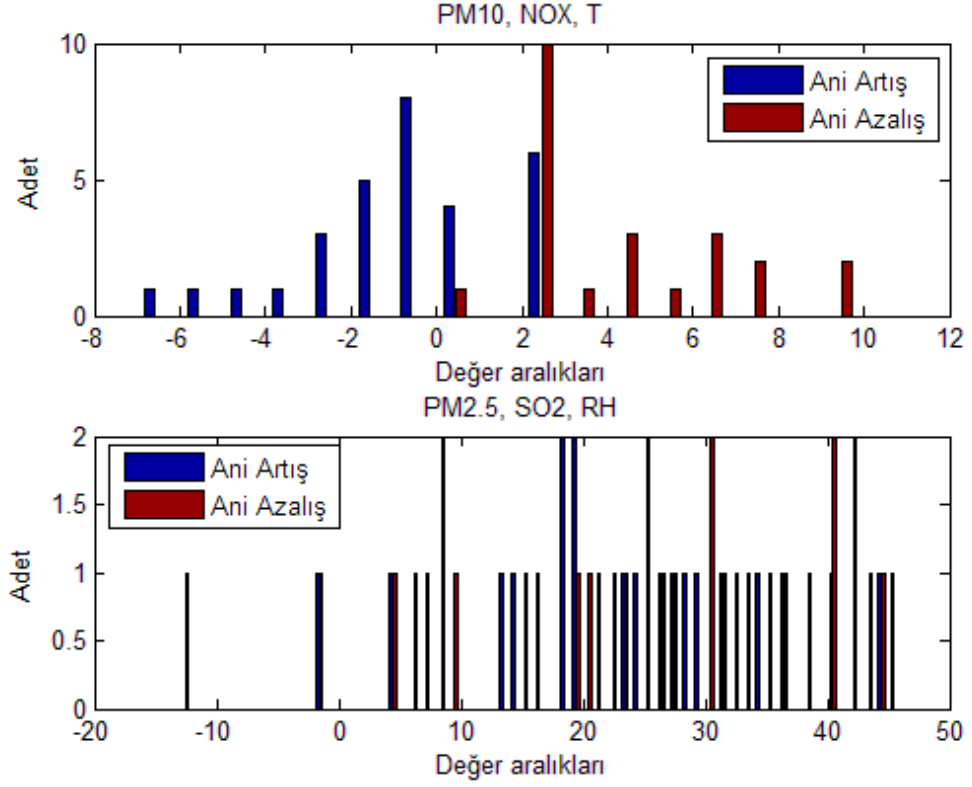
*analizi* algoritmaları kullanılmıştır. Bu yöntemlerin ilk üçü gözetimli öğrenme sonucusu ise gözetimsiz öğrenme yapmaktadır. Bu kısımda yukarıda adı geçen dört makine öğrenmesi yöntemi ile elde edilen bulgular paylaşılmaktadır.

Doğrusal ayırtaç analizi yöntemi, ozon seviyesindeki değişimin belirli bir eşik değerden daha fazla veya az olmasında rol oynayan değişkenlerin hangileri olduğunun incelenmesinde kullanıldı. Ozon seviyesindeki değişimin olduğu gün çifleri  $t$  ve  $(t + 1)$  gün olmak üzere sırasıyla günleri veri setinde kullanıldı. Değişkenlerin 3'lü alt kümeleri arasında bu iki sınıfı birbirinden en iyi ayıracak değişken kombinasyonu arandı. Ani artan ve azalan günleri birbirinden en iyi ayıran 3'lü değişken kombinasyonlarının belirlenmesinde Bölüm 3.5'te anlatılan  $J$  katsayısı kullanıldı. Tüm 3'lü kombinasyonlardan  $J$  katsayısı yüksek olan kombinasyon bu iki sınıfı en iyi ayıran değişken grubu olarak seçildi. Eşik değer 20 seçildiğinde en büyük  $J$  katsayısı değerine sahip ilk üç değişken grubu ve bu gruplara ait katsayı değerleri Tablo 4.3'te gösterilmektedir.

**Tablo 4.3:** DAA'nın bulunduğu ozon seviyesi ani değişimlerinin en fazla bağlı olduğu değişken grupları (th=20).

Değişken Grupları			$J$
PM <sub>10</sub>	NO <sub>x</sub>	T	1.184
PM <sub>10</sub>	NO	T	1.149
PM <sub>10</sub>	CO	T	1.091

Şekil 4.5'te iki farklı gün grubunun birbirinden ayrı olduğunu belirleyebilen ve belirleyemeyen değişken gruplarına ait histogramlar gösterilmektedir. Üst paneldeki histogram bu farklılığı belirleyebilen değişken grubu [PM<sub>10</sub>, NO<sub>x</sub>, T]'na ait, alt panel ise bu farklılığı belirleyemeyen grup olan [PM<sub>2.5</sub>, SO<sub>2</sub>, RH]'a ait histogramı göstermektedir. Şekil 4.6'te bu iki gruba ait iz düşümler ki gibi görünmektedir. Şekil 4.6'da üstteki panel en yüksek  $J$  değerine sahip değişkenlere ait, alttaki panel ise en düşük  $J$  değerine sahip değişkenlere aittir.



Şekil 4.5: En büyük ve en küçük  $J$  değerine sahip olan gruplara ait histogramlar ( $th=20$ ).

Tablo 4.4: DAA'nın bulunduğu ozon seviyesi ani değişimlerinin en fazla bağlı olduğu değişken grupları ( $th=30$ ).

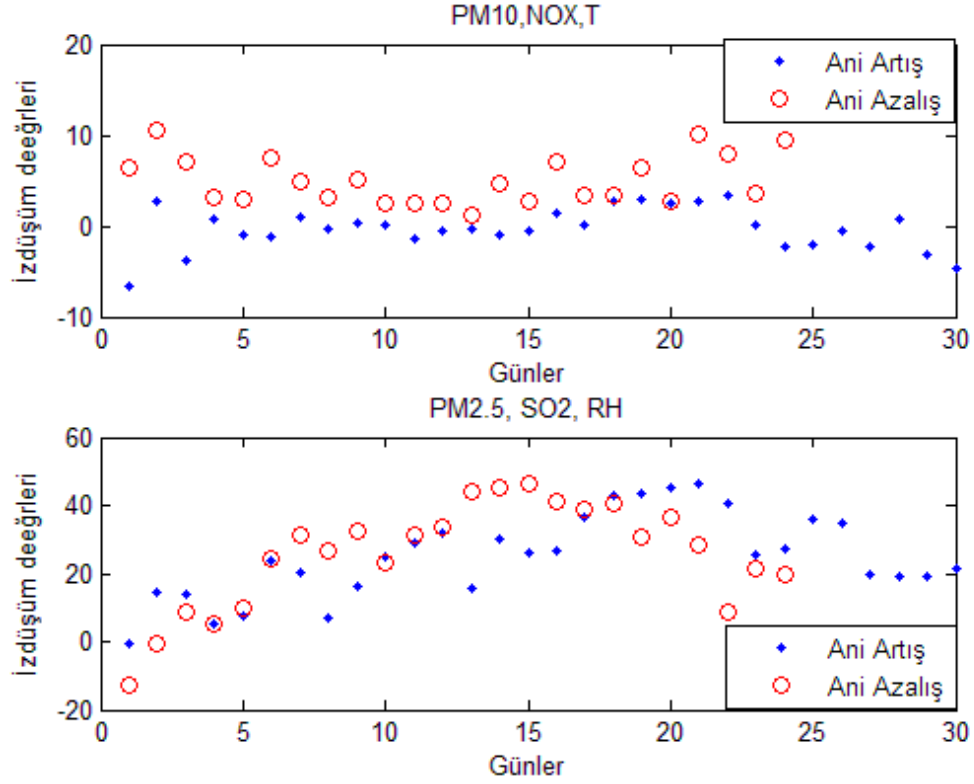
Değişken Grupları			$J$
PM <sub>10</sub>	PM <sub>2.5</sub>	CO	1.351
PM <sub>2.5</sub>	CO	BN	1.261
PM <sub>2.5</sub>	NO <sub>x</sub>	BN	1.237

Eşik değer 30 seçildiğinde büyük  $J$  katsayısı değerine sahip ilk üç değişken grubu ve bu gruplara ait katsayı değerleri Tablo 4.4'te gösterilmektedir. DAA eşik değer yüksek seçildiğinde sıcaklığı etkin değişken olarak seçmemektedir.

