İSTANBUL TEKNİK ÜNİVERSİTESİ ★ FEN BİLİMLERİ ENSTİTÜSÜ

ZAMAN SERİLERİNİN ÖNGÖRÜSÜ İÇİN GKA TABANLI DVR METODLARI

YÜKSEK LİSANS TEZİ

Bahadır BİCAN

Bilgisayar Mühendisliği Anabilim Dalı

Bilgisayar Mühendisliği Programı

MAYIS 2014

İSTANBUL TEKNİK ÜNİVERSİTESİ ★ FEN BİLİMLERİ ENSTİTÜSÜ

ZAMAN SERİLERİNİN ÖNGÖRÜSÜ İÇİN GKA TABANLI DVR METODLARI

YÜKSEK LİSANS TEZİ

Bahadır BİCAN
(504111501)

Bilgisayar Mühendisliği Anabilim Dalı

Bilgisayar Mühendisliği Programı

Tez Danışmanı: Yrd. Doç. Dr. Yusuf YASLAN

MAYIS 2014

İTÜ, Fen Bilimleri Enstitüsü'nün 504111504 numaralı Yüksek Lisans Öğrencisi **Bahadır BİCAN**, ilgili yönetmeliklerin belirlediği gerekli tüm şartları yerine getirdikten sonra hazırladığı **"ZAMAN SERİSİ ÖNGÖRÜSÜ İÇİN GKA TABANLI DVR METODLARI"** başlıklı tezini aşağıda imzaları olan jüri önünde başarı ile sunmuştur.

Tez Danışmanı : **Yrd. Doç. Dr. Yusuf YASLAN**
İstanbul Teknik Üniversitesi

Jüri Üyeleri : **Doç. Dr. Zehra ÇATALTEPE**
İstanbul Teknik Üniversitesi

Yrd. Doç. Dr. Fatih AMASYALI
Yıldız Teknik Üniversitesi

Teslim Tarihi : **05 Mayıs 2014**
Savunma Tarihi : **28 Mayıs 2014**

Aileme,

ÖNSÖZ

Öncelikle yüksek lisans eğitimim boyunca her koşulda yardımını esirgemeyen, kapısını her çaldığımda yardımlarıyla karşılaştığım, bana her türlü desteği sunan, rehberliği ile bilimsel olarak eksik olduğum konuları tamamlamama yardımcı olan sevgili hocam Yrd. Doç. Dr. Yusuf YASLAN'a teşekkür ve şükranlarımı sunmak isterim.

Öğrenimim süresince derslerini aldığım ve tez süresince danıştığım, üzerimde emekleri olan tüm hocalarıma, aynı dersleri alarak tanışma fırsatı bulduğum, gece&gündüz çalıştığımız, isimlerini teker teker yazamayacağım İTÜ'deki değerli arkadaşlarıma ve öğrencilik süreci dahil mesleki yaşantımda da yanımda olan, tez aşamasında çeşitli konularda yardımcı olan Y. Müh. Kd. Ütğm. Ümit SOYLU ve Dz. Kd. Ütğm. Yücel AKTAŞ'a teşekkürlerimi borç bilirim.

Ayrıca, tüm hayatım boyunca her zaman yanımda olan, bana her türlü desteği sağlayan, hiçbir zaman desteklerini esirgemeyen sevgili aileme sonsuz teşekkür ederim.

Mayıs 2014

Bahadır BİCAN
(Deniz Subayı)

İÇİNDEKİLER

Sayfa

KISALTMALAR

DAr : Doğru Artış
DAz : Doğru Azalış
DVM : Destek Vektör Makineleri
DVR : Destek Vektör Regresyonu
DVS : Destek Vektör Sınıflandırıcısı
GKA : Görgül Kip Ayrışımı
GSMH : Gayri Safi Milli Hasıla
IMF : İçkin Mod Fonksiyonu
İMKB : İstanbul Menkul Kıymetler Borsası
KKT : Karush-Kuhn-Tucker
KOSPİ : Güney Kore Borsa Endeksi
OKH : Ortalama Karesel Hata
OMHY : Ortalama Mutlak Hata Yüzdesi
TMSF : Tasarruf Mevduatı Sigorta Fonu
VC : Vapnik-Chervonenkis
YD : Yön Doğruluğu
YSA : Yapay Sinir Ağları
YRM : Yapısal Risk Minimizasyonu
YTF : Yarıçap Temelli Fonksiyon

ÇİZELGE LİSTESİ

Sayfa

ŞEKİL LİSTESİ

Sayfa

ZAMAN SERİLERİNİN ÖNGÖRÜSÜ İÇİN GKA TABANLI DVR METODLARI

ÖZET

Zaman serileri, bir değişkene ait zamana bağlı gözlemlerin eşit zaman aralıklarında oluşturduğu serilerdir. Zaman serisi öngörüsü ise, geçmişe dair gözlem verileri bilinen bir olayın geleceğe dair ne olabileceği hakkında tahminde bulunmanın kavramsal modelidir. Bu öngörüler ekonomi, endüstri gibi birçok alanda yapılacak yatırımlar ya da verilecek kararlar için büyük önem taşımaktadır. Geleceğe ait yüksek başarımlı öngörüler, zaman serilerinde bulunan geçmişe ait verilerin davranışlarının doğru tanımlanması ile gerçekleştirilir. Bu sebeple zaman serisi öngörüsü için kullanılan matematiksel model, zaman serisinin özelliklerine ve içindeki değişkenlere bağlı olmalıdır. Doğru ve yüksek başarıma sahip bir öngörü elde edebilmek için modellemenin yapısı önemli bir yer teşkil eder.

Zaman serisi öngörüsünün doğruluğu bir çok alanda verilecek kararlar ya da yapılacak yatırımların geleceğini ve büyüklüğünü etkilemektedir. Bu durum zaman serisi öngörüsünü önemli bir konu haline getirmiş, daha doğru ve güvenilir öngörü için bir çok teknik ve uygulama geliştirilmiştir. Literatürde zaman serisi öngörüsü için bir çok çalışma bulunmaktadır. Bu çalışmalar kullanım alanları ve tahmin başarımları açısından farklılık gösterebilmektedir. Ancak bütün öngörü çalışmalarında ortak olarak ortaya çıkan unsur belirsizliktir. Belli bir zamandaki şartlar biliniyor ve bilgiler doğru olarak toplanabiliyor ise öngörü işlemi daha güvenilir ve doğru olarak gerçekleştirilebilir. Ancak şartların ve etkenlerin bilinmediği bir ortamda öngörü işlemi zorlaşacaktır.

Bu tez çalışmasında, zaman serisi öngörüsü için daha önce kullanılan modeller incelenmiş ve başarılı modellerin başarımı artıran özellikleri tek bir modelde kullanılmak üzere birleştirilmiştir. Tez kapsamında oluşturulan modellerde, doğrusal ve doğrusal olmayan zaman serileri üzerinde yüksek bir tahminleme başarımına sahip olan Destek Vektör Regresyonu (DVR) algoritması kullanılmıştır.

Öncelikli olarak, DVR tahmin başarımını en yüksek seviyede tutmak için kullanılacak parametreler ve çekirdekler (kernel) incelenmiştir. Ayrıca zaman serisi özelliğine bağlı olarak öznitelik seçiminin tahmin başarımına etkisi araştırılmış, hangi özniteliklerin kullanılması gerektiği belirlenmiştir.

Son yıllarda, tekli algoritmalara ek olarak tahminleme başarımını artırmak için bütünleşik modeller de sıkça kullanılmaktadır. Bu çalışmada DVR tahmin başarımını artırmak için Görgül Kip Ayrışımı (GKA) tabanlı DVR modelleri üzerinde çalışılmıştır. Daha önce önerilmiş GKA tabanlı DVR modellerinin çoğunun test kümesini de eğitim kümesi ile birlikte kullandığı ve GKA'nın son nokta problemini görmezden geldiği saptanmıştır. Gerçek uygulamalarda kullanılamayacak olan bu yöntem yerine GKA tabanlı DVR modeli üzerinde 2 yaklaşım önerilmiştir.

Bu çalışma kapsamında önerilen ilk modelde, GKA ile yinelemeli (iterative) olarak elde edilen İçkin Mod Fonksiyonları (IMF) üzerinde birbirinden bağımsız olarak DVR ile tahminlemeler yapılmış, daha sonra bileşen tahminleri toplanarak genel tahmin elde edilmiştir. İkinci modelde ise veri kümesi GKA ile gürültüden arındırılmış ve elde edilen yeni veri kümesi asıl (orjinal) veri kümesi ile birlikte öznitelik olarak kullanılarak tahmin sonuçları alınmıştır.

Önerilen modellerin tahminlemeleri farklı alanlardan, farklı büyüklükte ve özellikte 7 veri kümesi üzerinde Ortalama Karesel Hata (OKH), Ortalama Mutlak Hata Yüzdesi (OMHY), Yön Doğruluğu (YD), Doğru Artan (DAr) ve Doğru Azalan (DAz) ölçütleri kullanılarak tekli DVR modeli tahminlemeleri ile karşılaştırılmıştır. İlk model olan yinelemeli GKA-DVR modelinin periyodiklik özelliği taşıyan veri kümeleri üzerinde yön tahminlerinde DVR'den daha başarılı sonuçlar verdiği, ikinci modelin ise bütün veri kümeleri için noktasal başarı ölçümlerinde (OKH, OMHY) DVR tahminlemesinden daha başarılı olduğu görülmüştür.

EMD BASED SVR METHODS FOR TIME SERIES PREDICTIONS

SUMMARY

Time series are the series that belong to time dependent observations of a variable which are formed in equal time intervals. Time series prediction is the conceptual model of what might happen on the future for an event where the observation data about the past is known. These predictions have great importance for giving decisions or making investments on many areas such as economy or industry. The rising up or falling down of the values can support researchers, economists or investors while giving their important decisions. The ability to know the correct directions of the future values contributes to the management of daily/yearly operations. High performance predictions can be performed by the correct identification of the time series behavior which contain historical data. For this reason, the mathematical models used for time series prediction, depends on variables and characteristics of the time series. In order to obtain higher success and accurate predictions, modeling structure constitutes an important topic.

The accuracy of time series prediction affect the future of the decisions or investments to be made in many areas. This case has made the time series prediction a major issue and lots of techniques and applications have been developed. for more accurate and reliable forecasting. In the literature, there are lots of studies for time series prediction and different machine learning algorithms have been utilized for time series forecasting. These studies may vary in terms of usage and performance of prediction. However, a common element of all these studies is the uncertainty. If the conditions at a particular time are known and the data can be collected accurately, forecasting process can be applied more reliable and accurate. However, in an environment where conditions and factors are not known, the prediction process will be difficult.

In this thesis study, the models that are used before are analyzed and the enhancing characteristics of successful models are combined in order to use them in one single

model. In the constructed models, Support Vector Regression (SVR) is used which has a higher prediction performance on time series without distinguishing the datasets are linear or nonlinear. In order to obtain highest performance, parameters and kernels to be used are analyzed firstly. Then, the effect of the feature selection depending on time series are analyzed and which features should be used are determined.

In recent years, hybrid models are commonly used besides single prediction algorithms to improve prediction performance. In this study, Empirical Mode Decomposition (EMD) based SVR models have been studied in order to improve the performance of the SVR predictions. EMD is a nonlinear signal processing method which divides the data into several simple subtasks. It offers a new way to deal with non-linear and non-stationary signals and decomposes original signal into several Intrinsic Mode Functions (IMF) and a residue component. It's realized that the most of the previously proposed EMD based SVR algorithms use the test dataset with the training dataset and ignoring the end point problem of the EMD. Instead of this method, which can not be used for real applications, 2 approaches have been proposed.

In the first proposed model, the predictions with SVR on each iteratively obtained component by EMD have obtained independently, and then they are summed to get the final prediction. In the second model, the noise reduction is performed with EMD, and final predictions are obtained with the noise-free data set which is used with the original dataset as a feature.

The prediction performance of proposed models are compared with single SVR model by using Mean Square Error (MSE), Mean Absolute Percentage Error (MAPE), Direction Accuracy (DA), Correct Up (CP) and Correct Down (CD) performance citerias on 7 different datasets from different areas.

In the experimental results, it's seen that the first proposed iterative EMD-SVR method is compared with single SVR. It is observed that the proposed method outperforms the single SVR performance on direction measurements including Direction Accuracy, Correct Up and Correct Down trends for the datasets which have periodicity characteristics.

In the second model, it’s seen that the proposed model has better performance than SVR on pointwise (MSE, MAPE) performance criterias for all data sets that have been used.

1. GİRİŞ

Zaman serisi, ilgilenilen bir konunun geçmişe dair değerlerinin sıralı ölçüm kümesidir. Örnek olarak, bir ülkedeki para biriminin başka bir para birimine oranı, bir bölgeye gelen turist sayısının miktarı, bir şehirdeki nüfus artış/azalışı, altın fiyatlarının günlük kapanış değeri veya bir nehirin yıllık akış hacmi (debisi) verilebilir. Gözlenen verilerin düzenli zaman aralıklarında olması doğru analiz açısından önemlidir. Zaman aralıkları her seride farklı olabilir. Buna örnek olarak saatlik, günlük, aylık, haftalık ya da daha farklı zaman aralıklarına sahip zaman serileri gösterilebilir.

Zaman serisi öngörüsü ise geçmişe ait verilerin analizi ile geleceğe ait değerlerin ne olacağını bulmayı amaçlar. Örneğin altın fiyatının önceki kapanış değerlerine bakarak bir gün sonraki kapanış fiyatını öngörmek finans alanında zaman serisi tahminidir.

Zaman serisi öngörüsü ekonomi, biyoloji, bilgisayar bilimleri, mühendislik, matematik, endüstri, iş dünyası gibi farklı alanlarda kullanılmaktadır. Öngörü için kurulan zaman serisi modeli geçmişe dair istatiksel veri üzerine kurulur ve bir ürünün ne kadar daha üretileceği, şehrin ne kadar daha büyüyeceği gibi stratejik planlamalarda karar mekanizması sağlar. Örneğin bir şehirdeki elektrik tüketiminin 10 yıl sonra ne olacağı hakkında bir fikrimiz varsa, yeni bir elektrik santraline gereksinimin olup olmayacağı planlanabilir.

Zaman serisi üzerinde yapılan öngörü işlemleri geçmiş veriyi kullandığından, verilerin doğruluğu önem kazanmaktadır. Doğru veriler üzerinde yapılan analiz ile geçmişteki davranışlar saptanır ve böylece tahminin daha doğru yapılması amaçlanır. Bir zaman serisinin geleceği, kendi geçmiş verileri ile öngörülebileceği gibi, başka değişkenlere de bağlı olabilir. Verinin nelerden etkilendiğini bilmek, öngörüyü kolaylaştıracaktır. Örneğin, elektrik tüketimi aynı zamanda hava sıcaklığına da bağlıdır ve kış mevsimlerinde tüketim artar. Etkenleri kolay saptanabildiğinden elektrik tüketimini tahmin etmenin başka alanlardaki veri kümelerine göre daha

kolay olduğu söylenebilir. Herhangi bir etkene bağlı olmayan ya da etkenleri zor belirlenen verilerde ise öngörü daha da zorlaşır. Örneğin finansal veriler (GSMH, İMKB endeksi) farklı durumlarda farklı değişkenlerden etkilenebilir. Finansal bir verinin etkenleri belirlenmiş olsa bile dikkat edilmesi gereken husus, bu etkenlerin de zaman içinde değişebileceğidir. Bu yüzden öngörü için oluşturulan model, elektrik öngörüsünde kullanılan hava sıcaklığı gibi sadece sabit etkenler üzerine değil, zamanla değişebilecek değişkenler/etkenler göz önüne alınarak kurulmalıdır.

t anındaki bir zaman serisi, $x(t)$ ile gösterilir. Öngörüde amaç; elde bulunan $x(1), x(2), \dots, x(t)$ verilerini kullanarak gelecek verilerin $x(t+1), x(t+2)$ değerlerini bulmaktır.

Zaman serisi analizlerinde kullanılan en önemli etken, serinin durağan (stationary) olup olmadığıdır. Bu ayrım, serideki değerlerin ortalama ve varyansına göre yapılmaktadır [1]. Serideki değerlerin ortalama ve varyansı zaman içerisinde farklılık göstermiyor ve belirli zaman aralıklarında alınan zaman serilerinin kovaryansları değişmiyor ise seri durağandır. Ortalama ve varyansın zamana bağlı olarak değiştiği seriler ise durağan olmayan (non-stationary) serilerdir.

Zaman serisi analizinde durağan serilerin tahminlemesi daha öngörülebilir durumdadır. Ancak durağan olmayan serilerde verilerin davranışı hakkında genelleme yapılamadığından serinin geleceğine dair tahminde bulunmak daha zordur.

1.1 Zaman Serisi Bileşenleri

Zaman serilerini oluşturan farklı unsurları tanımak, doğru analizler yapmayı sağlar. Zaman serisinin özelliklerini açıklayabilmek için seri içindeki etkili bileşenlerin belirlenmesi gerekir. Bileşenlerin belirlenmesi ve bunların gelecekteki etkilerinin saptanarak tahminlemenin daha doğru yapılması, kısa ve uzun süreli planlarda büyük yararlar sağlar. Zaman serisi verileri dört bileşenden oluşur [1];

1.1.1 Eğilim (Trend)

Zaman serisindeki değerlerin uzun bir süreç için gösterdiği artma ya da azalma durumudur. Verinin artış ya da azalış eğilimlerinin tahmin edilmesi, yatırım ya da tüketim gibi alanlar için önemlidir.

1.1.2 Mevsimsellik

Zaman serisindeki verilerin mevsimlere göre değişkenlik gösterdiği durumdur. Bir yıl ya da daha kısa süreli zaman serilerinde gözlemlenebilir. Mevsimin etkisi ile değerler artış ya da azalış gösterebilir. Örnek olarak yaz mevsiminde kömür satışlarının azalması, bir ülkedeki elektrik kullanımının kış mevsiminde artması ya da yaz mevsiminde dondurma satışlarının artması gösterilebilir.

1.1.3 Çevrimsel Bileşen

Daha çok ekonomi üzerinde, mevsimsellik etkisine bağlı olmayan, belirli bir zaman aralığında gözlenen değişimlerdir. Örneğin, ekonomide kısa süreli yaşanan ekonomik krizler ya da sektörlerin kârındaki refah ya da daralma dönemleri örnek olarak gösterilebilir.

1.1.4 Düzensiz Bileşen

Diğer unsurlar gibi belirli olmayan, doğal ya da sosyo-ekonomik nedenlerle ortaya çıkabilen değişmelerdir. Doğal afetlerin bir sektör ya da ekonomi üzerindeki etkisi örnek olarak gösterilebilir.

1.2 Zaman Serileri Üzerinde Yapılan Çalışmalar

Zaman serisi tahminlemesi üzerinde modellemeler halen akademik, endüstri, finans alanlarında gelişmekte olan bir konudur. Zaman içinde gelişen modellerin hacmi ciddi boyutlara ulaşmıştır [2].

Zaman serisi üzerinde çalışılan modellerden başlıcaları AR (Auto Regressive), ARMA (Auto Regressive Moving Average), ARIMA (Auto Regressive Integrated Moving Average) [3-4], ARCH (Autoregressive Conditional Heteroskedasticity) [5], Yapay Sinir Ağları (YSA) [6], Bulanık Mantık [7], Genetik Algoritma ve Destek Vektör Makineleridir (DVM).

Yapılan çalışmalarda ARIMA modelinin zaman serisi verilerde etkin bir yöntem olduğu gösterilmiştir [8]. Ancak AR, ARMA ve ARIMA modellerinin zaman serilerini durağan olarak kabul ediyor olması ve durağan olmayan modellerdeki düşük tahminleme başarısı göstermesi diğer algoritmaların kullanımını gerektirmiştir.

Bu modellere alternatif olarak kullanılan modellerden biri olan YSA; zaman serisinin belirli bir dağılıma sahip olduğunu varsaymayarak doğrusal olmayan zaman serileri üzerinde ARIMA modeline göre avantaj kazanmaktadır. M-Competition adlı yarışmada YSA modelinin, 104 zaman serisi kümesi üzerinde, içinde ARMA modeli de bulunan farklı 6 modelden aylık tahminlerde daha başarılı olduğu gösterilmiştir [9]. Ancak yıllık tahminlerde ARIMA modelinin YSA ile karşılaştırılabilir derecede iyi sonuçlar verdiği görülmüştür.

Doğrusal olmayan model olan YSA'nın tahminleme başarısının yüksek oluşu, halen bir çok zaman serisi tahminlemelerinde kullanılma sebebidir. Ancak YSA'nın yerel en küçük değere (minimum)/en büyük değere (maximum) takılma, büyük eğitim verisine gereksinim duyma, fazla uyum/eksik uyum (overfitting/underfitting) gibi problemleri mevcuttur. Literatürde bu problemleri ortadan kaldırmak için ise Destek Vektör Makinesi (DVM) tabanlı modeller kullanılmaktadır [10]. DVM yöntemi Vapnik [10] tarafından önerilmiş, yapısal risk minimizasyonu temelli ve çok yaygın olarak kullanılan bir yöntemdir. DVM'nin yerel en küçük değere takılmama garantisi, doğrusal olmayan (non-linear) ve durağan olmayan (non-stationary) veriler üzerindeki başarısı ve tahminlemedeki yüksek doğruluğu bir çok çalışma alanında kullanılma sebebidir.

Çeşitli çalışmalarda DVM ile elde edilen tahmin doğruluğunun diğer yöntemlerle elde edilen tahmin doğruluğundan daha üstün olduğu gösterilmiştir. Örneğin finansal bir veri kümesi olan KOSPI endeksi üzerinde yapılan çalışmada DVM modelinin daha iyi sonuçlar verdiği gösterilmiştir [11].

Algoritmaların karşılaştırılmasında çeşitli çalışmaların yanında, zaman serisi tahminlemesi üzerine yapılan yarışmalar da önemli katkı sağlamaktadır. Bu kapsamda 2001 yılında EUNITE adlı (European Network on Intelligent Technologies) Doğu Slovakya Elektrik Şirketi tarafından desteklenen elektrik yükü kullanımı tahminleme yarışması [12] zaman serisi modellerin karşılaştırmalarını sağlamıştır. Yarışmaya 21 ülkeden 56 yarışmacı farklı modeller ile katılım göstermiş, Chang ve arkadaşlarının [13] DVM yöntemi ile sunduğu model, aralarında YSA'nın da bulunduğu diğer modeller [14-15] arasında en iyi tahminleme yaparak yarışma birincisi olmuştur. Bu yarışmada dikkat çeken bir diğer durum ise, iyi bir tahminlemede iyi bir algoritma kullanmanın yanında, verilerin kullanılma şeklinin de modellemede önemli olduğudur. Buna örnek olarak, bu yarışmada başka DVM

modellerinin de olduğu, ancak tahminleme başarısının Chang ve arkadaşlarının [13] tahminleme başarısı kadar yüksek olmadığı gösterilebilir.

DVM halen rüzgar hızı [16], elektrik yükü [17], finans [18-19], ürün talebi, turist/ziyaretçi sayısı tahminlemeri gibi bir çok alanda kullanılmaktadır.

Son yıllarda, tekli tahmin yöntemlerinin uygulanmasının yanısıra, tahmin başarımını artırmak amacı ile bütünleşik modeller de sıkça kullanılmaktadır. Görgül Kip Ayrışımı (GKA) modelinin durağan ya da durağan olmayan verileri durağana yakın sonlu sayıda bileşenlere ayrıştırabilmesi ve her bileşenin kendi içinde farklı özelliklere sahip olması özelliği yaygın olarak kullanılmasını sağlamıştır [20]. Medikal, finans, su altı akustiği, sinyal işleme, akışkanlar mekaniği gibi birçok alanda kullanılmaktadır. GKA'nın bu özelliği aynı zamanda zaman serisi verileri üzerinde çalışılan modeller ile bütünleşik olarak kullanılması fikrini ortaya çıkarmıştır. Literatürde GKA tabanlı ARIMA [21], YSA [22-24], DVM [25] modelleri farklı alanlarda kullanılmıştır. Tekli modeller içinde DVM algoritmasının yüksek başarımlı oluşu sebebi ile sinyal işleme [26], finans, rüzgar hızı, uydu momentum tekeri sıcaklığı tahminlemesi [27] gibi bir çok alanda GKA tabanlı DVM algoritması üzerinde çalışmalar yapılmıştır.

Ancak GKA algoritmasının zaman serisi verileri ayrıştırmasında yaşanan son nokta (end point) problemi, GKA tabanlı bileşik modellerde tahminleme başarısını düşürmektedir. Zaman serisi tahminlemesi üzerinde GKA tabanlı çalışmaların büyük bir bölümü [21-27] bu sorun üzerinde durmayarak, gerçekte var olan problemlere çözüm getirmemiştir. Bu çalışmalar, veri kümesine ait test kısmını da eğitim kümesi ile birlikte kullandıklarından gerçekte var olan son nokta problemi ile karşılaşmamış, bu durumda yapılan tahminlemeler başarılı da olsa gerçek uygulamalarda kullanılamayan bir model ile elde edilmiştir.

GKA tabanlı DVM modeli için kullanılan bir diğer yaklaşımda ise, çok değişkenli (multivariate) zaman serilerinin modellenmesinde gürültü süzme için GKA kullanılmış ve gürültüsü süzülen değişkenler asıl (orjinal) veri kümesi ile birlikte kullanılmıştır. Öznitelik artırılarak elde edilen modelin diğer modellerden daha başarılı olduğu gösterilmiştir [5].

Bu tez çalışması kapsamında öncelikli olarak, zaman serisi tahminlemelerinde DVM algoritmasının farklı veri kümelerinde nasıl modellenmesi gerektiği incelenmiştir.

Modellemede kulllanılan parametrelerin ve özniteliklerin hangi veri kümeleri üzerinde etkili olduğu saptanmış, veri kümeleri üzerinde grafiksel ve matematiksel ifadeler ile gösterilmiştir.

GKA-DVM modelinde yapılan incelemelerde ise öncelikle halihazırda kullanılan yöntemler incelenmiş, aynı modelleme teknikleri kullanılarak bu tez çalışmasındaki veri kümeleri üzerinde sonuçları alınmıştır. Ancak bu yaklaşımın başarılı sonuçlarına rağmen neden gerçek uygulamalarda kullanılamayacağı açıklanmıştır. Ayrıca GKA tabanlı DVM algoritmasının gerçek zaman serisi uygulamaları için 2 adet model önerilmiştir. Önerilen modellerden ilkinde GKA ile elde edilen İçkin Mod Fonksiyonları tahminlemeleri üzerinde çalışılmıştır. İkincisinde ise GKA ile tek değişkenli zaman serilerinde gürültü süzme işlemi gerçekleştirilmiş ve gürültü süzülerek elde edilen değişkenler DVM modelinin oluşturulmasında kullanılmıştır. Test çalışmaları 7 adet veri kümesi üzerinde farklı başarım ölçütleri kullanılarak gerçekleştirilmiştir.

Bu tez çalışmasındaki ikinci bölümde Destek Vektör Makinalarının tanımlanması yapılmıştır. Üçüncü bölümde Görgül Kip Ayrışımı metodu ve bileşenlerin çıkarılış biçimi anlatılmıştır. Dördüncü bölümde tez çalışmasında kullanılan başarı ölçütleri ve veri kümeleri hakkında bilgiler verilmiştir. Beşinci bölümde Destek Vektör Regresyonu (DVR) algoritmasının zaman serisi verilerin özelliklerine göre nasıl modellenmesi gerektiği hakkında yapılan çalışmalar anlatılmıştır. Altıncı bölümde literatürde yer alan ve önerilen GKA tabanlı DVR modelleri elde edilen tahminlemelerin DVR ile karşılaştırmalı sonuçları yer almaktadır. Yedinci bölümde ise tezde yapılan çalışmaların sonuçları anlatılmıştır.

