TARIM SİGORTALARINDA KONUMSAL KÜMELEME ÜZERİNE BİR ÇALIŞMA

A STUDY ON SPATIAL CLUSTERING IN AGRICULTURAL INSURANCE

İSMAİL GÜR

DOÇ. DR. ŞAHAP KASIRGA YILDIRAK
Tez Danışmanı

Hacettepe Üniversitesi
Lisansüstü Eğitim-Öğretim ve Sınav Yönetmeliğinin
Aktüerya Bilimleri Anabilim Dalı için Öngördüğü
YÜKSEK LİSANS TEZİ olarak hazırlanmıştır.

2017

İSMAİL GÜR 'ün hazırladığı **"Tarım Sigortalarında Konumsal Kümeleme Üzerine Bir Çalışma"** adlı bu çalışma jüri tarafından **AKTÜERYA BİLİMLERİ ANABİLİM DALI** 'nda **YÜKSEK LİSANS TEZİ** olarak kabul edilmiştir.

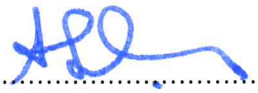
Prof. Dr. Fatih TANK

Başkan ..... 

Doç. Dr. Şahap Kasırga YILDIRAK

Danışman ..... 

Doç. Dr. Sevtap KESTEL

Üye ..... 

Yrd. Doç. Dr. Uğur KARABEY

Üye ..... 

Yrd. Doç. Dr. Könül BAYRAMOĞLU KAVLAK

Üye ..... 

Bu tez Hacettepe Üniversitesi Fen Bilimleri Enstitüsü tarafından YÜKSEK LİSANS tezi olarak onaylanmıştır.

Prof. Dr. Menemşe GÜMÜŞDERELİOĞLU

Fen Bilimleri Enstitü Müdürü

YAYINLAMA VE FİKRİ MÜLKİYET HAKLARI BEYANI

Enstitü tarafından onaylanan lisansüstü tezimin/raporumun tamamını veya herhangi bir kısmını, basılı (kağıt) ve elektronik formatta arşivleme ve aşağıda verilen koşullarla kullanıma açma iznini Hacettepe üniversitesine verdiğimi bildiririm. Bu izinle Üniversiteye verilen kullanım hakları dışındaki tüm fikri mülkiyet haklarım bende kalacak, tezimin tamamının ya da bir bölümünün gelecekteki çalışmalarda (makale, kitap, lisans ve patent vb.) kullanım hakları bana ait olacaktır.

Tezin kendi orijinal çalışmam olduğunu, başkalarının haklarını ihlal etmediğimi ve tezimin tek yetkili sahibi olduğumu beyan ve taahhüt ederim. Tezimde yer alan telif hakkı bulunan ve sahiplerinden yazılı izin alınarak kullanması zorunlu metinlerin yazılı izin alarak kullandığımı ve istenildiğinde suretlerini Üniversiteye teslim etmeyi taahhüt ederim.

☐ **Tezimin/Raporumun tamamı dünya çapında erişime açılabilir ve bir kısmı veya tamamının fotokopisi alınabilir.**
(Bu seçenekle teziniz arama motorlarında indekslenebilecek, daha sonra tezinizin erişim statüsünün değiştirilmesini talep etseniz ve kütüphane bu talebinizi yerine getirse bile, tezinin arama motorlarının önbelleklerinde kalmaya devam edebilecektir.)

☒ **Tezimin/Raporumun** 12.07.2020 **tarihine kadar erişime açılmasını ve fotokopi alınmasını (İç Kapak, Özet, İçindekiler ve Kaynakça hariç) istemiyorum.**
(Bu sürenin sonunda uzatma için başvuruda bulunmadığım taktirde, tezimin/raporumun tamamı her yerden erişime açılabilir, kaynak gösterilmek şartıyla bir kısmı ve ya tamamının fotokopisi alınabilir)

☐ **Tezimin/Raporumun ..... tarihine kadar erişime açılmasını istemiyorum, ancak kaynak gösterilmek şartıyla bir kısmı veya tamamının fotokopisinin alınmasını onaylıyorum.**

☐ **Serbest Seçenek/Yazarın Seçimi**

21 / 06 / 2017

İSMAİL GÜR

ETİK

Hacettepe Üniversitesi Fen Bilimleri Enstitüsü, tez yazım kurallarına uygun olarak hazırladığım bu tez çalışmasında;

- tez içindeki bütün bilgi ve belgeleri akademik kurallar çerçevesinde elde ettiğimi,
- görsel, işitsel ve yazılı tüm bilgi ve sonuçları bilimsel ahlak kurallarına uygun olarak sunduğumu,
- başkalarının eserlerinden yararlanılması durumunda ilgili eserlere bilimsel normlara uygun olarak atıfta bulunduğumu,
- atıfta bulunduğum eserlerin tümünü kaynak olarak gösterdiğimi
- kullanılan verilerde herhangi bir tahrifat yapmadığımı,
- ve bu tezin herhangi bir bölümünü bu üniversitede ya da bir başka üniversitede başka bir tez çalışması olarak sunmadığımı

beyan ederim.

21/06/2017

İSMAİL GÜR

ÖZET

TARIM SİGORTALARINDA KONUMSAL KÜMELEME ÜZERİNE BİR ÇALIŞMA

İsmail GÜR
Yüksek Lisans, Aktüerya Bilimleri Bölümü
Tez Danışmanı: Doç. Dr. Ş. Kasırga YILDIRAK
Haziran 2017, 84 sayfa

Bu çalışmanın amacı, buğday üretim yapılan ve hasar geçmişi bulunan konumlar için adil prim hesabı yapmaktır. Çalışmada konum, hasar gerçekleşme oranı ve daha sonra hasar gerçekleşme olasılığı değişkenlerini katmanlar şeklinde kümeleme yöntemi izlenmiştir. İlk olarak analizin birinci katmanını oluşturan konumsal özellikler incelenmiştir. Bu aşamada dağılım bazlı kümeleme esaslı sonlu karma von-Mises Fisher dağılımı ve uzaklık bazlı kümeleme yaklaşımı olarak küresel k-ortalamalar algoritması kullanılarak hasar bölgeleri için, optimal kümeleme sonuçları elde edilmiştir. İkinci aşamada her bir konum kümesi için üretimi yapılan tarımsal ürünün yüzde kaçının hasarlı olduğunu gösteren hasar gerçekleşme oranı değişkeninin dağılımı incelenmiştir. Ardından da farklı konum kümeleri içinde yer alan hasar gerçekleşme oranına göre yapılandırılmış alt kümeler için hasar gerçekleşme olasılığına dayalı kümeleme çalışması yapılmıştır. Her iki değişkenin de incelemesinde sonlu karma Beta dağılımı modeli kullanılmıştır. Her iki kümeleme yöntemi ile aktüeryal risk primi hesabı yapılmıştır.

Çalışmanın sonucunda sonlu karma von-Mises Fisher dağılımı ve küresel k-ortalamalar algoritması ile hesaplanan primlerin, konumsal özelliklere göre değişkenlik gösterdiği saptanmıştır. Buna ek olarak,sonlu karma von-Mises Fisher dağılımı ile hesaplanan

primin hasar gerçekleşme oranı ve olasılığındaki değişimi daha doğru bir şekilde ölçebildiği gösterilmiştir.

Anahtar Kelimeler: Yönsel Veri, Konumsal K-ortalamalar Kümeleme Algoritması, Sonlu Karma von-Mises Fisher Dağılımı, Sonlu Karma Beta Dağılımı modeli

ABSTRACT

A STUDY ON SPATIAL CLUSTERING IN AGRICULTURAL INSURANCE

İsmail GÜR
Master of Sciences, Department of Actuarial Sciences
Supervisor: Assoc. Prof. Dr. Ş. Kasırga YILDIRAK
June 2017, 84 pages

The purpose of this study is to make a fair premium estimation for wheat production. We prefer layer based clustering model where location, damage ratio and damage probability compose the layers. For the first layer, claim regions' location properties are examined. Optimal clustering results are obtained by using the spherical k-means algorithm, which is a distance-based clustering approach, and finite mixture von-Mises Fisher distribution which makes a density-based clustering. For the second layer, given the claim clusters, damage ratio values are fitted and for the third layer, given the first two layers, we model the damage probability obtained from claim data. Finite mixture of Beta distribution model is used for modelling the second and the third layers. The actuarial risk premium account is calculated by both clustering methods.

The result of the study shows that the premiums calculated by finite mixtures of von-Mises Fisher distribution and spherical k-means algorithm vary according to the locational characteristics. In addition, the change in damage ratios and probabilities can be more accurately measured by the premium using finite mixtures of von Mises Fisher distribution .

Key words: Directional Data, Spherical K-means Clustering Algorithm, Finite Mixtures of von-Mises Fisher Distribution, Finite Mixtures of Beta Distribution Model

TEŞEKKÜR

Tez çalışmamın her aşamasında değerli katkılarıyla ve eleştirileriyle yol gösteren, sonsuz sabırla beni her zaman çalışmaya teşvik eden ve güven veren danışmanım Sayın Doç. Dr. Şahap Kasırga YILDIRAK'a,

önemli yorum ve değerlendirmeleri ile katkıda bulunan jüri üyelerim Sayın Prof. Dr. Fatih TANK'a, Sayın Doç. Dr. Sevtap KESTEL'e, Sayın Yrd. Doç. Dr. Şeref HOŞGÖR'e, Sayın Yrd. Doç. Dr. Uğur KARABEY'e, Sayın Yrd. Doç. Dr. Könül BAYRAMOĞLU KAVLAK'a, Sayın Yrd. Doç. Dr. Şule ŞAHİN'e

ve tez çalışmamın gelişmesine yoğun katkısı bulunan Sayın Prof. Dr. Ashis SENGUPTA 'ya,

her türlü desteği esirgemeyen ve çalışmamın her aşamasında manevi olarak yanımda olan bölüm başkanımız Prof. Dr. Meral SUCU, bölüm başkan yardımcımız Yrd. Doç. Dr. Murat BÜYÜKYAZICI'ya ve Yrd. Doç. Dr. Yasemin GENÇTÜRK'e, mesai arkadaşlarım Arş. Gör. Mustafa Asım ÖZALP'e ve Arş. Gör. Çiğdem LAZOĞLU'na, Arş. Gör. Murat KIRKAĞAÇ'a, Arş. Gör. Betül Zehra KARAGÜL, Arş. Gör. Güven Şimşek'e, Arş. Gör. Ezgi NEVRUZ'a, Uzman Furkan YILDIRIM'a, oda arkadaşlarım Arş. Gör. Dr. Funda KARAMAN KUL'a, Arş. Gör. Samet GENÇGÖNÜL'e ve Arş. Gör. Müge YELDAN'a,

her zaman yanımda olan aileme içtenlikle teşekkür ederim.

İÇİNDEKİLER

Sayfa

ÇİZELGELER

Sayfa

ŞEKİLLER

Sayfa

1. GİRİŞ

Tarım sigortası, Tarım Sigortaları Havuzu (TARSİM) ile sistemi güvence altına alınmış, yaşanan risklere maruz kalan üreticilerin ürünleri teminat altına alınarak, üretimin sürdürülebilir olmasını sağlayan bir sigorta türüdür [1]. Tarım sigortası ürünleri, üç ana başlıkta incelenmektedir. Bunlar; tazminat bazlı tarım sigortaları, endeks bazlı tarım sigortaları ve ürün gelir ya da verim sigortalarıdır [2]. Bu çalışmada tazminat bazlı tarım sigortaları ele alınmıştır.

Tarım sigortalarının değerlemesinde üretilen ürünün fiyat değişim riski ve üretim rekolte riskinin dikkate alınması gerekmektedir. Bu risk türleri, diğer risk türlerinin aksine, meteorolojik ve coğrafik olaylardan, diğer bir deyişle, konumsal özelliklerden daha fazla etkilenmektedir. Konumsal özelliklerdeki değişimi daha iyi açıklayabilmek için yönsel veri analizine ihtiyaç duyulmaktadır. Tarım sigortalarında adil net risk priminin elde edilebilmesi için ise risk türlerine göre optimal risk sınıflandırılmasının yapılması gerekmektedir.

Bu çalışmada, 2010-2014 yılları arasında dolu riskine bağlı buğday ürünü tarım sigortasına ilişkin 5 yıllık hasar verisi; enlem-boylam bilgisi (il, ilçe, köykod, ada-parsel), hasar gerçekleşme oranı, hasar gerçekleşme olasılığı değişkenleri dikkate alınarak çok aşamalı konumsal kümeleme yöntemi ile optimal risk sınıflandırılması çalışması yapılmıştır.

Ortalama hasar gerçekleşme oranı, Eş. 1.1'de ve hasar gerçekleşme olasılığı Eş. 1.2 'de tanımlanmıştır.

$$\text{Hasar Gerçekleşme Oranı} = \frac{\text{Gerçekleşen Hasar Tutarı}}{\text{Tarımsal Üretim Yapılan Arazinin Ekonomik Değeri}} \tag{1.1}$$

$$\text{Hasar Gerçekleşme Olasılığı} = \frac{\text{Hasar Gerçekleşmiş Poliçe Sayısı}}{\text{Teminat Verilmiş Toplam Poliçe Sayısı}} \tag{1.2}$$

Çok aşamalı konumsal kümeleme çalışması aşamaları, Şekil 1.1'de gösterilmiştir. Bu çalışmanın ilk aşamasında hasar bölgelerinin konumsal özellikleri, uzaklık bazlı (küresel k-ortalamalar algoritması) ve dağılım bazlı (sonlu karma von-Mises Fisher dağılımı) kümeleme yöntemleri ile ayrı ayrı incelenmiştir. Bu yöntemlerle risk grupları belirlenmiş ve risk gruplarının konumsal farklılıkları incelenmiştir. Ardından her bir risk grubu

için, sırasıyla, ortalama hasar gerçekleşme oranı ve ortalama hasar gerçekleşme olasılığı değişkenlerine göre aşama aşama dallandırma çalışması yapılmıştır.