Ozonun ani artış ya da azalışını etkileyen değişkenlerin belirlenmesinde sınıflandırma araçlarında kullanılabilir. Ozonun ani arttığı günler ve ani azaldığı günler iki ayrı sınıfın üyeleri olsun. Bu durumda ozon harici değişkenler kullanılarak bu iki sınıfı başarılı bir şekilde sınıflandıran değişken grupları ozonun bu ani hareketine neden olan



değişkenlerdir. Sınıflandırma aracı olarak en bilinen yöntemlerden *K-en yakın komşu* ve *destek vektör makineleri* bu amaçla kullanıldı. Sınıflandırma geçerliliğinin değerlendirilmesinde *birini-dışarıda-bırak (leave-one-out)* çarpaz geçerleme uygulandı. Model başarıları Mathews çarpaz bağıntısı (Mathews Cross Correlation, MCC), f skor test ve başarı (accuracy, ACC) katsayıları kullanıldı.

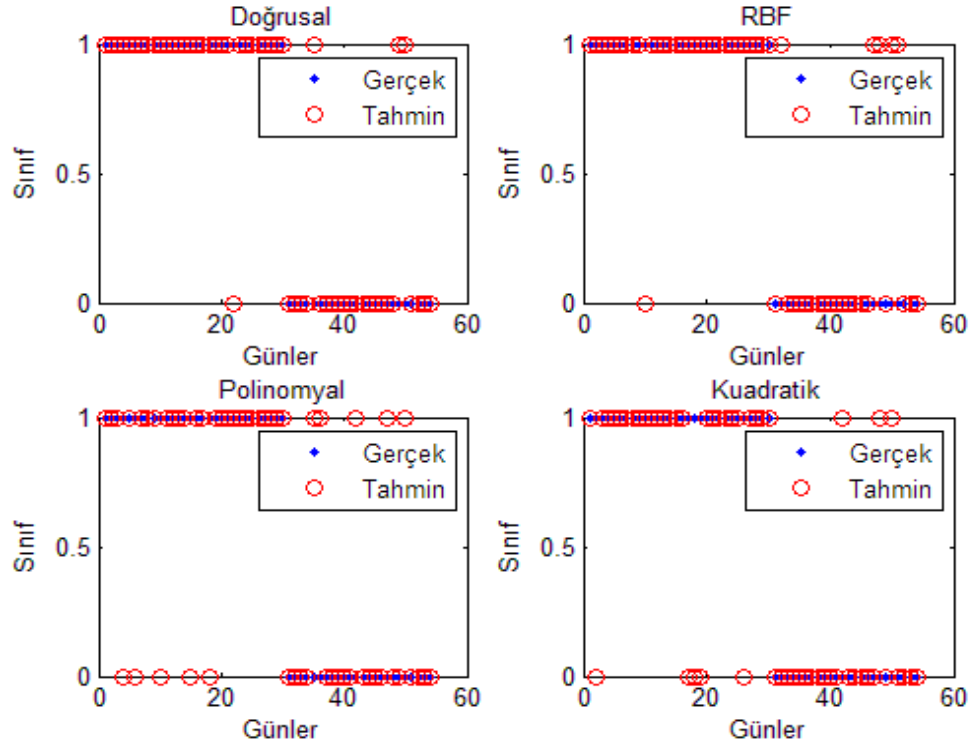


**Şekil 4.6:** DAA'nın ürettiği ani artış ve azalışı en iyi ve en kötü ayıran gruplara ait iz düşümler (th=20).

Destek Vektör Makinelerinin doğrusal, rbf, polinomial ve kuadratik çekirdek makineleri kullanılarak sınıflandırma yapıldı. Bu kısımda da değişkenler üzerli gruplar olarak ele alındı. Tablo 4.5'te destek vektör makinelerinin dört değişik çekirdek fonksiyonu kullanılarak seçilen değişkenler gösterilmektedir. Eşik değer olarak 20 seçildi. En yüksek ilişkili değişken grubu doğrusal çekirdek fonksiyonu ile bulundu. Doğrusal fonksiyonun seçtiği değişken grubu NO<sub>x</sub>, T, RH değişkenlerinden oluşmaktadır. Şekil 4.7'te dört çekirdek fonksiyonuna ait en başarılı tahminler gösterilmektedir.

**Tablo 4.5:** DVM'nin dört fonksiyonu kullanılarak seçilen değişkenler (th=20).

Th=20	Doğrusal				RBF			
	Değişkenler			Değer	Değişkenler			Değer
<b>MCC</b>	NO <sub>x</sub>	T	RH	0.851	NO	NO <sub>x</sub>	RH	0.780
<b>Fskor</b>	NO <sub>x</sub>	T	RH	0.913	NO	NO <sub>x</sub>	RH	0.863
<b>ACC</b>	NO <sub>x</sub>	T	RH	0.925	NO	NO <sub>x</sub>	RH	0.888
	Polinomyal				Kuadratik			
	Değişkenler			Değer	Değişkenler			Değer
<b>MCC</b>	NO <sub>x</sub>	RH	BN	0.625	SO <sub>2</sub>	T	BN	0.704
<b>Fskor</b>	NO <sub>x</sub>	RH	BN	0.791	SO <sub>2</sub>	T	BN	0.840
<b>ACC</b>	NO	CO	HB	0.814	SO <sub>2</sub>	T	BN	0.851

**Şekil 4.7:** Dört çekirdek fonksiyonunun en başarılı tahminleri (th=20).

Eşik değer olarak 30 seçilmesi durumunda destek vektör makinelerinin dört değişik çekirdek fonksiyonu kullanılarak seçilen değişkenler Tablo 4.6'da gösterilmektedir. En yüksek ilişkiyi doğrusal ve kuadratik çekirdek fonksiyonları verdi. Doğrusal fonksiyonun seçtiği değişken grubu PM<sub>10</sub>, NO, NO<sub>x</sub> değişkenlerinden oluşurken

kuadratik çekirdek fonksiyonunun seçtiği değişken grubu NO<sub>x</sub>, T, RH değişkenlerinden oluşmaktadır.

**Tablo 4.6:** DVM'nin 4 fonksiyonu kullanılarak seçilen değişkenler (th=30).

Th=30	Doğrusal				RBF			
	Değişkenler			Değer	Değişkenler			Değer
<b>MCC</b>	PM <sub>10</sub>	NO	NO <sub>x</sub>	0.833	T	BN	HB	0.666
<b>Fskor</b>	PM <sub>10</sub>	NO	NO <sub>x</sub>	0.916	T	BN	HB	0.833
<b>ACC</b>	PM <sub>10</sub>	NO	NO <sub>x</sub>	0.916	T	BN	HB	0.833
	Polinomyal				Kuadratik			
	Değişkenler			Değer	Değişkenler			Değer
<b>MCC</b>	NO <sub>2</sub>	T	RH	0.585	NO <sub>x</sub>	T	RH	0.833
<b>Fskor</b>	NO <sub>2</sub>	T	RH	0.800	NO <sub>x</sub>	T	RH	0.916
<b>ACC</b>	NO <sub>2</sub>	T	RH	0.791	NO <sub>x</sub>	T	RH	0.916

En yakın k-komşu Sınıflandırma yöntemi veri setindeki günler en yakın k=7 komşuya göre sınıflandırıldı. Eşik değer 20 seçildiğinde en etkili olduğu belirlenen değişkenler ve bunlara ait performans katsayıları Tablo 4.7'de görülmektedir.

