TÜRKİYE CUMHURİYETİ
ANKARA ÜNİVERSİTESİ
SAĞLIK BİLİMLERİ ENSTİTÜSÜ

FONKSİYONEL VERİ ANALİZİ İLE YAŞA ÖZEL DOĞURGANLIK HIZLARININ TAHMİN EDİLMESİ

Mustafa ERDOĞAN

BİYOİSTATİSTİK ANABİLİM DALI
YÜKSEK LİSANS TEZİ

DANIŞMAN
Doç. Dr. Beyza DOĞANAY ERDOĞAN

ANKARA
2020

TÜRKİYE CUMHURİYETİ
ANKARA ÜNİVERSİTESİ
SAĞLIK BİLİMLERİ ENSTİTÜSÜ

FONKSİYONEL VERİ ANALİZİ İLE YAŞA ÖZEL DOĞURGANLIK HIZLARININ TAHMİN EDİLMESİ

Mustafa ERDOĞAN

BİYOİSTATİSTİK ANABİLİM DALI
YÜKSEK LİSANS TEZİ

DANIŞMAN
Doç. Dr. Beyza DOĞANAY ERDOĞAN

ANKARA
2020

Etik Beyan

Ankara Üniversitesi

Sağlık Bilimleri Enstitüsü Müdürlüğü'ne,

Yüksek Lisans tezi olarak hazırlayıp sunduğum "Fonksiyonel Veri Analizi ile Yaşa Özel Doğurganlık Hızlarının Tahmin Edilmesi" başlıklı tez; bilimsel ahlak ve değerlere uygun olarak tarafımdan yazılmıştır. Tezimin fikir/hipotezi tümüyle tez danışmanım ve bana aittir. Tezde yer alan çalışma tarafımdan yapılmış olup, tüm cümleler, yorumlar bana aittir.

Yukarıda belirttiğim hususların doğruluğunu beyan ederim.
Öğrencinin Adı Soyadı: Mustafa ERDOĞAN
Tarih: 11 Aralık 2020
İmza

KABUL VE ONAY

İÇİNDEKİLER

Etik Beyan	ii
Kabul ve Onay	iii
İçindekiler	iv
Önsöz	v
Simgeler ve Kısaltmalar	vi
Şekiller	vii
Çizelgeler	viii
1.GİRİŞ	1
1.1. Fonksiyonel Veri Analizi	2
1.2.Fonksiyonel Veri	3
1.3.Fonksiyonel Verinin Düzgünleştirilmesi	5
1.3.1.Baz Fonksiyonları	6
1.3.1.1.Fourier Baz Fonksiyonları	7
1.3.1.2.Splayn Baz Fonksiyonları	8
1.3.1.3.B-Splayn Baz Fonksiyonları	9
1.3.2.Tahmin Yöntemleri	11
1.3.2.1.En Küçük Kareler Yöntemi	11
1.3.2.2.Pürüzlü Ceza Yaklaşımı	12
1.4.Klasik İstatistiksel Yöntemlerin Fonksiyonel Verilere Uyarlanması	13
1.4.1.Fonksiyonel Temel Bileşenler Analizi	14
1.4.2.Fonksiyonel Doğrusal Modeller	16
2.GEREÇ VE YÖNTEM	19
2.1.Uygulama Verileri	20
2.2.Yöntem	20
3.BULGULAR	28
4.TARTIŞMA	33
5.SONUÇ VE ÖNERİLER	36
ÖZET	37
SUMMARY	38
KAYNAKLAR	39
EKLER	42
Ek-1 Analize Ait R Kodları	42
ÖZGEÇMİŞ	44

ÖNSÖZ

Sağlık alanındaki çalışmalarda sıklıkla kullanılması önerilen fonksiyonel veri analizini Türkiye verileri üzerinde uygulayan az sayıda çalışma bulunmaktadır. Bu çalışma, Türkiye yaşa bağlı doğurganlık hızı verileri için fonksiyonel veri analizi kullanılarak tahmin edilmiştir.

Yüksek lisans eğitimim süresince ve tez aşamasında, benden ilgisini ve desteğini eksik etmeden sabırla rehberlik eden değerli danışman hocam Doç. Dr. Beyza DOĞANAY ERDOĞAN'a,

İlgi ve alakalarından dolayı Prof. Dr. Atillah Halil Elhan'a, Doç. Dr. Derya Gökmen'e, ve Doç. Dr. Serdal Kenan Köse'ye

Ayrıca destekleri ile her zaman yanımda olan ve bugünlere gelmemde büyük pay sahibi sevgili aileme ve dostlarıma teşekkür ederim.

SİMGELER VE KISALTMALAR

ABD	Amerika Birleşik Devletleri
AEKK	Ağırlıklı En Küçük Kareler
ANOVA	Tek Değişkenli Varyans Analizi
ARIMA	Otoregresif Tümleşik Hareketli Ortalamalar Modeli
$\boldsymbol{c}$	Katsayı Değerleri
CHKT	Cezalandırılmış Hata Kareler Toplamı
EKK	En Küçük Kareler
Φ	Baz Fonksiyonu
FANOVA	Fonksiyonel Tek Değişkenli Varyans Analizi
FTBA	Fonksiyonel Temel Bileşenler Analizi
FVA	Fonksiyonel Veri Analizi
GCV	Genelleştirilmiş Çapraz Geçerlilik
HKT	Hata Kareler Toplamı
ISFE	Entegre Kare Tahmin Hatası
K	Baz Sayısı
λ	Düzgünleştirme Parametresi
LC	Lee-Carter Yöntemi
m	Mertebe Sayısı
MERNİS	Merkezi Nüfus İdaresi Sistemi
MISE	Ortalama Entegre Kare Hatası
$\boldsymbol{PEN}$	Cezalandırma Parametresi
TBA	Temel Bileşenler Analizi
TÜİK	Türkiye İstatistik Kurumu
YÖDH	Yaşa Özel Doğurganlık Hızı

ŞEKİLLER

Şekil 1.1. Örnek Fonksiyonel Veri 5
Şekil 1.2. Fourier Bazı Örneği 7
Şekil 1.3. B-Splayn Baz Fonksiyonu Örneği 10
Şekil 3.1. Türkiye Yaşa Göre Doğurganlık Hızları (2001-2018) 29
Şekil 3.2. Türkiye Yaşa Göre Düzgünleştirilmiş Doğurganlık Hızları (2001-2018) 29
Şekil 3.3. Türkiye Yaşa Göre Uydurulmuş Doğurgalık Hızları 30
Şekil 3.4. Uydurulmuş Baz Fonksiyonları ve Katsayıları 30
Şekil 3.5. Tahmin Edilen Yaşa Göre Doğurganlık Hızları (2019-2023) 31
Şekil 3.6. Uydurulmuş ve Tahmin Edilen Baz Fonksiyonları ve Katsayıları 32

ÇİZELGELER

Çizelge 1.1. Fonksiyonel Doğrusal Regresyon Modelleri 16

1. GİRİŞ

Teknolojinin gelişmesiyle birlikte farklı tiplerde daha fazla veri toplanması mümkün hale gelmektedir. Yeni veri tipleri ise yeni istatistiksel analiz yöntemlerini gerektirmektedir. Bu kapsamda fonksiyonel veri analizi (FVA) önem kazanmakta ve yapılan çalışmalarda kullanımı artmaktadır. En genel tanımıyla fonksiyonel veri analizi, bir süreklilik üzerinde tanımlanan ya da sürekli yapıya sahip olan degişkenlerin analizi için kullanılan yöntemleri ifade etmektedir. Söz konusu verilerin analizinde klasik çok değişkenli yöntemlerin kullanılması da mümkün olmakla birlikte fonksiyonel veri analizi, fonksiyonel veride yer alan ve klasik yöntemler ile elde edilemeyen bilgilerin de kullanılmasına olanak sağlamaktadır. Aslında, örneklem uzayının sonsuz boyutlu olması sebebiyle fonksiyonel veri analizi, istatistiksel teorinin yeni bir dalı olarak kabul edilmektedir.

Tıp, ekonomi, doğa bilimleri gibi çoğu araştırma dalının yanı sıra, sağlık alanında da veriler farklı zaman noktalarında ya da bir alan üzerindeki farklı noktalardan tekrarlı ölçümler yapılarak elde edilebilmektedir. Benzer şekilde, fonksiyonel veri analizinin önem kazandığı uygulama alanlarından biri de sağlık, biyomedikal ve biyoinformatik alanlarıdır. Bu kapsamda, bu tez çalışmasının amacı, fonksiyonel veri analizini ve temel araçları tanıtmak, bu analiz yöntemini kullanarak Türkiye'de yaşa özel doğurganlık hızı verilerinin tahmini uygulaması yapmaktır. Çalışmada 2001-2018 yılı yaşa bağlı doğurganlık hızı verileri kullanılacaktır.

Sağlık alanındaki çalışmalarda sıklıkla kullanılması önerilen fonksiyonel veri analizini Türkiye verileri üzerinde uygulayan az sayıda çalışma bulunmaktadır. Bu çalışma, Türkiye sağlık verileri için bir tahminin fonksiyonel veri analizi kullanılarak yürütüleceği ilk çalışmalardan olup yöntemin kullanılabilirliğini gösterme ve ileriki çalışmalar için öncülük etme amacındadır.

1.1. Fonksiyonel Veri Analizi

Fonksiyonel veri analizi terimi Ramsay (1982) ve Ramsay ve Dalzell (1991) tarafından ortaya atılmış olsa da aslında bu çalışma alanı çok daha eskiye dayanmaktadır (Grenander (1950), Rao (1958), Wang ve ark. 2015). Fonksiyonel veri analizi, nokta olarak tanımlanan gözlemlerle ilgilenen klasik istatistiksel analizden farklı olarak, sürekli fonksiyonlar şeklinde tanımlanan gözlemlerle ilgilenmektedir. Diğer bir deyişle, bu analizde, gözlenen veri fonksiyonlarının yalnızca ayrık gözlem değerlerinin bir dizisi gibi değil, örneklemdeki tekil elemanlar olarak ele alınması gerekir (Ramsay ve Silverman, 2005). Analizde kullanılan örneklem uzayı da vektör uzayı değil, sonsuz boyutlu bir fonksiyonel uzaydır. Örneklem uzayının sonsuz boyutlu olması sebebiyle fonksiyonel veri analizi, istatistiksel teorinin yeni bir dalı olarak kabul edilmektedir (Cuevas, 2014).

Ramsay ve Silverman (2005), fonksiyonel veri analizinin temel amaçlarını aşağıdaki şekilde tanımlamıştır:

- Verinin ileri analizlere uygun şekilde işlenmesi/dönüştürülmesi,
- Veriye ait önemli karakteristikler ortaya çıkacak şekilde verinin görselleştirilmesi,
- Verideki örüntü ve değişimin çalışılması,
- Bir yanıt ya da bağımlı değişkendeki değişkenliğin bağımsız değişkenlere ait bilgiler kullanılarak açıklanması,
- Aynı ya da farklı fonksiyonlar içeren iki ya da daha fazla sayıda veri setinin ortak tekrar kümeleri açısından karşılaştırılması.

Fonksiyonel veri analizi temel olarak önce zaman serisi oluşturan ayrık/noktasal gözlemlerin bir fonksiyon oluşturacak şekilde ifade edilmesi, sonrasında ise oluşturulan fonksiyonel veri seti üzerinde çok değişkenli analiz yöntemlerinden uygun olanların uygulanarak modelleme ve tahmin çalışmaları yürütülmesini kapsamaktadır.

Bu çerçevede, çok değişkenli istatistiksel analizden fonksiyonel veri analizine uyarlanan bazı yöntemler fonksiyonel tanımlayıcı istatistik, fonksiyonel temel bileşenler analizi, fonksiyonel doğrusal modeller ve fonksiyonel kanonik korelasyon analizidir (Ramsay ve Silverman, 2005). Ancak, yürütülecek bir çalışmada FVA'nın bu özelliklerinin tamamının kullanılması gerekmemektedir; hangi özelliklerinin kullanılacağı verinin yapısı ve araştırmanın niteliğine uygun şekilde belirlenmelidir.

1.2. Fonksiyonel Veri

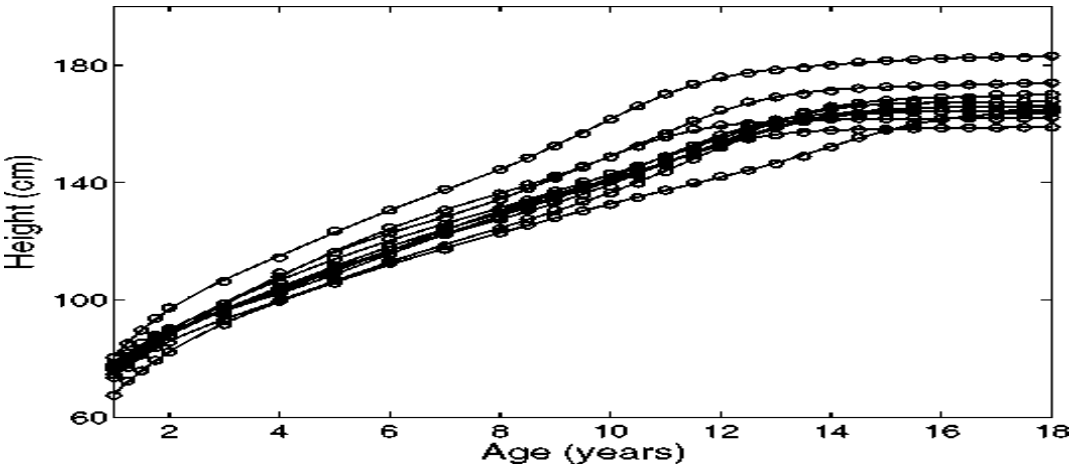
Fonksiyonel veri setinin temel öğesi bir fonksiyondur; rastgele bir örneklemdeki her gözlem için bir veya birkaç fonksiyon kaydedilmektedir (Wang et al, 2015). Başka bir deyişle, bir fonksyonel veri, tek bir gözlem değil; bir uzay üzerinde kaydedilen ve tek bir eleman olarak kabul edilen ölçümler setidir. Burada uzay genelde zaman olmakla birlikte farklı hatta çok boyutlu da olabilmektedir (Levitin ve ark., 2007).