2. DESTEK VEKTÖR MAKİNELERİ

2.1 Tanımı

Destek Vektör Makineleri (DVM), Vladimir Vapnik ve arkadaşları tarafından önerilmiş eğiticili (supervised) bir öğrenme yöntemidir [10]. Temeli İstatistiksel Öğrenme Teorisi (Vapnik-Chervonenkis (VC) teorisi) ve yapısal risk minimizasyonuna (YRM) dayanmaktadır ve verilerin doğrusal olması ya da olmamasına göre bir ayırım yapmadan iyi bir genelleme yapmaktadır [28]. İstatistiksel Öğrenme Teorisi önceleri sadece verilen bir veri kümesine uyan fonksiyon tahmini probleminin teorik analizi için kullanılmış, daha sonra bu teoriye dayanan DVM yönteminin önerilmesiyle teorik analizlerin yanında, yeni öğrenme algoritmalarının oluşturulmasında da kullanılabileceği görülmüştür [10]. YRM ise en iyi genelleme performasını elde etmek için kullanılan bir optimizasyon yöntemidir.

DVM yönteminin akademik ortamda gerçek anlamda kullanılması, çekirdek (kernel) kullanımı ile farklı alanlarda diğer yöntemlere göre daha iyi sonuçlar verdiğinin farkedilişi ile başlamıştır [29]. Halihazırda DVM, sınıflandırma, regresyon ve yoğunluk tahmini çözümlemelerinde kullanılan güçlü bir yöntemdir.

DVM, veri seti dağılımı hakkında varsayımı ya da bilgisi olmaması özelliği ile diğer klasik yaklaşımlardan ayrılmaktadır. DVM ile eğitim kümesindeki girdi (input) ve çıktıların (output) eşlenmesi ile test kümesindeki girdileri sınıflayacak karar fonksiyonları elde edilir. Eğitim verisi üzerinde öğrenme işlemi gerçekleştirilerek yeni veri üzerinde tahminleme yapılır.

Karmaşık veri kümelerini sınıflandırabilmesi ve çözümlerde iyi bir başarı sergilemesi DVM'nin diğer modellere göre avantajları arasındadır. DVM üzerinde çalışılan sınıflandırma işlemleri DVS (Destek Vektör Sınıflandırma), regresyon işlemleri ise DVR (Destek Vektör Regresyon) olarak adlandırılır.

DVM'lerin avantajlarının yanında aynı zamanda dezavantajlarından da bahsedilebilir. Örneğin büyük veri sayısına sahip veri kümelerinde DVM'nin eğitim

işlemi (öğrenmesi) zaman alır. Ancak doğrusal ya da doğrusal olmayan veriler için ayırım yapmaması yaygın olarak kullanılmasını sağlar.

DVM'ler ile verilerin karmaşıklığı doğru bir şekilde belirlenebilir [30]. Doğrusal olarak ayrılabilen verilerde, verileri ayırabilecek ve sınıfların sınırları arasını en büyük değerde tutacak olan doğru seçilir. Doğrusal olarak ayrılamayan verilerde ise örnek uzayı, daha yüksek boyutlu bir örnek uzaya aktarılarak, örnekler arasındaki en büyük sınırın bulunması ile sınıflarına ayrılır. Diğer bir ifade ile, sınıfları en iyi ayırabilen hiper düzlemin tanımlanması işlemi gerçekleştirilir.

DVM en büyük değerli aralık sınıflandırıcıları arasındadır. Aralıklar arasındaki bağlantının kurulması önemlidir. n noyutlu bir uzayda verileri sınıflandırmak için aşağıdaki ifadede belirtilen bir ayırıcı fonksiyon kullanılır.

$$f(x) = w \cdot x_i + b = 0 \quad \textbf{(2.1)}$$

Bu fonksiyonda, w ağırlık vektörünü, b eğilim değerini ifade eder. Fonksiyon, ağırlık vektörünün (w) bulunması ile hesaplanabilir. n adet eğitim verisi olan bir veri kümesinde w ve b değerleri

$$|w \cdot x_i + b| = 1, \quad i = 1, \dots, n \quad \textbf{(2.2)}$$

olacak şekilde, ayırıcı düzleme en yakın veri noktaları değerlerini içeren bir formda düzenlenebilir. Veriler farklı sınıflardan olduğudan, bazı veriler eşitliği

$$w.x_i + b = +1 \quad \textbf{(2.3)}$$

durumunda sağlarken, farklı sınıfta olan diğer veriler

$$w.x_i + b = -1 \quad \textbf{(2.4)}$$

durumunda sağlar. Diğer bir anlatım ile $+1$ ve -1 değerleri sınıfları ifade eder.

Ağırlık vektörü değerinin küçük olması, sınır aralığının büyük değerde olmasını (istenilen durum) sağlar. Bu sebeple DVM sınıflandırıcısına en büyük (maksimum) sınır sınıflandırıcısı da denir. DVM, sınır aralıklarını en büyük değerde tutarak YRM prensibini uygulamaya çalışır. DVM üzerinde karşılaşılan eksik uyum sorunu ise farklı çekirdek fonksiyonları ile aşılır.

DVM üzerinde sınır aralığını belirlemek, çalışılan veri kümesine göre farklılıklar gösterir. Örneğin eğitim verilerinin doğrusal olarak ayrılabildiği ve üzerinde hata bulunmadığı durumlarda DVM, VC değerini en küçük değere ulaştırmaya çalışarak

beklenen risk olasılığını azaltır. Böylece doğrusal bir ayrıcı fonksiyon ile başarılı bir genelleme yapılır ve yüksek başarı elde edilmiş olunur.

Verilerin doğrusal olarak ayrılabildiği ancak eğitim verilerinin üzerinde bir miktar hatanın mevcut olduğu durumlarda ise DVM, yine iyi bir karar fonksiyonunun belirlenmesini amaç edinir. Bu durumda en iyi genellemeyi sağlayabilmek için ayırıcı düzlemin seçiminde bir miktar hataya izin verilir.

Doğrusal olmayan veri seti için DVM sınıflandırıcısı, sınıflandırma işlemini gerçek (orjinal) uzay üzerinde gerçekleştiremez. Bu nedenle veriler, çekirdek fonksiyonları ile gerçek uzaydan daha yüksek boyutlu bir uzaya aktarılır. Daha sonra yeni uzay üzerinde sınıflandırma işlemini yapmak üzere ayırıcı fonksiyon oluşturulur.

DVM, belirlenecek karar sınırı için, beklenen risk olasılığını olabildiğince en küçük değerde tutarak aralık değerini en büyük değerde tutar. Ancak büyük değerdeki aralık için risk olasılığı artmaktadır. Bu yüzden iyi bir genelleme için aralık değerinin en iyi şekilde (optimum) belirlenmesi önemlidir.

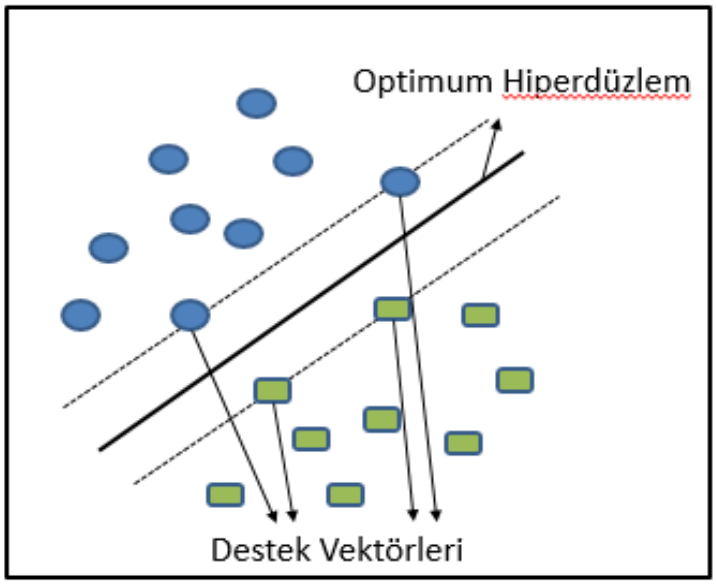
DVR metodunda ise verilerin karakteri mümkün olduğunca gerçeğe yakın bir şekilde yansıtma yapılır ve en iyi şekilde (optimum) ayırma işlemini yapan ayırıcı fonksiyonun bulunması amaçlanır. Sınıflandırmada olduğu gibi regresyonda da doğrusal olmayan verilerin işlenebilmesi için çekirdek fonksiyonları kullanılır.

2.2 Doğrusal Olarak Ayrılabilen Veriler için DVM (Sert Marjin)

DVM ile sınıflandırma işleminde oluşturulmaya çalışılan ayırıcı düzlem, eğitim verisi ile elde edilen bir karar fonksiyonu ile bulunur. Doğrusal ayrılabilen verilerde kullanılan karar fonksiyonu, eğitim verisini hatasız olarak ayırabilme özelliğine sahiptir [31]. Örneğin Şekil 2.1 (a)'da olduğu gibi n adet örnek barındıran bir veri setinde iki sınıfı ayırabilen birçok hiper-düzlem çizilebilir. Ancak DVM için asıl amaç, belirlenecek hiper-düzleme en yakın noktalar arasındaki uzaklığı en büyük değere (maksimum) çıkarmaktır [32]. Böylece karar fonksiyonu kullanılarak eğitim verisini en uygun şekilde ayırabilecek hiperdüzlem bulunur. Hiper düzlemden en yakın veri noktasına olan en küçük (minimum) uzaklığa sınır (marjin) denir. Daha büyük sınır, daha başarılı genelleme ile ayırma işlemi yapar.

Şekil 2.1.(b)'de görüldüğü üzere sınırı en büyük değere çıkararak en uygun ayrımı yapan hiper-düzleme optimum hiper-düzlem denir. Optimum hiper düzleme paralel

olarak çizilmiş iki doğru arasındaki uzaklığa ise sınır (marjin) adı verilir. Sınır üzerinde oluşan veri noktaları destek vektörleri adını alır. Destek vektörleri karar yüzeyine en yakın veri noktaları olduğu için en zor sınıflandırılan ve karar yüzeyinin yapısını belirleyen veri noktalarıdır.



Şekil 2.1 : Doğrusal olarak ayrılabilen veriler için ayırıcılar.

İki sınıflı, n elemeanlı bir eğitim verisinin doğrusal olarak ayrılabildiği ve tanımının

$$(x_1, y_1), \ldots, (x_n, y_n), x \in R^d, y \in \{+1, -1\} \quad \textbf{(2.5)}$$

olduğu kabul edilirse, optimum hiperdüzleme ait eşitsizlikler aşağıdaki şekilde olur:

$$w \cdot x_i + b \geq +1 \text{ her } y_i = +1 \text{ için} \quad \textbf{(2.6)}$$

$$w \cdot x_i + b \geq -1 \text{ her } y_i = -1 \text{ için} \quad \textbf{(2.7)}$$

Burada $x \in R^N$ olup N boyutlu bir uzayı, $y \in \{-1, +1\}$ ise sınıf etiketlerini, w ağırlık vektörünü (hiper-düzlemin normali) ve b eğilim (bias) değerini göstermektedir [32-33]. En iyi hiper düzlem, en iyi genellemeyi yaparak sınıflar arasındaki ayırma işlemini en iyi şekilde gerçekleştiren hiper-düzlemdir. Sınıfların sınırını oluşturan en iyi hiper düzeleme paralel düzlemler ise destek vektörlerinin bulunması ile çizilir.

Hiper düzleme ait eşitsizlikler tek eşitliğe indirmek istenilirse

$$w \cdot x_i + b \geq \pm 1 \quad \textbf{(2.8)}$$

şeklinde ifade edilirler.

Bu eşitlik, eğitim verileri içinde tanımlanabilecek hiper düzlemler için sağlanmış olur. Buradaki önemli kısım, hiperdüzlemin eğitim verileri içinde belirlenmesidir. Sınıflandırma probleminin doğrudan çözümü için hiper düzlem formülleştirilir. Eğitim verileri içindeki en iyi ayırıcı düzlemin fonksiyonu, daha sonra test verileri üzerinde kullanılır.

Sınırı/marjini en büyük değerde tutan en iyi hiperdüzlemin bulunması için eşitlik 2.7'deki w ve b değerlerinin hesaplanması gerekir.

$$w \cdot x + b = +1 \quad \textbf{(2.9)}$$

denklemini sağlayan örnekler için Şekil 2.2'deki H_1 hiperdüzleminin orijine dik olan uzaklığı

$$|1 - b| / \|\mathrm{w}\| \quad \textbf{(2.10)}$$

ve

$$w.x + b = -1 \quad \textbf{(2.11)}$$

denklemini sağlayan örnekler için Şekil 2.2'deki H_2 hiperdüzleminin orijine dik olan uzaklığı

$$|-1 - b| / \|\mathrm{w}\| \quad \textbf{(2.12)}$$

olacaktır. Bu durumda H_1 ve H_2 düzlemlerinin optimum ayırıcı hiper düzleme uzaklıkları, ya da destek vektörlerinin ayırıcı fonksiyona uzaklığı

$$1 / \|\mathrm{w}\| \quad \textbf{(2.133)}$$

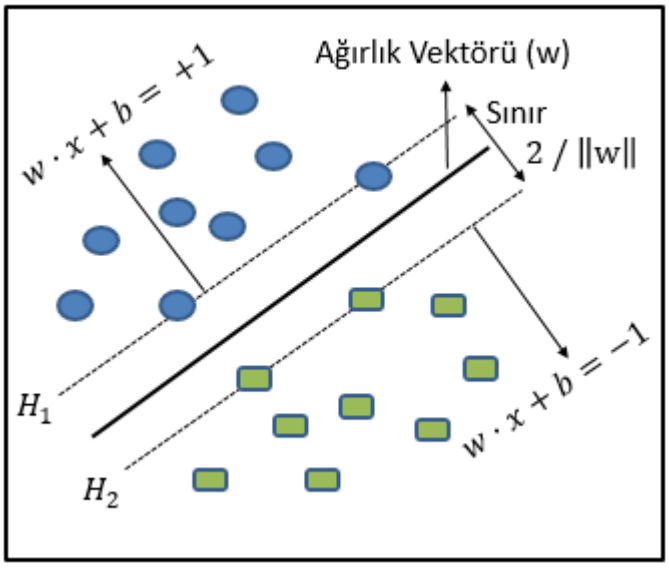
olur. Her iki düzlem arasındaki mesafe ise (belirlenen sınırın uzunluğu)

$$2 / \|\mathrm{w}\| \quad \textbf{(2.14)}$$

olarak tanımlanır. DVM'deki amaç, iyi bir genelleme yapabilmek için sınırın en büyük değere çıkarılmasıdır. Bunun için ise $\| w \|$ ifadesinin en küçük değer haline getirilmesi gerekir.

Bu durumda sınıfları en iyi şekilde ayırabilen hiperdüzlemin belirlenebilmesi için aşağıdaki sınırlı optimizasyon probleminin çözülmesi gerekir.

$$min\left[\frac{1}{2} \| w \|^2\right] \quad \textbf{(2.15)}$$



Şekil 2.2 : Doğrusal verilerin ayrımında kullanılan hiper düzlem.

Bu optimizasyon problemine bağlı sınırlamalar ise;

$$y_i\,(w.x_i + b) - 1 \geq 0 \; ve \; y_j \in \{-1, 1\} \qquad \textbf{(2.16)}$$

şeklinde ifade edilir [29]. Belirtilen eşitliklerdeki optimizasyon problemi, ikinci dereceden programlama tekniği olan Lagrange denklemleri kullanılmasıyla çözülebilir. Problemin Lagrange fonksiyonu ise,

$$L_P = \frac{1}{2} \| w \|^2 - \sum_{i=1}^{k} \left(\alpha_i \; y_i (w.x_i + \mathrm{b}) + \sum_{i=1}^{k} \alpha_i \right) \qquad \textbf{(2.17)}$$

şeklindedir.

Bu eşitlikte $\alpha_i \geq 0$ değerleri Lagrange çarpanları olarak adlandırılır. Burada amaç, daha önce belirtildiği gibi w değerlerini en küçük yapan noktayı bulmaktır. Bu aynı zamanda $\alpha_i \geq 0$ değerlerine göre en büyük yapan noktadır. Eşitlik 2.17, dışbükey (konveks) ikinci dereceden bir optimizasyon problemidir.

Bu problemi çözmek için ise; eşitlikteki w parametresini sadece α_i parametresi cinsinden ifade edebilmeyi sağlayacak olan Karush-Kuhn-Tucker (KKT) koşulları kullanılır. Bu işlemden sonra denklem, sadece Lagrange çarpanlarına göre maksimizasyon gerektiren ikili (dual) probleme dönüşür [34].

Bu problem için KKT koşulları:

$$\frac{\delta L_P}{\delta w} = 0 \text{ ise } w = \sum_i \alpha_i \; y_i \; x_i \qquad \textbf{(2.18)}$$

$$\frac{\delta L_P}{\delta b} = 0 \text{ ise } \sum_i \alpha_i \ y_i = 0 \quad \textbf{(2.19)}$$

şeklindedir. Optimizasyon probleminin ikili probleme dönüşmesi için belirtilen deklemler yerine yazılır ve problem ikili probleme dönüşmüş olur. Bu dönüşüm, yüksek boyutlu uzaylarda problemin daha kolay çözülmesine yardımcı olur [35]. Elde edilen ikili problem ise:

$$L_D = \sum \alpha_i - \frac{1}{2}\sum_{i,j} \alpha_i \alpha_j \ y_i y_j \ x_i x_j \ , \ \alpha_i > 0 \ \forall i \quad \textbf{(2.20)}$$

şeklindedir. Eşitlik 2.20'deki çözüm, L_P'nin minimizasyonu ya da L_D'nin maksimizasyonu ile çözülür. [36]. Burada her eğitim örneği için bir Lagrange çarpımı bulunur. Çözüm elde edildiğinde çoğu Lagrange çarpanlarının sıfır olduğu, yalnızca küçük bir kısmının da sıfırdan büyük olduğu görülür ($\alpha_i \geq 0$). Sıfırdan büyük olan Lagrange çarpanlarına karşılık gelen x_i örnekleri Destek Vektörleridir ve ayırıcı hiper düzleme paralel olan sınır doğrusu üzerinde yer alırlar.

Bu durumda en iyi (optimum) ayırıcı hiper düzlem:

$$f(x) = \sum(\alpha_i \ y_i \ x_i) + b = 0 \quad \textbf{(2.21)}$$

şeklinde belirlenmiş olur.

Ayrıca n veri kümesindeki örnek sayısını göstermek üzere , $E_n$beklenen hata (E_n) aşağıdaki sınırı sağlamaktadır [37]:

$$E_n[Hata\ Oranı] \leq \frac{E_n[Destek\ Vektörlerinin\ Sayısı]}{n}$$

2.3 Doğrusal Olarak Belirli Oranda Hata ile Ayrılabilen Veriler için DVM (Yumuşak Marjin)

Verilerin doğrusal olarak ayrılabildiği, ancak hatasız olmadığı durumlarda hatasız bir hiper ayırıcı bulunamaz. Veri kümesi gürültü içerdiğinde veri kümelerinde sınıflandırma için kullanılan sınır çok karmaşık olabilir. Bu karmaşıklığı azaltarak daha basit karar fonksiyonları elde etmek amacı ile gürültü kısmı gözardı edilerek daha basit karar fonksiyonları belirlenebilir. Bunu sağlamak ve optimum hiper düzlemi bulabilmek amacı ile ayırıcı düzlemin biraz hata yapmasına izin verilir. Hatalara izin verilmesinin amacı, küçük hataları tolere ederek daha yüksek genellemeye sahip olan ayırıcı hiper düzlemi bulabilmektir. Bunu sağlayabilmek için

ise esnek değişken (ξ_i) kullanılır. Sınır fonksiyonlarını çok fazla zorlayarak ayırıcı düzlemi çok karmaşık hale gitiren değişkenler gürültü olarak kabul edilerek gözardı edilir. DVM'nin yeniden tasarlanacağı bu yaklaşımda Eşitlik 2.6 ve Eşitlik 2.7 yeniden tanımlandığında:

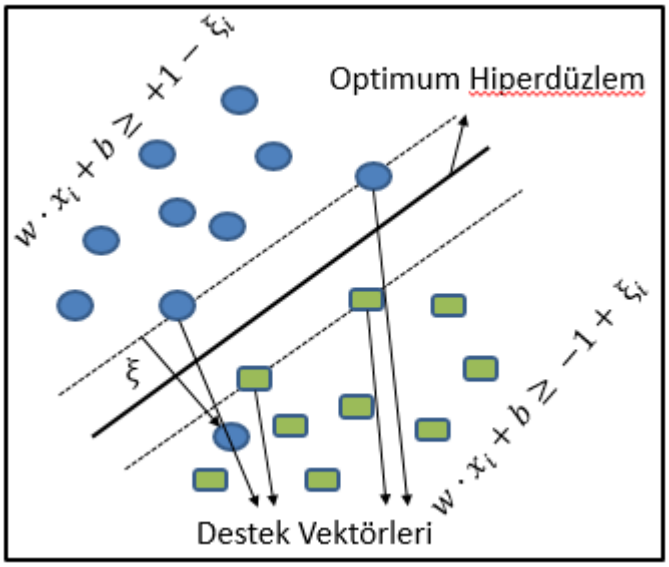
$$w \cdot x_i + b \geq +1 - \xi_i \text{ her } y_i = +1 \text{ için} \quad \textbf{(2.22)}$$

$$w \cdot x_i + b \geq -1 + \xi_i \text{ her } y_i = -1 \text{ için} \quad \textbf{(2.23)}$$

şeklinde olur. Burada (ξ_i) hata terimlerini ifade eder. (ξ_i) = 0 olması durumunda x_i örneği doğrusal olarak ayrılabilir durumda olduğu gibi doğru sınıflandırılır.

Bu yöntemde sınır içinde yer alan örneklerde iki tür sapma gözlenir. Birincisi, örneklerin ayırıcı hiperdüzlemin yanlış tarafında olduğu durumdur ve yanlış sınıflandırılırlar. ($\xi_i \geq 1$)

İkincisi ise, örneklerin ayırıcı hiperdüzlemin doğru tarafında olduğu ancak sınır bölgesinde yer aldığı durumdur. ($0 < \xi_i < 1$)



Şekil 2.3 : Doğrusal olarak ayrılamayan verilerde kullanılan parametreler.

Şekil 2.3'te doğrusal olarak ayrılamayan durumlarda optimum ayırıcı hiperdüzlem görülmektedir. Böyle bir durumda, sınırın en fazla haline (maksimum) getirilmesi ve yanlış sınıflandırma hatalarının en az haline (minimum) getirilmesi arasındaki denge C ile gösterilen ve ceza parametresi olarak aklandırılan bir parametre ile sağlanır. Ceza parametresi ve esnek değişken kullanılarak doğrusal olarak ayrılamayan veriler için optimizasyon problemi:

$$\min\left[\frac{1}{2} \| w \|^2 + C \cdot \sum_{i=1} \xi_i\right] \quad \textbf{(2.24)}$$

haline dönüşür. C parametresi ∞ değerini aldığında yanlış sınıflandırma yapılmasına izin verilmemesine sebep olur. Bu durumda problem sadece $0 < C < \infty$ tanımlaması ile kontrol edilebilir. C parametresi burada, sınırı en büyük değere getirme isteği ile en az hatalı sınıflandırma isteği arasında bir denge sağlar. C parametresinin alacağı değer, sonucu ciddi anlamda etkiler. C değerinin yüksek olması hatanın da yüksek olacağı anlamına gelir [36].

Bu yaklaşımda Lagrange formülasyonu:

$$L_P = \frac{1}{2} \| w \|^2 - \mathrm{C} \sum_{\mathrm{i}=1} \xi_i - \sum_{\mathrm{i}=1} \alpha_i \{ \mathrm{y}_i (\mathrm{x}_i w + b) - 1 + \xi_i \} - \sum_{\mathrm{i}=1} \mu_i \xi_i \quad \textbf{(2.25)}$$

şeklini alır. Bu denklemin çözümünün kolaylıştırılması, doğrusal olarak ayrılabilen verilerde olduğu gibi problemin ikili probleme dönüştürülmesi amacı ile KKT koşulları uygulanarak sağlanır. Bu durumda;

$$\frac{\delta L_P}{\delta w} = w - \sum_{\mathrm{i}} \alpha_i \mathrm{y}_i \mathrm{x}_i \quad \textbf{(2.26)}$$

$$\frac{\delta L_P}{\delta \mathrm{b}} = - \sum_{\mathrm{i}} \alpha_i \mathrm{y}_i = 0 \quad \textbf{(2.27)}$$

$$\frac{\delta L_P}{\delta \xi_i} = \mathrm{C} - \alpha_i - \mu_i \quad \textbf{(2.28)}$$

$$\mathrm{y}_i (\mathrm{x}_i w + b) - 1 + \xi_i \geq 0 \quad \textbf{(2.29)}$$

$$\xi_i \geq 0 \quad \textbf{(2.30)}$$

$$\alpha_i \geq 0 \quad \textbf{(2.31)}$$

$$\mu_i \geq 0 \quad \textbf{(2.32)}$$

$$\alpha_i \{ \mathrm{y}_i (\mathrm{x}_i w + b) - 1 + \xi_i \} = 0 \quad \textbf{(2.33)}$$

$$\mu_i \xi_i = 0 \quad \textbf{(2.34)}$$

ifadeleri elde edilir. Belirtilen ifadeler eşitlik 2.25'te yerlerine yazılırsa

$$L_D = \sum_i \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j \mathrm{y}_i \mathrm{y}_j \mathrm{x}_i \mathrm{x}_j \quad \textbf{(2.35)}$$

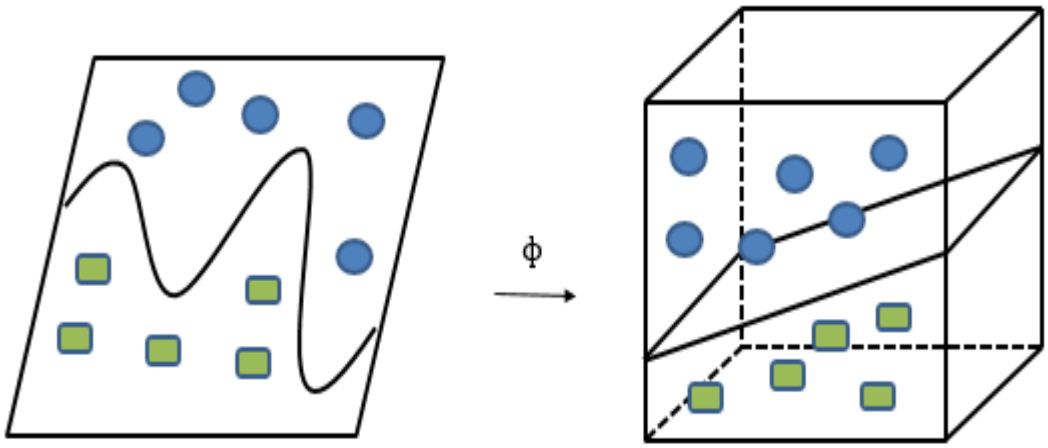
eşitliği elde edilir ve buna bağlı kısıtlar ise:

$$\sum_i \alpha_i \mathrm{y}_i = 0 \; ve \; 0 \leq \alpha_i \leq C, \forall i \quad \textbf{(2.36)}$$

şeklindedir. Bu çözümde $\alpha_i > 0$ olan ifadeler Destek Vektörlerin bulunmasını sağlarlar. Burada α_i değeri doğrusal olarak ayrılabilen verilerde yapılan çözümlemeden farklıdır. Doğrusal olarak ayrılabilen çözümlemede α_i değerinin sıfırdan büyük olma koşulu var iken, bu yaklaşımda α_i değerlerinin üst sınırının C olma koşulu vardır.