Şekil 1.1: Çok Aşamalı Konumsal Kümeleme Çalışması

Dünya'da tarım sigortaları, ilk kez 1949 yılında Halcrow [3] tarafından aktüeryal açıdan incelenmiştir. İlk aktüeryal prim hesabı çalışması ise 1986 yılında Skees ve Reed [4] tarafından yapılmıştır. Risk sınıflarının belirlenmesinin ardından hasar bölgeleri için ortalama sigorta bedeli, ortalama hasar gerçekleşme oranı, ortalama hasar gerçekleşme olasılığı değerleri kullanılarak ve bu değişkenlerin birbirinden bağımsız olduğu varsayımı altında aktüeryal net risk prim hesabı yapılmıştır. Hem uygun kümeleme modeli hem de optimal küme sayısı belirlenmesi amacıyla, ROC analizi kullanılmıştır. Analiz sonucunda 10 adet kümenin en iyi açıklanabilirlik düzeyine sahip olduğu sonucu elde edilmiştir.

Hasar bölgelerinin, konumsal özelliklerinin modellemesi için von-Mises Fisher dağılımından yararlanılmıştır. Bu dağılım, en temel küresel dağılım olmakla birlikte, R^3 uzayında çok değişkenli Gaussian dağılımı ile benzerlik göstermektedir. Ayrıca Türkiye dünya üzerinde sadece sınırlı bir bölümü kapsadığından, bu alandaki konumsal değişimi ölçmek için von-Mises Fisher dağılımın yeterli olduğu düşünülmüştür.

Hasar gerçekleşme oranı ve olasılığı değişkenleri için sınır koşulları atıldığında $(0, 1)$ aralığında tanım kümesine sahip olmaktadır. Bu durumda bu değişkenlerin değişiminin incelenmesinde sonlu karma Beta dağılımı kullanılmıştır.

İzleyen kesimde, yönsel veri kümelemesi ve tarım sigortalarında aktüeryal prim hesaplaması için yapılan çalışmalar kısaca özetlenmiştir.

1.1. Literatür Taraması

Yönsel veriler için kümeleme çalışmaları, uzaklık bazlı ve dağılım bazlı olmak üzere ikiye ayrılmaktadır.

Uzaklık bazlı kümeleme çalışmaları, ilk kez Salton ve McGill [5] tarafından yapılmıştır. Bu çalışmada uzaklık bazlı kümeleme algoritmasının temelini oluşturan "kosinüs benzerliği" tanımlanmıştır. Ardından bu tanım Rasmussen [6] tarafından k-ortalamalar kümeleme algoritmasında kullanılmıştır. Bu çalışma konumsal k-ortalamalar kümeleme çalışmalarına ışık tutmuştur.

Dhillon ve Modha [7] 2001 yılında konumsal k-ortalamalar kümeleme çalışmasını yapmışlardır. Dhillon ve Modha bu çalışmasında "kosinüs benzerliği", konsept vektörleri tanımlamalarını ve k-ortalamalar algoritması kullanarak büyük metin verilerini kümelemeyi başarmışlardır.

2010 yılında Maitra ve Ramler [8] tarafından çalışmada farklı uygulama alanları denenmiştir. Maitra ve Ramler'in çalışmasında hem tomurcuklanan maya için gen ifadeleme verisi kullanılırken hem de metin verisi kullanılarak farklı alanlardan uygulamalar yapılmıştır.

Yönsel verilere ait dağılım bazlı kümeleme çalışmalarına bakıldığında, Banerjee ve diğerlerinin [9] 2003 ve 2005 yıllarında kümeleme çalışmaları dikkati çekmektedir. Bu çalışmalarda yönsel dağılım olarak von-Mises Fisher dağılımı kullanılmıştır [10]. Ayrıca her iki çalışmada da ağırlıklı atama ve kesin atama algoritmaları tanıtılmıştır. Çalışmanın uygulama kısmında, bu algoritmaları küçük ve büyük sayıdaki verisetinde uygulayarak karşılaştırma yapmışlardır.

Hasar gerçekleşme oranı ve hasar gerçekleşme olasılığı değişkenleri gibi oran veya göreceli olasılık oranları değişkenleri için kullanılmak üzere beta regresyonu 2004 yılında Ferrari ve Cribato-Neto [11] tarafından ele alınmıştır. Bu çalışmanın ardından, Espinheira ve diğerleri [12] 2008 yılında, bu regresyon modeli tanımlayıcıları için artıkları incelemiştir.

Dünya'da tarım sigortaları değerleme çalışmaları ilk kez 1986 yılında Skees ve Reed [4] tarafından yapılmıştır.

Miranda [13], 1991 yılında alan verim sigortası üzerine alternatif bir model olarak komşu bölgelerin verimini de hesaba katmıştır.

Goodwin [14],1994 yılındaki çalışmasında çok bileşenli tarım sigortasını ele alarak ters seçim üzerine çalışmalar ortaya koymuştur.

1997 yılında Knight [15], 1980'den itibaren ABD'deki çok bileşenli tarım sigortasındaki ahlaki tehlike ve ters seçilim üzerine çalışma yapmıştır.

Skees ve diğerleri [16] 1997 yılında ABD'de yer alan Grup Risk Planı adlı planı ortaya koymuşlardır. Bu plan kapsamında alan verimliliği tekrar incelenmişlerdir.

Goodwin ve Ker [17], 1998 yılında parametrik olmayan yöntemler ile alan-verim hesaplamaları yapmışlardır.

Goodwin [18] 2001 yılında gerek ahlaki tehlike gerekse ters seçilim gibi unsurlar nedeniyle özel tarım sigortasının pek de mümkün olamayacağını göstermiştir. Arıca Goodwin [19] 2004 yılında tarımsal üretimin risk modellemesi üzerine çalışmalar ortaya koymuştur.

Bu çalışmalara ek olarak, Nelson [20] 1990 yılında tarım sigortalarında fiyatlama için farklı dağılımları deneyerek beta dağılımının bitkisel ürün sigortasında kullanabileceğini göstermiştir.

Turvey [21] tarım sigortasına hava durumuna bağlı türev ürünü bakışı kazandırarak farklı bir çalışma ortaya koymuştur.

Son yıllardaki tarım sigortaları üzerine yapılan çalışmalara bakıldığında, tarım sigortaları uygulamalarının tüm dünyada yaygınlaştığı görülmektedir.

2017 yılında Farzaneh ve diğerleri [22] İran için pamuk ürününü ele alırken, 2013 yılında Lou ve diğerleri [23] Çin Halk Cumhuriyeti için çay bitkisinin dolu sigortası için hava durumu endeksi ve alan verim anlayışını harmanlayarak fiyatlama çalışması yapmıştır. Çin Halk Cumhuriyeti için yapılan diğer çalışmalar Zhao ve diğerleri [24], Zhang ve diğerlerinin [25] yapmış olduğu Çin'in belirli bölümlerini inceleyen bölgesel çalışmalardır. 2014 yılında Farrin ve Murray [26], Zambiya için endeks sigortası çalışmasını ortaya koyarken 2015 yılında ise Ahmed ve Serra [27], İspanya için elma ve portakal ürünleri için tarım gelir sigortası çalışması yürütmüşlerdir. Bu çalışmaların yanında, Tack ve diğerleri [28], Ghimire ve diğerleri [29] iklim değişikliği düşüncesi ele alınarak değişken meteorolojik etkilerin sigorta üzerine etkisini incelemiştir.

Türkiye çerçevesinde tarım sigortaları kapsamında yapılan çalışmalara bakıldığında ise, endeks bazlı, tazminat bazlı ve verim bazlı sigorta türleri kullanılarak tarım sigortaları aktüeryal fiyatlama çalışmaları yapıldığı görülmektedir. Şahin ve diğerleri [30], buğday ürününü ele alarak bitkisel ürün sigortasına tazminat bazlı yaklaşarak prim hesabı çalışması yapmışlardır. Binici ve diğerleri [31] ise yine buğday ürününü ele alarak Konya ili için verim bazlı aktüeryal prim hesaplama çalışmaları ortaya koymuştur. Evkaya [32] doktora tezi çalışmasında Orta Anadolu illeri için hava durumu endeksli bitkisel ürün sigortası prim hesaplama çalışması yapmıştır.

1.2. Çalışma Planı

Tezin ikinci bölümünde, yönsel veri ve özellikleri anlatılacak ve yönsel verinin modellenmesi için kullanılan von-Mises Fisher dağılımı incelenecek ve bu dağılımı dikkate alan Sonlu Karma Modeli açıklanacaktır. Sonlu Karma von-Mises Fisher dağılımının parametre tahmini, EM (expectation maximization -beklentiençoklama) algoritması ile elde edilecektir.

Üçüncü bölümde, uzaklık bazlı kümeleme yöntemi olan küresel k-ortalamalar algoritması öncelikle "kosinüs benzerliği" tanımı yapılarak; konsept vektörü, hedef fonksiyonu ve algoritmanın teorik ve görsel olarak açıklaması yapılarak anlatılacaktır.

Dördüncü bölümde hasar gerçekleşme oranları ve hasar gerçekleşme olasılıklarını modellemek için kullanılacak Beta dağılımı anlatılacaktır. Ayrıca, bu bölümde parametre tahmin yöntemleri ve kümeleme çalışması için temel oluşturan Sonlu Karma Beta dağılımı açıklanacaktır.

Beşinci bölümde karma modellerin ve kümeleme algoritmalarının geçerliliklerin ve uygunluğunun test edilmesi için gerekli olan Silhoutte (Gölge) metodu açıklanacaktır.

Tezin uygulamasını içeren son bölümde ise, TARSİM (Tarım Sigortaları Havuzu) bünyesinde yer alan 2010-2014 yılları arasında dolu riskine bağlı buğday ürün sigortası için, sırasıyla, enlem-boylam bilgisi, hasar gerçekleşme oranı ve hasar gerçekleşme olasılıkları kullanılarak üç aşamalı kümeleme çalışması yapılacaktır.

Risk gruplarına ayrılan hasar bölgeleri için buğday ürünü baz alınarak dolu riskine bağlı saf aktüeryal prim hesabı yapılacaktır. En uygun kümeleme yöntemi ve küme sayısını bulmak için ROC Analizi kullanılacaktır.

2. YÖNSEL VERİ

Yönsel veri, açısal değerler ve yönsel büyüklük barındıran bir veri türüdür [37]. Bu veri türü, hem sosyal bilimlerinde hem de fen bilimlerinde kullanılabilmektedir. Örneğin, yönsel veri, biyoloji alanında, kuşların göç yollarını; meteoroloji alanında, rüzgar yönlerini, jeoloji alanında, depremlerin merkezlerini ya da arkeoloji alanında geçmişten günümüze uygarlıkların merkezlerini tespit etmek için kullanılmaktadır. olabilmektedir.

Zaman kavramının da yönsel veri olarak tanımlandığı çalışmalar bulunmaktadır. Zaman birimi gün, saat, ay olarak incelendiği zaman döngüsel süreçler olmuşmaktadır. Bu bakış açısıyla zaman kavramı ile ilgilenen bilim dalları da yönsel istatistikten yararlanabilmektedir.

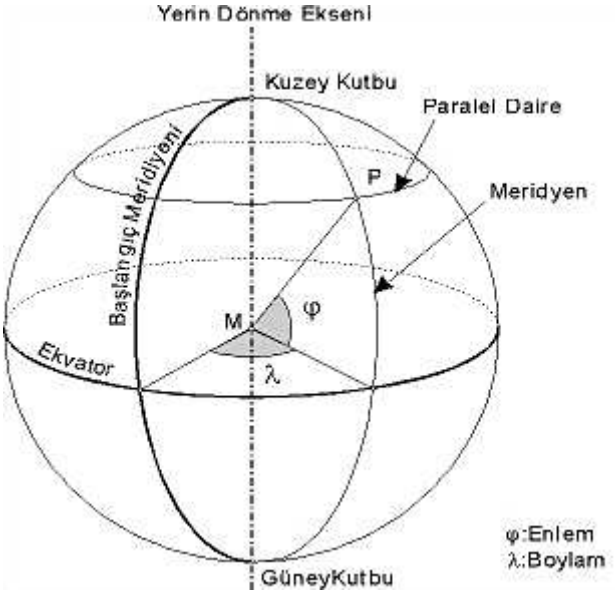
Yönsel veri, örneklerde görüldüğü gibi iki boyutlu yani dairesel ya da üç boyutlu yani küresel olabilmektedir. İki boyutlu yönsel veri, seçilen uygun bir başlangıç noktası ve dönüş yönü ele alınarak tanımlanır. Başlangıç noktasına örnek olarak Kuzey-Güney-Doğu-Batı gibi temel yönler olabilmekle birlikte belirli bir açı veya yön de olabilmektedir. Dönüş yönü ise, elde edilecek pozitif değerli açının saat yönünde ya da saat yönünün tersinde hareket etmesi anlamına gelmektedir.

Yönsel veri, ele alınan yön ya da açı değerinin herhangi bir büyüklüğü olmadığından dolayı, birim çemberinin çevresindeki noktalar ya da birim çemberinin orijininden bu noktalara giden birim vektörler şeklinde gösterilir. Bu dairesel gösterim sayesinde iki boyutlu yönsel veri içeren gözlemlere dairesel veri denilmektedir.

Benzer şekilde, üç boyutlu yönsel veri içeren gözlemlerde de iki açı, birim çemberin yüzeyinde yer alan noktalar ya da birim kürenin orijininden bu noktalara giden birim vektörler şeklinde temsil edilebilmektedir. Bu durumun en büyük örneği gezegenimizdir.

Dünya üzerindeki herhangi bir yerleşim yeri, enlem ve boylam koordinatları kullanılarak belirlenir. Şekil 2.1'de görüldüğü üzere enlem ve boylam değerlerinin belirlenmesinde Ekvator ve Greenwich standart alınır [33]. Bulunmak istenen konum bu standart kullanılarak elde edilir.

Yönsel veri, istatistiksel modellemelerde kendine özgü özelliklere sahiptir. Örnek olarak, iki boyutlu bir yönsel verinin matematiksel gösterimi ele alınırsa, çeşitli temellendirmelere göre farklı gösterimler elde edildiği görülür.