Fonksiyonel veri örneklemi, I=[0,T] gibi bir kompakt küme üzerinde tanımlanmış bağımsız gerçel $x(t_1), x(t_2), \dots, x(t_n)$ fonksiyonlarından oluşmaktadır. Fonksiyonel gözlemler ise n tane (t_j, y_j) çiftleri olarak ifade edilir (j=1,...n). Bu veriler kullanılarak $x(t_j)$'nin ölçülmemiş zaman noktalarındaki fonksiyon veya türev değerleri gibi karakteristikleri tahmin edilmektedir (Levitin et al, 2007). Burada t_j, ölçümün alındığı zaman noktasını; $x(t_j)$, arkaplandaki sürekli fonksiyonun, diğer bir deyişle karakteristiklerin değerini; y_j ise $x(t_j)$ fonksiyonunun genellikle hata içeren bir anlık görüntüsünü ifade etmektedir. Yukarıda bahsedildiği üzere, ölçümlerin alındığı t_j argümanlarının zaman noktası olması zorunlu değildir; yaş, boy, konum gibi devamlılık gerektiren durumlarda da ölçümler alınabilir. Ayrıca, t_j değerleri tüm ölçüm noktaları için aynı olabileceği gibi kayıttan kayda değişiklik de gösterebilir (Ramsay ve Silverman, 2005).

Ramsay ve Silverman (2002) alana sert sınırlar çizmemek üzere fonksiyonel veri analizinin genel bir tanımını yapmaktan özellikle kaçındıklarını belirtmiş; sadece fonksiyonel verinin yaygın özelliklerini tanımlamışlardır. Buna göre;

- Fonksiyonel veri süreklidir.
- Tekil verinin kendisi, bir noktadaki ölçüm değeri değil, bir fonksiyondur. Bu durumda fonksiyonel veriler çoğunlukla birbirinden bağımsız olacaktır ancak aynı fonksiyonel veri içinde yer alan değerlerin bağımsızlığına ilişkin bir varsayım bulunmamaktadır.
- Bazı durumlarda veri zamana bağlı bir fonksiyon olabilecektir. Ancak bu, özel bir durum değildir; verinin tek boyutlu değişkenin bir fonksiyonu olduğu durumlardaki çoğu varsayım doğrudan çok boyutlu değişkenlerin fonksiyonlarına uygulanabilmektedir.
- Verinin düzgün olması genel bir gereklilik olmamakla birlikte bu durum analizlerin önemli bir unsurudur. Bazı çalışmalarda fonksiyonun türevi önemli olurken bazı durumlarda ise verinin kendisi düzgün olmasa bile ortalamaların ya da modellemede kullanılan diğer fonksiyonların düzgün olması şartı aranabilir.

Şekil 1.1'de fonksiyonel verilere bir örnek verilmektedir. Veri, 10 kız çocuğunun 0-18 yaş arasındaki 31 farklı yaş noktasındaki boy ölçümlerinden oluşmaktadır. Yaş noktaları eşit aralıklı değildir; örneğin 1 yaş için dört çocuğun ölçümü bulunurken 2-8 yaş arası için yıllık ölçümler bulunmaktadır. Devamında ise iki yılda bir ölçüm yapılmıştır. Her bir ölçüm ayrık bir değeri göstermekle birlikte değerlerin tümü bir boy fonksiyonu olarak değerlendirilebilecek şekilde düzgün değişim göstermektedir. Bu kapsamda, veri seti 10 adet fonksiyonel gözlemden oluşmaktadır (Ramsay ve Silverman, 2005).



Şekil 1.1. Örnek Fonksiyonel Veri (Ramsay ve Silverman, 2005)

1.3. Fonksiyonel Verinin Düzgünleştirilmesi

Fonksiyonel veri analizinin ilk adımı ham verilerin fonksiyonel objelere, diğer bir deyişle fonksiyonel gözlemlerin kesikli noktalarının düzgün bir fonksiyona dönüştürülmesidir.[1] Bu işlem için, kesikli verilerin hatasız olduğu durumlarda interpolasyon yöntemi kullanılırken hatanın gözlendiği durumlarda düzleştirme (smoothing) yöntemleri kullanılır (Ramsay ve Silverman, 2005).

Kesikli verilerin düzgün bir fonksiyona dönüştürülmesinde yaygın olarak baz açılımları kullanılmaktadır. Baz fonksiyonları $f_k(t)$ ile gösterilen ve düzgün bir eğriyi temsil edecek şekilde özellikle seçilen fonksiyon kümesidir (Levitin ve ark, 2007).

$$x_i(t) = c_{i1}\,\Phi_1(t) + \ldots + c_{ik}\,\Phi_k(t) + \cdots + c_{iK}\,\Phi_K(t) \quad \text{i=1,…,N} \qquad \text{(Eşitlik 1.1)}$$

Burada K baz fonksiyonlarının sayısını, N birey sayısını, $\Phi_k(t)$ baz fonksiyonlarını, x(t) ise tahmin edilmek istenen fonksiyonu, c_{ik} ise her bir baz fonksiyonuna karşılık gelen ağırlık katsayılarını göstermektedir.

[1] Düzgünlük, bir fonksiyonu oluşturan iki komşu noktanın bağlantılı olduğu ve birbirinden çok farklı olmadığını ifade etmektedir (Ramsay ve Silverman, 2005).

Baz fonksiyonu açılımı ile ifade edilen eğrinin (fonksiyonun) tahmini c_{ik} parametrelerinin tahmin edilmesine karşılık gelmektedir. Katsayıların tahmininde ise en çok tercih edilen yöntemler En Küçük Kareler (EKK) ve Pürüzlü Ceza Yaklaşımıdır.

1.3.1. Baz Fonksiyonları

Bir fonksiyonu tahmin etmede kullanılabilecek farklı baz fonksiyon sistemleri mevcuttur. Bunlardan en sık kullanılanlar Splayn fonksiyonları ve Fourier serileridir. Bunun dışında, Sabit, Monomial ve Dalgacık bazlar da kullanılmaktadır. Baz fonksiyonu seçiminde verinin özellikleri dikkate alınmaktadır. Genel olarak, veri eğer periyodik yapı gösteriyorsa Fourier baz fonksiyonları, periyodik olmayan veriler için Splayn baz fonksiyonları ve süreksizlik içeren veriler için Dalgacık baz fonksiyonları kullanılmaktadır. Burada dikkat edilmesi gereken nokta, kullanılan baz fonksiyonlarının verinin özelliklerini temsil etme derecesinin yüksek olmasıdır. Böylelikle, x(t) fonksiyonunun tahmini de daha doğru olabilecektir. Diğer bir deyişle, gerçek fonksiyonların tahmininin doğruluğu baz fonksiyonlarının doğru seçimine bağlıdır (Yaraee, 2011).

Ramsay ve Silverman (1997) ise en az sayıda baz fonksiyonu ile veriyi en iyi tanımlayacak olan baz türünün seçilmesi gerektiğini belirtmiştir. Literatürdeki çalışmalara bakıldığında, verideki belirgin görünümleri özetlemede 20-30 bazın yeterli olduğunun belirtildiği görülmektedir (Şafak, 2019). Ölçüm sayısının fazla olduğu çalışmalarda ise baz fonksiyon sayısının belirlerlenmesinde ölçüm sayısının karekökünden faydalanıldığı görülmektedir (Ertaş ve Keser, 2008).

Ullah ve Finch (2013), 11 elektronik veritabanı üzerinde yapılan ve 1995-2010 arası yayınlanmış çalışmaları inceledikleri makalede, B-splayn fonksiyonu kullanılması yönteminin en çok kullanılan düzleştirme yöntemi olduğunu belirtmiştir. Bu durumun temel sebebi olarak da metodun sade ve esnek olması öngörülmüştür.

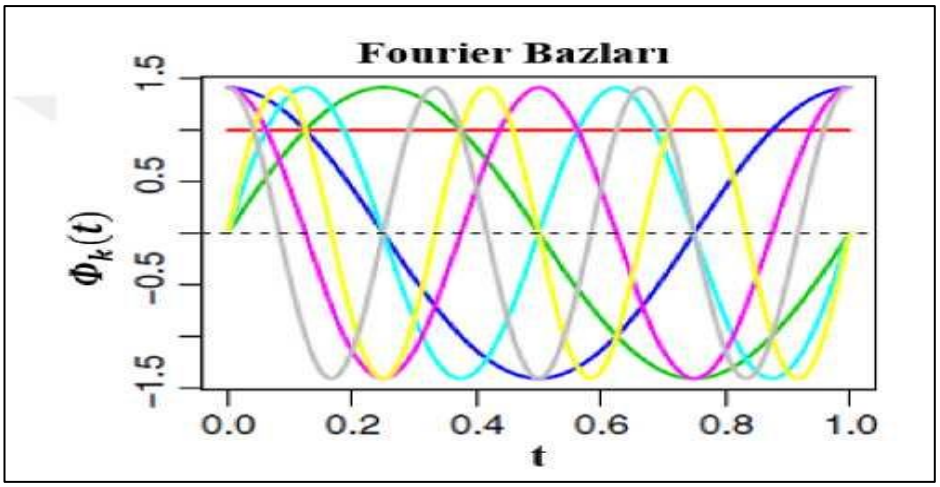
Bunun dışında en çok kullanılan yöntemler Fourier serileri, regresyon ve kernel düzgünleştirme yöntemleridir.

1.3.1.1. Fourier Baz Fonksiyonları

Bu baz fonksiyonu periyodiktir ve periyot w parametresi kullanılarak 2n/w olarak hesaplanmaktadır (Ramsay ve Silverman, 2005). Düzgünleştirme için kullanılacak baz fonksiyon sayısı K=1+2m eşitiliği kullanılarak ile belirlenir. Burada m, Eşitlik 1.2'de wt'nin çarpanı olarak alınabilecek en yüksek sayıyı (üst limiti) ifade etmektedir (Ramsay ve ark., 2009). Fourier Serisi aşağıdaki gibidir:

$$X(t) = c_0 + c_1 \sin wt + c_2 \cos wt + c_3 \sin 2wt + c_4 \cos 2wt + \cdots \quad \text{(Eşitlik 1.2)}$$

Fourier serileri, eğriliğin her yerde aynı dereceden olmaya meyilli olduğu durağan fonksiyonlar için özellikle kullanışlıdır. İdeal olarak ise Fourier serisinin periyodikliği belli bir miktarda veride de olmalıdır. Sıcaklık verisi bu verilere bir örnektir. Bu kapsamda, Fourier serileri belirli bir miktar düzensizlik içeren veriler için kullanışlı değildir (Ramsay ve Silverman, 2005). Şekil 1.2, K = 7 baz fonksiyonu ile bir tanesi sabit ve bire eşit olan Fourier bazlarını göstermektedir.



Şekil 1.2. Fourier Bazı Örneği (Soransen ve ark., 2013)

1.3.1.2. Splayn Baz Fonksiyonları

Bir splayn fonksiyonu her aralıkta tanımlanan sabit dereceli bir polinomdur. M-Splayn, B-Splayn ve budanmış güç fonksiyonları (truncated power functions) gibi farklı baz fonksiyonları vardır. Splayn fonksiyonlarını tanımlayan üç temel özellik bulunmaktadır (Ramsay ve ark, 2009):

- Kırılma noktaları (break points)
- Düğümler (knots)
- Polinomial kesimlerin derecesi (degree) veya mertebesi (order)

Bir splayn fonksiyonu oluşturmanın ilk adımı, ölçümlerin alındığı zaman aralığının belirli sayıda alt aralığa bölünmesidir. Bu alt aralıkların sınırlarının belirlendiği noktalara kırılma noktası denir. Her bir alt aralıkta yer alan splayn fonksiyonu sabit dereceli bir polinomdur. "Derece" polinomdaki en yüksek üs şeklinde tanımlanırken, "mertebe" ise polinomun derecesinin bir fazlasıdır (Ramsay ve ark., 2009). Örneğin, düz bir doğruda[2] yer alan x değişkeninin polinomdaki en yüksek üs değeri bir olduğu için derecesi bir, mertebesi ise ikidir (Şafak, 2019). Mertebe değeri bir ise sıfırıncı dereceden bir fonksiyon, iki ise parçalı doğrusal bir fonksiyon, üç ise kuadratik, dört ise kübik bir polinom ile ifade edilir. Uygulamada hem düzgün olmaları hem de birinci ve ikinci dereceden türevlerinin bulunması sebebiyle dördüncü sıradan splaynlar kullanılır. Ramsay ve ark (2009), düzgün türev fonksiyonlarına ihtiyaç olduğu durumlarda mevcut en yüksek türev derecesenin en az iki fazlası mertebeye sahip polinomun kullanılmasını önermektedir.

