



MARMARA ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ



AĞ ANALİZİ VE ZAMAN SERİSİ YÖNTEMLERİ İLE PATENT VERİSİ ANALİZİ

HATİCE IŞIK ÖZATA

YÜKSEK LİSANS TEZİ

Bilgisayar Mühendisliği Anabilim Dalı

Bilgisayar Mühendisliği Yüksek Lisans Programı

DANIŞMAN

Dr. Öğr. Üyesi Önder DEMİR

EŞ-DANIŞMAN

Dr. Öğr. Üyesi Buket DOĞAN

İSTANBUL, 2020



MARMARA ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ



AĞ ANALİZİ VE ZAMAN SERİSİ YÖNTEMLERİ İLE PATENT VERİSİ ANALİZİ

HATİCE IŞIK ÖZATA

523617002

YÜKSEK LİSANS TEZİ

Bilgisayar Mühendisliği Anabilim Dalı

Bilgisayar Mühendisliği Yüksek Lisans Programı

DANIŞMAN

Dr. Öğr. Üyesi Önder DEMİR

EŞ-DANIŞMAN

Dr. Öğr. Üyesi Buket DOĞAN

İSTANBUL, 2020

TEŞEKKÜR

Tez çalışmam boyunca benden bilgisini, desteğini ve sabrını esirgemeyen hocalarım Dr. Öğr. Üyesi Önder Demir'e ve Dr. Öğr. Üyesi Buket Doğan'a, hayatımın her evresinde bana destek olan aileme, uzmanlıklarıyla beni yönlendiren Marmara Üniversitesi İnovasyon ve Teknoloji Transfer Uygulama ve Araştırma Merkezi (MİTTO)'ya ve Marmara Üniversitesi BAP Birimi'ne (Proje No: FEN-C-YLP-130319-0063) verdiği desteklerden dolayı teşekkürü bir borç bilirim.

Kasım, 2020

Hatice IŞIK ÖZATA

İçindekiler

1. GİRİŞ	1
1.1 Patent Nedir?	2
1.2. Patent Analizi Nedir?	3
1.3. Amaç ve Önem	3
1.4. Literatür Araştırması	4
2. MATERYAL VE YÖNTEM	14
2.1 Patent Veri Tabanlarında Arama Yapmak	15
2.1.1 Patent Uzmanlarından Tavsiye Almak	16
2.1.2 Bilgi Güvenliği Alanında Lider Firmaları Belirlemek	16
2.1.3 Bilgi Güvenliği Alanında En Popüler IPC Sınıflarını Belirlemek	17
2.1.4 Anahtar Kelimeleri Belirlemek için Gerekli Sorgu Cümlelerini Oluşturmak	18
2.1.5 Anahtar Kelimeleri Belirlemek	18
2.1.6 Veri Setini Oluşturmak	20
2.2 Metin Ön İşleme	21
2.2.1 Veri Setinin Yıl Bazında İncelenmesi	22
2.2.2 Veri Setinin Ülke Bazında İncelenmesi.....	23
2.2.3 Veri Setinin Başvuran Firma Bazında İncelenmesi.....	24
2.2.4 Veri Setinin CPC Kodlarına Göre İncelenmesi	25
2.2.5 Veri Setinin IPC Kodlarına Göre İncelenmesi	29
2.3 Kullanılan Yöntemler	31
2.3.1 Ağ Analizi	31
2.3.1.1 Graf Veri Yapısı ve Bileşenleri	32
2.3.1.2 Ağ Ölçütleri	34
2.3.2 Zaman Serisi Analizi	34

2.3.2.1 Zaman Serisi Bileşenleri Nelerdir?	35
2.3.2.2 Zaman Serilerinde Regresyon Modelleri	35
2.3.2.3 Zaman Serilerinde Ayırıştırma Yöntemleri	36
2.3.2.4 Zaman Serilerinde Üstel Düzleştirme Yöntemleri	37
2.3.2.5 Zaman Serilerinde ARIMA Modelleri	38
2.3.2.6 Zaman Serilerinde Yapay Sinir Ağları Modelleri	39
3. BULGULAR VE TARTIŞMA	40
3.1 Ağ Analizi – Patentlerin Görsel Analizi	40
3.1.1 Yıl Bazında Ağ Analizi	40
3.1.2 Son 10 Yılın Verileriyle Ağ Analizi	44
3.2 Zaman Serisi Analizi – Patent Sınıflarına Ait Patent Sayılarının Mevsimsel Tahminlemesi	45
3.2.1 Zaman Serisi Regresyonu ile Patent Sınıflarına Ait Patent Sayılarının Tahminlenmesi	51
3.2.2 Ayırıştırma Yöntemleri ile Patent Sınıflarına Ait Patent Sayılarının Tahminlenmesi	59
3.2.3 Üstel Düzleştirme Yöntemi ile Patent Sınıflarına Ait Patent Sayılarının Tahminlemesi	61
3.2.4 ARIMA Modelleri ile Patent Sınıflarına Ait Patent Sayılarının Tahminlenmesi	66
3.2.5 Yapay Sinir Ağları Modeli ile Patent Sınıflarına Ait Patent Sayılarının Tahminlenmesi	76
4. SONUÇLAR	79
4.1 ÖNERİLER	83
KAYNAKLAR	85

ÖZET

AĞ ANALİZİ VE ZAMAN SERİSİ YÖNTEMLERİ İLE PATENT VERİSİ ANALİZİ

Hem özel hem kamu kurumları stratejik yatırımlar yapabilmek ve çalışmalarını hangi yönde ilerletmeleri gerektiğini öğrenebilmek için gelişmekte olan teknolojilere ilişkin ön bilgilere büyük ilgi duyarlar. Gelişmeleri izlemek ve yönelimleri takip etmek için patent dokümanlarından teknoloji tahminleri yapmak son zamanlarda ortaya çıkan ve parlayan bir araştırma alanıdır. Bugüne kadar farklı alanlarda teknoloji tahmini yapıp sonuçlarını paylaşan birçok çalışma yapılmıştır. Ancak günümüzün en önemli konularından olan bilgi güvenliği için teknoloji tahmini alanında yapılan çalışmalar kısıtlıdır.

Son zamanlarda bilgi güvenliği hem yazılım, hem donanım hem de insanı ilgilendiren karmaşık bir yapı haline gelmiştir. Bunun sonucunda, bilgi güvenliği alanında ortaya çıkan yeni teknolojiler hızla artmaktadır.

Bu çalışmada da son on yılda alınan patentlerden bilgi güvenliğini ilgilendirenler veri seti olarak hazırlanmış ve çalışmanın devamında veri setindeki patentler ağ analizi ve zaman serisi analizi yöntemleriyle incelenmiştir.

Ağ analiziyle yıl bazında analizler yapıp hangi yıl hangi ülkenin hangi patent sınıfıyla bu alana yöneldiği izlenmiştir. Aynı zamanda ağ analizi tüm veri seti için de yapılmış olup, son on yılın bilgi güvenliği alanında en çok patent alan ülkesi ve en çok patente sahip sınıfı belirlenmiştir. Patentler sınıflandırma sistemi sayesinde teknik alanlara göre ayrılmaktadır.

Zaman serisi analizinde frekans değerine farklı bir bakış açısıyla yaklaşp, veri setinde en çok kullanılan dokuz patent sınıfı frekans olarak kullanılmış ve bu şekilde yapılan zaman serisi analizi sonuçları paylaşılmıştır. Zaman serisi analizi için farklı yöntemler denenmiş, 2019 yılı için en yakın tahmini yapan metodun 2020 yılı için yaptığı tahminleri dikkate alınmıştır. Değerlendirmemiz önerilen zaman serisi analizi yönteminin 2019 yılı ilk 6 ayı için bilgi güvenliği alanında alınan patentlerin patent sınıflarına ait patent sayılarını %98 doğrulukla hesaplayabileceğini ortaya koymaktadır.

Kasım, 2020

Hatice IŞIK ÖZATA

ABSTRACT

PATENT DATA ANALYSIS WITH NETWORK ANALYSIS AND TIME SERIES METHODS

Both private and public institutions are very interested in prior knowledge of emerging technologies in order to make strategic investments and learn how to advance their work. Technology forecasting from patent documents to follow developments and follow trends is a recent and shining field of research. Until today, many studies have been conducted in different fields that make technology forecasts and share their results. However, studies in the field of technology forecasting for information security, which is one of the most important issues of today, are limited.

Recently, information security has become a complex structure that concerns both software, hardware and people. As a result, new technologies emerging in the field of information security are increasing rapidly.

In this study, among the patents obtained in the last decade, those concerning information security were prepared as a data set, and the patents in the data set were examined by network analysis and time series analysis methods.

With the network analysis, year-based analyzes were made and it was monitored which year which country and which patent class directed to this field. At the same time, the network analysis was performed for the entire data set, and the country with the most patents and the class with the most patents in the field of information security of the last decade was determined. Patents are divided according to technical areas thanks to the classification system.

In the time series analysis, approaching the frequency value from a different perspective, the most used nine patent class frequencies in the data set were used and the results of the time series analysis made in this way were shared. Different methods were tried for time series analysis, and the estimates made for 2020 by the method that made the closest estimate for 2019 were taken into account. Our evaluation reveals that the proposed time series analysis method can calculate the number of patents belonging to patent classes of

patents received in the field of information security for the first half of 2019 with 98% accuracy.

November, 2020

Hatice IŞIK ÖZATA



KISALTMALAR

IPC	:Uluslararası Patent Sınıflandırması (International Patent Classification)
CPC	:Ortak Patent Sınıflandırması (Cooperative Patent Classification)
USPTO	:United States Patent and Trademark Office (Amerika Birleşik Devletleri Patent ve Ticari Marka Ofisi)
INPADOC	:Uluslararası Patent Belgeleri (International Patent Documentation)
EPO	:European Patent Office (Avrupa Patent Ofisi)
OPS	:Açık Patent Hizmetleri (Open Patent Services)
ARIMA	:Otoregresif Hareketli Ortalamalar (Autoregressive Integrated Moving Average)
STL	:Loess kullanarak Mevsimsel ve Trend ayrıştırma (Seasonal and Trend decomposition using Loess)
SEATS	:ARIMA Zaman Serilerinde Mevsimsel Ekstraksiyon (Seasonal Extraction in ARIMA Time Series)
ACF	:Otokorelasyon Fonksiyonu (Autocorrelation Function)
PACF	:Kısmi Otokorelasyon Fonksiyonu (Partial Autocorrelation Function)
ME	:Ortalama Hata (Mean Error)
RMSE	:Kök Ortalama Kare Hatası (Root Mean Square Error)
MAE	:Ortalama Mutlak Hata (Mean Absolute Error)
MPE	:Ortalama Yüzde Hata (Mean Percentage Error)
MAPE	:Ortalama Mutlak Yüzde Hata (Mean Absolute Percentage Error)
MASE	:Ortalama Mutlak Ölçekli Hata (Mean Absolute Scaled Error)
AIC	:Akaike Bilgi Kriteri (Akaike Information Criterion)
Yy	:Yüzyıl
ArGE	:Araştırma ve Geliştirme

ŞEKİL LİSTESİ

SAYFA

Şekil 1-1 Patent Belgesi Örneği	2
Şekil 2-1 Gerçekleştirilen Çalışmanın İşlem Adımları	15
Şekil 2-2 Kelime Grupları Çaprazlaması	19
Şekil 3-1 Yıl Bazında Ağ Analizi.....	43
Şekil 3-2 Veri Seti Ağ Analizi	44
Şekil 3-3 Patent Sınıflarına ait Patent Sayıları Zaman Serisi Grafiği	47
Şekil 3-4 Patent Sınıfları ve Sınıflara Ait Patent Sayıları Zaman Serisine Uygulanan "ggseasonplot" Fonksiyonu Grafiği.....	48
Şekil 3-5 Patent Sınıflarına Ait Patent Sayıları Zaman Serisine Uygulanan "ggsubseriesplot" Fonksiyonu Grafiği.....	49
Şekil 3-6 Patent Sınıflarına ait Patent Sayıları Zaman Serisinin Otokorelasyon Grafiği50	
Şekil 3-7 Doğrusal Regresyon Modellemesinin Özeti	52
Şekil 3-8 Doğrusal Regresyon Modelinin Grafikleri	53
Şekil 3-9 “Breusch-Godfrey” Testi Sonuçları.....	54
Şekil 3-10 Doğrusal Regresyon Modelinin Patent Sınıflarına ait Patent Sayıları Zaman Serisi Üzerindeki Performans Grafiği	54
Şekil 3-11 Doğrusal Olmayan Kübik Regresyon Modeli Grafikleri.....	55
Şekil 3-12 Doğrusal Olmayan Kübik Regresyon Modelinin Patent Sınıflarına ait Patent Sayıları Zaman Serisi Üzerindeki Performansı	56
Şekil 3-13 R Programlamayla Regresyon Modellerinin Karşılaştırılması.....	58
Şekil 3-14 Regresyon Modellerinin Patent Sınıflarına Ait Patent Sayıları Zaman Serisi Üzerindeki Tahmin Performansları Grafiği.....	58
Şekil 3-15 R Programlamada STL Ayırıştırma Yöntemi ile Patent Sayısı Zaman Serisinin Ayırıştırılması ve Tahminlerin Alınması.....	60
Şekil 3-16 STL Ayırıştırma Yöntemi ile Patent Sınıflarına Ait Patent Sayıları Zaman Serisi Üzerinde Tahminleme Grafiği.....	61
Şekil 3-17 Patent Sınıflarına ait Patent Sayıları Zaman Serisine Uygulanan Üstel Düzleştirme Modelleri ve Hata Metrikleri	62
Şekil 3-18 Patent Sınıflarına Ait Patent Sayıları Zaman Serisi Üzerinde Üstel Düzleştirme Modellerinin Oluşturulması ve Hata Metrikleri	63
Şekil 3-19 Patent Sayıları Zaman Serisine Uygulanan Üstel Düzleştirme Model Tahminleri	64
Şekil 3-20 Holt-Winters Mevsimsel Yöntemi Model İncelemesi ve Tahminleri	65

Şekil 3-21 Patent Sınıflarına Ait Patent Sayısı Zaman Serisinin R Programlama ile Durağanlaştırılması.....	66
Şekil 3-22 Durağanlaştırılmış Patent Sınıflarına Ait Patent Sayısı Zaman Serisi Grafiği	67
Şekil 3-23 Patent Sınıflarına ait Patent Sayısı Zaman Serisi Fark-Otokorelasyon Fonksiyonu Grafiği.....	68
Şekil 3-24 Durağanlaştırılmış Patent Sınıflarına Ait Patent Sayıları Zaman Serisi, ACF ve PACF Grafikleri.....	69
Şekil 3-25 Patent Sınıflarına Ait Patent Sayıları Zaman Serisine Farklı ARIMA Modellerinin Uygulanması	70
Şekil 3-26 ARIMA Modellerinin Test Verileri Üzerindeki Performansı.....	71
Şekil 3-27 ARIMA(4,1,1)(1,0,1)[9] Modelinin Test Verisi Üzerindeki Performans Grafiği.....	72
Şekil 3-28 Tüm Veri Üzerinde Oluşturulan ARIMA(2,1,1)(0,0,0)[9] Modelinin Üç Senelik Tahmin Grafiği	73
Şekil 3-29 ARIMA Modelinin İncelemesi R Kodu	73
Şekil 3-30 Auto.arima Fonksiyonunun Patent Sınıflarına Ait Patent Sayıları Zaman Serisine Uygulanması ve İncelenmesi	74
Şekil 3-31 Auto.arima Modelinin Patent Sınıflarına Ait Patent Sayıları Zaman Serisi Üzerindeki Test Tahminleri Grafiği	75
Şekil 3-32 Auto.arima Modelinin Patent Sınıflarına Ait Patent Sayıları Zaman Serisi Üzerindeki Gelecek Üç Senelik Tahmin Grafiği.....	76
Şekil 3-33 Patent Sayıları Zaman Serisi Üzerinde Yapay Sinir Ağları Modelleme ve Simülasyonları.....	77
Şekil 3-34 Yapay Sinir Ağları Modellerinin Patent Sınıflarına Ait Patent Sayıları Zaman Serisi Üzerindeki Simülasyonu	77
Şekil 3-35 NNAR(9,1,5)[9] Yapay Sinir Ağları Modelinin Patent Sınıflarına Ait Patent Sayıları Zaman Serisi Üzerindeki Tahminleri	78
Şekil 3-36 Yapay Sinir Ağları Modeli Hata Metrikleri.....	78
Şekil 4-1 Patent Sınıflarına Ait Patent Sayıları Zaman Serisine Uygulanan Modellerin Tahminleri	81

TABLO LİSTESİ

	SAYFA
Tablo 2-1 Seçilen IPC Sınıflarının Tanımları	18
Tablo 2-2 Veri Setinin Yıllara Göre Sıralaması	22
Tablo 2-3 Veri Setinin Ülkelere Göre Sıralaması	23
Tablo 2-4 Veri Setinin Firmalara Göre Sıralaması	25
Tablo 2-5 Veri Setinin CPC Sınıflarına Göre Sıralaması	26
Tablo 2-6 Veri Setinin CPC Alt Sınıflarına Göre Sıralaması	27
Tablo 2-7 Veri Setinin CPC Alt Gruplarına Göre Sıralaması	28
Tablo 2-8 Veri Setinin IPC Sınıflarına Göre Sıralaması	29
Tablo 2-9 Veri Setinin IPC Alt Sınıflarına Göre Sıralaması	29
Tablo 2-10 Veri Setinin IPC Alt Gruplarına Göre Sıralaması	30
Tablo 3-1 Yıl Bazında Ağların İstatistik Verileri	42
Tablo 3-2 Tüm Veriler İstatistik Verileri	45
Tablo 3-3 Patent Sınıfları ve Sayıları Zaman Serisi	46
Tablo 4-1 Patent Sınıflarına Ait Patent Sayıları Zaman Serisine Uygulanan Modellerin Hata Kriterleri	80
Tablo 4-2 Tahminleme Modellerinin İki Senelik Patent Sayıları Öngörülleri	81

1. GİRİŞ

Teknolojinin hızlı bir tempoda gelişmesinden dolayı işletmeler yatırım yapmak için en iyi ve en uygun teknolojiye karar verirken zorlanmaktadırlar. Teknolojik gelişmeleri takip etmek; şirketlerin rakiplerine karşı güçlü durabilmeleri ve yatırım yapılması gereken alanları bilip, çalışmalarını hangi yönde ilerletmelerini görmeleri için gereklidir. Patent dokümanları daha fazla bilgiyi elde etmeyi sağlayan teknolojik bilgileri içerir. Ayrıca patentlerdeki açıklamaların büyük bir bölümü başka bir şekilde yayınlanmadığı için patent veri tabanı üzerinde olası teknolojileri keşfetmek teknoloji fırsatlarını yakalamak için önemli bir yoldur.

Bilgi güvenliği kavramı ilk defa 1970'lerde bilgisayar korsanlarının bedava telefon görüşmesi yapma teşebbüslerinden sonra ortaya çıkmıştır. 1980'ler zararlı yazılımların ilk olarak ortaya çıktığı zamanlardır. 1990'larla birlikte modern bilgi güvenliği endüstrisi ortaya çıkmaya başlamıştır. 21. yy'nin ilk 10 yılında ise kötü niyetli internet faaliyetleri para kazanma amacına dönüşmeye başlamıştır [1].

Günümüzde bilgi güvenliği hem yazılım, hem donanım hem de insanı ilgilendiren karmaşık bir yapı haline gelmiştir. Son 10 yılda bilgi güvenliği alanındaki çalışmalar çok hızlı bir şekilde ilerlemiş ve yeni teknolojiler hızla artmıştır.

Bu çalışmada, işletmelerin bilgi güvenliği alanında hangi teknolojilere yatırım yapmaları konusunda veriye dayalı kararlar vermesini desteklemek ve yeni teknolojileri tahmin etmek için patent dokümanlarının analizi yapılmıştır. Çalışma kapsamında öncelikle INPADOC veri tabanındaki son 10 yılda bilgi güvenliği kapsamında alınan patentlerden veri seti hazırlanmış ve veri seti hem ağ analizi hem de zaman serisi analizi için kullanılmıştır.

Ağ analizinde veri setinin ağ yapısının görselleştirmesi yapılırken zaman serisi analizinde frekans değerine veri setinde en çok kullanılan dokuz IPC sınıfı atanmış ve zaman serisi analizi yöntemine yeni bir bakış açısı getirilmiştir. Bu frekans değeriyle farklı zaman serisi metotları denenmiş ve en iyi tahmin yapan yöntemle geçmiş değerleri yorumlayarak gelecek için teknolojinin yönü tahmin edilmiştir.

Bu çalışma farklı alanlarda teknolojik tahmin yapmak için kullanılabileceği gibi bloglar ve sosyal ağlar gibi diğer veri kaynaklarından yararlanılarak da genişletilebilir.

1.1 Patent Nedir?

Teknolojik gelişmeleri izlemek için birden çok veri kaynağı kullanılabilir. Yayınlar, sosyal medya ve patent dosyaları bu veri kaynaklarından sadece birkaçı olarak söylenebilir. Elde edilen verinin önemine dayanarak patent dosyaları bu yöntemler arasında en kapsamlı veriyi sunan yöntemlerdendir. Bu yüzden teknolojik yönelimleri takip etmek için patent dokümanlarını incelemek, gelecek tahminleri yapabilmek adına önemli bir yoldur. Patent dosyasında bulunan başvuru tarihi, başvuran ünvanı ve uluslararası patent kodu (IPC) gibi bilgilerle yeni nesil teknolojiler kolaylıkla tanımlanabilmektedir ve bir patent belgesi yenilikle ilgili bütün bilgileri içermektedir. Patent belgesinin bir örneği aşağıda görülmektedir.

	
US 20180131716A1	
(19) United States	
(12) Patent Application Publication	(10) Pub. No.: US 2018/0131716 A1
Chantz	(43) Pub. Date: May 10, 2018
(54) BIOLOGY BASED TECHNIQUES WITH COGNITIVE SYSTEM ANALYSIS FOR HANDLING INFORMATION SECURITY AND PRIVACY	(52) U.S. Cl. CPC .. H04L 63/1441 (2013.01); G06F 2221/2101 (2013.01); G06F 21/55 (2013.01)
(71) Applicant: International Business Machines Corporation	(57) ABSTRACT Mechanisms are provided in which an electronic corpus of content specific to a domain of computer system attacks is ingested and cognitively analyzed to identify at least one instance of a reference to a computer attack and at least one corresponding characteristic of the computer attack. The mechanisms generate an attack profile data structure for the computer attack based on the at least one corresponding characteristic of the computer attack identified by the cognitive analysis of the content of the electronic corpus. A local segment analysis and security (LSAS) engine, associated with a segment of a distributed computing system, is configured based on the attack profile data structure. The LSAS engine determines a security response action to implement based on the attack profile data structure and transmits a control message to at least one computing resource to implement the determined security response action in response to detecting the computer attack.
(72) Inventor: Hyman D. Chantz, Scarsdale, NY (US)	
(21) Appl. No.: 15/345,755	
(22) Filed: Nov. 8, 2016	
Publication Classification	
(51) Int. Cl. H04L 29/06 (2006.01) G06F 21/55 (2006.01)	

Şekil 1-1 Patent Belgesi Örneği

Şekil 1-1'deki patent belgesinin ilk sayfasından alınan örnekte görülebileceği gibi her patentin bir patent numarası vardır. Şekildeki patent belgesi 10 Mayıs 2018 tarihinde alınmıştır. 54 numaralı bölümde dokümanın konusu, 71 numaralı bölümde patente başvuran şirketin adı ve 72 numaralı bölümde de icadı yapanın adı yazmaktadır. 51 numaralı bölümden bakarak patent belgesinin hangi IPC sınıfına ait olduğu görülebilir.

Her ülkenin bir patent kurumu vardır ve bu kurumların görevleri patent alımına yardımcı olmak, ülkedeki patent haklarını korumaktır. Kurumlar yurt dışındaki benzer kuruluşlarla işbirliği içerisinde. Türkiye’de bu işlemleri Türk Patent Enstitüsü yürütmektedir. Bazı patent kuruluşları veri tabanlarını ücretsiz olarak kullanıma açmışlardır. ESPACENET [2] , GooglePatents [3] ve USPTO [4] bu kuruluşlara örnek olarak verilebilir. Bu çalışmada veri setini hazırlamak için INPADOC [5] uluslararası patent koleksiyonu kullanılmıştır. Bu veri tabanı Avrupa Patent Ofisi (EPO) [6] tarafından hazırlanıp, haftalık olarak güncellenmektedir ve erişimi ücretsiz olduğu gibi internet servis aracılığıyla da veri toplamaya imkan sağlamaktadır [7].

1.2. Patent Analizi Nedir?

Patent verilerini işlemek; kullanılabilir verileri çıkartmak için ön hazırlık gerektiren uzun ve zorlayıcı bir prosedürdür. Başvuru numarası, tarih, IPC, soyut ve iddia gibi bölümlerini analiz ederek farklı amaçlar için veri seti hazırlanabilir. Piyasada bu işlemleri yapabilmek için hazırlanan araçlar bulunmakta fakat bu araçlar pahalı olduğu için her şirket tarafından satın alınamamaktadır. Aşağıdaki listede patent analizi için kullanılan araçlardan birkaç örnek mevcuttur.

- a) Thomson Reuters
- b) Patent iNSIGHT PRO
- c) Patent INTEGRATION
- d) PatBASE Analytics

Patent analizi, açılan bir davayı sonuçlandırmak, patentleri kıyaslamak ve doğru araştırma ve geliştirme (ArGe) yatırımları yapmak için kullanılmaktadır. Patent analizi sayesinde şirketler önünü rahatça görebilmektedir. Sağlıklı bir ArGe’nin olmazsa olmazları arasında patent analizi şart olarak düşünülmektedir.

1.3. Amaç ve Önem

Yapılacak olan tez çalışmasında toplanan patent verilerinden yola çıkarak bilgi güvenliği alanındaki eğilimlerin tespit edilmesi amaçlanmaktadır. Kullanılacak patent verileri toplanacak olup bilgi güvenliği alanında alınan patentlerden oluşan veri seti çalışmanın en başında hazırlanmıştır.

Hazırlanan veri setiyle ağ analizi ve zaman serisi analizi yöntemleri uygulanmış ve sonuçlara göre en yüksek performanslı yöntem seçilerek patent verileri anlaşılmaya çalışılmıştır.

Günümüzde patentlerden teknoloji tahmininde bulunma yöntemi hem firmaların araştırma ve geliştirme yatırımlarına yön vermek için kullandığı bir yol hem de rakip firmaların hangi alanlara eğilim yaptığını öğrenmek için başvurulmuş bir yöntem haline gelmiştir.

Bu çalışmada öncelikle veri setini oluşturmak için bir çalışma yapılmış ve elde edilen patent belgelerinin bibliyo dokümanlarından teknoloji tahmini alanında analiz yapmak için yararlanılmıştır. Çalışma bilgi güvenliği alanında yapılan ilk çalışma olduğu için önem kazanmaktadır. Çalışmanın sonucu bilgi güvenliği konusunda teknolojinin hangi alanlarda ilerlediğini göstermektedir.

Aynı zamanda çalışmada kullanılacak olan patent analizi yöntemi dünyada son zamanlarda popüler olarak yönelinen bir alan olmasına karşın Türkçe yayın sayısının az olması ve bilgi güvenliği alanında teknoloji tahmininin yapılmamış olması çalışmanın önemini artırmaktadır.

1.4. Literatür Araştırması

Veri madenciliği günümüzde sıklıkla kullanılan, büyük boyutlu veri setlerinden istenilen bilgiye ulaşma yöntemidir. Veri madenciliği bilgisayar biliminin alt dalıdır ve yapay zekadan, istatistiğe, veri biliminden veritabanı çalışmalarına kadar pek çok alanda kullanılmaktadır.

Bu çalışmada veri madenciliği bilgi güvenliği alanında alınan patent dokümanlarının analiz edilmesinde kullanılacaktır. Bugüne kadar farklı alanlarda alınan patentleri analiz edip sonuçlarını paylaşan birçok çalışma yapılmıştır. Bu başlık altında, patent analizine ait önemli çalışmalara yer verilmektedir.

Xin Li ve arkadaşları [8] perovskit güneş pili teknolojisi alanında patent analiziyle birlikte twitter üzerindeki verilerle veri madenciliği yaparak bu alandaki yeni gelişmeleri ve teknolojileri belirleme ve izleme üzerine 2019 yılında bir çalışma yayınlamışlardır. Patent analizi için metin madenciliğinde konu tabanlı kümeleme tekniklerinden biri olan Lingo Algoritmasını kullanmışlardır. Lingo Algoritması anlamlı gruplar oluşturabilmek için

ortak arama ifadelerini gizli semantik indeksleme teknikleriyle birleştirir. Diğer kümeleme algoritmalarına göre daha rafine konular ve konu etiketlerini otomatik bir şekilde üretebilir. Twitter üzerindeki verileri alabilmek için Twitter Application Programming Interface kullanılmışlardır. Bu makalede Twitter kullanıcılarının Twitter'da tweetlerde yer alan gelişen teknolojiye olan duygusal tepkilerini değerlendirmek ve Twitter kullanıcılarının, zamanla gelişen teknolojiye ilişkin beklentilerinin değişen özelliklerini incelemek için bir Twitter duyarlılık analiz aracı olan “Sentiment 140”ı kullanmışlardır. Twitter kullanıcılarının beklentileri ve beklentileri ile ilgili bu analiz, gelişmekte olan teknolojinin gelecekteki gelişim yönelimleri hakkında bilgi sahibi olmalarına yardımcı olmuştur. Ayrıca, bu yazıda, Twitter kullanıcılarının Twitter'daki tweetlerini ve yeni çıkan teknolojiyle ilgili tweetleri analiz etmek için “author-topic over time” (ATOT) modelini kullanmışlardır. ATOT modeli, yazarların araştırma ilgi alanlarının gelişimini dinamik bir bakış açısıyla incelemek için üç boyutlu bir analiz modeli oluşturmak üzere, konuları, yazarları ve belgenin zaman bilgisini birleştirir. Çalışmada öncelikle Derwent Innovations Index (DII) veritabanından perovskit güneş pili teknolojisiyle ilgili patentler indirilmiş ve veri seti oluşturulmuştur. Sonra, veri seti temizlenmiş ilişkisiz olanlar silinmiş, çalışmaya dahil edilecek patentler metin madenciliği için metin formatına dönüştürülmüştür. Daha sonra, hazırlanan veri seti üzerinde Lingo Algoritması çalıştırılmıştır. Sonuç olarak, konuyla ilgili sonuçlar çıkarılmıştır. Twitter ile çalışma yapmak için öncelikle Twitter verileri toplanmış ve ilgili veriler indirilmiştir. Gereksiz ve tekrarlı tweetler silinmiş ve veriler zamana göre sınıflandırılmıştır. Twitter verilerini analiz etmek için Natural Language Toolkit (NLTK) ve Stanford NLP kullanılmışlardır. Aynı zamanda, Twitter kullanıcılarının profesyonelliklerini ve ilgi alanlarını da analiz etmişlerdir. Çalışmada Twitter kullanıcılarının ortaya çıkan teknolojiye yönelik algı, tepki ve beklentilerini tanımlamışlardır. Sonuç olarak, patentlerden ve Twitter verilerinden edinilen sonuçlar karşılaştırılmıştır. Güneş pili teknolojisi alanında 2012 yılından itibaren edinilen patentlere karşılık, Twitter’da ise bu konu 2013 yılında konuşulmaya başlanmıştır. 2013 yılında atılan tweetler tavan yapmış ve ilerleyen zamanlarda azalarak devam etmiştir. Alınan patentler 2012’den 2016’ya kadar güneş pili teknolojisi için yıllar geçtikçe farklı alanlar içermeye başlamıştır. Örneğin 2012 yılında organik malzemeler kategorisi için

patent alınırken bu başlık 2016 yılında organik-inorganik hibrit malzemeler olarak değişmiştir.

Jingfeng Chen ve arkadaşlarının [9] 2017 yılında yayımladığı bu makalede medikal veri madenciliği yöntemleriyle elektronik sağlık kayıtları alanında 2000-2016 yılları arasında yapılan araştırmaların analizi yapılmıştır. Toplamda 2516 kayıt incelenmiştir ve bu kayıtlar PubMed, the Web of Science (WOS) Core Collection ve Association for Computing Machinery (ACM) veritabanlarından elde edilmiştir. Yönelimleri belirlemek için patent Dirichlet Allocation (LDA) and Topics Over Time (TOT) modelleri kullanılmıştır. TOT modeli önemli konuları ayıklamak için kullanılırken, LDA modeli konu oluşturma, bölme, birleştirme ve TOT'u analiz etmek için kullanılmıştır. Çalışmanın sonucunda Elektronik Sağlık Kayıtları için veri madenciliği genel araştırma çerçevesinde veri, yöntemler, bilgi ve karar seviyeleri de dahil olmak üzere konu-oluşum ilişkilerini ve alan bilgisini birleştirerek sunmuşlardır. Bu çalışma, elektronik sağlık kayıtları alanında çalışan akademisyenler için üst düzey bir bilgi sağlayabilir ve sağlık bilgisi keşfi, tıbbi karar desteği ve halk sağlığı yönetimi konularında tıbbi veri madenciliği teknikleri konusunda seçimlerini yönlendirebilir.