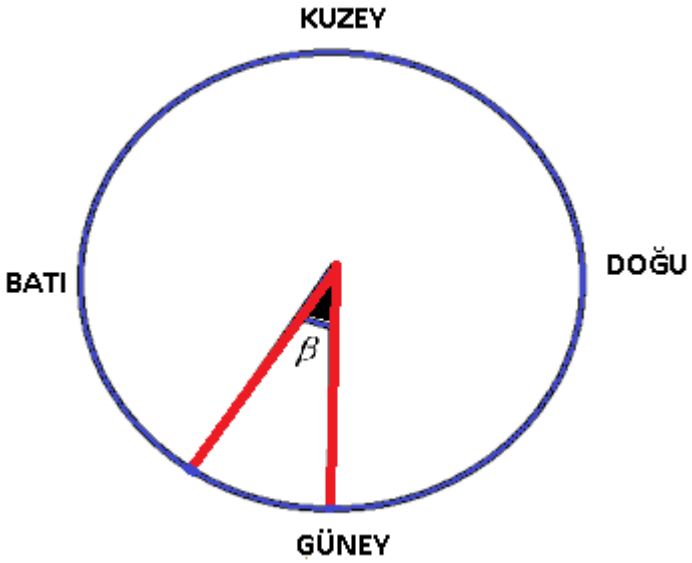
Şekil 2.1: Enlem ve Boylam Gösterimi

Şekil 2.2'de yer alan açısal değer herhangi bir rüzgar yönünü temsil ediyor olsun. Bu rüzgar yönünün başlangıç noktasına ve dönüş yönüne göre matematiksel gösteriminde değişiklikler gözlenecektir. İlk olarak başlangıç noktasını ve dönüş yönünü sırasıyla Güney yönü ve saat yönünde alınırsa, elde edilen açı β açısı olacaktır.

Başlangıç noktasını ve dönüş yönünü sırasıyla Batı yönü ve saat yönününün tersine alındığında ise açı $90 - \beta$ olacaktır, örnekleri çoğaltıp, kuzey ve doğu gibi diğer yönleri de ele aldığımızda daha farklı sonuçlar elde edilebilir. Sonuç olarak kullanılan veri değişmese de yapılan varsayımlara göre elde edilen sonuçlar ve buna bağlı olarak yapılacak yorumlar farklı olacaktır.

İki boyutta gerçekleşen bu farklılıkların benzeri üç boyutlu verilerde de görülmektedir. Bu nedenle bu veriler için kullanılacak istatistik hesaplamalardan elde edilen sonuçların gözlem verisiyle uyumlu sonuç vermesi ve rastlantısal seçilen başlangıç noktası ve dönüş yönünden bağımsız olması çok önemli bir unsurdur. Aksi takdirde hesaplamalarda kullanılan açısal veri ile modelleme sonucunda yapılan yorum tutarsız olacaktır. Bu nedenle, yönsel veriler analizinde en temel amaçlardan biri, yapılacak istatistiksel hesaplamaları rastgele seçilen başlangıç noktasına veya dönüş yönüne bağlı olmadan yapabilmektir.

Örneğin yönsel veri içeren gözlemler arasında sıralama yapılmak istendiğinde, rastlantısallık nedeniyle bu gözlemler arası sıralama yapılabilmesi mümkün olmayacaktır. Bunun nedeni bir yönsel veri ele alınırken başlangıç noktasının neresi olduğunun yanı



Şekil 2.2: Rüzgar Yönü Örneği

sıra dönüş yönünün, saat yönünde ya da saat yönünün tersinde olması durumuna bağlı olarak farklı sonuçlar elde edileceğinden, kullanılan model hatalı tanımlanabilecek ve sıralama hatası yapılabilecektir. Tanımlayıcı istatistikler olarak kullanılan ve bir veri için en temel açıklayıcı değerleri barındıran, örneklem ortalaması, varyansı, momentleri vb. gibi diğer istatistiklerde hatalı sonuçlar ortaya çıkabilecektir.

Ayrıca yönsel verilerin bir diğer temel özelliği döngüsel olmasıdır. Bu verilerin başlangıç ve bitiş noktalarında çakışma yaşanmaktadır yani yönsel verilerde $0 = 2 * \pi$ olduğundan dolayı periyodiklik söz konusudur. Örneğin bir β açısı,n tam sayı olmak koşuluyla, aynı zamanda derece ölçüsünde bakıldığında $\beta + p * 360$'a, radyan ölçüsünde bakıldığında ise $\beta + 2 * \pi$'ye eşit olabilmektedir. Bu nedenle, iki açının birbirine olan uzaklığını ele alan istatistiksel hesaplamalar yapılırken bu özellik dikkate alınmalıdır [34].

Ayrıca yönsel veriler girdi olarak kullanıldığında, çok değişkenli veri analizinde kullanılan standart Pearson korelasyon veya doğrusal regresyon gibi istatistiksel modellerde de anlamsız sonuçlar elde edilebilmektedir [35]. Bu nedenle gerek tanımlayıcı istatistikler gerekse kullanılmak istenen diğer istatistiksel yöntemler yönsel veriler için tekrar tanımlanmıştır.

2.1. Tanımlayıcı İstatistikler

Yönsel verilerin birim çemberin çevresinde yer alan noktalar ya da açılar olarak tanımlanabileceği bu bölümün başlangıcında belirtilmişti. İki boyutlu yönsel verilerin birim çemberdeki değerleri, polar koordinat düzlem veya dikdörtgensel koordinat sistemi ile tanımlanabilmektedir. Birim çember üzerindeki bu noktalar tanımlanırken O orijin değeri olan, (X, Y) düzlemine sahip dikdörtgensel koordinat sisteminden faydalanılacaktır.

Birim çemberde herhangi bir A noktası dikdörtgensel koordinat sistemindeki değerleri (X, Y) ile ya da polar koordinat sistemde (r,β) şeklinde belirtilebilmektedir. Polar koordinat sistemdeki r, orijin değerine olan uzaklığı, β ise açısal yönünü göstermektedir.

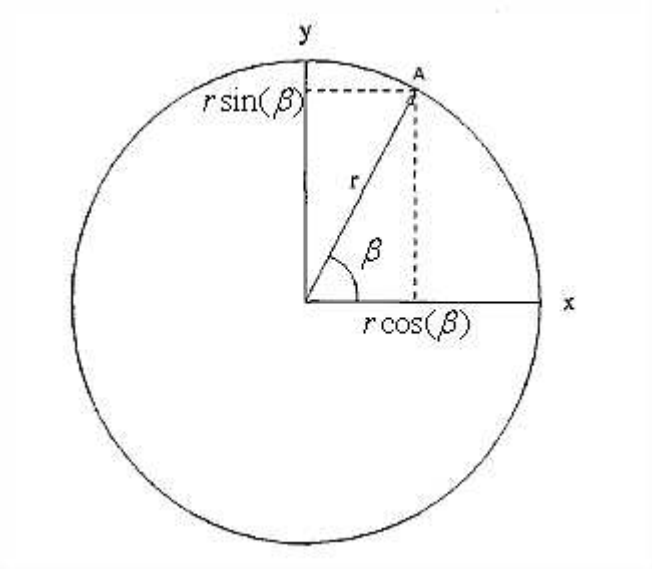
Polar koordinat sistem ile dikdörtgensel koordinat sistemi arasındaki geçişler, sinüs ve kosinüs trigonometrik fonksiyonları ile sağlanmaktadır.

Herhangi bir A noktasının polar koordinatları (r,β) olsun. Bu durumda dikdörtgensel koordinat sistemine geçişler Eş. 2.1 ve Eş. 2.2 şeklinde bulunur.

$$x = r\cos(\beta) \tag{2.1}$$

$$y = r\sin(\beta) \tag{2.2}$$

Yönsel verilerin analizinde, vektör büyüklüğünün değil yönlerin önemli olmasından dolayı bütün vektörler kullanım kolaylığı açısından, orijinden birim çembere doğru giden 1 birimlik (r = 1) vektörler olarak ele alınacakır. Böylece analiz edilecek her bir yön, aslında birim çember üzerindeki herhangi bir A noktasına denk gelecektir ve Şekil 2.3'teki gibi sadece açı ile gösterilebilecektir.



Şekil 2.3: Birim Çember Üzerindeki Herhangi A noktası

Birim çemberin çevresinde yer alan herhangi bir A noktasının, polar koordinat ve dikdörtgensel koordinat gösterimi Eş. 2.3 şeklindedir.

$$(1, \beta) \Leftrightarrow (x = \cos(\beta), y = \sin(\beta)) \tag{2.3}$$

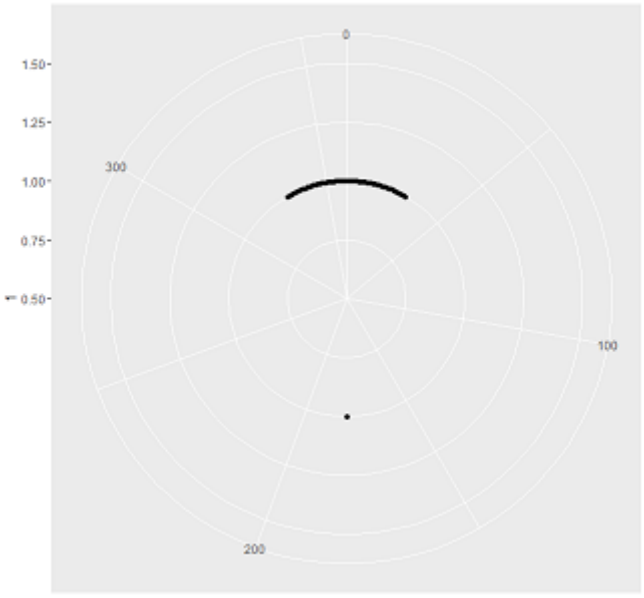
2.1.1 Ortalama Yön, Bileşke Uzunluğu, Ortalama Bileşke Uzunluğu

Verilen bir dairesel veri setinde ortalama yönü bulmak için, doğrusal ya da çoklu istatistiksel analizde olduğu gibi veri setindeki açıların aritmetik ortalamasını hesaplamak yanlış olacaktır; çünkü gerek örneklem ortalaması, gerekse örneklem standart sapması ve momentlerinin hesaplanmasında, yönsel verilerin, başlangıç noktası ve dönüş yönü seçimleri yapılan tüm hesaplamaları değiştirmektedir. Bu nedenle örneklem ortalaması hesaplanmasında aritmetik ortalama kullanmak uygun olmamaktadır.

Bu durum çeşitli örneklerle ele alınsın. Örneğin, ilk örnekte, örneklem ortalaması hesaplamasında neden aritmetik ortalama kullanmanın yanlış olduğu ikinci örnekte ise, örneklem ortalamasının başlangıç noktasına göre değişebileceği anlatılacaktır.

Örnek 1: Bir meteorolog araştırmasında Kuzey yönünü başlangıç noktası ve saat yönünü de dönüş yönü olarak belirleyip bir bölgeye esen rüzgarların yönlerini kaydetmiştir.

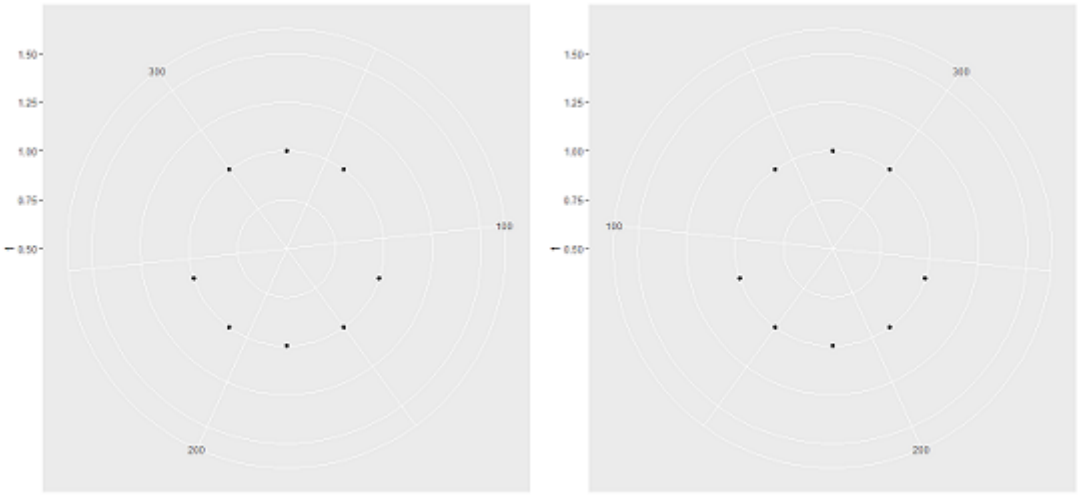
Şekil 2.4'te görüleceği üzere, tüm rüzgar yönleri kuzey, kuzeydoğu, kuzeybatı yönlerinde esmekte iken, aritmetik ortalaması alındığında, ortalama rüzgarın güney yönünde es-



Şekil 2.4: Örnek 1

tiği düşünülmektedir. Ancak bu değer, veri seti ile karşılaştırıldığında tam ters yönde kalmaktadır. Sonuç olarak, yönlerin ortalamasını bulmak için aritmetik ortalama kullanmak yanlış olacaktır.

Örnek 2: Bir biyolog Kuzey yönünü başlangıç noktası ve saat yönünü de dönüş yönü olarak belirleyip kuşların uçuş yönleri hakkında bir araştırma yapmak istemektedir. Kuşların uçuş yönleri sırasıyla 30, 150, 210, 330, 60, 120, 240, 330 derece olsun.



Şekil 2.5: Örnek 2

Şekil 2.5'te görüleceği üzere, tüm uçuş yönleri birbirine simetrik yönlerdedir, bu nedenle örneklem ortalaması başlangıç noktasına göre sürekli değişkenlik gösterecektir. Bu örnekte 180 derecelik bir başlangıç noktası farkı alınarak iki farklı grafik elde edilmiştir. Sonuç olarak, aritmetik ortalama hesaplanmasında başlangıç noktası ve dönüş yönü farklılıları farklı sonuçlar doğurabilmektedir.

Örneklerden de anlaşılacağı üzere, doğrusal veriler için temel istatistiksel analiz olan ve yoğun bir şekilde kullanılan aritmetik ortalama, gözlenen yönsel verilerin merkezinin ölçüsü anlamına gelmemektedir. Bu ortalama, başlangıç noktası ve dönüş yönünün seçimine bağlı bir fonksiyondur [36].