2.4 Doğrusal Olarak Ayrılamayan Veriler için DVM

Doğrusal olarak belirli oranda hata ile ayrılabilen verilerde esneklik parametresi kullanılarak belli bir ölçüde hatalı sınıflandırmaya izin verilmişti. Ancak veri kümesinin büyük olçüde doğrusal olarak ayrılamadığı durumlarda, çok fazla hata olacağından bu yöntem çözüm getirmez. DVM için bu sorunun çözümü, asıl (orjinal) girdi uzayını daha yüksek boyutlu bir uzaya taşıyarak burada sınıflandırma işlemini yapmaktır.



Şekil 2.4 : Verilerin daha yüksek boyutlu bir uzaya aktarımı.

Bu sayede doğrusal olarak ayrılamayan veriler, çok yüksek boyutlu bir uzaya bile rahatlıkla dönüştürülebilir ve burada sınıflarına ayrılabilirler.

$$x \in R^N \rightarrow \phi(x) \in R^f \qquad \textbf{(2.37)}$$

Şekil 3.4'te görüleceği üzere orijinal girdi uzayından yüksek boyutlu uzaya taşıma işlemini yapan fonksiyon ϕ ile temsil edilir. Bu durumda doğrusal olarak ayrılabilen

DVM'lerde yapılan işlemlerden tek farkı x yerine $\phi(x)$ kullanılmasıdır. Böylece dönüştürülmüş uzayda karar fonkisyonu:

$$w \cdot \phi(x_i) + b = 0 \qquad \textbf{(2.38)}$$

olarak tanımlanır. Hiperdüzlemin ayırdığı veriler ise

$$w \cdot \phi(\mathrm{x}_i) + b \geq +1 \text{ her } y_i = +1 \text{ için} \quad \textbf{(2.39)}$$

$$w \cdot \phi(\mathrm{x}_i) + b \geq -1 \text{ her } y_i = -1 \text{ için} \quad \textbf{(2.40)}$$

şeklinde tanımlanır. Doğrusal olarak ayrılabilen verilerde olduğu gibi çözülmesi gereken optimizasyon problemi

$$min\left[\frac{1}{2} \parallel w \parallel^2\right] \quad \textbf{(2.41)}$$

ve bu optimizasyon problemine bağlı sınırlamalar ise;

$$y_i\,(w.\phi(x_i) + b) - 1 \geq 0 \; ve \; y_j \in \{-1, 1\} \quad \textbf{(2.42)}$$

olarak tanımlanır.

Eldeki eşitlikler edinildikten sonra iki problemin varlığı dikkat çekmektedir. Birincisi nasıl bir haritalama fonksiyonu kullanılacağı, ikincisi ise optimizasyon probleminin çözümü için yüksek boyutlu uzaya aktarımın zor ve karmaşık olacağıdır. Bu problemlerin üstesinden gelmek için ise çekirdek fonksiyonları kullanılır.

Bu durumda w ve b parametreleri,

$$w = \sum \alpha_j \; \mathrm{y}_i \, \phi(\mathrm{x}_i) \quad \textbf{(2.43)}$$

$$f(x) = (\sum \alpha_j \; \mathrm{y}_i \, \phi(\mathrm{x}_i)\phi(\mathrm{x}_j)) + b \quad \textbf{(2.44)}$$

eşitlikleri ile hesaplanır ve bu eşitlikler yüksek boyutlu uzaydaki vektörlerin çarpımlarını içerir. Burada dikkat edilmesi gereken nokta, orijinal girdi uzayını daha yüksek boyutlu uzayına haritalama yaparken, yeni uzayın boyutu çok fazla olduğundan hesaplamaları (çarpımları) yapmanın zorlaşacağıdır. Bu yüzden işlemi daha da basit hale getirebilmek için çekirdek (kernel) fonksiyonları kullanılır.

2.4.1. Çekirdek Fonksiyonları

DVM'de örneklerin yüksek boyutlu bir uzaya aktarımında, bütün değerlerin teker teker çarpım değerlerinin hesaplanarak bulunması işlemleri zorlaştıracaktır. Bunun yerine, doğrudan çekirdek fonksiyonu kullanılarak veri kümesi üzerindeki örneklerin asıl girdi uzayından belirli bir kurala göre taşınması ve yüksek boyutlu uzaydaki değerinin bulunması sağlanır. Böylece çekirdek fonksiyonunu kullanan yöntemler,

yeni uzayın çok fazla boyutlu olması durumunda bile kolay bir şekilde işlem yaparak karmaşık halde bulunan verileri ayırabilme özelliğine sahip olurlar.

Bu durumda eğitim algoritması, yeni uzaydaki verilerin $\phi(x_i) \cdot \phi(x_j)$ iç çarpanlarına bağlı olacaktır. Çekirdek fonksiyonu,

$$K(x_i, x_j) = \phi(x_i) \cdot \phi(x_j) \qquad \textbf{(2.45)}$$

olarak tanımlanır.

Çekirdek fonksiyonlarında "Çekirdek hilesi (kernel trick)" adı verilen yöntem ile haritalama işleminin (ϕ) ne olduğu bilgisine ihtiyaç duyulmaz. Çekirdek fonksiyonu ile eşitlik 2.44'teki ϕ fonksiyonunu hesaplamaya gerek kalmadan örnekler yeni bir uzaya aktarılabilir. Yeni uzayda iç çarpım, girdi uzayında eşdeğer bir çekirdeğe sahiptir. Veri kümesinin test kısmındaki örneğin alacağı değer karar fonksiyonu ile belirlenir. Bu durumda karar fonksiyonu:

$$f(x) = (\sum \alpha_j \, y_i \, \phi(x_i) \, \phi(x_j)) + b \;\; = \sum a_i \, y_i K(x_i, x_j) + b \qquad \textbf{(2.46)}$$

olarak tanımlanır.

Bir fonksiyonun çekirdek fonksiyonu olarak kabul edilebilmesi için, Mercer koşullarını yerine getirmesi gerekir [29]. Mercer Koşulları yeni (dönüştürülmüş) uzayda iki asıl (orjinal) vektörün iç çarpım olarak ifade edilmesini sağlar.

Çekirdek fonksiyonları kullanılırken hangi durumda hangi çekirdek fonksiyonunun daha başarılı olduğunu gösteren analitik bir yöntem yoktur. Bu nedenle uygun çekirdek fonksiyonu, üzerinde çalışılan veri kümesinin eğitim kısmında deneme yaparak bulunmaya çalışılır.

Çizelge 2.1'de en sık kullanılan çekirdek fonksiyonları isimleri, matematiksel ifadeleri ile birlikte sunulmuştur.

Çizelge 2.1 : Çekirdek fonksiyonları.

Çekirdek İsimleri	Matematiksel İfadesi
Doğrusal Çekirdek	$\mathrm{K}(\mathrm{x},\mathrm{y}) = x \cdot y$
Polinom Çekirdeği	$\mathrm{K}(\mathrm{x},\mathrm{y}) = \big((x \cdot y) + 1\big)^{d}$
Normalleştirilmiş Polinom Çekirdeği	$\mathrm{K}(\mathrm{x},\mathrm{y}) = \dfrac{\big((x \cdot y) + 1\big)^{d}}{\sqrt{\big((x \cdot x) + 1\big)^{d}\big((y \cdot y) + 1\big)^{d}}}$
Yarıçap Tabanlı Fonksiyon Çekirdeği	$\mathrm{K}(\mathrm{x},\mathrm{y}) = e^{-\frac{\|x-\mathrm{y}\|^2}{2\sigma^2}}$
Hiperbolik Algılayıcı (Sigmoid)	$\mathrm{K}(\mathrm{x},\mathrm{y}) = \tanh(kx \cdot y)$
Güç Çekirdeği	$\mathrm{K}(\mathrm{x},\mathrm{y}) = -\|x - y\|^{d}$
Log Çekirdeği	$\mathrm{K}(\mathrm{x},\mathrm{y}) = -\log(\|x - y\|^{d} + 1)$
Dalga Çekirdeği	$\mathrm{K}(\mathrm{x},\mathrm{y}) = \dfrac{\phi}{\|x - y\|} - \sin(-\dfrac{\|x - y\|}{\phi})$
İkinci Dereceden Çekirdek	$\mathrm{K}(\mathrm{x},\mathrm{y}) = \sqrt{\|x - y\|^2 + c^2}$
Ters İkinci Dereceden Çekirdek	$\mathrm{K}(\mathrm{x},\mathrm{y}) = \dfrac{1}{\sqrt{\|x - y\|^2 + c^2}}$
Laplace Çekirdeği	$\mathrm{K}(\mathrm{x},\mathrm{y}) = e^{-\frac{\|x-\mathrm{y}\|}{\sigma}}$

Çekirdek fonksiyonlarının başarısı problem tipine göre değişebiilir ve bu yüzden farklı problem tiplerinde farklı çekirdek fonksiyonları kullanmak gerekebilir. Yüksek dereceden çekirdek fonksiyonu olan Yarıçap Temelli Fonksiyon (Radial Basis Function), yapılan çalışmalarda en iyi çalışma başarımı veren çekirdek tipi olarak gösterilmekte [38] ve zaman serisi tahminlemelerinde yaygın olarak kullanılmaktadır.

2.5 Destek Vektör Makinelerinde Regresyon

Zaman serisi veriler için regresyon, gerçek değerli bir veri kümesinde öğrenilme işleminin yapılması ve geleceğe dair yine gerçek değerli bir tahminde bulunulması işlemidir. Regresyon tahmininde temel amaç, uygun parametreleri belirleyerek veriler üzerinde öğrenmeyi en iyi şekilde yapmak ve en iyi hedef değerlerini tahmin etmektir. Destek Vektör Makinaları ile yapılan regresyon Destek Vektör Regresyonu (DVR) olarak isimlendirilir. Regresyon çözümlerinde, sınıflandırma çözümlerinde ayırıcı olarak kullanılan

$$f(x) = w.x + b \tag{2.47}$$

fonksiyonu ile birlikte hata parametresi (ε) değişkeni kullanılır. Hata parametresi (ε) değişkeni hedefteki sapma değerini gösterir. Regresyon için öğrenilen ve tahmin edilen değerler gerçek değerlere sahiptir. Öğrenilen değerler (girdi) x_i ve tahmin edilen (çıktı) değerler y_i olmak üzere:

$$x_i \in R^N,\ y_i \in R^N \quad i = 1, \dots, n \tag{2.48}$$

şeklinde tanımlanır. Hedefteki sapma değerini (ε) her iki yönde de temsil etmek için karar verici fonksiyon:

$$y_i - w\,x_i - b \leq \varepsilon \tag{2.49}$$

ve

$$-y_i + w\,x_i + b \leq \varepsilon \tag{2.50}$$

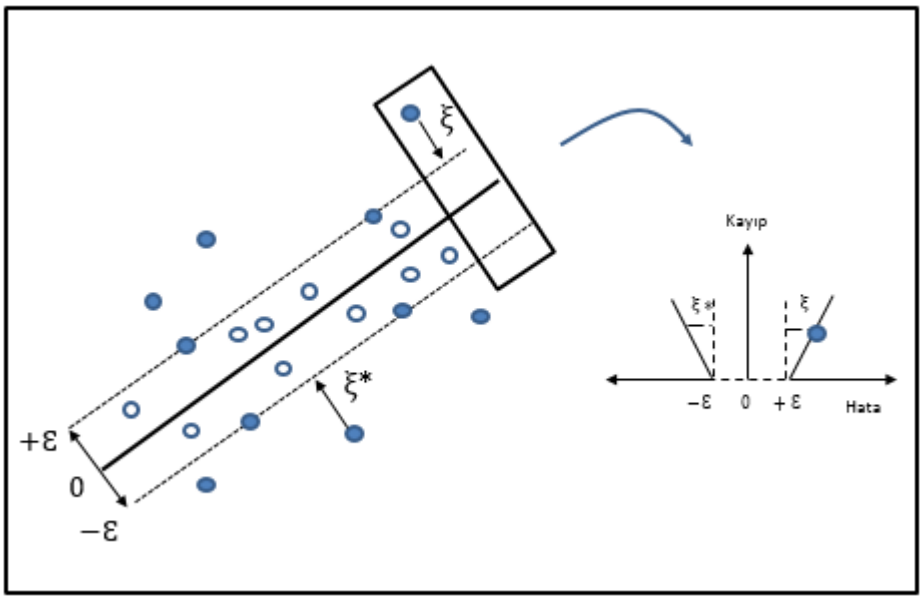
olarak kullanılır.

DVR'de hedef değişkenleri etrafında bir genişlik olduğunu düşünebiliriz. Yapılan tahminde bu genişliğin içinde kalanların hatasının olmadığını, genişliğin dışında kalanların ise genişliğin uzaklığına bağlı olarak hatasının olduğunu söyleyebiliriz.

Hata parametresi (ε) Vapnik [29][37] tarafından aynı zamanda duyarsız kayıp fonksiyonu olarak da adlandırılmıştır. (ε) parametresi küçük ve büyük sapmalar arasındaki genişliği kontrol eder.

Hata parametresinin (ε) büyük olduğu durumlarda destek vektörlerinin sayısı azalır ve böylece öğrenme süresi kısalmış olur. Ancak başarımda düşüş gözlenir.

Hata parametresinin (ε) küçük olduğu durumlarda ise destek vektörlerinin sayısı artar ve bu yüzden öğrenme süresi uzamış olur. Ancak başarımda artış gözlenir.



Şekil 2.5 : Regresyon probleminde hata parametresinin kullanımı.

Şekil 2.5'te, duyarsız kayıp fonksiyonunun regresyon doğru için kullanımı görülmektedir.

Şekil 2.5'in sol tarafında bir veri kümesi içindeki doğrusal DVM, sağ tarafında ε - duyarsız kayıp foksiyonu gösterilmektedir. Sağ tarafdaki fonksiyon, solda gösterilen sınırların dışında uzanan noktaları cezalandırmak amacı ile kullanılmaktadır.

Destek vektör regresyonunda $f(x)$ fonksiyonu için hedefe uygun w ve b parametreleri belirlenerek tahminde bulunulur.

$$min\frac{1}{2}\ \|w\|^2 + C\sum_1^n(\xi_i + \xi_i^*) \qquad \xi_i, \xi_i^* \geq 0 \qquad i = 1 \dots n \quad \textbf{(2.51)}$$

$$(w\, x_i + b) - y_i \leq \varepsilon + \xi_i \qquad i = 1 \dots n \quad \textbf{(2.52)}$$

$$y_i - (w\, x_i + b) \leq \varepsilon + \xi_i \qquad i = 1 \dots n \quad \textbf{(2.53)}$$

Lagrange çarpanları ile dual problemin tanımı:

$$\sum_1^n y_i(\alpha_i - \alpha_j) - \epsilon\sum_1^n y_i(\alpha_i - \alpha_j) - \frac{1}{2}\sum_{i=1}^n\sum_{j=1}^n(\alpha_i - \alpha_j)(\alpha_i - \alpha_j)\, x_i^T x_j \quad \textbf{(2.54)}$$

$$\sum_{i=1}^n(\alpha_i - \alpha_j) = 0 \qquad 0 \leq \alpha_\iota, \alpha_i \leq C \qquad i = 1 \dots n \quad \textbf{(2.55)}$$

Şeklinde olur. Eşitlik 2.55 bir optimizasyon problemidir ve bu durumda sınıflandırma probleminde olduğu gibi çekirdek fonksiyonu kullanılabilir. DVR'nin genel fonksiyonu

$$f(x) = \sum_{i=1}^{n}(\alpha_i - \alpha_j)\ \mathrm{K}(\mathrm{x}_i, \mathrm{x}_j) \quad \textbf{(2.56)}$$

olarak kullanılır.

2.6 Parametre Arama İşlemi ve Çapraz Doğrulama

DVR'lerin başarımı parametrelere bağlı olduğundan parametre arama işlemi genelleştirme başarımı için önemlidir. Bir problemin çözümünde hangi parametrelerin kullanılacağı problemin tipine (sınıflandırma, regresyon), hangi çekirdek fonksiyonlarının kullanılacağına (Polinom, Lineer vb.) bağlı olarak değişir. Örneğin Yarıçap Tabanlı Fonksiyon (YTF) (Radial Basis Function) çekirdeği kullanan bir sınıflandırma probleminde ceza parametresi C ve çekirdek parametresi olan (γ) parametresi kullanılırken, aynı çekirdeği kullanan regresyon probleminde bunlara ek olarak hata parametresi (ε) kullanılır.

Bu tez çalışmasında yaygın olarak kullanılan ve daha önce yapılmış çalışmalarda en yüksek başarıma sahip çekirdek olarak kabul edilen [38] YTF çekirdeği kullanılmıştır. Ayrıca diğer çekirdek fonksiyonları (laplace, polinom vb.) ile de sonuçlar alınmış ancak YTF çekirdeğinden daha başarılı olmadıkları gözlenmiştir.

Yapılan regresyon tahminlemelerinde başarımı artırmak amacı ile ceza parametresi (C), YTF çekirdeği parametresi (γ) ve hata parametresi (ε) (regresyon başarımı) için değişmeler gözlemlenmiştir. Eğitim kısmında her bir parametrenin başarımı diğer iki parametrenin değişimleri ile elde edilmiş, en iyi başarımı veren parametreler test kısmında kullanılmıştır. Parametre aralıkları

$$C = [2^{-3}, \dots, 2^{8}] \quad \textbf{(2.57)}$$

$$\gamma = [2^{-9}, \dots, 2^{1}] \quad \textbf{(2.58)}$$

$$\varepsilon = [2^{-11}, \dots, 2^{-5}] \quad \textbf{(2.59)}$$

olarak alınmıştır. Parametre aralıklarının artması optimum parametrelerin bulunma süresini artırmaktadır. Ayrıca belirtilen parametre aralıkları ile optimum parametreyi belirleyerek en iyi başarıyı bulmak amacı ile çapraz doğrulama (cross validation) yöntemi kullanılmıştır.

Çapraz doğrulama yönteminde eğitim kümesi eşit büyüklükte ve birbirinden bağımsız v adet alt kümeye ayrılır. Alt kümeler

$$(K_1, K_2, \dots, K_v) \tag{2.60}$$

ve

$$i = 1 \dots v \tag{2.61}$$

olmak üzere, i. alt küme küme test olarak seçilir, geri kalan v-1 küme eğitim kümesi olarak kullanılır. Çapraz doğrulamanın kullanılmasının amacı aşırı uyum (overfit) sorununu ortadan kaldırmasıdır. Alt küme sayısının artması işlem süresini uzatır ancak doğruluğu artırır. Bu yöntem bütün alt kümeler için kullanılır ve test hatası alt küme hatalarının ortalaması olarak alınır.

3. GÖRGÜL KİP AYRIŞIMI

Görgül Kip Ayrışımı (GKA) (Empirical Mode Decomposition), bir veri kümesi ya da sinyali durağana yakın sonlu sayıda bileşenlere ayrıştırabilen bir sinyal işleme metodudur [39]. GKA'nın en önemli avantajı, veri kümesinin/sinyalin özelliğine bakmaksızın ve dolayısı ile bir ön bilgi gerekmeden veri kümesini bileşenlerine kolayca ayrıştırabilmesidir ve bu özelliği ile diğer algoritmalardan ayrılır. Bazı çalışmalar GKA'nın, Fourier ve dalgacık dönüşümleri gibi diğer dönüşüm algoritmalarından daha başarılı olduğunu göstermektedir [39]. GKA ile orijinal veri kümesi İçkin Mod Fonksiyonları (IMF) (Özgün Kip İşlevleri) (Intrinsic Mode Functions) ve bir adet Kalıntı (Residue) adı verilen bileşenlere ayrıştırılır.

$x(\mathrm{t})$ zaman serisi veri kümesi olmak üzere:

$$x(\mathrm{t}) = \sum_{i=1}^{n} IMF_i(t) + r\,(t) \qquad \textbf{(3.1)}$$

ifadesindeki $IMF(t)$ ayrışıtırılan IMF adlı bileşenleri, $r(t)$ ise kalıntı bileşenini ifade eder.

Ayrıştırılan veri kümesinin bileşenleri (IMF'ler ve Kalıntı) tekrar toplandığında ise tekrar asıl (orijinal) veri kümesininin değerini verir. Bu yüzden GKA ile istenildiği zaman asıl veri kümesine dönülmüş olunur.

Ayrıştırma işleminin oluşturduğu bileşenler 2 özelliğe sahiptir:

1. Veri kümesindeki uç noktaların (yerel maksimum ile yerel minimum) sayıları ile bunların arasındaki sıfır geçişlerinin sayıları arasındaki fark en fazla birdir.
2. Yerel en büyük değer (maksimum) ile yerel en küçük değerin (minimum) ortalaması daima sıfırdır.

3.1 Kabuk Soyma İşlemi

GKA'da bileşenlere ayrıştırma işlemini yapan işleme Kabuk Soyma (Sifting) adı verilir. $x(t)$, t zamanlı bir veri kümesi olmak üzere GKA üzerinde bileşenlerine (IMF'ler ve kalıntı) ayırmayı amaçlayan kabuk soyma işlemi aşağıdaki adımlarla gerçekleştirilir

1. Veri kümesinin yerel en büyük ve yerel en küçük değerleri bulunur.
2. Yerel en büyük ve yerel en küçük değerleri birleştirilerek üst zarf (x_{ust}) ve alt zarflar (x_{alt}) elde edilir.
3. Alt ve üst zarfların ortalaması hesaplanır,

$$m_1(t) = (x_{ust}(t) + x_{alt}(t))/2 \quad \textbf{(3.2)}$$

elde edilen ortalama orijinal veri kümesinden çıkartılır.

$$h_1(t) = x(t) - m_1(t) \quad \textbf{(3.3)}$$

4. Elde edilen $h_1(\mathrm{t})$ veri kümesinin GKA bileşenlerinin sağlaması gereken 2 özelliği taşıyıp taşımadığı kontrol edilir.
5. Şartlar sağlanmadığı taktirde yeni veri kümesi $h_1(\mathrm{t})$ üzerinden 1. adımdan tekrar başlanarak işlemler şartlar sağlanana kadar tekrar edilir.
6. Şartlar sağlandığı taktirde eldeki veri kümesi

$$h_1(\mathrm{t}), i = 1, \ldots n = \mathrm{IMF}_i \quad \textbf{(3.4)}$$

olmak üzere IMF olarak kabul edilir.

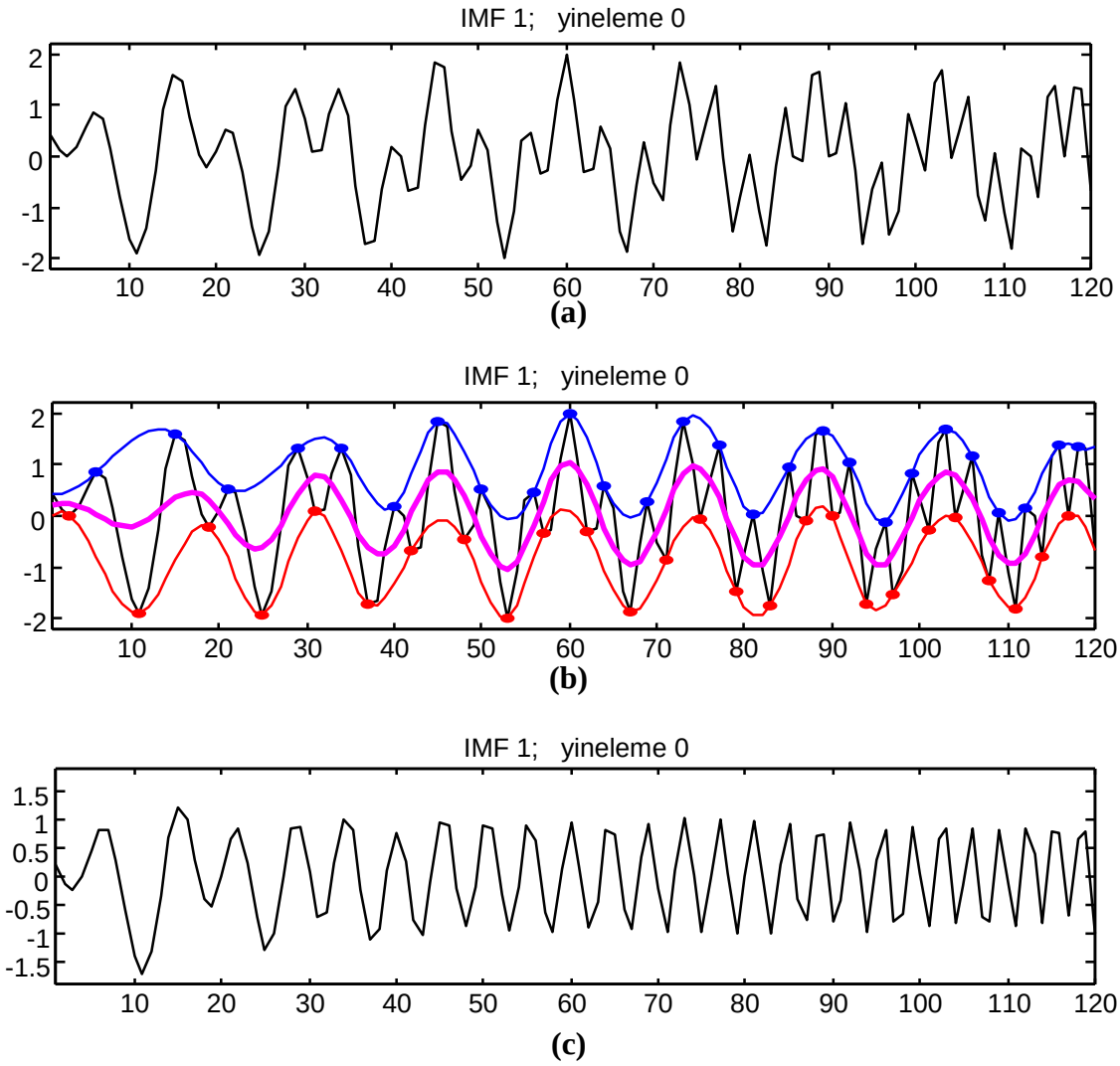
7. Yeni veri kümesi

$$r_1(t) = x(\mathrm{t}) - h_1(\mathrm{t}) \quad \textbf{(3.5)}$$

olmak üzere ilk 6 adım yeni IMF'ler elde edilmek üzere yinelemeli (iterasyon) olarak tekrar edilir. Bu işlem belirli bir sayıda IMF elde edilene kadar (belirtilmiş ise) ya da $r_1(\mathrm{t})$ tek düze (monoton) bir değere ulaşana kadar tekrar eder. İşlem sonlandığında elde bulunan $r_1(\mathrm{t})$ veri kümesi Kalıntı $r(\mathrm{t})$ olarak adlandırılır. Böylece veri kümesi IMF'ler ve kalıntı olmak üzere bileşenlerine ayrılmış olur.