Bir splayn bazını tanımlarken, bir dizi düğüm noktası yerleştirilir. Düğüm noktaları olarak kırılma noktaları kullanılabilir. Ayrıca, herhangi bir kırılma noktasına birden fazla düğüm noktası tanımlanabilir.

[2] $y = \mathrm{a} + \mathrm{bx}$

Düğüm noktalarını yerleştirmenin farklı yolları vardır. Düğüm noktaları ölçümlerin alındığı her zaman noktasına yerleştirilebileceği gibi eşit aralıklarla da yerleştirilebilir. Bahsi geçen ilk yöntem daha çok tekrarlı ölçüm boyutu çok büyük olmadığında tercih edilirken ikinci yöntem her bir eğri için çok sayıda gözlem noktası bulunuyorsa ve ölçümler yaklaşık olarak eşit zaman noktalarında ise tercih edilmelidir. Böylelikle daha fazla eğrilik ve daha seyrek değerler bulunan yerlere daha fazla düğüm yerleştirilebilecektir (Yaraee, 2011). Splayn sistemdeki baz fonksiyon sayısı "K= mertebe + tanımlanan aralık içindeki düğüm sayısı -2" eşitliği kullanılarak hesaplanmaktadır.

1.3.1.3. B-Splayn Baz Fonksiyonları

Boor (2001) tarafından geliştirilmiş olan B-Splayn fonksiyonları periyodik olmayan fonksiyonel veri ya da parametrelerin düzgünleştirilmesinde en yaygın olarak kullanılan yöntemdir. B-splayn baz sistemi oluşturmadan önce Ramsay ve Silverman (2005) aşağıdaki özellikleri haiz bir baz fonksiyon sistemi tanımlamaktadır:

- Her taban fonksiyonunun kendisi de K düğüm sayısına sahip, m dereceli bir splayn fonksiyonudur.
- Birden fazla splaynların toplamı, farkı, çarpımı ve dolayısıyla doğrusal kombinasyonu yine bir splayn fonksiyonudur.
- K düğüm sayısına sahip, m dereceli herhangi bir splayn fonksiyonu bu baz fonksiyonlarının doğrusal bir kombinasyonu olarak ifade edilebilir.

B-splayn baz fonksiyonları kullanılarak elde edilen düzgünleştirilmiş fonksiyon şu şekilde ifade edilir (Şafak, 2019):

$$x(t) = \sum_{i=1}^{K} c_i \, \Phi_i^d \qquad \text{(Eşitlik 1.3)}$$

Burada $\{\Phi_i^d\}_{i=1}^K$ terimi B-splaynları göstermektedir. B-splayn fonksiyonu, d dereceden oluşan bir sürekli parçalı fonksiyon olarak kabul edilmektedir. Bu kapsamda, derecesi pozitif tam sayı olan polinomlar için mertebesi m olan b-splayn bazı aşağıdaki tekrarlı ilişki ile gösterilebilir (Jiang ve ark., 2016).

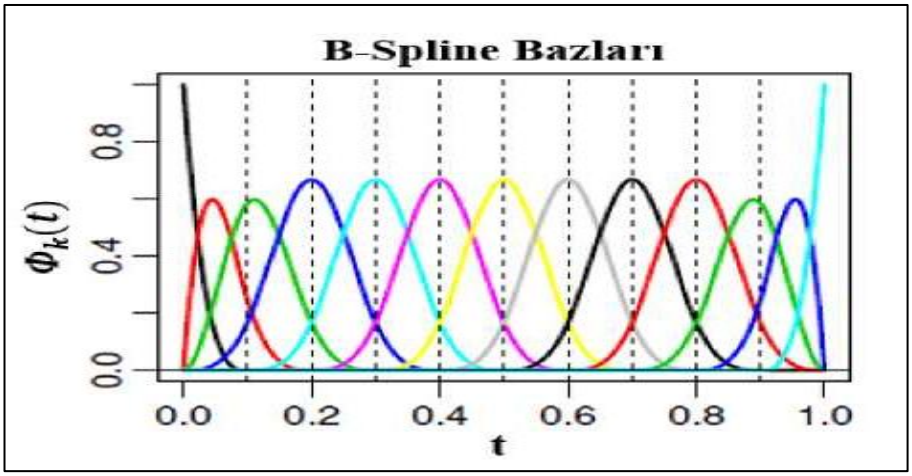
$$\Phi_j^d(t) = \frac{t-t_j}{t_{j+d}-t_j}\,\Phi_j^{d-1}(t) + \frac{t_{j+1+d}-t}{t_{j+1+d}-t_{j+1}}\,\Phi_{j+1}^{d-1}(t) \qquad \text{(Eşitlik 1.4)}$$

Burada t_j terimi bağımsız değişkenleri ifade etmekte olup gözlem aralığını alt kesişenlere bölmek için kullanılan düğümleri göstermektedir.

Derecesi sıfır olan polinomlar için mertebesi m olan B-splayn bazı ise aşağıdaki gibidir:

$$\Phi_j^d(t) = 1_{(t_j,t_{j+1})} = \begin{cases} 1 & t_j \geq t \geq t_{j+1} \\ 0 & diğer \end{cases} \qquad \text{(Eşitlik 1.5)}$$

Şekil 1.3'te verilen örnekte 11 eşit düğüm noktası ile 4. mertebe[3], K = 13 baz fonksiyonu ile B-splayn baz fonksiyonu gösterilmektedir (Soransen ve ark., 2013).



Şekil 1.3. B-Splayn Baz Fonksiyonu Örneği (Soransen ve ark., 2013)

[3] kübik polinomlar

1.3.2. Tahmin Yöntemleri

Yukarıda bahsedildiği üzere, uygun baz fonksiyonu kullanılarak bir açılım elde edildikten sonraki aşama açılımda yer alan c_{ik} parametrelerinin tahmin edilmesidir. Bu amaçla kullanılan yöntemlerden en sık tercih edilenler En Küçük Kareler ve Pürüzlü Ceza Yaklaşımıdır.

1.3.2.1. En Küçük Kareler Yöntemi

Bu yöntem ile hata kareler toplamı (HKT) minimize edilerek katsayılar elde edilmektedir. Eşitlik 1.1'de verilen fonksiyon için hata kareler toplamı aşağıdaki gibidir:

$$HKT = \sum_{i=1}^{n}[y_i - \sum_{k=1}^{K} c_k \, \Phi_k(t_i)]^2 \qquad \text{(Eşitlik 1.6)}$$

Bu ifade matris formunda şu şekilde yazılabilir:

$$\mathrm{HKT} = (\mathrm{y} - \Phi\mathrm{c})'(\mathrm{y} - \Phi\mathrm{c}) \qquad \text{(Eşitlik 1.7)}$$

Bu ifadenin c'ye göre türevi alınıp çözüldüğünde ĉ olarak ifade edilen en küçük kareler tahmini elde edilir:

$$\hat{\mathrm{c}} = (\Phi' \Phi)^{-1} \Phi' \mathrm{y} \qquad \text{(Eşitlik 1.8)}$$

EKK yaklaşımının, gerçek hataların ortalamasının sıfır ve varyansının sabit olup istatistiksel olarak bağımsız olduğu durumlarda kullanılması uygundur. Bu sebeple, eğer otokorelasyon ya da değişen varyans söz konusu ise verinin düzleştirilmesinde Ağırlıklı En Küçük Kareler (AEKK) yönteminin kullanılması daha uygun olur (Ramsay ve Silverman, 2005). Bu durumda, hata kareler toplamı ve ĉ, simetrik ve

pozitif bir W matrisinin eklenmesi yoluyla aşağıdaki gibi hesaplanır. W ağırlık matrisi hataların varyans kovaryans matrisinin tersine karşılık gelmektedir.

$$\text{HKT} = (y - \Phi c)'W(y - \Phi c) \qquad \text{(Eşitlik 1.9)}$$

$$\hat{c} = (\Phi'W\Phi)^{-1}\Phi'Wy \qquad \text{(Eşitlik 1.10)}$$

1.3.2.2. Pürüzlü Ceza Yaklaşımı

Bu yöntem, fonksiyonun komplekslilik (pürüzlülük) seviyesini cezalandırırken aynı zamanda fazla sayıda baz fonksiyonu kullanımına olanak vermektedir. Böylelikle, fonksiyonun pürüzlülüğünün bir ölçüsü kullanılanarak eğri pürüzlülüğünü veriye uyumsuzlaştıran ölçüt en aza indirilmektedir. Bir fonksiyonun yaygın olarak kullanılan pürüzlülük ölçüsü ise fonksiyonun toplam eğriliğini gösteren ikinci türevinin karesinin integralidir ve PEN_2 şeklinde ifade edilir (Ramsay ve ark, 2009):

$$PEN_2 = \int [D^2 x(t)]^2 \, dt \qquad \text{(Eşitlik 1.11)}$$

Pürüzlülük ölçüsü, veriye ve düzleştirmenin amacına göre seçilmelidir. Örneğin ikinci türevlerin düzleştirilmesi amaçlanıyorsa pürüzlülük ölçüsü hesaplanırken ikinci türevin eğikliğinin kullanılması daha iyi sonuç verir:

$$PEN_4 = \int [D^4 x(t)]^2 \, dt \qquad \text{(Eşitlik 1.12)}$$

Fonksiyonel veri analizinde temel amaçlardan biri kullanılan eğrinin verilere maksimum uyumunu sağlayıp iyi uyum gösteren bir kestirim yapmanın yanı sıra bu amaca kısmen ters düşerek çok fazla iniş çıkış göstermeyen bir eğri kestirimi elde etmektir (Ertaş ve Keser, 2008). En küçük kareler yöntemi tercih edildiğinde fonksiyonun düzgünlük seviyesine ilişkin bir kontrol yapılmamakta; pürüzlü ceza yaklaşımında ise hatalara ilişkin bir kontrol bulunmamaktadır. Eğrinin veriye uyumu

ile düzgünlüğü arasında bir denge sağlamak amacıyla her ikisini de bir arada bulunduran cezalandırılmış hata kareler toplamı (CHKT) kullanılabilir:

$$CHKT_{\lambda} = HKT + \lambda\, PEN_2 \qquad \text{(Eşitlik 1.13)}$$

Bu yöntem kullanıldığında hem kullanılan eğri hata kareler toplamına göre iyi bir uyum sağlar hem de düzgünlük derecesi kontol edilmiş olur. Burada λ düzgünleştirme parametresini ifade etmektedir. Düzgünleştirme parametresinin tercihinde ise sıklıkla Genelleştirilmiş Çapraz Geçerlilik (Generalized Cross Validation-GCV) kriteri kullanılmaktadır (Ramsay ve ark, 2009).

$$GCV(\lambda) = \left(\frac{n}{n-df(\lambda)}\right)\left(\frac{EKK}{n-df(\lambda)}\right) \qquad \text{(Eşitlik 1.14)}$$

Burada $df(\lambda)$ serbestlik derecesini ifade etmektedir. λ'nın artması, tahmin edilen eğrinin doğrusal regresyona yaklaşmasına; benzer şekilde azalması ise, eğrinin daha inişli-çıkışlı bir hale gelmesine sebep olmaktadır (Ertaş ve Keser, 2008).

1.4. Klasik İstatistiksel Yöntemlerin Fonksiyonel Verilere Uyarlanması

Daha önce belirtildiği üzere klasik istatistiksel yöntemin fonksiyonel veri analizine göre uyarlandığı birçok çalışma bulunmaktadır. Fonksiyonel verilere uyarlanan ilk istatistiksel yöntemler temel bileşenler analizi ve doğrusal regresyon modelleridir (Shang, 2014). Bunların dışında Tek Değişkenli Varyans Analizi (ANOVA), genelleştirilmiş doğrusal modeller ve kanonik korelasyon analizi gibi yöntemler de uyarlanmıştır.

1.4.1. Fonksiyonel Temel Bileşenler Analizi

Temel Bileşenler Analizi (TBA) çok değişkenli verilerde değişkenliğin incelenmesi ve boyut indirgeme amacıyla kullanılan bir yöntemdir. Yöntemin amacı, birbiriyle ilişkili çok sayıdaki değişkenden oluşan bir veri setinin, değişkenliğin öenmli bir kısmını koruyarak, bu değişkenleri temsil edebilen daha az sayıda değişkene indirgemek ve değişkenlere etki eden genel nedensel faktörleri tespit etmektir. Bu indirgeme, verilerin "temel bileşen" adı verilen bağımsız ve daha az sayıdaki değişkene dönüştürülmesi yoluyla elde edilir. [4] Analiz sonucunda elde edilen temel bileşenler, orijinal değişkenlerin birbirinden bağımsız ve varyansları toplam sistem varyansını mümkün olabilecek en fazla biçimde açıklayan doğrusal birleşimleridir. Temel bileşenler yöntemi kısaca aşağıdaki basamaklardan oluşmaktadır:

- Veri matrisinin standartlaştırılması
- Standartlaştırılmış veri matrisinin korelasyon matrisinin bulunması
- Korelasyon matrisinin özdeğerlerinin ve standartlaştırılmış öz vektörlerinin hesaplanması
- Öz değerlerden temel bileşenlerin toplam varyansı açıklama oranlarının bulunması
- Her bir öz vektörün devrik vektörü ile standartlaştırılmış veri matrisinin çarpılarak temel bileşen değerlerinin bulunması (Dinçer ve ark, 2003).