Tek bir yön doğrultusunda yığılmalı ya da tek tepe noktası bulunan yönsel veri kümesi için uygun ve anlamlı bir ortalama bulunmalıdır. Bu hesaplama için, veri setindeki her gözlem birer birim vektör olarak düşünülür ve bu vektörlerinin bileşkesi alınarak bileşke yön vektörü kullanılır.

Veri kümesinde yer alan açısal veriler için bileşke vektörünün yönü hesaplaması şu şekildedir: $\beta_1, \beta_2, \beta_3 ... \beta_n$ açı cinsinden verilmiş dairesel gözlem kümesindeki veriler olsun. Her bir gözlem için polar koordinat sistemden dikdörtgensel koordinat sistemine dönüşümü Eş. 2.4'teki gibi ele alınır.

$$(\cos(\beta_i), \sin(\beta_i)), \quad i = 1, 2, 3...n \tag{2.4}$$

n adet birim vektörün bileşke vektörü ise Eş. 2.5 ile bulunur:

$$R = \left(\sum_{i=1}^{n} \cos \beta_i, \sum_{i=1}^{n} \sin \beta_i \right) = (C, S) \tag{2.5}$$

$$R = \| R \| = \sqrt{C^2 + S^2} \tag{2.6}$$

Eş. 2.6 ile elde edilen R bileşke vektörü için dairesel merkezi yön olarak açıklanan açısal değer $\bar{\beta}_0$ olarak gösterilsin. Bu açısal değerin polar koordinat sistemde gösterimi Eş. 2.7 şeklindedir.

$$\bar{\beta}_0 = \arg \left(\sum_{j=1}^{n} \cos \beta_j + i \sum_{j=1}^{n} \sin \beta_j \right) \tag{2.7}$$

Bu açısal değerin kosinüs ve sinüs denklemleri Eş. 2.8 ve Eş. 2.9'da verilmiştir.

$$\cos \bar{\beta}_0 = \frac{C}{R} \tag{2.8}$$

$$\sin \bar{\beta}_0 = \frac{S}{R} \tag{2.9}$$

Kosinüs ve sinüs denklemlerinin çözümünde Eş. 2.10'da yer alan arctanjant fonksiyonu kullanılır.

$$\bar{\beta}_0 = \arctan^* \left(\frac{S}{C} \right) \tag{2.10}$$

$$\bar{\beta}_0 = \left\{ \begin{array}{cc} \arctan(S/C) & \text{eğer } S \geq 0, C \geq 0 \\ \pi/2 & \text{eğer } C = 0, S > 0 \\ \arctan(S/C) + \pi & \text{eğer } \quad C < 0 \\ \arctan(S/C) \ + 2\pi & \text{eğer } C \geq 0, S < 0 \\ \text{tanımsız} & \text{eğer } S = 0, C = 0 \end{array} \right\} \tag{2.11}$$

Denklem çözümünün sonucunda $\bar{\beta}_0$ değeri, ortalama yön olarak adlandırılmaktadır. Tanjant fonksiyonunun tersi alırken dikkat edilmesi gereken $\tan(\beta) = \tan(\beta+\pi)$ olması durumudur. Bu nedenle herhangi β açısının 2 adet tersi elde edilmektedir. Ancak Eş. 2.11 kullanıldığında, C ve S ifadelerinin işaretlerine göre $[0, 2\pi]$ aralığındaki açının tanjant fonksiyonuna göre tersi bir adet olacak şekilde elde edilir.

Bu yöntemle, $\bar{\beta}_0$ değeri veri setinde yer alan yönsel verilerin merkezi olarak gösterilebilir ve başlangıç noktasından ve dönüş yönünden bağımsız olarak elde edilebilir.

Birim vektörlerin bileşke vektörünün uzunluğu $(0, n)$ aralığında değerler almaktadır. Ortalama bileşke uzunluğu ise ortalama yön, $\bar{\beta}_0$ ile ilişkilidir ve Eş. 2.12 ile bulunur:

$$\bar{R} = \frac{R}{n} \tag{2.12}$$

Bileşke vektörün tanımlanabilmesi için uzunluğunun sıfırdan büyük olması gerekmektedir. Bileşke vektör pozitif uzunlukta ise bu vektör yönü olarak hesaplanan $\bar{\beta}_0$, ortalama yön olacaktır. Ayrıca bileşke vektörün örneklem sayısına eşit olması durumunda bütün noktalarının aynı açısal değeri gösterdiği yani çakıştığı anlamına gelmektedir. [37].

Bileşke vektörün uzunluğun sıfıra eşit olması durumu, veri setinde dairesel noktaların daire üzerinde eşit uzaklıklarla dağıldıklarını ve verilerin herhangi bir yöne yığılma göstermediğini anlatmaktadır.

2.1.2 Yoğunlaşma Parametresi

Yoğunlaşma parametresi, yönsel verinin ortalama yön yakınlarında ne kadar yoğunlaştığını gösteren bir parametredir. Bu parametre κ ile gösterilir. Yoğunlaşma parametre-

sinin en çok olabirlik tahmini $\hat{\kappa}$, Eş 2.13 şeklinde bulunmaktadır.

$$A_1(\hat{\kappa}) = \frac{R}{n} = \bar{R} \tag{2.13}$$

Eş 2.13'te yer alan $\bar{R}$ ifadesi bileşke vektörünün ortalama uzunluğu olarak tanımlanır.$A_1(\kappa)$ fonksiyonu, sürekli artan bir fonksiyondur ve $[0, \infty)$ aralığındaki değer kümesini, [0,1) değer kümesine tanımlamaktadır. Eş 2.14'te yer alan $A_1()$ fonksiyonu iki Bessel fonksiyonunun oranını göstermektedir.

$$A_1(x) = \frac{I_1(x)}{I_0(x)} \tag{2.14}$$

Örneğin $\kappa = 0$ olduğunda, yönsel veriler tekdüze dağılım özelliği göstermekte iken $\kappa = \infty$ olduğunda yönsel veriler, ortalama yön üzerinde tepe olan bir nokta dağılım özelliği göstermektedir [38].

$A_1()$ fonksiyonu bilgisayar destekli yinelemeli çözümler ile bulunabilmektedir. Bu çözümler, tezin Kappa(Yoğunlaşma) Parametresi Tahmin Yöntemleri bölümünde detaylı incelenecektir.

2.1.3 Dairesel Varyans

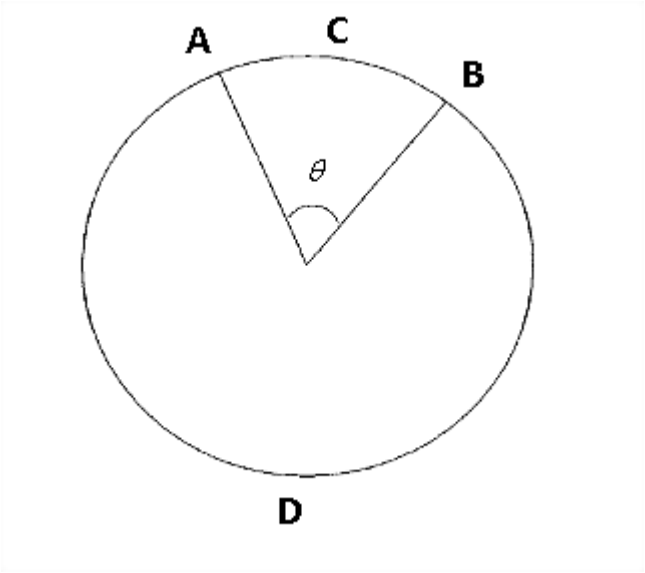
R bileşke vektörünün açısal değeri, $\bar{\beta}_0$ gözlem değerleri için ortalama yönü göstermekle birlikte, bileşke vektörün uzunluğu da tek bir yön doğrultusunda yığılmış ya da tek tepe noktası bulunan yönsel veri kümesi için gözlem değerlerinin merkezi etrafında ne kadar yoğunlaştığını gösterebilmektedir.

Eğer tüm birim vektörler aynı açısal değere sahip olursa, R bileşke vektörü n adet birim vektörün toplamı haline gelir ve örneklem sayısına eşit olur. Tam tersi, eğer veri seti çember üzerinde yoğunluk göstermeden eşit şekilde yayılırsa, R bileşke vektörü 0 değerine sahip olur.

Örneklem varyansı çember üzerinde uygun bir uzaklık ölçüsü şeklinde gösterilirse, $1-\bar{R}$ değerine eşit olur. İki nokta arasında çembersel uzaklık almanın bir yolu, bu noktalar arasında yer alan iki yay uzunluğunun kısa olanını almak olarak düşünülebilir. Yani, herhangi iki açı (α, β) değeri düşünüldüğünde, kısa olan uzaklık Eş. 2.15 ile bulunur.

$$\mathrm{d}_0\{\alpha,\beta) = \min(\alpha - \beta, 2\pi-(\alpha - \beta)) = \pi - |\pi - |\alpha - \beta|| \tag{2.15}$$

Örneğin, Şekil 2.8'de yer alan A ve B noktaları arasındaki uzaklık, ACB yayı ya da ADB yayı uzunluğu olabilir. Ancak ACB yay uzunluğu, ADB yay uzunluğuna göre daha



Şekil 2.6: ACB Yayı

kısa olduğundan ACB yay uzunluğu çembersel uzaklık olarak tanımlanır. Bir çember üzerindeki iki noktanın uzaklık değeri $[0, \pi]$ aralığındadır. A ve B noktaları arasındaki çembersel uzaklığın başka bir şekildeki tanımı ise Eş. 2.16'daki gibidir.

$$\mathrm{d}(\alpha,\beta) = (1 - \cos(\alpha - \beta)) \tag{2.16}$$

Çembersel uzaklık denkleminde yer alan α ve β değerleri A ve B gözlem değerlerinin sahip olduğu açısal değerlerdir. Eğer θ açısı A ve B gözlem değerlerinin arasında açısal değeri gösteriyor ise çembersel uzaklık denklemi bu parametreye bağlı monoton artan bir fonksiyon olacaktır. Eğer iki gözlem değeri arasındaki açı 0 ise uzaklık değeri de 0 olacaktır, iki gözlem değeri arasındaki açı π değerine eşit olması durumunda çembersel uzaklık değeri 2 olacaktır.

Doğrusal istatistiksel analizdeki örneklem varyansı s^2 ile benzer olarak yönsel veriler için yayılım ölçüsü olarak $1 - \bar{R}$ kullanılabilir.

Gözlem değerleri birim vektörler şeklinde olacak şekilde, $\{u_i, i = 1, 2, ..., n\}$ olarak tanımlansın. $D_z(u_1, u_2, ..., u_n)$ ise rastgele seçilen $z = (a, b)$ birim vektörüne göre örneklem yayılımını göstersin. Gözlem değeri u_i ile rastgele seçilen v vektörü arasındaki açı θ_i olsun. Bu açı $0 \leq \theta_i \leq \pi$ koşulunu sağlamak durumundadır. Eş. 2.16 kullanılarak, n adet gözlem değerinin v vektörüne olan dairesel uzaklıklarının ortalaması Eş. 2.17 şeklinde

bulunur.

$$\begin{aligned} \mathrm{D}_z(\mathrm{u}_1, \mathrm{u}_2, ..., \mathrm{u}_n) &= \frac{1}{n}\sum_{i=1}^{n} d(z, u_i) \\ &= \frac{1}{n}\left[n - \sum_{i=1}^{n}\cos(\theta_i)\right] \\ &= \frac{1}{n}\sum_{i=1}^{n}\left[1 - \cos(\theta_i)\right] \end{aligned} \tag{2.17}$$

Doğrusal istatistiksel modellerde olduğu gibi herhangi m değeri çevresinde yayılım ölçüsü $\sum (x_i - m)^2$ değerinin en küçük olması, m değerinin örneklem ortalamasına eşit olması ile mümkün olabilmektedir. Yönsel veriler için de aynı mantık çerçevesinde bir yaklaşıma gerek vardır. Yönsel veriler için Eş. 2.18'de yer alan D fonksiyonunu en küçük yapacak birim çember üzerindeki noktanın araştırılması gerekmektedir. Buna göre, z^* normalleştirilmiş bileşke vektörü olması durumunda $D_{z^*}(u_1, u_2, ..., u_n)$ yayılım ölçüsü en küçük değeri alır ve normalleştirilmiş bileşke vektörü $z^* = \left(\frac{C}{R}, \frac{S}{R}\right)$ olarak bulunur ve örneklem varyansı $1 - \bar{R}$ olarak bulunur [39].

2.1.4 Dairesel Standart Sapma

Doğrusal istatistiksel analizde kullanılan standart sapma tanımına benzer bir yapıda olan dairesel standart sapma fonksiyonu için, $\bar{R}$ fonksiyonunun bir dönüşümü kullanılmaktadır. Bu dönüşüm Eş. 2.18'de verilmiştir.

$$v = \left[-2\log_e(1 - V)\right]^{\frac{1}{2}} = \left[-2\log_e(\bar{R})\right]^{\frac{1}{2}} \tag{2.18}$$

Bu dönüşümde kullanılan $\bar{R}$, ortalama bileşke uzunluğun tanım aralığı $(0, 1)$ iken standart sapma(v) değerinin tanım aralığı $(0, \infty)$ olmaktadır.

2.1.5 Trigonometrik Momentler

Trigonometrik momentlerin elde edilmesi için, açıların kosinüs ve sinüs bileşenleri ayrı ayrı incelenmektedir. C ve S tanımlamaları sırasıyla kosinüs ve sinüs bileşenleri toplamı olarak gösterilmektedir ve kosinüs bileşeninin ortalaması Eş. 2.19'da, sinüs bileşeninin ortalaması ise Eş. 2.20'de verilmiştir.

$$\bar{C}_n = \frac{1}{n}\sum_{i=1}^{n}\cos(\beta_i) \tag{2.19}$$

$$\bar{S}_n = \frac{1}{n}\sum_{i=1}^{n}\sin(\beta_i) \tag{2.20}$$

$e^{i\beta} = (\cos\beta + i\sin\beta)$ eşitliğinden dolayı, kosinüs ve sinüs bileşenlerinin ortalamasının kutupsal formda gösterimi Eş. 2.21 şeklinde olacaktır.

$$\bar{C}_n + i\bar{S}_n = \frac{1}{n}\sum_{j=1}^{n} e^{i\beta_j} \tag{2.21}$$

Benzer şekilde, yüksek dereceden momentler elde edilebilmek için p fonksiyonu tanımlanır. p tam sayı olmak üzere, Eş. 2.22 edilir.

$$e^{ip\beta} = (\cos p\beta + i\sin p\beta) \tag{2.22}$$

Açısal değerler, birden n değerine kadar alındığında ise sırasıyla Eş. 2.23 ve Eş. 2.24 elde edilir.

$$\frac{1}{n}\sum_{j=1}^{n}\left(e^{i\beta_j}\right) = \frac{1}{n}\sum_{j=1}^{n}\cos(p\beta_j) + i\frac{1}{n}\sum_{j=1}^{n}\sin(p\beta_j) \tag{2.23}$$

$$\frac{1}{n}\sum_{j=1}^{n}\left(e^{i\beta_j}\right) = \bar{C}_n(p) + i\bar{S}_n(p) \tag{2.24}$$

$(\bar{C}_n(p), \bar{S}_n(p))$ değerleri örneklemin p. trigonometrik momentleri olarak tanımlanır.