Han Pil Cho ve arkadaşları [10] yüksek katlı bina yapımında gelecek vaat eden teknolojiyi tahmin etmek için patent analizi üzerine makale yayımlamışlardır. Bu çalışmada Amerika Birleşik Devletleri (ABD), Avrupa, Kore, Çin ve Japonya'dan 1995-2013 yılları arasında yüksek katlı binalar ile ilgili alınan 2875 makale incelenmiştir. Teknoloji tahmini için OS Matrix metodu kullanılmıştır. OS Matrix metodu teknoloji tahminlerinde kullanılan bir yöntem olup x eksenini çözülmesi beklenen durumu y-ekseni ile ilgili problemi göstermektedir. Yaptıkları çalışmanın sonuçlarına bakıldığında Kore ve Çin bu teknoloji alanında gelişme göstermişlerdir. Hatta Çin bu konuda şiddetli büyüme gösteren ülke olmuştur. Çalışmanın sonucunda ortaya çıkan ve önerilen alanlar olarak belirtilen alanlar ise şöyledir; yüksek katlı bina inşaatının verimliliğini artırmak için izleme teknolojisi, bilgi modelleme teknolojisine dayalı inşaat aşamasında enerji tüketiminin azaltılması ve güvenlik yönetimi teknolojisidir. Bu çalışma gelecekteki yüksek katlı bina yapımı çalışmaları için yönlendirmeler yapmayı hedeflemiştir.

Gabjo Kim ve Jinwoo Bae'nin 2017 yılında yaptıkları çalışmada [11] patent analizini kullanarak sağlık alanında yeni teknolojileri tahmin etmede yeni bir yöntem önermişlerdir. Önerilen yöntemin genel süreci üç adımdan oluşmaktadır. İlk adım;

teknoloji kümelerini oluşturmak için ortak patent sınıflandırması (CPC) temelinde sınıflandırma yapmaktır. İkinci olarak, ilk adımda oluşturulan kümelerin yapıları incelenmiştir. Son olarak, oluşturulan patent kümelerinin umut verici olup olmadığı değerlendirmek için üçlü patent aileleri, bağımsız patentler gibi patent göstergeleri analiz edilmiştir. Patent veritabanı olarak USPTO kullanılmıştır. 2002-2009 yılları arasındaki verileri eğitim kümesi, 2010-2013 yılları arasındaki verileri de test kümesi olarak kullanmışlardır. Sonuç olarak, eğitim setinin gelecek vadeden bir teknoloji kümesi olarak göreceli olarak daha yüksek gösterge değerlerine sahip olan teletıp teknolojisini seçmişlerdir. Gelecek vaat eden diğer teknoloji kümelerini sağlık bakımı iş sistemi, veri yönetimi ve teletıp teknolojileri olarak belirlemişlerdir. Bu kümelerin hem eğitim hem de test kümelerinde olduğu görülmüştür. Sonuç olarak, oluşturulan gruplardan anlaşılmıştır ki sağlık alanında artan bir şekilde yenilikler yapılmaktadır. Aynı zamanda sağlık endüstrisi alanında geçmiş ve gelecek alanında oluşan gruplarda farklılık görülmediği de ortaya konmuştur.

Sunghae Jun ve Sang Sung Park yaptıkları çalışmada [12] Apple firmasının yaptığı yenilikleri tahmin etmek amacıyla Apple'ın aldığı patentleri incelemişlerdir. Bu yöntemi Apple'ın gelişmelerini takip etmek için objektif bir yaklaşım olarak görmektedirler. Bu makalede yazarlar, Apple'ın teknolojik açıdan bakıldığında inovasyonunu incelemeye yönelik objektif bir yaklaşım olan patent başvurularını analiz ederek Apple'ın teknolojik yeniliklerini ve tahminini incelemeyi amaçlamışlardır. Apple'ın teknolojik yeniliklerini incelemek için yazarlar başvuru alan tüm patentleri incelemişler ve üç farklı yaklaşıma göre de modeller oluşturmuşlardır. Öncelikle, zaman serisi regresyonu ve çoklu linear regresyon metodlarıyla teknoloji haritası çizmek için istatistiksel modeller kullanmışlardır. İkinci olarak, Apple'ın boşta kalan teknoloji alanlarını keşfetmek için tüm patentleri kümelemişlerdir. Son olarak, Apple'ın geleceğinin merkezindeki teknolojileri analiz etmek için sosyal ağ analizinden faydalanmışlardır. Bu araştırmada, Apple'ın keşfedilmemiş teknoloji alanlarıyla birlikte odaklandığı teknolojiler de bulunmuştur. Sonuçlarda Apple'ın en temel teknolojisinin G06F; elektrik dijital veri işleme, olduğu görülmüştür. Patentler beş kümeye bölünmüştür ve beş küme içerisinde iki tanesinin eleman sayısı az olduğu için henüz boş teknoloji alanları seçilmişlerdir. Oluşturulan kümeler araştırmacılar tarafından incelenmiştir ve G06F, H04B, H04N IPC kodlarının tüm kümelerde olduğu görülmüştür. G06F, H04B ve H04N teknolojileri

“elektriksel dijital veri işleme”, “iletim” ve “resimsel iletişim; televizyon gibi” teknolojilerdir. Apple’ın patentlerinin çoğunun, bu kodların teknolojileri alanında geliştirildiği bu çalışmayla öğrenilmiştir. Henüz boş sayılabilecek kümenin patent alanları incelendiğinde; “elektronik müzik ve konuşmanın iletişim ve işleme teknolojileri” alanlarının olduğunu görmüşlerdir. Bu nedenle, Apple’ın bu teknolojileri diğerlerinden daha fazla geliştirmesi gerekiyor sonucuna bu çalışmayla varılmıştır.

Evert Nijhof’un hazırladığı makalede [13] dünya çapında kapsama alanı olan ve on milyonlarca kayıt içeren bir patent veritabanında en ilişkili kayıtları nasıl bulacağımıza odaklanmaktadır. Aramaları yaparken kısıtlı bir zamanda sonuçlara ulaşılacağı düşünülmüştür. Kısıtlı zamanda en ideal sonuçlara ulaşmak için yapılması gerekenleri yazar şu şekilde sıralamıştır;

- Tarihe göre en yeniden eskiye doğru sıralamak gereklidir,
- “Title” ve “abstract” bölümlerinde arama yapmak tüm dokümanı aratmaktan daha mantıklıdır,
- Konuya özel anahtar kelimeler kullanılmalıdır,
- Aynı şekilde çok spesifik sınıflar kullanılmalıdır.

Makalede aynı zamanda verimli olmak için KISS (Keep It Simple and Straightforward) prensibinin de uygulanması gerektiği belirtilmiştir. Ayrıca, oluşturulan sorguların da kısa tutulması gerektiği belirtilmiştir. Patent araması yapılırken esnek olunmalı ve eldeki konuya göre arama şekline kolay adapte olmaları gereklidir.

Bilal ALP’in yaptığı yüksek lisans tezinde [14] patent analizi yaparak bulut bilişimi ile ilgili gelişmeleri takip etmeyi hedeflemiştir. Kullanılan patent veritabanları; USPTO, Free Patent Online ve WIPO’dur. 60000’e yakın patent analiz için kullanılmıştır. İlk bölümde, patentlerin yazı işleme teknikleriyle nasıl bulunduğu ve nasıl işlendiği detaylıca bahsedilmiştir. Bu bölümde en verimli yöntem term-variance seçilmiştir. İkinci bölümde k-means algoritması ile nasıl kümeleme yapıldığı anlatılmıştır. Benzerlik metriği olarak kosinüs benzerliği kullanılmıştır. Son bölümde ise bulut bilişim teknolojinin ilgili teknolojileri ve ilişkili olduğu teknoloji alanları ortaya çıkmıştır.

Fatma Altuntaş’ın yüksek lisans tezi için hazırladığı çalışmada [15] İnsansız Deniz Aracı Teknolojileri (İDAT) ile ilgili patentler kullanılarak bu teknolojiye ait teknoloji ağları oluşturulmuştur. Analiz için seçilen anahtar kelimeler; insansız deniz aracı, patent analizi, S-eğrisi, doğrusal regresyon analizi ve teknoloji ağıdır. Veritabanı olarak USPTO

seçilmiştir. AcclaimIP yazılım ile patentler elde edilmiştir. S-eğrisi ve lineer regresyon teknikleriyle veriler düzenlenmiştir. Teknolojik ilişki haritası sosyal network yazılımı ile görselleştirilmiştir. Çalışmanın sonucunda insansız deniz aracı teknolojileriyle ilgili patentlerden yola çıkarak patentlerin IPC kodlarının arasındaki ilişki ortaya çıkarılmıştır. Örneğin; teknelerin üzerindeki saldırı ya da savunma düzenleri, mayın döşeme, mayın tarama, denizaltılar, uçak gemileri teknolojileriyle radyo dalgalarıyla yön bulma teknolojileri ve elektrikli olmayan değişkenlerin ayarlanması ya da kontrolü için sistemler ilişkili çıkmıştır.

P.S Aithal ve arkadaşının yaptığı çalışmada [16] zaman geçtikçe değişen araştırma metodlarına yeni eklenen yöntemlerden birisi olan patent analizinden bahsedilmiştir. Patent analiz metodunun analiz, tanımlama ve yorumlama basamaklarından bahsedilmiştir. Yeni teknoloji ve eğitimleri tanımlayan patentler çok büyük bir veri olarak düşünülmektedir ve son 10 yılda 2500000 patent tüm dünyada oluşturulduğundan bahsedilmiştir. Makalede patent başvurularının yıldan yıla arttığından da bahsedilmiştir. Örneğin; 2001 yılında 1457000 olan rakam 2016 yılında 3127900'e kadar çıkmıştır. Makalede farklı konularda uzun yıllar boyunca bolca patentin yayınlandığını, araştırmacıların bunları analiz edip yorumlayabilmeleri için çok fazla fırsat olduğunu böylece yeni bilgilere ulaşabildikleri yazmaktadır. Çalışmada patent analizi yapmanın nedenleri şu şekilde sıralanmıştır;

- a. Yeni teknolojileri tahmin etmek
- b. Teknolojik değişimi haritalamak
- c. Patent ihlali riskini değerlendirmek
- d. Teknoloji yönelimlerini tanımlamak
- e. Teknolojinin etkilerini tahmin etmek
- f. Teknolojik konuları tahmin etmek
- g. Yönelimleri izlemek
- h. Teknolojiyi ölçmek
- i. Yeni icatları izlemek
- j. İş planlaması yapmak
- k. Yeni fikirleri takip etmektir.

José Carlos Rodríguez ve Mario Gómez yaptıkları çalışmada [17] 1965 – 2012 yılları arasında Latin Amerika ülkeleri tarafından yapılan patent başvurularını analiz edilerek

teknolojik alandaki yeni eğilimler belirlenmiştir. World Intellectual Property Organization (WIPO) veritabanı olarak kullanılmıştır. Arjantin, Brezilya, Şili ve Meksika'nın verileri incelenmiştir. Sonuçlara baktığımızda, Arjantin'de 1990'larda patentlerde ara verildiği görülmüştür. Arjantin'de bu yıllarda yaşanan rejim değişikliği buna sebep olmuş olabilir. Brezilya ve Şili'de ise bu aralar 2000 yıllardadır. Buradan Brezilya ve Şili'nin arasında bir rekabet olduğu görülebilmektedir. Ayrıca Brezilya ve Şili diğer Latin Amerika ülkelerine göre teknolojik alanda daha fazla yenilikler yapmaktadır. Sonuçlara bakıldığında Meksika'nın ABD'deki mülkiyet ve diğer politikalardan daha fazla etkilendiği görülmektedir. Brezilya ve Şili ise diğer ülkelere oranla politikalar açısından daha özgür davranmaktadır. Bu çalışma Latin Amerika ülkelerinde bilim, teknoloji ve yenilik (STI) politikası ve yenilikçi yetenek gelişimi arasındaki ilişkiyi incelemek için ilk katkı olarak düşünülmektedir ve alternatif politikaların diğer birçok ülkedeki etkisini anlamak için daha fazla araştırma yapılması gerektiği belirtilmektedir.

Heena Girdher ve Poonam Gaur'un yaptıkları çalışmada [18] metin madenciliği yöntemleri hakkında bilgi verilmiştir. Metin madenciliğinin tanımını yapılandırılmış veya yapılandırılmamış verilerden anlamlı bilgiler çıkarmak için kullanılan teknoloji olarak yapılmıştır. Metin madenciliği için dokümanı okuyan ve kullanıcı için anlayan yazılım olarak bahsedilmiştir. Anahtar kelime aratmanın metin madenciliğinden farkı burada görülmektedir çünkü kullanıcı anahtar kelime aratarak bulduğu tüm dokümanları kendisi okuyup anlamalıdır. Metin madenciliği basamakları şu şekilde sıralanmıştır;

- Metin dokümanlarını toplama.
- Toplanan dokümanlar gereksiz verilerden kurtulmak için temizlenmelidir.
- Metin belgesi madencilik için uygun formlara dönüştürülür.
- İlişkisiz özellikler metinden çıkartılır.
- Metin madenciliği teknikleri veri madenciliği teknikleriyle birleştirilir.
- Elde edilen patentler değerlendirilir.

Özetle bu derlemede metin madenciliği süreci tanımlanmıştır. Bilgi edinme, bilgi çıkarma, özetleme, sınıflandırma, kümeleme gibi metin madenciliği teknikleri anlatılmıştır.

Kisik Song ve arkadaşının yaptığı çalışmada [19] mevcut patent analizi yöntemlerine alternatif olarak “post hoc patent analysis” adında yeni bir metod önermişlerdir çünkü mevcut yöntemde kullanılan patentin referans bilgisine bakılması metodu yakın zamanda tescilli patentlere uygulanamamaktadır. Ek olarak çalışmada yeni teknoloji için müşterilerin geri dönüşleri de toplanmıştır. Müşteri geri dönüşlerinden gelen verilerden anahtar kelimeler belirlenmiştir. Çalışmada bulunan anahtar kelimelerle ilgili mevcut olan teknolojiler için patent bilgileri toplanmıştır. Müşterilerden ve patentlerden alınan verilerle bir portföy oluşturulmuştur. Sonuçta toplanan verilerden oluşturulan portföy ile yeni teknolojilerin özellikleri belirlenmeye çalışılmıştır. Belirlenen yeni teknolojiler üç gruba ayrılmıştır bunlar; yüksek fayda veren, uygulanabilirliği yüksek olan ve süresi uzun olan teknolojilerdir. Yüksek faydaya sahip teknolojiler, daha fazla sayıda katkı sağlayan patentlerin teknolojileridir. Örneğin mucitler ve başvuru sahiplerinin patentleri gibi. Güçlü teknolojik temelleri olan teknolojiler ise daha fazla uygulanabilir olan teknolojiler olarak belirtilmiştir. Bu grupta belirtilen patent referans sayıları diğerlerine göre daha fazladır. Çalışmanın sonucunda araç kapı sistemleri ile ilgili patent gruplarının teknoloji ve pazarlama alanında değerlendirme puanları belirlenmiştir. Hem teknolojik hem de pazarlanabilirlik açısından ilk 10’a girebilen beş patent hakkında bilgi verilmiştir. Bu teknolojiler; kapıları açıp kaparken araç içi ışıkları açan bir teknoloji, bir sürücüyü arabanın içinde koruyan bir kapı kilitleme sistemi, araç kapı panellerinde kullanılmak üzere bir ısıtma cihazı, araç kapı panelleri içindeki hareketli şanzıman parçalarının bağlantısını kolaylaştırmak için kullanılan teknoloji ve araç kapısı açma / kapama aygıtı ve araç kapı kilitlerinin birden fazla yerine izin veren kilitleme teknolojisidir.

Assad Abbas ve arkadaşlarının yapmış olduğu çalışmada [20] patent analizinde en son teknolojiyi sunmaktadır. Ayrıca piyasada patent analizi için kullanılan tekniklerin özelliklerini, zayıf yönlerini vurgulamışlardır. Patent analizi için ScienceDirect, ACM digital library, IEEE digital library ve CiteSeerX veritabanları kullanılmıştır. Anahtar kelime olarak şunlar seçilmiştir; “patent analysis tools and techniques”, “visualization approaches in patent analysis” ve “text mining and patent analysis”. Son yıllarda alınan patentler seçilmiş olup literatür taraması için sonuçta 22 makale seçilmiştir. Patent yazılarından anlamlı veriler elde etmek için metin analizi yapılmış ve sonuçlar için görselleştirme yaklaşımlarına başvurulmuştur. Bu yaklaşımlarda patentleri kümelemek için k-means algoritmasına başvurulmuştur. Burada yapılan literatür taraması,

araştırmacılara birleşik formdaki patent analiziyle ilgili en son araştırma çabalarını bulmada yardımcı olacaktır.

Serkan Altuntaş ve arkadaşlarının yaptığı çalışmada [21] patent verisine dayalı teknoloji başarısını tahmin etmek için yeni bir yöntem önerilmiştir. Teknoloji tahmini için dört kriter; teknoloji yaşam döngüsü, difüzyon hızı, patent gücü ve genişleme potansiyeli göz önünde bulundurulmuştur. Belirtilen dört farklı kriterlerden elde edilen sonuçları birleştirmek için veri füzyon algoritması uygulanmıştır. Araştırmanın sonucunda bulunan tahmin metodunun kullanışlılığını ve potansiyelini ölçmek için ABD'nin üç teknoloji alanında alınan patentlerinden faydalanılmıştır. Teknoloji alanları; TFT-LCD ekran, flaş bellek sistemi ve kişisel dijital asistan kullanılarak gösterilmiştir. Bu patentlerden elde edilen sonuçlar, kişisel dijital asistan teknolojisinin diğer teknolojilere göre tercih edildiğini göstermektedir. TFT-LCD ekran ve flash bellek sistemi teknolojilerindeki yatırımlar eşit önceliğe sahiptir. Çalışmada USPTO veritabanı kullanılmıştır. Toplanan verilen S-Eğrisi çizilmiştir. Bulunan sonuçlar şu şekildedir; kişisel dijital asistan teknolojisi teknolojik yayılma hızı olarak 1. sıradadır ve aynı zamanda bu teknoloji en yüksek patlama potansiyeline sahip teknolojidir. Flaş bellek teknolojisi yayılma hızı olarak 2. sıradadır. TFT-LCD ekran teknolojisi en çok patent alınan alandır fakat yayılma hızı en düşük teknoloji olarak çıkmıştır. Çalışmanın sonuncunda önerilen yöntemle aşağıdaki gibi sorulara cevaplar verilmiştir: yatırım için hangi teknolojiler uygundur ve teknolojiler arasında önceliklendirme nedir?

Cemil Başpınar'ın Türk Patent Enstitüsü'nde yaptığı uzmanlık tezinde [22] Türkiye'nin 1980–2007 yıllarına ait patent verileri ve 1995–2007 yıllarına ait faydalı model verileri, Fraunhofer Institute for Systems and Innovation Research, Observatoire des Sciences et des Techniques ve University of Sussex, Science and Policy Research Unit tarafından hazırlanmış olan patent-endüstriyel sınıflandırma uyum tablosu kullanılarak sektörel bazda (NACE Rev. 1.1) hesaplanmıştır. Hazırlanan tezde öncelikle patent verilerine ilişkin literatürden bahsedilmiş 2. bölümde ise patentlerin dünyada nasıl sınıflandırıldığı anlatılmıştır. Türk Patent Enstitüsü veri tabanında kayıtlı bulunan veriler, Uluslararası Patent Sınıflandırma Sistemi'ne (IPC) göre sınıflandırıldığı belirtilmiştir ve hâlihazırda IPC'nin 8. versiyonu kullanılmaktadır denmiştir. IPC'nin sekiz bölümünden bahsedilmiş ve bölümler şu şekilde sıralanmıştır;

- A- İnsan İhtiyaçları
- B- İşlemlerin Uygulanması, Taşıma
- C- Kimya, Metalürji
- D- Tekstil, Kâğıt
- E- Sabit İnşaatlar
- F- Makine Mühendisliği, Aydınlatma, Isıtma, Silahlar, Tahrip Malzemeleri
- G- Fizik
- H- Elektrik

Bölümlerden sonra sınıf, alt sınıf, ana grup ve alt gruplar numaralandırılmıştır. 3. bölümde istatistiklerin nasıl elde edildiği anlatılmıştır. Bu çalışma kapsamında 62931 adet patent ve faydalı model başvurusu üzerinde çalışılmıştır. Farklı alanlar için yerli patent ve faydalı model sayıları tezde grafiklerle belirtilmiştir.

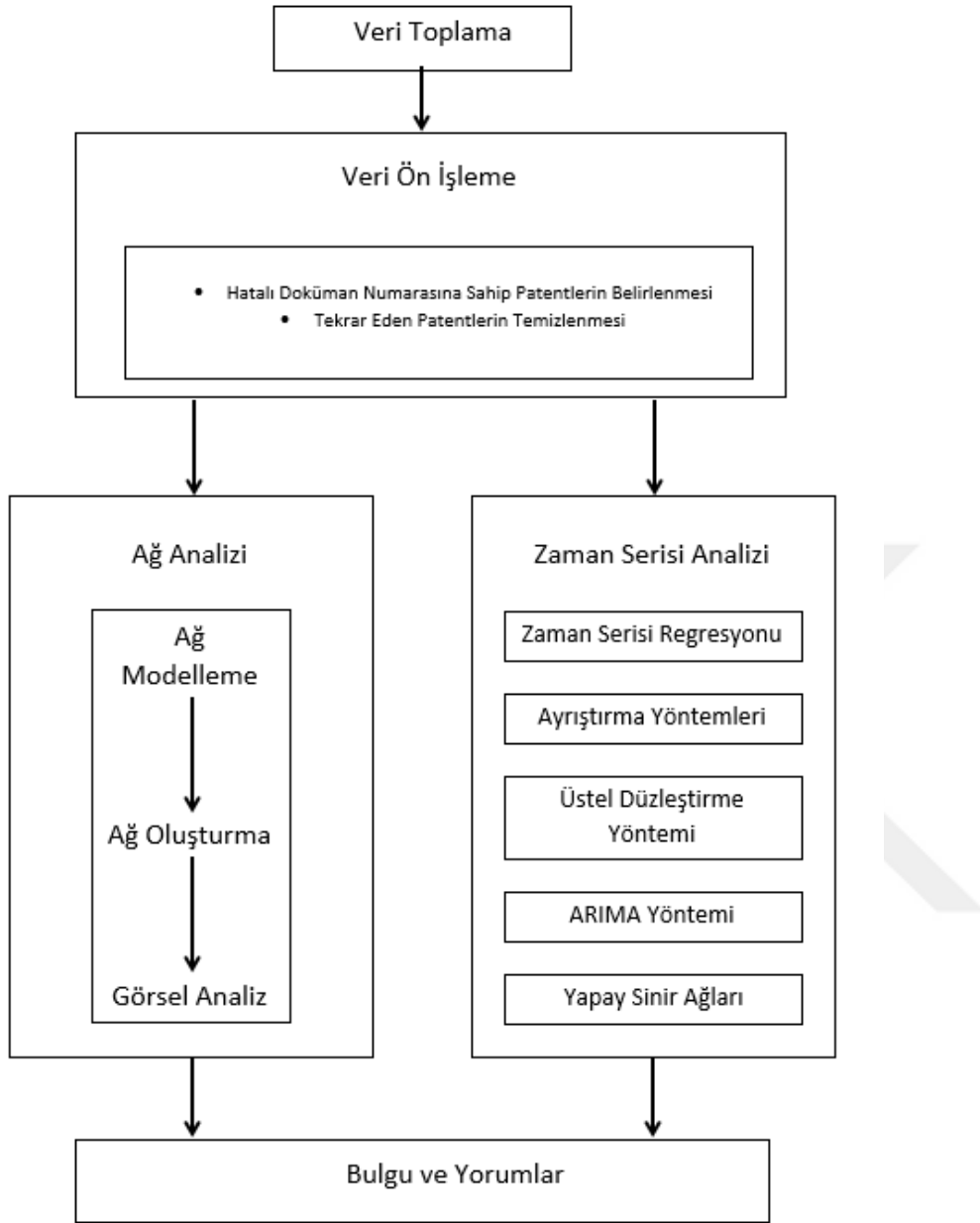
2. MATERYAL VE YÖNTEM

Patent belgeleri teknolojiyle ilgili alanları analiz etmek için eşsiz kaynaklardır. Bu belgeleri kullanarak pek çok çalışma yapılabilir. Patent belgeleri sadece ilgili teknoloji alanını içermeyip aynı zamanda o teknolojinin yönlendirdiği diğer alanları da ifade eden bilgileri kapsarlar. Bu analizi yapabilmek için öncelikle patent dokümanlarının yapısını ve bölümlerini iyi anlamak gereklidir çünkü bu bölümler çıkarılabilecek alanla ilgili pek çok ipucu verecektir.

Bu tez çalışmasında amaç patent analizi yaparak bilgi güvenliği alanında teknoloji tahmini yapmaktır. Patentleri işlemeyen önce nasıl bir yöntem uygulandığı belirtilmelidir. Bu çalışmada kullanılan yöntem, Şekil 2-1’de gösterilen adımları içermektedir.

Tüm işlemlere başlamadan önce teknoloji alanı seçilmelidir. Günümüzde bilgi güvenliği alanı pek çok teknoloji alanını etkilemekte birlikte her geçen gün hızla büyümektedir. Bu çalışmada bilgi güvenliği alanındaki teknolojilerin arasındaki ilişkilerin, farklılıkların ve teknolojik geçişlerin görülmesi planlanmaktadır. Çalışmada tüm süreç öncelikle patentleri toplayarak başlamıştır.

Bu tez çalışmasının amacı patent veri seti oluşturulması ve sosyal ağ analizi, zaman serisi analizi çalışmalarını ortak bir paydada buluşturarak bilgi güvenliği alanındaki patentlerden teknolojinin yönünü ortaya koymaktır. Tez kapsamında iki aşamalı bir araştırma süreci izlenmiş ve analizler için Gephi, Cytoscape yazılımlarından ve R programlamadan faydalanılmıştır.



Şekil 2-1 Gerçekleştirilen Çalışmanın İşlem Adımları

2.1 Patent Veri Tabanlarında Arama Yapmak

İnternet üzerinden patent dokümanlarına erişilebilecek pek çok site mevcuttur. Ancak bu sitelerin bazıları ücretliyken bazıları da internet sayfalarından sadece ilk 500 sonuca erişilmesine izin vermektedir. Çevrimiçi ve ücretsiz bir şekilde patent dokümanlarına erişebilmek için bu çalışmada INPADOC veritabanı ve OPS internet servisi kullanılmıştır. INPADOC uluslararası patent doküman servsidir, patentlerin

bibliyografik verilerinin bulunduğu veri tabanıdır ve haftalık olarak Avrupa Patent Ofisi tarafından güncellenmektedir. Avrupa Patent Ofisi, ESPACENET veri tabanı ile orantılıdır ve INPADOC veri tabanına erişim sağlamaktadır [23]. ESPACENET dünya çapında kapsama alanına sahip olan 1782’den günümüze kadar olan buluşlar ve teknik gelişmeler hakkında bilgilere ücretsiz erişim sunmaktadır. Bu veri tabanı günlük olarak uzmanlar tarafından güncellenmekte olup dünyanın dört bir yanından 110 milyondan fazla patent belgesine ilişkin veriler içermektedir [24]. Çalışmada verileri bir yazılım programı aracılığıyla toplayabilmek için Avrupa Patent Ofisi veritabanının internet servisi (OPS) kullanılmıştır. OPS verilere erişimi sağlayan bir internet ağı hizmetidir.

2.1.1 Patent Uzmanlarından Tavsiye Almak

Çalışmada kullanılacak veri setini toplamak için internet servisiyle gönderilecek sorguların oluşturulması gereklidir ve bu sorgularda kullanılacak olan anahtar kelimelerin doğru seçilmesi önemlidir. Patent uzmanları anahtar kelimeleri belirlemek için farklı yöntemler önerebilirler. Çünkü patent dokümanlarında arama yapmak uzmanlık gerektiren bir iştir. Bu konuda deneyimli kişilerden yardım almak doğru veri setini oluşturmak adına çok önemlidir. Herhangi bir kombinasyon yapmadan sadece tek bir kelime ile yapılan patent aramaları sonucunda çok farklı dokümanlara da erişilebilmektedir. Örneğin; “security” (güvenlik) kelimesini patent dokümanlarının başlık kısmında aratıldığında sadece bilgi güvenliği ile ilgili patentler değil aynı zamanda şapka, anahtar gibi farklı güvenlik alanlarını da kapsayan patentler bulunabilmektedir. Bu sebepten dolayı bu çalışmada patent uzmanlarının tavsiye ettiği şekilde patent araması yapılmış aşağıdaki bölümlerde detaylandırıldığı gibi uzmanların belirlediği rotadan ilerlenerek tutarlı bir veri seti oluşturulmaya çalışılmıştır.

2.1.2 Bilgi Güvenliği Alanında Lider Firmaları Belirlemek

Bilgi güvenliği alanında alınan patentlere erişebilmek için patent arama aşamasında doğru anahtar kelimelerin kullanılması şarttır. Doğru kelimeleri seçmek için de önceden bu alanda alınmış patentlerden yararlanmak en uygun yol olacaktır. Önceden bilgi güvenliği alanında alınan patentlere erişebilmek için bu alanda lider olan firmaların aldığı patentler incelenmiştir. Böylelikle bu firmaların patent dokümanlarında sıklıkla kullandıkları

kelimeler belirlenebilecektir. İnternette yapılan arařtırmalar sonucunda son yıllarda bilgi güvenliđi alanında en fazla yatırım yapan firmaların řunlar olduđu bulunmuřtur;

1. CISCO
2. IBM
3. MICROSOFT
4. AMAZON
5. SYMANTEC

Veri setini oluřturmak iin patent arama ařamasında kullanılacak anahtar kelimelere yukarıda adları geen beř firmanın son  yılda aldıđı patent dokmanları incelenmiřtir.

2.1.3 Bilgi Gvenliđi Alanında En Popler IPC Sınıflarını Belirlemek

2.1.2.'de bahsedilen firmaların bilgi güvenliđi alanında aldıkları patentlere eriřebilmek iin sorgu cmlesine IPC sınıflarının da eklenmesi gereklidir. Bu sayede dođrudan ilgili alandaki patentlere eriřilebilecektir. Bu alıřmada anahtar kelime belirlemek iin oluřturulan sorgularda IPC sınıflarından H04L9 ve G06F21 kodları kullanılmıřtır. Bu sınıflar IPC sınıflarının tanımları okunarak analiz edilmek istenen alana en yakın gelen sınıf seilerek bulunmuřtur. H04L9; elektrik blmn temsil eden H sınıfının, elektrik haberleřme tekniđi alt sınıfının (H04), sayısal bilgilerin iletimi alanının (H04L), gizli veya güvenli iletiřim gerekli dzenlemeler (H04L9) ile ilgili patentleri iermektedir. G06F21 ise fizik (G) sınıfının alt grubu olan bilgi iřlem, hesaplama veya sayma (G06) sınıfının elektrik diđital veri iřleme alanının (G06F), bilgisayarları, bunların bileřenlerini, programları veya verileri yetkisiz faaliyetlere karřı korumak iin gvenlik dzenlemeleri (G06F21) ieren patentleri kapsamaktadır. (Tablo 2-1)

Tablo 2-1 Seçilen IPC Sınıflarının Tanımları

IPC Sınıfı	Teknoloji Alanı
H	Elektrik
H04	Elektriksel Haberleşme Tekniği
H04L	Sayısal Bilgilerin İletimi
H04L9	Gizli veya Güvenli İletişim için Düzenlemeler
G	Fizik
G06	Bilgi İşlem, Hesaplama veya Sayma
G06F	Elektriksel Dijital Veri İşleme
G06F21	Verileri Yetkisiz Faaliyetlere Karşı Korumak için Yapılan Güvenlik Düzenlemeleri

2.1.4 Anahtar Kelimeleri Belirlemek için Gerekli Sorgu Cümlelerini Oluşturmak

Bölüm 2.1.2. ve 2.1.3.'te belirtilen yöntemlerle belirlenen bilgi güvenliği alanında lider firmalar ve popüler IPC sınıflarıyla 2017 ile 2019 yılları arasında alınan patentlerde en sık kullanılan kelime ve kelime gruplarını belirlemek için sorgu cümleleri oluşturulmuştur.