Fonksiyonel temel bileşenler analizde (FTBA) ise temel amaç, klasik temel bileşenler analizi ile aynı şekilde verilerdeki değişimi açıklamak olup fonksiyonel veri analizi kapsamında en sık kullanılan yöntemlerden biridir (Ullah ve Finch (2012). Bu amaçla, verilerdeki değişimi önemli bir bilgi kaybı olmaksızın tanımlayan birkaç ortogonal fonksiyon elde edilir. FTBA ile TBA analizi arasındaki temel fark FTBA'da vektörler yerine eğrilerle çalışılmasıdır. Dolayısıyla, TBA'da kullanılan kovaryans matrisi ve özvektörler yerine FTBA'da doğrusal operatörler ve öz fonksiyonlar kullanılır.

[4] TBA konusunda daha detaylı bilgi için bkz: Johnson ve Wichern (2002).

FBTA'da öncelikle her bir temel bileşen için, fonksiyonel veri ile aynı t aralığı üzerinde tanımlı temel bileşen ağrlık fonksiyonları tanımlanır. Ağırlık fonksiyonlarının ɣ(t) ile gösterildiği doğrusal kombinasyon aşağıdaki gibidir (Ertaş ve Keser, 2008):

$$Y_j = \int \gamma_j(t)\big[x(t) - E\big(x(t)\big)\big]dt \qquad \text{(Eşitlik 1.15)}$$

Burada Y_j, her bir x(t) için [x(t) – E x(t)]'nin γ_j üzerine izdüşüm (projection) miktarına karşılık gelmektedir. Ağırlık fonksiyonları fonksiyonel temel bileşen analizinin her aşamasında maksimum değişimi açıklayabilecek şekilde oluşturulur. Oluşturulan bu ağırlık fonksiyonları, ortogonal baz fonksiyonlar setidir. Devamında, eşitlik 1.15'te verilen doğrusal bileşenin varyansı maksimize edilmesi problemi TBA'dakine benzer şekilde ancak özdeğer ve özfonksiyon ayrımında çözülür.

Analiz sonucunda kaç tane temel bileşenin dikkate alınacağı hususunda literatürde en sık kullanılan yöntemlerden biri yamaç eğimi (scree plot) yöntemidir. Bu yöntem görsel bir yöntem olup özdeğerlerin büyükten küçüğe doğru sıralanması ve bu sıralamalara karşılık gelen bir çizgi grafiği ile gösterilmesini kapsamaktadır. Uygun bileşen sayısına ise bu çizgi grafiğinin dirsek noktasına bakılarak, geri kalan özdeğerlerin görece küçük ve hemen hemen aynı olduğu noktaya kadar alınarak karar verilir (Johnson ve Wichern, 2002).

Bunun yanında, FTBA, aykırı değerlere karşı duyarlıdır. Bu çerçevede, aykırı değerler ve verinin görsel analizi kapsamında kullanılan yöntemler açıklayıcı kutu diyagramları[5] ya da FTBA'nin sağlam (robust) versiyonlarının[6] kullanılmasıdır.

[5] Örnek bir çalışma için bkz. Hyndman ve Shang, 2010
[6] Örnek bir çalışma için bkz Kraus ve Panaretos, 2012

1.4.2. Fonksiyonel Doğrusal Modeller

Farklı değişkenler arasındaki ilişkiyi incelemek ve ileriye yönelik tahminlerde bulunmak amacıyla kullanılan doğrusal modeller de fonksiyonel veriye uyarlanmış; fonksiyonel doğrusal modeller (FDM) adı altında çeşitli çalışmalar yapılmıştır. Çizelge 1.1'de özetlendiği üzere üç farklı FDM türü olabilir:

- En az bir açıklayıcı değişkenin fonksiyonel olduğu modeller
- Yanıt değişkeninin fonksiyonel olduğu modeller
- Hem açıklayıcı hem de yanıt değişkenlerinin fonksiyonel olduğu modeller

Çizelge 1.1. Fonksiyonel Doğrusal Regresyon Modelleri (Acar Denizli 2016)

Model Adı	Yanıt Değişkeni	Açıklayıcı Değişkenler	Fonksiyonel Doğrusal Model
Fonksiyonel Yanıt Modeli	Fonksiyonel	Skaler	$\gamma_i(t) = \beta(t)X_i + \varepsilon_i(t)$
Tam Fonksiyonel Model	Fonksiyonel	Fonksiyonel	$\gamma_i(t) = \int \beta(t,s)\,\chi_i(s)ds + \varepsilon_i(t)$
Skaler Yanıt Modeli	Skaler	Fonksiyonel	$Y_i = \int \beta(s)\,\chi_i(s)ds + \varepsilon_i$

Çizelge 1.1'de özetlenen fonksiyonel modeller arasında en basit olanı yanıt değişkeninin fonksiyonel, açıklayıcı değişkenlerin skaler olduğu fonksiyonel yanıt modelidir. Aslında bu model klasik regresyon analizinin her t değeri için hesaplanması durumudur. Dolayısıyla burada regresyon katsayıları da zamana bağlı fonksiyonlardır. Bu regresyon katsayısı fonksiyonları, her t zamanı için, açıklayıcı değişkenin yanıt değişkenini nasıl etkilediğini göstermektedir (Levitin ve ark, 2007).

Hem yanıt hem açıklayıcı değişkenlerin fonksiyonel olduğu modellerde ise katsayılar tek boyutlu fonksiyon yerine iki boyutlu fonksiyonel yüzeylerdir ve $\beta(t,s)$ ile gösterilir. Bu yüzeyler $x_i(s)$ eğrisinin her bir zaman noktasının, $y_i(t)$ eğrisinin her bir zaman noktasını nasıl etkilediğini göstermektedir. Dolayısıyla bu model gecikmeli

etkilere de izin verir; veri bir zaman noktası için açıklayıcı değişkenlerin daha önceki davranışı yanıt değişkeni eğrisinin o andaki durumunu etkileyebilir. Benzer şekilde, açıklayıcı değişkenlerin ileri dönemlerdeki davranışı yanıt değişkeni eğrisini herhangi bir zaman noktası için etkileyebilir (Levitin ve ark, 2007).

Yukarıda açıklanan yanıt değişkeninin fonksiyonel olduğu iki modelde yanıt değişkenini tahmin etmek için fonksiyonel tek değişkenli varyans analizi (functional analysis of variance) (FANOVA) ya da fonksiyonel çoklu regresyon (functional multiple regression) analizi kullanılır. Bu analizler klasik yöntemlerin basit uyarlamaları şeklinde uygulanabilmektedir; temel değişiklik regresyon katsayılarının artık birer fonksiyon olmasıdır (Ramsay ve Silverman, 2005).

Yanıt değişkeninin skaler olduğu bir fonksiyonel doğrusal regresyon modeli ise aşağıdaki gibidir:

$$Y = \beta_0 + \int \chi(s)\beta(s)\,ds + \varepsilon \qquad \text{(Eşitlik 1.16)}$$

Burada, modeldeki açıklayıcı değişkenlerden en az biri, vektör değil, karesi integrallenebilen fonksiyonlardır. Bu durumda, katsayılar yine fonksiyondur ve her t noktası için, açıklayıcı değişkenlerin paterninin yanıt değişkenine katkısını tanımlamaktadır (Levitin ve ark, 2007).

Bu fonksiyonel modelin, yanıt değişkenini mükemmel derecede tahmin eden sonsuz sayıda çözümü vardır. Bu sebeple, $\beta(s)$ fonksiyonu için yorumlanabilir bir kestirim bulmak için taban fonksiyonları veya düzgünleştirmeye dayalı farklı yöntemler önerilmektedir (Ramsay ve Silverman, 2005). Önerilen bu yöntemlerden en sık kullanılanlar klasik taban açılımı temsili, boyut indirgeme amacıyla kullanılan fonksiyonel temel bileşenler analizi ve fonksiyonel en küçük kareler analizi yöntemleridir.

Fonksiyonel temel bileşenler analizi ve fonksiyonel en küçük kareler analizi kullanılarak oluşturulan modeller kurulurken iki farklı yaklaşım kullanılabilir. Bunlar taban açılımı yaklaşımı ve Febrero-Bande ve Oviedo de la Fuente (2012) tarafından önerilen ve verinin öz yapısından elde edilen fonksiyonel bileşenlere dayalı olan yaklaşımlardır (Acar Denizli, 2016).

2. GEREÇ VE YÖNTEM

Önceki bölümde görüldüğü üzere, fonksiyonel veri analizi, zaman serisi oluşturan ayrık/noktasal gözlemlerin bir fonksiyon oluşturacak şekilde ifade edilmesi, oluşturulan fonksiyonel veri seti üzerinde çok değişkenli analiz yöntemlerinden uygun olanların uygulanarak modelleme ve tahmin çalışmaları yürütülmesini kapsamaktadır. Yürütülecek çalışmalarda FVA'nın özelliklerinin tamamının kullanılması gerekmemektedir; hangi özelliklerinin kullanılacağı verinin yapısı ve araştırmanın niteliğine uygun şekilde belirlenmelidir.

Bu kapsamda, bu çalışmada kullanılacak yöntem belirlenirken öncelikle doğurganlık hızı, ölüm hızı, göç hızı gibi demografik göstergelerin analizi ve tahminine ilişkin kullanılan yöntemler taranmıştır. Söz konusu yöntemlerden Hyndman ve Booth (2008) tarafından önerilen yöntem bu çalışma kapsamında kullanılmıştır. Yöntem, Hyndman ve Ullah (2007) tarafından öne sürülen ve birçok çalışmada tercih edilen yöntemi bazı açılardan geliştirilerek yeni bir uygulama yapmaktadır. [7]

Yöntemin uygulanmasında yazarlar tarafından erişime açık olarak sunulmuş olan kodlar Türkiye İstatistik Kurumu (TÜİK) tarafından yayınlanan Türkiye yaşa özel doğurganlık hızı verileri ve bu çalışma kapsamında modifiye edilerek kullanılmıştır. Uygulama, açık kodlu R programında gerçekleştirilmiş olup programda yer alan "readxl" (Wickham ve ark. 2019), "demography" (Hyndman ve ark. 2019), "rainbow" (Shang ve Hyndman 2019), "cobs" (Pin ve Maechler 2020), "mgcv" (Wood 2020), "strucchange" (Zeileis ve ark. 2019), "RCurl" (CRAN TEAM ve Lang 2020) ve "ftsa" (Hyndman ve Shang 2020) kütüphanelerinden faydalanılmıştır.

[7] Detaylı bilgi 2.2. Yöntem bölümünde sunulmaktadır.

2.1. Uygulama Verileri

Çalışmada Türkiye İstatistik Kurumu tarafından 2001-2018 yılları için yayınlanmış yaşa özel doğurganlık hızı verileri kullanılmıştır. Yaşa özel doğurganlık hızı belli bir yaş grubunda bin kadın başına düşen ortalama canlı doğan çocuk sayısını ifade etmektedir:

$$Y\ddot{O}DH = \frac{D_i}{N_i} \text{ x } 1000 \qquad \text{(Eşitlik 2.1)}$$

Burada D_i i'nci yaş grubundaki doğum sayısını, N_i ise i'nci yaş grubundaki yıl ortası kadın nüfusu göstermektedir. Veri kaynakları İçişleri Bakanlığı, Nüfus ve Vatandaşlık İşleri Genel Müdürlüğü, Merkezi Nüfus İdaresi Sistemi (MERNİS) ve TÜİK'tir.