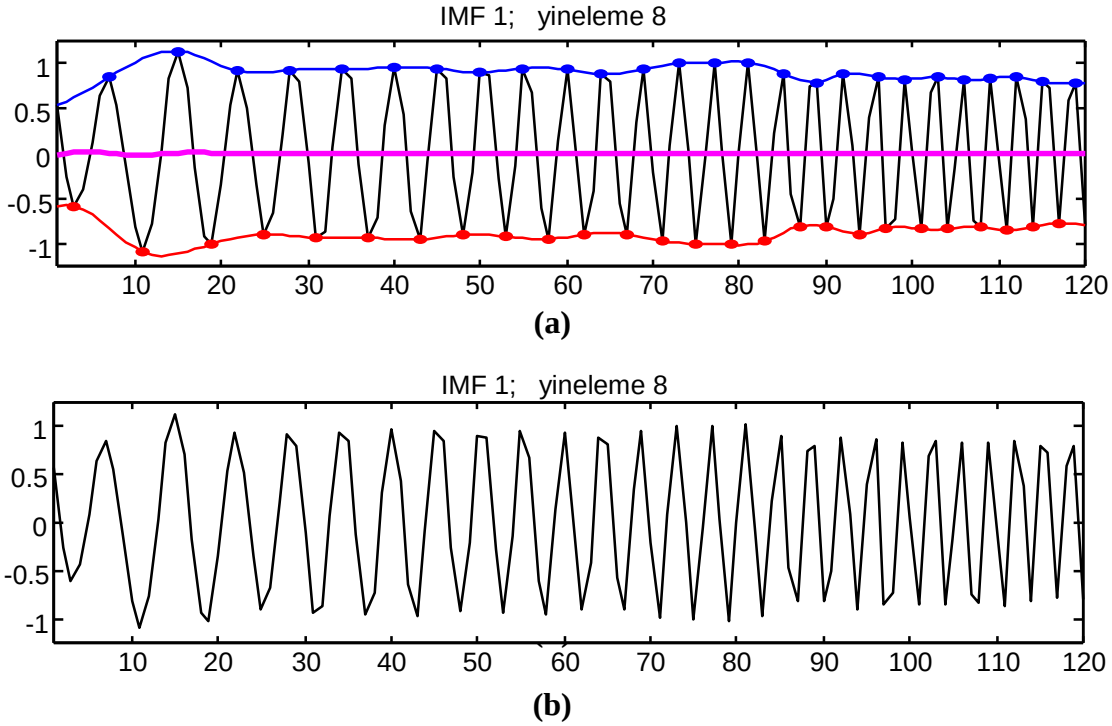
GKA yöntemi birçok alandaki uygulamalarda başarılı sonuçlar vermektedir [40]. GKA ile durağan olmayan veri kümelerinin periyodiğe yakın bileşenlere ayrılabilmesi ve her bileşenin diğerlerinden bağımsız özellikler taşıması bir çok alanda kullanılmasını sağlamıştır. Elde edilen bileşenler incelendiğinde, ilk bileşenlerin daha yüksek frekanslı, son bileşenlerin ise daha düşük frekanslı olduğu görülmektedir [40].

Şekil 3.1'de GKA yönteminin bileşenlere ayırma işlemi grafik üzerinde gösterilmiştir.



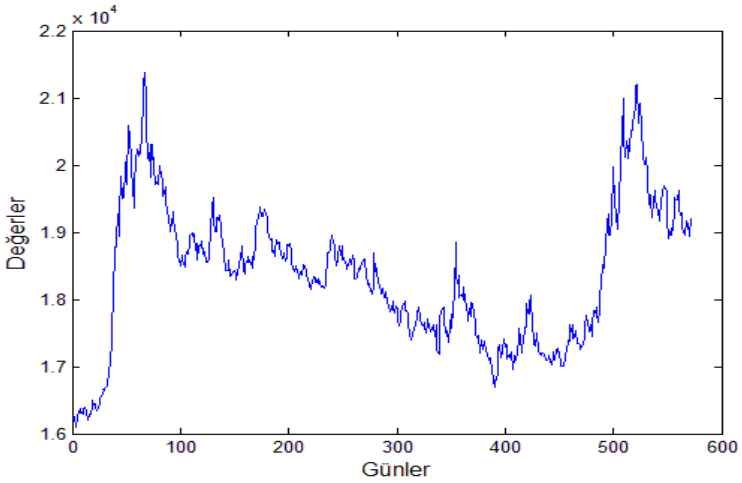
Şekil 3.1 : Kabuk soyma işlemi.

Şekil 3.1.(a) asıl (orjinal) veri kümesinin zamana göre değerlerini göstermektedir. (b)'deki mavi çizgi üst zarfı, kırmızı çizgi alt zarfı, pembe çizgi ise üst zarf ve alt zarfın ortalamasını (c)'deki grafik ise orjinal veri kümesinden ortalamanın çıkarılması işleminden kalan değerleri göstermektedir. İlk yinelemede (iterasyon) IMF şartları sağlanmamıştır. Bu durumda kabuk soyma işleminin 1. Adımından başlanarak işlemler tekrar edilecektir. Şekil 3.2'de kabuk soyma işleminde ilk IMF'nın elde edildiği 8. yinelemem görülmektedir.

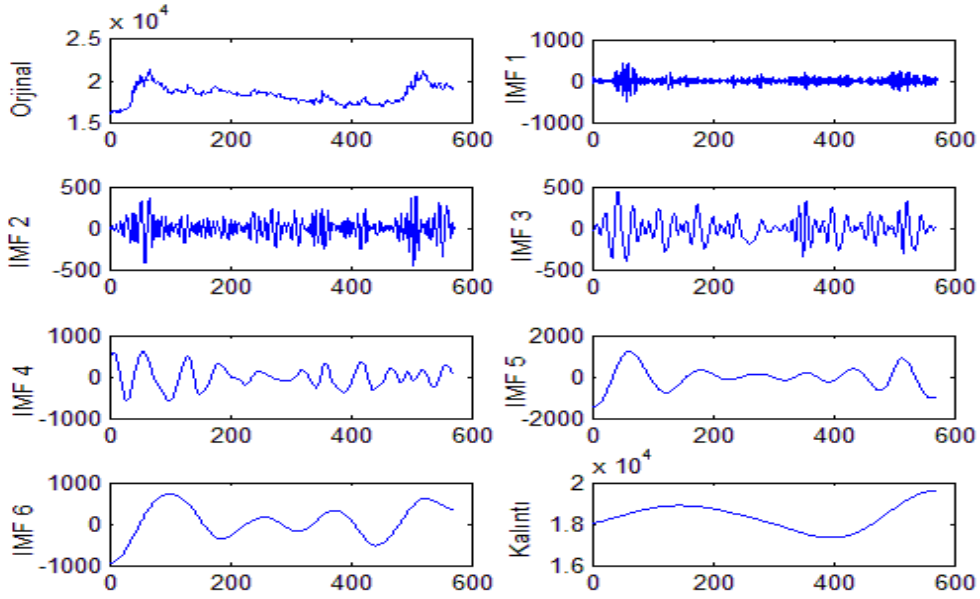


Şekil 3.2 : Kabuk soyma işlemi ile ilk IMF'nin elde edilmesi.

Şekil 3.2 (a)'da mavi çizgi üst zarfı, kırmızı çizgi alt zarfı, pembe çizgi ise üst zarf ve alt zarfın ortalamasını, (b)'de asıl veri kümesinden ortalamanın çıkarılması işleminden kalan değerleri göstermektedir. 8. yinelemede (iterasyon) IMF şartları sağlanmıştır. Bu durumda ilk IMF değeri, asıl veri kümesinden çıkarılacak ve yeni veri kümesi üzerinden bir sonraki IMF elde edilmek üzere yinelemeli olarak adımlar tekrar edilecektir.

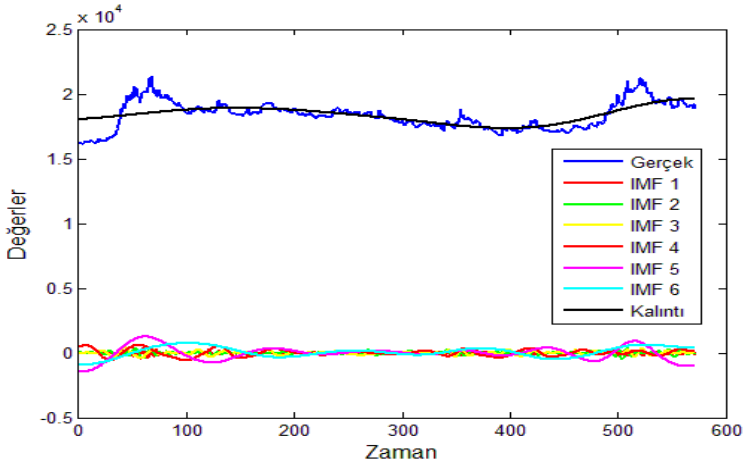


Şekil 3.3 : TL/Avro zaman serisi.



Şekil 3.4 : Durağan olmayan veri kümesine ait GKA bileşenleri.

Şekil 3.3'te durağan olmayan bir zaman serisi veri kümesi (21 Mart 2006-26 Haziran 2008 tarihleri arası günlük TL/Avro kapanış fiyatları), Şekil 3.4'te ise aynı veri kümesinin GKA ile ayrıştırılmış bileşenleri (6 adet IMF ve 1 adet Kalıntı) gösterilmektedir. Veri kümesinde düzensiz değişimlerin (ani çıkış ve inişler) olması, GKA'nın durağana yakın bileşenler çıkarmasını engellememiştir. Şekil 3.4'teki t zamanının bileşenleri incelendiğinde, Kalıntı bileşeni değerlerinin diğer bileşenlerden çok daha büyük olduğu, asıl veri kümesi değerlerine yakın olduğu gözlenmiştir. Şekil 3.5'te Şekil 3.4'te görülen bileşenlerin gerçek veri kümesi üzerindeki değerleri görülmektedir.



Şekil 3.5 : GKA Bileşenlerinin asıl veri kümesi üzerinde gösterimi.

Bileşenlerin ayrıştırılmasında durma kriteri olarak kalıntı bileşeninin tekdüze (monoton) olması ya da bileşen sayısının belirtilmesi yöntemi ile belirlenebilir. Bu tez çalışmasında 6. bölümde anlatılacağı üzere her ikisi de kullanılmıştır.

GKA üzerinde yapılan çalışmalarda, elde edilen bileşenlerin karakteristikleri de incelenmiştir [40]. Gürültü çıkarımı için hangi IMF'nin gürültü olduğu tespit edilerek, gürültü barındıran IMF hariç diğerleri toplanarak gürültüsüz bir veri kümesi/işaretin elde edildiği çalışmalar mevcuttur [41-43]. Örneğin Premanode, finansal verilerde gürültülü işaretin kalıntıdan bir önceki bileşen (son IMF) olduğunu göstermektedir [43].

Bu tez çalışmasında GKA'nın zaman serisi veriler için önceki çalışmalarda kullanımı incelenmiş, ayrıca gerçek uygulamalarda kullanılmak üzere DVR algoritması ile bütünleşik olarak iki yaklaşım için önerilmiştir. İlk yaklaşımda, DVR tahminlemeleri asıl veri kümesi yerine GKA bileşenleri üzerinde yapılarak tüm bileşenlerin tahminlemeleri toplanmıştır. İkinci yaklaşımda ise GKA bileşenleri kullanılarak gürültü süzme tekniği kullanılmıştır. Gürültü süzülerek elde edilen yeni veri kümesi, DVR yönteminde öznitelik olarak kullanılmıştır. Her iki yaklaşım ile elde edilen sonuçlar tekli DVR yöntemi ile karşılaştırılmıştır.

4. KULLANILAN VERİ KÜMELERİ VE BAŞARI ÖLÇÜMLERİ

4.1 Veri Kümeleri ve Analizleri

Zaman serisi veriler üzerinde yapılan tahminlerde bir modelin başarılı olduğunun doğrulanabilmesi için oluşturulan modelin farklı büyüklükteki/özellikteki/alanlardaki veri kümeleri üzerinde uygulanması gerekir. Bir modelin bir veri kümesi tahminlemesi için başarılı sonuçlar vermesi, o modelin her zaman başka veri kümeleri için de başarılı sonuçlar vereceği anlamına gelmez.

Bu tez çalışmasında önerilen modellerin başarı ölçümü toplam 7 adet veri kümesi üzerinde uygulanmıştır. Veri kümeleri içinde durağan ve durağan olmayan özellikte veriler olduğu gibi, büyüklük olarak 1- 10 yılları arasında farklı büyüklükte veriler de bulunmaktadır. Başarı ölçümlerinin etkin bir şekilde görülebilmesi amacı ile test kümelerinin boyutları da farklı veri kümelerinde farklı büyüklüklerde kullanılmıştır. Kullanılan 1., 2. ve 3. veri kümeleri durağan, 4., 5., 6.ve 7. veri kümeleri ise durağan olmayan özelliğe sahiptir. Veri kümelerinin özellikleri Çizelge 4.1'de gösterilmiştir.

Çizelge 4.1 : Kullanılan veri kümeleri.

Veri Kümesi	Toplam Adet	Eğitim	Test	Ortalama	Standart Sapma	En Büyük Değer	En Küçük Değer	Kullanım Alanı
1	374	343	31	752.609	46.782	876	612	Endüsri
2	1601	1400	201	0.966	0.1639	0.618	1.344	Endüstri
3	396	365	31	5214.95	577.831	6654.5	3803	Endüstri
4	1034	1002	32	1.738	0.188	2.324	1.395	Finans
5	752	602	150	1.662	0.142	1.919	1.395	Finans
6	3098	2500	598	1.187	0.195	1.601	0.828	Finans
7	1400	1120	280	9.889	2.607	6.797	1400	Oşinografi

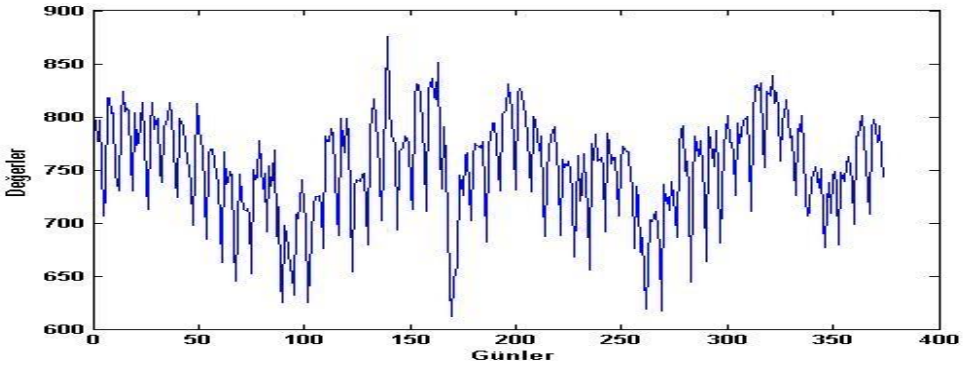
- Veri Kümesi 1 [12]

Slovakya Doğu Bögesi Elektrik kullanım değerlerini içerir. Aynı zamanda EUNITE [13] elektrik yükü tahminleme yarışmasında kullanılan veri kümesidir.

Yarışmada, eğitim verisi olarak 1997 ve 1998 yıllarının yarım saatlik elektrik kullanım değerleri verilmiş, 1999 Ocak ayına ait her gündeki en fazla elektrik kullanım değerlerinin tahmin edilmesi istenmiştir. Bu yüzden eğitim kümesinde yer alan yarım saatlik elektrik kullanım değerlerinin gün içindeki en fazla olanı, o günün değeri olarak alınmıştır. Aynı zamanda veri kümesinde kış dönemi elektrik tüketiminin yaz dönemine göre daha yüksek olduğu gözlenmiştir (mevsimsel etki). Tahmin yapılacak ayın kış mevsiminde olmasından dolayı, yaz aylarına ait değerler eğitim kümesinden çıkarılmış, yalnızca Ocak-Mart ve Ekim-Aralık değerleri kullanılmıştır. Böylece veri kümesi üzerinde küçültme yapılmış, veri adedi 376'ya indirilmiştir.

Eğitim Kümesi: 1997 – 1998 yıllarına ait günlük elektrik tüketim yükü (MW)

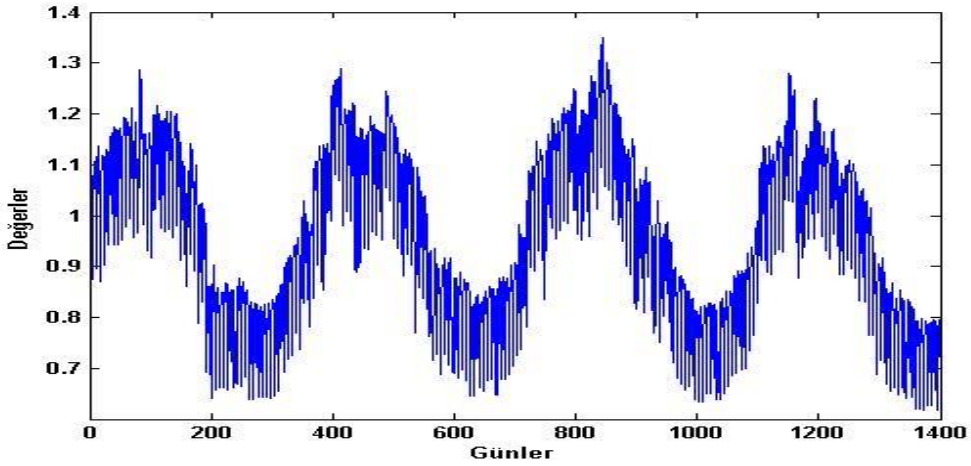
Test Kümesi: Ocak 1999 ayına ait günlük elektrik tüketim yükü (MW)



Şekil 4.1 : Veri Kümesi 1 Zaman/Değer Grafiği.

- Veri Kümesi 2 [44]

1990'lı yıllardaki Polonya ülkesine ait elektrik kullanım değerlerini içerir. Değerlerin birimi ve hangi günlere ait olduğu bilgileri verilmemiştir [44].



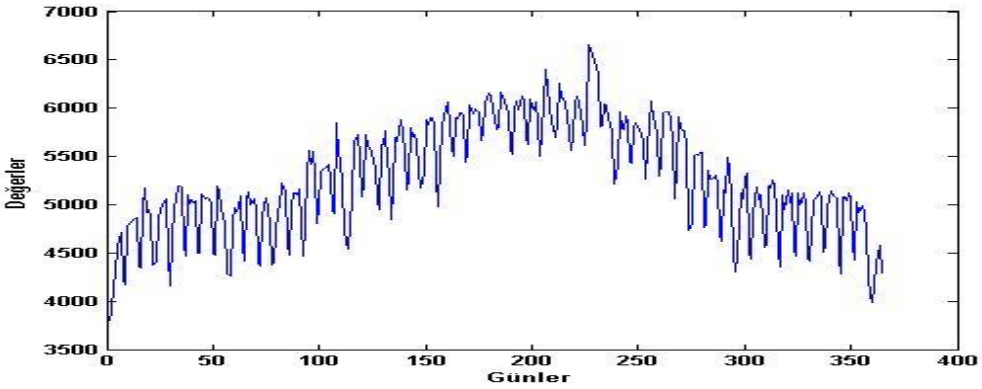
Şekil 4.2 : Veri Kümesi 2 Zaman/Değer Grafiği.

- Veri Kümesi 3 [45]

Yeni Zelanda ülkesine ait milli elektrik kullanım değerlerini içerir.

Eğitim Kümesi: 2011 yılına ait günlük elektrik tüketim yükü (MW)

Test Kümesi: Ocak 2012 ayına ait günlük elektrik tüketim yükü (MW)



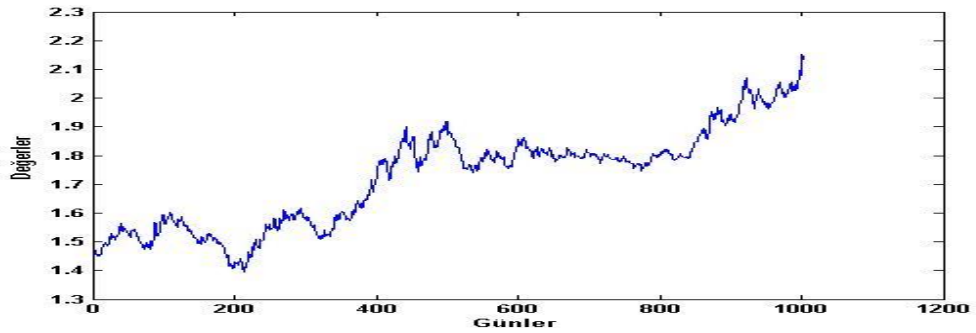
Şekil 4.3 : Veri Kümesi 3 Zaman/Değer Grafiği.

- Veri Kümesi 4 [46]

Türk Lirası (TL)/ Amerikan Doları (USD) paritesinin değerlerini içerir. Haftasonları işlem olmadığından değerler haftaiçi bilgilerine aittir.

Eğitim Kümesi: 2010-2013 yıllarına ait günlük TL/USD paritesi

Test Kümesi: 1 Ocak-14 Şubat 2014 tarihleri arasındaki günlük TL/USD paritesi



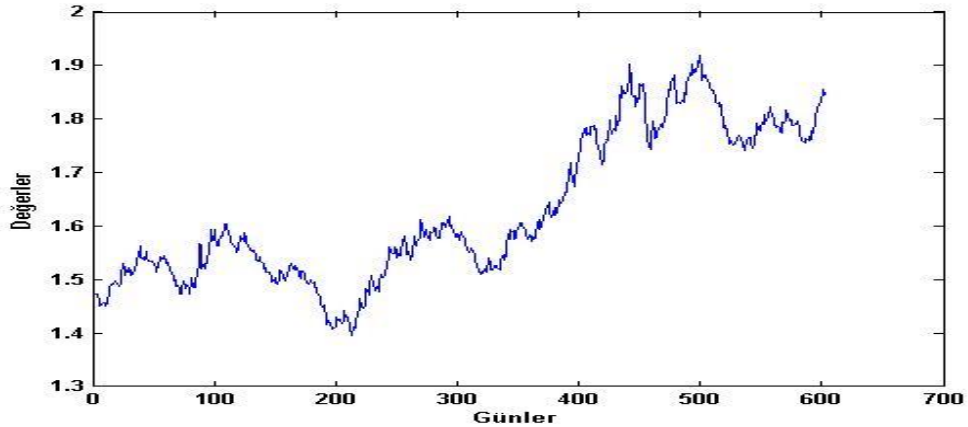
Şekil 4.4 : Veri Kümesi 4 Zaman/Değer Grafiği.

- Veri Kümesi 5 [46]

Türk Lirası (TL)/ Amerikan Doları (USD) paritesinin değerlerini içerir. Haftasonları işlem olmadığından değerler haftaiçi bilgilerine aittir.

Eğitim Kümesi: 1 Ocak 2010-23 Mayıs 2012 yıllarına ait günlük TL/USD paritesi

Test Kümesi: 24 Mayıs 2012 – 31 Aralık 2012 tarihleri arasındaki günlük TL/USD paritesi



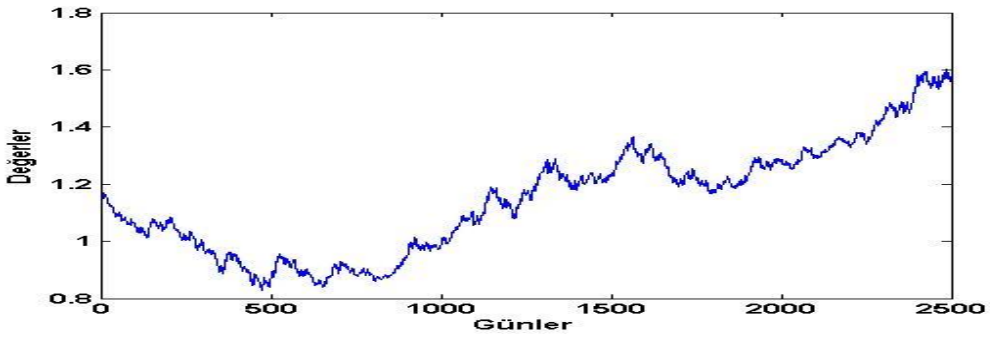
Şekil 4.5 : Veri Kümesi 5 Zaman/Değer Grafiği.

- Veri Kümesi 6 [45]

Avro (EU) / Amerikan Doları (USD) paritesinin değerlerini içerir. Haftasonları işlem olmadığından değerler haftaiçi bilgilerine aittir.

Eğitim Kümesi: 28 Temmuz 1998-3 Temmuz 2008 yıllarına ait günlük EU/USD paritesi

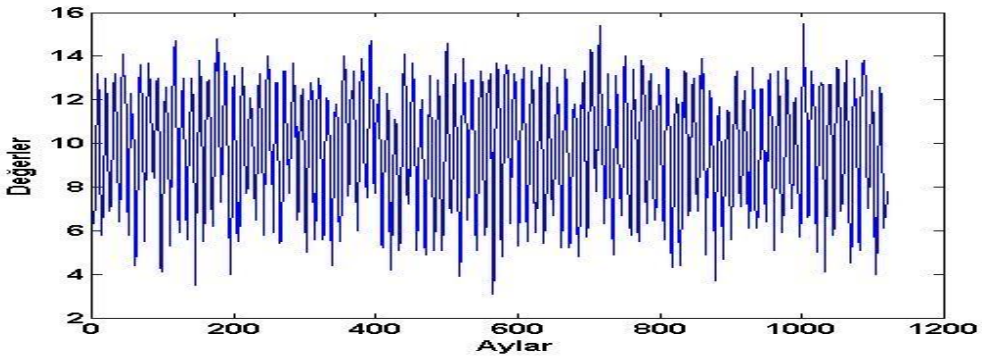
Test Kümesi: 3 Temmuz 2008 – 20 Kasım 2010 tarihleri arasındaki günlük EU/USD paritesi



Şekil 4.6 : Veri Kümesi 6 Zaman/Değer Grafiği.

- Veri Kümesi 7 [47]

1882 ile 1998 yıllarındaki her ayın deniz seviyesi basınç değerlerini içerir. Veri kümesinin %80'i eğitim, %20'si test kümesi olarak kullanılmıştır.



Şekil 4.7 : Veri Kümesi 7 Zaman/Değer Grafiği.

4.2 Başarı Ölçümleri

Zaman serisi tahminlemelerinde başarımı değerlendirebilmek amacı ile kurulan modelden alınan tahmin değeri ile gerçek değer arasındaki başarım fonksiyonlar kullanılarak ölçülür ve elde edilen veriler istatiktiklere dayandırılır. Ölçüm teknikleri problemin cinsine göre değişiklik gösterebilir. Bir verinin artıp ya da azalacağı bir sınıflandırma problemi iken, verinin hangi rassal değişkeni alacağı bir regresyon problemidir.

Bir veri kümesinden doğru bir tahmin başarımı elde edebilmek için farklı teknikler ile farklı ölçümler yapılmalıdır. Tek bir ölçüm değeri bir tahminin her zaman başarılı olduğu anlamına gelmeyebilir. Örneğin yapılan bir tahmin, noktasal olarak çok

başarılı olarak değerlendirilebilirken, artış ya da azalış yönünden başarılı olmayabilir. Bu yüzden tahminlerin başarım ölçümünde tek bir kritere güvenilmemelidir.

Zaman serisi veriler için tahmin yapılacak n kadar örnek var ise, hata terimi sayısı da n olacaktır. Bu yüzden ölçümlerin her biri zaman serilerinin test kümesi uzunluğuna bağlı olarak değişir. Aynı zamanda değerlerin büyüklüğü de hata oranlarını doğru orantılı olarak etkileyecektir. Örneğin elektrik kullanım yükü tahminlerinin noktasal olarak hata oranları bir kasabada düşük iken, bir ülke genelinde büyük olacaktır.

Bu tez çalışmasında noktasal ve yön tahminlerini içeren 5 adet ölçüm kriteri kullanılmış, matematiksel ifadeleri çizelge 4.2'de gösterilmiştir.

Çizelge 4.2 : Başarı ölçümleri ve matematiksel ifadeleri.