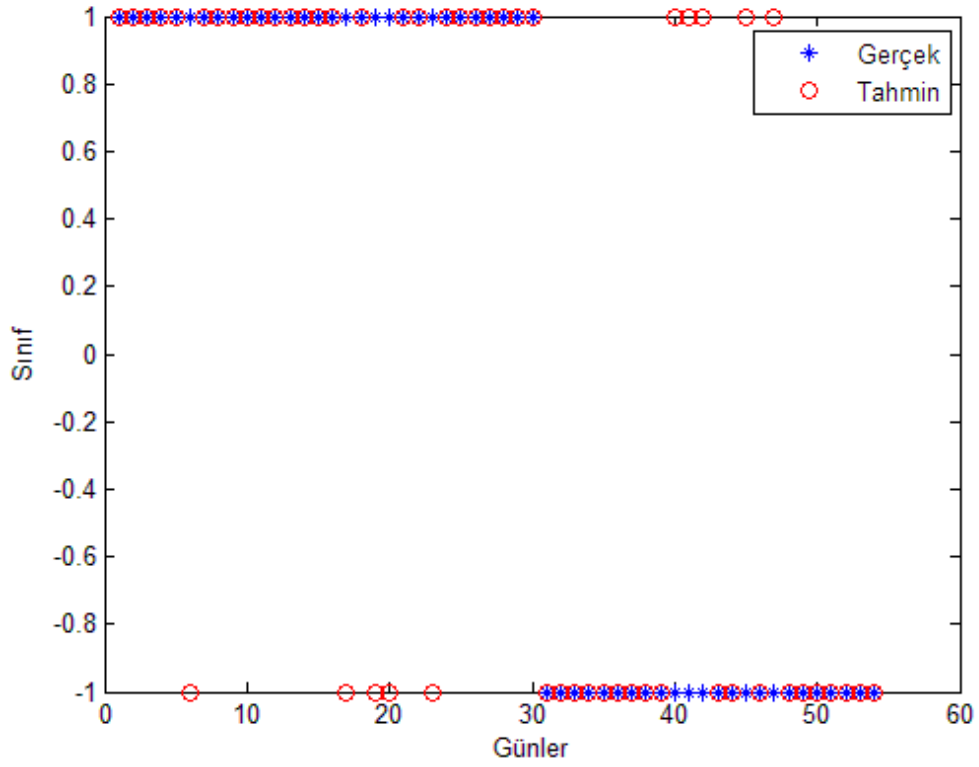
**Tablo 4.7:**EYKK'nın bulduğu en etkili değişken grubu (th=20).

Th=20	Seçilen Değişkenler			Değer
<b>MCC</b>	PM <sub>10</sub>	NO	BN	0.625
<b>F score</b>	PM <sub>10</sub>	NO	BN	0.791
<b>ACC</b>	PM <sub>10</sub>	NO	BN	0.814

**Tablo 4.8:** EYKK'nın bulduğu en etkili değişken grubu (th=30).

Th=30	Seçilen Değişkenler			Değer
<b>MCC</b>	SO <sub>2</sub>	T	BN	0.430
<b>F score</b>	PM <sub>2.5</sub>	T	BN	0.695
<b>ACC</b>	PM <sub>2.5</sub>	T	BN	0.708

Tablo 4.8'de eşik değer 30 seçildiğinde EYKK'nın bulduğu en yüksek ilişkili değişken grupları gösterilmektedir. Tablodan görüldüğü üzere T ve PM<sub>2.5</sub> yüksek değişimlerde ozon seviyesi üzerinde etkilidir. Bu değişken grubuna ait sınıflandırma başarısı Şekil 4.8'de gösterilmektedir.



**Şekil 4.8:** EYKK'nın bulduğu en etkili değişken grubu ile doğru tahmin edilen günler (th=20).

Ozon seviyesinin, ardışık iki gün için belli bir eşik değerden fazla değişmesine neden olan ve olmayan etmenlerin belirlenmesi için bağıntı analizi kullanıldı. Bağıntı katsayısı, değişken grubunun ozon miktarı değişimindeki etkisini göstermektedir. Yüksek bağıntı, bu değişimdeki etkili olan grupları, aynı şekilde düşük bağıntı ise değişimde etkili olmayan ya da az etkili olan grupları ifade etmektedir. Değişken gruplarından, kanonik bağıntı analizi ile elde edilen iz düşümleriyle ozon miktarının bağıntıları yüksek olan ve düşük olanlar sırasıyla Tablo 4.9 ve Tablo 4.11'de gösterilmektedir. Tablo 4.9'e göre orta seviyedeki ani değişimlerde NO, T, ve RH değişkenleri ozon değişiminde etkili iken daha yüksek seviyede ozon değişiminde PM, SO<sub>2</sub>, T ve BN değişkenlerin etkili olduğu görülmektedir. Ani değişimlerde etkisiz gruplara incelendiğinde ise genel olarak sadece kirleticilerin yer aldığı gruplar görülmektedir.

**Tablo 4.9:**Eşik değerlere göre ozon değişimine neden olan etmen grupları.

Eşik Değer=15					Eşik Değer=20				
Seçilen Değişkenler			Bağıntı	Kanonik	Seçilen Değişkenler			Bağıntı	Kanonik
NO	T	RH	0,814713	0,851517	NO	T	RH	0,773083	0,858175
NO	T	BN	0,812189	0,80897	SO <sub>2</sub>	NO	T	0,772752	0,837977
SO <sub>2</sub>	T	BN	0,811875	0,860761	PM <sub>2.5</sub>	T	BN	0,760343	0,820049
SO <sub>2</sub>	NO	T	0,807028	0,849361	PM <sub>10</sub>	T	BN	0,754325	0,842749
PM <sub>2.5</sub>	T	BN	0,80341	0,778785	NO <sub>2</sub>	T	RH	0,745065	0,915189
NOX	T	BN	0,794886	0,705246	NO	T	BN	0,711586	0,878924
SO <sub>2</sub>	NO <sub>x</sub>	T	0,783736	0,86427	T	RH	HB	0,709315	0,938329
PM <sub>2.5</sub>	NO	T	0,777779	0,764581	NO <sub>x</sub>	T	BN	0,700616	0,729637
NO	NO <sub>2</sub>	T	0,771244	0,856487	T	RH	BN	0,700266	0,887271
PM <sub>10</sub>	NO	T	0,762535	0,767637	PM <sub>10</sub>	NO	T	0,695049	0,783933
Eşik Değer=25					Eşik Değer=30				
Seçilen Değişkenler			Bağıntı	Kanonik	Seçilen Değişkenler			Bağıntı	Kanonik
PM <sub>2.5</sub>	T	BN	0,755355	0,813586	PM <sub>10</sub>	SO <sub>2</sub>	T	0,801667	0,898021
PM <sub>10</sub>	SO <sub>2</sub>	T	0,736105	0,878664	PM <sub>2.5</sub>	SO <sub>2</sub>	T	0,754565	0,921556
PM <sub>2.5</sub>	SO <sub>2</sub>	T	0,729503	0,916679	SO <sub>2</sub>	T	RH	0,69269	0,991444
PM <sub>2.5</sub>	NO <sub>2</sub>	T	0,698224	0,624188	T	RH	BN	0,688371	0,843158
SO <sub>2</sub>	NO	T	0,693857	0,894118	SO <sub>2</sub>	NO	T	0,681904	0,949041
SO <sub>2</sub>	NO <sub>x</sub>	T	0,683187	0,905289	SO <sub>2</sub>	T	BN	0,677045	0,8909
T	RH	HB	0,678322	0,922963	PM <sub>2.5</sub>	T	BN	0,67692	0,783784
SO <sub>2</sub>	T	RH	0,669218	0,850232	NO <sub>x</sub>	T	RH	0,66849	0,949175
NO	NO <sub>x</sub>	T	0,665705	0,900377	SO <sub>2</sub>	NO <sub>x</sub>	T	0,66753	0,960517
T	RH	BN	0,662474	0,865451	PM <sub>2.5</sub>	SO <sub>2</sub>	RH	0,656184	0,799146