2.2. Yöntem

Doğurganlık hızı, ölüm hızı, göç hızı gibi diğer demografik göstergelerin analizi ve tahminine ilişkin kullanılan yöntemler arasında en popüler olan ve birçok çalışmaya öncülük eden yöntem Lee ve Carter (1992) tarafından ortaya atılmıştır. Lee ve Carter (1992), tekil değer ayrışımı (Singular Value Decomposition)[8] metodunu kullanarak Amerika Birleşik Devletleri (ABD) için 1933-1987 dönemi ölüm hızını analiz etmiş ve 2065 yılına kadar bir tahmin oluşturmuştur. Yöntem, izleyen yıllarda sıklıkla kullanılmış ve çeşitli değişiklik ya da geliştirmeler önerilmiştir. Örneğin, Li ve ark (2004), düzensiz aralıklar üzerinde az sayıda gözlem olduğu durumlar için bir yöntem geliştirerek Çin ve Güney Kore için bir uygulama gerçekleştirmiştir.[9] Lee-Carter (LC)

[8] Tekil değer Ayrışımı, karmaşık bir matrisi ortonormal bir matris, köşegen bir matris ve ortonormal bir matristen oluşan üçlü bir çarpıma ayrıştıran bir algoritmadır.
[9] Diğer örnekler için bkz: Lee (1993), Booth ve ark (2002), Renshaw ve Haberman (2003), Wolf (2004).

yöntemi, eksik verinin çok olduğu ya da zaman serisinin yeterince uzun olmadığı durumlarda kullanışlı olmamaktadır (Fang ve Härdle, 2015).[10]

LC yönteminin geliştirilmesi sonucunda ortaya çıkan metotlardan en sık kullanılanı ise Hyndman ve Ullah (2007) tarafından öne sürülen yöntemdir. Çalışma, Fransa ölüm hızı ile Avusturalya doğurganlık hızı verilerinin fonksiyonel yaklaşımla ele alındığı bir tahmin yöntemi sunmaktadır. Çalışmada fonksiyonel veri analizi, parametrik olmayan düzleştirme yöntemleri ile sağlam istatistik yöntemlerinin birlikte kullanıldığı ve Lee-Carter yönteminin bir genelleştirmesi olarak nitelendirilebilecek yeni bir yöntem önerilmektedir. Yöntem yaşa bağlı düzgün fonksiyonlar elde edilmesini sağlamakta olup aykırı değerlere karşı da sağlamdır.[11]

Hyndman ve Ullah (2007) tarafından öne sürülen yöntemde öncelikle aşağıdaki değişkenler tanımlanmıştır:

$$y_t(x_i) = f_t(x_i) + \sigma_i(x_i)\varepsilon_{t,i} \qquad t = 1, \dots, n;\ i = 1, \dots . p \qquad \text{(Eşitlik 2.2)}$$

Burada, $y_t(x)$ modellenmek istenen değişkenin (ölüm ya da doğurganlık hızı gibi) logaritmasını, $f(.)$ x'e bağlı düzgün bir fonksiyonu, $\varepsilon_{t,i}$ standart hata terimini, $\sigma_i(x_i)$ x ile birlikte değişkenlik gösteren gürültüyü, t zamanı, i ise yaş gruplarını ifade etmektedir. Çalışmanın amacı ise $y_t(x)$'i $[x_1, x_p]$ ve $t = n+1, \dots, n+h$ için tahmin etmektir. Çalışmanın aşamaları aşağıda özetlenmiştir:

1) $\{x_t, y_t(x_i)\}$'nin kullanılarak $x \in [x_1, x_p]$ için $f_t(x)$'in tahmin edilmesi amacıyla; her bir t için verinin, parametrik olmayan düzleştirme yöntemleri kullanılarak düzleştirilmesi, $i = 1, \dots, p$

[10] LC yöntemi ve diğer bazı yöntemlerle karşılaştırması için bkz Booth ve ark (2006), Shang ve ark (2011), Shang (2012).
[11] Hyndman ve Ullah (2007) yöntemi esas alınarak yapılan çalışmalara örnek olarak Erbas, Hyndman, ve Gertig (2007), Shang ve ark (2013), Fang ve Härdle (2015) verilebilir.

2) Uydurulmuş eğrilerin aşağıdaki baz fonksiyonu kullanılarak ayrıştırılması:

$$f_t(x) = \mu(x) + \sum_{k=1}^{K} \beta_{t,k}\, \Phi_k(x) + e_t(x) \qquad \text{(Eşitlik 2.3)}$$

Burada $\mu(x)$, $f_t(x)$ için bir lokasyon ölçütü, $\{\Phi_k(x)\}$ ortonormal baz fonksiyonlar setini göstermektedir. Ayrıca, $e_t(x) \sim N(0, v(x))$'tir.

3. Tek değişkenli zaman serisi modellerinin her bir $\{\beta_{t,k}\}$ katsayısıya uydurulması, $k = 1, \dots, K$

4. Uydurulmuş zaman serisi modellerinin kullanılarak $\{\beta_{t,k}\}$ katsayısılarının tahmin edilmesi, $k = 1, \dots, K, t = n+1, \dots, n+h$

5. Katsayıların tahminlerinin 2 numaralı aşamada verilen eşitlikte kullanılarak $f_t(x)$ için tahmin elde edilmesi. Burada $f_t(x)$ tahminleri aynı zamanda $y_t(x)$'in tahminleridir (bkz 1. aşama).

6. 1 ve 2 nci aşamalarda verilen işlemlerden elde edilen hata terimlerinin ölçülen varyansı kullanılarak tahminler için tahmin aralıklarının hesaplanması

Şimdi yöntemi daha detaylı inceleyelim.[12] $p_t(x)$, x yaşındaki kadınlar ve t yılı için, bin kadın başına düşen doğurganlık hızını; $N_t(x)$ ise t yılı ortasındaki x yaşındaki kadın nüfus büyüklüğünü göstersin. Bu durumda Taylor yaklaşımı kullanılarak $p_t(x)$'in varyansı aşağıdaki gibi bulunur:

$$\hat{\sigma}_t^2(x) \approx [1000 - p_t(x)] N_t^{-1}(x) p_t^{-1}(x) \qquad \text{(Eşitlik 2.4)}$$

Düzleştirme için kısıtlı ağırlıklandırılmış pürüzlü cezalı regresyon splayn kullanılmıştır. Burada doğru tahmin elde edebilmek için bazı kısıtlar koymak gereklidir. Doğurganlık hızı için Hyndman ve Ullah (2007) uydurulmuş eğrilerin

[12] Yöntem Hyndman ve Ullah (2007) çalışması takip edilerek, doğurganlık hızı özelinde anlatılmıştır.

konkav olduğu kısıtını kullanmış; bu varsayımın tüm doğurganlık verileri için geçerli olduğunu belirtmiştir.

İkinci aşamada öngörülen ayrıştırma işlemi için ise temel bileşenler analizi uygulanmıştır. Veri bir K için, baz fonksiyonları $\{\Phi_k(x)\}$ aşağıda verilen ortalama entegre kare hatasını (mean integrated squared error) (MISE) minimize eder:

$$MISE = \frac{1}{N} \sum_{t=1}^{N} \int e_t^2(x)dx \qquad \text{(Eşitlik 2.5)}$$

Hyndman ve Ullah (2007), sağlam ölçümler elde etmek için ise $\mu(x)$ için ölçülen düzgün eğriler $\{\hat{s}_1(x), \ldots, \hat{s}_n(x)\}$'in L_1 medyanını kullanmıştır:

$$\hat{\mu}(x) = arg \min_{\theta(x)} \sum_{t=1}^{n} \|\hat{s}_t(x) - \theta(x)\| \qquad \text{(Eşitlik 2.6)}$$

Burada $\|g(x)\| = \{\int g^2(x)dx\}^{1/2}$ g fonksiyonunun L_2 normunu ifade etmektedir. Devamında, $\{\hat{s}_t^*(x)\}$ üzerinde ağırlıklandırılmış fonksiyonel temel bileşenler analizi uygulanmıştır. Burada $\{\hat{s}_t^*(x)\}$ medyana göre düzeltilmiş veriyi göstermektedir:

$$\hat{s}_t^*(x) = \hat{s}_t(x) - \hat{\mu}(x) \qquad \text{(Eşitlik 2.7)}$$

Ağırlıklar FTBA esnasında iki aşamalı bir yöntemle[13] hesaplanmış ve uzak yıllar için 0, diğerleri için 1 olarak kabul edilmiştir.

$\Phi_k(x)$'lerin seçilme yönteminden dolayı, $k \neq l$ olduğu sürece $\hat{\beta}_{t,k}$ ve $\hat{\beta}_{t,l}$ korele olmayan değişkenlerdir. Bu durumda, zamana bağlı değişen $\beta_{t,k}$'lerin tahmininde otoregresif tümleşik hareketli ortalamalar modeli (ARIMA) gibi tek değişkenli zaman serisi modelleri kullanılabilir.

[13] Detaylar için bkz Hyndman ve Ullah (2007).

Eşitlik (2.2) ve eşitlik (2.3) birleştirildiğinde aşağıdaki eşitlik elde edilir:

$$y_t(x) = \mu(x) + \sum_{k=1}^{K} \beta_{t,k}\Phi_k(x) + e_t(x) + \sigma_t(x)\varepsilon_t \qquad \text{(Eşitlik 2.8)}$$

Bu eşitliği $y_{n+h}(x)$ 'nin tahmini için yazarsak:

$$\hat{y}_{n+h}(x) = E[y_{n+h}(x)] = \hat{\mu}(x) + \sum_{k=1}^{K} \hat{\beta}_{n,k,h}\, \widehat{\Phi}_k(x) \qquad \text{(Eşitlik 2.9)}$$

Burada $\hat{\beta}_{n,k,h}$, $\beta_{n,k,h}$'nin $\hat{\beta}_{1,k}, \ldots, \hat{\beta}_{n,k}$ kullanılarak bulunmuş h aşama sonraki tahminidir. Devamında ise her bir bileşen ortogonal olduğundan, toplam varyans bileşenlerin varyanslarının toplamı olarak hesaplanır.

Eşitlik 2.9'da verilen tahmin için ise hata terimi aşağıdaki gibi ifade edilir:

$$e_{n,h}(x) = y_{n+h}(x) - \hat{y}_{n,h}(x) \qquad \text{(Eşitlik 2.10)}$$

Bu durumda entegre kare tahmin hatası (integrated squared forecast error) (ISFE) ise aşağıdaki gibi tanımlanır:

$$ISFE_n(h) = \int_x e_{n,h}^2(x)dx \qquad \text{(Eşitlik 2.11)}$$

Burada K'yı seçebilmek için ISFE test örneklemi için tekrarlayan şekilde minimize edilir. Yani, model bir t'ye kadar uydurulduktan sonra gelecek m period için tahminler yapılıp $ISFE_t(h)$ hesaplanır $(h = 1, \ldots, m)$. Daha sonra K, $\sum_{t=n}^{n-h}\sum_{h=1}^{m} ISFE_t(h)$ ifadesini minimiz edecek şekişde seçilir. Burada N, modeli uyduruabilmek için kullanılan minimum gözlem sayısını göstermektedir. Hyndman ve Ullah (2007), uygulamada N=20 kullanmıştır.

Hyndman ve Ullah (2007) tarafından öne sürülen ve birçok çalışmada tercih edilen bu yöntem Hyndman ve Booth (2008) çalışmasında bazı açılardan geliştirilerek yeni bir uygulama yapılmıştır. Bu çalışmanın önceki çalışmadan üç temel farkı bulunmaktadır:

1. Hyndman ve Ullah (2007) çalışması modellenmek istenen verinin logaritmasını kullanırken Hyndman ve Booth (2008) daha genel bir dönüşüm yöntemi olan bir Box-Cox dönüşümü kullanmaktadır:

$$y_t(x) = \begin{cases} \frac{1}{\lambda}\left([y_t^*(x)]^\lambda - 1\right) & 0 < \lambda \leq 1 \\ \log_e y_t^*(x) & \lambda = 1 \end{cases} \qquad \text{(Eşitlik 2.12)}$$

Burada λ'nın değeri dönüşümün gücünü göstermektedir. $\lambda = 0$ olduğunda logaritmik dönüşüm yapılmış; $\lambda = 1$ olduğunda ise hiç dönüşüm yapılmamış demektir.

2. Hyndman ve Ullah (2007) çalışması modeli doğurganlık ve ölüm hızı verilerine uygularken Hyndman ve Booth (2008) ilave olarak net göç hızı verilerine de uygulamıştır.
3. Hyndman ve Ullah (2007) çalışmasında tahminin varyansını hesaplamak üzere kullanılan metot, gözlenen veriler üzerinde daha iyi kalibrasyon sağlayacak şekilde modifiye edilmiştir:

Hyndman ve Ullah (2007)'ye göre tahmin varyansı aşağıdaki gibidir:

$$V_h(x) = Var\left[s_{n+h}(x)|I, \Phi\right] = \hat{\sigma}_\mu^2(x) + \sum_{k=1}^{K} u_{n+h,k}\widehat{\Phi}_k^2(x) + v(x) \qquad \text{(Eşitlik 2.13)}$$

Burada I tüm gözlenen verileri göstermekte olup $u_{n+h,k} = Var(\beta_{n+h,k}|\beta_{1,k}, \ldots, \beta_{n,k})$ şeklinde zaman serisi modelinden; $\hat{\sigma}_\mu^2(x)$ kullanılan düzleştime yönteminden elde edilir. $v(x)$ ise $\hat{e}_t^2(x)$'in her bir x için ortalaması

bulunarak hesaplanır. Birinci terim düzleştirmeden kaynaklanan hatayı, ikinci terim tahminden kaynaklanan hatayı, üçüncü terim ise açıklanamayan dinamik varyansı göstermektedir. Gözlemsel hata da ilave edildiği zaman aşağıdaki eşitlik elde edilir:

$$Var\,[y_{n+h}(x)|I,\Phi = \; V_h(x) + \sigma_t^2(x) \qquad \text{(Eşitlik 2.14)}$$

Küçük h değerleri için $V_h(x)$'in geçerliliği örneklem içi ampirik tahmin varyansını hesaplayarak test edilebilir:

$$W_h(x) = \frac{1}{n-h+m+1} \sum_{t=m}^{n-h}[s_{t+h}(x) - \hat{s}_{t,h}]^2 \qquad \text{(Eşitlik 2.15)}$$

Burada m modeli uydurabilmek için gerekli en küçük gözlem sayısıdır. Pratikte $W_h(x)$ ile $V_h(x)$ arasında ciddi farklar gözlenebilmektedir. Bu sebeple Hyndman ve Booth (2008), aşağıdaki düzeltilmiş varyans ifadesini kullanmıştır:

$$Var\,[y_{n+h}(x)|I,\Phi = \frac{V_h(x)W_1(x)}{V_1(x)} + \sigma_t^2(x) \qquad (2.16)$$

Böylelikle varyans, tek aşamadaki tahmin varyansı ile örneklem içi ampirik tek aşamadaki tahmin varyansı eşleştirilmiş olur. Hyndman ve Booth (2008), bu düzeltme kullanıldığında çoğu yaş grubu doğurganlık ve ölüm hızları için daha küçük varyans elde edildiğini belirtmiştir.