Ölçüm Kriteri Adı	Matematiksel Formülasyonu
Ortalama Karasel Hata (OKH)	$\frac{1}{n}\sum_{i=1}^{n}(T_i - G_i)^2$
Ortalama Mutlak Hata Yüzdesi (OMHY)	$\frac{100}{n}\sum_{i=1}^{n}\left\|\frac{T_i - G_i}{T_i}\right\|$
Yön Doğruluğu (YD)	$\frac{100}{n}\sum_{i=1}^{n} d_i$ $d_i = \begin{cases} 1 & (T_i - T_{i-1})\,(G_i - G_{i-1}) > 0 \\ 0 & \text{diğeri} \end{cases}$
Doğru Artış (DAr)	$\frac{100}{Ar}\sum_{i=1}^{n} d_i$ $d_i = \begin{cases} 1 & (G_i - G_{i-1}) > 0 \text{ ve } (T_i - T_{i-1})\,(G_i - G_{i-1}) > 0 \\ 0 & \text{diğeri} \end{cases}$
Doğru Azalış (DAz)	$\frac{100}{Az}\sum_{i=1}^{n} d_i$ $d_i = \begin{cases} 1 & (G_i - G_{i-1}) < 0 \text{ ve } (T_i - T_{i-1})\,(G_i - G_{i-1}) > 0 \\ 0 & \text{diğeri} \end{cases}$

Çizelgede 4.2'de G Gerçek Değer, T Tahmin Değeri, n veri kümesindeki tahmin yapılacak değer sayısı, Ar gerçekte artan değer sayısı, Az gerçekte azalan değer sayısıdır.

4.2.1 Ortalama karasel hata

Ortalama Karesel Hata (OKH), her nokta için toplam tahmin değerinin gerçek değerden ortalama ne kadar uzakta olduğunu gösterir. OKH değerinin 0 olması, yapılan tahmin başarımının gerçek değer ile aynı ve mükemmel olması anlamına gelir. Ölçülen hataların karesinin alınması noktasal olarak büyük hataların daha fazla cezalandırılmasını sağlar böylece eğitim işleminde yüksek hataya sahip modeller seçilmez. Ölçümlerde yüksek başarımlı bir tahmin için düşük OKH değeri hedeflenir.

4.2.2 Ortalama mutlak hata yüzdesi

Ortalama Mutlak Hata Yüzdesi (OMHY), yüzdelik hataların mutlak değerleri toplamlarının ortalamasına eşittir. Yüzdelik hata değeri tahmin başarısını daha anlaşılır hale getirmektedir. Örneğin OKH hatasının 100 ya da 1000 olması veri kümesini incelemeden bir şey ifade etmezken, OMHY hatasının %2 olması tahmin başarısını daha anlaşılır biçimde gösterir. OKH'da olduğu gibi OMHY'da da hata değerinin küçük olması yüksek başarımı gösterir.

4.2.3 Yön tahminleri

Tahmin değerlerinin yön başarısı özellikle finans alanında önemlidir. Bu çalışmada bir sonraki değerin artacak ya da azalacak olması tahmini Yön Doğruluğu (YD) ölçümü ile yapılmıştır. Gerçekte artan değerlerin yüzdesel olarak ne kadar başarılı bilindiğini gösteren ölçüm Doğru Artış (DAr), gerçekte azalan değerlerin yüzdesel olarak ne kadar başarı ile bilindiğini gösteren ölçüm ise Doğru Azalış (DAz) olarak adlandırılmıştır. Doğru Artış ve Doğru Azalış değerlerinin gösteriliş amacı, Yön Doğruluğu ölçümünün doğru olup olmadığının gözlemlenmesidir. Örneğin %70 oranında artışa sahip bir veri kümesinde bütün yönlerin artışta olduğunu tahmin etmek, Yön Doğruluğu ölçümünde %70 başarı gösterecektir. Ancak bu durumda Doğru Artış değeri %100, Doğru Azalış değeri ise %0 göstereceğinden oluşan sorunu tespit etmek kolaylaşacaktır.

Tez kapsamında incelenen bazı çalışmalarda yön başarı ölçümlerinin küçük veri kümeleri üzerinde yapılmış oluğu görülmüştür. Az sayıda test değeri içeren bir veri kümesinde, yön doğruluğunun başarısını doğru bir şekilde ölçmek mümkün değildir. Örneğin 25 adet test değeri içeren bir veri kümesinde 2 model karşılaştırıldığında, 3 adet daha fazla yön doğruluğuna (artış/azalış) sahip modelin YD başarımı %12 daha fazla olacaktır. Ancak test kümesinin küçüklüğünden kaynaklanan bu başarı, çok sayıda test değeri içeren bir veri kümesinde görülemeyecek ve yanıltıcı olacaktır.

4.2.4 Karışıklık matrisi

Karışıklık Matrisi (Confusion Matrix), sınıflandırma problemlerinde yapılan tahminin gerçekte ne kadar başarılı olduğunun ölçümünü örneklem sayıları ile gösteren matristir. Bu matrise bakarak modelin hangi alanda ne kadar doğrusu ya da yanlışı olduğu görülebilir. Çizelge 4.3'te gösterilen karışıklık matrisinde, artan değerler pozitif, azalan değerler negatif olarak tanımlanmıştır.

Çizelge 4.3 : Karışıklık Matrisi.

		Tahmin	
Gerçek		Pozitif(P)	Negatif (N)
	Pozitif(P)	Doğru Pozitif (DP)	Yanlış Negatif (YN)
	Negatif (N)	Yanlış Pozitif (YP)	Doğru Negatif (DN)

Bu tez çalışmasında, yön tahminlerinde kullanılan Yön Doğruluğu, Doğru Artış ve Doğru Azalış başarı ölçümleri, aynı zamanda karışıklık matrisi kullanılarak da elde edilmiş ve bu şekilde Çizelge 4.2'de matematiksel ifadeleri yer alan YD, DAr veDAz ölçümlerinin sağlaması yapılmıştır. Çizelge 4.4'te, karışıklık matrisindeki ifadelerin kullanılması ile yön ölçümlerinin tanımlanması yapılmıştır.

Çizelge 4.4 : Yön başarı ölçümlerinin karışıklık matrisi kullanılarak elde edilmesi.

Ölçüm Kriteri Adı	Matematiksel Formülasyonu
Yön Doğruluğu (YD)	$\frac{DP + DN}{DP + DN + YP + YN}$
Doğru Artış (DAr)	$\frac{DP}{DP + DN}$
Doğru Azalış (DAz)	$\frac{DP}{DP + DN}$

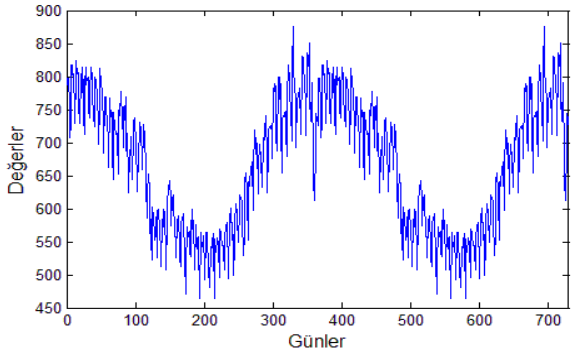
5. ZAMAN SERİSİ VERİLERİN DESTEK VEKTÖR MAKİNELERİ ÜZERİNDE MODELLENMESİ

2001 yılında yapılan EUNITE [12] yarışmasını en iyi tahmin başarımı ile kazanan çalışma, DVR yöntemi başarısının yanısıra, modellenmenin önemini de ortaya koymuştur [13]. Bu bölümde, DVR algoritmasının zaman serisi veriler üzerindeki başarısını artırmak için yapılan işlemler incelenmiştir.

Zaman serisi verilerde öncelikle üzerinde tahmin yapılacak veri kümesi incelenir. Veri kümesinin durağan olması ya da olmaması durumunda zaman serisinin modellemesi farklılık gösterebilir.

5.1 Veri Kümesinin Analizi ve Modelin Hazırlanışı

Bir veri kümesinin durağan ya da periyodik oluşu, tahmin başarımını artırır. Örneğin Chen ve arkadaşları [13], en iyi tahminlemeyi yaparak birinci oldukları EUNITE yarışmasında verilen veri kümesini incelemiş, tahmin başarısını artırmak için veri kümesinin özelliğine göre modelleme yapmışlardır. Yarışmada, elektrik yükünün değerlerini içeren bir veri kümesi üzerinde çalışmışlardır.



Şekil 5.1 : EUNITE yarışmasında kullanılan veri kümesi.

Veri kümesini incelediklerinde, yaz aylarında elektrik kullanımının daha az, kış aylarında ise daha fazla olduğunu tespit etmişler, dolayısı ile mevsimsel etki gözlemlemişlerdir. Yarışmada Ocak 1999 ayındaki günlük en fazla (maksimum) değerlerinin tahmin edilmesi istendiğinden ve Ocak ayı bir kış ayı olduğundan,

eğitim kümesinde yalnızca kış aylarını kullanarak, yaz aylarının değerlerini çıkarmak başarımı artırmıştır [13]. Bu çalışmada, aynı zamanda bahsi geçen yarışmanın veri kümesi olan Veri Kümesi 1'in üzerinde, bu etki göz önüne alınarak sonuçlar alınmıştır.

Veri kümesinin modellenmesinde kullanılan diğer özellikler aşağıdaki maddelerde gösterilmiştir.

5.1.1 Takvim bilgisi

Bir veri kümesinde, değerler periyodik ya da yarı periyodik özellik gösterebilir. Bu tez çalışmasındaki elektrik kullanım değerlerini içeren ilk 3 veri kümesinde, haftasonlarına ait elektrik kullanım değerleri hafta içine göre daha düşük, hatta çoğu durumda pazar gününe ait değerler en düşük seviyededir. Bu durumda kullanımın haftalık olarak periyodik olduğu gözlemlenmiştir.

Periyotluk bu tez çalışmasındaki ilk 3 veri kümesinde haftalık olarak gözlenebileceği gibi, farklı veri kümelerinde aylık ya da yıllık bazda da görülebilir. Örneğin, yazın artış gösteren aylık dondurma satışlarını içeren, kışın artış gösteren akarsu debisi değerlerini içeren ya da yazın artış gösteren hava alanı yolcu sayısını içeren veri kümeleri yıllık bazda periyodiktir.

Bu değerlendirmeler dikkate alınacak olursa, modelin belirli günler ya da aylarda artış ya da azalışı öğrenmesi amacı ile takvim bilgisi kullanılır. Bu çalışmada yine elektrik kullanımı değerlerini içeren 1., 2. ve 3. Veri Kümelerinde de haftalık periyot gözlendiğinden veri modellemesinde bu özellik dikkate alınmıştır.

5.1.2 Olağan dışı bilgi

Veri kümesi içinde rutin değerlerin yanında olağan dışı bir etki görüldüğünde kullanılır. Örneğin elektrik verileri üzerinde yapılan çalışmalarda tatil günleri elektrik kullamının düşük olduğu saptanmıştır. Tatil günü olan 1 Ocaktaki değerin bir çarşamba günü (hafta içi) de olsa düşük seviyede olduğu gözlenmiştir. Bu durum, olağan dışı bir gelişmedir ve öznitelik olarak tatil günlerini kullanmayı gerektirir.

Bir diğer örnek ise hava alanına inen yolcu sayılarının tahmininde verilebilir. Bir şehirdeki hava alanına gelen yolcu sayısının, o şehirde yapılan festival günlerinde artış gösterdiği gözlenmiş olsun. Bu durum, bir sonraki festivalde gelen yolcu

sayısının artacağını gösterir ve bir sonraki festival gününe ait yolcu sayısı tahmini bu özellik dikkate alındığında daha doğru yapılacaktır.

Olağan dışı bilgi, bu tez çalışmasında elektrik kullanımı değerlerini içeren 1, 2. ve 3. Veri Kümelerinde kullanılmıştır.

5.1.3 Geçmiş değerler

Zaman serisi veriler için geleceğe dair bir tahminleme yapıldığında geçmişteki değerler de önemlidir. Bu yüzden modellemede geçmiş değerler de kullanılmalıdır. Örneğin i. gün üzerinde tahmin yapılacak olduğunu düşünürsek, y tahmin edilecek değer olmak üzere,

$$y_i = x_{i-1}, x_{i-2}, \dots, x_{i-n} \quad \textbf{(5.1)}$$

geçmiş bilgilerini de kullanmak yararlı olacaktır. Geçmiş kaç günün kullanılacağı ise optimize edilmesi gereken bir parametredir. Bu parametre, üzerinde çalışılacak veri kümesine göre değişebilir. Yarışmadaki elektrik kullanımı veri kümesinde haftalık periyot gözlendiğinden Chen ve arkadaşları tarafından her bir gün için geçmiş 7 gün kullanılmıştır [13].

5.2 Veri Temsili ve Öznitelik Kullanımı

Zaman serisi verilerin gösterimi genelde

$$y(t) = f(x_t) \quad \textbf{(5.2)}$$

şeklinde gösterilir. Burada $y(t)$, t. güne ait değer, $f(x_t)$ ise t. güne ait geçmiş değerlerin bulunduğu kümedir. Diğer bir gösterim ile

$$y_t = x_{t-1}, x_{t-2}, \dots, x_{t-n} \quad \textbf{(5.3)}$$

olarak tanımlanır. Ancak verilerin zaman içinde farklı değerler aldığı durağan olmayan zaman serileri için bu gösterim uygun değildir. Zaman serileri, her *t* zamanında içinde farklı karakteristikleri içerdiğinden ve veri kümesi analizinde belirlenen özellikler öznitelik olarak kullanılacağından,

$$y_i(t) = f_{i(t)}(x_t) \quad \textbf{(5.4)}$$

gösterimi kullanılacaktır. Bu durumda $y_i(t)$, $t = 1, \dots, n$ zamanlarındaki bütün öznitelikleri içerir.

Diğer bir gösterim ile y_i, i. güne ait değer, x_i ise i. güne ait özniteliklerin bulunduğu kümedir. DVR algoritmasında, özniteliklerin bulunduğu x_i kısmı kullanılarak y_i değerleri öğrenilir. Tahminlemede ise yine tahminde bulunulacak güne ait x_t kısmı kullanılarak y_t değeri tahmin edilir. Veri kümeleri üzerinde Bölüm 5.1'de tespit edilen analizlerin DVR üzerinde modellenebilmesi için bu analizler öznitelik (feature/attribute) olarak kullanılır. Bu çalışmada x_i kısmında kullanılacak öznitelikler:

- Takvim bilgisi (seçime bağlı)
- Olağan dışı bilgi (seçime bağlı)
- Geçmiş değerler

olarak belirlenmiştir.

Diğer bir tanım ile $d_1, \dots, d_k$ takvim bilgisini, od olağan dışı bilgiyi, $x(t), \dots, x(t-k)$ geçmiş verileri, k kullanılacak zaman (örn: gün) sayısını içeren öznitelikler olmak üzere:

$$y(t) = f_t(d_1, \dots, d_k, od, x(t), x(t-1), \dots, x(t-k)) \qquad \textbf{(5.5)}$$

olarak gösterilir.

Bu tez çalışmasında, takvim ve olağan dışı bilgilerin öznitelikleri temsili için ikili (binary) değerler kullanılmıştır. Pazartesi – Cumartesi günleri için 1, Pazar günü için 0 kullanılmıştır.Aynı zamanda i. gün eğer bir olağan dışı gün ise (örneğin tatil) 1 ile, rutin bir değer ise 0 ile temsil edilmiştir. Geçmiş veriler ise gerçek değerleri ile kullanılmıştır.

Çizelge 5.1 : Özniteliklerin kullanımı.

x_i												y_i
P	Slı	Çar	Per	Cu	Cmt	Pz	Tatil	G_{i-7}	G_{i-6}	…	G_{i-1}	G_i
0	0	1	0	0	0	0	1	818	730	…	796	800

Örnek olarak Çizelgede 5.1'de Veri Kümesi 1'in i. gününe ait öznitelikler görülmektedir. i. günün bir Çarşamba günü ve resmi tatil olduğunu varsayarsak,

Çarşamba gününü temsil eden öznitelik için 1, diğer günleri temsil eden öznitelikler için 0 kullanılır. i. gün eğer resmi tatil ise Çizelge 5.1'de olduğu gibi tatil (olağan dışı) özniteliğine 1 yazılır. Tatil olmasa idi 0 ile temsil edilecekti. Geçmiş değerlere ait öznitelikler (G_{i-d}) için ise i. zamanın (günün) geçmiş değerleri yazılarak x_i kullanılır. Böylece x_i değerleri modellenmiş olunur. Öğrenilecek/tahmin edilecek veri için ise y_i ifadesi kullanılır. Bu işlem bütün günler için yapılarak model oluşturulur.

DVR kulanımında ayrıca geçmiş verileri içeren öznitelikler 0 ve 1 arasına ölçeklenir ve bu işleme normalizasyon adı verilir. Veri kümesindeki en büyük değer 1, en küçük değer ise 0 olmak üzere tüm değerler 0 ve 1 arasına normalize edilir. Bu tez çalışmasında alınan sonuçlar MATLAB üzerinde LIBSVM [48] kütüphanesi kullanılarak elde edilmiştir.

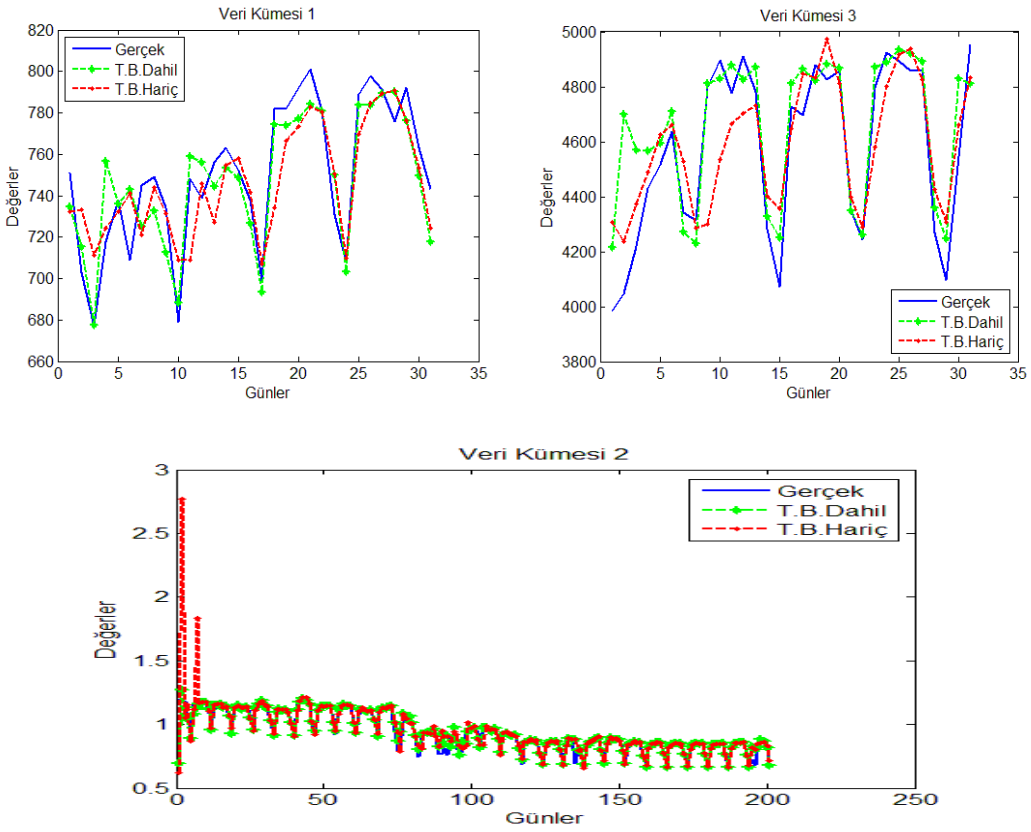
5.2.1 Takvim bilgisi özniteliğinin tahmin başarımına etkisi

Chen ve arkadaşları [13] kullandıkları elektrik kullanım değerlerini içeren veri kümesinde takvim bilgilerini kullanmadan da sonuçlar edinmiş, ancak en yüksek başarımın, takvim bilgisi öznitelikleri ile elde edildiğini göstermişlerdir.

Takvim bilgisi etkisini gözlemleyebilmek için, bu tez çalışmasında da elektrik kullanım değerlerini içeren (yarı periyodik) 1, 2, ve 3. veri kümelerinin, takvim bilgisi içeren ve içermeyen DVR tahmin sonuçları sonuçları incelenmiştir.

Çizelge 5.2 : Takvim bilgisi özniteliğinin tahmin başarımına etkisi.

Veri Kümesi	Metod	Takvim Bilgisi	OKH	OMHY (%)
1	DVR	Evet	**245.6424**	**1.7368**
	DVR	Hayır	409,4291	2,1922
2	DVR	Evet	**0.0025**	**2.4856**
	DVR	Hayır	0.0179	3.6375
3	DVR	Evet	**28825.70**	**2.6295**
	DVR	Hayır	31467.18	3.1094

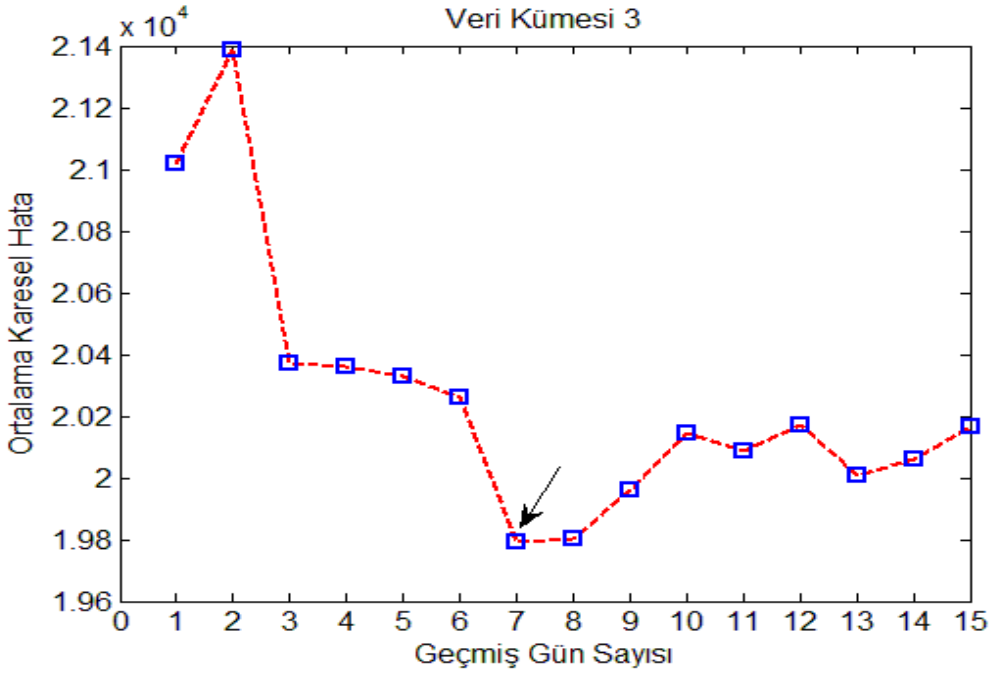


Şekil 5.2 : Takvim bilgisi özniteliğinin tahmin başarımına etkisi.

Çizelge 5.2 ve Şekil 5.2'de görüldüğü üzere hafta bazında yarı-periyodik bir yapı içeren elektrik veri kümelerinde (Veri Kümesi 1,2 ve 3) takvim bilgisini kullanmak tahmin başarımını artırmaktadır. Ancak peiyodiklik içermeyen diğer veri kümelerinde takvim bilgisini kullanmak başarımı daha da düşürmektedir. Bu yüzden bu tezdeki çalışmada yarı periyodiklik içeren ilk 3 veri kümesi için takvim bilgisi özniteliği kullanılmıştır.

5.2.2 Geçmiş gün değerleri özniteliğinin tahmin başarımına etkisi

Tahmin edilen/öğrenilen bir y_i değerine ait özniteliklerinin bulunduğu x_i kısmında geçmişe ait kaç günün bulunduğu da belirlenmesi gereken bir parametredir. Bunu belirleyebilmek için ise eğitim kümesinde en iyi tahmin başarımı sağlayan geçmiş gün sayısının belirlenmesi gerekir. Örnek olarak şekil 5.3'te Veri Kümesi 3'ün eğitim kümesine ait tahmin başarımının (OKH) geçmiş gün sayıları kullanılarak elde edilmiş hata değerleri gösterilmektedir.



Şekil 5.3 : Geçmiş gün bilgisi özniteliğinin tahmin başarımına etkisi.

Eğitim hatalarına bakıldığında, bir y_i değeri için, Veri Kümesi 3'te geçmiş 7 gün kullanılmasının en küçük OKH dğeri ile en iyi başarımı verdiği görülmektedir. Bu yüzden 3. Veri Kümesi için her günün eğitilmesi ve tahmin edilmesi (test) işleminde geçmiş 7 günün değerini kullanmak daha yüksek bir tahmin başarımı verecektir.

5.3 DVR Algoritmasında Kullanılan Parametreler

DVR parametreleri, tahmin başarımı üzerinde önemli bir etkendir. Kullanılan parametreler arasında aynı zamanda çekirdek parametreleri de yer almaktadır. Literatürdeki çoğu çalışmada YTF çekirdeğinin en iyi başarımı verdiği gösterilmiş [38], bir çalışmada ise laplace çekirdeğinin çok başarılı sonuçlar verdiği belirtilmiştir [43]. Ancak belirtilen çalışmadaki model kurularak laplace çekirdeği ile uygulandığında, o veri kümesi üzerinde çok başarılı sonuçlar verdiği, farklı veri kümesinde hatta aynı veri kümesinin test kısmı adedi değiştirildiğinde sonuçların aynı başarımda olmadığı gözlenmiştir. Bu sebeple bu tez çalışmasında YTF çekirdeği kullanılmış ve parametre optimize etme işleminde YTF parametresi (ɣ) kullanılmıştır.

Literatürde DVR ile elde edilen bazı zaman serisi tahminlemeleri çalışmalarında [13], ε parametresi sabit alınarak (genellikle 0.01) C ve ɣ parametreleri optimize edilmiş ve buna göre sonuçlar alınmıştır. Yapılan incelemede ε parametresini

sabitleyerek optimize etme işlemini 2 parametre (C ve γ) için yapmanın parametre bulma süresini kısalttığı, ancak 3 parametre (C γ ve ε) ile kullanılarak yapılan tahminlere göre tahmin başarımını düşürdüğü görülmüştür.