**Tablo 4.10:** Bağıntısı en yüksek gruplar için KBA'nın ürettiği katsayılar.

EşikDeğer =15		EşikDeğer =20		EşikDeğer=25		EşikDeğer =30	
Değişkenler	1.Görü	Değişkenler	1.Görü	Değişkenler	1.Görü	Değişkenler	1.Görü
NO	0,463718	NO	0,046879	PM <sub>2.5</sub>	0,544404	PM10	0,275098
T	-0,81386	T	-0,72938	T	-0,79919	SO <sub>2</sub>	0,528214
RH	0,350141	RH	0,682501	BN	0,254788	T	-0,80331

En yüksek bağıntıya sahip değişken gruplarına ait, kanonik bağıntı analizinin bulduğu katsayılar ve kanonik bağıntıları Tablo 4.10'da gösterilmektedir. En düşük bağıntıya

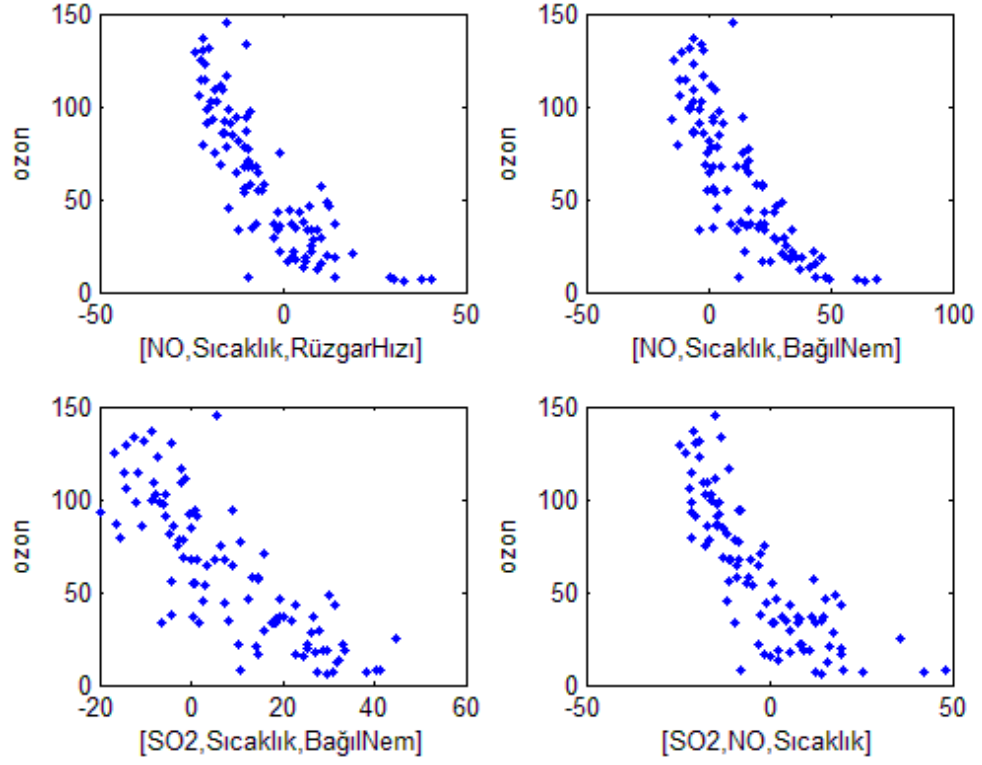
sahip deęişken gruplarına ait, kanonik baęıntı analizinin bulduęu katsayılar ve kanonik baęıntıları ise Tablo 4.12’te gösterilmektedir. Deęişkenlere ait katsayılara bakıldığında grup içindeki en önemli deęişkenin sıcaklık olduęu görülmektedir.

**Tablo 4.11:**Seçilen eşik değere göre ozon ani deęişimi ile baęıntılı olmayan deęişken grupları.

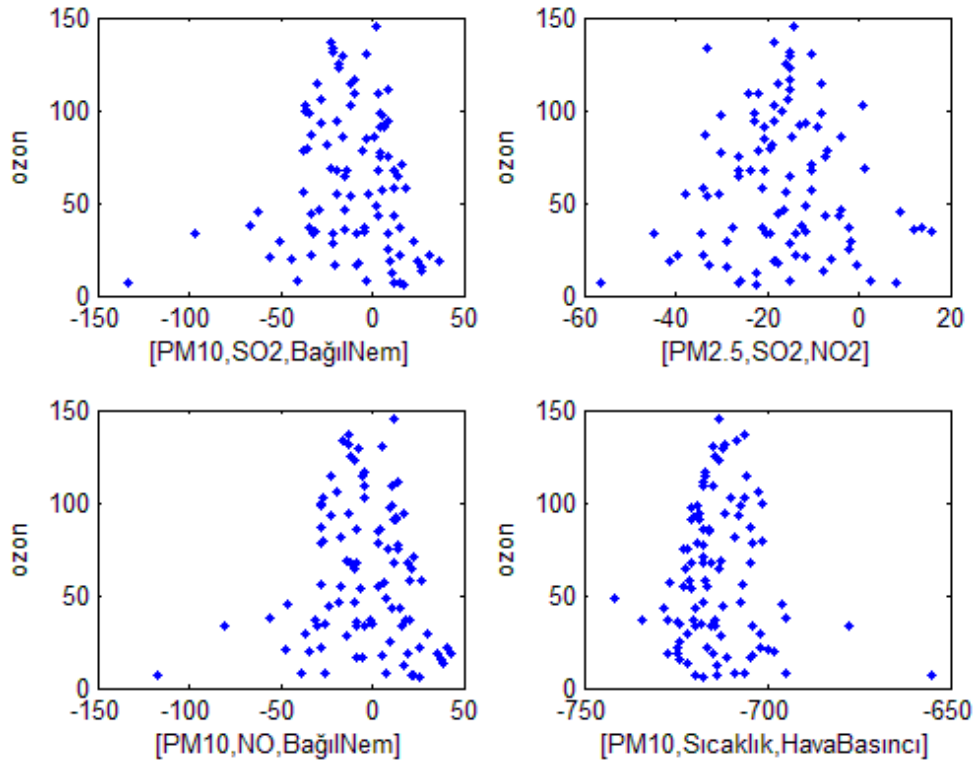
Eşik Deęer=15					Eşik Deęer=20				
Seçilen Deęişkenler			Baęıntı	Kanonik	Seçilen Deęişkenler			Baęıntı	Kanonik
PM <sub>10</sub>	SO <sub>2</sub>	BN	0,007319	0,574149	PM <sub>10</sub>	PM <sub>2,5</sub>	SO <sub>2</sub>	0,007186	0,287381
PM <sub>2,5</sub>	SO <sub>2</sub>	NO <sub>2</sub>	0,015494	0,629656	NO	NO <sub>x</sub>	RH	0,037268	0,670575
PM <sub>10</sub>	NO	BN	0,027609	0,597099	PM <sub>10</sub>	SO <sub>2</sub>	NO <sub>2</sub>	0,050914	0,615309
Eşik Deęer=25					Eşik Deęer=30				
Seçilen Deęişkenler			Baęıntı	Kanonik	Seçilen Deęişkenler			Baęıntı	Kanonik
PM <sub>10</sub>	PM <sub>2,5</sub>	BN	0,00369	0,715778	PM <sub>2,5</sub>	SO <sub>2</sub>	NO	0,025806	0,632917
PM <sub>10</sub>	NO <sub>x</sub>	BN	0,006074	0,779761	PM <sub>10</sub>	PM <sub>2,5</sub>	NO <sub>2</sub>	0,06172	0,707238
PM <sub>10</sub>	SO <sub>2</sub>	BN	0,01869	0,666079	PM <sub>2,5</sub>	NO	BN	0,099345	0,733364