Örnek sorgu cümleleri;

'pa="CISCO TECH INC" and pd within "2017 2019" and (ipc="H04L9")'

'pa="CISCO TECH INC" and pd within "2017 2019" and (ipc="G06F21")'

Yukarıdaki örnek sorgularda geçen “pa” başvuran parametresini filtrelerken, “pd” yayınlanma tarihi ve “ipc” ise IPC sınıfını temsil etmektedir.

2.1.5 Anahtar Kelimeleri Belirlemek

Anahtar kelimeleri belirlemek için gerekli sorguları internet servisi aracılığıyla çalıştırdıktan sonra beş firmanın son üç yılda, belirlenen iki IPC sınıfıyla ilgili aldığı toplam patent sayısı 3283 olarak belirlenmiştir. 3283 eşsiz patentin başlık kısmında sıklıkla kullanılan kelimeler ve kelime grupları, kombinasyonları bulunmuştur. Bu kombinasyonları belirlemek için oluşturulan ön veri setinin başlık kısmında geçen 34025

kelime analiz edilmiştir. Analiz sonucunda sıklıkla kullanılan kelime grupları belirlenmiştir. Böylelikle veri setini oluşturmak için kullanılacak olan sorgularda patentlerin başlık kısımlarında bu kelimeler aratılmıştır.

Analiz sonuçlarına göre çıkan kelime grupları şu şekildedir;

GRUP A -1 : hardware*, software*, malware*

GRUP A-2: serv*, comput*, execut*

GRUP A-3: virtual*, digital*, data*, system*

GRUP A-4: inform*, network*, device*, method*, applicat*

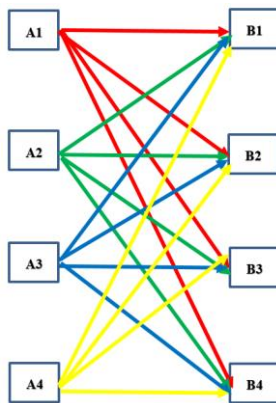
GRUP B-1: verificat*, authenticat*, autoriz*, block* chain*, blockchain*

GRUP B-2: crypto*, encrypt*, password*, encod*, key*

GRUP B-3: protect*, secre*, secur*

GRUP B-4: priva*, safe*, sure*, policy*

Şekil 2-2’de gösterildiği gibi A ve B grupları kendi aralarında çaprazlanarak son 10 yıl için tüm anahtar kelimeleri kapsayacak şekilde sorgular oluşturulmuştur. Örneğin şekildeki ilk çaprazlama olan A1 ve B1 gruplarının çaprazlamasının sonucunda şu şekilde bir sorgu cümlesi elde edilmiştir; (ti any "hardware* software* malware*") and (ti any "verificat* authenticat* autoriz* block* chain* blockchain*") and (pd = "20090101 20090104 ").



Şekil 2-2 Kelime Grupları Çaprazlaması

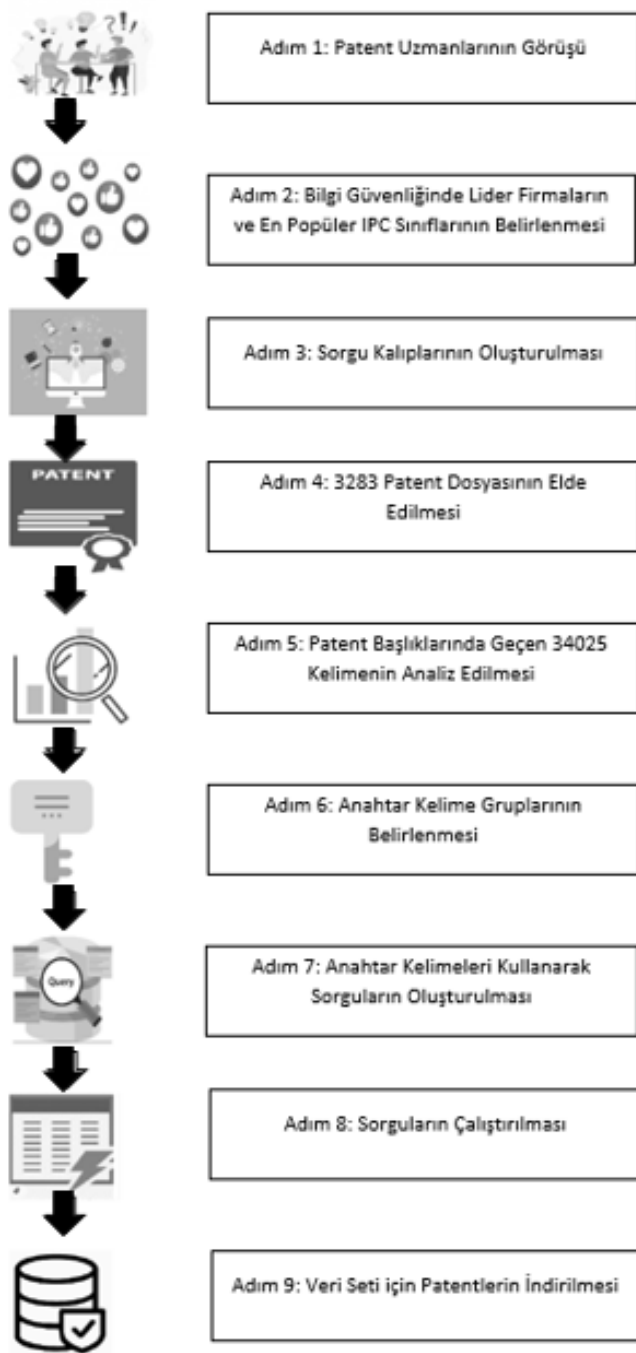
2.1.6 Veri Setini Oluşturmak

Başlık 2.1.4'te bahsedilen anahtar kelimeler kullanılarak patent başlık bölümünde son 10 yılda alınan patentlerden oluşan bir veri seti oluşturulmuştur. Bu çalışmada EPO internet servisi kullanıldığından ve internet servisi tek seferde en fazla 100 patenti indirmeye izin verdiğinden dolayı sorgular gelen sonuca göre hafta hafta ya da gün bazlı olarak hazırlanmıştır. Bu çalışmada veri setini oluşturmak için toplamda 27290 sorgu çalıştırılmıştır. Hazırlanan sorgular internet servisi aracılığıyla otomatik olarak çalıştırılıp patentlerin bibliyografik verilerinin bulunduğu xml dosyaları indirilmiştir.

```
<?xml version="1.0" encoding="UTF-8"?><?xml-stylesheet type="text/xsl" href="/3.2/style/exchange.xsl"?>
<ops:world-patent-data xmlns="http://www.epo.org/exchange" xmlns:ops="http://ops.epo.org" xmlns:xlink="http://www.w3.org/1999/xlink">
  <exchange-documents>
    <exchange-document system="ops.epo.org" family-id="41669507" country="US" doc-number="2011142239" kind="A1">
      <bibliographic-data>
        <publication-reference>
          <document-id document-id-type="docdb">
            <country>US</country>
            <doc-number>2011142239</doc-number>
            <kind>A1</kind>
            <date>20110616</date>
          </document-id>
          <document-id document-id-type="epodoc">
            <doc-number>US2011142239</doc-number>
            <date>20110616</date>
          </document-id>
        </publication-reference>
        <classifications-ipcr>
          <classification-ipcr sequence="1">
            <text>H04K 1/ 00 A I</text>
          </classification-ipcr>
        </classifications-ipcr>
        <patent-classifications>
          <patent-classification sequence="1">
            <classification-scheme office="" scheme="CPC"/>
            <section>H</section>
            <class>04</class>
            <subclass>W</subclass>
            <main-group>12</main-group>
            <subgroup>02</subgroup>
            <classification-value>I</classification-value>
          </patent-classification>
        </patent-classifications>
      </bibliographic-data>
    </exchange-document>
  </exchange-documents>
</ops:world-patent-data>
```

Şekil 2-3 US2011142239.A1 No'lu Patente Ait Bibliyo Belgesinden Bir Kesit

Bölüm 2.1.5 Anahtar Kelimeleri Belirlemek başlığının altında anlatılan yöntemle oluşturulan sorgular çalıştırılmış ve sonuçta 284972 patent dosyası indirilmiştir. İndirilen patent dosyalarından US2011142239.A1 numaralı patentin bibliyo belgesinden alınan bir kesit Şekil 2-3'te gösterilmektedir. Şekilde patentin doküman numarası, türü, tarihi, ülkesi ve dahil olduğu sınıfların birkaçı vardır. İndirilen patent bibliyografik dosyaları xml dosyası ayrıştırma yöntemleriyle analiz edilip çalışmalara devam edilmiştir.



Şekil 2-4 Veri Seti Toplama Aşamaları

2.2 Metin Ön İşleme

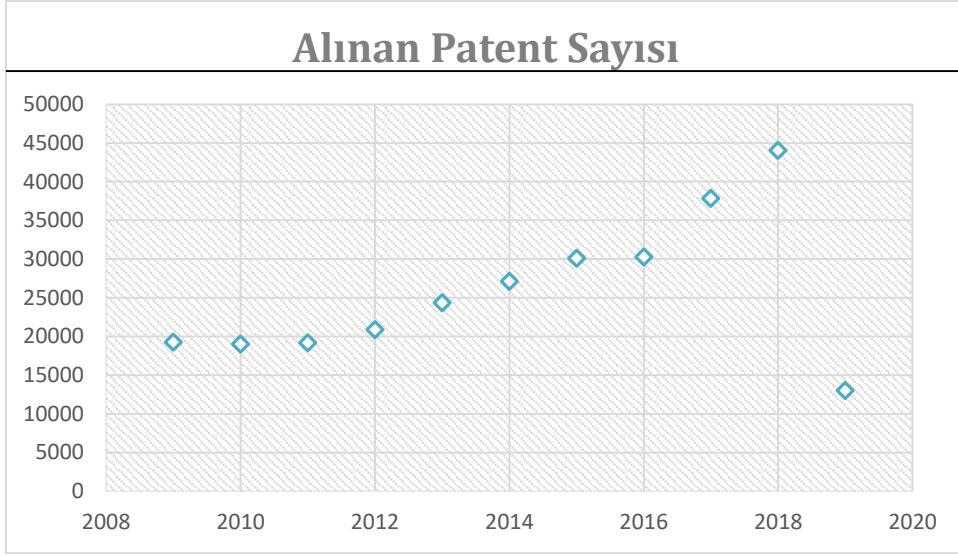
İndirilen xml dosyaları patent belgesi bölümlerine dayanarak ayrıştırıldığında yıllara, ülkelere ve kodlara dayanan farklı sonuçların çıktığı görülmüştür.

2.2.1 Veri Setinin Yıl Bazında İncelenmesi

Ocak.2009-Mayıs.2019 tarihleri arasında bilgi güvenliği alanında alınan patent adetlerini incelediğimizde sonuçların yıllar geçtikçe doğru orantılı olarak arttığı görülmektedir.

Tablo 2-2 Veri Setinin Yıllara Göre Sıralaması

Yıl	Alınan Patent Sayısı	Sıralama
2009	19262	8
2010	19011	10
2011	19153	9
2012	20875	7
2013	24333	6
2014	27130	5
2015	30096	4
2016	30254	3
2017	37819	2
2018	44023	1
2019	13013	11



Şekil 2-5 Veri Setinin Yıllara Göre Grafiği

2.2.2 Veri Setinin Ülke Bazında İncelenmesi

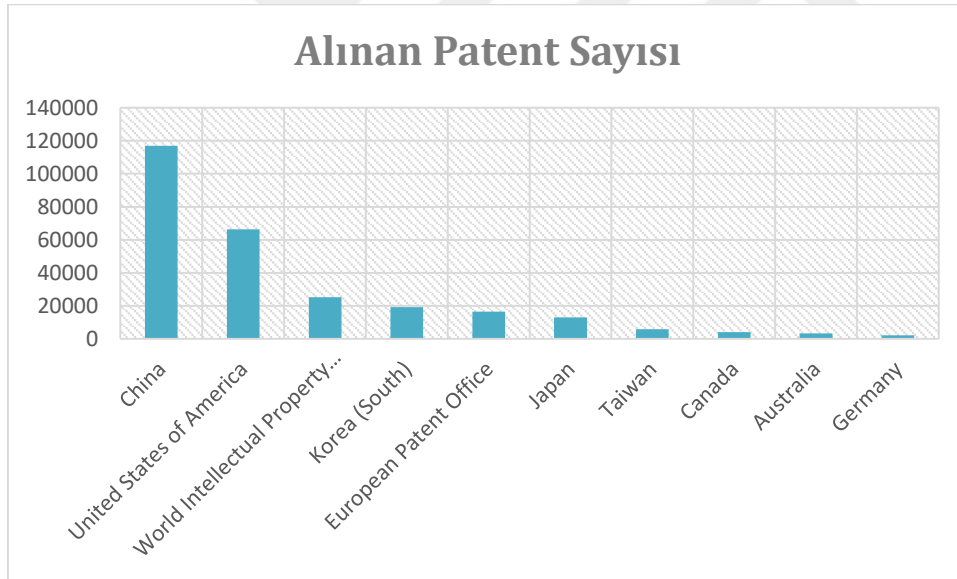
Veri seti ülke bazında incelendiğinde 56 farklı ülke tarafından son 10 yılda bilgi güvenliği alanında patent alındığı görülmüştür. Bu ülkelerden Çin, ABD ve Kore ilk üç sırada yer almaktadır.

Tablo 2-3 Veri Setinin Ülkelere Göre Sıralaması

Ülke	Ülke Kodu	Alınan Patent Sayısı	Sıralama
Çin	CN	116907	1
ABD	US	66397	2
World Intellectual Property Organization (WIPO)	WO	25352	3
Güney Kore	KR	19245	4
European Patent Office	EP	16650	5
Japonya	JP	13149	6

(Tablo 2-3 Devam)

Tayvan	TW	5954	7
Kanada	CA	4196	8
Avustralya	AU	3411	9
Almanya	DE	2243	10
Rusya	RU	2208	11
Birleşik Krallık	GB	1794	12



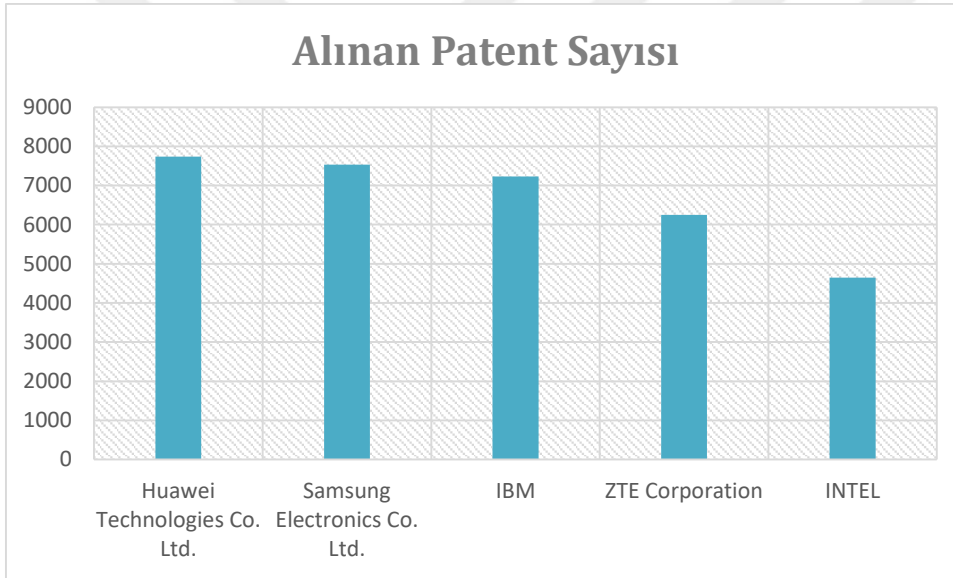
Şekil 2-6 Veri Setinin Ülkelere Göre Grafiği

2.2.3 Veri Setinin Başvuran Firma Bazında İncelenmesi

Toplanan veri seti başvuru yapan firma bazında incelendiğinde bazı firmaların diğerlerine göre açık ara fark attığı görülmektedir. Sıralamadaki en büyük patent sayısı değerine sahip firma Huawei Technologies Co. Ltd. iken, bu firmayı Samsung Electronics Co. Ltd. ve IBM firmaları yakın farklarla takip etmektedir.

Tablo 2-4 Veri Setinin Firmalara Göre Sıralaması

Firma Adı	Alınan Patent Sayısı	Sıralama
Huawei Technologies Co. Ltd.	7739	1
Samsung Electronics Co. Ltd.	7532	2
IBM	7233	3
ZTE Corporation	6252	4
INTEL	4649	5



Şekil 2-7 Veri Setinin Firma Bazında Sıralaması

2.2.4 Veri Setinin CPC Kodlarına Göre İncelenmesi

Bu çalışmada hazırlanan veri setinde alınan patentlerin ait oldukları CPC sınıfları ana sınıf, alt sınıf ve alt gruplar olmak üzere üç ayrı kategoride değerlendirilmiştir. CPC kodlarına göre analiz sonuçlarına Tablo 2-5 ve Şekil 2-8'den bakılabilir.

Tablo 2-5 Veri Setinin CPC Sınıflarına Göre Sıralaması

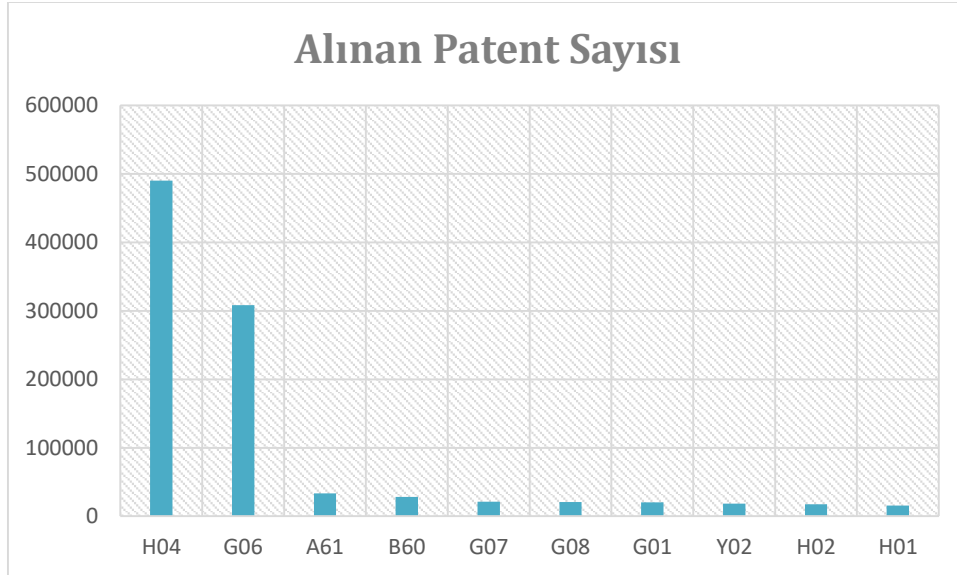
CPC	Tanım	Alınan Patent Sayısı	Sıralama
H04	Electric Communication Technique	490128	1
G06	Computing; Calculating; Counting	308290	2
A61	Medical or Veterinary Science; Hygiene	33308	3
B60	Vehicles in General	28348	4
G07	Checking-Devices	21140	5
G08	Signalling	20887	6
G01	Measuring; Testing	20413	7
Y02	Technologies or Applications for Mitigation or Adaption Against Climate Change	18414	8
H02	Generation; Conversion or Distribution of Electric Power	17669	9
H01	Basic Electric Elements	15823	10

Tablo 2-6 Veri Setinin CPC Alt Sınıflarına Göre Sıralaması

CPC	Tanım	Alınan Patent Sayısı	Sıralama
H04L	Transmission of Digital Information	306581	1
G06F	Electric Digital Data Processing	195895	2
H04W	Wireless Communication Networks	81370	3
G06Q	Data Processing Systems or Methods	76363	4
H04N	Pictorial Communication	72796	5
G06K	Recognition of Data	25931	6
G08B	Signalling or Calling Systems	15321	7
H04M	Telephonic Communication	13642	8
A61B	Diagnosis; Surgery; Identification	13137	9
G07C	Time or Attendance Registers	12492	10

Tablo 2-7 Veri Setinin CPC Alt Gruplarına Göre Sıralaması

CPC	Alınan Patent Sayısı	Sıralama
H04W12/06	12283	1
H04L63/08	9187	2
H04L63/0428	8421	3
H04L63/20	5761	4
G06F21/32	5714	5
G06F21/31	4957	6
H04L63/10	4897	7
H04L63/083	4895	8
H04W12/04	4781	9
H04L9/3247	4750	10



Şekil 2-8 Veri Setinin CPC Sınıflarına Göre İncelemesi

2.2.5 Veri Setinin IPC Kodlarına Göre İncelenmesi

Hazırlanan veri setinde yapılan incelemeler sonucunda IPC kodlarına göre gruplama işlemi bir ana ve üç alt sınıf olarak toplamda dört gruba ayrılmıştır. Grupların adları ve detaylarına ilişkin bilgiye Tablo 2-8 ve Şekil 2-9'dan erişilebilir.

Tablo 2-8 Veri Setinin IPC Sınıflarına Göre Sıralaması

IPC	Tanım	Alınan Patent Sayısı	Sıralama
H04	Electric Communication Technique	9923	1
G06	Computing; Calculating; Counting	6959	2
A61	Medical or Veterinary Science; Hygiene	1183	3
G01	Measuring; Testing	738	4
G08	Signalling	710	5
H02	Generation; Conversion or Distribution of Electric Power	703	6
G07	Checking-Devices	554	7
B60	Vehicles in General	522	8
E04	Building	495	9

Tablo 2-9 Veri Setinin IPC Alt Sınıflarına Göre Sıralaması

IPC	Tanım	Alınan Patent Sayısı	Sıralama
H04L	Transmission of Digital Information	5257	1
G06F	Electric Digital Data Processing	4230	2

(Tablo 2-9 Devam)

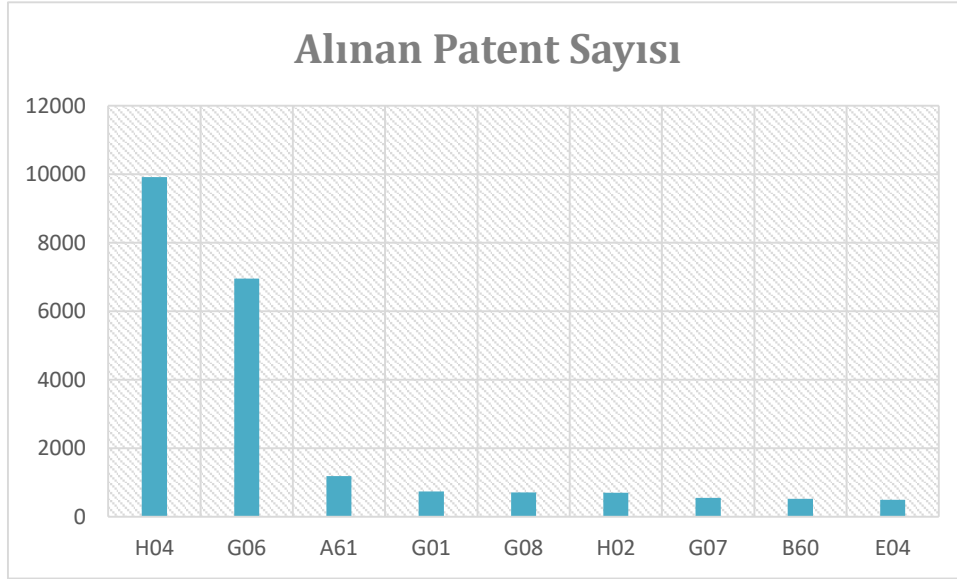
H04W	Wireless Communication Networks	2323	3
G06Q	Data Processing Systems or Methods	2051	4
H04N	Pictoral Communication	1616	5
G08B	Signalling or Calling Systems	551	6
G06K	Recognition of Data	536	7
C12N	Microorganisms or Enzymes	403	8
G10L	Engineering; Culture Media Speech Anaylsis or	399	9

Tablo 2-10 Veri Setinin IPC Alt Gruplarına Göre Sıralaması

IPC	Alınan Patent Sayısı	Sıralama
H04L29/06	38027	1
H04L9/32	21179	2
H04L9/08	10518	3
H04L29/08	10467	4
H04W12/06	8328	5
G06F17/30	8084	6

(Tablo 2-10 Devam)

G06F21/62	7134	7
G06F21/00	5852	8
G06F21/60	5404	9



Şekil 2-9 Patentlerin IPC Sınıflarına Göre Sıralaması

2.3 Kullanılan Yöntemler

Bu bölümde ağ analizi ve zaman serisi analizi konuları üzerinde durulmuştur. Ağ analizi nedir, bileşenleri nelerdir açıklanmasının ardından ağ ölçütlerinden bahsedilmiştir. Zaman serisi analizi kısmında zaman serisi tahminleme modellerinden regresyon, ayırıştırma üstel düzeltirme, ARIMA ve yapay sinir ağları modelleri anlatılmıştır. Bu modeller detaylı bir şekilde incelenmiştir.

2.3.1 Ağ Analizi

Bu bölümde graf nedir, bileşenleri nelerdir ve graf gösterim şekilleri nelerdir kavramlarının açıklaması yapılmıştır.

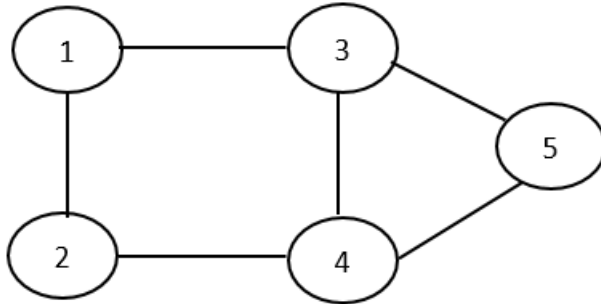
2.3.1.1 Graf Veri Yapısı ve Bileşenleri

Düğüm ve kenarlardan oluşan ağ yapılarına graf denir. Çizge ve çizit isimleriyle de bilinmektedir. Gerçek hayatta karşılaşılan yapıları bilgisayar dilinde temsil etmek amacıyla kullanılırlar. Örnek olarak iki şehir arasındaki en kısa mesafenin hesaplamak, sosyal medyada kimin kiminle arkadaş olduğunu göstermek, firmaların tedarik zincirlerini görüntülemek için graf yapılarından faydalanılabilir. Graf düğümleri ve bu düğümleri birbirine bağlayan kenarlardan oluşan ağ yapılarıdır. Matematikte graf düğüm ve kenarlardan oluşan iki küme ile ifade edilir [25].

Graf yapısının en temel nesnesi düğümdür (node). Köşe ya da vertex isimleriyle de anılırken sosyal ağlarda aktör ifadesi de düğüm için kullanılmaktadır. Graf yapısındaki düğümlerin sayısı grafın boyutunu vermektedir. Örneğin n tane düğüm içeren bir grafın boyutu n 'dir ve $|V| = n$ şeklinde gösterilir.

Düğümleri birbirlerine bağlayan bağlantılara kenar adı verilir. Graf yapısındaki ilişkileri gösterir.

$G = (D, K)$ Burada G grafı, D düğümlerin K ise kenarlar kümesini ifade etmektedir.



Şekil 2-10 Graf Örneği

Yukarıdaki şekilde gösterilen graf matematiksel olarak şu şekilde ifade edilir.

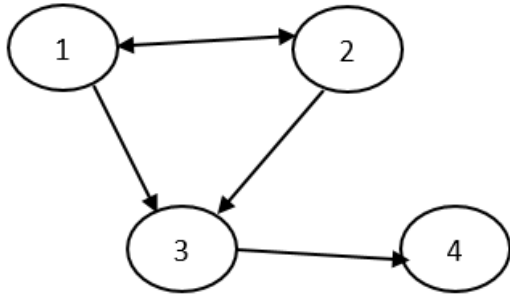
$V: \{1,2,3,4,5\}$

$E: \{\{1,2\}, \{1,3\}, \{2,4\}, \{3,4\}, \{3,5\}\}$

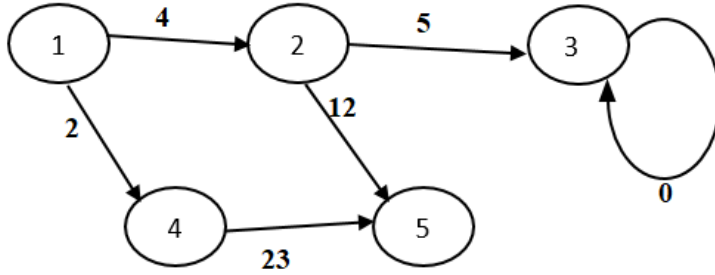
$G = (V, E)$

Grafları kenarlarının yönlü olup olmamasına göre Yönlü Graflar ve Yönsüz Graflar olarak 2'ye ayırabiliriz. Bir graftaki kenarların yön belirtmemesi durumunda bu grafa

yönsüz graf denir. Şekil 2-5'te gösterilen graf yönsüz grafa örnektir. Yönsüz grafa örnek olarak Facebook verilebilir çünkü Facebook'ta bir insanla arkadaş olmak iki insanın da birbirleriyle arkadaş olması demektir. Her iki insan da birbirlerinin gönderilerini görebilirler. Bir grafın kenarlarının yön belirtmesi durumunda bu grafa yönlü graf denir. Burada yön bir düğümden diğerine giden yolu göstermektedir. Yönlü grafın en iyi örneklerinden biri Instagram olarak gösterilebilir. Instagram'da bir kişiyi takip edebilirsiniz ancak bu o kişinin sizi takip ettiği anlamına gelmez. Takip edenin karşıdaki kişi ile ilişkisi olmasına rağmen onun takip edenle bir ilişkisi yoktur.



Şekil 2-11 Yönlü Graf Örneği



Şekil 2-12 Yönlü Ağırlıklı Graf Örneği

Bir grafın yapısındaki kenarların değerleri eşit değilse, her biri farklı bir değer alabiliyorsa bu graflara ağırlıklı ya da maliyetli graf adı verilir. Yol haritalarını temsil eden graflar ağırlıklı graflara örnek olarak verilebilir. Bir graftaki tüm kenarlara ait ağırlıkların toplamı o grafın ağırlığını yani maliyetini vermektedir. Matematikte ağırlıklı graflar $G = (V, E, W)$ şeklinde gösterilir. W E 'deki her bir kenara ait ağırlık değerinin kümesidir.

2.3.1.2 Ağ Ölçütleri

Bir ağdaki düğümlerin önemini tanımlamak için ağ analiz setinden farklı ölçütler kullanılmaktadır.

Bir düğümün temel ölçütü derecedir. Derece yönsüz ağlarda bir düğüme bağlı kenarların sayısı iken yönlü ağlarda iç ve dış derecelerin toplamı o düğümün toplam derecesini verir. İç derece düğüme gelen kenarların sayısını, dış derece ise düğümden çıkan kenarların sayısını gösterir.

Merkezlilik bir ağın gücünün göstergesidir. Ağ analizinde ölçüm yaparken “Arasındalık”, “Yakınlık” ve “Derece” ölçütlerinden faydalanılır.

Derece merkeziliği düğümlerin derecelerine bakarak önemini ortaya koyan ölçüttür. Bir düğümün derecesi ne kadar yüksek ise o düğüm o kadar önemlidir.

Yakınlık merkeziliği bir düğümden ulaşılabilen tüm düğümlerin ortalama en kısa mesafesidir.

Arasındalık merkeziliği bir düğümün ağdaki diğer düğümler arasında ne kadar uzaklıkta olduğunu belirtir [26].

Kümeleme Katsayısı bir ağdaki düğümlerin kümelenme derecesinin ölçüsüdür. Sıfır ve bir arasında değerler alır.

Yoğunluk sıkça kullanılan ağ ölçütlerindendir. Bir ağda tüm düğümler birbiriyle ilişki içerisinde olmayacaktır. Ağın yoğunluğu bize sosyal ağın “kapsayıcılığı” verir. Ağ büyüdükçe yoğunluk azalır çünkü büyük ağlarda düğümlerin ilişkili olma durumları daha zordur. Bu yüzden farklı büyüklükteki ağların yoğunlukları karşılaştırılmaz [27].

2.3.2 Zaman Serisi Analizi

Bir zaman serisi, zaman içinde sırayla alınan bir gözlem dizisidir. Bir zaman serisinin içsel bir özelliği, tipik olarak, bitişik gözlemlerin bağımlı olmasıdır. Zaman serisi analizi, bu bağımlılığın analizi için teknikler ile ilgilidir [28].