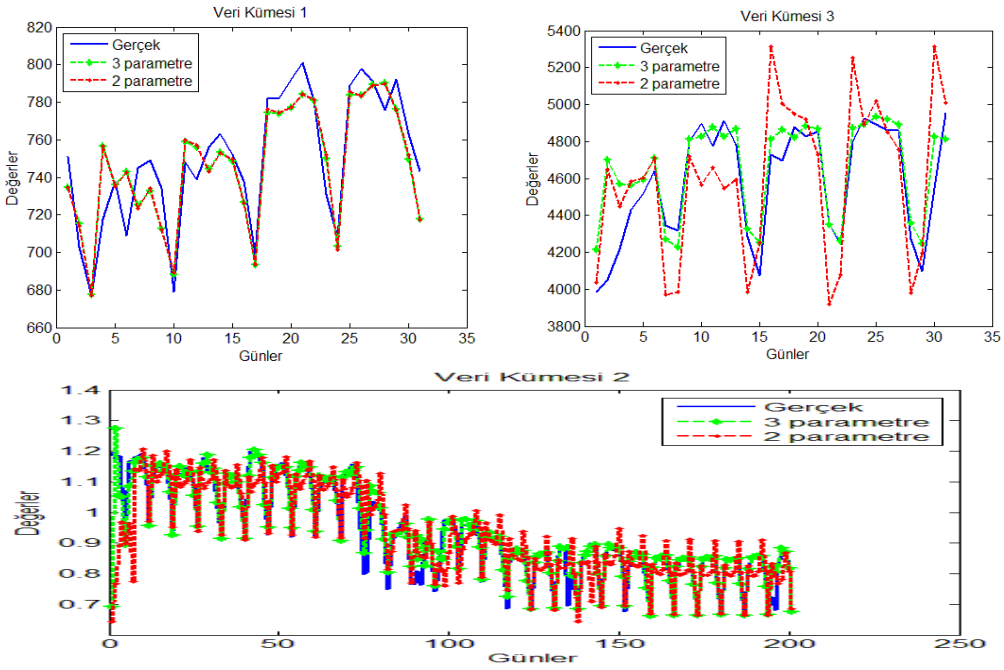
Çizelge 5.3 ve Şekil 5.4'te bu tez çalışmasındaki ilk 3 veri kümesi için gösterilen tahmin başarı ölçümlerinde, 3 parametre (C γ ve ε) ile alınan sonuçların 2 parametre (C ve γ) ile alınan sonuçlardan daha başarılı olduğu görülmektedir. Çizelge 5.3 incelendiğinde, Veri Kümesi 1 için başarım farkı oldukça az olmasına rağmen 3 parametre ile bulunan tahminlemenin yine de daha başarılı olduğu görülmektedir. Diğer veri kümelerinde ise tahmin başarımı farkı oldukça yüksektir.

Ayrıca Çizelge 5.3'te görüldüğü üzere 2 ve 3 parametre kullanılarak elde edilen optimizasyon sürelerinde 3 parametre kullanımının süreyi artırdığı görülmektedir. Sonuçlar Intel(R) Core (TM) i7 2.20 GHz işlemci, 8 GB RAM, Windows 8 (64 Bit) işletim sistemi ve Matlab R2011a (64 Bit) ortamında elde edilmiştir.

Optimizasyon süresinin artmasına ragmen, daha yüksek başarımlı tahmin sonuçları vermesi sebebi ile bu tez çalışmasındaki DVR algoritmalarının sonuçları, 3 parametre optimizasyonu ile elde edilmiştir.

Çizelge 5.3 : 2 ve 3 parametre optimizasyonunun tahmin başarımına etkisi.

Veri Kümesi	Metod	Parametre Sayısı	OKH	OMHY (%)	Zaman (Sn)
1	DVR	3	**245.6424**	**1.7368**	52.36
	DVR	2	249.9442	1.7503	50.44
2	DVR	3	**0.0025**	**2.48**	394.03
	DVR	2	0.0074	5.74	208.53
3	DVR	3	**28825.70**	**2.62**	53.73
	DVR	2	87659.21	5.13	34.98



Şekil 5.4 : 2 ve 3 parametre optimizasyonunun tahmin başarımına etkisi.

5.4 DVR Algoritmasının AR Modeli ile Karşılaştırılması

DVR Algoritması ile alınan sonuçlarda en iyi başarımı elde etmek için öznitelik seçimi, parametre kullanımı ve çekirdek tipi etkileri incelenmesi sonucu Çizelge 5.4'te belirtilen bilgiler kullanılmıştır.

Çizelge 5.4 : Destek vektör regresyonu modellemesi.

Destek Vektör Regresyonu		
Öznitelik Seçimi	Bütün Veri Kümeleri	Geçmiş Gün Özniteliği
	Periyodiklik İçeren Veri Kümeleri	Olağan Dışı Durum
		Takvim Bilgisi
Parametre Kullanımı	Bütün Veri Kümeleri	Ceza Paremetresi (C)
		YTF Çekirdek Paremetresi (γ)
		Hata Parametresi (ε)
Çekirdek Kullanımı	Bütün Veri Kümeleri	Yarıçap Temelli Fonksiyon Çekirdeği

Bu tez çalışmasında ayrıca Çizelge 5.4'te kullanılan bilgiler ile kurulan DVR modelleri, literatürde yaygın olarak kullanılan Auto Regressive (AR) ve Auto Regressive Moving Average (ARMA) modelleri ile karşılaştırılmıştır. AR ve ARMA modellerinin parametreleri her veri kümesi için en iyi tahmin başarımını sağlayan değerler ile belirlenmiştir. Çizelge 5.5'te görüldüğü üzere DVR modeli 6. veri kümesi hariç bütün veri kümelerinde yüksek tahmin başarımına sahiptir. 6. veri kümesinde ise ARMA modeli ile DVR modeli tahmin başarıları çok yakın olmakla beraber ARMA modeli az bir farkla daha başarılı olarak görülmektedir. Ancak DVR modelinin diğer veri kümelerinde daha başarılı tahminler vermesi, DVR'nin kullanılması için tercih sebebidir.

Çizelge 5.5 : DVR ve AR modellerinin karşılaştırılması.

Veri Kümesi	Model	OKH	OMHY (%)
1	DVR	**245.642**	**1.73**
	AR	8404.7	11.78
	ARMA	561.710	2.70
2	DVR	**0.00258**	**2.48**
	AR	0.0060	4.70
	ARMA	0.0041	3.99
3	DVR	**28825.7**	**2.62**
	AR	51631.13	3.86
	ARMA	30794.55	3.06
4	DVR	**0.000333**	**0.68**
	AR	0.00047	0.86
	ARMA	0.000335	0.69
5	DVR	**5.98e-05**	**0.32**
	AR	9.28e-05	0.42
	ARMA	6.19e-05	0.33
6	DVR	0.00014	0.65
	AR	0.00015	0.67
	ARMA	**0.00013**	**0.64**
7	DVR	**1.12708**	**8.66**
	AR	2.04850	12.32
	ARMA	1.3334	9.57

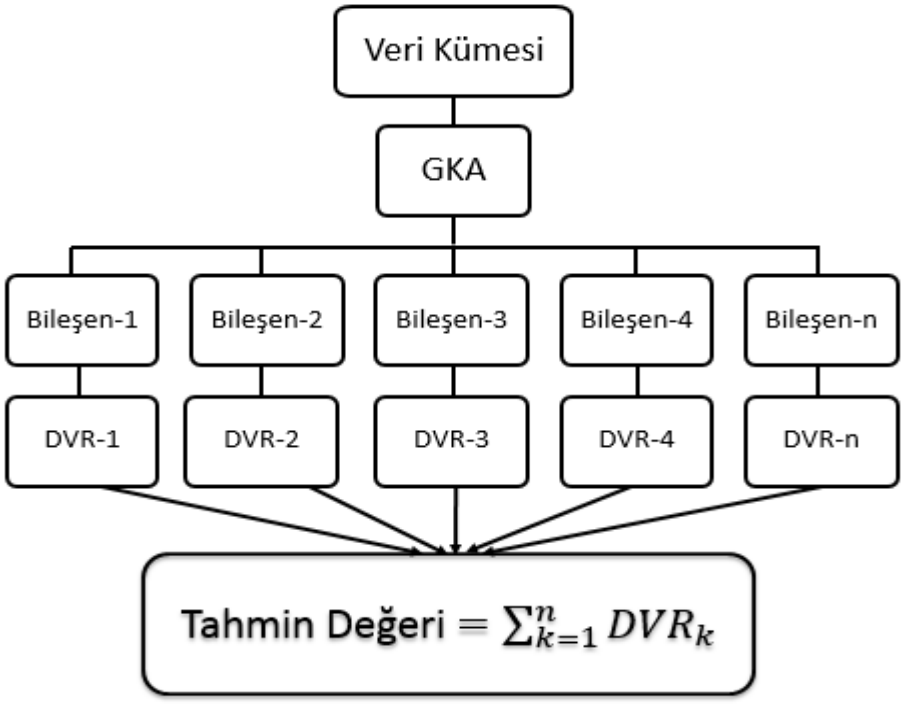
6. GKA TABANLI DVR (GKA-DVR) BÜTÜNLEŞİK MODELİ

DVR algoritmasının zaman serisi veriler için oluşturulan modelin başarısı [13] ve GKA algoritmasının durağan olması ya da olmaması şartını aramaksızın zaman serisi verileri durağana yakın bileşenlere ayırması, GKA ile DVR algoritmalarının birleşmesi fikrini ortaya çıkarmıştır.

Literatürde yer alan çoğu GKA-DVR bütünleşik modellerinin temeli, DVR algoritmasını orjinal veri kümesi üzerinde uygulamak yerine GKA ile elde edilen bileşenler üzerinde uygulayarak, her bir bileşen için yapılan tahminlerin toplanması işlemine dayanmaktadır. Bu yaklaşımda $p(t), t$ zamanındaki tahmin değeri ve $p_{IMF_i(t)}$ ile $p_{r(t)}$, t zamanındaki bileşen tahmin değerleri, n IMF sayısı olmak üzere:

$$p(t) = \sum_{i=1}^{n} p_{IMF_i(t)} + p_{r(t)} \quad \textbf{(5.6)}$$

ifadesi kullanılır.



Şekil 6.1 : GKA-DVR bütünleşik modeli.

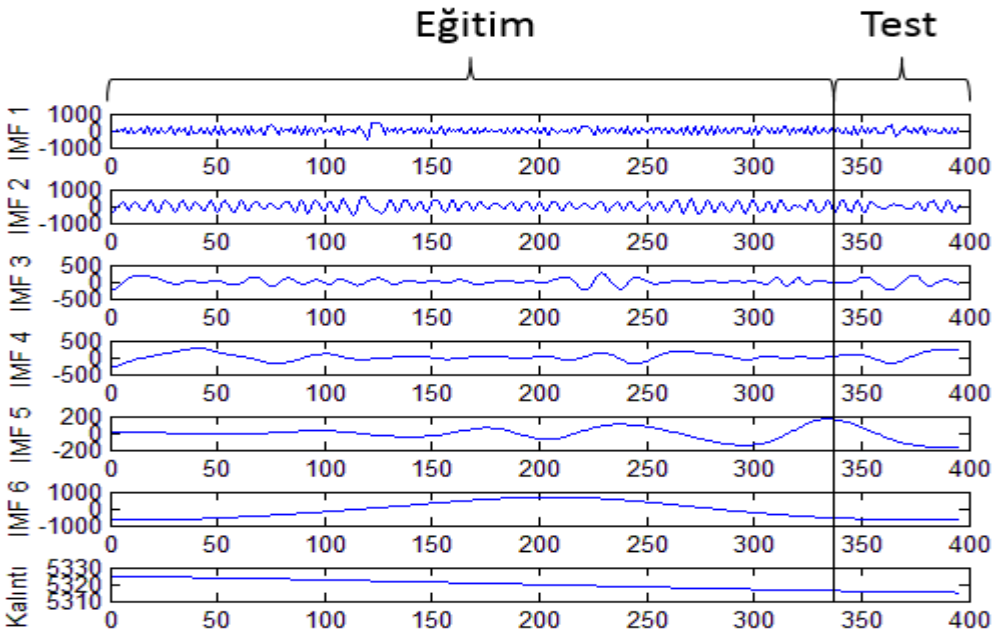
Şekilden 6.1'den anlaşılacağı üzere, GKA-DVR bütünleşik modelinde;

1. Üzerinde çalışılan zaman serisi veri kümesi olan $x(t)$, GKA algoritması ile bileşenlerine ayrılır.
2. Her bileşenin kendi eğitim kümesi kullanılarak, DVR algoritması ile her bileşen için bağımsız olarak bir model oluşturulur. Buradaki önemli nokta her bir bileşenin özelliği farklı olacağından, DVR algoritması ile oluşturulan modelin farklı bileşenler üzerinde farklı parametreler/öznitelikler/çekirdekler kullanılabilmesidir.
3. Her bileşenin test kümesi için, kendi bileşen modeli kullanılarak tahminleme işlemi yapılır.
4. Bileşenlerin tahmin değerleri toplanarak orjinal veri kümesine ait asıl tahmin değeri elde edilir.

Bu tez çalışmasında GKA-DVR bütünleşik modelini zaman serisi veriler üzerinde uygularken, en iyi tahmin başarımını alabilmek için Bölüm 5'te anlatılan öznitelikler kullanılmıştır.

6.1 GKA– DVR Modelinin Önceki Çalışmalarda Kullanımı

GKA bileşenlerinin özellikleri ve her bir bileşen üzerinde uygulanan DVR algoritmasının tahmin başarıları, elde edilecek toplam başarı üzerinde etkili olmaktadır. Bu yüzden en iyi tahmini elde edebilmek için, modelleme işleminde her bileşenin test kümesini en iyi tahminleyecek DVR modeli uygulanmalıdır. Her bileşen tahminlemesinin diğer bileşenlerden bağımsız olması, bileşenler için oluşturulan modellerde farklı parametre, çekirdek ve öznitelik kullanılması imkanını sunmaktadır. Bileşen tahminlemesinin yüksek başarımlı olabilmesi için aynı zamanda bileşene ait eğitim kümesi ile test kümesinin de aynı özellikte olması gerekir. Bu şekilde her bileşenin eğitim kümesi üzerinde kurulacak model ile aynı özellikteki test kümesi doğru tahminlenebilecek ve bileşenlerin tahminleri toplandığında yüksek başarımlı bir genel tahmin elde edilebilecektir.

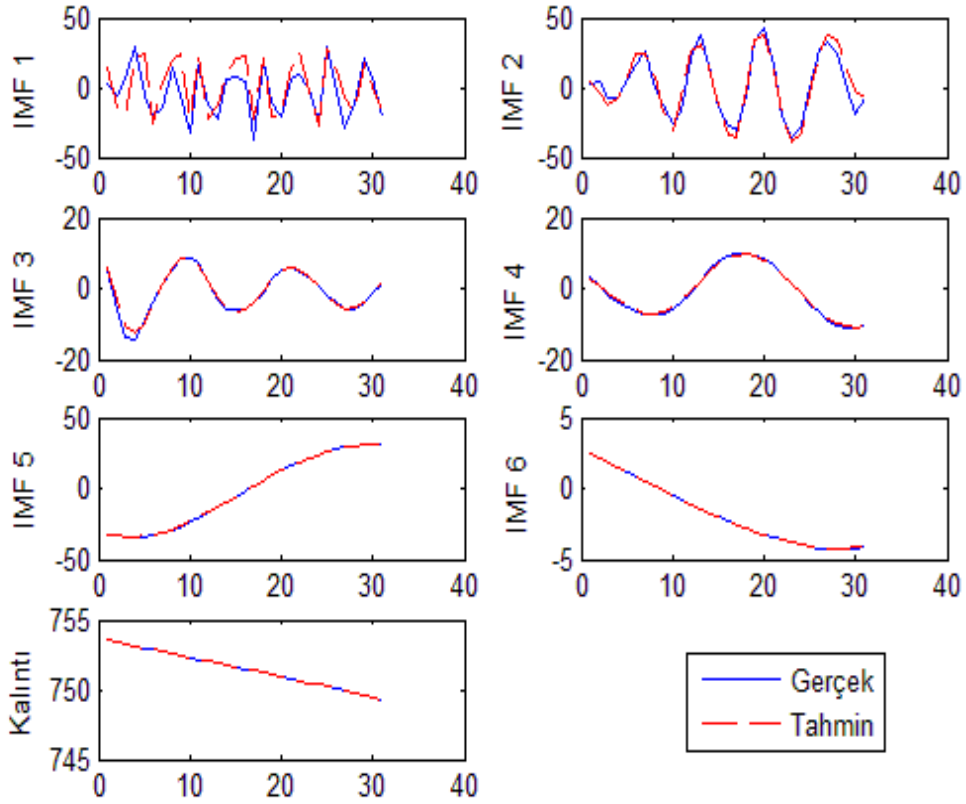


Şekil 6.2 : GKA bileşenleri (Veri Kümesi 3).

Zaman serisi verilerde kullanılan çoğu GKA-DVR çalışmalarında [26-27], veri kümesinin GKA ile bileşenlerine ayırma işlemi, veri kümesinin tümü (test kümesi dahil) kullanılarak yapılmıştır. Bu şekilde ayrılan bileşenler, test kümesinin de elde bulunulduğu varsayılarak elde edilmiştir. GKA bileşenleri tüm veri kümesi kullanılarak elde edildikten sonra Şekil 6.2'de olduğu gibi eğitim ve test kümelerine ayrılmaktadır. Bu yaklaşımla, veri kümesinin tümünün bileşenleri Şekil 6.2'de olduğu gibi durağana yakın bir özellikte ortaya çıkmaktadır. Bileşenlerin eğitim ve test kümeleri aynı özellikleri taşıdığından, her bileşenin test kümesi için DVR ile tahmin yapıldığında başarılı sonuçlar alınabilmektedir. Örnek olarak Şekil 6.3'te Veri Kümesi 3'ün test bileşenleri ve bu bileşenler üzerinde DVR ile yapılan tahminler görülmektedir.

Bileşenler için kurulan modellerde her bileşen için en iyi başarımı veren farklı C, γ ve ε parametreleri ile YTF çekirdeği kullanılmıştır. Ayrıca periyodiklik özelliği gösteren 1.2. ve 3. Veri Kümelerinde olağan dışı durum ve takvim bilgisi de öznitelik olarak eklenmiştir. İlk 3 veri kümesinde olağan dışı durum ve takvim bilgisi öznitelik olarak eklenmediğinde daha düşük tahmin başarımı elde edildiği gözlenmiştir. Şekil 6.3'te görüldüğü üzere bileşen tahminlerinde en kötü başarım ilk bileşendedir. Bunun sebebi ise ilk bileşenin diğerlerine göre daha yüksek frekansta

olmasıdır. Verinin büyük bir kısmını içeren Kalıntı bileşeni tahmini ise oldukça yüksek bir başarıma sahiptir. Böylece bu yöntem ile verinin büyük bir kısmı sadece kalıntı bileşeni üzerinde yapılan tahmin ile yüksek bir başarımla tahmin edilmiş olunur.



Şekil 6.3 : GKA test bileşenleri ve üzerinde yapılan tahminler (Veri Kümesi 1).

Bileşen tahminleri elde edildikten sonra tahminleri toplamı, veri kümesi için genel tahmin değerini vermektedir.

Bu yaklaşım, bu tez çalışmasında yer alan bütün veri kümeleri üzerinde uygulanmıştır. Çizelge 6.1 ve Şekil 6.4 ile 6.5'te tüm veri kümeleri için bu yaklaşımla (GKA-DVR %100) yapılan tahminler ile DVR tahminleri başarılarının matematiksel ve grafiksel olarak karşılaştırılmaları görülmektedir.

Çizelge 6.1 : GKA-DVR (%100) modeli ile tekli DVR modeli tahminleri karşılaştırmaları.

Veri Kümesi	Metod	OKH	OMHY (%)	YD (%)	DAr (%)	DAz (%)
1	DVR	245.642	1.73	70	61.53	81.25
1	GKA-DVR (%100)	**185.684**	**1.50**	**83.33**	**84.61**	**87.50**
2	DVR	0.00258	2.48	75.50	77.22	73.73
2	GKA-DVR (%100)	**0.00098**	**1.9806**	**77.50**	**79.20**	**75.75**
3	DVR	28825.7	2.62	60	56.25	64.28
3	GKA-DVR (%100)	**16794.3**	**2.16**	**63.33**	**62.50**	**64.28**
4	DVR	0.00033	0.68	48.38	56.25	40
4	GKA-DVR (%100)	**0.00010**	**0.37**	**83.87**	**87.50**	**80**
5	DVR	5.98e-05	0.32	43.62	42.02	45.56
5	GKA-DVR (%100)	**2.89e-05**	**0.23**	**68.45**	**65.21**	**72.15**
6	DVR	0.00014	0.65	49.24	50	48.78
6	GKA-DVR (%100)	**0.00012**	**0.61**	**80.40**	**79.41**	**82.00**
7	DVR	1.12708	8.66	71.68	71.73	75.37
7	GKA-DVR (%100)	**0.35651**	**4.88**	**86.73**	**87.68**	**90.29**

Çizelge 6.2 : GKA-DVR (%100) modeli ile tekli DVR modeli tahminleri Karışıklık Matrisleri.

Yöntem		Veri Kümesi 1		
			Tahmin	
			P	N
1*	Gerçek	P	8	5
		N	3	14
2*		P	11	2
		N	3	14

Yöntem		Veri Kümesi 2		
			Tahmin	
			P	N
1*	Gerçek	P	78	23
		N	26	73
2*		P	80	21
		N	24	75

Yöntem		Veri Kümesi 3		
			Tahmin	
			P	N
1*	Gerçek	P	9	7
		N	5	9
2*		P	10	6
		N	5	9

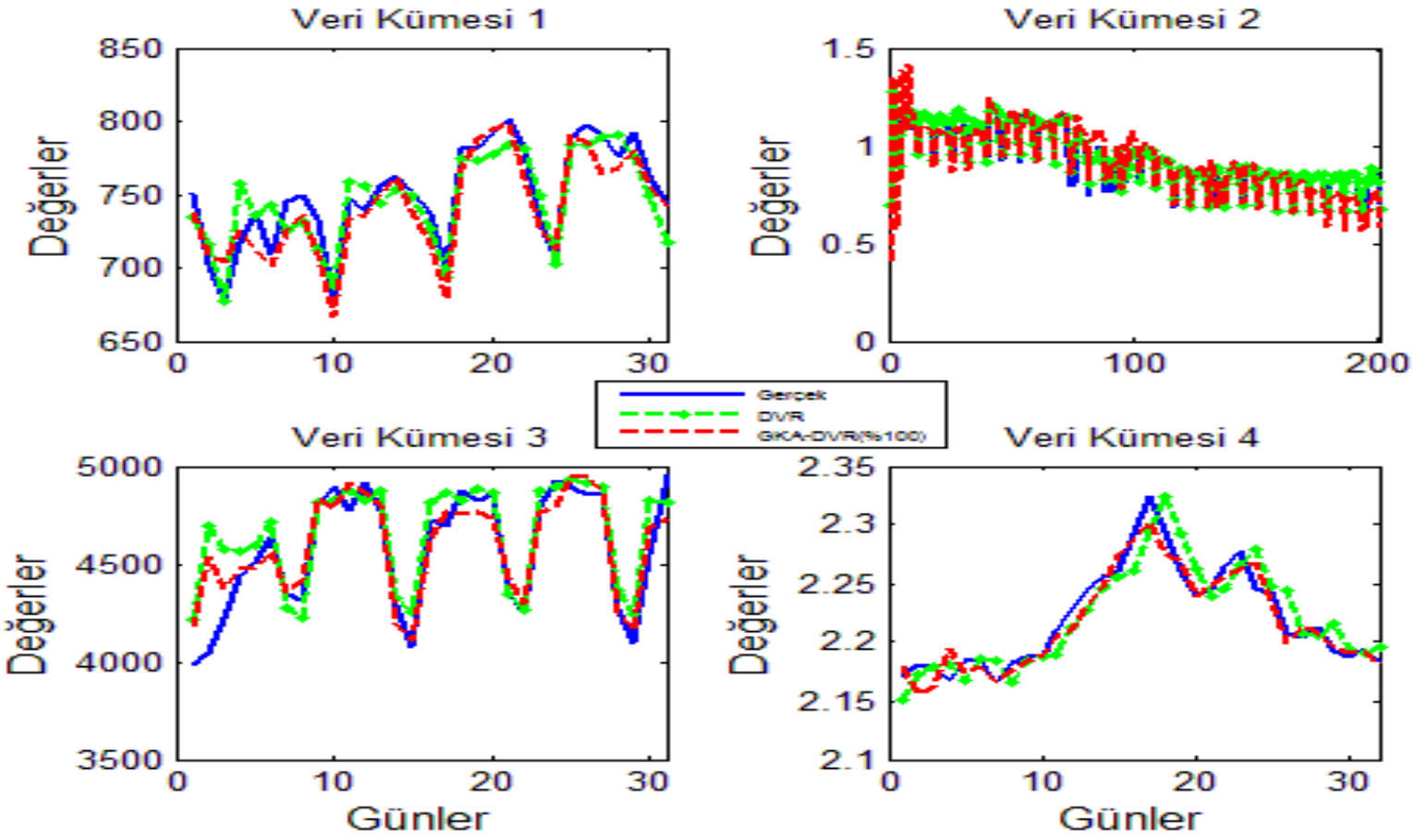
Yöntem		Veri Kümesi 4		
			Tahmin	
			P	N
1*	Gerçek	P	9	7
		N	9	6
2*		P	14	2
		N	3	12

Yöntem		Veri Kümesi 5		
			Tahmin	
			P	N
1*	Gerçek	P	32	37
		N	43	37
2*		P	45	24
		N	22	58

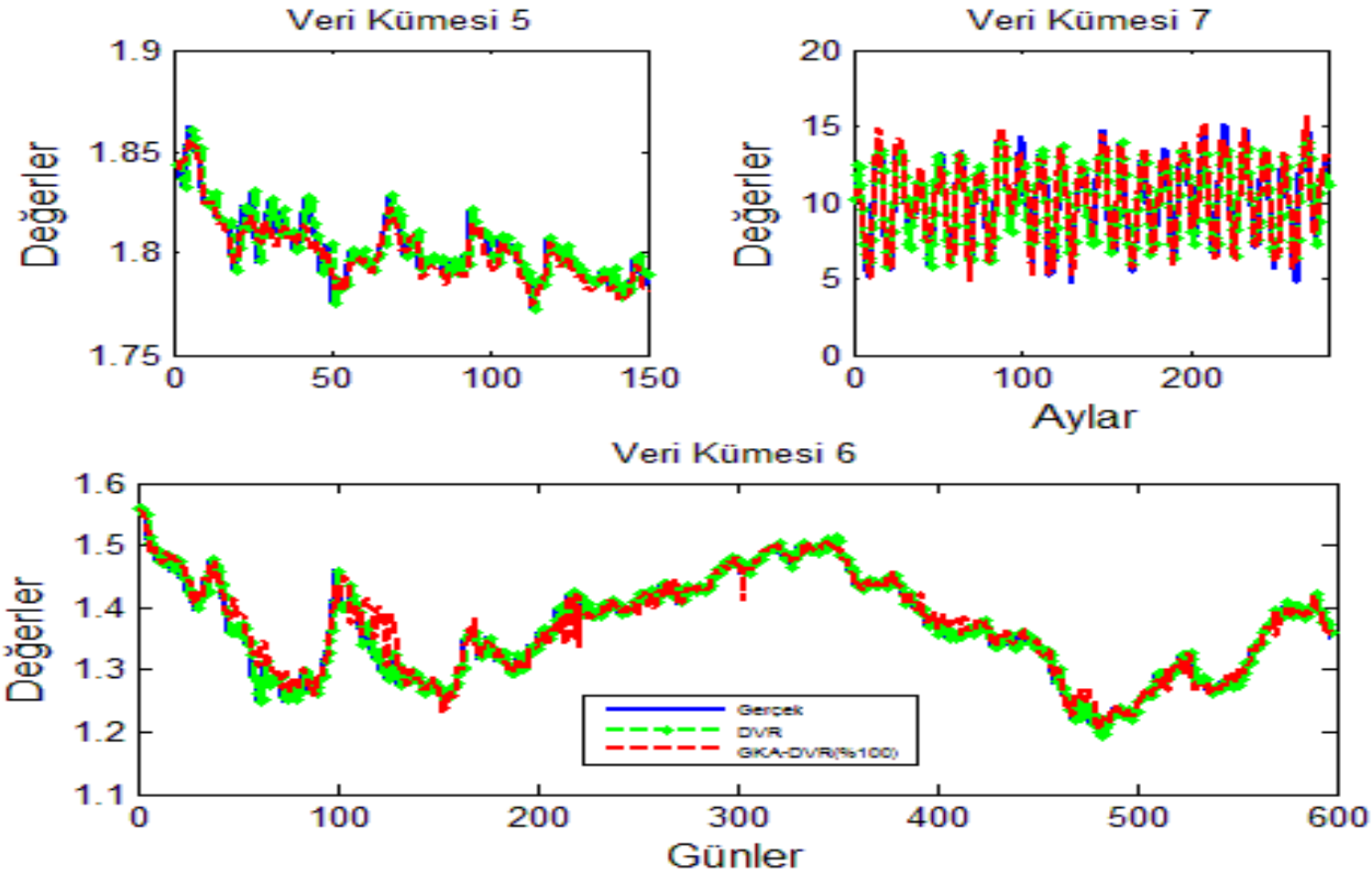
Yöntem		Veri Kümesi 6		
			Tahmin	
			P	N
1*	Gerçek	P	153	153
		N	150	141
2*		P	243	63
		N	52	239

Yöntem		Veri Kümesi 7		
			Tahmin	
			P	N
1*	Gerçek	P	99	39
		N	37	104
2*		P	121	17
		N	17	124

Yöntem 1: Tekli DVR Modeli. Yöntem 2: DKA-DVR (%100) Modeli.