3. SONLU KARMA VON-MISES FISHER DAĞILIMLARI YÖNTEMİ

Sonlu karma dağılımlar, her bir bileşeni uygun parametrik dağılım özelliği göstermesi varsayımı kullanılarak gözlem değerlerinin kümelenmesine olanak sağlamaktadır. Sonlu karma dağılımlar, sonlu sayıda, birden çok dağılımın dışbükey(konveks) birleşmesi ile türetilmektedir [40].

Sonlu karma von-Mises Fisher dağılımları yöntemi, standartlaştırılmış uzaklıkları yani birim küre üzerinde yer alan noktalar kullanılarak bu noktaların üzerinde dağılım bazlı kümeleme yöntemi kullanılmasına imkan sağlamaktadır. Bu yöntem ile sonlu karma von-Mises Fisher dağılımınına uygun örneklemler üretilebilmektedir [9]. Herhangi bir yönsel veri için en çok olabilirlik yöntemi tahmini ile EM algoritması kullanılarak sonlu karma von-Mises Fisher dağılımı parametreleri bulunabilmektedir [10].

Eğer veri kümesi olarak birim kürede yer alan noktalar incelenmek istenirse,uygulanacak karma dağılım için von-Mises Fisher dağılımı doğal bir seçenek olacaktır. Benzer şekilde, R^2 kümesinde yer alan veriler, diğer bir deyişle, birim çemberde yer alan gözlemler için ise von-Mises dağılımı uygun olabilecektir ve birim çemberdeki gözlem değerleri bu dağılım ile ifade edilebilecektir.

3.1. von-Mises Fisher Dağılımı

x herhangi bir d boyutlu birim vektörü göstermektedir. S^{d-1}, d boyutlu bir hiperküre (hypersphere) olsun. Yani $S^{d-1} = \{x \in R^d : \|x\| = 1\}$ olacaktır. Eş. 3.1 'de görüleceği üzere S^{d-1} üzerindeki olasılık öğesi dS^{d-1} ile gösterilsin ve S^{d-1} üzerinde polar koordinatlar (r, θ) ile ifade edilsin. $r = 1$ ve $\theta = [\theta_1, \theta_2, ..., \theta_{d-1}]$ olduğu kabul edilsin. Sonucunda $x_j = \sin\theta_1 .... \sin\theta_{d-1} \cos\theta_d, 1 < j < d$ ve $x_d = \sin\theta_1 .... \sin\theta_{d-1}$ olacaktır [41].

$$dS^{d-1} = \left(\prod_{k=2}^{d-1} \sin^{d-k}\theta_{k-1} \right) d\theta \tag{3.1}$$

3.1.1 von-Mises Fisher Dağılımı Olasılık Yoğunluk Fonksiyonu

Herhangi bir rastgele birim vektör $C_d(\kappa)e^{\kappa\mu^T x}dS^{d-1}$ olasılık ögesine sahip ise, bu birim vektörün, d boyutlu von-Mises Fisher dağılımına sahip olduğu söylenebilir. Bu dağılım

için normalleştirme parametresi $C_d(\kappa)$ Eş. 3.2'de verilmiştir.

$$C_d(\kappa) = \frac{\kappa^{d/2-1}}{(2\Pi)^{d/2} I_{d/2-1}(\kappa)} \tag{3.2}$$

Bu normalleştirme parametresinde yer alan I fonksiyonu birinci tip düzeltilmiş Bessel fonksiyonudur. Bessel fonksiyonunun genel formu Eş. 3.3'te gösterilmektedir. Eş. 3.3'de gösterilen Bessel fonksiyonundaki toplamın ilk bölümünün paydasında yer alan fonksiyon, Γ, gamma fonksiyonudur.

$$I_d(\kappa) = \sum_{k \geq 0} \frac{1}{\Gamma(k+d+1)k!} \left(\frac{\kappa}{2}\right)^{2k+p} \tag{3.3}$$

von Mises-Fisher dağılımında yer alan normalleştirme parametresinin hesabı için integrasyon ölçüsünün, tekdüze ölçüm normalleştirilmesi kullanılabilir [41]. Bu nedenle $C_d(\kappa)$ yerine $\frac{C_d(\kappa)(2\Pi)^{d/2}}{\Gamma(d/2)}$ ifadesi kullanılır. Bunun sonucunda von-Mises Fisher olasılık yoğunluk fonksiyonu Eş. 3.4 gibi elde edilir.

$$p(x, \mu, \kappa) = C_d(\kappa) e^{\kappa \mu^T x} \tag{3.4}$$

Bu olasılık yoğunluk fonksiyonunda μ ortalama yönü, κ ise yoğunlaşma parametresini göstermektedir. Bilindiği üzere, κ değeri, von-Mises Fisher dağılımından elde edilen birim vektörlerin, sahip oldukları ortalama yönün yakınlarında yoğunlaşıp yoğunlaşmadığını gösteren bir parametredir.

Örneğin $\kappa = 0$ olduğunda, bu olasılık yoğunluk dağılımı tekdüze dağılım özelliği göstermekte iken $\kappa = \infty$ olduğunda bu dağılım, ortalama yön üzerinde tepe olan bir nokta dağılım özelliği göstermektedir.

von-Mises Fisher dağılımı, yönsel istatistiğin en temel dağılımlarından biridir. Bu dağılım, R^d üzerinde yer alan veriler için kullanılan çok değişkenli Gaussian dağılımı ile benzer özelliklere sahiptir.

3.1.2 von-Mises Fisher Dağılımı ile En Çok Olabilirlik Yöntemi ile Parametre Tahmini

von-Mises Fisher dağılımına sahip n tane birim vektör örnekleminin log-olabilirlik fonksiyonu Eş. 3.5'te yer almaktadır.

$$n \log(C_d(\kappa)) + \kappa \mu^T r \tag{3.5}$$

Log-olabilirlik fonksiyonunda yer alan r, herhangi birim vektörlerin bileşke vektörüdür, $r = \sum_{i=1}^{n} x_i$. Log-olabilirlik fonksiyonunun, $\mu^T\mu = 1$ ve $\kappa \geq 0$ kısıtları altında, parametrelere göre türevlenmesinin ardından elde edilen fonksiyonun sıfıra eşitlenip çözülmesi ile en çok olabilirlik tahmini edicileri Eş. 3.6 ve Eş. 3.7'deki gibi elde edilir. Eş. 3.6'da yer alan $\hat{\mu}$ ortalama yönü gösterirken, Eş. 3.7'te yer alan ρ ifadesi ortalama bileşke yönü göstermektedir.

$$\hat{\mu} = r / \|r\| \tag{3.6}$$

$$\rho = -\frac{C'_d(\kappa)}{C_d(\kappa)} = \frac{\|r\|}{n} \tag{3.7}$$

Log-olabilirlik fonksiyonunun türevi alınırken, $\frac{1}{C_d(\kappa)}$ ifadesinin logaritmik türevi aldıktan sonra $A_d(\kappa) = -\frac{C'_d(\kappa)}{C_d(\kappa)}$ olarak yazıldığında ve ortalama bileşke vektörü $\rho = \frac{\|r\|}{n}$ olarak bulunur ve κ, yoğunlaşma parametresi için en çok olabilirlik tahmin edicisi $A_d(\hat{\kappa}) = \rho$ olarak elde edilir. $A_d(\kappa)$ fonksiyonu, Bessel fonksiyonlarına ait yineleme özelliği kullanılarak ifade edilebilir [37].

$$A_d(\hat{\kappa}) = -\frac{C'_d(\kappa)}{C_d(\kappa)} = \frac{I_{d/2}(\kappa)}{I_{d/2-1}(\kappa)} \tag{3.8}$$

Eş. 3.8'de yer alan $A_d(\kappa)$ fonksiyonu, sürekli artan bir fonksiyondur ve $[0, \infty)$ aralığındaki değer kümesini, [0,1) değer kümesine dönüştürmektedir. Ayrıca bu fonksiyon, Riccati eşitliğini de sağlamaktadır [42]. Riccati eşitliği Eş. 3.9'da verilmiştir.

$$A_d{}'(\kappa) = 1 - A_d(\kappa)^2 - \frac{d-1}{\kappa} A_d(\kappa) \tag{3.9}$$

κ parametresinin en çok olabilirlik yöntemi tahmin edicisini bulabilmek için çeşitli yaklaşımlar ve bilgisayar destekli yinelemeli yöntemler kullanılır.

3.1.3 Kappa (Yoğunlaşma Parametresi) Tahmin Yöntemleri

1. İlk tahmin yöntemi olarak, Mardia ve Jupp [43], düşük boyutlu veriler için Eş. 3.10 gibi basit bir yaklaşım kullanmışlardır. Ancak bu yaklaşım hem $\frac{\kappa}{d}$ oranı birden büyük veriler için hem de çok boyutlu veriler için doğru sonuçlar vermemektedir.

$$A_d(\kappa) \approx \frac{\kappa}{d} \tag{3.10}$$

2. Sonrasında, Banerjee ve diğerleri ise, κ tahmini için, $A_d(\kappa)$, kesirli gösteriminde kısaltmalara gitmişlerdir ve elde ettikleri bu yeni fonksiyonun çözümü ile tahmin

yapmışlardır. Bu kesme işlemi ile yoğunlaşma parametresini Eş. 3.11 şeklinde tahmin etmişlerdir. Bu tahmin yöntemi, Mardia ve Jupp'ın tahmin yöntemine göre daha doğru sonuçlar vermişlerdir [10].

$$\hat{\kappa} = \frac{\bar{R}d - \bar{R}^3}{1 - \bar{R}^2} \tag{3.11}$$

3. Ardından, Tanabe ve diğerleri ise, κ, yoğunlaşma parametresi için sınırlar elde etmişlerdir. Bu sınırlar, $A_d(\kappa)$ çözümünde yer alan Bessel eşitsizliği ile elde edilmiştir. Ayrıca Tanabe ve diğerleri, bu çalışmalarına ek olarak, κ yaklaşık çözümü için Eş. 3.12'de görülen sabit nokta iterasyonu yöntemi algoritması geliştirmişlerdir [44]. Eş. 3.12'de yer alan κ_l ve κ_u ifadeleri Eş. 3.13 de gösterilmektedir. Ayrıca yoğunlaşma parametrelerinin tahminde yer alan Φ_{2d} ifadesi ise Eş. 3.14'te verilmiştir.

$$\hat{\kappa} = \frac{\kappa_l \Phi_{2d}(\kappa_u) - \kappa_u \Phi_{2d}(\kappa_l)}{\Phi_{2d}(\kappa_u) - \Phi_{2d}(\kappa_l) - (\kappa_u - \kappa_l)} \tag{3.12}$$

$$\kappa_l = \frac{\bar{R}d - 2\bar{R}}{1 - \bar{R}^2} \leq \hat{\kappa} \leq \kappa_u = \frac{\bar{R}d}{1 - \bar{R}^2} \tag{3.13}$$

$$\Phi_{2d}(\kappa) = \bar{R}\kappa A_d(\kappa)^{-1} \tag{3.14}$$

4. Son yöntem olarak, Banerjee ve diğerlerinin kullandığı tahmin yöntemi ile Newton metodu birleşiminden elde edilen "Kesilmiş Newton Yaklaşımı" yöntemi daha kesin ve daha çabuk bir tahmin yöntemi olarak elde edilmiştir. Bu yöntemde iki adımlı Newton iterasyonu kullanılmaktadır [41].

$$A_d{}'(\kappa) = 1 - A_d(\kappa)^2 - \frac{d-1}{\kappa} A_d(\kappa) \tag{3.15}$$

$A_d(\kappa) - \bar{R} = 0$ çözümü kullanılarak iterasyonlar Eş. 3.15'de olduğu gibi tekrarlanmıştır. Newton yöntemine göre ilk adım Eş. 3.16'da verilmiştir.

$$\kappa_1 = \kappa_0 - \frac{A_d(\kappa_0) - \bar{R}}{1 - A_d(\kappa_0)^2 - \frac{d-1}{\kappa_0} A_d(\kappa_0)} \tag{3.16}$$

İkinci iterasyon sonucunda κ tahmini elde edilmiştir. Bu tahmin yöntemi, Tanabe ve diğerlerinin kullandığı modelle benzerlik göstermektedir. Her iki modelde iki adet $A_d(\kappa)$ fonksiyonuna ihtiyaç duyar. Ancak bu model, Tanabe ve diğerlerinin kullandığı modele göre daha hızlı cevap vermektedir.

3.2. Sonlu Karma von-Mises Fisher Dağılımında EM Algoritması

Karma dağılımlar, sonlu sayıda dağılımın belirli ağırlıklarla konveks bileşimi ile elde edilmektedir. Buna göre K bileşene sahip sonlu bir karma dağılım Eş. 3.17 şeklinde gösterilir.

$$f(x|\Theta) = \sum_{h=1}^{K} \alpha_h f_h(x|\theta_h) \tag{3.17}$$

Eş. 3.17'de yer alan f fonksiyonu, karma dağılımın olasılık yoğunluk fonksiyonunu, Θ ise α, bileşim olasılıklarından ve θ, dağılım parametrelerininden oluşan vektörü göstermektedir. f_h fonksiyonu, parametreleri θ_h olan von-Mises Fisher dağılımının olasılık yoğunluk fonksiyonunu göstermektedir. Karma dağılımlarda, bileşen ağırlıklarının sıfırdan büyük ve bileşen ağırlıkları toplamlarının bire eşit olması gerekmektedir. Sonlu karma von-Mises Fisher dağılımının en çok olabilirlik yöntemi ile parametre tahmini için EM algoritması kullanılmaktadır [9].