**Tablo 4.12:** Baęıntısı en düşük deęişken grupları için KBA’nın ürettięi katsayılar.

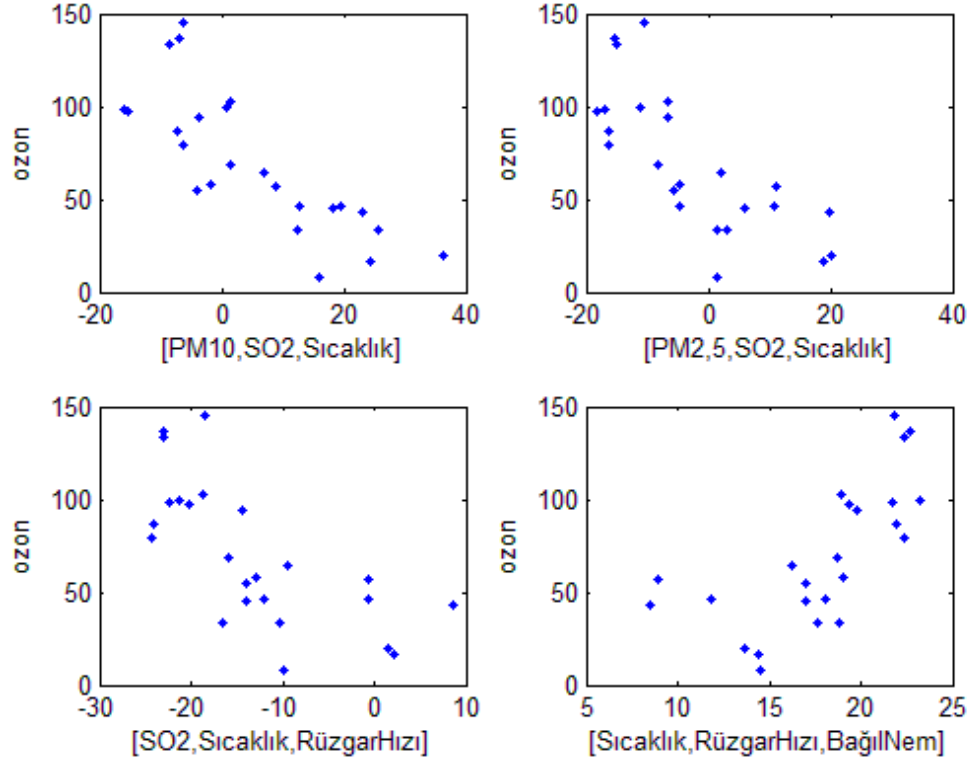
EşikDeęer =15		EşikDeęer =20		EşikDeęer =25		EşikDeęer =30	
Deęişkenler	1.Görü	Deęişkenler	1.Görü	Deęişkenler	1.Görü	Deęişkenler	1.Görü
PM <sub>10</sub>	-0,77164	PM <sub>10</sub>	-0,59713	PM <sub>10</sub>	0,385393	PM <sub>2,5</sub>	-0,80623
SO <sub>2</sub>	0,154662	PM <sub>2,5</sub>	0,780826	PM <sub>2,5</sub>	0,430685	SO <sub>2</sub>	0,291319
BN	0,616974	SO <sub>2</sub>	-0,18369	BN	-0,81608	NO	0,514908



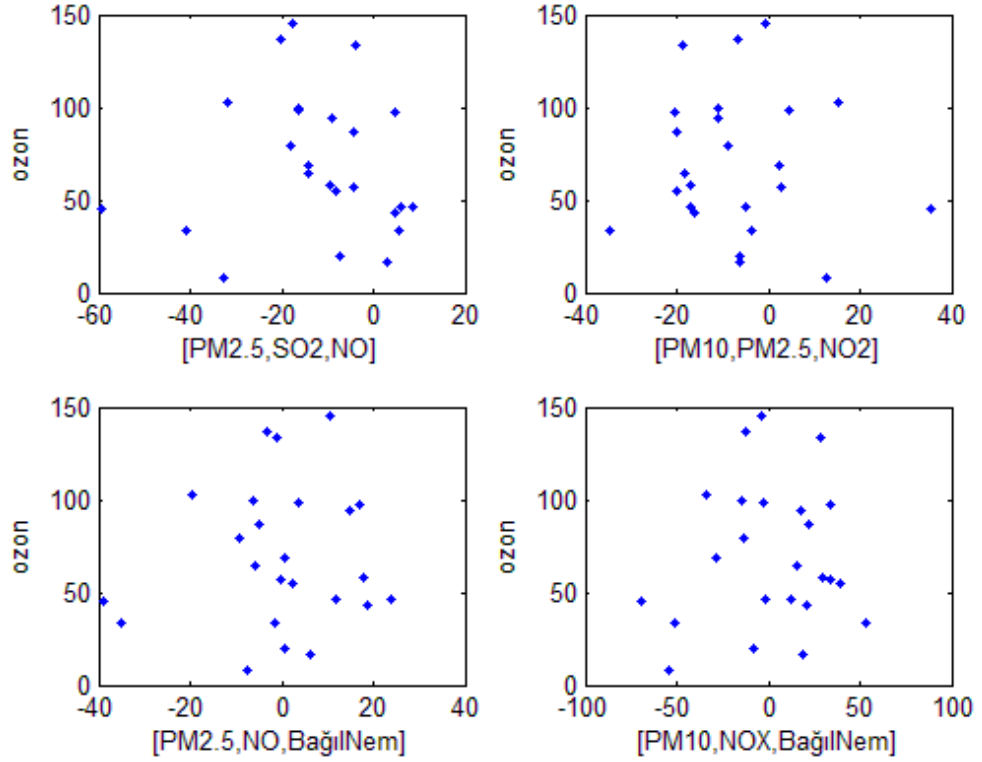
Şekil 4.9: Ozonun ilişkili olduğu değişken grupları ile bağıntısının görünümü (th=15).



Şekil 4.10: Ozonun ilişkili olmadığı değişken grupları ile bağıntısının görünümü (th=15).

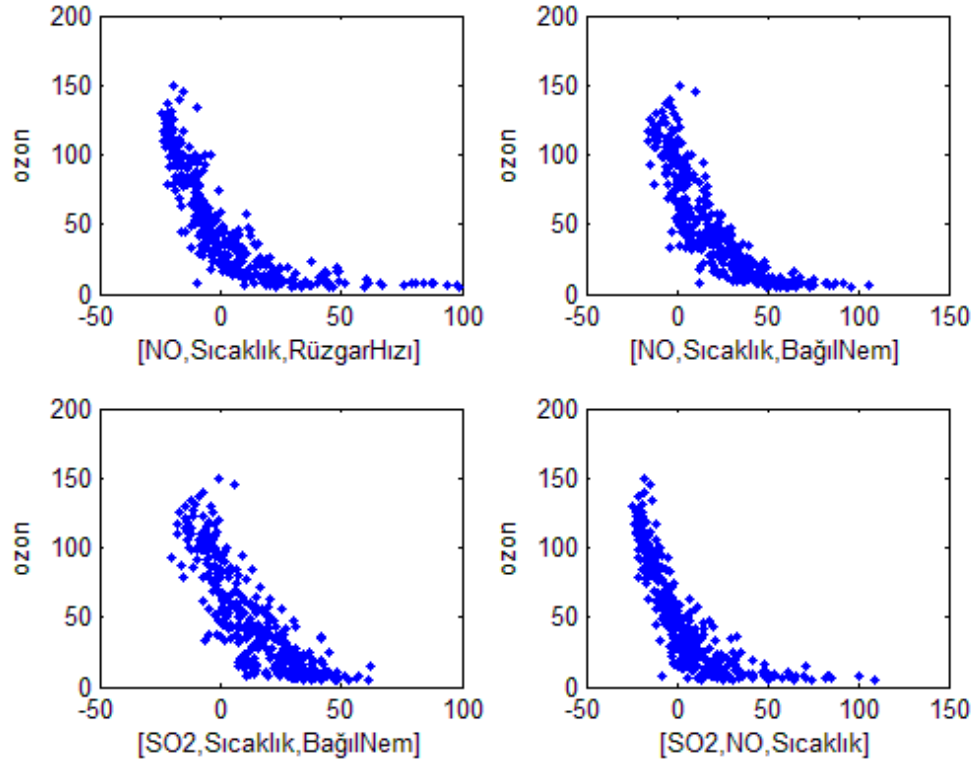


Şekil 4.11: Ozonun ilişkili olduğu değişken grupları ile bağıntısının görünümü (th=30).