Şekil 6.4 : GKA-DVR (%100) modeli ile tekli DVR modeli tahminleri-1.



Şekil 6.5 : GKA-DVR (%100) modeli ile tekli DVR modeli tahminleri-2.

Her iki modelin karşılaştırılmasında, tüm veri kümesini (test kümesi dahil) (%100) kullanarak bileşenlere ayırma işlemini yapan GKA-DVR (%100) modelinin bütün veri kümelerinde her başarı ölçümü için çok daha başarılı olduğu görülmektedir. Ancak test kümesinin gerçek uygulamalarda var olmadığı açıktır ve GKA ile bileşenlerine ayırma işleminde bütün veri kümesi kullanılamayacaktır. Böylece eğitim ve veri kümesi bileşenleri Şekil 6.2'de görüldüğü üzere aynı özellikte olmayacak ve yapılan tahminler Şekil 6.3'teki başarımda olamayacktır. Bu sebeple bu yaklaşım gerçek uygulamalar için sonuç alınamayacak bir yöntemdir.

Bu yaklaşımın gerçek uygulamalar için uygulanabilmesi amacı ile bileşenler test kümesi kullanılmadan eğitim bileşenleri üzerinden elde edilmeli, daha sonra test bileşenleri kullanılmalıdır. Ancak bu durumda GKA üzerinde son nokta problem ile karşılaşılmakta, Bölüm 6.2'de açıklanacağı üzere test bileşenleri ile eğitim bileşenleri aynı özellikte olmamaktadırlar. Literatürde son nokta problemine kesin olarak çözüm getiren bir bulunmamakla birlikte, bir çalışmada [49] 4 adet model önerilmiştir. Çalışmada en başarılı çözümün Rato [50] metodu ile gerçekleştirebildiği gösterilmiştir. Ancak yapılan tahmin başarımının sadece Simetrik-OMHY ölçümü ile yapılması, Simetrik-OMHY ölçümünün zaman serisi tahminlerinde gerçek başarımı göstermeyeceği [51], Rato metodunun bu tez çalışmasındaki başarı ölçümlerinde yüksek başarımlar vermemesi göz önüne alındığında bu çalışma da kesin çözüm içermemektedir.

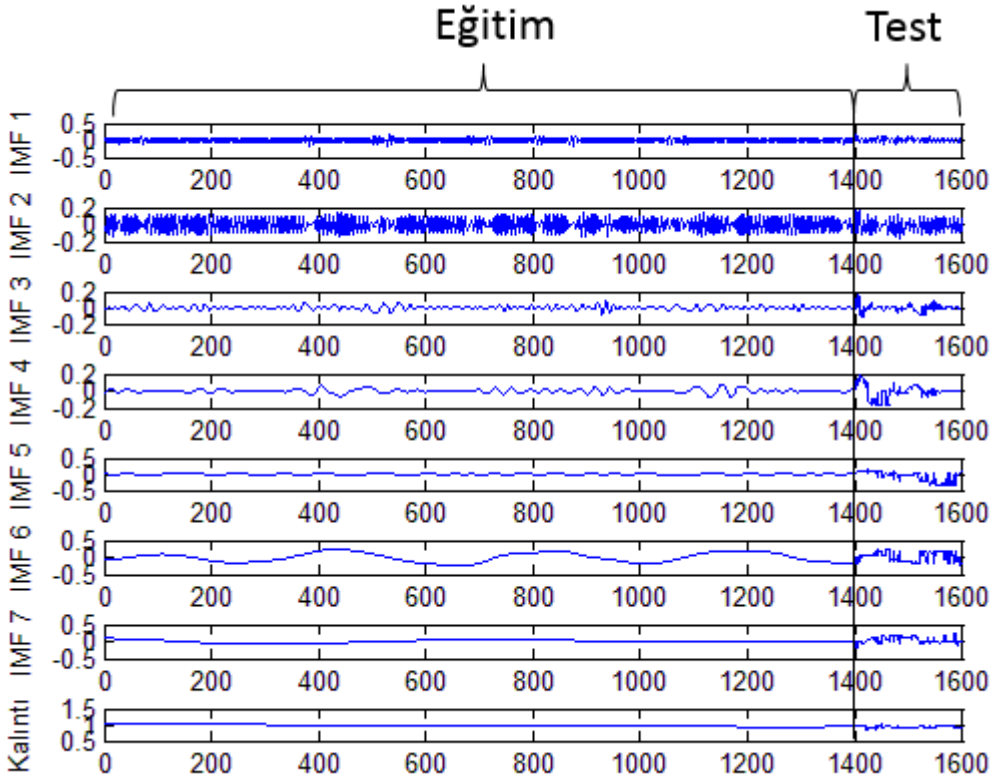
6.2 Yinelemeli GKA-DVR Modeli

Tahminleme modelleri bir sonraki günün, haftanın ya da ayın değerlerini edinmek için kurulabilir. Örneğin bir sonraki adımın (one-step ahead) tahminlendiği bir modelde dikkat edilmesi gereken nokta, yapılan tahminin test kümesinin tümünü aynı anda değil, bir sonraki adımını (one-step ahead) teker teker tahmin ediyor oluşudur.

Oluşturulan model bir sonraki günü tahmin etmekte ise, test kümesinde bulunan 30 günlük veri için 30 kez günlük tahmin yapılmakta ve modelin ne kadar başarılı olduğu bu şekilde ölçülmektedir. Örneğin test kısmının birinci günü tahmin edildikten sonra, yine 1. günün gerçek değeri kullanılarak 2. günün tahmini, 2. günün gerçek değeri kullanılarak 3. günün tahmini yapılır. Bu işleme yinelemeli olarak

eldeki son test verisi tahmin edilene kadar devam edilir. Son değer olan 30. günü tahmin ederken, ondan önceki 29 günün gerçek değeri kullanılmış olur.

GKA-DVR modelinde yapılan tahminin ne kadar başarılı olduğunu karşılaştırabilmek için ise test kümesi değelerini yine gün bazında (teker teker) tahmin etmek gerekir. Bu durumda GKA modeli ile bileşenlere ayırma işlemi daha önceki çalışmalarda olduğu gibi test kümesinin tümünü kullanarak değil, gün bazında (teker teker) alınarak yapılmalıdır. Bu şekildeki bir yaklaşım ile GKA alma işleminde, test kümesinin 1. değeri veri kümesi eğitim kısmına katılarak bileşenler alınır ve test kümesinin 1. değerine ait bileşenler saklanır. Bu işlem yinelemeli olarak test kümesinin son değerine kadar devam edilir. Her eklenen günde alınan bileşen sayısı farklı olabileceğinden bileşen sayısı sınırlanır. Şekilde 6.6'da Veri Kümesi 2'ye ait olan ve GKA ile yinelemeli olarak alınan test kümesi bileşenleri görülmektedir. Bileşenlerin sayısı 8 ile sınırlandırılmıştır.



Şekil 6.6 : Yinelemeli GKA ile alınan test bileşenleri (Veri Kümesi 2).

Ancak bu yaklaşımda yinelemeli olarak alınan test bileşenleri, eğitim bileşenleri ile aynı özellikte olmamaktadır. Şekil 6.6'da görüldüğü üzere test kümesi bileşenleri üzerinde gürültü ve düzensiz değişimler mevcut olmakla birlikte, bu bileşenler

durağanlığa yakın olma özelliklerini kaybetmişlerdir. Bu durumda eğitim kümesindeki durağana yakın bileşenler ile test kümesindeki durağan olmayan bileşenler tahmin edilemeyecektir.

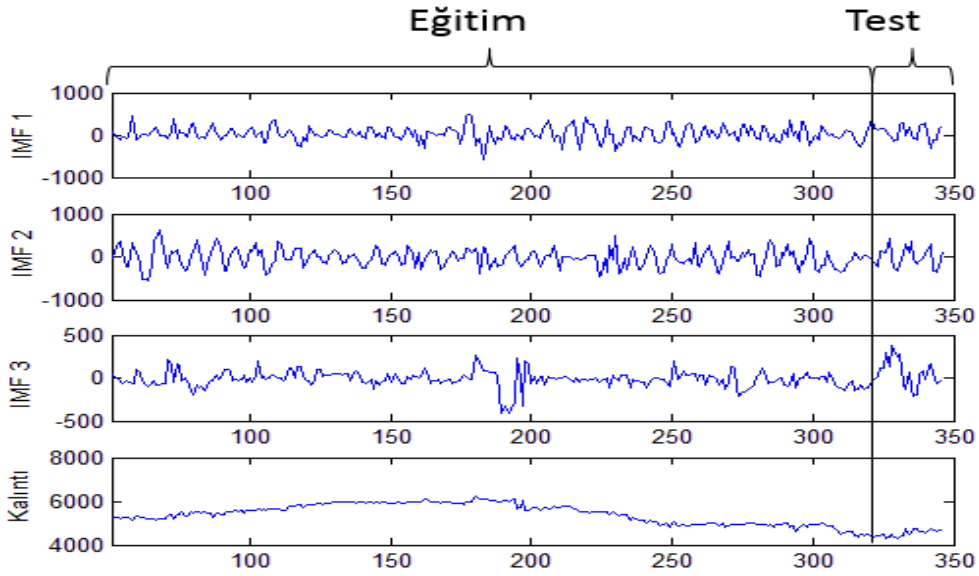
Öğrenme başarımını artırabilmek amacı ile eğitim kümesi bileşenleri ile test kümesi bileşenleri aynı özelliklere sahip olmalıdır. Bu amaçla eğitim kümesi bileşenleri de yinelemeli olarak alınmıştır. Ancak bu yaklaşımda 2 durum ile karşılaşılmıştır:

- Zaman serisindeki ilk verilerin (örneğin ilk 5) GKA algoritmasında bileşenlerine ayrılamaması
- Veri kümesine her değer (gün) eklendiğinde elde edilen bileşen sayısının değişmesi

Bu 2 durum, belirlenmesi gereken 2 parametreyi ortaya çıkarmıştır:

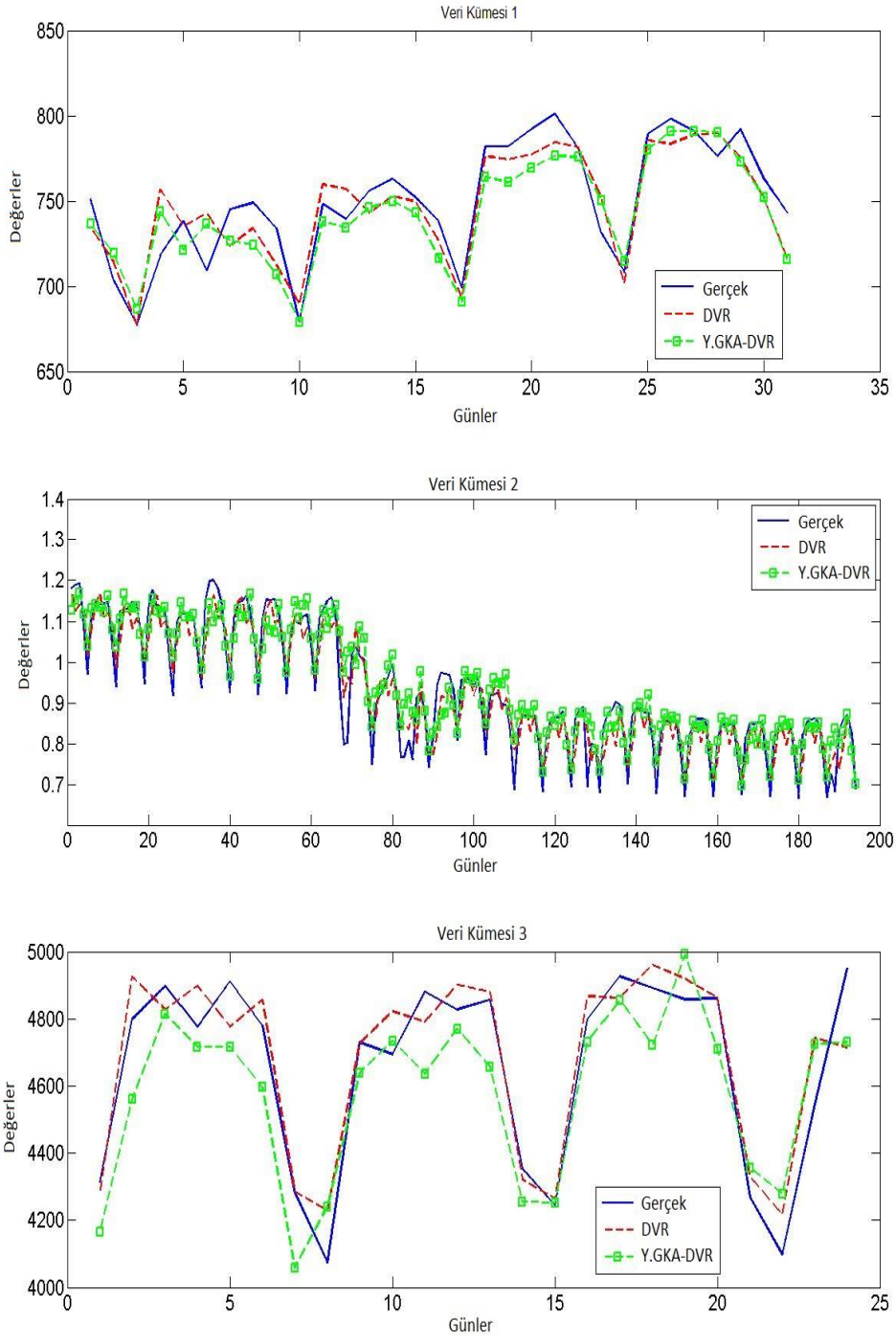
- GKA Algoritması bileşenlerinin alınabilmesi için veri kümesinin kaçıncı verisinden başlanacağı (Parametre 1)
- GKA Algoritmasının kaç bileşenden oluşacağı (Parametre 2)

Bu iki parametre ise eğitim kümesi içinde farklı sayılar ile optimize edilmiştir. Eğitim kümesi tahminlemesinde en yüksek başarımı veren bu 2 parametre, test kümesi üzerine uygulanmıştır. Şekil 6.7'de örnek olarak Veri Kümesi 3 için 50. veriden başlanarak 4 bileşen alınması ile oluşan yeni bileşenler gösterilmektedir.



Şekil 6.7 : Yinelemeli olarak elde edilen GKA bileşenleri.

Elde edilen yeni bileşenler üzerinde, her birinin test kümesi için tahminlemeler yapılarak toplandığında, gerçek veri kümesi için tahmineleme değerleri alınmıştır. Şekil 6.8'de Yinelemeli GKA-DVR modelinin 1., 2. ve 3. veri kümeleri üzerinde yapılan tahminlerinin DVR modeli ile karşılaştırılmış grafikleri, Çizelge 6.3'te ise başarı ölçümleri gösterilmiştir.



Şekil 6.8 : Yinelemeli GKA-DVR modeli ile tekli DVR modeli karşılaştırması.

Çizelge 6.3 : Yinelemeli GKA-DVR (%100) yaklaşımı ile tekli DVR tahminlerinin karşılaştırılmaları.

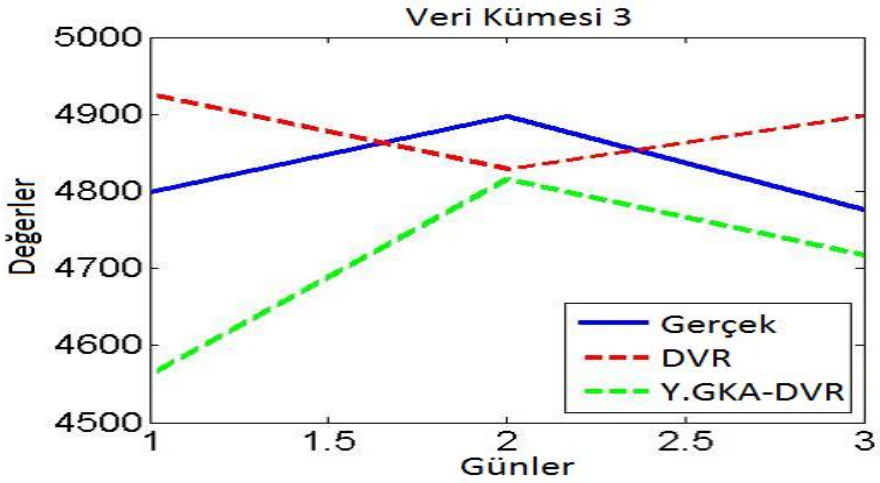
	Metod	OKH	OMHY (%)	YD (%)	DAr (%)	DAz (%)	Par. 1	Par. 2
1	DVR	**245.642**	**1.73**	70	61.53	81.25	147	3
	Y.GKA-DVR (%100)	288.78	2.00	**80.64**	**69.23**	**93.75**		
2	DVR	**0.0025**	**2.62**	75.50	77.22	73.73	50	4
	Y.GKA-DVR (%100)	0.0042	4.35	**76.11**	**78.21**	**74.74**		
3	DVR	**28825.7**	**2.62**	60	56.25	64.28	50	4
	Y.GKA-DVR (%100)	49267.2	3.34	**64.51**	**62.50**	**71.42**		

Elde edilen sonuçlarda, Yinelemeli GKA-DVR modelinin DVR modeline göre başarıları ve bileşenlerin düzensiz değişimleri sahip olmasından kaynaklanan başarısızlıklarının olduğu gözlenmiştir.

Çizelge 6.3'te görüldüğü üzere Yinelemeli GKA-DVR modeli DVR modeline göre OKH ve OMHY hata ölçümlerinde başarısız, ancak yön başarı ölçümlerinde başarılıdır. Bu durumda yeni yöntemin noktasal olarak değerin ne olacağı hakkında bir tahminin olduğu, ancak bu tahminin DVR modeli tahmininden daha iyi olmadığı görülmektedir. Başarılı olduğu kısım ise bir sonraki adımın artacak ya da azalacak olduğunun doğruluğudur. Çizelge 6.3'de görüldüğü üzere Yinelemeli GKA-DVR modeli tüm yön başarımı ölçülerinde (DA, DAr, DAz) daha yüksektir.

Yön başarım doğruluğunu daha iyi gözlemlemek amacı ile Şekil 6.9'da Veri Kümesi 3'ün tahminleme grafiği ayrıntılı görülebilecek şekilde ölçeklenmiştir. Yeni Zelanda'nın elektrik kullanım değerlerini içeren Veri Kümesi 3'te, 2.gün için gerçek değer 4900 MW'dır. DVR'nin tahmini 4820 KW ile Yinelemeli GKA-DVR yöntemine göre daha başarılıdır. Ancak 1. gün ile 2. gün arasında gerçekte yükselen

bir kullanım ve 2. gün ile 3. gün arasında ise gerçekte azalan bir kullanım var iken, DVR tam tersini tahmin etmiş, Yinelemeli GKA-DVR modeli ise doğru bir şekilde tahmin etmiştir.



Şekil 6.9 : Yinelemeli GKA-DVR modeli ile tekli DVR modeli tahminlerinin karşılaştırılması.

Bu yaklaşımda, OKH ve OMYH başarım ölçümlerinin düşük olmasına ragmen YD, DAr ve DAz ölçümlerinin yüksek başarıma sahip olmasının 2 sebebi bulunmaktadır. Birincisi, GKA algoritmasının yinelemeli olarak bileşenlere ayırma işleminde, eklenen son noktanın yönünün (artış ya da azalış), yeni çıkarılan bileşen yönleri ile aynı olmasıdır. İkinci sebebi ise veri kümelerinin periyodiklik (haftalık) göstermesidir. Örneğin Pazar günleri en düşük değerde olan elektrik kullanımı değeri cumartesi günlerine göre azalış göstermekte, azalış aynı zamanda yine her iki günün bileşenleri arasında da görülmektedir. Ayrıca Pazar gününe ait gözlenen bu azalış her Pazar (periyodik olarak) gerçekleşmektedir. Böylece yapılan tahminin noktasal olarak değeri büyük hata ile tahmin edilse bile yön doğruluğu korunmaktadır.

6.3 Ortalama IMF ile GKA-DVR Modeli

Bu bölümde, GKA algoritması ile elde edilen bileşenler üzerinde gürültüyü tanımlayan çalışmalar incelenmiştir. Örneğin bir veri kümesi GKA algoritması ile bileşenlere ayrıldığında, elde edilen bileşenlerin beyaz gürültü (White Gaussion Noise) özelliği taşıyıp taşımadığı bilgisi gürültü süzme tekniğinde yararlı olacaktır.

Bu konuda daha once yapılan çalışmalarda, hangi bileşenin beyaz gürültü taşıdığı araştırılmıştır [42-43] [52-53]. Örneğin Qian, [49] yapmış olduğu çalışmada, veri

kümesi için ayırdığı bileşenlerdeki gürültüyü tek tek analiz etmiş ve 1. bileşenin (1. IMF) en yüksek gürültüyü içerdiğini belirtmiştir. Bir diğer çalışmada ise, Kalıntı bileşeninden bir önceki bileşenin (son IMF) gürültü içerdiği tespit edilmiştir [45]. Bu çalışmalarda gürültü içeren bileşen hariç diğer bileşenlerin toplanması yöntemi ile veri kümesi gürültüden arındırılmış ve tahminleme yeni (gürültüsüz) veri kümesi üzerinden yapılmıştır.

Ancak GKA algoritması ile zaman serisi bileşenleri elde ediminde son problem mevcuttur. Bu durum, Şekil 6.6'da olduğu gibi test bileşenlerinin eğitim bileşenleri ile aynı özelliğe sahip olmaması ve test bileşenlerinin durağana yakın olma özelliğini kaybetmesine yol açmaktadır. Örneğin bir veri kümesinin eğitim kümesinde 1. bileşen gürültü içerirken, test kümesinin 1. bileşeni artık aynı seviyede gürültü içermeyebilir. Bu durum, gürültü süzüldüğünde elde edilen yeni (gürültüsüz) veri kümesinin, tam olarak gürültüden arındırılamamasına yol açmakta ve yapılan tahmin başarımını düşürmektedir.

GKA bileşenlerinden Kalıntı hariç diğerlerinin (IMF'ler) ortalamasının alınması ile yüksek bir gürültü bulunduğu çıkarılmıştır [45]. Bu çalışmanın, spesifik olarak bir bileşen belirtmeyerek bileşenlerin ortalamasını alması ve bu ortalamanın gürültü özelliğini güçlü bir şekilde taşıması, gürültü süzme işlemini kolaylaştırmaktadır. Çalışmada uygulanan gürültü süzme işlemi çok değişkenli (multivariate) bir veri kümesi üzerinde yapılmıştır. Bu tez çalışması kapsamında ise ortalama IMF'ye bağlı gürültü süzme tekniği tek değişkenli (univariate) veri kümelerinde kullanılarak DVR'nin başarımı artırılmaya çalışılmıştır. Zaman serisi verilerde test kümesinin GKA bileşenleri eğitim kümesi bileşenleri ile aynı özelliği taşımadığından, gürültü süzme işlemi yalnızca eğitim kümesinde yapılmıştır.

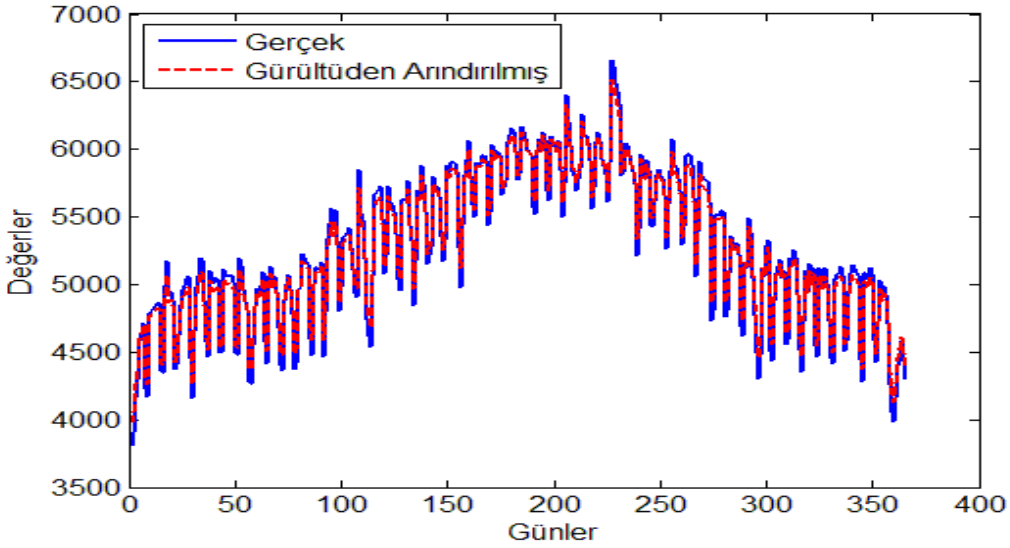
Bu durumda IMF'lerin ortalaması IMF_{ort}, ve IMF sayısı n olmak üzere;

$$IMF_{ort}(t) = \frac{\sum_{i=1}^{n} IMF_i(t)}{n} \quad \textbf{(6.1)}$$

Veri kümesi gürültüden arındırıldığında ortaya çıkan veri kümesi ise:

$$x_y(t) = x(t) - IMF_{ort}(t) \quad \textbf{(6.2)}$$

olarak tanımlanır. Şekil 6.10'da ve Çizelge 6.3'te Veri Kümesi 3'e ait eğitim kümesi ($x(t)$) ile gürültüden arındırılmış olan yeni veri kümesi ($x_y(t)$) karşılaştırılmaları gösterilmiştir.



Şekil 6.10 : Gürüntüden arındırılmış değerler ile gerçek değerlerin karşılaştırılması.

Çizelge 6.4 : Gürüntüden arındırılmış değerler ile gerçek değerlerin karşılaştırılması.

Veri Kümesi	YD (%)	DAr (%)	DAz (%)
$x_y(t)$	100	100	100

Her iki veri kümesi karşılaştırıldığında, gürültüden arındırılmış veri kümesinin asıl değerler ile aynı yönlerde olduğu görülmektedir. Gürültüden arındırma işlemi tamamlandıktan sonra, ortaya çıkan yeni veri kümesi, asıl veri kümesi ile birlikte kullanılarak tahminleme işlemi yapılmıştır.