Bunun yanında, daha önce bahsedildiği üzere, gözlenen varyans veri setinin yapısına bağlıdır. Hyndman ve Booth (2008) doğurganlık için ortalaması $f_t(x)E_t^F(x)$ olan Poisson dağılım varsayımını kullanmışlardır. Böylelikle:

$$\sigma_t^2(x) \approx \; f_t(x)^{2\lambda-1}E_t^{-1}(x) \qquad \text{(Eşitlik 2.17)}$$

Ayrıca, uygulamada K'nın belirlenmesi gerekmektedir. Hyndman ve Booth (2008), Hyndman ve Ullah (2007) tarafından kullanılan ve ortalama entegre kare hatasının minimize edilmesini öngören metotun, K'nın yeterince büyük seçilmesi durumunda, K seçimine duyarsız olduğunu belirtmiştir. Bu nedenle de $K = 6$ seçiminin yeterince büyük olduğu ve çalışmaları için seçildiği ifade edilmiştir.

Bu tez çalışmasında ise Hyndman ve Booth (2008) tarafından önerilen yöntem kullanılmaktadır. Yöntemin uygulanmasında yazarlar tarafından erişime açık olarak sunulmuş olan kodlar[14] Türkiye verisi ve bu çalışma kapsamında modifiye edilerek kullanılmıştır. Uygulama açık kodlu R programı kullanılarak gerçekleştirilmiştir.

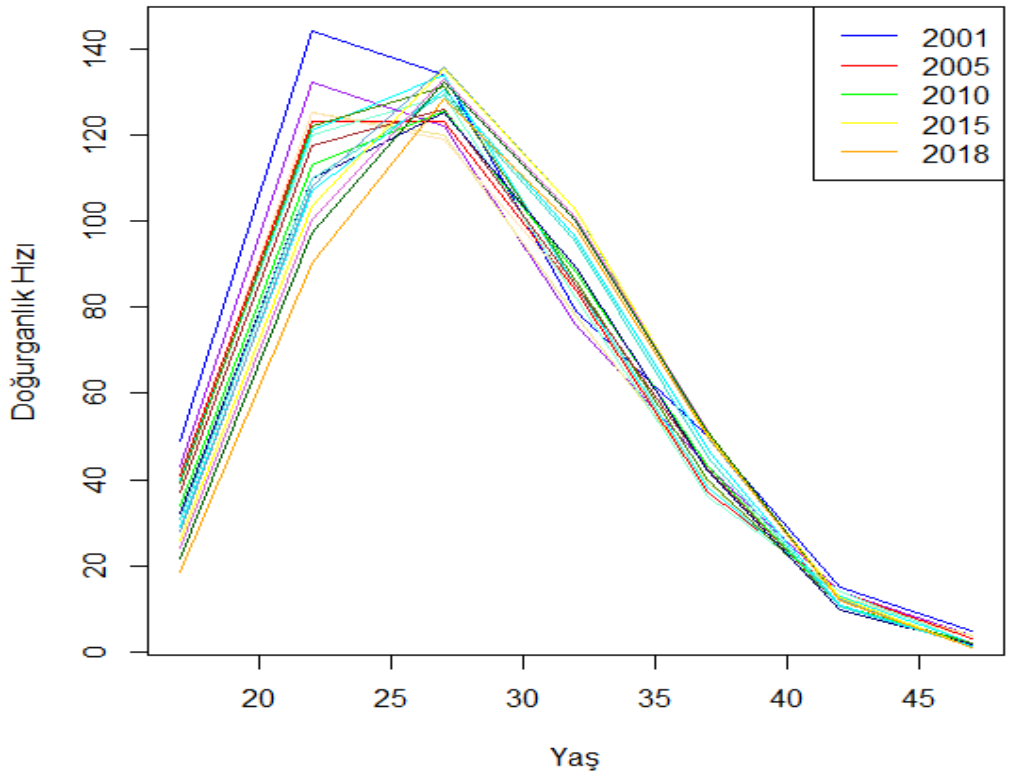
[14] https://cran.r-project.org/web/packages/demography/index.html

3. BULGULAR

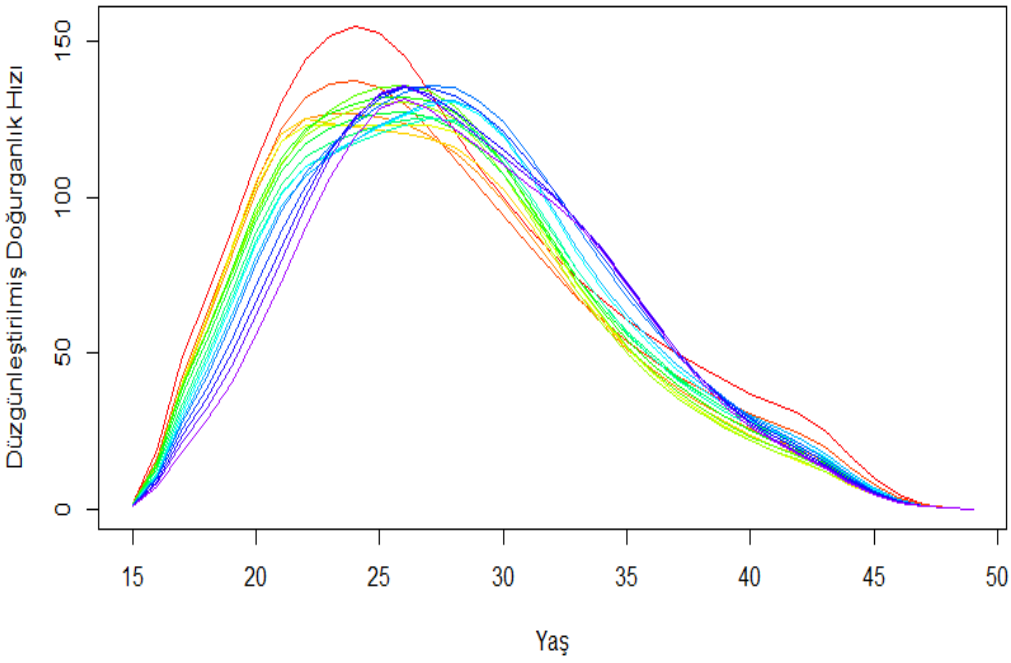
Bölüm 2.2'de açıklanan yöntem Türkiye için yaşa özel doğurganlık hızı verilerine uygulanmıştır. Doğurganlık verilerinin dönüştürülmesinde kullanılacak yöntem konusunda literatürde bir fikir birliği bulunmamaktadır (Hyndman ve Booth, 2008). Bu çalışmada, Hyndman ve Booth (2008) tarafından izlenilen yönteme paralel şekilde, örneklem dışı tahmin hatasını minimize eden ve en küçük tahmin aralığı sunan λ değeri seçilmiştir. Bunun yanında, benzer çalışmalarda kullanılan Akaike bilgi kriterine göre de karşılaştırma yapılmıştır. Sonuç olarak, $\lambda = 0$ bu kriterleri sağlayan değer olarak bulunmuştur.

Daha önce belirtildiği üzere baz fonksiyonu sayısı K yeterince büyük seçilirse, metot K seçimine duyarsız hale gelmektedir. Benzer şekilde bu çalışmada da $K = 6$ seçilerek işlem yapılmış, ancak $K = 5$ seçildiğinde toplam varyansın tamamının açıklandığı görülmüştür. Ayrıca $\lambda = 0$ seçimi ile birlikte farklı K değerleri için yapılan değerlendirmede $K = 5$ seçiminin en düşük hata ve Akaike bilgi kriterini sağladığı görülmüş olup bu seçimle devam edilmiştir.

Öncelikle, 2001-2018 yıllarına ait doğurganlık hızı verileri Şekil 3.1'de verilmektedir. Şekilden görüldüğü üzere, doğurganlık hızının en yüksek olduğu yaş aralığı zamanla birlikte artmaktadır. 2001 yılında doğurganlık hızı en yüksek 20-24 yaş grubunda iken 2018 yılında 25-29 yaş grubundadır. Sonrasında düzgünleştirme işlemi uygulanmış olup Şekil 3.2 elde edilmiştir.

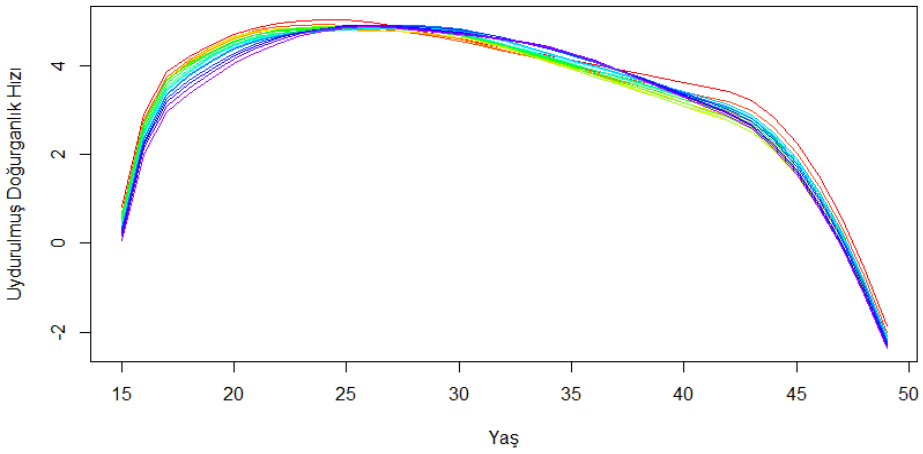


Şekil 3.1. Türkiye Yaşa Göre Doğurganlık Hızları (2001-2018)



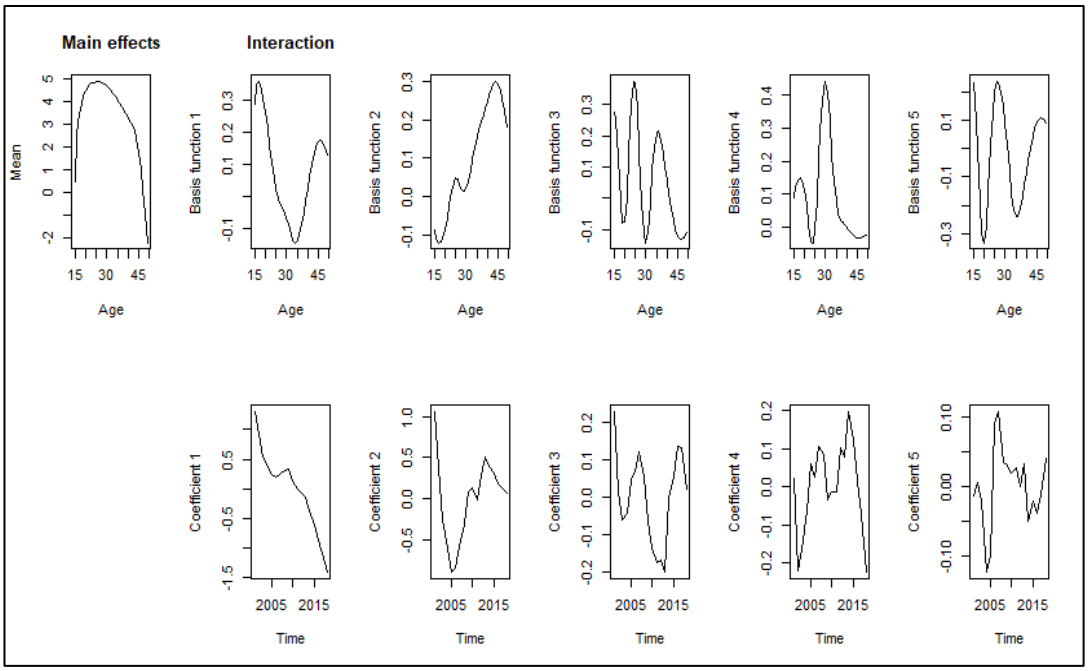
Şekil 3.2. Türkiye Yaşa Göre Düzgünleştirilmiş Doğurganlık Hızları (2001-2018)

Son aşamada elde edilen uydurulmuş veriler ise Şekil 3.3'te sunulmaktadır.