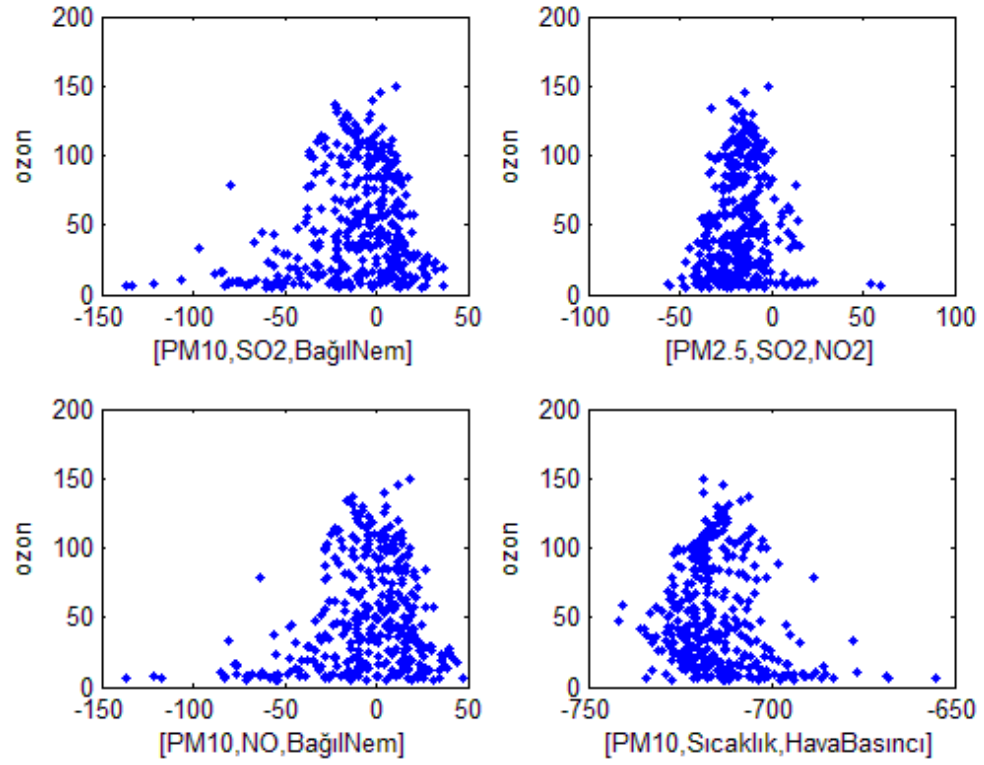


Şekil 4.12: Ozonun ilişkili olmadığı değişken grupları ile bağıntısının görünümü (th=30).

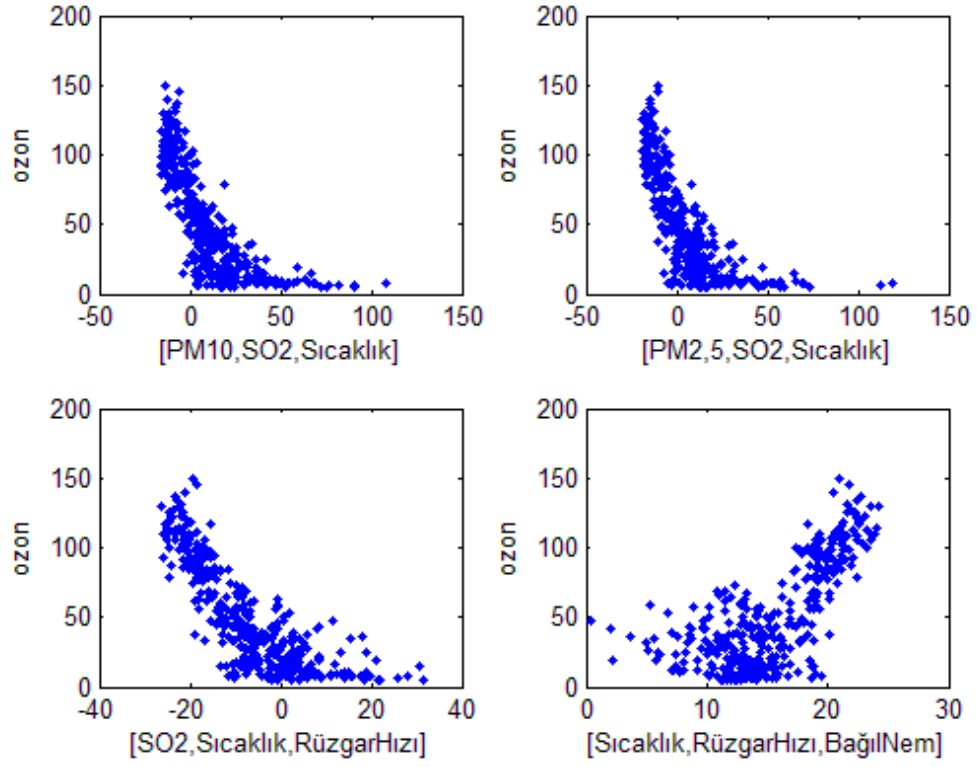




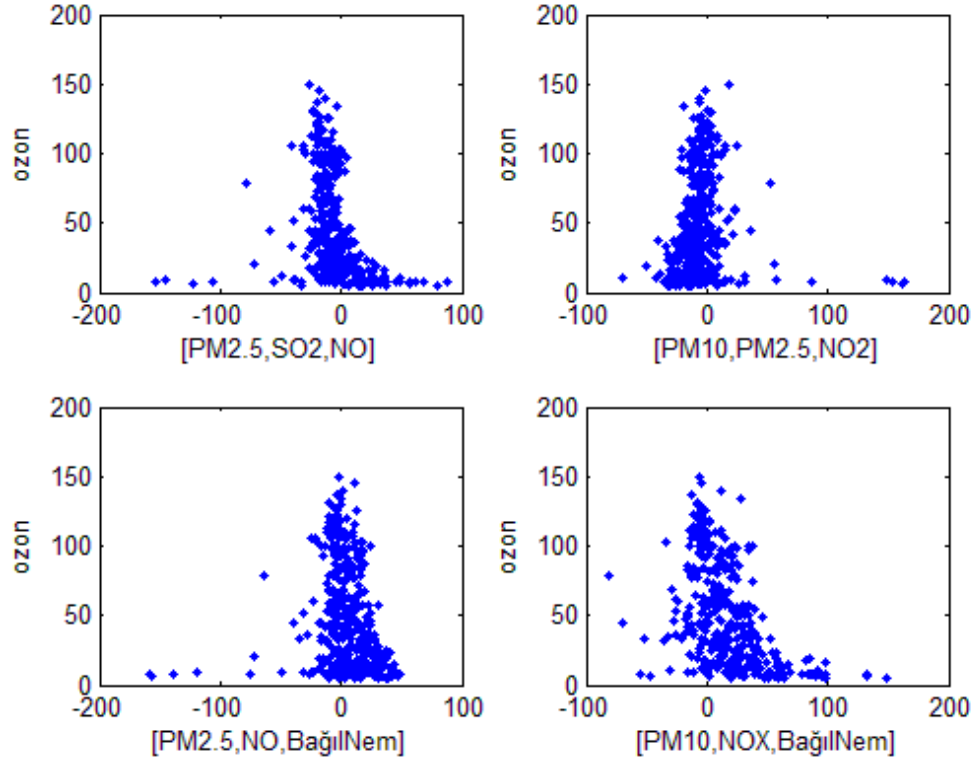
**Şekil 4.13:** Ozonun ilişkili olduğu değişkenlerden elde edilen katsayıların tüm veri setine uygulanması ile elde edilen ilişkinin görünümü (th=15).



**Şekil 4.14:** Ozonun ilişkili olmadığı değişkenlerden elde edilen katsayıların tüm veri setine uygulanması ile elde edilen ilişkinin görünümü (th=15).