Bu yaklaşımla, $d_1, \dots, d_7$ takvim bilgisini, od olağan dışı bilgiyi, $x(t), \dots, x(t-6)$ orjinal veri kümesinin geçmiş bilgilerini, $x_y(t), x_y(t-1)$ gürültüden arındırılmış veri kümesinin geçmiş bilgilerini içeren veriler olmak üzere, öznitelik kullanımı;

$$y(t) = f_t(d_1, \dots, d_7, od, x(t), x(t-1), \dots, x(t-6), x_y(t), x_y(t-1)) \quad \textbf{(6.1)}$$

şeklinde gerçekleşmiştir. $d_1, \dots, d_7$ gün bilgisini içeren öznitelikler yarı- periyodiklik içeren ilk 3 veri kümesinde, Olağan dışı (od) bilgisi yalnızca Veri Kümesi 1'de kullanılmıştır. Gürültüden arındırılmış veri kümesinin sadece 2 günlük geçmiş değerlerinin kullanılma sebebi, eğitim kümesi içinde en iyi öğrenmeyi yaparak en başarımlı tahminlemede bulunmasıdır. Ancak bazı veri kümelerinde gürültüden

arındırılmış veri kümesine ait geçmiş değerin yalnızca 1 değeri ile daha yüksek başarımlı tahminler elde edildiği görüldüğünden, geçmiş kaç günün kullanılacağı da bir parametre (P) olarak alınmıştır. P parametresi ise eğitim kümesi üzerinde en iyi başarımı veren geçmiş gün sayısı ile tespit edilmiştir.

Çizelge 6.4 ile Şekil 6.7 ve Şekil 6.8'de bu tez çalışmasındaki tüm veri kümeleri için bu yaklaşımla (gürültüden arındırılmış) yapılan tahminler ile DVR tahminleri başarılarının karşılaştırılmaları görülmektedir.

Tahminler karşılaştırıldığında gürültüden arındırılmış modelin 7 veri kümesi üzerinde de DVR modeline göre daha başarılı sonuçlar verdiği görülmektedir. Gürültüden arındırılmış veri kümesi ile yapılan tahminlerde noktasal olarak başarımların yanında YD alanında da başarı daha yüksektir. Örneğin TL/USD paritesi tahmini YD değeri DVR modelinde %48.38 iken, yeni modelde %61.29 olarak gözlenmiştir. Böylece elde edilen yaklaşımın, DVR modelinden daha yüksek başarımlı tahminlemelerde bulunduğu 7 adet veri kümesi üzerinde de gösterilmiştir.

Çizelge 6.5 : Tekli DVR modeli ve ortalama IMF ile gürültüden arındırılmış DVR modeli karşılaştırmaları.

Veri Kümesi	Metod	OKH	OMHY (%)	YD (%)	DAr (%)	DAz (%)	P
1	DVR	245.642	1.73	70	61.53	81.25	
1	GKA-DVR (IMFort)	**227.35**	**1.70**	**73.33**	**69.23**	**81.25**	1
2	DVR	0.00258	2.48	75.50	77.22	73.73	
2	GKA-DVR (IMFort)	**0.00244**	**2.56**	75	**75.24**	**74.74**	2
3	DVR	28825.7	2.62	60	56.25	64.28	
3	GKA-DVR (IMFort)	**23708.3**	**2.48**	**63.33**	**68.75**	**57.14**	1
4	DVR	0.00033	0.68	48.38	56.25	40	
4	GKA-DVR (IMFort)	**0.000324**	**0.67**	**61.29**	**68.75**	**53.33**	2
5	DVR	5.98e-05	0.32	43.62	42.02	45.56	
5	GKA-DVR (IMFort)	**5.86e-05**	**0.30**	**45.63**	**43.47**	**48.10**	2
6	DVR	0.000142	0.65	49.24	50	48.78	
6	GKA-DVR (IMFort)	**0.000141**	0.64	49.24	50	48.78	2
7	DVR	1.127	8.66	71.68	71.73	75.37	
7	GKA-DVR (IMFort)	**1.121**	8.63	70.25	68.84	75.37	1

Çizelge 6.6 : Tekli DVR modeli ve ortalama IMF ile gürültüden arındırılmış DVR modeli Karışıklık Matrisleri.

Yöntem		Veri Kümesi 1		
			Tahmin	
			P	N
1	Gerçek	P	8	5
		N	3	14
2		P	9	4
		N	4	13

Yöntem		Veri Kümesi 2		
			Tahmin	
			P	N
1	Gerçek	P	78	23
		N	26	73
2		P	76	25
		N	25	74

Yöntem		Veri Kümesi 3		
			Tahmin	
			P	N
1	Gerçek	P	9	7
		N	5	9
2		P	11	5
		N	6	8

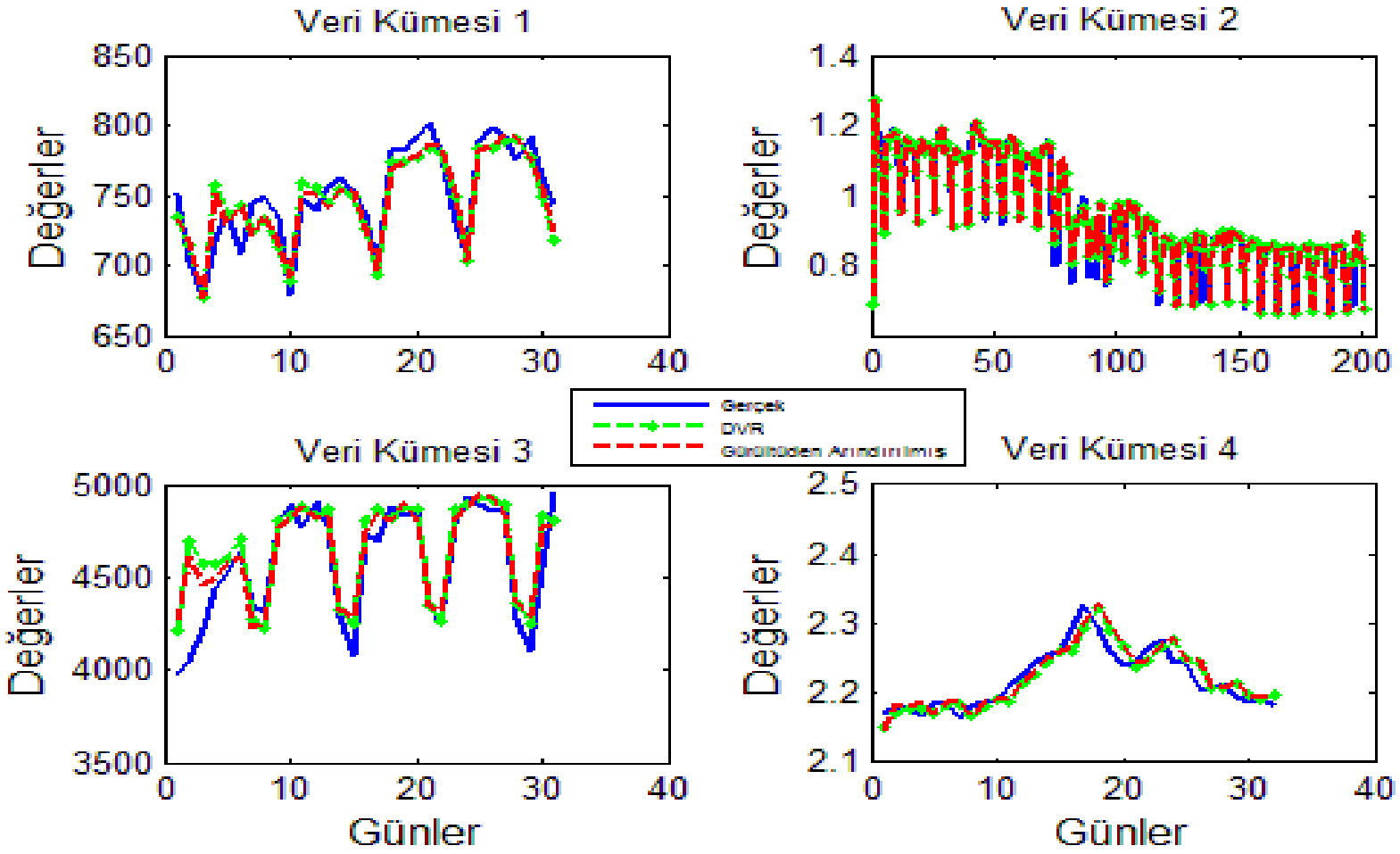
Yöntem		Veri Kümesi 4		
			Tahmin	
			P	N
1	Gerçek	P	9	7
		N	9	6
2		P	11	5
		N	7	8

Yöntem		Veri Kümesi 5		
			Tahmin	
			P	N
1	Gerçek	P	32	37
		N	43	37
2		P	30	39
		N	41	39

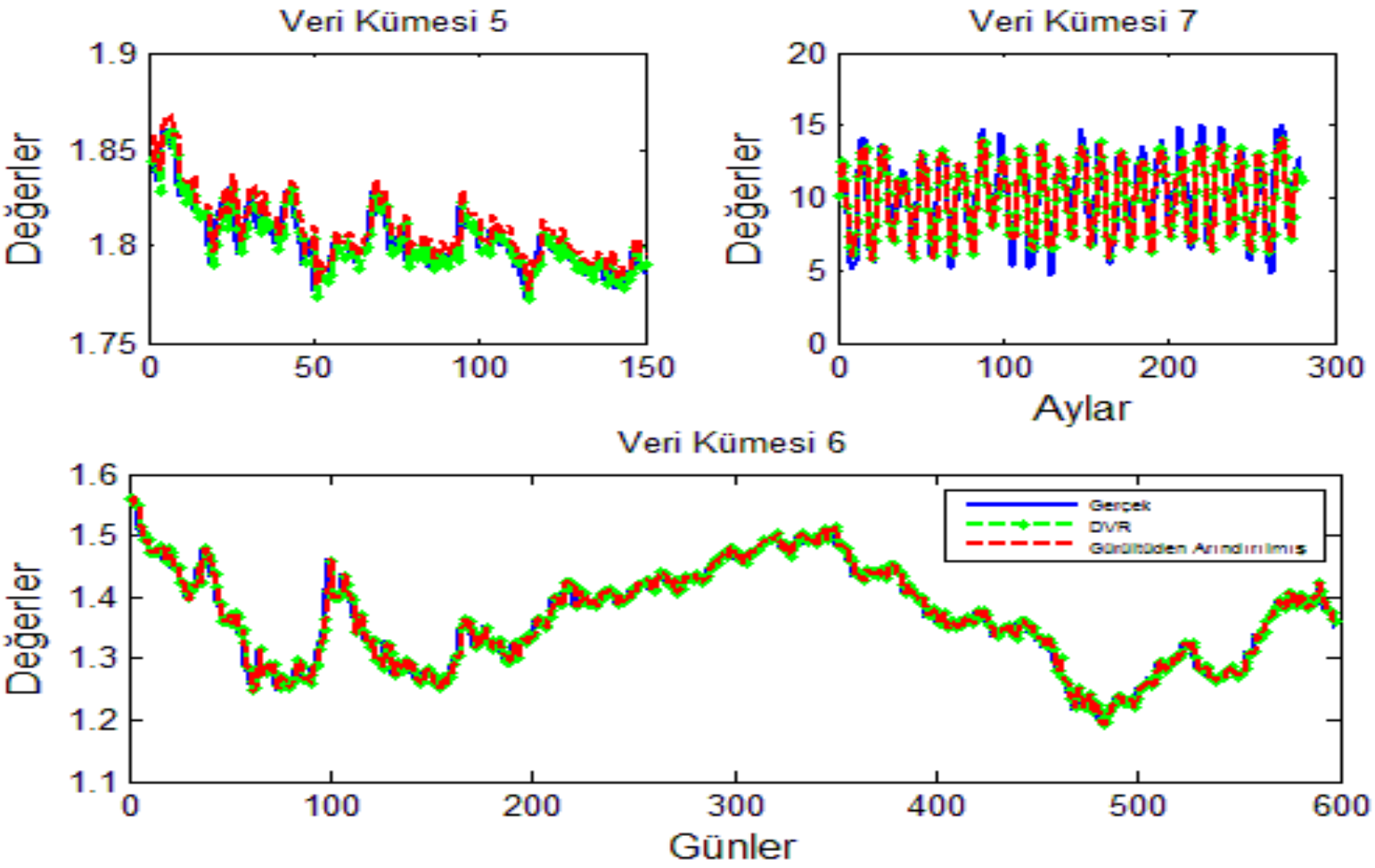
Yöntem		Veri Kümesi 6		
			Tahmin	
			P	N
1	Gerçek	P	153	153
		N	150	141
2		P	153	153
		N	150	141

Yöntem		Veri Kümesi 7		
			Tahmin	
			P	N
1	Gerçek	P	99	39
		N	37	104
2		P	95	43
		N	38	103

Yöntem 1: DVR Modeli. Yöntem 2: GKA-DVR (IMFort) Modeli.



Şekil 6.11 : Tekli DVR modeli ile ortalama IMF ile gürültüden arındırılmış DVR modeli karşılaştırmaları-1.



Şekil 6.12 : Tekli DVR modeli ile ortalama IMF ile gürültüden arındırılmış DVR modeli karşılaştırmaları-2.

7. SONUÇ

Zaman serisi öngörüsü geçmişe ait verilerin kullanımı ile geleceğe ait değerlerin bulunmasını amaçlamaktadır. Literatürde tek değişkenli zaman serisi öngörüsü için Destek Vektör Regresyonu (DVR) algoritmasının yüksek başarıma sahip olduğu gösterilmiştir. Ayrıca DVR algoritması ile kurulan modellerde öznitelik seçimi, parametre kullanımı, çekirdek tercihi, algoritmaların bütünleşik kullanılmaları gibi durumlar öngörü başarımını etkilemektedir.

Bu tez çalışmasında öncelikli olarak DVR algoritmasının zaman serilerinde modellenmesi incelenmiş olup, çekirdek kullanımı, parametre optimizasyonu, takvim bilgisi, geçmiş veriler, olağan dışı durumları içeren öznitelik kullanımı gibi etkenlerin hangi veri kümelerinde etkili olduğu ve yüksek başarım için ne şekilde kullanılabileceği belirlenmiştir. Çalışma kapsamında başarımlar farklı alanlarda, farklı büyüklükte test ve eğitim kümelerine sahip toplam 7 veri kümesi üzerinde test edilmiştir. Kurulan modellerde, veri kümelerinin eğitim kümelerinde öğrenme işlemi yaptırıldıktan sonra test kümeleri üzerinde tahminlemeler yapılmıştır. Veri kümelerinin farklı alanlardan alınma sebebi, bir veri kümesinde yüksek başarımlı olan bir modelin başka bir veri kümesinde her zaman başarılı olamayabileceğidir. Test kümelerinin farklı sayıda alınmasının sebebi ise, özellikle yön doğruluğu ölçümünün gerçek başarısını edinebilmektir. Test sonuçlarında takvim bilgisi ve olağan dışı özniteliği kullanımının periyokliğe sahip veri kümelerinde, DVR'nin YTF çekirdeği kullanılarak ceza parametresi (C), YTF çekirdeği parametresi (γ) ve hata parametresi (ε) birlikte optimize edildiğinde ise bütün veri kümelerinde başarımı artırdığı tespit edilmiştir.

Son zamanlarda bütünleşik modeller kullanılarak tekli algoritmalardan daha iyi başarı sağlayan modeler geliştirilmektedir. Bu kapsamda, DVR algoritmasından daha iyi sonuçlar veren Görgül Kip Ayrışımı tabanlı DVR modelleri üzerinde çalışılmıştır. Halihazırda mevcut GKA-DVR modeli üzerinde yapılan çalışmaların çoğu, GKA bileşenlerini ayrıştırırken test kümesini de kullandığından gerçek uygulamalarda kullanılamayacak bir modelleme tekniği özelliği taşımakta ve son nokta problemi ile

karşılaşmamaktadır. Daha önce yapılmış test kümesini ayrı değerlendiren bir çalışmada ise son nokta problemi için farklı yöntemler ile çözümlemeler önerilmektedir. Fakat son nokta problemi için kesin bir çözüm bulunmadığından ve önerilen yöntemler sadece tek bir başarım ölçümü üzerinden değerlendiğinden, DVR yöntemine göre her başarım ölçümünde yüksek bir başarı gösterememektedir. GKA tabanlı bütünleşik yöntemlerin başarımı kullanılan son nokta yönteminin başarımına bağlıdır.

Bu tez çalışmasında, GKA tabanlı DVR algoritması üzerinde farklı iki model önerilmiştir. İlk modelde GKA algoritmasına ait İçkin Mod Fonksiyonları üzerinde tahminlemeler yapılmış ve üzerinde periyodiklik görülen veri kümeleri için yön doğruluğu ölçümlerinde DVR'ye gore daha başarılı sonuçlar elde edilmiştir.

İkinci modelde ise GKA kullanılarak asıl (orjinal) zaman serileri üzerinden gürültü ayrıştırma gerçekleştirilmiş ve elde edilen yeni veri kümesi DVR eğitiminde başarımı artırmak için kullanılmıştır. Bu modelde ise kullanılan tüm başarım ölçütlerinde ve tüm veri kümelerinde DVR'den daha yüksek sonuçlar alınmıştır.

İleriki çalışmalarda son nokta problemine çözüm getiren algoritmaların farklı başarım ölçütleri ile değerlendirilmesi yapılacaktır. Ayrıca son nokta problemi için önerilmiş farklı yöntemler ile gürültü süzülerek öznitelik zenginleştirilmesinin başarıma etkileri incelenecektir.

KAYNAKLAR

[1] **Newbold, Paul**. İşletme ve İktisat için İstatistik, çev. Ümit Şenesen, Literatür Yayınclık., 2000. s. 777-785.

[2] **NA, Sung Hoon ve So Young SOHN** (2011), "Forecasting Changes in Korea composite Price Index (KOSPI) Using Association Rules", Expert Systems with Applications, 38:9046-9049.

[3] **Y.S.Lee, L.I.Tong**. Forecasting timeseries using a methodology based on autoregressive integrated moving average and genetic programming, Knowledge- based Systems 24 (2011) 66–72.

[4] **Erdogdu E.** Electricity demand analysis using cointegration and ARIMA modelling: a case study of Turkey. Energy Policy 2007;35(2):1129–46.

[5] **C.W.Cheong.** Modeling and forecasting crude oil markets using ARCH-type models, Energy Policy 37 (2009) 2346–2355.

[6] **C.M. Lee, C. N. Ko.** Time series prediction using RBF neural networks with a nonlinear time-varying evolution PSO algorithm, Neurocomputing 73 (2009) 449–460.

[7] **Kucukali S, Baris K.** Turkey's short-term gross annual electricity demand forecast by fuzzy logic approach. Energy Policy 2010;38(5):2438–45.

[8] **Box G.E.P, Jenkins G.M, Reinsel G.C**. "Time series analysis: forecasting and control", Englewood Clis,NJ:Prentice-Hall, 1994.

[9] **J. V. Hansen, J. B. McDonald, R. D. Nelson.** Time Series Prediction With Genetic-Algorithm Designed Neural Networks: An Empirical Comparison With Modern Statistical Models, Computational Intelligence, v.15, no.3, 1999.

[10] **V. N. Vapnik.** "The Nature of Statistical Learning Theory", Springer-Verlag, New York, 1995.

[11] **Kim, Kyoung-jae** (2003), "Financial Time Series Forecasting Using Support Vector Machines", Neurocomputing, 55:307-319.

[12] **Url-1** <*http://neuron-ai.tuke.sk/competition/*> alındığı tarih: 15.12.2012.

[13] **M-W. Chang, B-J Chen, C-J Lin.** "EUNITE Network Competition:Electricity Load Forecasting", November 2001. Winner of EUNITE world wide competition on electricity load prediction.

[14] **D. Esp.** "Adaptive Logic Networks for East Slovakian Electrical LoadForecasting", EUNITE World-wide competition, November 2001.

[15] **W.Brockmann, Steffen Kuthe.** "Different models to forecast electricityloads", EUNITE World-wide competition, November 2001.

[16] **Mohandes MA, Halawani TO, Rehmam S, Hussain AA.** (2004) Support vector machines for wind speed prediction. Renew. Energy 29: 939-947.

[17] **Pai PF, Hong WC.** (2005). Support vector machines with simulated annealing algorithms in electricity load forecasting. Energy. Convers. Manage. 46:2626-2669.

[18] **Cao LJ (2003**). Support vector machines experts for time series forecasting. Neurocomputing 51:321-339.

[19] **Cao LJ, Tay FEH.** (2001). Financial forecasting using support vector machines. Neural Comput. Appl. 10:184-192.

[20] **N. E. Huang, Z. Shen, S. R. LONG. ve diğ.** "The Empirical Mode Decomposition and the Hilbert Spectrum for Nonlinear and Non–stationary Time Series Analysis", Proceedings of the Royal Society, London, 1998, pp. 903-995.

[21] **X. Xu, Y. Qi, Z. Hua.** Forecasting demand of commodities after natural disasters, Expert Systems with Applications 37 (2010) 4313–4317.

[22] **C.F.Chen, M. C. Lai, C. C. Yeh.** Forecasting tourism demand based on empirical mode decomposition and neuralnetwork, Knowledge-based Systems 26 (2012) 281–287.

[23] **L. A. Yu, S. Y. Wang, K. K. Lai.** Forecasting crude oil price with an EMD-based neural network ensemble learning paradigm, Energy Economics 30 (2008) 2623–2635.

[24] **Z. Guo, W. Zhao, H. Lu, J. Wang.** Multi-step forecasting for wind speed using a modified EMD-based artificial neural network model, Renewable Energy 37 (2012)241–249.

[25] **L. Tang, L. Yu, S. Wang, J. Li, S. Wang.** A novel hybrid ensemble learning paradigm for nuclear energy consumption forecasting, 93 (2012) 432–443.

[26] **Chang-Jun Yu, Yuan-Yuan He, and Tai-Fan Quan.** "Frequency Spectrum Prediction Method Based on EMD and SVR," 2008 Eighth International Conference on Intelligent Systems Design and Applications, pp. 39–44, Nov.2008.

[27] **Jun Fan and Yanzhen Tang.** "An EMD-SVR method for non-stationary time series prediction," 2013 International Conference on Quality, Reliability, Risk, Maintenance, and Safety Engineering (QR2MSE), pp. 1765–1770, July 2013.

[28] **Wang, L.** 2005, Support Vector Machines: Theory and Applications, Springer, New York, 1434-9922.

[29] **Huang, Te-Ming, Kecman, Vojislav, Kopriva, Ivica.** Kernel Based Algorithms for Mining Huge Data Sets: Supervised, Semi-supervised, and Unsupervised Learning, Springer, 2006.

[30] **Erastö, P.** (2001). Support Vector Machines-Backgrounds and Practice, Academic Dissertation for The Degree of Licentiate of Philosophy. Rolf Nevanlinna Institute, Helsinki.

[31] **Cortes, Corinna, Vapnik Vladimir.** "Support-vector Networks", Machine Learning,C:20, No:3,1995, s.273–297.

[32] **Kavzoğlu, T. ve Çölkesen.** İ. (2010). Destek Vektör Makineleri ile Uydu Görüntülerinin Sınıflandırılmasında Kernel Fonksiyonlarının Etkilerinin İncelenmesi. Harita Dergisi, 144, 73-82.

[33] **Osuna, E.E., Freund, R., Girosi, F.** 1997, Support Vector Machines: Training and Applications, A.I. Memo No. 1602, C.B.C.L. Paper No. 144, Massachusetts Institute of Technology and Artificial Intelligence Laboratory, Massachusetts.

[34] **Abe, Shigeo.** Support Vector Machines for Pattern Classification, London, Springer, 2010.

[35] **Hamel, Lutz H.** Knowledge Discovery with Support Vector Machines, New Jersey, Wiley-Interscience, 2009.

[36] **Burges, C. J. C.** (1998). A Tutorial on Support Vector Machines for Pattern Recognition, Data Mining and Knowledge Discovery, Kluwer Academic Publishers, Boston, USA, 2, 121-167.

[37] **Vapnik, Vladimir N.** Statistical Learning Theory, New York, Wiley-Interscience, 1998.

[38] **Aygoğan Ü.** (2010), "Destek vektör makinalarında kullanılan çekirdek fonksiyonların sınıflama performanslarının karşılaştırılması", Hacettepe Üniversitesi, (Yüksek Lisans Tezi) Ankara.

[39] **Huang N E, Shen Z, Long S R.** "A New View of Nonlinear Water Waves: The Hilbert Spectrum," Annual Review of Fluid Mechanics, vol. 31, pp. 417-457, 1999.

[40] **Baykut S.**, (2012). "Doğal süreçlerde gürültü analizi ve sinyal modellemeleri," İstanbul Teknik Üniversitesi, (Doktora Tezi), İstanbul.

[41] **Baykut S., Akgül T., Ergintav S.** "GPS Konum Verilerinin GKA tabanlı Analizi ve Gürültüden Arındırılması," Signal Processing and Communications Applications Conference, 2009. SIU 2009. IEEE 17th, 2009, pp. 644–647.

[42] **Premanode, B., Vonprasert, J., & Toumazou**, C. (2013). Prediction of exchange rates using averaging intrinsic mode function and multiclass support vector regression. In Artificial Intelligence Research (Vol. 2, pp. 47–61).

[43] **B. Premanode, C. Toumazou.** "Improving prediction of exchange rates using Differential EMD," in Expert Systems with Applications, 2013, vol. 40, no. 1, pp. 377–384.

[44] **Url-2** <*http://research.ics.aalto.fi/eiml/datasets.shtml*> alındığı tarih:5.01.2014

[45] **Url-3** <*http://www.ea.govt.nz/about-us/what-we-do/our-history/archive/ operations -archive/monitoring*> alındığı tarih: 28.01.2014

[46] **Url-3** <*http://fx.sauder.ubc.ca*> alındığı tarih: 15.02.2014

[47] **Url-4** < *http://www.isds.duke.edu/~mw/ts_data_sets.html*> alındığı tarih: 15.02.2014

[48] **C.-C. Chang and C.-J.** Lin. "{LIBSVM}: A library for support vector machines," ACM Trans. Intell. Syst. Technol., vol. 2, no. 3, pp. 27:1–27:27, 2011.

[49**] *T*. Xiong, Y. Bao, and Z. Hu**. "Does restraining end effect matter in EMD-based modeling framework for time series prediction? Some experimental evidences," Neurocomputing, vol. 123, pp. 174–184, Jan. 2014.

[50] **R. Rato, M. Ortigueira, A. Batista.** "On the HHT ,it's problems, and some solutions", Mechanical Systems and Signal Processing 22 (2008) 1374–1394.

[51] **Hyndman, R. J. & Koehler, A. B.** "Another look at measures of forecast accuracy", International Journal of Forecasting. (2006).

[52] **Y. Qian, J. Xia, K. Fu, and R. Zhang.** "Network traffic forecasting by support vector machines based on empirical mode decomposition denoising," in 2012 2nd International Conference on Consumer Electronics, Communications and Networks (CECNet), 2012, pp. 3327–3330.

[53**] J. Cheng, X. Sun, D. Liu, J. Bian, and D. H. S. Zou.** "Reducing measurement noise in rock bolts detecting," in 2009 9th International Conference on Electronic Measurement & Instruments, 2009, vol. 1, pp. 4–527–4–531

ÖZGEÇMİŞ

Ad Soyad:	Bahadır BİCAN
Doğum Yeri ve Tarihi:	İzmit, 1985
Adres:	Yeni mah. Üçevler Sit. Örnekevler Cad. No: 12 GÖLCÜK/KOCAELİ
E-Posta:	bicanb@itu.edu.tr
Lisans:	Deniz Harp Okulu
Yüksek Lisans:	İstanbul Teknik Üniversitesi
Yayın ve Patent Listesi:	A Hybrid Method For Time Series Prediction Using EMD and SVR

TEZDEN TÜRETİLEN YAYINLAR/SUNUMLAR

Bican B., Yaslan Y., 2014 : "A Hybrid Method For Time Series Prediction Using EMD and SVR." The 6th International Symposium on Communications, Control, and Signal Processing (ISCCSP 2014), May 21-23, 2014, Athens, Greece.