Zaman serisi analizi, seçilen değişkenlerin zaman içindeki değişiklikleri tanımlar, açıklar, tahmin eder ve kontrol eder. Eldeki verilerle zaman serisi analizi yapılabilmesi için verilerin düzenli şekilde toplanması gereklidir. Verilerden çıkarılan anlamlı sonuçlar istatistiksel yöntemler kullanılarak elde edilir.

Zaman serisi analizini kullanmanın en büyük avantajı geçmiş davranışları anlayarak gelecek hakkında tahmin yapabilmesidir [29]. Günümüzde hisse senedi fiyatları, sanayi tahminleri ve mevsimsel değişiklikler gibi pek çok alanda zaman serisi analizi kullanılmaktadır. Zaman serisi analizinin kullanılmasının amacı geçmişi anlamak ve geleceği tahmin etmektir.

2.3.2.1 Zaman Serisi Bileşenleri Nelerdir?

Zaman serilerinde tahmin algoritmalarını seçebilmek amacıyla bileşenlerin sistematik ve sistematik olmayan bileşenler olarak ikiye ayrılır. Sistematik bileşenler tanımlanabilir, tutarlı tekrarlanabilir bileşenler iken, sistematik olmayan bileşenler doğrudan modellenemeyen bileşenlerdir. Sistematik bileşenler trend, mevsimsellik iken gürültü sistematik olmayan bileşen olarak adlandırılmaktadır [30].

- **Trend:** Zamana göre gözlemlenen bir değişkenin uzun dönemde gösterdiği artış veya azalışa denir. Doğrusal ve doğrusal olmayan trend olarak ikiye ayrılır [31]. Zaman serisinin eğilimidir.
- **Mevsimsellik:** Zaman serilerinin belirli aralıklarla düzenli bir seyir izlediği görülmektedir. Düzenli aralıklarla meydana gelen değişikliklere “mevsimsellik” ya da “mevsimsel değişme” adı verilir [32].
- **Gürültü:** Zaman serilerinde tahmin edilemeyen değişimlere verilen addır.
- **Kalıntı:** Gözlemler ile karşılık gelen uygun görülen değerler arasındaki farktır [33].

2.3.2.2 Zaman Serilerinde Regresyon Modelleri

Regresyon analizi bağımlı değişkenin, bir ya da birden fazla bağımsız değişken arasındaki ilişkinin matematiksel bir fonksiyon şeklinde yazılmasıdır. Regresyon denklemi bu fonksiyonun adıdır [34].

Regresyon modelleri, bir zaman serisini öngörücü olarak kullanıp farklı bir zaman serisinin tahminlenmesidir. Örneğin reklam harcamalarından yola çıkarak aylık satışların tahmin edilmesi gibi.

Doğrusal regresyon modeli, bağımlı değişken olan Y’nin tek bir bağımsız değişken olan X ile aralarındaki ilişkiyi doğrusal fonksiyonla ifade eden modeldir. Formülü şu şekildedir;

$$y = a + bx + e \quad (2.1)$$

Yukarıdaki denklemdeki a fonksiyonun sabiti olup, b doğrunun eğimini gösterir.

Doğrusal regresyon modelinde a ve b katsayılarını bulmak için en küçük kareler yöntemi ve maksimum benzerlik yöntemleri kullanılabilir [35]. En küçük kareler yöntemi katsayıları bulmak için sıklıkla başvurulanan yöntemlerdendir. Bu yöntemin amacı minimum hatayı veren denklemi bulmaktır. Maksimum benzerlik yöntemi ise parametreler arasından gözlemlenen örneği elde etme olasılığı büyük olan değerleri seçer.

Doğrusal olmayan regresyon modeli, bağımlı değişken olan Y'nin bağımsız değişken X ile arasındaki ilişkinin doğrusal olmaması durumudur. Bu model ikinci derece veya daha üst derecelerden fonksiyon ile gösterilir.

$$y = a + bx + cx^2 + e \quad \text{2. Derece Fonksiyon} \quad (2.2)$$

$$y = a + bx + cx^2 + dx^3 + e \quad \text{3. Derece Fonksiyon} \quad (2.3)$$

Yukarıdaki denklemde a sabit sayıyı, b fonksiyonun eğimini, c eğimin değişme hızını, d ise bükümün derecesindeki değişme oranını gösterir [34].

Kübik regresyon modeli, bağımlı değişken ile bağımsız değişkenler arasındaki ilişkinin polinom şeklinde olduğu 3. dereceden bir regresyon modelidir.

2.3.2.3 Zaman Serilerinde Ayırıştırma Yöntemleri

Zaman serilerinde bileşenlerin tahmininden serinin öngörüsünü hesaplayan yonteme ayırıştırma yöntemi denir. Serinin özelliklerini ortaya koymak için kullanılan başlıca yöntemlerden olup serinin bileşenlerine ayrılması temeline dayanan en eski yöntemlerdendir.

STL (Seasonal and Trend decomposition using Loess) bir zaman serisini trend, mevsimsellik ve kalan bileşenlere ayırmak için kullanılan filtreleme prosedürüdür. STL yöntemiyle çok uzun zaman serileri hızlıca ayrıştırılabilir [36].

Diğer ayırıştırma yöntemleri olan klasik, SEATS ve X11 ayırıştırma yöntemlerine göre birçok avantajı vardır:

- Yalnızca aylık ve üç aylık verileri değil her türlü mevsimselliği işleyebilir.
- Mevsimsellik bileşeninin zaman içinde değişmesine izin verir.

- Trend döngüsü kullanıcı tarafından kontrol edilebilir.
- Zaman zaman gözlemlenen aykırı değerler trend döngüsünü ve mevsimselliği etkilememektedir.

Bunların yanı sıra STL'nin işlem gününü veya takvim değişikliklerini otomatik olarak işlememesi ve sadece ek ayrıştırmalar için olanak sağlaması bu yöntemin dezavantajı olarak sayılır [37].

2.3.2.4 Zaman Serilerinde Üstel Düzleştirme Yöntemleri

Üstel düzleştirme yöntemleri zaman serilerinde tahmin için kullanılan en başarılı yöntemlerdendir ve geçmiş gözlemlerin ağırlıklı ortalamasına eşittir. Ağırlıklar gözlemler eskidikçe azalır. Başka bir deyişle gözlem ne kadar yeni ise o gözlemin tahmine etkisi o kadar yüksektir. Bu çalışma prensibiyle üstel düzleştirme yöntemleri hızlı bir şekilde çok çeşitli zaman serileri için güvenilir tahminler üretirler. Bu yönü üstel düzleştirmenin en önemli avantajlarından olup, endüstrideki uygulamalar için büyük önem taşımaktadır [38].

Basit üstel düzleştirme yöntemi; trende ve mevsimsel dalgalanmaya sahip olmayan sadece bir ortalama düzey etrafında hareket eden serilerin analizinde uygulanmaktadır.

$$\text{Formülü şu şekildedir; } \hat{Z}_{T+1} = S_T = \alpha Z_T + (1 - \alpha)S_{T-1} \quad (2.4)$$

Formüldeki S_0 başlangıç değeri için serinin ortalamasıdır, α tahmin hatasını minimum yapan sabit bir değerdir [39].

Holt doğrusal trend yöntemi; trende sahip mevsimsel dalgalanması olmayan serilerin tahmininde holt doğrusal trend yöntemi kullanılmaktadır. Formülü şu şekildedir;

$$\hat{Z}_{T+F} = a_T + b_T(F) \quad (2.5)$$

$$a_T = \alpha Z_T + (1 - \alpha)(a_{T-1} + b_{T-1}) \quad (2.6)$$

$$b_T = \gamma(a_T - a_{T-1}) + (1 - \gamma)b_{T-1} \quad (2.7)$$

Yukarıdaki formülde α ortalama düzeyin düzleştirme katsayısını, γ eğimin düzleştirme katsayısını, F öngörü yapılacak dönem sayısını göstermektedir [39].

Holt-winters mevsimsel yöntemi; trend ve mevsimsel dalgalanmalara sahip serilerin tahmininde holt-winters mevsimsel yöntemi kullanılır. Bu yöntem önce serinin eğimine, ortalama düzeyine, sonrasında mevsimsel bileşenine uygulanmaktadır. Formülü şu

şekildedir;

$$a_T = \alpha (Z_T - M_T(T - s)) + (1 - \alpha)(a_{T-1} + b_{T-1}) \quad (2.8)$$

Formülde a_T T dönemindeki yeni düzleştirme yöntemi tahminini, α ortalama düzeyin düzleştirme katsayısını, $Z_T - M_T(T - s)$ T dönemindeki mevsimsellikten arınmış veriyi, a_{T-1} T-1 dönemindeki eski düzleştirme tahminini ve b_{T-1} T-1 dönemindeki eğimin eski düzleştirme tahminini gösterir [39].

2.3.2.5 Zaman Serilerinde ARIMA Modelleri

ARIMA modelleri durağan hale getirilebilen bir zaman serisini tahmin etmek için en genel model sınıfıdır. Durağan zaman serilerinde istatistiksel özelliklerinin tümü zaman içinde sabittir. Durağan serinin eğilimi yoktur. ARIMA(AutoRegressive Integrated Moving Average) otoregresif bütünleşik hareketli ortalamanın kısaltmasıdır. ARIMA modellemenin tahmin denklemindeki durağanlaştırılmış serinin gecikmeleri otoregresif terimler olarak adlandırılır [40].

Farklılaşmayı otoregresyon ve hareketli ortalama modeli ile birleştirirsek, mevsimsel olmayan bir ARIMA modeli elde edilir [41]. Mevsimsel olmayan bir ARIMA modeli, “ARIMA (p, d, q)” modeli olarak sınıflandırılır, burada:

p, otoregresif terimlerin sayısıdır,
d, durağanlık için gereken mevsimsel olmayan farklılıkların sayısıdır ve
q, tahmin denklemindeki gecikmeli tahmin hatalarının sayısıdır. Mevsimsel olmayan ARIMA (p, d, q) modelinin tahmin denklemi şu şekildedir;

$$\hat{y}_t = \mu + \phi_1 y_{t-1} + \dots + \phi_p y_{t-p} - \theta_1 e_{t-1} - \dots - \theta_q e_{t-q} \quad (2.9)$$

Denklemden $\hat{y}_t$ farklılaştırılmış seriyi, sağ taraftaki parametreler gecikmeli değerleri ve hataları temsil etmektedir [40].

Mevsimsel ARIMA modelleri, mevsimsel olmayan ARIMA modellerine mevsimsel terimler eklenerek oluşturulmaktadır. Bu modeller “ARIMA (p, d, q) (P, D, Q)_m” olarak gösterilir. Burada m yıllık gözlem sayısını ifade etmektedir. Küçük harfle gösterilen kısım mevsimsel olmayan bölümü, büyük harfle gösterilen kısım ise mevsimsel olan bölümü ifade etmektedir. Mevsimsel ARIMA modellerinde modelleme prosedürü mevsimsel olmayan verilerle hemen hemen aynıdır [42].

2.3.2.6 Zaman Serilerinde Yapay Sinir Ağları Modelleri

Yapay sinir ağları beynin temel öğrenme işlevini gerçekleştiren modellere dayanan tahmin yöntemidir. Eksik veriyle çalışabilme, bulunduğu ortama adapte olabilme belirsizlik durumunda tahmin yapabilme gibi pek çok avantajı vardır. Ayrıca hata toleransı olduğu için birçok alanda başarılı bir şekilde uygulanmaktadır. Optimizasyon, örüntü tanıma, sınıflandırma gibi alanlarda kullanılmaktadır [43].

Bir sinir ağı katmanlar halinde düzenlenmiş bir “nöron” ağı olarak düşünülebilir. Nöronlar yapay sinir ağına gelen bilgiyi işleyen birimlerdir. Girdiler alt katmanı, çıktılar yani tahminler ise üst katmanı oluşturmaktadır. Aradaki diğer katmanlar gizli katmandır. Doğru tahmin yapılabilmesi için ağı eğitilmesi gerekir. Eğitim işleminden sonra ağı test veri seti gösterilir [44] eğitim sırasındaki amaç bulunması gereken en doğru tahmini yapabilmektedir. Çıktı ile hedef değerler arasındaki fark hatadır. Ağ bu hatayı kabul edilebilir sınırlara indirebilmek için sürekli olarak işlemi tekrarlar. Bu yüzden yapay sinir ağlarında ağı öğrenmesinde yani eğitilmesinde kullanılacak yöntem çok önemlidir [45].

3. BULGULAR VE TARTIŞMA

Bu bölümde, çalışma kapsamında kullanılan yöntemler ve elde edilen sonuçlar gösterilmektedir.

3.1 Ağ Analizi – Patentlerin Görsel Analizi

Bu çalışmada Gephi ve Cytoscape yazılımları yardımıyla ağın öncelikle yıl bazında görsel analizi ve devamında tüm yılları kapsayan bir görsel analiz yapıp ağda yapısal desenler aranmıştır.

Patent verilerinin ülke ve IPC sınıflarına göre dağılımlarının görsel analizi Gephi ile yapılmıştır. Görselleştirme yaparken kenarların üst üste binmesini engelleyen ve dengeli dağıtımı sağlayan Fruchterman-Reingold yerleşimi seçilmiştir. Ülke adları ve IPC değerleri düğüm etiketi olarak eklenmiştir. Ülkelerin düğüm etiketleri için ortak bir renk seçilirken, IPC değerlerinin düğüm etiketleri için de ayrı ortak bir renk seçilmiştir. Bu renklendirme ülke ve IPC değerlerini grafikte kategorize etmek için yapılmıştır.

3.1.1 Yıl Bazında Ağ Analizi

Bu bölümde 2009'dan 2019 yılına kadar bilgi güvenliği alanında alınan patentlerin ağ analiz sonuçlarına yer verilmiştir. Her yıl için Gephi ile yapılan görsel analizde ülkeler ve onların yöneldiği IPC sınıfları gözlemlenebilmektedir. 2009'dan 2019'a kadar yapılan görselleştirme çalışmaları Fruchterman-Reingold yerleşimi üzerinde yapılmıştır. Şekil 3-1'de yıl bazlı görsel analizler mevcuttur.

Ağlarda bulunan düğümlerin önemini bulmak için Gephi kullanılarak merkezilik analizi yapılmıştır. Bir ağdaki bir düğümün önemi o düğüme ait merkezilik değeri ile ölçülür. Merkezilik analizi Gephi'de derece merkezilik ölçütüyle yapılmıştır. Şekil 3-1'de bulunan ağların merkezilik ölçütlerine ait değerler Tablo 3-1'de verilmiştir. Merkezilik değerlerinde ağların maksimum iç derece ve dış dereceye ait düğümleri, ortalama derece ve ortalama ağırlıklı derece değerleri verilmiştir.

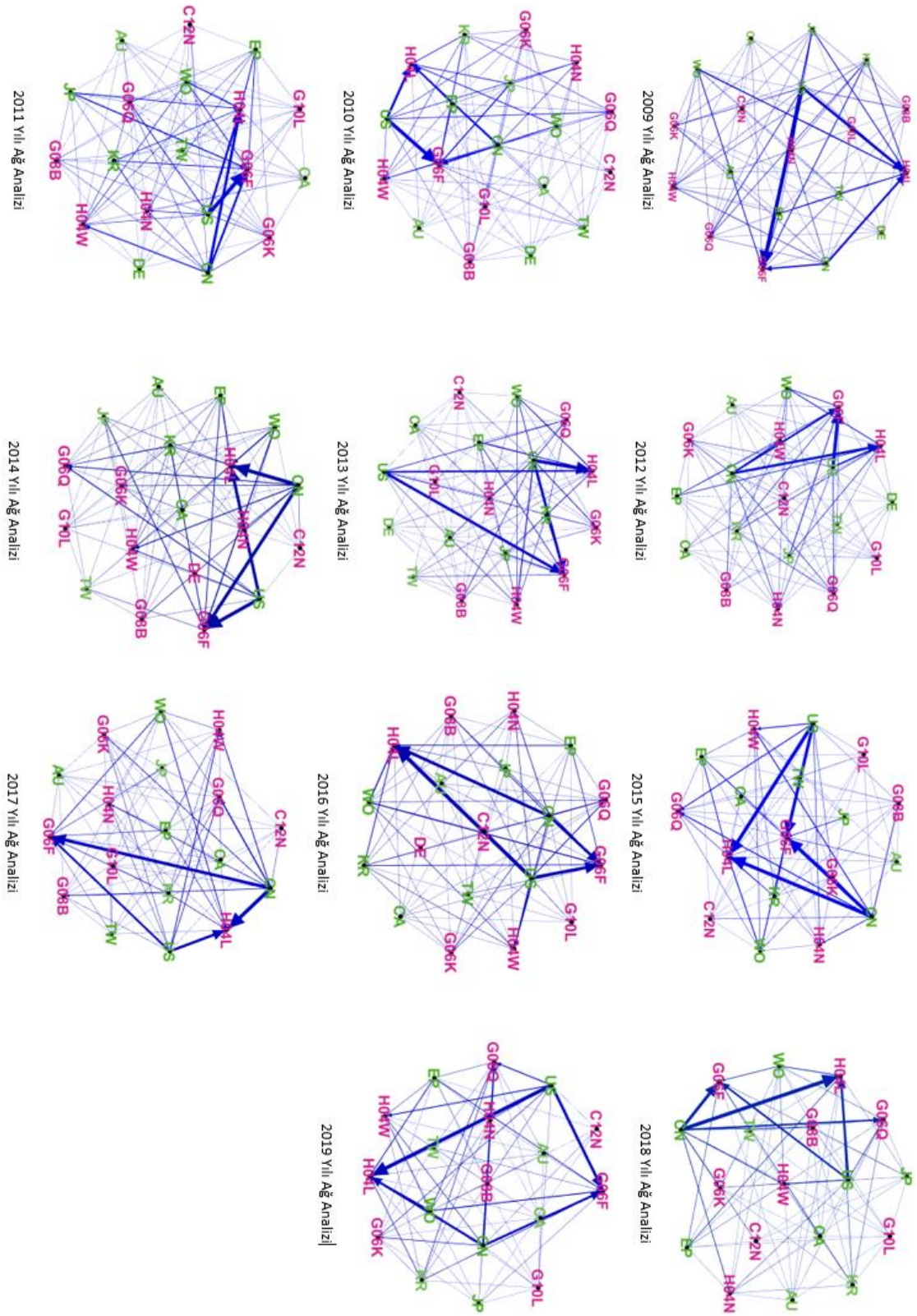
Şekil 3-1'de bulunan ağların düğüm sayısı, kenar sayısı, kümelenme kat sayısı, karakteristik yol uzunluğu, bağlı komponentler, çap, yarıçap, ortalama komşu sayısı gibi ölçümleri Cytoscape yazılımı ile elde edilmiştir. Tablo 3-1 ağların istatistik verilerini göstermektedir.

Tablo 3-1 incelendiğinde hangi yıl hangi ülkenin bilgi güvenliği alanında aldığı patentlerle önde olduğu ve hangi yılda hangi sınıftan daha fazla patent aldığı görülmektedir. 2009’dan 2019’a kadar ülkeleri incelediğimizde ABD ve Çin’in en fazla patent alan ülkeler olduğu görülmektedir. Ayrıca ülkelerin bilgi güvenliği alanında aldığı patentlerin sayısı yıl geçtikçe artmaktadır. Ağırlıklı dış derece düğüm değerlerine baktığımızda Çin’in ABD’den açık ara önde olduğu görülmektedir. Alınan patentleri sınıf bazında incelediğimizde de G06F ve H04L sınıflarının en çok tercih edilen sınıflar olduğu görülmektedir. İlk yıllarda eğilim G06F sınıfında iken son yıllarda H04L sınıfına eğilim artmaktadır. Son yıllarda alınan patent sayısına orantılı olarak bilgi güvenliği ile ilgili olarak H04L sınıfından G06F sınıfına göre daha fazla patent alınmıştır.



Tablo 3-1 Yıl Bazında Ağların İstatistik Verileri

	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019
Düğüm Sayısı	19	19	19	19	19	18	18	18	18	18	18
Kenar Sayısı	88	86	88	85	88	79	80	80	79	78	77
Kümeleme Katsayısı	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Karakteristik Yol Uzunluğu	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
Yoğunluk	0.257	0.251	0.257	0.249	0.257	0.231	0.234	0.234	0.231	0.228	0.225
Bağlı Komponentler	1	1	1	1	1	1	1	1	1	1	1
Çap	1	1	1	1	1	1	1	1	1	1	1
Yarıçap	1	1	1	1	1	1	1	1	1	1	1
Ortalama Komşu Sayısı	9.263	9.053	9.263	8.947	9.263	8.778	8.889	8.889	8.778	8.667	8.556
Maksimum Dış Derece-İlgili Düğüm	5374-US	5572-US	4884-US	5174-CN	6643-CN	7067-CN	8868-CN	8399-US	14507-CN	18350-CN	4420-US
Maksimum İç Derece-İlgili Düğüm	5718-G06F	5848-G06F	5586-G06F	5857-G06F	5942-G06F	7082-G06F	7890-G06F	8551-H04L	11209-H04L	12676-H04L	4050-H04L
Ortalama Derece	4.632	4.562	4.632	4.474	4.632	4.158	4.211	4.211	4.158	4.105	4.053
Ortalama Ağırlıklı Derece	878	884.684	877.632	886.789	960.789	1093.789	1274.895	1350.158	1776.632	2006.842	652.158

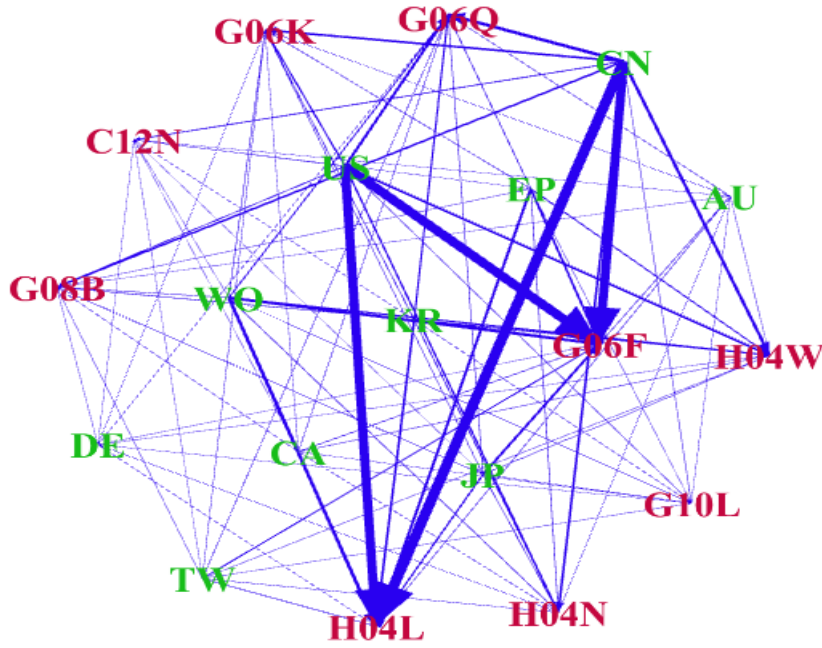


Şekil 3-1 Yıl Bazında Ağ Analizi

3.1.2 Son 10 Yılın Verileriyle Ağ Analizi

Bu bölümde veri setinin tamamını kullanarak oluşturulan ağ analiz sonuçlarına yer verilmiştir. Görsel analiz Gephi yazılımı ile yapılmıştır ve Gephi üzerindeki Fruchterman-Reingold yerleşimi kullanılmıştır. Şekil 3-2’de oluşturulan ağ ve Tablo 3-2’de ağa ait istatistik verileri gösterilmektedir. Ağın merkezilik analizleri için de Gephi kullanılmıştır. Ağın istatistik verilerini hesaplatmak için Cytoscape yazılımı kullanılmıştır. Tablo 3-2’de ayrıca ağın maksimum iç ve dış dereceye ait düğümleri, ortalama derece değeri ve ortalama ağırlıklı derece değerleri mevcuttur.

Tablo 3-2 incelendiğinde Çin’in en fazla patent alan ülke olduğu ve H04L sınıfının en çok tercih edilen sınıf olduğu görülmektedir.



Şekil 3-2 Veri Seti Ağ Analizi

Tablo 3-2 Tüm Veriler İstatistik Verileri

	01.2009-05.2019
Düğüm Sayısı	19
Kenar Sayısı	90
Kümeleme Katsayısı	0.0
Karakteristik Yol Uzunluğu	1.0
Yoğunluk	0.263
Bağlı Komponentler	1
Çap	1
Yarıçap	1
Ortalama Komşu Sayısı	9.474
Maksimum Dış Derece-İlgili Düğüm	83628-CN
Maksimum İç Derece-İlgili Düğüm	76923-H04L
Ortalama Derece	4.737
Ortalama Ağırlıklı Derece	12642.368

3.2 Zaman Serisi Analizi – Patent Sınıflarına Ait Patent Sayılarının Mevsimsel Tahminlemesi

Zamana göre sıralanan veri dizilerine zaman serisi denir. Diğer bir ifadeyle, bir zaman serisi, zaman içinden alınan ardışık bir dizi gözlemdir [46]. Verilerin matematiksel modellerinin oluşturulup analiz edilmesi işlemine de zaman serisi analizi denir. Zaman serisi analizine geçmiş verilere bakarak geleceğe dönük tahminler yapmak için başvurulur.

Bu tez çalışmasında bilgi güvenliği alanında ilerleyen teknolojinin yönünü tahmin edebilmek amacıyla patent sınıflarının patent sayılarına zaman serisi analizi yapılmıştır. Patent sınıfları ve sınıflara ait patent sayıları olarak tezin ilk aşamasında toplanan verilerden elde edilmiştir. Veri setinde 2019 yılına ait veriler tam olmadığı için dahil edilmemiştir. 2009-2018 yıllarına ait sınıf bazında elde edilen patent sayılarına Tablo 3-3'ten ulaşılabilir. Patent sınıflarına ait patent sayılarının zaman serisinin analizi ve tahminleme modellerinin çıkarıldığı çalışma, R programlama üzerinde gerçekleştirilmiştir.

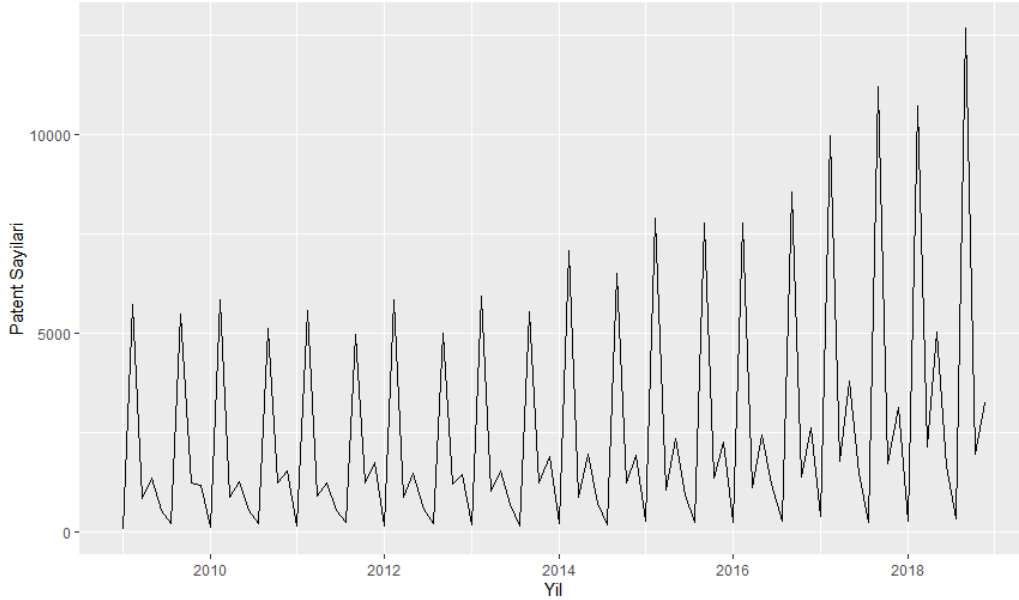
Tablo 3-3 Patent Sınıfları ve Sayıları Zaman Serisi

YIL	Patent Sınıflarına ait Patent Sayıları								
	C12N	G06F	G06K	G06Q	G08B	G10L	H04L	H04N	H04W
2009	70	5718	849	1353	557	211	5492	1241	1191
2010	142	5848	874	1263	541	212	5139	1255	1535
2011	156	5586	907	1256	540	251	4990	1243	1746
2012	166	5857	894	1472	575	230	4996	1214	1445
2013	180	5942	1046	1549	700	153	5541	1231	1913
2014	222	7082	897	1959	746	197	6519	1229	1931
2015	291	7890	1075	2372	947	249	7760	1378	2261
2016	250	7759	1138	2440	1223	278	8551	1393	2621
2017	398	9973	1788	3798	1472	261	11209	1712	3145
2018	293	10723	2149	5034	1669	345	12676	1967	3274

R programlamada ilk olarak yüklenmesi gereken kütüphaneler indirilmiş ve “supply” fonksiyonu ile kullanıma hazır hale getirilmiştir. Daha sonra veri olarak Tablo 3-3’de

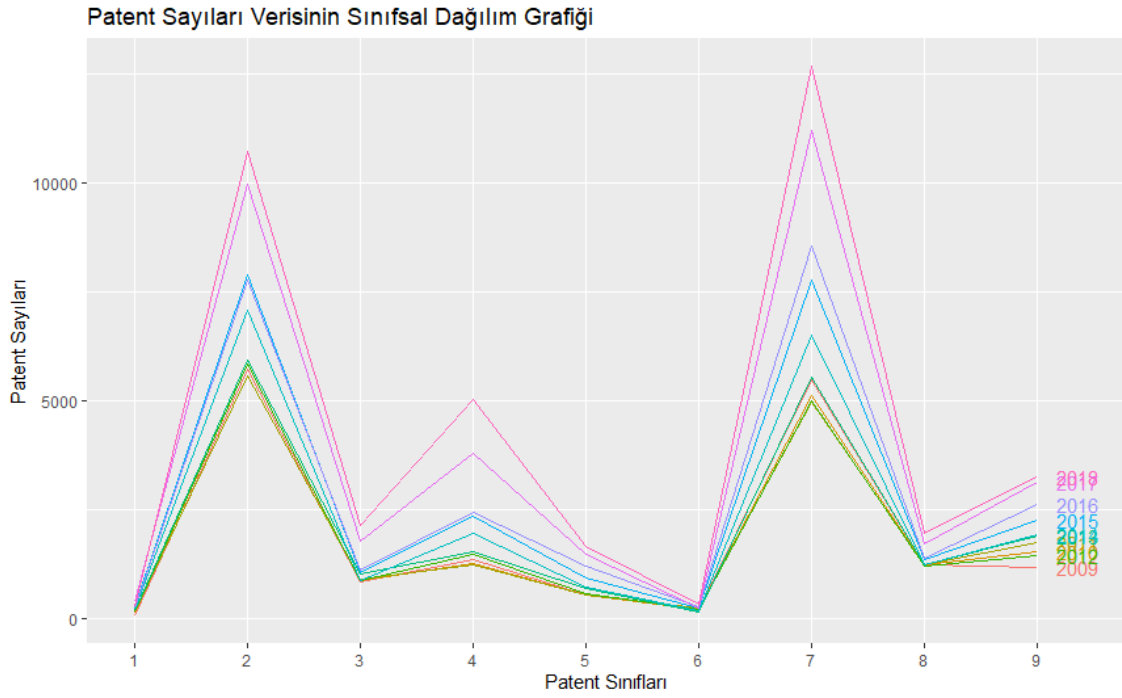
gösterilen veriler okunmuş ve “ts” fonksiyonuyla zaman serisi nesnesine dönüştürülmüştür. Ts fonksiyonunu çağırırken veri setinin başlangıç yılı olan 2009 seçilmiş ve verilerde dokuz patent sınıfı olduğu için zaman serisinin frekansı dokuz olarak seçilmiştir. Frekans birim zaman başına düşen gözlem sayısıdır. Aylık olarak atandığında 12, günlük olarak atandığında yedi, mevsimsel olarak atandığında dört ve yıllık olarak atandığında bir yazılır. Burada birim yıl ve yıl başına düşen gözlem sayısı sınıf sayısına eşit olduğu için frekans değeri bu şekilde seçilmiştir.

Veri setinin R programlamada gidişatını görebilmek için “autoplot” fonksiyonuyla grafik çizdirilmiştir. Elde edilen grafik Şekil 3-3’te gösterilmekte olup çizgi grafiğidir. Grafiğin X koordinatı zamanı, Y koordinatı ise patent sayılarını göstermektedir. Grafik incelendiğinde yukarı doğru bir eğilim olduğu gözlemlenmektedir. Bu durum zaman serisinin artan ve doğrusal olmayan bir trendi olduğunu göstermektedir. Patent sayılarının zaman serisi grafiğindeki dalgalanmalar sınıfsal farklardan kaynaklı olabilir fakat farklı etkenlerden de kaynaklı olabilir. Değişimlerin sınıfsal olduğunu düşünüp incelenmek istenildiğinde R programlama üzerinden hazır fonksiyonlar kullanılıp görseller çizilebilir.