**Şekil 4.15:** Ozonun ilişkili olduğu değişkenlerden elde edilen katsayıların tüm veri setine uygulanması ile elde edilen ilişkinin görünümü (th=30).



**Şekil 4.16:** Ozonun ilişkili olmadığı değişkenlerden elde edilen katsayıların tüm veri setine uygulanması ile elde edilen ilişkinin görünümü (th=30).

Eşik değeri 15 seçildiğinde ozonla bağımlı deęişkenlerin KBA ile izdüşümlerinin ozonla olan ilişkisi Şekil 4.9'da, ozonla bağımlı olmayan deęişkenlerin KBA ile izdüşümlerinin ozonla ilişkisi ise Şekil 4.10'da görölmektedir. KBA'nın ürettięi katsayıların veri setinde ki tüm günlere uygulanması ile ozonla bağımlı ve bağımlı olmayan deęişkenlerin ozonla ilişkileri sırasıyla Şekil 4.13 ve Şekil 4.14'te görölmektedir.

Eşik değeri 30 seçildiğinde Şekil 4.11 ve Şekil 4.12 sırasıyla ozonla bağımlı ve bağımlısız deęişken gruplarının KBA ile izdüşümlerinin ani deęişim günlerinde ozonla deęişimini göstermektedir. Şekil 4.15 ve Şekil 4.16 ise KBA'nın ani deęişim günlerini kullanarak elde ettięi katsayıların tüm veriye uygulanmasını göstermektedir.

## 5. TARTIŞMA VE SONUÇ

Zaman serisi analizlerinde, deęişkenler belirli zaman aralıklarında kaydedilerek zamanla nasıl deęiştikleri incelenmektedir ve birbiri ile bağlantılı birçok deęişkenin biraraya getirilmesi ile zaman serisi verisetleri oluşturulmaktadır. Bu veri setlerinin analizleri sonucunda deęişkenlerin birbirleri ile ilişkileri nicel olarak derecelendirilmektedir. Böylece ilgilenilen deęişkenin etkilendięi dięer deęişkenler tespit edilebilmekte ve buna göre karar verme mekanizmaları geliştirilmektedir. Bu tez kapsamında da ozon zaman veri serisi kullanılarak ozonun en çok etkilendięi deęişkenler tespit edilmiştir. Ozon seviyesini ani deęişimlerinin veri setinde bulunan kirleticiler ve meteorolojik faktörlerin oluşturduęu alt kümelerin hangileri ile bağlantılı olduęu deęişik teknikler kullanılarak gösterilmiştir.

Troposferik ozon, insan saęlığı ve çevre üzerindeki zararlı etkilerinden dolayı çalışılmaya devam edilen bir konudur. Ozon yoğunluk haritaları, geleceęe yönelik ozon tahmin sistemleri ve ozonun etkilendięi deęişkenlerin kestirimi bu konuda çalışılan genel başlıklardır. Özellikle büyük şehirlerde ozon tahmin sistemlerinin geliştirilmesi solunum yolu rahatsızlıkları olanlar için çok önemlidir. Ankara Türkiye'nin en önemli ve yoğun nüfuslu büyük şehirlerinden olmasına karşın bu il için böyle bir tahmin sistemi üzerinde henüz çalışılmamıştır. Bu açığa yönelik Ankara ili için ozon tahmin modelleri geliştirilmiştir ve ozonun ani deęişimine neden olan deęişken grupları tespit edilmiştir. Ozonun etkilendięi deęişkenlerin belirlenmesi, bu deęişkenleri atmosfere salan kaynakların tespiti için önemlidir. Dięer bir önemli nokta ise tespit edilen bu deęişkenlerin zaman içindeki tutumu incelenerek ozon seviyesi için bir kestirimde bulunulabilir. Bu tez çalışmasında troposferik ozon seviyesinin seçilen dört eşik deęer üzerine çıkmasına neden olan deęişken gruplarının tespiti için makine öğrenmesi yöntemlerinden kanonik bağıntı analizi, isomap ve doğrusal ayırtaç yöntemi kullanılmıştır. Bu tekniklerden kanonik bağıntı analizi ve isomap gözetimsiz öğrenme yapmakta, doğrusal ayırtaç yöntemi ise gözetimli öğrenme yapmaktadır. Bu yöntemler

kullanılarak ozon seviyesini etkileyen deęişkenler üçerli alt gruplar halinde ele alınmış ve gruptaki deęişkenlerin beraberce ozon deęişimi üzerindeki etkisi incelenmiştir.

Oluşturulan tahmin modelinde günlük ortalamalar kullanılarak bir sonraki günün ortalama ozon seviyesi tahmini yapılmıştır. Önceki günün ozon seviyesini kullanmak tahminin başarısını büyük oranda etkilemektedir. Tahminin kesinliği için kullanılan karesel hata katsayısı (RMSE) günlük ortalamalar üzerinden yaklaşık olarak 13 bulunmuştur. Karesel hatanın düşük olması tahminin doğruluęu için yeterli düzeydedir (İnal, 2010). Kullanılan teknikler ile elde edilen ozon ile baęıntısı yüksek gruplar içerisinde nitrik asitler, sıcaklık, rüzgar hızı bulunmaktadır. Eşik deęerin yüksek olduęu durumlarda partikülür maddeler ve SO<sub>2</sub> de ozonun ani deęişimine etki etmektedir. Bu deęişkenler başka bölgelerde yapılan araştırmalar ile uyuşmaktadır ((Şakar, 2011), (İnal, 2010)). Dięer bir önemli nokta ise, genel olarak sadece birincil kirleticilerin olduęu gruplar ozon ile düşük baęıntıya sahiptir. Yüksek baęıntıya sahip deęişken grupları hem kirletici hem de meteorolojik faktörler ihtiva etmektedir. Buradan, ozon seviyesindeki deęişim ortamda biriken kirleticilerle beraber uygun meteorolojik faktörlerin oluşmasına da baęlı olduęu sonucu çıkarılabilir.

Ozonun geleceęe yönelik tahmini yapılırken en etkin deęişkenler NO, NO<sub>2</sub>, NO<sub>x</sub> ve rüzgar hızı olduęu görölmektedir. Ancak ani deęişimlerin gerçekleştięi günler arasında ani artış mı yoksa ani azalma mı olduęu kestiriminde sıcaklık ve baęıl nemin etkinliği artmaktadır. Ozon seviyesindeki daha yüksek ani deęişimlerin olduęu günler ele alındığında ise artma/azalma sınıflandırma başarısı ile özellikle partikül madde ve SO<sub>2</sub> deęişkenleri daha yüksek baęıntı göstermektedir. Ozon seviye kestiriminin mevsimsel ve genel deęişiklikleri yakaladıęı için o tip bir çalışma ile seçilen etkili deęişkenlerin ani deęişiklikleri kestirmede aynı derecede efektif kullanılmadıęı görölmüştür. Dolayısıyla, ozon seviyesinin kestirimi ve ozon seviyesindeki ani deęişimin hangi yönde olacaęının kestirimi gibi iki farklı çalışma sonucunda farklı deęişken gruplarının etkili olduęu gösterilmiştir.