Sonlu karma von-Mises Fisher dağılımları özelliği gösteren herhangi bir noktadan örneklem oluşturulması için öncelikle herhangi α_h olasılıkla rastgele h. von-Mises Fisher dağılımı seçilir, ardından bu dağılıma ait $f_h(x|\theta_h)$ olasılık yoğunluk fonksiyonuna sahip bir nokta oluşturulur.

$X = \{x_1, x_2, ..., x_n\}$ ifadesini örneklemden çekilmiş n tane bağımsız noktadan oluşan bir veri seti olsun. $Z = \{z_1, z_2, ..., z_n\}$ ifadesi ise X veri setine uyumlu saklı rastlantı değişkenleri kümesini oluşturmaktadır. Bu Z rastlantı değişkeni kümesi, örneklemde oluşturulan noktaların hangi von-Mises Fisher dağılımından oluşturulduğunu göstermektedir. Örneğin, herhangi x_i noktası $f_h(x_i|\theta_h)$ olasılık yoğunluk fonksiyonuna sahip h. von-Mises Fisher dağılımdan çekilirse, $z_i = h$ olmaktadır. Bu rastlantı değişkeni, saklı raslantı değişkeni olarak da isimlendirilebilir.

Z kümesinde yer alan değerlerin bilindiği varsayımı altında, örneklemden çekilen noktalara ait log-olabilirlik fonksiyonu Eş. 3.18 şekildedir. En çok olabilirlik tahmin edicileri bu fonksiyonun maksimizasyonu sonucunda elde edilmektedir.

$$\ln(P(X, Z|\Theta) = \sum_{j=1}^{n} \ln(\alpha_{z_j} f_{z_j}(x_j|\theta_{z_j}) \tag{3.18}$$

Her bir z_j değerinin bilinmesi durumunda, en çok olabilirlik yöntemi ile tahmin edicileri bulmak kolay olacaktır. Ancak sonlu karma von-Mises Fisher dağılım modeline

böyle bir durum söz konusu değildir. Eş. 3.18'te belirtilen log-olabilirlik fonksiyonu Z rastlantı değişkeninin dağılımına bağlı olarak değişkenlik gösterecektir. Bu şekildeki log-olabilirlik fonksiyonlarına, tam veri log-olabilirlik fonksiyonu denmektedir.

Bu modelde, (X, Θ) değişkenlerine ait bilgilere sahip olunduğunda, en olası $(Z|(X, \Theta))$ koşullu dağılımına ait tahminler yapılabilmektedir. Bu tahmin, EM algoritmasının, beklenen değer adımını oluşturmaktadır. Sonlu karma von-Mises Fisher dağılımına ait EM algoritmasının beklenti adımında iki farklı yol izlenebilir ve bu yollar vasıtasıyla iki farklı şekilde saklı raslantı değişkenlerine ait dağılım tahmini yapılabilmektedir. Bu tahmin yöntemindeki farklılıklar iki farklı algoritma ile sağlanmaktadır. Bu algoritmalar sırasıyla ağırlıklı atama ya da kesin atama algoritmaları olacaktır [10].

3.2.1 Maksimizasyon Adımı

EM algoritmasının bu adımında parametre tahmini yapılmaktadır. Öncelikle,bu adım için $p(h|x_i, \Theta) = p(z_i = h|x = x_i, \Theta)$ eşitliği varsayılsın. Bir başka deyişle, tüm verilen noktalar ve h değerleri için sonsal olasılık değerlerinin bilindiği düşünülsün.

Aksi belirtilmediği sürece, tüm beklenen değer hesaplamaları $(Z|(X, \Theta))$ rastlantı değişkenin dağılımı üzerinden hesaplanacaktır. Aksi takdirde $(Z|(X, \Theta))$ rastlantı değişkenine ait her değişim, sonrasında yapılacak beklenen değer hesaplamalarında ve buna bağlı olarak parametre tahminlerinde değişikliğe neden olacaktır.

Bu varsayım altında tam veri log-olabirlik fonksiyonunun p dağılımı üzerinden elde edilen beklenen değer fonksiyonu Eş. 3.19'da ifade edilmiştir.

$$E_p[\ln(P(X, Z|\Theta) = \sum_{h=1}^{k} \sum_{i=1}^{n} (\ln \alpha_h + \ln f_h(x_i, \theta_h))\, p(h|x_i, \Theta) \tag{3.19}$$

Parametre tahmini adımında başka bir deyişle maksimizasyon adımında, Θ parametreleri, log-olabilirlik fonksiyonunun beklenen değerini maksimize edecek şekilde sürekli yeniden tahmin edilmektedir. Beklenen değeri maksimize etmek için, beklenen değer fonksiyonunun α_h ve θ_h terimlerini içeren bölümlerinin ayrı ayrı maksimize edilmesi gerekmektedir; çünkü α_h ve θ_h parametreleri aralarında yani bileşen parametreleri ile karma dağılım parametreleri arasında bağımlılık söz konusu değildir.

α_h, ağırlık değerlerini elde etmek için $\sum_{h=1}^{h} \alpha_h = 1$ kısıtı altında λ Lagrange çarpanı kullanılır. Lagrange amaç fonksiyonunun her bir α_h değerine göre kısmi türevleri alınıp

çözüldüğünde, Eş. 3.20 elde edilir.

$$\hat{\alpha}_h = \frac{\sum_{k=1}^{n} p(h|x_k, \Theta)}{n} \tag{3.20}$$

Sonrasında ${\mu_h}^T \mu_h = 1$ ve $\kappa_h \geq 0$ kısıtları altında ortalama yön ve yoğunlaşma parametresi tahmininin yapılabilmesi için $\lambda_1, \lambda_2 ..., \lambda_k$ Lagrange çarpanları kullanılarak, Lagrange fonksiyonundan yararlanılır. Bu Lagrange fonksiyonu Eş. 3.21 şekilde bulunur.

$$\begin{aligned} L(\{\mu_h, \kappa_h, \lambda_h\}_{h=1}^{k}) = & \sum_{h=1}^{k} \left[\sum_{i=1}^{n} \ln(C_d(\kappa_h)) p(k|x_i, \Theta) \right] \\ & + \sum_{h=1}^{k} \left[\sum_{i=1}^{n} \kappa_h \mu_h^T x_i p(h|x_i, \Theta) + \lambda_h \left(1 - \mu_h^T \mu_h\right) \right] \end{aligned} \tag{3.21}$$

Lagrange fonksiyonun $\mu_h, \kappa_h, \lambda_h$ parametrelerine göre kısmi türevleri alındıktan sonra sıfıra eşitlenmesi ile, ortalama yön değeri Eş. 3.22'te gösterilmiştir.

$$\hat{\mu}_h = \frac{\sum_{k=1}^{n} p(h|x_k, \Theta) x_k}{\left\| \sum_{k=1}^{n} p(h|x_k, \Theta) x_k \right\|} \tag{3.22}$$

Aynı yöntemle, yoğunlaşma parametresi, Eş. 3.23 kullanılarak bulunur [10].

$$A_d(\hat{\kappa}_h) = \frac{\left\| \sum_{j=1}^{n} x_j p(h|x_j, \Theta) \right\|}{\sum_{j=1}^{n} p(h|x_j, \Theta)} \tag{3.23}$$

κ, yoğunlaşma parametresin tahmininde kullanılan A_d fonksiyonu Bessel fonksiyonlarının oranı şeklindedir ve Eş. 3.24 şeklinde gösterilebilir.

$$A_d(\kappa) = \frac{I_{d/2}(\kappa)}{I_{d/2-1}(\kappa)} \tag{3.24}$$

Bu algoritmada yoğunlaşma parametresinin tahmini için $\rho = A_d(\kappa)$ eşitliği kullanılır. Yoğunlaşma parametresi tahmini yöntemlerinden, Banerjee ve diğerlerinin [9], Tanabe ve diğerlerinin [44] ya da kesilmiş Newton yaklaşımı yöntemlerinden [41] herhangi biri kullanılarak κ, yoğunlaşma parametresi elde edilir.

3.2.2 Beklenti Adımı: Dağılım Tahmini

Bu adımda, $(Z|(X, \Theta))$ dağılımını elde etmek için kullanılacak iki farklı yöntem bulunmaktadır. Bu iki farklı yöntem sayesinde iki farklı şekilde sonsal dağılımlar hesaplanmaktadır. Beklenti adımında bulunan sonsal dağılımlar kullanılarak, maksimizasyon

adımında veriye ait olabilirlik fonksiyonu elde edilir ve bu fonksiyon maksimize edilerek parametre tahminleri yapılabilmektedir.

Sonsal dağılımların tahmini için ilk yöntem, ağırlıklı atama yöntemidir. Standart EM algoritmasında, saklı rastlantı değişkenlerinin dağılımı Eş. 3.25'te verilen şekilde kullanılmaktadır. Saklı rastlantı değişkeninlerine ait bu dağılım ağırlıklı atama algoritmasında kullanılmaktadır [45].

$$p(h|x_i,\Theta) = \frac{\alpha_h f_h(x_i|\Theta)}{\sum_{j=1}^{k} \alpha_j f_j(x_i|\Theta)} \tag{3.25}$$

Dağılım tahminindeki ikinci yöntem ise kesin atama yöntemidir. Bu yöntemde saklı rastlantı değişkenlerine ait dağılım Eş. 3.26'da verilen şekilde kullanılır [46].

$$q(h|x_i,\Theta) = \left\{ \begin{array}{ll} 1, & \text{eğer } \; h = \arg\max_{h'} p(h'|x_i,\Theta) \\ 0, & \text{diğer} \end{array} \right\} \tag{3.26}$$

Banerjee ve diğerleri, 2005 yılındaki çalışmasında, kesin atama algoritmasında kullanılan q dağılımının konumsal kümeleme çalışması için optimal olduğunu göstermişlerdir [9]. Bu çalışmalarında, q dağılımı ile elde edilen beklenen değer fonksiyonunun, p dağılımı ile hesaplanan olabilirlik fonksiyonu için altsınır oluşturduğunu göstermişlerdir. Bu alt sınır dikkate alındığında, q dağılımı üzerinden hesaplanan beklenen değer, tam veri log-olabilirlik fonksiyonu ile sınırlandırıldığı için bu dağılımın kabul edilebilir ve makul olabileceğini söylemişlerdir [10].

3.2.3 Ağırlıklı atama algoritması

Algoritma 1 Ağırlıklı Atama Algoritması

Girdi: Enlem ve Boylam değerlerine sahip konumlar

Çıktı:Konumların ağırlıklı atama ile kümelenmesi

Başlat:$\alpha_h, \mu_h, \kappa_h, h = 1, 2..., k$

tekrar

EM algoritmasının Beklenen Değer kısmı

for $j = 1 : n$

 for $h = 1 : k$

  $f_h(x_j|\theta_h) \leftarrow C_d(\kappa_h) \exp(\kappa_h {\mu_h}^T x_j)$

 end

 for $h = 1 : k$

  $p(h|x_j, \Theta) = \frac{\alpha_h f_h(x_j|\Theta)}{\sum_{m=1}^{k} \alpha_m f_m(x_j|\Theta)}$

 end

end

EM algoritmasının Maksimizasyon kısmı

 for $h = 1 : k$

  $\alpha_h = \frac{1}{n} \sum_{m=1}^{n} p(h|x_m, \Theta)$

  $r_h = \sum_{m=1}^{n} x_m p(h|x_m, \Theta)$

  $\hat{\mu}_h = \frac{r_h}{\|r_h\|}$

  $\kappa_h \leftarrow {A_d}^{-1} \left(\frac{\left\| \sum_{m=1}^{n} p(h|x_m,\Theta) x_m \right\|}{\sum_{m=1}^{n} p(h|x_m,\Theta)} \right)$

 end

yakınsaklık sağlanana kadar

Ağırlıklı atama algoritması her bir noktaya olasılıksal etiketler atamaktadır. Bu olasılıklar, her bir küme için elde edilen sonsal olasılıklar ile bulunmaktadır. Sonsal olasılıklar $p(h|x_i, \Theta)$ ile gösterilmektedir.

Ağırlıklı atama algoritmasında uygulanan EM algoritmasının beklenti kısmında $p(h|x_i, \Theta)$ sonsal olasılık dağılımı kullanılmaktadır. Ardından EM algoritmasının maksimizasyon kısmına gelindiğinde bu aşamada en çok olabilirlik yöntemi ile parametre tahminleri yapılmaktadır.

Bu algoritmanın sonuç aşamasında, sonsal olasılık değerleri veri alınarak, herhangi

bir nokta için hangi kümeye ait sonsal olasılık değeri en büyükse, noktanın o kümeye ataması gerçekleştirilir [47].

3.2.4 Kesin atama algoritması

Algoritma 2 Kesin Atama Algoritması

Girdi: Enlem ve Boylam değerlerine sahip konumlar

Çıktı:Konumların kesin atama ile kümelenmesi

Başlat:$\alpha_h, \mu_h, \kappa_h, h = 1, 2..., k$

tekrar

EM algoritmasının Beklenen Değer kısmı

for $j = 1 : n$

for $h = 1 : k$

$f_h(x_j|\theta_h) \leftarrow C_d(\kappa_h) \exp(\kappa_h {\mu_h}^T x_j)$

end

for $h = 1 : k$

$$q(h|x_j, \Theta) = \left\{ \begin{array}{ll} 1, & \text{eğer } \ h = \arg\max_{h'} p(h'|x_j, \Theta) \\ 0, & \text{diğer} \end{array} \right\}$$

end

end

EM algoritmasının Maksimizasyon kısmı

for $h = 1 : k$

$$\alpha_h = \frac{1}{n} \sum_{j=1}^{n} q(h|x_j, \Theta)$$

$$r_h = \sum_{m=1}^{n} x_m q(h|x_m, \Theta)$$

$$\hat{\mu}_h = \frac{r_h}{\|r_h\|}$$

$$\kappa_h \leftarrow {A_d}^{-1} \left(\frac{\left\| \sum_{m=1}^{n} q(h|x_m, \Theta) x_m \right\|}{\sum_{m=1}^{n} q(h|x_m, \Theta)} \right)$$

end

yakınsaklık sağlanana kadar

Kesin atama algoritması vasıtasıyla q fonksiyonu ile elde edilen sonsal dağılımlar kullanılarak kümeleme yapılır. Kesin atama algoritması ile ağırlıklı atama algoritması arasındaki fark, beklenen değer adımlarında farklı sonsal dağılım kullanılmasından kaynaklanmaktadır. Kesin atama algoritmasında, ağırlıklı atama algoritmasında kullanılan p fonksiyonu yerine q fonksiyonu kullanılmaktadır. Kesin atama algoritmasında sonsal olasılıklar sadece 0 ve 1 değerlerini almaktadır.