Şekil 3-3 Patent Sınıflarına ait Patent Sayıları Zaman Serisi Grafiği

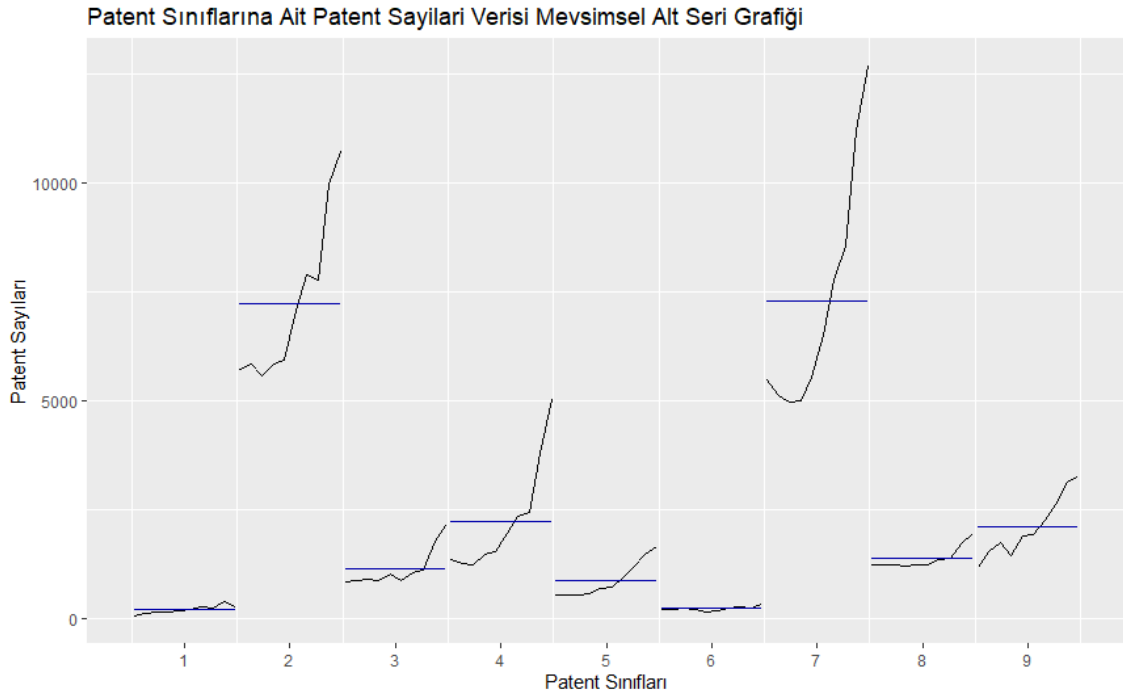
Zaman serisinin patent sınıf değerlerinin sınıfsal grafiği R programlamadaki mevsimsellik grafiği çizdirme fonksiyonu “ggseasonplot” fonksiyonu ile yapılmıştır. Bu fonksiyon ile dokuz ayrı sınıfa ait hareketler detaylı bir şekilde görülebilmektedir.



Şekil 3-4 Patent Sınıfları ve Sınıflara Ait Patent Sayıları Zaman Serisine Uygulanan "ggseasonplot" Fonksiyonu Grafiği

Şekil 3-4'teki grafik incelendiğinde her yılın benzer mevsimsel desene sahip olduğu görülür. Bu grafikte iki, dört ve yedi numaralı patent sınıfları; grafiğin en yüksek üç tepe noktası olup yıl bazında en çok patent alınan sınıflardır. Grafiğin X eksenindeki iki rakamı "G06F" sınıfını temsil etmekte iken dört rakamı "G06Q" sınıfını, yedi rakamı ise "H04L" sınıfını göstermektedir. Zaman serisinin mevsimselliğine başka bir mevsimsellik grafiği kullanarak da bakılabilir. "ggsubseriesplot" fonksiyonuyla serinin yıllık çeyreklerinin patent sayıları üzerindeki seyri çizilmiştir. Bu grafik bu veri setinde açık bir şekilde sınıfsal hareketin olduğunu göstermektedir. Bu grafik aynı zamanda bu çalışmadaki patent sınıflarına ait hareketlerin mevsimsel fonksiyonlar yardımıyla gözlemlenebileceğini göstermiştir.

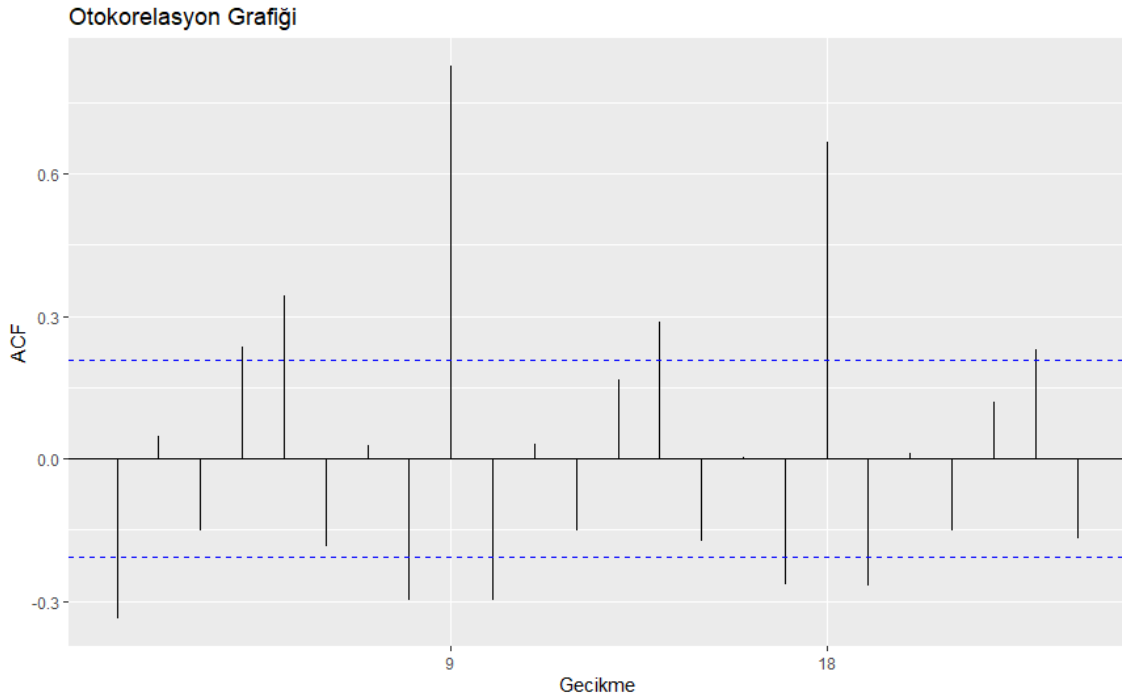
Patent sınıflarına ait değişim R programlamadaki farklı bir mevsimsellik grafiği yardımıyla da izlenebilir. Bu grafik mevsimsel alt serileri göstermek için kullanılırken burada sınıfsal alt seriler için kullanılmıştır ve "ggsubseriesplot" fonksiyonuyla grafik çizdirilmiştir.



Şekil 3-5 Patent Sınıflarına Ait Patent Sayıları Zaman Serisine Uygulanan "ggsubseriesplot" Fonksiyonu Grafiği

Şekil 3-5'teki grafik incelendiğinde çeyreklik periyotlardaki ortalama değerlerin mavi çizgiler ile gösterildiği görülmektedir. Bu grafik yıl içerisindeki her sınıf bazında mevsimsel paterni verdiği görülmektedir. Mevsimsellik bulunan zaman serilerinde ortalama çizgiler birbirinden uzaktır. Bu durumda bu seride mevsimsellik yani verilerin sınıflara göre dağılımı söz konusudur denilebilir. Hemen hemen her sınıf için zaman içerisinde artış söz konusudur.

Otokorelasyon zaman serisiyle bu serinin gecikmeli serileri arasındaki ilişkidir. R programlamada zaman serisinin otokorelasyon grafiği "ggAcf" fonksiyonuyla çizilmiştir. Otokorelasyon grafiği gecikme sayılarına karşılık otokorelasyon değerlerinin yer aldığı grafikdir. Korelogram olarak da adlandırılabilir.



Şekil 3-6 Patent Sınıflarına ait Patent Sayıları Zaman Serisinin Otokorelasyon Grafiği

Şekil 3-6'daki grafikte x eksenini gecikmeleri, y eksenini ise otokorelasyon değerlerini, mavi çizgi ise sınır değerlerini gösterir. R programlamada grafik çizdirirken geriye dönük 27 gecikme kullanması istediğimiz için 27 önceki değere kadar otokorelasyon hesaplaması yaptırılmıştır. Bu grafikte sınır değerlerinin üzerinde değerler olduğu için zaman serisinin durağan olmadığını söyleyebiliriz. Mevsimsel periyotlardaki yani dokuz ve dokuzun katsayısı döngülerde sınıfsal periyotlar görülmektedir ve o periyotlardaki otokorelasyon katsayılarının yüksek olması mevsimselliğin net olarak görülmesinden kaynaklıdır. Aynı zamanda o zamandaki gözlemler büyüklük olarak birbirine yakındır.

Tahminleme modeline geçmeden önce veri setinin trend, mevsimsellik ve gecikmelerle ilişkileri incelenmiştir. Son olarak veri üzerinde dönüşümün gerekli olup olmadığını incelenmiş olup verideki varyansı durağan hale getirerek daha doğru tahminler yapmak amaçlanmıştır.

Zaman serisi analizinde veriye uygulanan dönüşümlerin ve düzeltmelerin amacı doğru tahminlere ulaşmak için patternleri sadeleştirmektir. Şekil 3-3'deki zaman serisi grafiğine bakarsak verinin değişkenliğinin değiştiği görülmektedir. Bu değişkenliklerden dolayı pattern karmaşılaşmakta ve tahmin zorlaşmaktadır. Bu sebepten ötürü veri

sadeleştirilmeye çalışılacaktır. Dönüştürme işlemi için R programlamadaki “BoxCox” fonksiyonuyla box-cox dönüşümü uygulanmıştır. Kısaca box-cox dönüşümü “lambda” parametresine bağlıdır ve şu şekildedir;

$$\omega_t = \begin{cases} \log(y_t) & \lambda = 0 \text{ ise} \\ \frac{y_t^\lambda - 1}{\lambda} & \lambda \neq 0 \text{ ise} \end{cases} \quad (3.1)$$

Yukarıdaki formülde görülebileceği gibi “BoxCox” fonksiyonu “lambda” sıfır iken seriye “log” dönüşümü yapmaktadır. Burada önemli olan doğru bir “lambda” değerine karar vermektir. R programlamada “lambda”yı otomatik belirleyen “BoxCox.lambda” fonksiyonuyla patent sayıları serisi için uygun olan “lambda” değeri 0.0607704 olarak belirlenmiştir. Bu değer ile dönüşüm yapıldığında verinin logaritması alınacaktır [47]. Belirlenen “lambda” değeri üzerinden “BoxCox” fonksiyonuyla dönüşüm yapılmıştır. Tahminleme yaparken dönüştürülmüş seri kullanılacaktır.

3.2.1 Zaman Serisi Regresyonu ile Patent Sınıflarına Ait Patent Sayılarının Tahminlenmesi

Patent sınıflarına ait patent sayılarının zaman serisi ile tahminleme modelinin oluşturulması için zaman serisi regresyonu modeli oluşturulmuştur. Veri seti, modelleme için eğitim ve test olarak ikiye bölünmüştür. Eğitim örneklemiyle modelleme için gerekli katsayı ve bilinmeyenleri tahminleme yapılırken, test örneklemiyle eğitim örnekleminde tahmin edilen parametrelerin uyumu ve performansının ölçümü yapılır. Tez çalışmasında eğitim verileri 2008 yılından 2015 yılına kadar olan kısmı kapsarken, test verileri 2016’dan 2018’e kadar olan kısmı kapsamaktadır. Veri setinde 2019 yılı için tüm senenin verisi bulunmadığı için 2019 kapsam dışı bırakılmıştır. R programlamada eğitim ve test veri seti oluşturulması “window” fonksiyonuyla yapılırken, “tslm” fonksiyonuyla doğrusal model oluşturulmuştur. Kurulan modelin sonuçları “summary” fonksiyonuyla gösterilmiştir.

```

> summary(linearModel) #ozet bilgi

Call:
tslm(formula = train ~ trend + season)

Residuals:
    Min       1Q   Median       3Q      Max
-780.71 -214.14  -41.58   158.50 1600.56

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -221.616    172.563  -1.284   0.2046
trend         14.175      2.837   4.997 6.72e-06 ***
season2      6085.254    216.667  28.086 < 2e-16 ***
season3      730.936    216.723   3.373  0.0014 **
season4     1385.618    216.816   6.391 4.30e-08 ***
season5       426.014    216.946   1.964  0.0548 .
season6      -31.447    217.113  -0.145  0.8854
season7     5516.378    217.316  25.384 < 2e-16 ***
season8       981.346    217.557   4.511 3.62e-05 ***
season9     1428.743    217.834   6.559 2.31e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 405.3 on 53 degrees of freedom
Multiple R-squared:  0.9721,    Adjusted R-squared:  0.9674
F-statistic: 205.1 on 9 and 53 DF,  p-value: < 2.2e-16

```

Şekil 3-7 Doğrusal Regresyon Modellemesinin Özeti

Şekil 3-7’deki doğrusal regresyon modelinin özetini incelediğimizde ilk bölümde “Call” başlığı altında kullanılan R formülü görülmektedir.

Bir alt başlık olan “Residuals” bölümünde ise gerçek gözlemlenen değerlerle modelin öngördüğü değerler arasındaki farklar gösterilmektedir. Kalıntılar bölümü beş nokta için sonuç vermiştir. Modelin ne kadar iyi tahminde bulunduğunu değerlendirirken, bu noktaların sıfır(0) değerinde simetrik bir dağılım içinde olması gerektiği unutulmamalıdır. Fakat örneğimizde noktaların dağılımının simetrik görünmediği gözlemlenmektedir. Bu, modelin gözlemlenen gerçek noktalardan çok uzakta noktaları tahmin ettiğini gösterir. Bunu, görmek için kalıntıların grafiği çizdirilebilir.

Kalıntılar başlığından bir sonraki bölüm “Coefficients” modelin katsayılarından bahseder. Basit doğrusal regresyonda katsayılar kesişme(intercept) ve eğim(slope) terimlerini temsil eden iki bilinmeyen sabittir. Kalıntıların altında “Estimate” kolonu 10 satır içerir; ilki kesişme noktasıdır. Kesişme noktası doğrusal model çizgisinin nerede başladığını belirtir. Trend değişkeni için, seride sınıflar arasında patent sayısı olarak ortalama 14.175 adetlik artan bir eğilimin olduğu görülmektedir. Sınıfsal farklılıklar incelendiğinde ikinci sınıfta birinci sınıfa oranla ortalama 6085.245 adet daha fazla patent sayısına sahipken, son sınıf ilk sınıfa oranla ortalama 1428.743 adet daha fazla patent

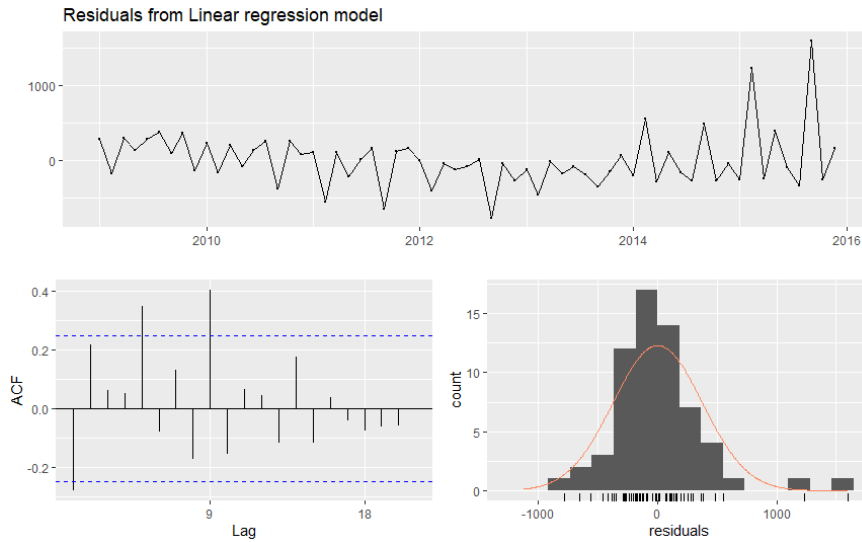
sayısına sahiptir. “Std. Error” kolonu katsayı tahminlerinin tahmin ettiği ortalama değerlerin, değişkenin ortalama değerinden farkını gösterir. “t value” kolonu katsayı tahminlerinin sıfırdan ne kadar standart sapma olduğunun ölçüsüdür. “Pr (>|t|)” kolonu t’ye eşit veya daha büyük herhangi bir değerin gözlemlenme olasılığıdır.

Doğrusal regresyon modelinin bir sonraki bölümü ise “Residual Standard Error” satırıdır ve doğrusal regresyonun kalitesinin ölçüsüdür. Bu satır, patent sayılarının 53 serbestlik derecesiyle hesaplandığında ortalama regresyon çizgisinden yaklaşık 405 adet sapma olabileceğini göstermiştir.

Bir alt satır olan “Multiple R-squared” değeri modele dahil edilen değişken sayısına bağlı olarak artacaktır. Aynı satırdaki “Adjusted R-squared” değeri ise modellemede kullanılan değişkenlerin bağımlı değişkendeki değişkenliğin %96,74’ünü kapsar uygunlukta olduğunu göstermektedir. R kare istatistikleri modelin gerçek verilere ne kadar iyi uyduğunun ölçüsüdür. R kare istatistiklerinden çoklu(multiple) R kare değeri dahil edilen değişkene göre değişeceği için ayarlanan(adjusted) R kare değeri daha çok tercih edilir.

En alt satır olan “F-statistics” değeri 1’den ne kadar fazlaysa o kadar iyidir.

Model oluşturma işlemi tamamlandıktan sonra “checkresiduals” fonksiyonuyla modelin ACF grafiği ve kalıntıların histogramı çizilmiştir.



Şekil 3-8 Doğrusal Regresyon Modelinin Grafikleri

Şekil 3-8’deki grafikler incelendiğinde kalıntıların ortalamasının sıfır olmadığı görülmektedir. Bu durum tahminlerin yanlış tahminler olmasına sebep olacaktır. ACF grafiğinden kalıntılar arasında otokorelasyon olduğu görülmektedir ve bu durum veride modelin kapsayamadığı kısımlar olduğunu gösterir.

```
> checkresiduals(linearModel)
```

```
Breusch-Godfrey test for serial correlation of order up to 13
```

```
data: Residuals from Linear regression model  
LM test = 40.361, df = 13, p-value = 0.0001209
```

Şekil 3-9 “Breusch-Godfrey” Testi Sonuçları

Kalıntılar arasındaki otokorelasyonun varlığı “Breusch-Godfrey” testi ile de görülebilir. Şekil 3-9’da görülen sonuçlardan p değeri 0.05’in altında olması istatistiksel olarak otokorelasyonun olduğunu göstermektedir.

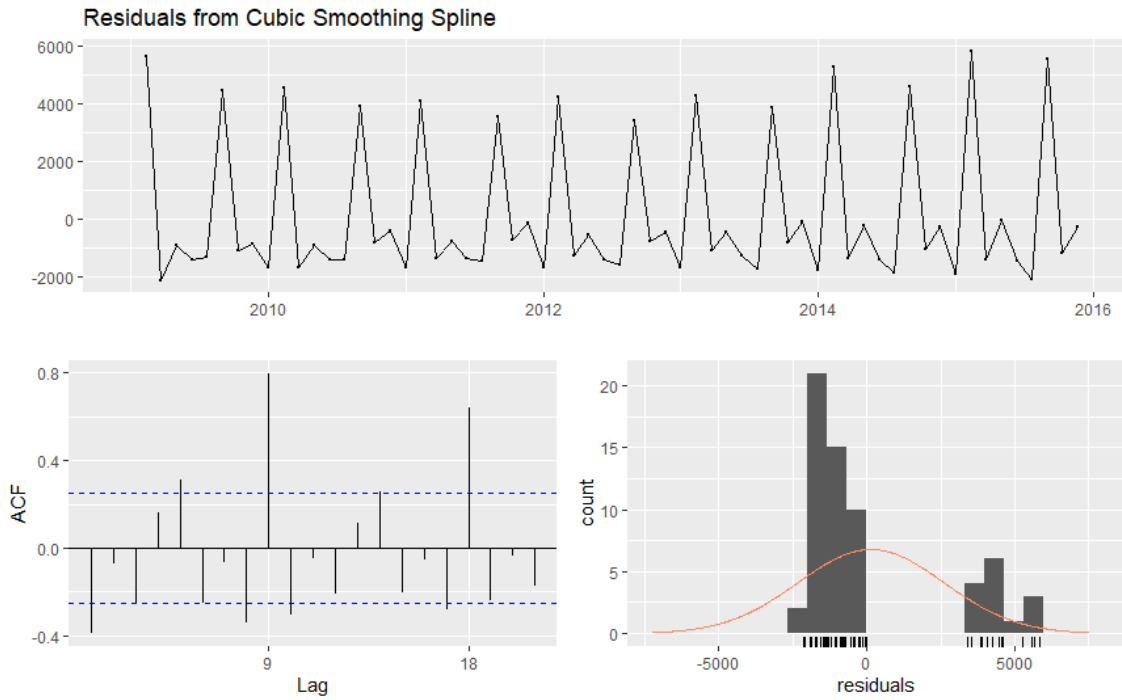


Şekil 3-10 Doğrusal Regresyon Modelinin Patent Sınıflarına ait Patent Sayıları Zaman Serisi Üzerindeki Performans Grafiği

Şekil 3-10’da gösterilen doğrusal regresyon modelinin patent sayıları zaman serisi üzerindeki performansı incelendiğinde tahminlenen verilerin eğitim verilerini

yakalamakta gayet başarılı olduğu görülmektedir. Model aynı zamanda trend olarak da eğitim verilerini yakalamıştır.

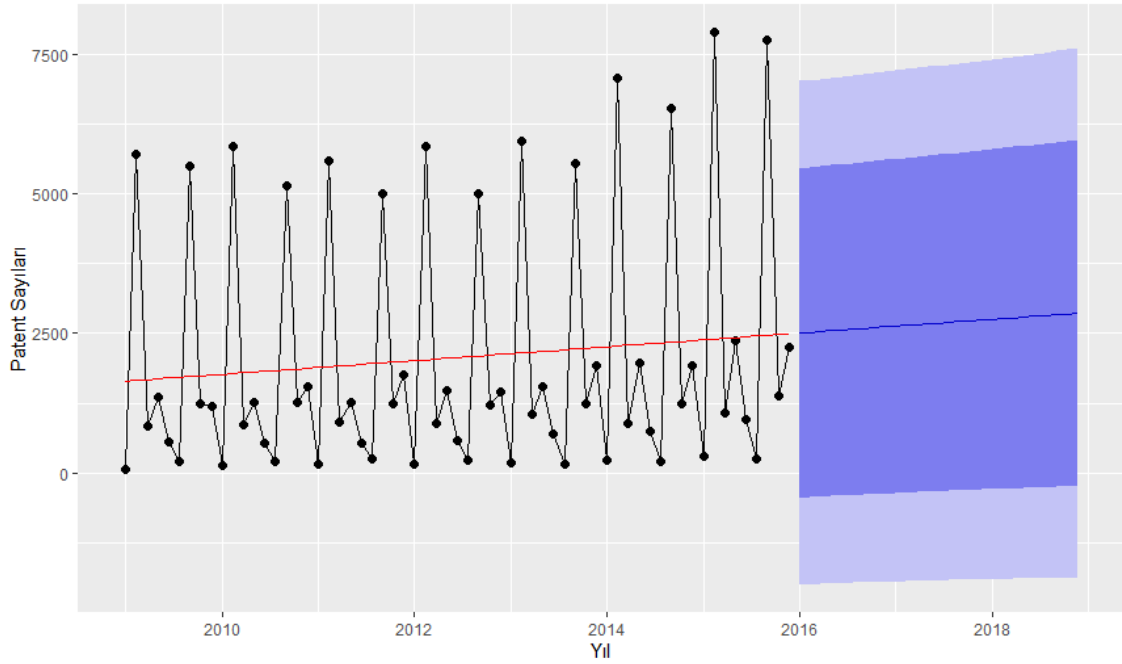
Doğrusal regresyon modelinden sonra “splinef” fonksiyonuyla doğrusal olmayan regresyon modeli oluşturulmuş ve zaman serisi ile uyumu incelenmiştir. Fonksiyonda “h” parametresi için 27 atanmıştır çünkü test verisi kadar tahmin istenmektedir. Doğrusal olmayan regresyon modelinde kübik düzleştirme çizgileri kullanılarak veriye uyum sağlanmaya çalışılmıştır.



Şekil 3-11 Doğrusal Olmayan Kübik Regresyon Modeli Grafikleri

Doğrusal ve doğrusal olmayan modellerin ACF grafiklerine baktığımızda kalıntıların sabit bir varyansla hareket ettiği görülmektedir. Fakat ACF grafiğinde doğrusal olmayan modelde doğrusal modele göre daha fazla sayıda anlamlı otokorelasyon katsayısı görülmektedir.

Bunlara ek olarak, tahminlenen değerler ile eğitim seti grafik üzerinde karşılaştırılmıştır.



Şekil 3-12 Doğrusal Olmayan Kübik Regresyon Modelinin Patent Sınıflarına ait Patent Sayıları Zaman Serisi Üzerindeki Performansı

Şekil 3-12’deki grafik incelendiğinde en sondaki mavi çizginin “splinef” fonksiyonundan dönen tahmin değerlerinden oluştuğu görülmektedir. Model tahmin değerlerinde trendi yakalamıştır. Doğrusal regresyon modelinin performans grafiğiyle karşılaştırıldığında daha az noktanın tahmin değeriyle kesiştiği görülmektedir. Bu durumda doğrusal olmayan kübik regresyon modeli doğrusal regresyon modelinden daha iyidir denilemez.

Doğrusal regresyon modeli ile doğrusal olmayan kübik regresyon modelinin performansları eğitim ve test verileriyle bazı hata kriterleri üzerinden ölçümlenerek karşılaştırılabilir. Karşılaştırma için kullanılan hata kriterleri ortalama hata, karekök ortalama hata, ortalama mutlak hata, ortalama yüzde hata, ortalama mutlak yüzde hata ve ortalama mutlak ölçekli hatadır.

- Ortalama Hata (ME, Mean Error): Model tarafından öngörülen tahminler ile gerçek değerler arasındaki ortalama hataya denir.

$$ME = \frac{1}{n} \sum_{j=1}^n e_j \quad (3.2)$$

- Karekök Ortalama Hata (RMSE, Root Mean Square Error): Tahmin hatalarının standart sapmasıdır. RMSE tahmin hatalarının ne kadar yayıldığıнын ölçüsüdür.

Regresyon analizinde yaygın olarak kullanılır.

$$RMSE = \sqrt{\frac{\sum_{j=1}^n e_j^2}{n}} \quad (3.3)$$

- Ortalama Mutlak Hata (MAE, Mean Absolute Error): Tüm mutlak hataların ortalamasıdır. Tahminlerin gerçek değerden ne kadar uzak olduğunu ölçer.

$$MAE = \frac{1}{n} \sum_{j=1}^n |e_j| \quad (3.4)$$

- Ortalama Yüzde Hata (MPE, Mean Percentage Error): gerçek değerle tahmin değerinin arasındaki farkın ortalama yüzdesidir.

$$MPE = \frac{100}{n} \sum_{j=1}^n \frac{e_j}{A_j} \quad (3.5)$$

- Ortalama Mutlak Yüzde Hata (MAPE, Mean Absolute Percentage Error): Zaman serisi modellerinde doğruluğu ölçmek için sıkça kullanılan ortalama mutlak yüzde hata, hata değerinin gerçek değerlerle oranını bulma temeline dayanır.

$$MAPE = \frac{100}{n} \sum_{j=1}^n \frac{|e_j|}{|A_j|} \quad (3.6)$$

- Ortalama Mutlak Ölçekli Hata (MASE, Mean Absolute Scaled Error): Tahminlerin doğruluğunun bir ölçüsüdür. Tahminlerin karşılaştırmalı doğruluğunu belirlemek için kullanılır.

$$MASE = \frac{\sum_{t=1}^n |e_t|}{\frac{n}{n-1} \sum_{t=2}^n |Y_t - Y_{t-1}|} \quad (3.7)$$

Ortalama mutlak hata ve karekök ortalama hata bu kriterlerden en yaygın olarak kullanılanlarıdır. Doğrusal ve doğrusal olmayan modeller bu iki kritere bakarak karşılaştırılır. Bu modellere ek olarak “naive metod” ve “ortalama metod” adında iki model daha karşılaştırma için kullanılmıştır. Ortalama metodu R programlamada “meanf” fonksiyonunu kullanırken, naive metodu “naive” fonksiyonunu kullanmaktadır. Meanf fonksiyonu zaman serisinde bulunan değerlerin ortalamasını alır ve bu ortalama değeri tahmin değeri olarak belirtir. Naive fonksiyonu ise zaman serisindeki son değeri tahmin değeri olarak belirtir. Doğrusal modelin tahminleri “forecast” fonksiyonuyla yapılmıştır. En sonunda modellerin eğitim ve test verileri üzerindeki hata değerleri “accuracy” fonksiyonuyla yazdırılmıştır.

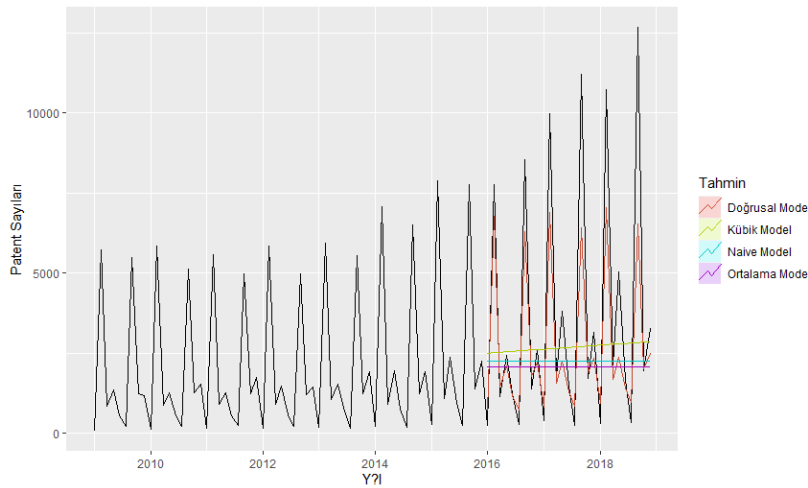
```

> accuracy(f_Linear, test)
              ME      RMSE      MAE      MPE      MAPE      MASE      ACF1 Theil's U
Training set -1.443995e-14 371.7559 261.0456  5.007233 36.47741  1.341881 -0.2795098      NA
Test set      9.068214e+02 1953.1295 1209.6612 -26.200567 58.11909  6.218154 -0.4143421 0.3780451
> accuracy(f_Cubic, test)
              ME      RMSE      MAE      MPE      MAPE      MASE      ACF1 Theil's U
Training set 140.2835 2441.829 1897.885 -219.9180 253.7809  9.755905 -0.3885857      NA
Test set      928.7213 3849.631 2702.513 -181.3536 221.2313 13.892023 -0.4631093 0.7600121
> accuracy(f_Mean, test)
              ME      RMSE      MAE      MPE      MAPE      MASE      ACF1 Theil's U
Training set 1.659645e-13 2225.330 1774.834 -280.9332 310.4855  9.123376 -0.3950438      NA
Test set      1.544698e+03 4056.164 2638.254 -118.6314 169.7488 13.561705 -0.4531847 0.8213965
> accuracy(f_Naive, test)
              ME      RMSE      MAE      MPE      MAPE      MASE      ACF1 Theil's U
Training set  35.33871 3738.268 2854.758 -190.5799 248.8004 14.67462 -0.6163597      NA
Test set     1351.55556 3986.612 2653.704 -139.0522 186.3357 13.64112 -0.4531847 0.8021462

```

Şekil 3-13 R Programlamayla Regresyon Modellerinin Karşılaştırılması

Şekil 3-13'te görüldüğü üzere; hata kriterlerinde eğitim verileri için MAE değeri doğrusal model için 261.0456, kübik model için 1897.885 ve naive model için 2854.758'dir. Aynı verilerin RMSE değerlerine bakıldığında doğrusal model 371.7559, kübik model için 2441.829 ve naive model için ise 3738.268'dir. Bu durumda MAE ve RMSE baz alındığında en iyi model doğrusal modeldir. Yukarıdaki doğrusal ve doğrusal olmayan kübik regresyon modellerinin performans grafiklerine bakıldığında da veriye en iyi oturan model doğrusal modeldir ve MAE ve RMSE bazında da daha iyi sonuçlar vererek tahminlemede daha başarılı olmuştur. Bunun yanı sıra kübik model farklı hata metriklerinde daha iyi sonuçlar çıkarmış olabilir. Fakat burada en sık kullanılan MAE ve RMSE değerlerine bakılıp karşılaştırma yapılmıştır.