Şekil 3.3. Türkiye Yaşa Göre Uydurulmuş Doğurgalık Hızları

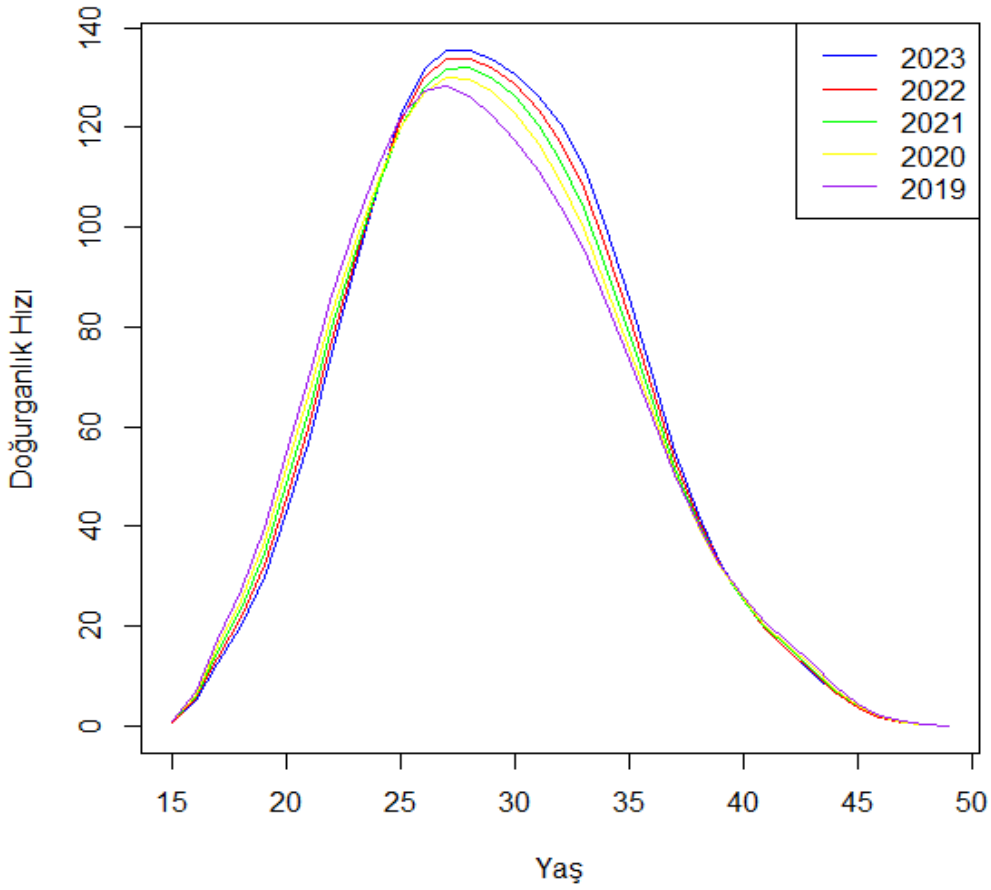
Şekil 3.4 ise elde edilen baz fonksiyonları ile katsayılarını göstermektedir. Birinci baz fonksiyonu toplam varyansın yüzde 62,1'ini, ikinci baz fonksiyonu ise yüzde 33,6'sını açıklamaktadır. Diğer üç baz fonksiyonu ise sırasıyla yüzde 2, yüzde 1,8 ve yüzde 0,5'ini açıklamaktadır. Bu şekilde, beş baz fonksiyonu kullanılarak varyansın tamamı açıklanmıştır.



Şekil 3.4. Uydurulmuş Baz Fonksiyonları ve Katsayıları

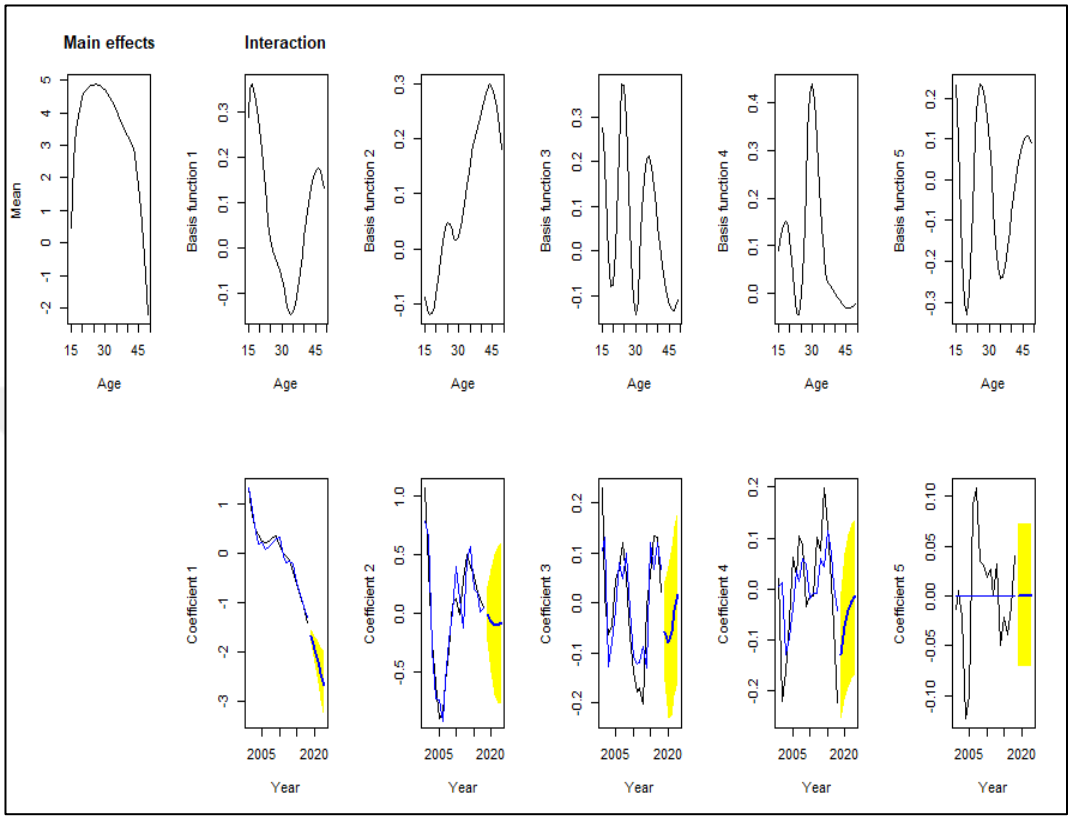
Birinci baz fonksiyonuna ait katsayıya bakıldığında doğurganlık hızının genel olarak tutarlı şekilde düştüğü görülmektedir. Bu düşüşü yalnızca 2006-2008 arasındaki hafif artış bozmaktadır. Baz fonksiyona göre ise, bu düşüş daha küçük ve daha büyük yaş gruplarında daha hızlı olmuştur. Diğer bir deyişle, küçük yaş grupları ile büyük yaş gruplarındaki doğurganlık hızı, ortalamaya göre daha hızlı düşmektedir.

Uygulama sonucunda elde edilen tahmin değerleri ise Şekil 3.5'te gösterilmektedir. Buna göre, 2019-2023 arası doğurganlık hızının en yüksek olduğu yaşın bir miktar artması tahmin edilmektedir. Bunun yanında, 25-40 yaş arası gruplarda doğurganlık hızının zamanla artması tahmin edilmiştir. Diğer bir deyişle, bu yaş aralığı için tahmin edilen doğurganlık hızları, 2019 yılından 2023 yılına doğru gittikçe artmaktadır. 20 yaş altı grup için ise tam tersi durum söz konusudur. 2019-2023 arasında 20 yaş altı grup için doğurganlık hızının azaldığı tahmin edilmektedir.



Şekil 3.5. Tahmin Edilen Yaşa Göre Doğurganlık Hızları (2019-2023)

Şekil 3.6 ise katsayıların tahminlerini göstermektedir. Burada kullanılan sarıyla taralı alan hesaplanan %80 tahmin aralığını ifade etmektedir.



Şekil 3.6. Uydurulmuş ve Tahmin Edilen Baz Fonksiyonları ve Katsayıları

4. TARTIŞMA

Fonksiyonel veri analizi temel olarak önce zaman serisi oluşturan ayrık gözlemlerin bir fonksiyon oluşturacak şekilde ifade edilmesi, sonrasında ise oluşturulan fonksiyonel veri seti üzerinde çok değişkenli analiz yöntemlerinden uygun olanların uygulanarak modelleme ve tahmin çalışmaları yürütülmesini kapsamaktadır. Bu çerçevede, çok değişkenli istatistiksel analizden fonksiyonel veri analizine uyarlanan bazı yöntemler fonksiyonel betimsel istatistik, fonksiyonel temel bileşenler analizi, fonksiyonel doğrusal modeller ve fonksiyonel kanonik korelasyon analizidir (Ramsay ve Silverman, 2005). Ancak, yürütülecek bir çalışmada FVA'nin bu özelliklerinin tamamının kullanılması gerekmemektedir; hangi özelliklerinin kullanılacağı verinin yapısı ve araştırmanın niteliğine uygun şekilde belirlenmelidir.

Fonksiyonel veri analizinin ilk adımı ham verilerin fonksiyonel objelere, diğer bir deyişle fonksiyonel gözlemlerin kesikli noktalarının düzgün bir fonksiyona dönüştürülmesidir. Bu işlem için, kesikli verilerin hatasız olduğu durumlarda interpolasyon yöntemi kullanılırken hatanın gözlendiği durumlarda düzleştirme (smoothing) yöntemleri kullanılır (Ramsay ve Silverman, 2005). Kesikli verilerin düzgün bir fonksiyona dönüştürülmesinde ise yaygın olarak baz açılımları kullanılmaktadır. Uygun baz fonksiyonu kullanılarak bir açılım elde edildikten sonraki aşama açılımda yer alan parametrelerin tahmin edilmesidir. Bu amaçla kullanılan yöntemlerden en sık tercih edilenler En Küçük Kareler (EKK) ve Pürüzlü Ceza Yaklaşımıdır.

Klasik istatistiksel yöntemin fonksiyonel veri analizine göre uyarlandığı birçok çalışma bulunmaktadır. Fonksiyonel verilere uyarlanan ilk istatistiksel yöntemler temel bileşenler analizi ve doğrusal regresyon modelleridir (Shang, 2014). Bunların dışında ANOVA, genelleştirilmiş toplanır modeller ve kanonik korelasyon analizi gibi yöntemler de uyarlanmıştır.

Doğurganlık hızı, ölüm hızı, göç hızı gibi diğer demografik göstergelerin analizi ve tahmininde de çeşitli fonksiyonel veri analizi yöntemleri kullanımaktadır. Bu yöntemler arasında en popüler olan ve birçok çalışmaya öncülük eden yöntem Lee ve Carter (1992) tarafından ortaya atılmıştır. Yöntem, izleyen yıllarda sıklıkla kullanılmış ve çeşitli değişiklik ya da geliştirmeler önerilmiştir. Örneğin, Li ve ark (2004), düzensiz aralıklar üzerinde az sayıda gözlem olduğu durumlar için bir yöntem geliştirerek Çin ve Güney Kore için bir uygulama gerçekleştirmiştir. LC yöntemi, eksik verinin çok olduğu ya da zaman serisinin yeterince uzun olmadığı durumlarda kullanışlı olmamaktadır (Fang ve Härdle, 2015).

LC yönteminin geliştirilmesi sonucunda ortaya çıkan metotlardan en sık kullanılanı ise Hyndman ve Ullah (2007) tarafından öne sürülen yöntemdir. Çalışma, Fransa ölüm hızı ile Avusturalya doğurganlık hızı verilerinin fonksiyonel yaklaşımla ele alındığı bir tahmin yöntemi sunmaktadır. Çalışmada fonksiyonel veri analizi, parametrik olmayan düzleştirme yöntemleri ile sağlam istatistik yöntemlerinin birlikte kullanıldığı ve Lee-Carter yönteminin bir genelleştirmesi olarak nitelendirilebilecek yeni bir yöntem önerilmektedir. Yöntem yaşa bağlı düzgün fonksiyonlar elde edilmesini sağlamakta olup aykırı değerlere karşı da sağlamdır.

Hyndman ve Booth (2008), Hyndman ve Ullah (2007) tarafından kullanılan ve ortalama entegre kare hatasının minimize edilmesini öngören metotun, K'nın yeterince büyük seçilmesi durumunda, K seçimine duyarsız olduğunu belirtmiştir. Bu nedenle de $K = 6$ seçiminin yeterince büyük olduğu ve çalışmaları için seçildiği ifade edilmiştir.

Bu tez çalışmasında ise Hyndman ve Booth (2008) tarafından önerilen yöntem ve Türkiye İstatistik Kurumu tarafından 2001-2018 yılları için yayınlanmış yaşa özel doğurganlık hız verileri kullanılmıştır. Hyndman ve Booth (2008) tarafından izlenilen yönteme paralel şekilde, örneklem dışı tahmin hatasını minimize eden ve en küçük tahmin aralığı sunan λ değeri seçilmiştir. Bunun yanında, benzer çalışmalarda

kullanılan Akaike bilgi kriterine göre de karşılaştırma yapılmıştır. Sonuç olarak, $\lambda = 0$ bu kriterleri sağlayan değer olarak bulunmuştur.