Kesin atama algorimasında, ağırlıklı atama algoritması ile benzer şekilde, maksimizasyon adımında yapılan parametre tahmininde en çok olabilirlik yöntemi kullanılmaktadır.

3.2.5 Küresel K-ortalamalar algoritması

Algoritma 3 Küresel K-ortalamalar Algoritması

Girdi: Enlem ve Boylam değerlerine sahip konumlar

Çıktı:Konumların K adet ayrı ayrı kümelenmesi

Başlat:$\mu_h, h = 1, 2..., k$

tekrar

EM algoritmasının Beklenen Değer kısmı

$X_h \leftarrow \phi, h = 1, ..., k$

for $j = 1 : n$

$X_h \leftarrow X_h \cup \{x_j\}, h = \arg\max_{h'} {x_j}^T \mu_{h'}$

end

EM algoritmasının Maksimizasyon kısmı

for $h = 1 : k$

$$\mu_h \leftarrow \frac{\sum\limits_{x \in X_h} x}{\left\| \sum\limits_{x \in X_h} x \right\|}$$

end

yakınsaklık sağlanana kadar

Küresel K-ortamalar algoritması, hem ağırlıklı atama hem de kesin atama algoritmasının özel bir durumu olarak düşünülebilir. Her iki algoritma için de belirli kısıtlar tanımlandığında küresel k- ortalamalar algoritmasına benzer özellikler taşımaktadır.

Bu algoritmanın, kesin atama algoritmasına benzemesi için ilk olarak her bir k. von-Mises Fisher dağılımının eşit yoğunlaşma parametresine sahip olduğu varsayımı kullanılmaktadır. Bu varsayıma ek olarak, karma dağılımda kullanılan ağırlıkların birbirine eşit olması durumu yani, $\alpha_h = 1/k$ kullanılmıştır.

Küresel k-ortalamalar algoritmasının, ağırlıklı atama algoritmasının özel bir durumu olması için ise, tüm bu varsayımlara ek olarak tüm bileşenler için yoğunlaşma parametresi, $\kappa_h = \kappa \rightarrow \infty, \forall h$ yani sonsuz olarak alınacaktır [9].

Bu varsayımlar altında, EM algoritmasında yer alan beklenti adımında, kümelenecek olan nokta ile küme içindeki noktalar arasında kosinüs benzerliği hesabına dayalı ola-

rak yakınlık tanımı yapılmıştır. Bu yakınlık tanımına bağlı olarak noktanın en yakın kümeye ataması öngörülmüştür [48].

Böylece, x_i noktası $h_i* = \arg\max_h {x_i}^T \mu_h$ kümesine atanabilecektir; çünkü $p(h*|x_i, \Theta) = \lim_{\kappa \to \infty} \frac{e^{\kappa {x_i}^T \mu_{h*}}}{\sum_{h=1}^{k} e^{\kappa {x_i}^T \mu_h}} \to 1$ ve $p(h|x_i, \Theta) \to 0, \forall h$ olacaktır.

Bu varsayımlar altında, ortak yoğunlaşma parametresinin tahmin edilmesine gerek kalmamaktadır; çünkü kesin atama fonksiyonu sadece ${x_i}^T \mu_h$ kosinüs benzerliğine bağımlı hale gelmiş olacaktır.

Küresel k-ortalamalar algoritmasında $\{\mu_h\}_{h=1}^{k}$, her bir küme için merkez değerler kümesi olarak verilmiş olsun. $X_h = \left\{x : x \in X, h = \arg\max_{h'} x^T \mu_{h'}\right\}$ olarak tanımlansın. $\{X_h\}_{h=1}^{k}$ kümesi, X verisetindeki noktalardan, küresel k-ortalamalar algoritması sonucu elde edilen k adet ayrık kümeyi göstermektedir.

Küresel K-ortalamalar algoritmasını daha detaylı incelemesi, bir sonraki bölümde yer almaktadır.

4. KÜRESEL K-ORTALAMALAR ALGORİTMASI

4.1. Kosinüs Benzerliği

$x_1, x_2, ..., x_n$, R^d kümesine ait birim küre üzerindeki noktaları göstermektedir. Bu noktalar için, iç çarpım hesabı, noktalar arası benzerliğin doğal bir ölçüsü olarak düşünülebilir. R^d içerisinde yer alan herhangi x ve y noktaları iki birim vektörünü, $\theta(x, y)$ ise aralarındaki açısal değeri göstermekte olsun. Buna göre, x ve y noktalarının iç çarpımı Eş. 4.1 'de verilmiştir.

$$x^T y = |x|\,|y| \cos\theta(x, y) = \cos\theta(x, y) \tag{4.1}$$

Bu eşitlik, $x^T y$ iç çarpımının aslında bu iki nokta arasındaki açının kosinüs değerine eşit olduğunu göstermektedir. Kosinüs benzerliği gerek hesaplama gerekse yorumlama açısından büyük kolaylık sağlamaktadır ve kümeleme işlemi gerektiren bir çok alanda kullanılmaktadır [49].

4.2. Konsept Vektörü

$x_1, x_2, ..., x_n$, R^d kümesinde yer alan birim vektörler olsun.$\Pi_1, \Pi_2, ..., \Pi_k$ ise k ayrık kümeye ayrılan bölümler olsun. Öyle ki, Eş. 4.2'de gösterildiği gibi, her bir kümenin birleşimi verisetini oluşturmaktadır ve Eş. 4.3'teki gibi, bu kümeler arasında ortak nokta yer almamaktadır [50].

$$\bigcup_{l=1}^{k} \Pi_l = \{x_1, x_2, ..., x_n\} \tag{4.2}$$

$$\Pi_j \bigcap \Pi_l = \emptyset \text{ eğer } j \neq l \tag{4.3}$$

Her bir $1 \leq l \leq k$ için ve Π_j kümesi için, ortalama vektör ya da birim vektörlerin ağırlık merkezi, Eş 4.4 'te verilen şekilde olacaktır.

$$m_l = \frac{\sum\limits_{x \in \Pi_l} x}{n_l} \tag{4.4}$$

Ağırlık merkezi hesaplamasında yer alan n_l ise Π_l kümesine ait eleman sayısını göstermektedir. Ortalama vektör m_l birim ölçüye sahip olmak zorunda olmakla birlikte, Konsept vektörü dikkate alınarak yön tayini yapılabilmektedir.

$$c_l = \frac{m_l}{\|m_l\|} \tag{4.5}$$

z, R^d içerisinde herhangi birim vektör olsun. Herhangi x vektörleri için Cauchy-Schwarz eşitsizliği kullanılarak Eşitsizlik 4.6 elde edilmektedir [51].

$$\sum_{x \in \Pi_l} x^T z < \sum_{x \in \Pi_l} x^T c_i \tag{4.6}$$

Bu sayede hesaplanan konsept vektörün, Π_l kümesine ait tüm noktalar için kosinüs benzerliğinde en yakın vektörü gösterdiği anlaşılmaktadır.

4.3. Hedef Fonksiyonu

Cauchy-Schwarz eşitsizliğinden yararlanarak, her bir Π_l kümesine ait konsept vektörün uygunluğunun ve kalitesinin ölçülmesi gerekir. Bu ölçüm ise $\sum_{x \in \Pi_l} x^T c_l$ toplamı ile yapılabilmektedir.

Verisetinde yer alan noktaların, tek bir kümede ve aynı değere sahip olması durumunda, bu kümeye ait ortalama uyumun kesin olasılığa yani 1 değerine sahip olabileceği düşünülür. Diğer taraftan, verisetinde yer alan noktaların, bir küme içerisinde çok değişkenlik göstermesi durumunda yani küme içerisinde yayılımın fazla olması durumunda, ortalama uyum düşük ve sıfır değerine yakın bir değer olması beklenir. $\sum_{x \in \Pi_l} x = n_l m_l$ ve $\|c_l\| = 1$ eşitlikleri düşünüldüğünde Eş. 4.7 elde edilmektedir.

$$\sum_{x \in \Pi_l} x^T c_l = n_l {m_l}^T c_l = n_l \|m_l\| {c_l}^T c_l = n_l \|m_l\| = \left\| \sum_{x \in \Pi_l} x \right\| \tag{4.7}$$

Eş. 4.7'de yer alan bu basit gösterim, her bir Π_l kümesinin kalitesinin, bu küme içerisinde yer alan birim vektörlerin L_2 normunda toplamı ile ölçülebileceğini göstermektedir.

Hedef fonksiyonu ise, bu toplamın her bir küme için toplanmasıyla oluşturulmaktadır. Buna göre ölçüt fonksiyonu Eş. 4.8 'deki gibi bulunur.

$$Q\left(\{\Pi_l\}_{l=1}^k\right) = \sum_{l=1}^{k} \sum_{x \in \Pi_l} x^T c_l \tag{4.8}$$

4.4. Küresel K-ortalamalar Algoritmasında Kullanılan Tanıtlar

Hedef fonksiyonunu maksimize eden, n adet konumsal vektörün k adet kümeye ayrılması planlanmaktadır [8]. Bu maksimizasyon probleminin çözümü için küresel k-ortalamalar algoritması ortaya konulmuştur. Eş. 4.9 ve Eş. 4.10'da küresel k-ortalamalar

algoritmasına ait hedef fonksiyonunun azalmayan bir fonksiyon olduğu gösterilmiştir.

$$\{\Pi_l\}_{l=1}^{k} = \arg\max_{\{\Pi_l\}_{l=1}^{k}} Q\left(\{\Pi_l\}_{l=1}^{k}\right) \tag{4.9}$$

Tanıt 1:Her bir $t \geq 0$ için,

$$Q\left(\left\{\Pi_l{}^{(t)}\right\}_{l=1}^{k}\right) \leq Q\left(\left\{\Pi_l{}^{(t+1)}\right\}_{l=1}^{k}\right) \tag{4.10}$$

Tanıt 1, küresel k-ortalamalar algoritmasında konsept vektör ve kümeler arasındaki ilişkiyi göstermektedir.Eş. 4.11'e göre, küresel k-ortalamalar algoritması adımlarında konsept vektörler ${c_i}^{(t)}$, $\Pi_i{}^{(t+1)}$ kümelerini meydana getirmekte iken, bu kümelerden de bir sonraki adım için konsept vektörler, ${c_i}^{(t+1)}$ oluşmaktadır [7].

İspat:

$$\begin{aligned}
Q\left(\left\{\Pi_l{}^{(t)}\right\}_{l=1}^{k}\right) &= \sum_{l=1}^{k}\left(\sum_{x\in\Pi_l{}^{(t)}} x^T {c_l}^{(t)}\right) \\
&= \sum_{l=1}^{k}\left(\sum_{j=1}^{k}\left(\sum_{x\in\Pi_l{}^{(t)}\cap x\in\Pi_j{}^{(t+1)}} x^T {c_l}^{(t)}\right)\right) \\
&\leq \sum_{l=1}^{k}\left(\sum_{j=1}^{k}\left(\sum_{x\in\Pi_l{}^{(t)}\cap x\in\Pi_j{}^{(t+1)}} x^T {c_j}^{(t)}\right)\right) \\
&= \sum_{j=1}^{k}\left(\sum_{l=1}^{k}\left(\sum_{x\in\Pi_l{}^{(t)}\cap x\in\Pi_j{}^{(t+1)}} x^T {c_j}^{(t)}\right)\right) \\
&= \sum_{j=1}^{k}\left(\sum_{x\in\Pi_j{}^{(t+1)}} x^T {c_j}^{(t)}\right) \\
&\leq \sum_{j=1}^{k}\left(\sum_{x\in\Pi_j{}^{(t+1)}} x^T {c_j}^{(t+1)}\right) = Q\left(\left\{\Pi_l{}^{(t+1)}\right\}_{l=1}^{k}\right)
\end{aligned} \tag{4.11}$$

Tanıt 2:Hedef fonksiyonu bir limite sahiptir.

$$\lim_{t\to\infty} Q\left(\left\{\Pi_l{}^{(t)}\right\}_{i=1}^{k}\right) \tag{4.12}$$

Tanıt 2'de küresel k-ortalamalar algoritmasının sonsuz kez tekrarlanması durumunda, hedef fonksiyon değerinin yakınsaklık göstereceği kanıtlanmıştır. Eş. 4.13'e göre hedef

fonksiyonu, artan bir seri olduğu ve sonunda sabit bir sayı ile sınırlandığı ispatlanmıştır [7].

$$Q\left(\left\{\Pi_l{}^{(t)}\right\}_{l=1}^{k}\right) = \sum_{l=1}^{k}\sum_{x\in\Pi_l} x^T c_l{}^{(t)} = \sum_{l=1}^{k} n_l{}^{(t)} \left\|m_l{}^{(t)}\right\| \leq \sum_{l=1}^{k} n_l{}^{(t)} = n \tag{4.13}$$

4.5. Küresel K-ortalamalar Algoritması İşleyişi

Adım 1:

Küresel k-ortalamalar algoritması, veri setinde yer alan noktaların rastgele bölümlenmesi ile başlanmaktadır, böylece $t = 0$ yani başlangıç zamanı için, $\{\Pi_i\}_{i=1}^{k}$, her bir küme belirlenmekte ve bu kümelere göre konsept vektörler,$\{c_i\}_{i=1}^{k}$ elde edilmektedir.