Şekil 3-14 Regresyon Modellerinin Patent Sınıflarına Ait Patent Sayıları Zaman Serisi Üzerindeki Tahmin Performansları Grafiği

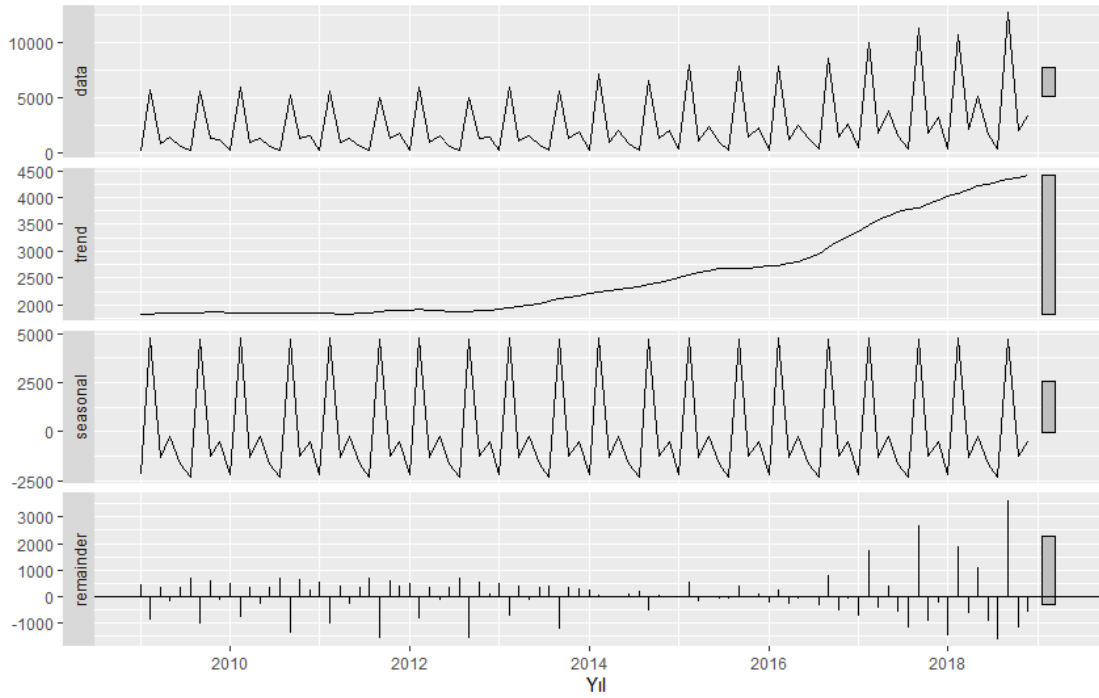
3 farklı modelin tahminlerini grafik üzerinde incelemek için “autoplot” fonksiyonuyla Şekil 3-13’te gösterilen hata kriterlerinin performans grafikleri R programlamada çizdirilmiştir. Tüm tahminlemelerin sonuçları grafik üzerine katmanlar halinde “autolayer” fonksiyonuyla eklenmiştir. Şekil 3-14’te gösterilen modellerin performans grafikleri incelendiğinde ortalama model eğitim verilerinin ortalamasını aldığı için en kötü performansı sergilemektedir. Dört model arasından doğrusal model artan trend göstererek hem trendi yakalayan hem de test setine en yakın tahmini yapan model olmuştur. Doğrusal model daha gerçek ve tutarlı hareket etmiştir. Genel olarak baktığımızda en iyi modelin doğrusal model olduğu söylenebilir.

3.2.2 Ayırıştırma Yöntemleri ile Patent Sınıflarına Ait Patent Sayılarının Tahminlenmesi

Bir zaman serisi mevsimsel dalgalanma, trend, döngüsel hareket ve düzensiz rastgele hareketlere sahip bileşenlerden oluşmaktadır. Ayırıştırma yöntemi; bu bileşenlerin tahmininden zaman serisinin öngörüsünü hesaplayan yöntemdir. Zaman serisinin seyri, geleceğe ilişkin davranış biçiminin belirlenmesinde fayda sağlayacaktır. Ayırıştırma yöntemi kısa dönemli tahminlerde yapılması ve anlaşılması en kolay yöntemlerden biridir. Ayırıştırma yönteminin zayıf yönlerinden biri sezgilere dayanıyor olmasıdır.

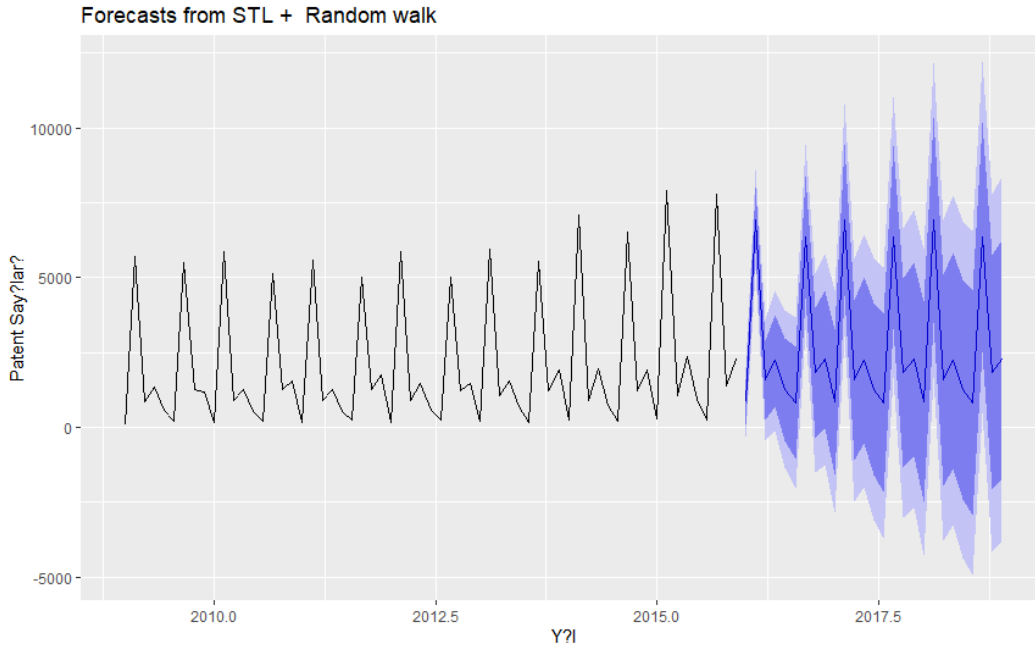
Zaman serisi modelleri toplumsal ve çarpımsal modeller olmak üzere iki sınıfa ayrılabilir. Çarpımsal model mevsimsel hareketin sabit dalgalanmalardan oluşmadığı durumlarda tercih edilir. Toplumsal modelde mevsimsel dalgalanmalar sabit kalmakta, zaman içinde değişmemektedir. Çarpımsal modelde ise serinin dalgalanması zaman içinde artar ya da azalır.

Toplamsal ve çarpımsal ayırıştırma yapabilen SEATS, STL ve X11 gibi pek çok yöntem mevcuttur. Bu çalışmada daha esnek bir yapıda olduğu için STL yöntemi kullanılmıştır. R programlamada “stl” fonksiyonuyla çağrılan bu yöntem trend, mevsimsellik ve kalıntı bileşenlerine ayrılmıştır. “forecast” fonksiyonuyla da bu yöntem üzerinde tahminlemeler yazılmıştır.



Şekil 3-15 R Programlamada STL Ayırıştırma Yöntemi ile Patent Sayısı Zaman Serisinin Ayırıştırılması ve Tahminlerin Alınması

Şekil 3-15'teki grafikler incelendiğinde patent sayıları zaman serisinin STL yöntemiyle bileşenlerine ayrıştırıldığında, trend bileşeninin zaman serisini düzgün bir şekilde takip ettiği görülür. Mevsimsellik bileşeninden sınıfsal dalgalara bakıldığında dalgaların her dönemde aynı desene sahip olduğu görülmektedir. Kalıntılar bileşeni trend ve mevsimsellik bileşenlerinden zaman serisinin çıkarılmasıyla oluşmaktadır ve 2018'deki H04L sınıfında yüksek değerler görülmektedir. Bu durum o sınıftaki gözlem değerlerinin yüksek olmasından kaynaklanmaktadır.



Şekil 3-16 STL Ayırıştırma Yöntemi ile Patent Sınıflarına Ait Patent Sayıları Zaman Serisi Üzerinde Tahminleme Grafiği

Şekil 3-16'daki grafikte STL ayırıştırma yöntemi ve naive metoduyla oluşturulan tahmin değerleri gösterilmektedir. Burada tahminler zaman serisinin trendini yakalayabilmiş ve sabit dalgalanmalar şeklinde devam etmiştir.

3.2.3 Üstel Düzleştirme Yöntemi ile Patent Sınıflarına Ait Patent Sayılarının Tahminlemesi

Üstel düzleştirme yöntemi verideki son değişimleri dikkate alarak tahminlerin sürekli güncellendiği bir yöntemdir. Yöntemin genel denklemi şu şekildedir [39];

$$Z_{T+1} = S_T = \alpha Z_T + \alpha(1 - \alpha)Z_{T-1} + \alpha(1 - \alpha)^2 Z_{T-2} + \dots + \alpha(1 - \alpha)^{T-1} Z_1 + \alpha(1 - \alpha)^T S_0 \quad (3.8)$$

$$Z_{T+1} = S_T = \alpha \sum_{j=0}^T (1 - \alpha)^j Z_{T-j} \quad (3.9)$$

S_T = Bir sonraki dönemin (T+1) tahmini

Z_T = Son döneme ait gerçek değer

α = Düzleştirme katsayısı ya da ağırlık

Denklem incelendiğinde tahminlerin gözlem değerlerinin ağırlıklarla çarpılarak elde edildiği görülmektedir. Ağırlıklar sıfır ile bir arasında bir değer olduğundan geçmiş gözlemlere doğru ağırlık değeri düşmektedir. Bu yüzden serinin tahmininde son döneme ait gözlemlerin etkisi daha büyüktür.

Başlıca üstel düzleştirme yöntemleri arasında Basit Üstel Düzleştirme, Holt Doğrusal Trend Yöntemi ve Holt-Winters Mevsimsel Yöntemi vardır. “Basit Üstel Düzleştirme Yöntemi” trend veya mevsimsellik bulunmayan seriler için kullanılırken “Holt Doğrusal Trend Yöntemi” trend bileşeninin olduğu mevsimsellik bileşeninin olmadığı zaman serileri için kullanılmaktadır. “Holt-Winters Mevsimsel Yöntemi” hem trend hem mevsimsellik içeren seriler için kullanılmaktadır. Patent sayıları zaman serisinde hem trend bileşeninin olduğu ve hem de mevsimsellik bulunduğu holt-winters mevsimsel yönteminin kullanılması daha uygun görülmektedir. Diğer üstel düzleştirme yöntemlerinin de performansları karşılaştırılmıştır.

```
> BDMModel <- tscv(mydata, ses, h = 27) #Basit Duzlestirme Fonksiyonu
> HoltModel <- tscv(mydata, holt, h = 27) #Holt Dogrusal Trend
> HoltModel_Damped <- tscv(mydata, holt, damped=TRUE, h = 27) #Holt Dogrusal Trend
> HoltWintersModel <- tscv(mydata, hw, seasonal='multiplicative', h = 27) #Holt Winters Mevsimsel Yontemi
> HoltWintersModel_Damped <- tscv(mydata, hw, seasonal='multiplicative', damped=TRUE, h = 27) #Holt Winters Mevsimsel Yontemi
> sqrt(mean(BDMModel^2, na.rm = TRUE))
[1] 3041.84
> sqrt(mean(HoltModel^2, na.rm = TRUE))
[1] 11598.5
> sqrt(mean(HoltModel_Damped^2, na.rm = TRUE))
[1] 11678.78
> sqrt(mean(HoltWintersModel^2, na.rm = TRUE))
[1] 4385.355
> sqrt(mean(HoltWintersModel_Damped^2, na.rm = TRUE))
[1] 4377.769
> mean(abs(BDMModel), na.rm = TRUE)
[1] 2008.21
> mean(abs(HoltModel), na.rm = TRUE)
[1] 4556.084
> mean(abs(HoltModel_Damped), na.rm = TRUE)
[1] 4580.236
> mean(abs(HoltWintersModel), na.rm = TRUE)
[1] 1448.79
> mean(abs(HoltWintersModel_Damped), na.rm = TRUE)
[1] 1488.917
```

Şekil 3-17 Patent Sınıflarına ait Patent Sayıları Zaman Serisine Uygulanan Üstel Düzleştirme Modelleri ve Hata Metrikleri

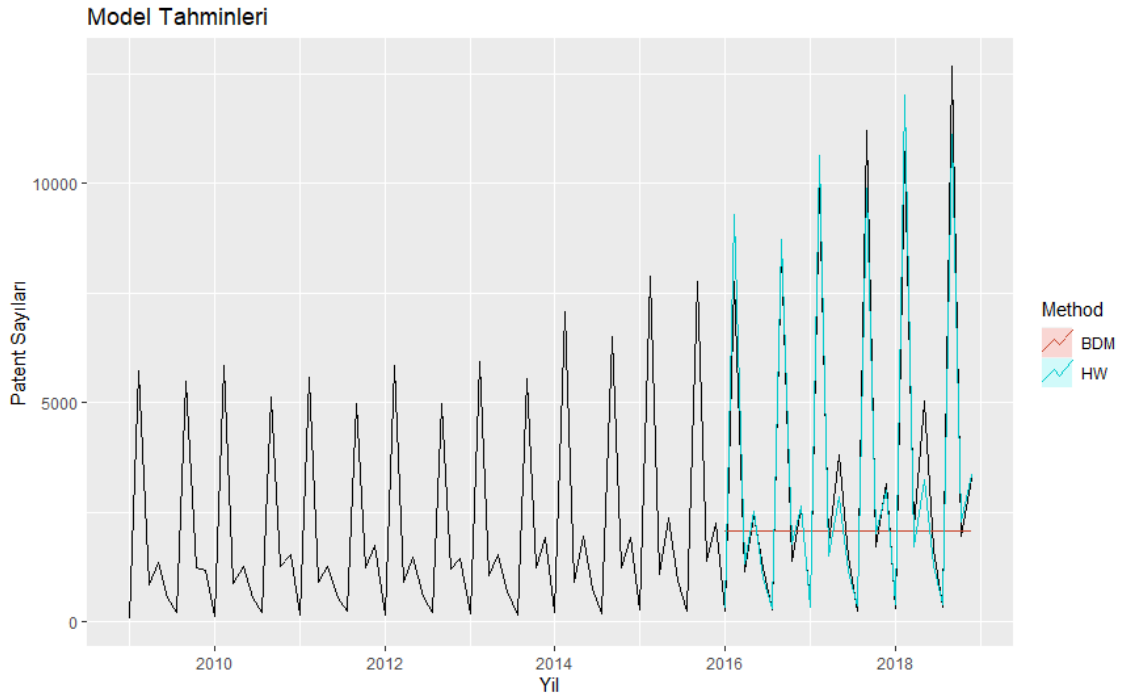
R programlama ile “tsCV” fonksiyonu kullanılarak zaman serisinin alt kümelerine tahmin fonksiyonu uygulayarak elde edilen tahmin hataları hesaplanmıştır. Basit düzleştirme fonksiyonu için “ses” parametresi verilirken, holt doğrusal trend yöntemi yöntemi için “holt” ve holt-winters mevsimsel yöntemi için “hw” parametresi kullanılmıştır. Öngörü ufku olarak 27 adet tahmin istenmiştir. Bazı modellerde damped parametresi “TRUE” olarak verilmiş olup serinin sürekli yükselen bir trend içinde olmayacağı belirtilmiştir. Holt-winters mevsimsel yönteminde “seasonal” parametresi “multiplicative” olarak

atanmış ve mevsimselliğin artarak seyrettiği söylenmiştir. Oluşturulan beş modelin ortalama kare hatası ve kök ortalama kare hataları hesaplatılmış ve bu modeller karşılaştırılmıştır. Şekil 3-17’de kök ortalama kare hata metriklerine baktığımızda 3041.8 ile basit düzleştirme fonksiyonu diğerlerine göre daha iyi sonuç vermişken ortalama kare hata metriklerinden holt-winters mevsimsel yöntemi diğerlerinden daha iyi sonuç vermişlerdir. Sonuç olarak basit düzleştirme fonksiyonu ve holt-winters mevsimsel yöntemi en başarılı modeller olarak görülmektedir. Karşılaştırma için elde edilen sonuçlar Şekil 3-18’den incelenebilir.

```
> accuracy(BDM, test)
      ME      RMSE      MAE      MPE      MAPE      MASE      ACF1 Theil's U
Training set  0.5552372 2225.441 1774.643 -280.8385 310.3974  9.12239 -0.3950465      NA
Test set     1544.7989193 4056.202 2638.250 -118.6208 169.7403 13.56169 -0.4531847 0.8214065
> accuracy(HW, test)
      ME      RMSE      MAE      MPE      MAPE      MASE      ACF1 Theil's U
Training set 38.71128 219.8872 166.0563 -4.847281 17.28965 0.8535973 -0.07100802      NA
Test set     74.27456 716.0642 476.1914 -1.838276 16.70514 2.4478189 0.02835037 0.1269762
```

Şekil 3-18 Patent Sınıflarına Ait Patent Sayıları Zaman Serisi Üzerinde Üstel Düzleştirme Modellerinin Oluşturulması ve Hata Metrikleri

En başarılı iki model olan basit düzleştirme fonksiyonu ve holt-winters mevsimel yöntemlerinin üzerinden tahminlemeler yapabilmek için modeller eğitim verileri üzerinden modellenmiştir ve “accuracy” fonksiyonuyla modellerin test verileri üzerindeki performansları ölçülmüştür. Sonuçlar incelendiğinde holt-winters mevsimsel modelinin performans olarak önde olduğu görülmektedir. Bu iki modelin tahminlemeleri daha sonra tüm veri ile beraber grafik haline getirilmiş ve Şekil 3-19’daki grafik ortaya çıkmıştır.



Şekil 3-19 Patent Sayıları Zaman Serisine Uygulanan Üstel Düzleştirme Model Tahminleri

Şekil 3-19'daki grafik incelendiğinde holt-winters mevsimsel yöntemi trend olarak grafiği yakalamış ve veri setiyle daha uyumlu ilerlemiştir. Basit düzleştirme fonksiyonu ise ne trendi yakalamış ne de doğru tahminler yapmıştır.

```

> Hw
      Point Forecast      Lo 80      Hi 80      Lo 95      Hi 95
2016.000      295.7025  213.5359  377.8691  170.039521  421.3655
2016.111      9290.8978  6704.2269 11877.5688  5334.926023 13246.8696
2016.222      1328.6336   957.1911  1700.0761   760.561378  1896.7058
2016.333      2511.0253 1804.0707  3217.9798 1429.831625  3592.2189
2016.444      1009.7930   722.5115  1297.0744   570.433827  1449.1521
2016.556      320.9631   228.3431   413.5832  179.312965   462.6133
2016.667      8715.3758  6154.2352 11276.5165  4798.449228 12632.3024
2016.778      1788.7294 1251.3024  2326.1564   966.805672  2610.6531
2016.889      2654.2660 1835.7726  3472.7593 1402.488407  3906.0435
2017.000      339.2971   220.1442   458.4501  157.068390   521.5259
2017.111     10639.9144  6800.9024 14478.9263  4768.652269 16511.1765
2017.222      1518.6768   954.1994  2083.1543   655.383089  2381.9706
2017.333      2864.9317 1765.4323  3964.4311 1183.392445  4546.4709
2017.444      1150.0612   693.4477  1606.6747   451.731089  1848.3913
2017.556      364.9141   214.7864   515.0418   135.313605   594.5146
2017.667      9892.1125 5669.7972 14114.4279  3434.638404 16349.5867
2017.778      2026.9113 1128.4459  2925.3767   652.827112  3400.9955
2017.889      3002.8986 1619.6396  4386.1575   887.386543  5118.4106
2018.000      384.3272   187.7873   580.8671    83.745344   684.9090
2018.111     12033.9351 5664.5592 18403.3111 2292.815200 21775.0551
2018.222      1715.1425   775.2266  2655.0584   277.665109  3152.6199
2018.333      3230.9516 1397.3264  5064.5767  426.663945  6035.2392
2018.444      1295.1912   533.9224  2056.4601  130.930980  2459.4515
2018.556      410.4074   160.5881   660.2268    28.341682   792.4732
2018.667     11110.6530 4107.4329 18113.8732  400.152038 21821.1540
2018.778      2273.6575   789.9908  3757.3242    4.585089   4542.7299
2018.889      3364.2172 1092.1253  5636.3090 -110.647441  6839.0818
> data.frame(TestSet=test, Hw=Hw$mean)
  weight      Hw
1      250  295.7025
2     7759  9290.8978
3     1138  1328.6336
4     2440  2511.0253
5     1223  1009.7930
6       278   320.9631
7     8551  8715.3758
8     1393  1788.7294
9     2621  2654.2660
10      398   339.2971
11    9973 10639.9144
12     1788  1518.6768
13     3798  2864.9317
14     1472  1150.0612
15       261   364.9141
16   11209  9892.1125
17     1712  2026.9113
18     3145  3002.8986
19       293   384.3272
20   10723 12033.9351
21     2149  1715.1425
22     5034  3230.9516
23     1669  1295.1912
24       345   410.4074
25   12676 11110.6530
26     1967  2273.6575
27     3274  3364.2172
  > Hw$model
Holt-winters' multiplicative method

Call:
hw(y = train, h = 27, seasonal = "multiplicative")

Smoothing parameters:
  alpha = 0.0316
  beta  = 0.0303
  gamma = 0.3251

Initial states:
  l = 1657.9426
  b = 23.0669
  s = 0.7752 0.6231 2.815 0.2342 0.2649 0.659
      0.3549 3.1643 0.1095

sigma: 0.2168

      AIC      AICC      BIC
966.7506 975.5006 996.7545

```

Şekil 3-20 Holt-Winters Mevsimsel Yöntemi Model İncelemesi ve Tahminleri

Şekil 3-20’de, başarılı tahminler veren holt-winters mevsimsel yöntemi modelinin tahminleri ve güven aralıkları yazdırılmış, modelin parametreleri, katsayıları ve ceza parametreleri incelenmiştir ve test verileriyle karşılaştırılmıştır. Aynı şekildeki “Initial States” başlığındaki l ve b ifadelerinin holt-winters mevsimsel yöntemindeki serinin seviyesi ve trend ifadelerinin değerleri olduğu görülmektedir. Bu denklemin içerisindeki katsayı değerleri ise “Smoothing Parameters” başlığı altındakiler olduğu görülmektedir. AIC, AICC ve BIC değerleri ceza parametreleri olup aynı bağımlı değişkenler üzerinden aynı şekilde dönüşüm yapılmış modellerin karşılaştırılması için kullanılır. Daha iyi performansa sahip modelde bu parametrelerin değerleri düşüktür.

3.2.4 ARIMA Modelleri ile Patent Sınıflarına Ait Patent Sayılarının Tahminlenmesi

Bu bölümde patent sınıflarına ait patent sayıları zaman serisi üzerinde ARIMA modelleri kullanarak tahminlemeler yapılmaya çalışılmıştır.

ARIMA modelleme durağan zaman serileri ile yapılmaktadır. Bir zaman serisinin durağan olması demek serinin zamandan bağımsız olması demektir. Yani gözlemlerin zamandan etkilenmemesi gereklidir. Bir serinin durağan olup olmadığı birim kök testleriyle anlaşılabilir. “adf” testi birim kök testlerinden bir tanesidir. Bu test, sıfır hipotezine göre modelin trendini de dikkate alıp serinin durağanlığı hakkında bilgi verir.

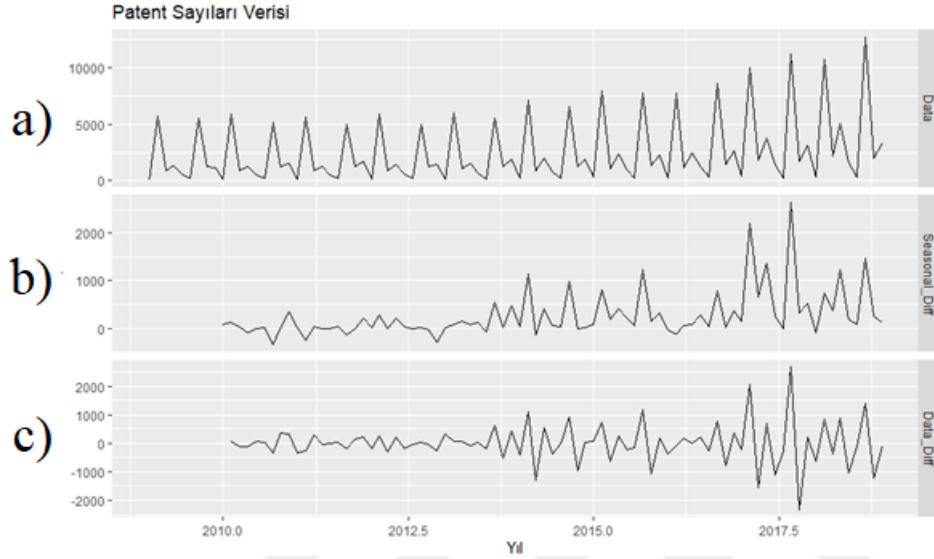
Bu çalışmadaki veri setinde artan bir trende sahip olduğu için durağan değildir denilebilir. Seriyeye fark alma veya mevsimsel fark alma gibi işlemler uygulanarak durağan olmayan veri seti ARIMA modelleme için durağan hale getirilebilir. Fark alma, zaman serisinin ardışık gözlemlerinin birbirinden çıkarılması işlemidir. Mevsimsel fark alma ise zaman serisinin bir önceki mevsimsel periyottaki değerlerinin birbirinden çıkarılması işlemidir. Son çeyrek verisinin bir önceki yılın son çeyrek verisinden çıkarılması gibidir. R programlama üzerinde fark alma işlemi için gerekli farkların sayısını tahmin etmek için “ndiffs”, mevsimsel farkların sayısını tahmin etmek için ise “nsdiffs” fonksiyonu kullanılır. Bu iki fonksiyonla bir zaman serisini durağan hale getirebilmek için gerekli fark alınması işlemlerinin sayısı bulunabilir.

```
> ndiffs(mydata)
[1] 1
> nsdiffs(mydata)
[1] 1
```

Şekil 3-21 Patent Sınıflarına Ait Patent Sayısı Zaman Serisinin R Programlama ile Durağanlaştırılması

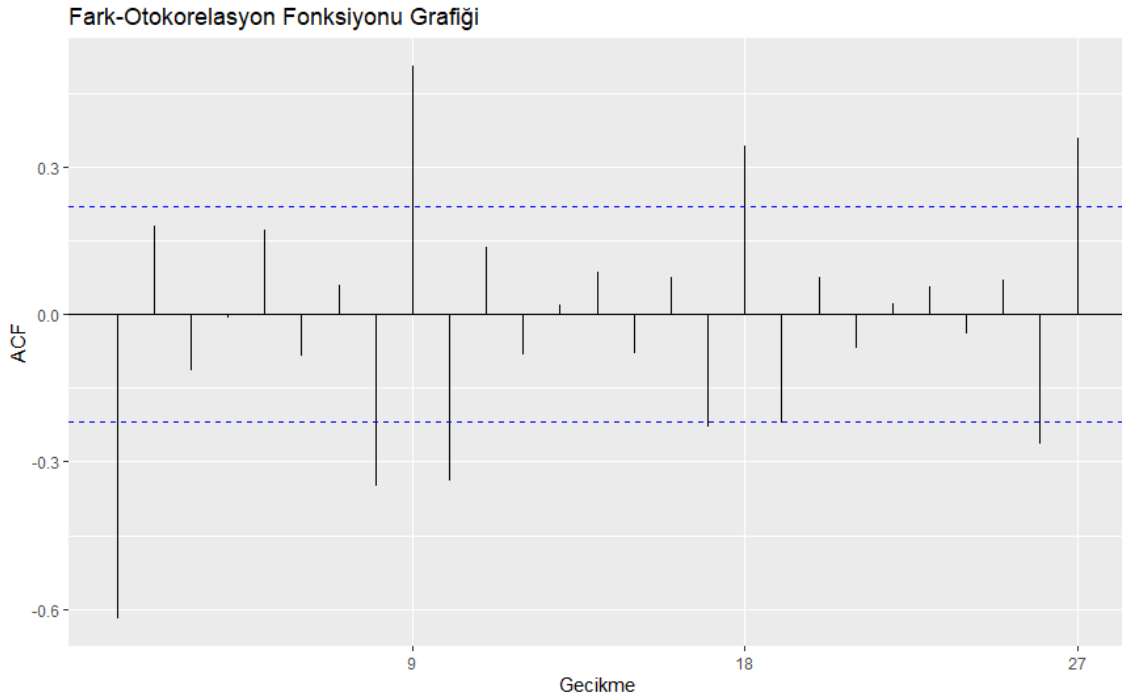
Şekil 3-21’deki sonuç zaman serisinde hem normal fark alma hem de mevsimsel fark almaya ihtiyaç olduğunu göstermektedir. Öncelikle veriye mevsimsel fark alma işlemi uygulanmıştır. Daha sonra mevsimsel fark alma işlemi uygulanan veriye normal fark alma işlemi yapılmıştır. Fark alma işlemleri “diff” fonksiyonuyla yapılmıştır ve fonksiyona normal fark alma işleminde parametre olarak bir, mevsimsel fark alma işleminde toplam dokuz sınıf olduğu için dokuz verilmiştir.

R programlamada “cbind” fonksiyonuyla matris yaratılmış, verinin kendisini, sınıfsal fark uygulanmış halini ve sınıfsal fark uygulanan verinin normal fark alma uygulanmış halini üç farklı kolon olarak bu matrisin içine yerleştirilmiştir. Devamında “autoplot” fonksiyonuyla grafiği çizdirilmiştir.



Şekil 3-22 Durağanlaştırılmış Patent Sınıflarına Ait Patent Sayısı Zaman Serisi Grafiği

Şekil 3-22’deki grafik incelendiğinde a bölümünde verinin kendisi artan bir trend ile ilerlemektedir. B ve c bölümlerinde mevsimsel fark alma işleminin sonucunda hem trend hem de mevsimsellik ortadan kalkmıştır ve verinin hareketler rastgele haline gelmiştir. Burası verinin büyük ölçüde durağanlaştığının göstergesidir.

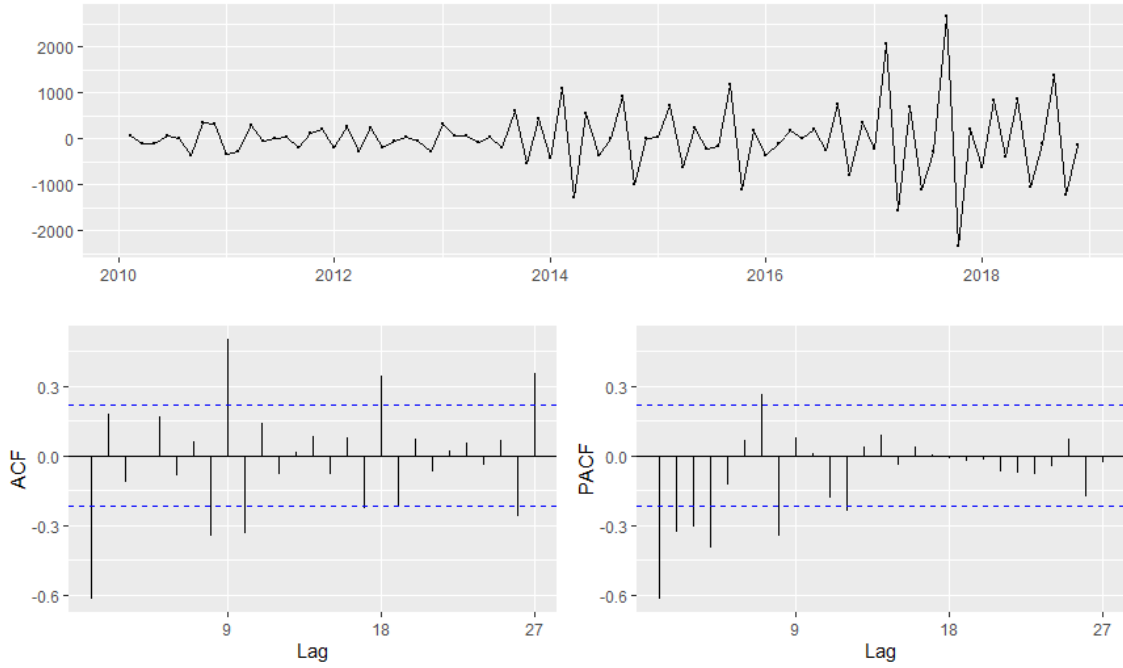


Şekil 3-23 Patent Sınıflarına ait Patent Sayısı Zaman Serisi Fark-Otokorelasyon Fonksiyonu Grafiği

Şekil 3-23'deki grafik veri setinin durağanlaştığından emin olmak için çizilmiştir. Fark alınmış serinin otokorelasyon görüntüsünde birbiriyle ilişkisi olmayan bir görünüm mevcuttur. Grafik incelendiğinde beyaz gürültü serisinin otokorelasyon grafiğine çok benzediğini söylenebilir çünkü katsayıların gidişatı tamamen rastgeledir. Bu grafikten yola çıkarak fark alınmış serinin durağan bir seri olduğu söylenebilir.