## KAYNAKLAR

- ABDUL-WAHAB S.A., AL-ALAWİ S.M., 2002, Assesment and prediciton of tropospheric ozone concentration levels using artificial neural networks.*Environmental modelling & Software*.Vol. 17, pp. 219-228
- AL-ALAWİ S.M., ABDUL-WAHAB S.A., BACKHEIT C.S., 2008, Combining principal component and artificial neural networks for more accurate prediciton of ground-level ozone.*Environmental modelling & Software*.Vol. 23, pp. 396-403
- ALPAYDIN E., 2010, Yapay Öğrenme. *Boğaziçi üniversitesi yayınevi*, İstanbul, 978-605-423849-1
- CAPE J.N., 2008, Surface ozone concentrarion and ecosystem healt: past trends and a guide to future projections. *Science of the Total Environment*. Vol. 400, PP. 257-269
- CARROLL C., 2006, Canonical Correlation Analysis: Assessing Links Between Multiplex Networks. *Social Networks*. Vol.28, PP. 310-330
- CHALOULAKOU A., SAISANA M., SPYRELLIS N., 2003, Comperative assessment of neural networks and regression model for forecasting summertime ozone in Athens. *The Sience of the Total Environment*. Vol. 313, PP. 1-13
- CHEN C.-S., CHENG M.-Y., WU Y.-W.,2012, Siesmic assessment of school buildings in Taiwan using the evolutionary support vector machine inference system. *Expert Systems with Applicaitons*. Vol. 39(4), PP. 4102-4110
- DEMİR G., ALTAY G., ŞAKAR C.O., ALBAYRAK S., ÖZDEMİR H., YALÇIN S.,2008, Prediction and evaluation of tropospheric ozone concentration in İstanbul using artificial neural network modelling according to time parameter.*Journal of Scientific & Industrial Research*.Vol. 17, pp. 674-679
- DUEÑAS C., FERNÁNDES M.C., CAÑETE S., CARRETERO J. LİGER E., 2005, Stochastic model to forecast ground-level ozone concentration at urban and rural areas.*Chemosphere*.Vol. 61, pp. 1379-1389
- GÓMEZ-SANCHIS J., MARTÍN-GUERRERO J.D., SORIA-OLIVAS E., VILLA-FRANCHÉS J., CARRASCO J.L., VALLE-TASCÓN S.D., 2006, Neural networks for analysing the relevance of input variables in the prediction of tropospheric ozone concentration. *Atmospheric Environment*. Vol. 40, PP. 6173-6180

- GÜMÜŞ E., KURŞUN O., SERTBAŞ A., ÜSTEK D., 2012, Application of Canonical Correlation Analysis for Viral Integration Preferences. *Bioinformatics*. Vol.28(5), PP. 651-655
- HARDOON D.R., SZEDMAK S., SHAW-TAYLOR J., 2004, Canonical Correlation Analysis: An Overview with Application to Learning Methods. *Neural Networks*. Vol.16, PP. 2639-2664.
- İNAL F. 2010, Artificial Neural Network prediction of tropospheric ozone concentrations in İstanbul, Turkey. *Clean-soil, Air, Water*. Vol. 38(10), PP. 897-908
- JAMMALAMADAKA S.R., LUND U.J., 2006, The effect of wind direction on ozone levels: a case study. *Environmental and Ecological Statistics*. Vol. 13, PP. 287-298
- KURŞUN O., ALPAYDIN E., FAVOROV O.V., 2011, Canonical Correlation Analysis Using Within-Class Coupling. *Pattern Recognition Letters*. Vol.32(2), PP. 591-599.
- MENEZES J. M. P., BARRETO G. A., 2008, Long-term time series prediction with the NARX network: An empirical evaluation. *Neurocomputing*. Vol. 71(16-18), PP. 3335-3343.
- MESHGİNİ S., AGHAGOLZADEH A., SEYEDARABİ H., 2013, Face recognition using gabor-based direct linear discriminant analysis and support vector machines. *Computer & Electrical Engineering*. Vol. 39(3), PP. 727-745.
- OUARDA T.B.M.J., GIRARD C., CAVADIAS G.S., BOBÉE B., 2001, Regional Flood Frequency Estimation with Canonical Correlation Analysis. *Journal of Hydrology*. Vol.254, PP. 157-173
- PASTOR-BÁRCENAS O., SORIA-OLIVAS E., MARTÍN-GUERRERO J.D., CAMPS-VALLS G., CARRASCO-RODRÍGUEZ J.L., VALLE-TASCÓN S.D., 2005, Unbiased sensitivity analysis and pruning techniques in neural networks for surface ozone modelling. *Ecological Modelling*. Vol. 182, PP. 149-158
- PIRES J.C.M., ALVIM-FERRAZ M.C.M., PEREIRA M.C., MARTINS F.G., 2011, Prediction of ozone concentrations: Application of methodology based on darwin's theory of evolution. *Expert System with Applications*. Vol. 38, pp. 1903-1908
- PISONI, E., FARINA M., CARNEVALE C., PIRODDI L., 2009, forecasting peak air pollution levels using narx models. *Engineering Applications of Artificial Intelligence*. Vol. 22(4-5), pp. 593-602
- SOUSA S.I.V., MARTINS F.G., PEREIRA M.C., ALVIM-FERRAZ M.C.M., 2006, Prediction of ozone concentrations in oporto city with statistical approaches. *Chemosphere*. Vol. 64, pp. 1141-1149

- SCHLINK U., HERBARTH O., RICHTER M., DORLING S., NUNNARI G., CAWLEY G., PELIKAN E., 2006, Statistical models to assess the health effects and to forecast ground-level ozone. *Environmental modelling & Software*. Vol. 21, pp. 547-558
- SUÁREZ-SÁNCHEZ A., GARCÍA NIETO P.J., RIESGO FERNÁNDEZ P., DEL COZ DÍAZ J.J., 2011, Application of an SVM-based regression model to the air quality study at local scale in the Aviles urban area. *Mathematical and Computer Modelling*. Vol. 54, pp. 1453-1466
- SUBASI A., GURSOY M. I., 2010, EEG signal classification using PCA, ICA, LDA and support vector machines. *Expert Systems with Applications*. Vol. 37, pp. 8659-8666
- ŞAKAR C.O., DEMİR G., KURŞUN O., ÖZDEMİR H., ALTAY G., YALÇIN S., 2011, Feature selection for the prediction of tropospheric ozone concentration using a wrapper method. *Intelligent automation and Soft Computing*. Vol. 17(4), pp. 403-413
- ŞAKAR C.O., KURŞUN O., KARAALİ, A., ERDEM Ç.E., Feature Extraction for Facial Expression Recognition by Canonical Correlation Analysis. *IEEE 20<sup>th</sup> Signal Processing and Application Conference*. 18-20 Nisan 2012 Muğla, 1-3
- TEMİYASATHİT C., KİM S.B., PARK S.K., 2009, Spatial prediction of ozone concentration profiles. *Computational Statistics and Data Analysis*. Vol. 53, PP. 3892-3906
- TENENBAUM J. B., DE SİLVA V., LANGFORD J. C., 2000, A global geometric framework for nonlinear dimensionality reduction. *Science*. Vol. 290, PP. 2323-2326
- TOZSİN G. Yer seviyesi ozon kirliliğine neden olan parametrelerin istanbul ili için incelemesi. *Yüksek lisans tezi*. İstanbul, 2003.
- TURHAN B. Nonlinear dimensionality reduction methods for pattern recognition. *Yüksek lisans tezi*. İstanbul, 2004.
- Meteoroloji Genel Müdürlüğü, <http://www.mgm.gov.tr/arastirma/ozon-ve-uv.aspx>[online], ziyaret tarihi (24.03.2013)
- Hava Kalitesi İzleme, <http://www.havaizleme.gov.tr/Default.ltr.aspx>[online], ziyaret tarihi (24.03.2013)



## **ÖZGEÇMİŞ**

Ahmet DEVELİ 23.04.1986 tarihinde Bayburt ilinde doğdu. İlk öğrenimini Şair Zihni İlköğretim Okulunda, orta öğrenimini ise Bayburt Anadolu Lisesinde tamamladı. 2005 yılında Dumlupınar Üniversitesi, Mühendislik Fakültesi, Bilgisayar Mühendisliği bölümünde lisans eğitime başladı, 2010 yılında Erciyes Üniversitesinin aynı adlı bölümünden mezun oldu. Aynı yıl öğretim görevlisi yetiştirme programı (ÖYP) kapsamında Bayburt Üniversitesine atandı. 2011 yılının Şubat ayında İstanbul Üniversitesi, Fen Bilimleri Enstitüsü, Bilgisayar Mühendisliği Ana Bilim Dalı'nda Yüksek Lisans Eğitime başladı ve aynı yılın Mayıs ayında ise bu bölümde ÖYP kapsamında Araştırma Görevlisi olarak göreve başladı. Halen Araştırma Görevlisi olarak çalışmalarına devam etmektedir.