Baz fonksiyonu sayısı K yeterince büyük seçilirse, metot K seçimine duyarsız hale gelmektedir. Benzer şekilde bu çalışmada da $K = 6$ seçilerek işlem yapılmış, ancak $K = 5$ seçildiğinde toplam varyansın tamamının açıklandığı görülmüştür. Ayrıca $\lambda = 0$ seçimi ile birlikte farklı K değerleri için yapılan değerlendirmede $K = 5$ seçiminin en düşük hata ve Akaike bilgi kriterini sağladığı görülmüş olup bu seçimle devam edilmiştir. Birinci baz fonksiyonuna ait katsayıya göre doğurganlık hızının genel olarak tutarlı şekilde düştüğü görülmektedir. Bu düşüşü yalnızca 2006-2008 arasındaki hafif artış bozmaktadır. Baz fonksiyona göre ise, bu düşüş daha küçük ve daha büyük yaş gruplarında daha hızlı olmuştur. Diğer bir deyişle, küçük yaş grupları ile büyük yaş gruplarındaki doğurganlık hızı, ortalamaya göre daha hızlı düşmektedir.

2019-2023 arası doğurganlık hızının en yüksek olduğu yaşın 2019 yılı öncesine göre bir miktar artması tahmin edilmekle birlikte doğurganlık hızının en yüksek olduğu yaş aralığının aynı (25-20 yaş) kalacağı öngörülmektedir. Bunun yanında, 25-40 yaş arası gruplarda doğurganlık hızının zamanla artması tahmin edilmiştir. Diğer bir deyişle, bu yaş aralığı için tahmin edilen doğurganlık hızları, 2019 yılından 2023 yılına doğru gittikçe artmaktadır. 20 yaş altı grup için ise tam tersi durum söz konusudur. 2019-2023 arasında 20 yaş altı grup için doğurganlık hızının azaldığı tahmin edilmektedir.

5. SONUÇ VE ÖNERİLER

Fonksiyonel veri analizinin önem kazandığı uygulama alanlarından biri sağlık, biyomedikal ve biyoinformatik alanlarıdır. Bu kapsamda, bu tez çalışmasında, fonksiyonel veri analizi ve temel araçları tanıtılmış, bu analiz yöntemini kullanarak Türkiye'de yaşa bağlı doğurganlık hızı verilerinin tahmini uygulaması yapılmıştır.

Çalışmada 2019-2023 yılları yaşa özel doğurganlık hızları, 2001-2018 verisi ile Hyndman and Booth (2008) tarafından önerilen yöntem kullanılarak tahmin edilmiştir. Bu kapsamda, örneklem dışı tahmin hatasını minimize eden ve en küçük tahmin aralığı sunan λ değeri seçilmiştir. Bunun yanında, benzer çalışmalarda kullanılan Akaike bilgi kriterine göre de karşılaştırma yapılmıştır. Sonuç olarak, $\lambda = 0$ bu kriterleri sağlayan değer olarak bulunmuştur. Ayrıca, $K = 5$ seçildiğinde toplam varyansın tamamının açıklandığı görülmüştür. Bunun yanında, $\lambda = 0$ seçimi ile birlikte farklı K değerleri için yapılan değerlendirmede $K = 5$ seçiminin en düşük hata ve Akaike bilgi kriterini sağladığı görülmüş olup bu seçimle devam edilmiştir. Çalışma sonucunda, 2019-2023 arası doğurganlık hızının en yüksek olduğu yaş ile 25-40 yaş arası gruplarda doğurganlık hızının artması tahmin edilmiştir.

Sağlık alanındaki çalışmalarda sıklıkla kullanılması önerilen fonksiyonel veri analizini Türkiye verileri üzerinde uygulayan az sayıda çalışma bulunmaktadır. Bu çalışma, Türkiye sağlık verileri için bir tahminin fonksiyonel veri analizi kullanılarak yürütüldüğü ilk çalışmalardan olup yöntemin kullanılabilirliğini gösterme ve ileriki çalışmalar için öncülük etme amacındadır. Söz konusu yöntem ve uygulama farklı göstergeler için de uyarlanabilir olup farklı sağlık istatistiklerinin analizi ve tahmini için de kullanılabilecek; sağlık alanında politika oluşturma süreçlerine katkı sağlanabilecektir.

ÖZET

Fonksiyonel Veri Analizi ile Yaşa Özel Doğurganlık Hızlarının Tahmin Edilmesi

Teknolojinin gelişmesiyle birlikte farklı tiplerde daha fazla veri toplanması mümkün hale gelmektedir. Yeni veri tipleri ise yeni istatistiksel analiz yöntemlerini gerektirmektedir. Bu kapsamda fonksiyonel veri analizi son zamanlarda önem kazanmaktadır. Sağlık alanındaki çalışmalarda sıklıkla kullanılması önerilen fonksiyonel veri analizini Türkiye verileri üzerinde uygulayan az sayıda çalışma bulunmaktadır. Bu tez çalışmasında, fonksiyonel veri analizini ve temel araçları tanıtılmış; fonksiyonel veri analiz yöntemleri kullanılarak Türkiye'de yaşa bağlı doğurganlık hızı verileri tahmin edilmiştir. Çalışma, yöntemin kullanılabilirliğini gösterme ve ileriki çalışmalar için öncülük etme amacındadır.

Anahtar Kelimeler: Doğurganlık Hızı, Fonksiyonel Veri Analizi, Fonksiyonel Zaman Serisi Analizi

SUMMARY

Estimation of Age-Specific Fertility Rates with Funcional Data Analysis

With the advance of technology, more data in different types are being recorded. New types of data requires new statistical analysis methods. In this context, functional data analysis has been gaining importance. There are few studies in Turkey appliying this method, which is often recommended for health sector. In this thesis, functional data analysis and its main tools are introduced; and age-specific fertility rates for Turkey were estimated using functional data analysis tools. This study aims to demonstrate the applicability of the method and guide further studies.

Keywords: Fertility Rate, Funcional Data Analysis, Funcional Time Series Analysis

KAYNAKLAR

BİRCAN E., B., HYNDMAN, R. J., ve GERTİG, D. M. (2007). Forecasting age-specific breast cancer mortality using functional data models. Statistics in Medicine, **26(2)**, 458−470.

BOOR C. (2001). A Practical Guide to Splines. Springer-Verlag: New-York

BOOTH, H., HYNDMAN, R.J., TİCKLE, L., DE JONG, P., 2006. Lee–Carter mortality forecasting: a multi-country comparison of variants and extensions. Demographic Research, Volume **15 (9)**, 289-310.

BOOTH, H., TİCKLE, L., SMİTH, L., 2005. Evaluation of the variants of the Lee–Carter method of forecasting mortality: a multi-country comparison. New Zealand Population Rev. **31 (1)**, 13–34.

BÜLENT D., METİN Ö., TANER K., İllerin ve Bölgelerin Sosyo-Ekonomik Gelişmişlik Sıralaması Araştırması (2003), DPT Yayın No.2671, Ankara, 2003.

CRAN Team and LANG D. T. (2020). RCurl: General Network (HTTP/FTP/...) Client Interface for R. R package version 1.98-1.2. https://cran.r-project.org/web/packages/RCurl/index.html

CUEVAS, A. (2014). A partial overview of the theory of statistics with functional data. Journal of Statistical Planning and Inference, **147**:1–23.

FANG, L. AND HÄRDLE, W. (2015). Stochastic population analysis: a functional data analysis. Technical report, SFB 649 Discussion Paper Series 2015.

H. L. SHANG, H. BOOTH AND R. J. HYNDMAN. (2011). Point and interval forecasts of mortality rates and life expectancy: A comparison of ten principal component methods. Demographic Research, **25**:173-214.

Hadley W., BRYAN J., KALİCİNSKİ M., VALERY K., LEİTİENNE C., COLBERT B., HOERL D., MİLLER E. (2019). readxl: Read Excel Files. R package version 1.3.1. https://cran.r-project.org/web/packages/readxl/index.html

HYNDMAN R. J. and SHANG H. L. (2020). ftsa: Functional Time Series Analysis. R package version 5.9. https://cran.r-project.org/web/packages/ftsa/index.html

HYNDMAN R.J., BOOTH H., TİCKLE L., MAİNDONALD J. (2019). demography: Forecasting Mortality, Fertility, Migration and Population Data. R package Version 1.22. https://cran.r-project.org/web/packages/demography/index.html

HYNDMAN RJ, SHANG HL. (2010). Rainbow plots, bagplots, and boxplots for functional data. Journal of Computational and Graphical Statistics **19**:29–45

HYNDMAN, R. J., H. BOOTH (2008). Stochastic Population Forecasts Using Functional Data Models for Mortality, Fertility and Migration, International Journal of Forecasting, No. **24(3)**, 323–342

JİANG X, WİNK S, WATER B, SCHNEİDER AK (2016). Functional analysis of high-content highthroughput imaging data. Journal of Applied Statistics, **44(11)**: 1903-1919.

JOHNSON, R. A. AND WİCHERN, D. W. (2002). Applied Multivariate Statistical Analysis. Fifth edition. New Jersey: Prentice Hail.

KADİR E., İSTEM KÖYMEN K. (2008). Düzgünleştirilmiş fonksiyonel ana bileşenler analizi ile imkb verilerinin incelenmesi. İstanbul Üniversitesi İktisat Fakültesi Ekonometri Ve İstatistik Dergisi, **8**: 132.

KRAUS D, PANARETOS VM. (2012). Dispersion operators and resistant second-order functional data analysis. Biometrika **99**:813–832

LEE R. D. AND L. R. CARTER. (1992). Modeling and forecasting U.S. mortality. Journal of the American Statistical Association, **87**:659-671.

LEE, R.D., 1993. Modeling and forecasting the time series of U.S. fertility: age distribution, range, and ultimate level. Internat. J. Forecasting **9**, 187–202.

LEVİTİN DJ, NUZZO RL, VİNES BW, RAMSAY JO (2007). Introduction to functional data analysis. Canadian Psychology, **48(3)**: 135-155.

Lİ N., R. LEE AND S. TULJAPURKAR. (2004). Using the Lee-Carter method to forecast mortality for populations with limited data. International Statistical Review, **72**:19-36.

NİHAN A. D. (2016). Uzaktan Algılama Verilerine Fonksiyonel Veri Analizi Yaklaşımının Uygulanması. Doktora Tezi. Mimar Sinan Güzel Sanatlar Üniversitesi Fen Bilimleri Enstitüsü.

Pin T. and MAECHLER M. (2020). cobs: Constrained B-Splines (Sparse Matrix Based). R package Version 1.3-4. https://cran.r-project.org/web/packages/cobs/index.html

RAMSAY JO, HOOKER G, GRAVES S. (2009). Functional Data Analysis with R and Matlab. Springer, New York.

RAMSAY, J. O. AND SİLVERMAN, B. W. (1997). Functional Data Analysis. Springer - Verlag, New York.

RAMSAY, J. O. AND SİLVERMAN, B. W. (2002). Applied Functional Data Analysis: Methods and Case Studies. Springer-Verlag, New York.

RAMSAY, J. O. AND SİLVERMAN, B. W. (2005). Functional Data Analysis. Springer, USA.

RENSHAW, A.E., HABERMAN, S., 2003. Lee–Carter mortality forecasting: a parallel generalized linear modelling approach for England and Wales mortality projections.Appl. Statist. **52 (1)**, 119–137.

SHANG H. L. and HYNDMAN R.J., (2019). rainbow: Bagplots, Boxplots and Rainbow Plots for Functional Data. R package version 3.6. https://cran.r-project.org/web/packages/rainbow/index.html

SHANG, H. L. (2012). Point and interval forecasts of age-specific fertility rates: a comparison of functional principal component methods, Journal of Population Research, **29(3)**, 249-267

SHANG, H. L. (2014). A survey of functional principal component analysis. AStA Advances in Statistical Analysis, **98(2)**:121–142.

ŞAFAK, Ç. (2019). Fonksiyonel Veri Analizi ve Sağlık Alanına Özgü Uygulama. Yüksek Lisans Tezi. Ankara Üniversitesi Sağlık Bilimleri Enstitüsü.

ULLAH S AND FİNCH CF (2013). Applications of functional data analysis: A systematic review. BMC Medical Research Methodology, **13**:43.

WANG JL, CHİOU JM, MÜLLER HG (2015). Review of functional data analysis. Annu. Rev. Statist: 1-41.

WOLF, D.A., 2004. Another variation on the Lee–Carter model. Paper Presented at the Annual Meeting of the Population Association of America, April 2004.

WOOD S. (2020). mgcv: Mixed GAM Computation Vehicle with Automatic Smoothness Estimation. R package version 1.8-33. https://cran.r-project.org/web/packages/mgcv/index.html

YARAEE K (2011). Functional Data Analysis with Application to MS and Cervical Vertebrae Data. Master thesis. University of Alberta. Edmonton.

ZEİLEİS A., LEİSCH F., HORNİK K., KLEİBER C., HANSEN B., and MERKLE E. C., (2019). strucchange: Testing, Monitoring, and Dating Structural Changes. R package version 1.5-2. https://cran.r-project.org/web/packages/strucchange/index.html

EKLER

Ek-1 Analize Ait R Kodları

#Kütüphaneler

```
library(readxl)
library(demography)
library(rainbow)
library(cobs)
library(mgcv)
library(strucchange)
library(RCurl)
library(ftsa)
```

#Fonksiyonlar

```
demogdata
plot.demogdata
lines.demogdata
points.demogdata
print.demogdata
summary.demogdata
set.upperage
is.el
get.series
smooth.demogdata
fert.curve
standardize
use.weight
```

fdm

print.fdm

summary.fdm

residuals.fdm

fdmMISE

ftsaMISE

forecast.fdm

print.fmforecast

plot.fmforecast

models

models.fmforecast