ARIMA modelleme otoregresif ve hareketli ortalama modellerinin entegrasyonudur. Otoregresif modellerde regresyon bağımlı değişkenin kendisi kullanılarak yapılır. Oto ifadesi de burdan gelmektedir. Hareketli ortalama modelleme ise geçmiş dönemdeki tahmin hatalarını kullanarak tahmin yapar.

ARIMA modellerinde p, d, q ve P, D, Q değerlerini bulabilmek için zaman serisine bakmak genellikle yeterli gelmemektedir. Bu yüzden ACF ve PACF grafiklerine bakmak gereklidir. P ve p değeri hangi sayıda gecikmeli gözlemin kullanılacağını, Q ve q değeri hangi sayıda tahmin hatalarının kullanılacağını, d ve D değeri ise ne kadar sayıda fark alma işlemi uygulanacağını gösterir. R programlamada ACF ve PACF grafikleri “`ggtsdisplay`” fonksiyonuyla oluşturulur.



Şekil 3-24 Durağanlaştırılmış Patent Sınıflarına Ait Patent Sayıları Zaman Serisi, ACF ve PACF Grafikleri

ACF ve PACF grafiklerindeki dokuz ve katları sınıfsal değerleri, diğer değerler sınıfsal olmayan değerleri göstermektedir. Şekil 3-24'deki ACF grafiği incelendiğinde 1. gecikmede anlamlı gösterge olduğundan modelleme esnasında dikkate alınabilir. 9., 18. ve 27. gecikmedeki çubuk mevsimsel MA(1) modelini tavsiye ediyor olabilir bu yüzden ACF grafiğindeki çubuklardan MA(1) modelinin göstergesi olabilir. Aynı şekilde PACF grafiğinde 2. çubuktan sonra 3., 4., 7., 8. ve 12. çubuklarda bir anlamlılık vardır. Buradan AR(1), AR(2), AR(3), AR(4), AR(7), AR(8) ve AR(12) modeli dikkate alınabilir. Bu değerlerle birlikte $ARIMA(2, 1, 1)(1, 0, 1)_9$, $ARIMA(3, 1, 1)(1, 0, 1)_9$, $ARIMA(4, 1, 1)(1, 0, 1)_9$, $ARIMA(7, 1, 1)(1, 0, 1)_9$, $ARIMA(8, 1, 1)(1, 0, 1)_9$ ve $ARIMA(12, 1, 1)(1, 0, 1)_9$ modelleri bu çalışmada kullanılmıştır.

```

> #ARIMA(12,1,1)(0,0,0)[9]
> Model23 <- Arima(y = diff_data,
+                 order = c(12, 1, 1),
+                 seasonal = list(order = c(0, 0, 0), period = 9),
+                 lambda = 0)

> list = list(Model11, Model12, Model13, Model14, Model15, Model16, Model17, Model18, Model19, Model10, Model11, Model12, Model13, Model14, Model15, Model16, Model17, Model18, Model19, Model20, Model21, Model22, Model23)
> for (i in list) {
+   print(i$aicc)
+ }
[1] 151.6592
[1] 148.8565
[1] 149.2893
[1] 147.1458
[1] 153.9943
[1] 151.2521
[1] 153.0282
[1] 150.104
[1] 152.2362
[1] 147.4834
[1] 156.1893
[1] 153.0375
[1] 153.147
[1] 150.2763
[1] 163.8813
[1] 159.9047
[1] 159.6949
[1] 157.5525
[1] 168.0923
[1] 164.0141
[1] 160.8771
[1] 161.6735
[1] 163.6483

```

Şekil 3-25 Patent Sınıflarına Ait Patent Sayıları Zaman Serisine Farklı ARIMA Modellerinin Uygulanması

Bahsedilen modeller birkaç farklı versiyonuyla beraber oluşturulup birbirleriyle karşılaştırılmışlardır. Karşılaştırma için AIC; Akaike Information Criterion – Akaike bilgi Kritesi değeri kullanılmıştır. En küçük AIC değerini veren model $ARIMA(2, 1, 1)(0, 0, 0)_9$ olmuştur.

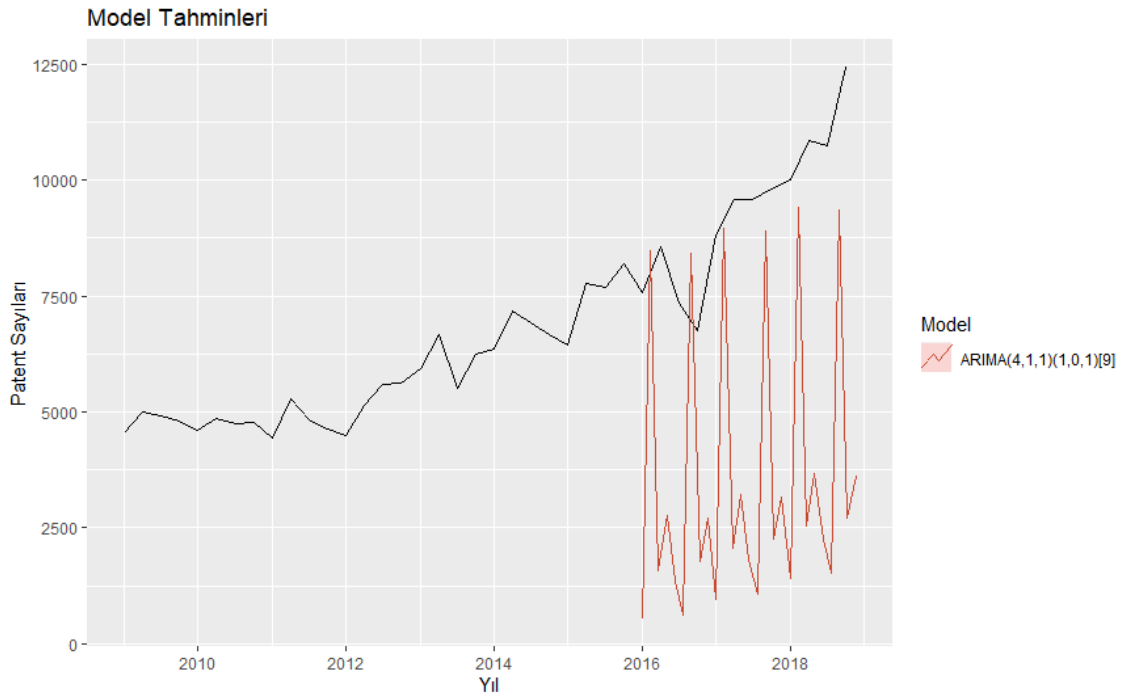
Daha sonra oluşturulan tüm modeller “Arima” fonksiyonuyla eğitim verileriyle modellenmiş, “forecast” fonksiyonuyla tahminler oluşturulmuş ve “accuracy” fonksiyonuyla da test verileri üzerindeki performansları ölçülerek karşılaştırma yapılmıştır.

Training set	ME	RMSE	MAE	MPE	MAPE	MASE	ACF1	Theil's U
Test set	17.6729	260.6781	186.1224	-6.697597	19.54296	0.9567453	0.08322497	NA
Training set	-141.3866	1020.4407	797.2397	-74.883481	81.93543	4.0981383	-0.45181453	0.1630987
Test set	ME	RMSE	MAE	MPE	MAPE	MASE	ACF1	Theil's U
Training set	37.78421	256.6396	164.1282	-5.453804	18.05454	0.843686	-0.05921764	NA
Test set	140.02524	1124.7360	754.9700	-58.572363	68.74045	3.880855	-0.39810893	0.202806
Training set	ME	RMSE	MAE	MPE	MAPE	MASE	ACF1	Theil's U
Test set	142.7727	1105.902	885.8596	-122.4485	153.7584	4.553681	-0.1060570	NA
Training set	1114.4541	3655.729	2331.4824	-124.9905	163.8611	11.984774	-0.4307057	0.7181159
Test set	ME	RMSE	MAE	MPE	MAPE	MASE	ACF1	Theil's U
Training set	227.4108	2008.166	1629.417	-240.3317	297.0359	8.375872	-0.1235966	NA
Test set	1327.0761	3989.499	2666.398	-144.3813	191.3234	13.706378	-0.4541298	0.8018805
Training set	ME	RMSE	MAE	MPE	MAPE	MASE	ACF1	Theil's U
Test set	20.30098	243.807	156.3899	-6.184327	15.48483	0.8039081	-0.03933177	NA
Training set	13.58313	1014.754	727.2000	-60.296939	68.52052	3.7381055	-0.41594118	0.1759751
Test set	ME	RMSE	MAE	MPE	MAPE	MASE	ACF1	Theil's U
Training set	157.6812	954.3927	748.4951	-100.2302	133.9846	3.847571	-0.0441446	NA
Test set	833.5634	3539.0407	2353.8484	-143.6089	177.8836	12.099744	-0.4382677	0.6862308
Training set	ME	RMSE	MAE	MPE	MAPE	MASE	ACF1	Theil's U
Test set	22.16164	240.3028	153.3775	-6.919776	17.01478	0.7884232	-0.04737741	NA
Training set	17.22889	1015.2388	715.5677	-59.378730	67.48157	3.6783109	-0.41713478	0.1780316
Test set	ME	RMSE	MAE	MPE	MAPE	MASE	ACF1	Theil's U
Training set	25.21096	276.8843	188.5394	-9.3845	27.98935	0.9691696	-0.04326021	NA
Test set	58.43743	1146.6932	791.7956	-67.7267	77.77741	4.0701536	-0.44158970	0.2026668
Training set	ME	RMSE	MAE	MPE	MAPE	MASE	ACF1	Theil's U
Test set	157.6507	953.3141	691.267	-93.38165	128.5661	3.553395	-0.1847633	NA
Training set	1007.7945	3488.4364	2220.316	-115.21292	160.0437	11.413334	-0.4556140	0.6837906
Test set	ME	RMSE	MAE	MPE	MAPE	MASE	ACF1	Theil's U
Training set	256.4805	1729.410	1204.392	-183.8036	231.3787	6.19107	-0.2218731	NA
Test set	1197.0037	3829.749	2540.142	-134.5344	178.4458	13.05737	-0.4692401	0.7531439
Training set	ME	RMSE	MAE	MPE	MAPE	MASE	ACF1	Theil's U
Test set	21.69590	236.570	154.3021	-6.489362	17.40624	0.7931762	-0.03797432	NA
Training set	16.97427	1014.454	727.2240	-57.831114	66.21712	3.7382290	-0.43304061	0.1764995
Test set	ME	RMSE	MAE	MPE	MAPE	MASE	ACF1	Theil's U
Training set	28.78941	243.3763	167.2583	-6.717894	20.47054	0.8597759	-0.04684418	NA
Test set	85.60477	1084.2899	760.3210	-57.094790	66.94410	3.9083610	-0.44280294	0.1913463
Training set	ME	RMSE	MAE	MPE	MAPE	MASE	ACF1	Theil's U
Test set	150.8841	877.4389	615.4882	-67.07266	110.3929	3.163861	0.04759369	NA
Training set	939.7475	3469.1108	2202.9331	-85.90228	130.1306	11.323978	-0.43649106	0.6782281
Test set	ME	RMSE	MAE	MPE	MAPE	MASE	ACF1	Theil's U
Training set	255.2305	1569.974	995.6792	-127.1885	170.4170	5.118198	0.04277057	NA
Test set	1118.8826	3796.703	2531.6306	-108.6334	151.9101	13.013618	-0.46192974	0.7499978
Training set	ME	RMSE	MAE	MPE	MAPE	MASE	ACF1	Theil's U
Test set	40.40969	244.883	175.5733	-7.707181	20.12093	0.9025186	-0.09199533	NA
Training set	438.23042	1217.390	731.7900	-20.642698	38.79112	3.7617003	-0.28909410	0.2281254
Test set	ME	RMSE	MAE	MPE	MAPE	MASE	ACF1	Theil's U
Training set	26.08607	254.2271	180.0300	-8.565265	26.76360	0.925428	0.03346959	NA
Test set	40.69982	1085.9336	776.1463	-64.165593	74.33498	3.989710	-0.49392978	0.190361
Training set	ME	RMSE	MAE	MPE	MAPE	MASE	ACF1	Theil's U
Test set	142.7421	856.7698	583.8569	-66.56451	103.8178	3.001263	-0.02226276	NA
Training set	975.2567	3395.6964	2221.0037	-77.63356	125.6786	11.416868	-0.48540010	0.653228
Test set	ME	RMSE	MAE	MPE	MAPE	MASE	ACF1	Theil's U
Training set	241.2871	1539.008	978.2883	-125.6318	168.3151	5.028802	-0.03942782	NA
Test set	1153.9098	3722.454	2587.6856	-103.4634	157.0126	13.301763	-0.51336006	0.7233336
Training set	ME	RMSE	MAE	MPE	MAPE	MASE	ACF1	Theil's U
Test set	33.74563	231.9407	163.3533	-6.870831	19.22397	0.8397029	-0.02718419	NA
Training set	222.55283	1050.8594	647.4126	-29.738027	40.64341	3.3279657	-0.37743123	0.1985762
Test set	ME	RMSE	MAE	MPE	MAPE	MASE	ACF1	Theil's U
Training set	90.44061	294.1152	185.5872	-0.9769404	14.79414	0.953994	-0.08788839	NA
Test set	796.47727	1419.0136	843.8761	14.2110253	20.74000	4.337869	-0.11940403	0.2620964
Training set	ME	RMSE	MAE	MPE	MAPE	MASE	ACF1	Theil's U
Test set	131.0380	337.845	209.5205	2.520129	16.01753	1.077021	0.07845406	NA
Training set	924.5653	1543.319	938.5268	21.527683	23.63689	4.824412	-0.15191754	0.2838936
Test set	ME	RMSE	MAE	MPE	MAPE	MASE	ACF1	Theil's U
Training set	38.37315	255.2557	173.5759	-9.050753	19.22468	0.8922513	-0.05817447	NA
Test set	279.96570	1111.2976	671.3672	-38.795330	50.35386	3.4511021	-0.30952406	0.2080831
Training set	ME	RMSE	MAE	MPE	MAPE	MASE	ACF1	Theil's U
Test set	45.13771	266.6311	182.7713	-10.21160	22.34154	0.9395194	-0.0653946	NA
Training set	240.04895	1143.6981	714.2793	-48.96129	60.15762	3.6716877	-0.3390699	0.212861

Şekil 3-26 ARIMA Modellerinin Test Verileri Üzerindeki Performansı

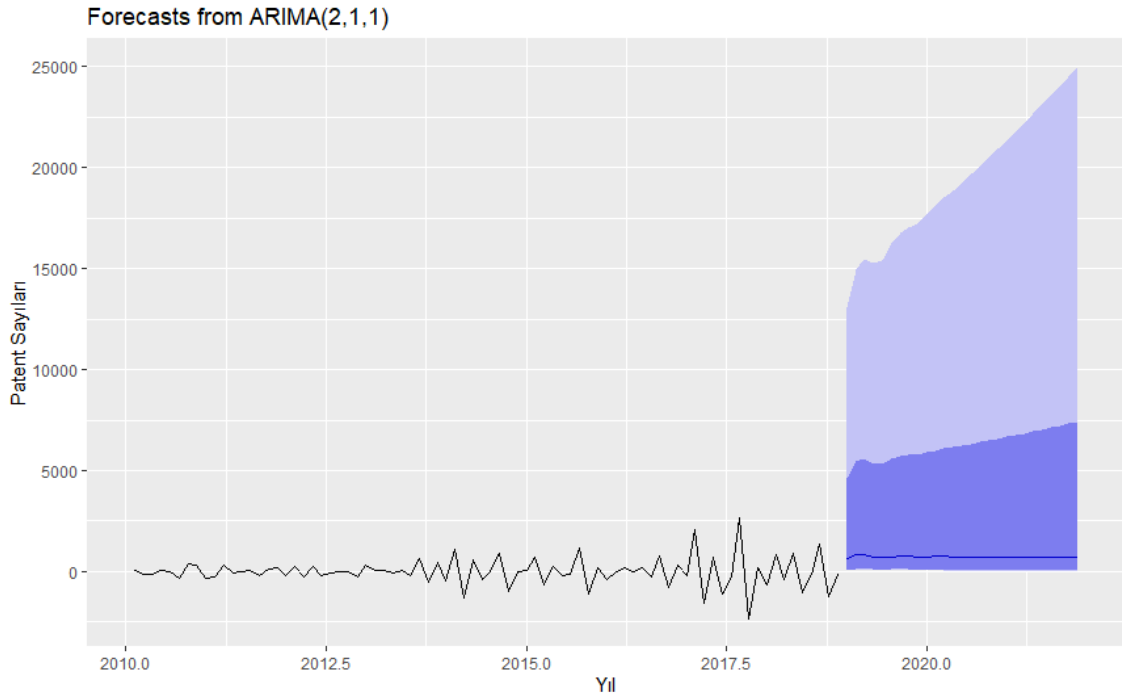
Şekil 3-26’da gösterilen farklı modellerin hata metriklerine bakıldığında eğitim verisi ile RMSE değerleri karşılaştırıldığında en başarılı model en düşük RMSE değeri ile $ARIMA(8, 1, 1)(1, 0, 1)_9$ modelidir. Test verileri üzerinden karşılaştırma yapıldığında en düşük RMSE değerine sahip en başarılı model $ARIMA(4, 1, 1)(1, 0, 1)_9$ olmuştur. $ARIMA(2, 1, 1)(0, 0, 0)_9$, $ARIMA(8, 1, 1)(1, 0, 1)_9$ ve $ARIMA(4, 1, 1)(1, 0, 1)_9$ modelleri tahminleme için kullanılabilir.

Test verisi üzerinde en başarılı performansa sahip olan $ARIMA(4, 1, 1)(1, 0, 1)_9$ modelinin test verileriyle uyumunu ve gelecek üç senelik tahminlerini görmek için R programlama ile bir grafik çizilmiştir.



Şekil 3-27 $ARIMA(4,1,1)(1,0,1)[9]$ Modelinin Test Verisi Üzerindeki Performans Grafiği

Şekil 3-27’deki grafik, Şekil 3-26’daki modellerden test verisi üzerindeki en başarılı seçilen modelin gelecek üç yıl için tahminlerini göstermektedir.



Şekil 3-28 Tüm Veri Üzerinde Oluşturulan ARIMA(2,1,1)(0,0,0)[9] Modelinin Üç Senelik Tahmin Grafiği

Tüm veri seti üzerinde oluşturulan modellerden en başarılı seçilen modelin gelecek üç yıl için tahminlerinin olduğu grafiğe Şekil 3-28'den ulaşılabilir. Bu grafikte kullanılan modelin detayları ve nokta tahminleri bir sonraki Şekil 3-29 üzerinden incelenebilir.

```
> Model14
Series: diff_data
ARIMA(2,1,1)
Box Cox Transformation: lambda= 0

Coefficients:
      ar1      ar2      ma1
    0.2438  -0.4872  -0.8209
s.e.  0.1759   0.1914   0.1149

sigma^2 estimated as 1.706: log likelihood=-68.98
AIC=145.97  AICC=146.52  BIC=155.34
> Model14 %>% forecast(h=27) -> Fit
> Fit$mean
Time Series:
Start = c(2019, 1)
End = c(2021, 9)
Frequency = 9
[1] 635.2811 809.8196 818.1770 728.7438 704.9265 739.8038 760.7837 748.1784 735.0538 737.8702 744.9539 745.3017 741.9250
[14] 740.9354 742.3347 743.1594 742.6777 742.1587 742.2667 742.5458 742.5613 742.4290 742.3893 742.4440 742.4767 742.4580
[27] 742.4375
>
```

Şekil 3-29 ARIMA Modellerinin İncelemesi R Kodu

Şekil 3-29'dan tahmin sonuçları incelendiğinde 2020 yılında ilk sınıf olan C12N için 737.8702 adet patent sayısı öngörülmüştür. Aynı şekilde 2012 yılında 9. sınıf olan H04W için 742.4375 adet patent sayısı tahmini yapılmıştır. Tahminlerde kullanılan AIC, BIC ve AICC değerleri ve hata katsayıları modelin detaylarında mevcuttur.

ARIMA modelleme yaparken manuel olarak belirlenen modellerden sonra R programlama üzerinde otomatik olarak belirlenen modeller ile de tahminlemeler yapılmıştır. “auto.arima” fonksiyonu ile yapılan modellemenin detayları incelenmiştir.

```
> Sfit_auto <- auto.arima(y=diff_data, seasonal=TRUE, stepwise = FALSE,
+ approximation = FALSE)
> Sfit_auto
Series: diff_data
ARIMA(1,0,1)(2,0,0)[9] with zero mean

Coefficients:
      ar1      ma1      sar1      sar2
-0.0447 -0.9196  0.2248  0.4916
s.e.    0.0684  0.0598  0.0427  0.0841

sigma^2 estimated as 165342: log likelihood=-596.3
AIC=1202.59 AICC=1203.41 BIC=1214.5
> Fit_auto_train <- Arima(y=train, order=c(1,0,1), seasonal=list(order = c(2,0,0), period = 9),
> Fit_auto_forecasts <- forecast(Fit_auto_train, h = 27)
> accuracy(Fit_auto_forecasts, test)

      ME      RMSE      MAE      MPE      MAPE      MASE      ACF1 Theil's U
Training set  62.27257 254.974 170.5534 -7.677674 19.54170 0.8767142 -0.0002984255 NA
Test set      537.25621 1079.830 675.6613  3.895147 22.79577 3.4731754 -0.1165923112 0.1937775

> autoplot(diff_data) +
+ autolayer(Fit_auto_forecasts, series="ARIMA(1,0,1)", PI = FALSE) +
+ ggtitle("Model Tahminleri") + xlab("Yıl") + ylab("Patent Sayıları") +
+ guides(colour=guide_legend(title="Model"))
> Sfit_auto %>% forecast(h = 27) %>% autoplot() + xlab("Yıl") + ylab("Patent Sayıları")

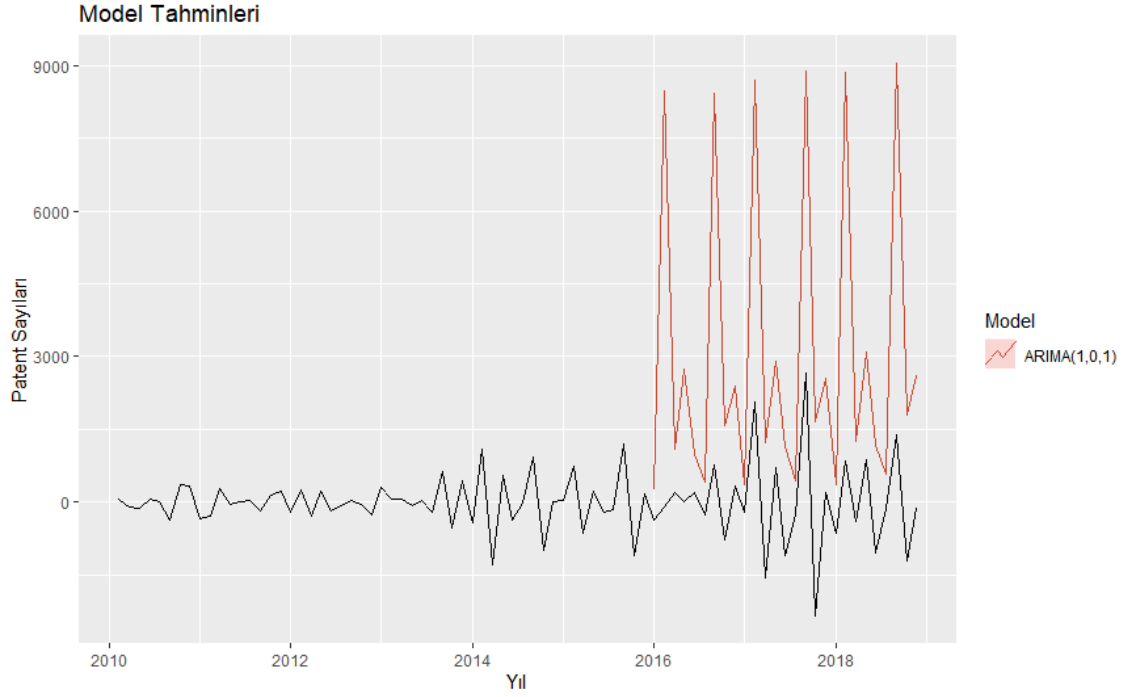
> Sfit_auto_f <- forecast(Sfit_auto, h = 27)
> cbind('Manuel Model Tahminleri'=Fit$mean,
+       'Otomatik Model Tahminleri'=Sfit_auto_f$mean)
Time Series:
Start = c(2019, 1)
End = c(2021, 9)
Frequency = 9

      Manuel Model Tahminleri Otomatik Model Tahminleri
2019.C12N      635.2811      225.05067
2019.G06F      809.8196      1186.85857
2019.G06K      818.1770      -855.42362
2019.G06Q      728.7438      544.71391
2019.G08B      704.9265      -778.76525
2019.G10L      739.8038      -156.17610
2019.H04L      760.7837      1626.00127
2019.H04N      748.1784      -1422.37363
2019.H04W      735.0538      72.46610
2020.C12N      737.8702      -258.65948
2020.G06F      744.9539      687.12362
2020.G06K      745.3017      -383.52160
2020.G06Q      741.9250      552.62261
2020.G08B      740.9354      -685.85952
2020.G10L      742.3347      -90.65919
2020.H04L      743.1594      1045.41685
2020.H04N      742.6777      -915.57673
2020.H04W      742.1587      -45.65899
2021.C12N      742.2667      52.50621
2021.G06F      742.5458      737.95508
2021.G06K      742.5613      -506.76706
2021.G06Q      742.4290      392.01722
2021.G08B      742.3893      -537.03450
2021.G10L      742.4440      -97.16030
2021.H04L      742.4767      1034.38977
2021.H04N      742.4580      -905.09359
2021.H04W      742.4375      25.36479
```

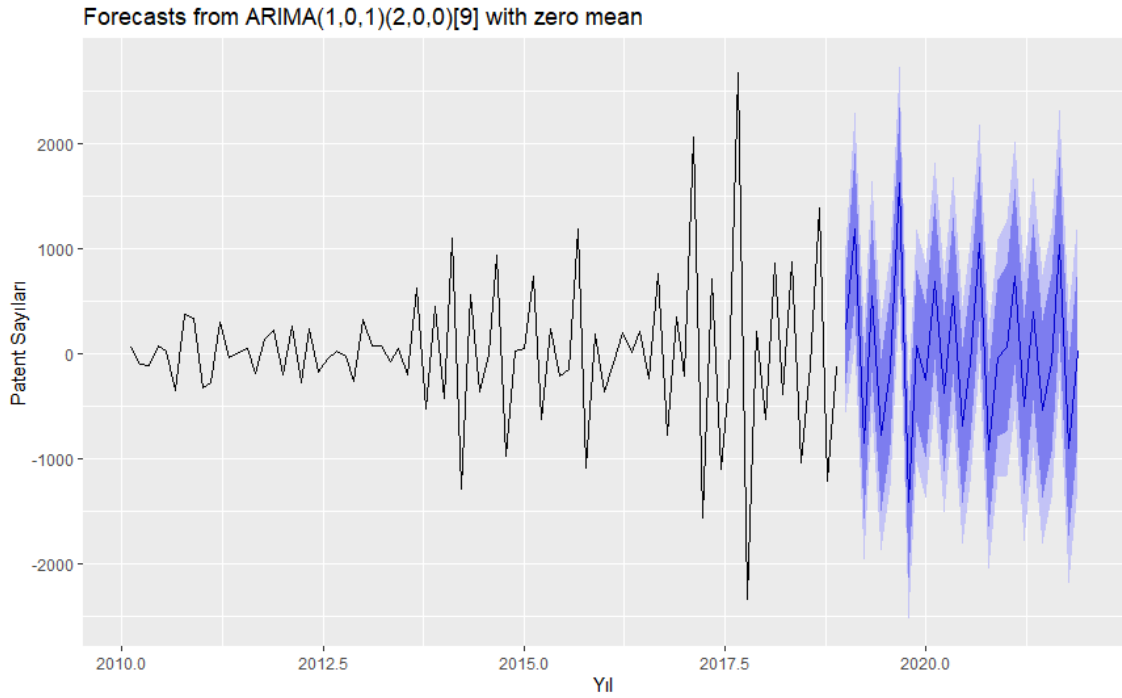
Şekil 3-30 Auto.arima Fonksiyonunun Patent Sınıflarına Ait Patent Sayıları Zaman Serisine Uygulanması ve İncelenmesi

ARIMA ile otomatik modellemede öncelikle “auto.arima” fonksiyonuyla veri seti üzerinde ARIMA modeli otomatik olarak oluşturulmuştur. Modeli oluştururken veri seti mevsimsel olduğu için “seasonal” parametresi kullanılmıştır. Otomatik oluşturulan modelin $ARIMA(1,0,1)(2,0,0)_9$ olduğu görülmektedir. Bu modelin test verileri üzerinde hata kriterlerinden RMSE değerine bakıldığında 1079.830 olduğu görülmektedir ve bu değer manuel modellemedeki RMSE değeri 1014.454 olan

$ARIMA(4, 1, 1)(1, 0, 1)_9$ modelinden daha performanslı olmadığı görülmektedir. Otomatik belirlenen modelin test verileri üzerindeki performansı ve gelecek üç sene için tahmin grafiği çizdirilmiştir. En son olarak otomatik ve manuel oluşturulan modellerin nokta tahminleri karşılaştırılmıştır.



Şekil 3-31 Auto.arima Modelinin Patent Sınıflarına Ait Patent Sayıları Zaman Serisi Üzerindeki Test Tahminleri Grafiği



Şekil 3-32 Auto.arima Modelinin Patent Sınıflarına Ait Patent Sayıları Zaman Serisi Üzerindeki Gelecek Üç Senelik Tahmin Grafiği

3.2.5 Yapay Sinir Ağları Modeli ile Patent Sınıflarına Ait Patent Sayılarının Tahminlenmesi

Patent sayıları zaman serisi üzerinde regresyon, ayrıştırma, ETS ve ARIMA modelleri ile tahminlemeler önceki bölümlerde yapılmıştır. Bu bölümde yapay sinir ağlarıyla bir model kurulup tahminlemeler yapılmaya çalışılmıştır. Yapay sinir ağlarıyla modellemeyi zaman serilerinde kullanırken ileri beslemeli yapay sinir ağı kullanılmaktadır. Bu sinir ağında gizli bir katman bulunmaktadır ve bu katmanda bulunan nöron sayısı önemlidir.

R programlamada zaman serisi üzerinde yapay sinir ağıyla modelleme yapmak için “nnetar” fonksiyonu kullanılmıştır. Bu fonksiyon zaman serisinin gecikmeli değerleri üzerinden uygun bir model oluşturur. Ortaya çıkan model otoregresif bir modeldir ve tahmin aralıkları için simülasyonlar kullanılır.

```

> neural_model <- nnetar(mydata)
> neural_model
series: mydata
Model: NNAR(9,1,5)[9]
Call:    nnetar(y = mydata)

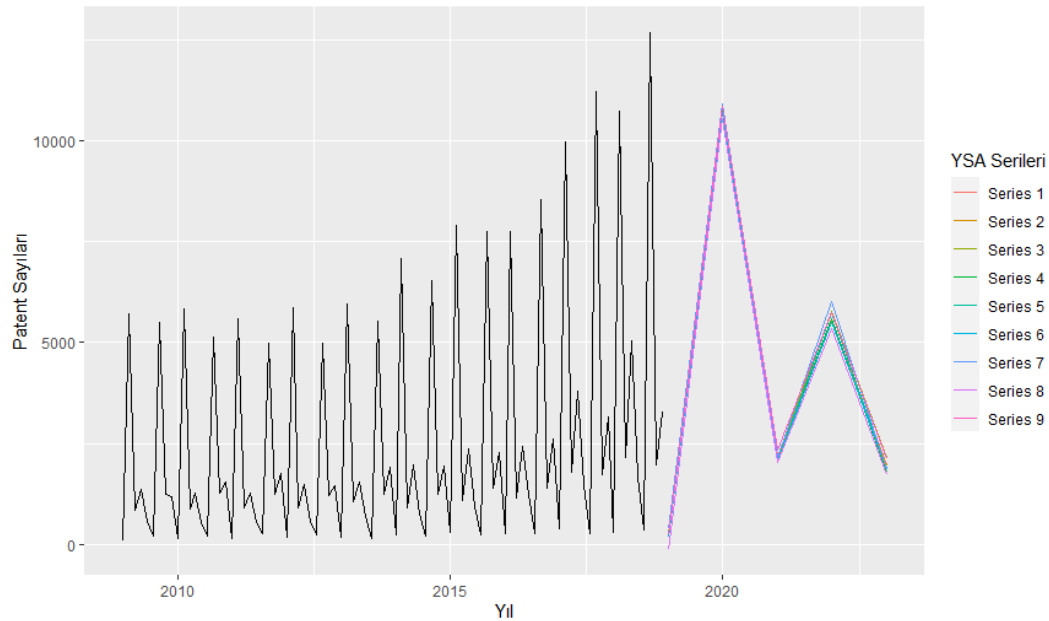
Average of 20 networks, each of which is
a 9-5-1 network with 56 weights
options were - linear output units

sigma^2 estimated as 14924
> sim <- ts(matrix(0, nrow=5, ncol=9), start=end(mydata)[1]+1)
> for(i in seq(9))
+   sim[,i] <- simulate(neural_model, nsim=5)
> autoplot(mydata) + forecast::autolayer(sim) + xlab("Yıl") + ylab("Patent Sayıları") + guides(colour=guide_legend(title="YSA Serileri"))

```

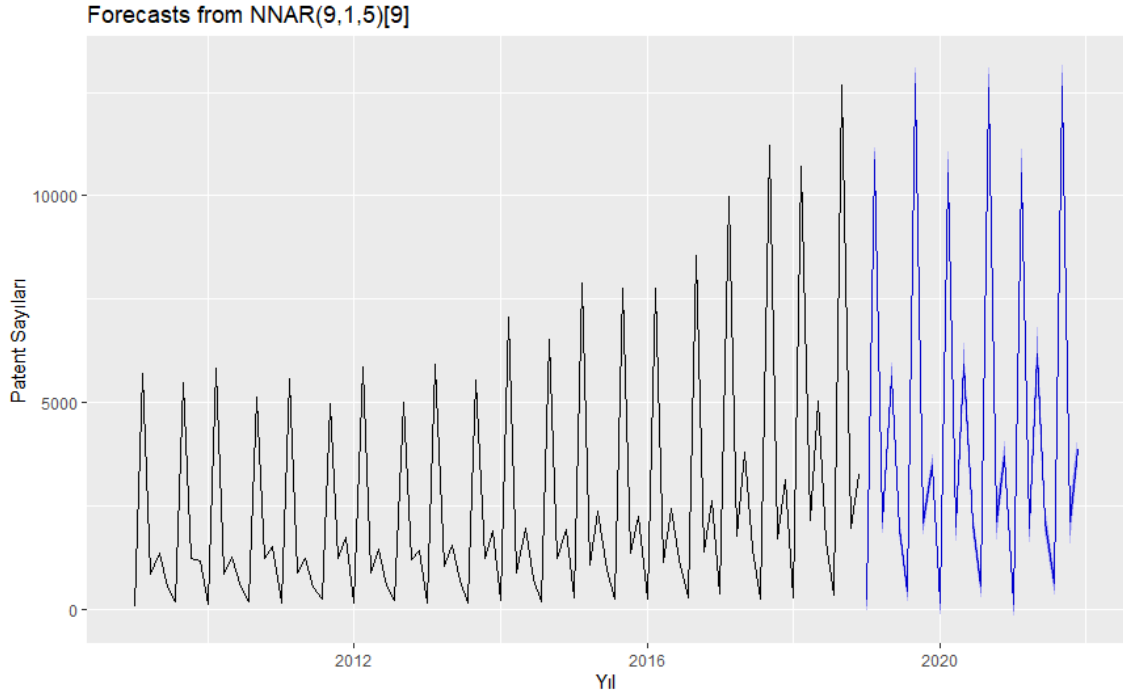
Şekil 3-33 Patent Sayıları Zaman Serisi Üzerinde Yapay Sinir Ağları Modelleme ve Simülasyonları

Şekil 3-33'deki grafik incelendiğinde “nnetar” fonksiyonu sonucunda serideki dokuzuncu gecikmenin girdi olarak alındığı, bir adet mevsimsel fark uygulanmış ve gizli katmanda beş nöronun olduğu NNAR(9, 1, 5) modeli ortaya çıkmıştır. R programlamada bu model ile gelecekteki örnek yol beş yıllık süreçlerle yinelemeli olarak simüle edilmiştir. Ayrıca veriler için gelecek dokuz olası yolun beş yıl sonrası simülasyonları Şekil 3-34'deki grafik ile gösterilmiştir. YSA serileri patent sınıflarını göstermektedir.



Şekil 3-34 Yapay Sinir Ağları Modellerinin Patent Sınıflarına Ait Patent Sayıları Zaman Serisi Üzerindeki Simülasyonu

R üzerinde “forecast” metoduyla modelin üç senelik tahminlemeleri alınmış ve güven aralıklarıyla birlikte aşağıdaki grafik oluşturulmuştur.



Şekil 3-35 NNAR(9,1,5)[9] Yapay Sinir Ağları Modelinin Patent Sınıflarına Ait Patent Sayıları Zaman Serisi Üzerindeki Tahminleri

Modelin kendisinin ve test verileri üzerindeki performansının çıktıları aşağıdaki şekilde gösterilmektedir.

```
> neural_model <- nnetar(train)
> accuracy(fitted(neural_model),train)
      ME      RMSE      MAE      MPE      MAPE      ACF1 Theil's U
Test set -0.4317373 309.7418 235.5784 -0.3427156 4.172775 -0.2618641 0.558653
> neural_model_test <- nnetar(test, model=neural_model)
> accuracy(fitted(neural_model_test), test)
      ME      RMSE      MAE      MPE      MAPE      ACF1 Theil's U
Test set 1836.834 2032.883 1836.834 17.3441 17.3441 0.2741976 1.355795
```

Şekil 3-36 Yapay Sinir Ağları Modeli Hata Metrikleri

Yapay sinir ağları modelinin test verileri üzerindeki performansı üstel düzleştirme, ARIMA ve regresyon modelleriyle RMSE ve MAE hata kriterleri dikkate alınarak karşılaştırıldığında ortalama başarıya sahip olduğu görülmektedir.

4. SONUÇLAR

Bu tez çalışmasında bilgi güvenliği alanında alınan patentlerin patent sınıflarına ait patent sayıları tahminlemesi yapılmıştır. Böylelikle gelecekte bilgi güvenliği alanında eğilimin hangi yöne doğru olduğuna dair doğru bir tahminleme için kullanılabilecek bir veri analizi çalışması gerçekleştirilmiştir.

Ağ analizi yapılırken bu çalışma kapsamında hazırlanan veri seti kullanılmıştır. Veri seti uygun bir biçimde CSV dosyalarına kaydedilmiş ve bu dosyalar Gephi ve Cytoscape'e aktararak amaçlar doğrultusunda ağlar oluşturulmuştur. Öncelikle 2009 yılından 2019 yılına kadar her yıl için ayrı ağlar oluşturulmuş olup sonrasında tüm yılların verilerini kapsayan genel bir ağ oluşturulmuştur. Elde edilen ağlarla çeşitli analizler yapılmıştır.

Yıl bazında yapılan ağ analizinde derece merkezilik ölçütünü kullanarak ağlardaki en merkezi düğümler saptanmıştır. Analiz sonuçlarına göre ABD ve Çin'in bilgi güvenliği alanında en çok patent alan ülkeler olduğu görülmektedir. Yıl geçtikçe bu alanda alınan patentlerin sayısı da artmaktadır. ABD ve Çin'i kendi aralarında karşılaştırdığımızda Çin'in önde olduğu görülmektedir. En çok tercih edilen patent sınıfları ise G06F ve H04L sınıflarıdır. Son on yılın ilk yıllarında G06F sınıfı en çok başvuru alan sınıf olurken son yıllarda H04L sınıfı daha çok tercih edilmiştir. İki sınıf arasında karşılaştırma yapıldığında H04L sınıfının açık ara önde olduğu görülmektedir.

Son on yılın verileriyle oluşturulan ağ analizi sonuçlarına göre Çin'in en fazla patent alan ülke, H04L sınıfının ise en çok patent alınan sınıf olduğu görülmektedir.

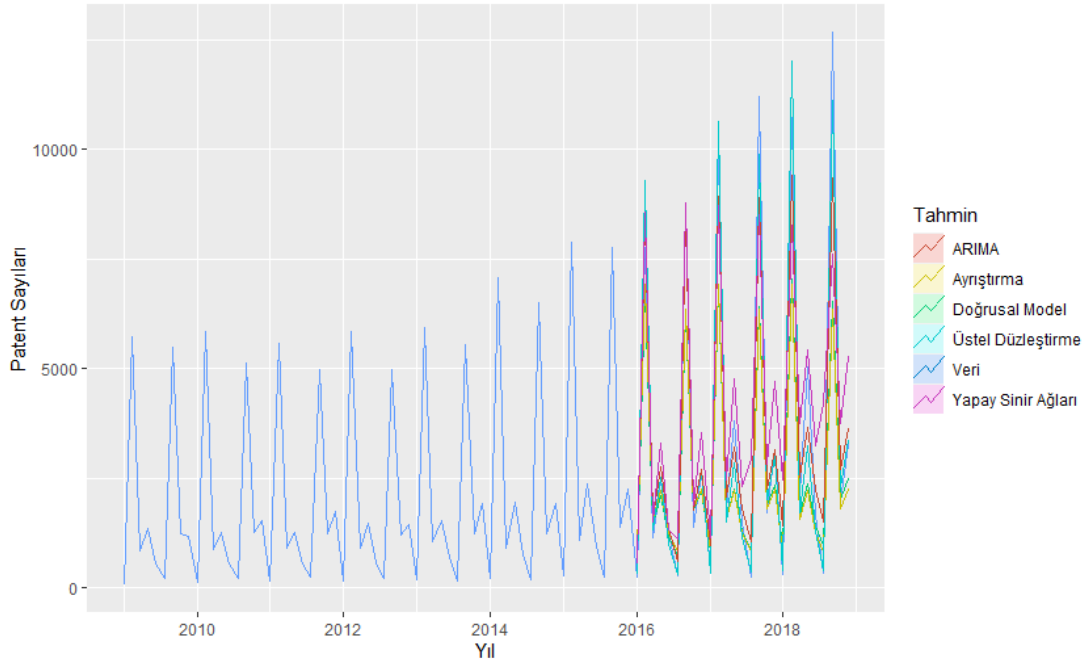
Patent sınıflarına ait patent sayıları tahminlemesi yapılırken zaman serileri üzerinde de durulmuş, tahminleme modellerinden şunlar kullanılmıştır; regresyon, ayırıştırma, ETS, ARIMA ve yapay sinir ağları. Bu modeller ile patent sınıflarına ait patent sayılarından geleceğe dönük tahminler yapılmıştır. Modellerin test verileri üzerindeki performansları, karekök ortalama hata (RMSE) ve ortalama mutlak hata (MAE) metrikleri değerleriyle beraber aşağıdaki tabloda verilmiştir.

Tablo 4-1 Patent Sınıflarına Ait Patent Sayıları Zaman Serisine Uygulanan Modellerin Hata Kriterleri

Kullanılan Model	Karekök Ortalama Hata	Ortalama Mutlak Hata
Doğrusal Regresyon Modeli	1. 953,1295	1.209,6612
Ayrıştırma Modeli	1.901,5270	1.562,0105
Holt-Winters Mevsimsel Yöntemi ile ETS Modeli	716,0642	476,1914
ARIMA(4,1,1)(1,0,1)[9] ARIMA Modeli	1.014,454	727,2240
ARIMA(1,0,1)(2,0,0)[9] ARIMA Modeli	1.079,830	675,6613
NNAR(9,1,5)[9] Yapay Sinir Ağları Modeli	2.032,883	1.836,834

Hata kriterlerine göre Holt-Winters Mevsimsel Yöntemi ile ETS modeli test verileri üzerinde en başarılı performansa sahip model olarak görülmektedir. Geride kalan modellerin performans sıralaması ARIMA, ayrıştırma, regresyon ve yapay sinir ağları şeklindedir. Yapay sinir ağları modeli performans olarak en başarısız model olmuştur.

Patent sınıflarına ait patent sayıları üzerinde uygulanan modellerin tahminlerine ait grafik aşağıdaki gibidir.



Şekil 4-1 Patent Sınıflarına Ait Patent Sayıları Zaman Serisine Uygulanan Modellerin Tahminleri

Modellerin test verileri üzerindeki performansı eğitim verileri üzerindeki performansı ile aynı olmayabilir. Eğitim ve test sınıflarının kapsadığı yıllar değiştirilip farklı sonuçlar elde edilebilir. Patent sınıflarına ait patent sayılarını tahminleme modellerinde aynı model tüm verileri kullanılarak eğitilmiş ve 2019, 2020 yıllarının dokuz ayrı sınıfına ait tahminler aşağıdaki tabloda verilmiştir.

Tablo 4-2 Tahminleme Modellerinin İki Senelik Patent Sayıları Öngörülleri

Yıl-Sınıf	Regresyon	Ayrıştırma	ETS	ARIMA	YSA
2019-C12N	1628.652	1601.431	424.9793	625.3799	467.9343
2019-G06F	8649.652	8592.518	13612.3315	11013.4725	11439.3282
2019-G06K	2573.552	2487.798	2513.4347	2547.0941	2345.4035
2019-G06Q	3661.452	3551.905	5463.7002	5579.9308	6365.9035
2019-G08B	2308.852	2177.305	1896.8086	2081.2479	1703.5651

(Tablo 4-2 Devam)

2019-G10L	1650.552	1499.831	387.5096	755.5892	697.6186
2019-H04L	8699.152	8519.238	14113.4164	13133.6708	13039.8644
2019-H04N	2798.152	2585.674	2225.5053	2381.5515	2347.4046
2019-H04W	3518.052	3274.000	3872.9867	3631.6293	3655.5321
2020-C12N	1885.352	1601.431	492.5803	1113.8703	570.7361
2020-G06F	8906.352	8592.518	15752.6026	11351.4564	11639.1518
2020-G06K	2830.252	2487.798	2904.1306	3006.8745	2256.0726
2020-G06Q	3918.152	3551.905	6303.4990	5994.3746	7548.3228
2020-G08B	2565.552	2177.305	2185.1510	2553.1409	1658.2150
2020-G10L	1907.252	1499.831	445.7791	1246.4462	1358.9073
2020-H04L	8955.852	8519.238	16213.0372	13443.8980	12947.4974
2020-H04N	3054.852	2585.674	2553.1175	2849.2044	2748.4048
2020-H04W	3774.752	3274.000	4437.2367	4081.4937	4477.6611

Çalışmanın ilk bölümünde toplanan veri setinde 2019 yılının ilk altı ayına ait patent sayıları C12N sınıfı için 126, G06F sınıfı için 7088, G06K sınıfı için 1256, G06Q sınıfı için 3698, G08B sınıfı için 896, G10L sınıfı için 288, H04L sınıfı için 8100, H04N sınıfı için 1140 ve H04W sınıfı için 2190 adettir. Bu sayıları en iyi yakalayan model ise ETS modelidir. Ortalama yüzde hata formülü ile modelin tahmin ettiği değerler ile gerçek değerler arasındaki farkın ortalama yüzdesi bulunmuş ve bu model ile 2019 yılı ilk altı ayı için bilgi güvenliği alanında alınan patentlerin patent sınıflarına ait patent sayılarının %98 doğrulukla hesaplanabildiği görülmüştür. Diğer modellerin performans sıralamaları yapay sinir ağları, regresyon, ayrıştırma ve ARIMA'dır. En başarısız model ARIMA olmuştur. En başarılı model olan ETS'nin 2020 yılı tahminlerine baktığımızda H04L sınıfının en çok başvurulacak sınıf olduğunu görülmektedir. G06F sınıfı ise H04L'den sonra en çok başvuru alacak sınıf olarak karşımıza çıkmaktadır.

Hem ağ analizi hem de zaman serisi analizi sonuçlarına baktığımızda bilgi güvenliği alanında alınan patentlerin sınıflarında ilk sırada H04L'nin ikinci sırada ise G06F'nin olduğu görülmektedir. H04L sınıfı dijital bilgilerin aktarımı ile ilgili patentlerin sınıflandırıldığı IPC kodu olup H elektrik ana sınıfı kodunun altında yer alan H04 elektrik iletişim tekniği grubunun bir alt grubudur. Bu alt sınıf dijital formatta sağlanan sinyalleri kapsar ve veri aktarımı, telgraf iletişimi ya da gözlemleme için kullanılan düzenlemeleri ya da methodları içerir. 14 alt sınıfı vardır. 14 alt sınıfın kapsadığı toplam 285 alt sınıf daha bulunmaktadır [48].

G06F sınıfı elektrikli dijital veri işleme ile ilgili patentlerin toplandığı sınıf olup G fizik sınıfının altındaki G06; bilgi işlem, hesaplama ve sayım sınıfına dahildir. Bu alt sınıf verilerin işlenmesi, taşınması ve veri işleme ekipmanları ile ilgili patentleri ilgilendirir. 12 alt sınıfı vardır. 12 alt sınıfın kapsadığı 362 alt sınıf daha bulunmaktadır [49].

Bu çalışmayla bilgi güvenliği alanında çalışan firmalar yatırımlarını geleceğin teknolojisi olan dijital bilgilerin aktarımı ile ilgili alanlara yönlendirebilirler. Örneğin; internet üzerinden bilgi aktarımı, telefon ve telgraf ile iletişim, yangın ve polis telgrafları, kendi başına şifreleme veya deşifre eden aparatlar, kod çözme ve kod dönüştürme gibi genellikle iletişim tabanlı alanlara yatırım yapabilirler.

4.1 ÖNERİLER

Patent analizi üzerine yapılan çalışmalar incelendiğinde amacın genel olarak bu çalışma ile aynı olduğu görülmektedir. Ayrıca bu çalışmada patentleri toplamak için seçilen veri tabanı kapsamından dolayı literatürde en çok tercih edilen veri tabanlarından bir tanesidir. INPADOC veritabanında yayınlanan patentleri kullanarak yapılan analizlerle bu çalışmada başarı elde edilmiştir. INPADOC veri tabanı ESPACENET veritabanıyla ilişkili olduğundan ve ESPACENET veri kapsamı ve müşteri desteği açısından diğer veri tabanlarına göre daha yüksek puana sahip olduğu için tercih edilmiştir.

Patentler sınıflandırma sistemi sayesinde teknik alanlarına göre ayrılmıştır ve böylelikle çeşitli alanlardaki teknolojik gelişmelerin değerlendirilmesi yapılabilmektedir. Daha doğru sonuçlar elde etmek için patent analizi çalışmalarında genellikle IPC kodları kullanılmıştır. Bu çalışmada da teknolojinin gitmesi muhtemel yönünü tahmin etmek için

patent sınıflarından yararlanılmıştır. Detaylı alt sınıflara ayrılan IPC yapısı, teknolojileri tespit etmede çalışmayı daha hassas hale getireceğine inanıldığı için seçilmiştir.

Mevcut çalışmalarda kullanılan yöntemlere baktığımızda genellikle ağ analizi [50], kümeleme yöntemleri [51] ve sınıflandırma ağacı yapısı [52] kullanılarak bir model geliştirilmiştir. Buralarda kullanılan yöntemler yoğunlaşmanın daha çok olduğu patentleri bulmakla beraber zaman serisi analizi gibi geçmişini anlayıp geleceği tahmin eden yöntemlerden değildir. Bu çalışmada sosyal ağ analizine ek olarak kullanılan zaman serisi analizi yöntemi tahminin doğruluğunu arttırmıştır. Ayrıca, bu çalışmada kullanılan zaman serisi analizi yönteminde frekans değerine sınıf sayısı vererek yapılan tahminleme diğer çalışmalardan farklılığını göstermektedir.

Bunlara ek olarak; patent analizi dünyada son yıllarda popüler bir yöntem olmasına rağmen Türkçe yayın sayısının azlığı bu çalışmaya ayrıca önem kazandırmaktadır.

Şimdiye kadar bilgi güvenliği alanında patent analizi yapan çalışmaların kısıtlı olması bu çalışmanın yeni bir bakış açısı getirmesi açısından önemlidir ve bu alandaki gelişen teknolojileri göstermiştir.

Patent analiziyle yapılan teknoloji tahmini şirketlere yatırım yapması gereken alanlar konusunda yön verebilir ve rakip firmalar birbirlerini izleyebilir. Ülkelerin kendilerine stratejik teknoloji ortakları bulması konusunda yardımcı olabilir. Hangi ülkenin hangi alanda daha çok teknolojik buluş yaptığını görerek bireylerin kariyerlerine yön çizmesine bile etki edebilecek küresel bir bakış açısı sunmaktadır.

Bunların yanısıra, tahmin edilen teknolojinin kapsamını daraltmak için IPC grup ya da alt grupları kullanılarak çalışmanın doğruluğu artırılabilir. Ayrıca çalışmaya ekstra fon sağlayarak ESPACENET'in kapsamadığı patentler ücretli veri tabanlarından toplanıp çalışmanın kapsamı genişletilebilir. Bunlara ek olarak farklı alanlar seçilip bu alanlardaki teknolojik tahminlemeler yapılabilir. Bloglar ve farklı sosyal ağlar gibi diğer veri kaynakları çalışmaya dahil edilerek kapsam genişletilebilir.

KAYNAKLAR

- [1] The Evolution of Information Security, <https://www.securityweek.com/evolution-information-security> Erişim Tarihi: 19/08/2019
- [2] Espacenet Patent Search, <https://www.epo.org/searching-for-patents/technical/espacenet.html> Erişim Tarihi: 10/05/2019
- [3] Google Patents, <https://patents.google.com/> Erişim Tarihi: 10/05/2019
- [4] United States Patent and Trademark Office, <https://www.uspto.gov/> Erişim Tarihi: 10/05/2019
- [5] EPO Worldwide Legal Event Data (INPADOC), <https://www.epo.org/searching-for-patents/data/bulk-data-sets/inpadoc.html> Erişim Tarihi: 10/05/2019
- [6] European Patent Office, <https://www.epo.org/> Erişim Tarihi: 10/05/2019
- [7] Lingua, D. G. (2005). INPADOC: 30 years of endeavours yet unmapped territories remain!. World Patent Information, 27(2), 105-111
- [8] Li, X., Xie, Q., Jiang, J., Zhou, Y., & Huang, L. (2019). Identifying and monitoring the development trends of emerging technologies using patent analysis and Twitter data mining: The case of perovskite solar cell technology. Technological Forecasting and Social Change, 146, 687-705.
- [9] Chen, J., Wei, W., Guo, C., Tang, L., & Sun, L. (2017). Textual analysis and visualization of research trends in data mining for electronic health records. Health Policy and Technology, 6(4), 389-400.
- [10] Cho, H. P., Lim, H., Lee, D., Cho, H., & Kang, K. I. (2018). Patent analysis for forecasting promising technology in high-rise building construction. Technological Forecasting and Social Change, 128, 144-153.
- [11] Kim, G., & Bae, J. (2017). A novel approach to forecast promising technology through patent analysis. Technological Forecasting and Social Change, 117, 228-237.
- [12] Jun, S., & Park, S. S. (2013). Examining technological innovation of Apple using patent analysis. Industrial Management & Data Systems.
- [13] Nijhof, E. (2018). Want to find? Break the rules!. World Patent Information, 52, 22-25.

- [14] Alp, B. (2016) A Technology Mining System Based On Patent Documents: A Case Study of Cloud Computing. Yüksek Lisans Tezi, Bahçeşehir Üniversitesi Fen Bilimleri Enstitüsü, İstanbul, Türkiye, 12-58.
- [15] Altuntaş, F., & YILMAZ, M. K. (2017). Patent analizi ile teknoloji ağlarının oluşturulması. Girişimcilik ve İnovasyon Yönetimi Dergisi, 6(2), 97-129.
- [16] Aithal, P. S., & Aithal, S. (2018). Patent Analysis as a New Scholarly Research Method. International Journal of Case Studies in Business, IT, and Education (IJCSBE),(2018), 2(2), 33-47.
- [17] Rodríguez, J. C., & Gómez, M. (2014). Innovation trends in Latin America countries. International Journal of Economics and Management Engineering, 8(3), 638-644.
- [18] Girdher, H., & Gaur, P. (2017). Text Mining Techniques-A Review. Global Journal of Engineering Science and Researchers: 4 pp, (6), 68-73.
- [19] Song, K., & Lee, S. (2017). Development of a Technology Assessment Model by Patents and Customers' Review Data. World Academy of Science, Engineering and Technology, International Journal of Computer, Electrical, Automation, Control and Information Engineering, 10(3), 663-70.
- [20] Abbas, A., Zhang, L., & Khan, S. U. (2014). A literature review on the state-of-the-art in patent analysis. World Patent Information, 37, 3-13.
- [21] Altuntas, S., Dereli, T., & Kusiak, A. (2015). Forecasting technology success based on patent data. Technological Forecasting and Social Change, 96, 202-214.
- [22] Başpınar, C. (2008). İmalat Sanayii Patent Aktivitelerinin Sektörel Analizi. Uzmanlık Tezi Türk Patent Enstitüsü Patent Dairesi Başkanlığı, Ankara.
- [23] European Patent Office, <https://patent-client.readthedocs.io/en/latest/epo.html> Erişim Tarihi: 25/05/2019
- [24] Espacenet: Free Access to Over 120 Million Patent Documents, <https://worldwide.espacenet.com/> Erişim Tarihi: 27/05/2019
- [25] Tunalı, V. (2016) Sosyal Ağ Analizine Giriş, 1. Baskı, . Nobel Yayınları, Ankara, Türkiye.
- [26] Çelik, S. (2019). Dünyadaki Çatışmaların Sosyal Ağ Analizi Yöntemiyle İncelenmesi. Öneri, 14(52), 236.

- [27] Gençer, M. (2017). Sosyal Ağ Analizi Yöntemlerine Bir Bakış. Yildiz Social Science Review, 3(2), 19-34.
- [28] Box, G. E., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. (2015). Time series analysis: forecasting and control. John Wiley & Sons.
- [29] Velicer, W. F., & Plummer, B. A. (1998). Time series analysis in historiometry: a comment on Simonton. Journal of personality, 66(3), 477–493.
- [30] Tezuran, M. (2019) Çağrı Merkezi Performans Analizi Ve Çağrı Sayılarının Mevsimsel Tahminlemesi. Yüksek Lisans Tezi, Marmara Üniversitesi Fen Bilimleri Enstitüsü, İstanbul, Türkiye, 8-9.
- [31] Zaman Seri Analizinde Temel Kavramlar, http://kisi.deu.edu.tr/hamdi.emec/UygEko/ZAMAN_SERI_ANALIZINDE_TEMEL_KAVRAMLAR.pdf Erişim Tarihi: 01/04/2020
- [32] Zorlutuna, Ş., & Bircan, H. (2019). Türkiye’ye Gelen Turist Sayısı Tahmininde Zaman Serileri Analizi Ve Yapay Sinir Ağları Yöntemlerinin Karşılaştırılması. Cumhuriyet Üniversitesi İktisadi ve İdari Bilimler Dergisi, 20(2), 164-185.
- [33] 3.3 Residual Diagnostics, <https://otexts.com/fpp2/residuals.html> Erişim Tarihi: 03/05/2020
- [34] Can, M., & Orhunbilge, D. N. (2009). İşletmelerde Zaman Serileri Analizi ile Tahmin. İstanbul Üniversitesi Sosyal Bilimler Enstitüsü Doktora Tezi, İstanbul.
- [35] Thomas, R. L. (1997) Modern Econometrics, Addison Wesley
- [36] Cleveland, R. B. STL: A seasonal-trend decomposition procedure based on loess. 1990. DOI: citeulike-article-id, 1435502.
- [37] STL Decomposition, <https://otexts.com/fpp2/stl.html> Erişim Tarihi: 15/05/2020
- [38] Exponential Smoothing, <https://otexts.com/fpp2/expsmooth.html> Erişim Tarihi: 17/05/2020
- [39] Üstel Düzleştirme Yöntemi, http://kisi.deu.edu.tr/hamdi.emec/Ekonometri%20Semineri/USTEL_DUZLESTIRME_YONTEMLER.pdf Erişim Tarihi: 20/05/2020
- [40] ARIMA Models for Time Series Forecasting, <https://people.duke.edu/~rnau/411arim.htm> Erişim Tarihi: 20/05/2020
- [41] Non-Seasonal ARIMA Models, <https://otexts.com/fpp2/non-seasonal-arima.html> Erişim Tarihi: 22/05/2020

- [42] Non-Seasonal ARIMA Models, <https://otexts.com/fpp2/seasonal-arima.html>
Eriřim Tarihi: 22/05/2020
- [43] Yapay Sinir Ağları, http://www.papatya.gen.tr/PDF/yapay_sinir_aglari.pdf
Eriřim Tarihi: 03/06/2020
- [44] Can, ř., & Gerřil, M. (2018). Manisa Pamuk Fiyatlarının Zaman Serisi Analizi Ve Yapay Sinir Ağı Teknikleri İle Tahminlenmesi Ve Tahmin Performanslarının Karşılaştırılması. Yönetim ve Ekonomi: Celal Bayar Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi, 25(3), 1017-1031.
- [45] Zorlutuna, ř., & Bircan, H. Türkiye'ye Gelen Turist Sayısı Tahmininde Zaman Serileri Analizi Ve Yapay Sinir Ağları Yöntemlerinin Karşılaştırılması. Cumhuriyet Üniversitesi İktisadi ve İdari Bilimler Dergisi, 20(2), 164-185.
- [46] Pekarčík, P., Gajdoř, A., & Sokol, P. (2020, November). Forecasting Security Alerts Based on Time Series. In International Conference on Hybrid Artificial Intelligence Systems, 546-557, Springer, Cham.
- [47] What is a Box Cox Transformation?, <https://www.statisticshowto.datasciencecentral.com/box-cox-transformation/> Eriřim Tarihi: 01/08/2020
- [48] H Electricity,
<https://www.wipo.int/classifications/ipc/en/ITsupport/Version20180101/transformations/ipc/20180101/en/htm/H04L.htm> Eriřim Tarihi: 03/09/2020
- [49] G Physics,
<https://www.wipo.int/classifications/ipc/en/ITsupport/Version20170101/transformations/ipc/20170101/en/htm/G06F.htm> Eriřim Tarihi: 03/09/2020
- [50] Kwon J., & Jeong D. (2014). Technology Relevance Analysis between Wind Power Energy - Fuel Cell - Green Car using Network Analysis, IPC Map. COLLNET Journal of Scientometrics and Information Management, 8:1, 109-121.
- [51] Kim, K. H., Han, Y. J., Lee, S., Cho, S. W., & Lee, C. (2019). Text Mining for Patent Analysis to Forecast Emerging Technologies in Wireless Power Transfer. Sustainability, 11, 6240.
- [52] Lee, S. J., & Jun, S. (2014). Key IPC codes extraction using classification and regression tree structure. In Soft Computing in Big Data Processing, 101-109, Springer, Cham.

ÖZGEÇMİŞ

Adı Soyadı : Hatice IŞIK ÖZATA
Doğum Yeri ve Tarihi : Kütahya - 1988
E-posta : hisikozata@gmail.com

Öğrenim Durumu

Derece	Bölüm/Program	Üniversite/Lise	Mezuniyet Yılı
Lise	Matematik/Fen	Tavşanlı İMKB Anadolu Öğretmen Lisesi	2006
Lisans	Bilgisayar Mühendisliği/Burslu (İngilizce)	İzmir Ekonomi Üniversitesi	2011
Yüksek Lisans	Bilgisayar Mühendisliği	Marmara Üniversitesi	-

İş Deneyimi

Yıl	Firma/Kurum	Görevi
2011-2014	Vestel Elektronik A.Ş.	Kıdemli Tasarım Doğrulama ve Test Mühendisi
2014 - 2015	Nafi Güral Yatırım A.Ş.	Yazılım Mühendisi
2017-	TÜBİTAK BİLGEM Kamu Sertifikasyon Merkezi	Araştırmacı / Lider Test Mühendisi