Adım 2:

Her bir nokta, $x_j, 1 < j < n$ için x_i noktasına kosinüs benzerliği bakımından en yakın konsept vektör bulunmaktadır. Eş. 4.14'te gösterildiği gibi, t zamanında elde edilen konsept vektörleri kullanılarak $t+1$ zamanında yer alacak yeni kümeler elde edilmektedir.

$$\Pi_i^{(t+1)} = \left\{x \in \{x_j\}_{i=1}^{n} : x^T c_i{}^{(t)} > x^T c_l{}^{(t)}, 1 < l < n, l \neq i\right\}, 1 \leq i \leq k \tag{4.14}$$

Eş. 4.14'te $\Pi_i^{(t+1)}$, $c_i{}^{(t)}$ konsept vektörlerine en yakın noktalar kümesini göstermektedir. Bu noktaların birden fazla konsept vektöre yakın olması durumunda, atama yakın kümeler arasında rastgele bir şekilde yapılmaktadır.

Adım 3:

Yeni konsept vektörler, Eş. 4.15'te gösterilen şekilde, yeni ortalama vektörlerin, bu ortalama vektörlerin uzunluğuna bölünmesi ile bulunmaktadır. Bu eşitlikte yer alan $m_i{}^{(t+1)}$ ifadesi, $Pi_i{}^{(t+1)}$ kümesinde yer alan noktaların ağırlık merkezi ya da ortalaması olarak düşünülebilir.

$$c_i{}^{(t+1)} = \frac{m_i{}^{(t+1)}}{\|m_i{}^{(t+1)}\|}, 1 \leq i \leq k \tag{4.15}$$

Adım 4: Durma Kriteri

$\Pi_i* = \Pi_i^{(t+1)}$, $c_i* = c_i^{(t+1)}, 1 \leq i \leq k$ olması durumunda durma kriterine ulaşılmakta ve algoritmadan çıkılmaktadır. Aksi takdirde, t birer artırılarak devam edilmekte ve 2. adımdan devam edilmektedir. Durma kriteri için Eş. 4.16, örnek olarak düşünülebilir.

$$\left| Q\left(\left\{\Pi_i{}^{(t)}\right\}_{i=1}^{k}\right) - Q\left(\left\{\Pi_i{}^{(t+1)}\right\}_{i=1}^{k}\right)\right| < \varepsilon \tag{4.16}$$

Bu kriterde, ε gibi bir eşik değeri belirlenmekte ve hedef fonksiyonunun $t, t+1$ zamanlarındaki değişimine bakıkmaktadır. Bu değişimin mutlak değeri belirlenen eşik değerinden düşük olması durumunda kümeleme algoritması durdurulmaktadır.

5. SONLU KARMA BETA DAĞILIMI MODELİ

5.1. Beta Regresyon Modeli

Beta regresyon modeli, yanıt değişkeni beta dağılımına sahip açıklayıcı değişkenler için kullanılmak üzere çeşitli kurallarla ortalama ve saçılım parametreleri ile ifade edilen bir regresyon modeli sunmaktadır.

Bu regresyon modelinde yanıt değişkeni, sürekli ve $(0, 1)$ aralığında tanımlı olan oran ya da olasılık değerleri olarak düşünülmektedir. Beta regresyon modelinde, regresyon katsayıları, yanıt değişkeninin ortalaması ya da logit bağ fonksiyonu kullanıldığında, olasılık oranı(odds ratio) ortalaması cinsinden yorumlanabilmektedir [11].

Beta regresyon modeli, yanıt değişkenine uygulanan çeşitli dönüşümlerle doğrusallaştırılarak bir regresyon modeli ortaya koymaktadır. Ayrıca beta regresyon modelinde parametre tahmin yöntemi olarak en çok olabilirlik yöntemi kullanılmaktadır [40].

Beta dağılımı, herhangi oran ya da olasılık değerlerinin modellenmesi için oldukça uygun bir dağılımdır. Beta dağılımı, dağılımın göstergeleri olan iki parametreye bağlı olarak, olasılık yoğunluk fonksiyonunda farklı değerlere sahip olabilmektedir.

Beta dağılımının olasılık yoğunluk fonksiyonu Eş 5.1'deki gibi gösterilmektedir.

$$f(z; p, q) = \frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)} z^{p-1}(1-z)^{q-1}, 0 < z < 1 \tag{5.1}$$

Bu olasılık yoğunluk fonksiyonunda p ve q parametreleri sıfırdan büyük olmalıdır. Olasılık yoğunluk fonksiyonununda Γ ifadesi ile gösterilen fonksiyon bir gamma fonksiyonu göstermektedir. Beta dağılımına sahip herhangi bir z bağımlı değişkeninin beklenen değeri ve varyansı sırasıyla Eş. 5.2 ve Eş. 5.3 şeklinde bulunmaktadır.

$$E(z) = \frac{p}{p+q} \tag{5.2}$$

$$V(z) = \frac{pq}{(p+q)^2(p+q+1)} \tag{5.3}$$

Beta dağılımında yer alan p ve q parametrelerinin birden büyük olması durumunda, z bağımlı değişkeninin modu, Eş. 5.4'te verilmiştir.

$$\text{mod}(z) = \frac{p-1}{p+q-2} \tag{5.4}$$

Beta dağılımının özel bir durumu olarak, p ve q parametreleri bire eşit olması durumunda, elde edilen beta dağılımı tekdüze dağılımına yakınsaklık göstermektedir [11].

5.2. Beta Regresyon Modelinde Parametre Tahmini

Beta regresyon modelinin başlıca amacı, beta dağılımına sahip bağımlı değişken için anlamlı bir regresyon modeli ortaya koymaktır ve girdi olarak kullanılan bağımsız değişkenlerin etkisini incelemektedir. Beta dağılımının olasılık yoğunluk fonksiyonu, p ve q parametreleri ile açıklanabilen bir fonksiyondur. Herhangi bir oranı ya da olasılığı modellemek için kullanılan bu regresyon modeli, bağımlı değişkenin ortalaması ve saçılımını modellemek için uygun olmalıdır.

Yanıt değişkeninin ortalama ve saçılım parametresi ile regresyon modeli elde etmek için, beta dağılımına uygun olacak şekilde p ve q parametrelerinin farklı bir şekilde yeniden tanımlanması gerekmektedir. Bu yeni tanımlamanın sonucunda sonucunda p ve q parametreleri yerini μ ve ϕ parametrelerine bırakacaktır ve bu parametreler $\mu = \frac{p}{p+q}$ ve $\phi = p+q$ olacaktır. Ayrıca bu yeni değişkenlerin dönüşümü $p = \mu\phi$ ve $q = (1-\mu)\phi$ eşitlikleri elde edilecektir. Buna göre, y bağımlı değişkeninin ortalamasının ve varyansının yeni parametrelerle gösterimi sırasıyla Eş. 5.5 ve Eş. 5.6'da verilmiştir.

$$E(z) = \mu \tag{5.5}$$

$$V(z) = \frac{V(\mu)}{1+\phi} \tag{5.6}$$

Eş. 5.6'da yer alan varyans fonksiyonunun payı olan $V(\mu)$ ifadesi $\mu(1-\mu)$ çarpımına denk olacaktır. Beta dağılımına uyumlu yanıt değişkeninin ortalama değeri değişmediği varsayımında, kesinlik parametresindeki, ϕ herhangi artış varyansın düşmesine neden olacaktır. Bu durumda değişken varyanslılıktan söz edilebilmektedir. Elde edilen yeni parametreler kullanılarak beta dağılımı olasılık yoğunluk fonksiyonu, Eş. 5.7'deki gibi elde edilir.

$$f(z,\mu,\phi) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)} z^{\mu\phi-1}(1-z)^{(1-\mu)\phi-1}, 0 < z < 1 \tag{5.7}$$

Beta olasılık yoğunluk fonksiyonunda yer alan yeni parametreler için $0 < \mu < 1$ ve $\phi > 0$ kısıtları dikkate alınmalıdır. Ayrıca, beta dağılımı, ortalaması 0.5 olması durumunda simetrik bir dağılım özelliği gösterirken, buna ek olarak kesinlik parametresi 2 olması durumunda, standart tekdüze dağılıma uyumlu hale gelecektir. Ayrıca bu dağılım, p ve q parametreleri birbirine eşit olduğunda ve limit durumunda standart normal

dağılıma yakınsama göstermektedir.

Beta regresyon modelinde, $z_1, z_2, ..., z_n$ birbirinden bağımsız rastlantı değişkenleri olmak üzere, herbir z_i, μ_i ve ϕ parametrelerine sahip beta dağılımına ait olsun. z_i'nin ortalama değeri aşağıdaki ifade şeklinde yazılabildiğinde, bu rastlantı değişkeni için beta regresyon modeli elde edilebilmektedir.

$$g(\mu_i) = \sum_{j=1}^{k} x_{ij}\beta_i = \eta_i \tag{5.8}$$

Eş. 5.8'de $\beta = (\beta_1, \beta_2, ..., \beta_k)$ bilinmeyen regresyon parametrelerini göstermektedir. x_{ij} ise k adet gözleme ait eşdeğişkeni göstermektedir. Eş. 5.8'de yer alan g fonksiyonu, monoton ve iki kez türevlenebilir bir bağ fonksiyonudur [52]. Bu fonksiyon $(0, 1)$ aralığında yer alan değerleri R kümesine taşımaktadır.

Bu modelde kullanılacak g, bağ fonksiyonu logit, probit, birikimli dağılım fonksiyonu, tamamlayıcı log-log bağ fonksiyonu ya da log-log bağ fonksiyonu olabilir. Bağ fonksiyon çeşitleri Çizelge 5.1'de yer almaktadır [53].

Çizelge 5.1: Bağ Fonksiyon Çeşitleri ve Fonksiyon İfadeleri

Bağ fonksiyon çeşitleri	Fonksiyon İfadeleri
Logit fonksiyonu	$log((\mu/(1-\mu))$
Probit fonksiyonu	$\Phi^{-1}(\mu)$
Tamamlayıcı log-log fonksiyonu	log(-log(1-μ)
Log-log fonksiyonu	-log(-log(1-μ))

Beta regresyon modelinde bağ fonksiyonu kullanılması iki farklı amaca hizmet etmektedir. İlk olarak, regresyon denkleminin sağ ve sol eşitlikler farklı tanım kümesine sahip olduğundan, bu denklemlerde doğrusallık olmayacaktır. Modelde, yanıt değişkeninin ortalama değerine uygulanan bu bağ fonksiyonu yoluyla regresyon denkleminin her iki tarafı da reel kümelerde tanımlı hale geleceği varsayılır. İkinci amaç ise, uygulayıcıya farklı bağ fonksiyonları kullanımı imkanı sağlayarak, herhangi bir regresyon modeli için, özellikle fazla yayılım gösteren veri setlerinde, veriye en iyi şekilde uyum sağlayacak bağ fonksiyonunun bulunması yoluyla kullanıcı açısından veri analizinde kolaylık sağlamaktadır [54].

Herhangi bir yanıt değişkeni için, bu değişkenin varyansı aynı zamanda ortalamasının bir fonksiyonu olduğundan, bu regresyon modeli değişken varyanslılık özelliğine

sahip olacaktır. Bu durumda, saçılım ortalamaya bağlı hale gelecek ve ortalama artığında varyans değerinde de artış beklenebilecektir.

$$V(z) = \frac{\mu(1-\mu)}{1+\phi} = \frac{g^{-1}({x_i}^T\beta)[1-g^{-1}({x_i}^T\beta)]}{1+\phi} \tag{5.9}$$

Beta dağılıma ait Log-olabilirlik fonksiyonu $l(\beta,\phi)$ olarak gösterilsin. Buna göre, $l(\beta,\phi) = \sum_{i=1}^{n} l_i(\mu_i,\phi)$ olacaktır. Herhangi n gözlem için log-olabilirlik fonksiyonu Eş. 5.10 şeklinde olacaktır.

$$l_i(\mu_i,\phi) = \log\Gamma(\phi) - \log\Gamma(\mu_i\phi) - \log\Gamma((1-\mu_i)\phi) + (\mu_i\phi - 1)\log z_i + \{(1-\mu_i)\phi - 1\}\log(1-z_i) \tag{5.10}$$

Beta regresyon modeli içerisinde yer alan herhangi i. özniteliğe ait ortalama hesabı için bağ fonksiyonunun tersi Eş. 5.11'deki gibi kullanılır.

$$\mu_i = g^{-1}(x_i^T\beta) \tag{5.11}$$

Bu regresyon modelinde parametre tahmini için en çok olabilirlik yöntemi kullanılarak tahmin yapılmaktadır. Bu modelin devamı niteliğinde olan, Smithson ve Verkuilen'in [55] 2006 yılında ortaya koyduğu ve Simas ve diğerleri [56] tarafından 2010 yılında geliştirilen bir başka beta regresyon modeli ise, değişken saçılımlı beta regresyon modelidir.

Değişken saçılımlı beta regresyon modelinde, kesinlik parametresi her bir gözlem için sabit kabul edilmeyip, modelde kesinlik parametresinin değişkenliği ele alınmıştır. Bu model için iki farklı bağ fonksiyonu tanımlanmıştır. Bu bağ fonksiyonları, Eş. 5.12 ve Eş. 5.13'te gösterildiği gibi, hem ortalama hem de kesinlik parametrelerini açıklamak için kullanılmaktadır. β ve γ sırasıyla ortalama ve kesinlik parametreleri için kullanılan regresyon katsayıları vektörünü göstermektedir.

$$g_1(\mu_i) = \eta_{1i} = x_i^T\beta \tag{5.12}$$

$$g_2(\phi_i) = \eta_{2i} = y_i^T\gamma \tag{5.13}$$

Bu bağ fonksiyonları, monotonluk özelliğine sahip olmak durumundadır. Buna ek olarak g_1 fonksiyonu, herhangi bir beta dağımlı bağımlı değişkenin ortalama değerini yani $(0,1)$ aralığına sahip herhangi bir değeri reel sayılar kümesine, g_2 fonksiyonu ise, $(0,\infty)$ aralığında değere sahip kesinlik parametresini reel sayılar kümesine tanımlayan bir fonksiyon olacaktır [57]. Eş. 5.12 ve Eş. 5.13'te yer alan η_{1i} ve η_{2i} parametreleri ise doğrusal tahmin edicileri göstermektedir.

sabit kesinlik parametreli beta regresyonuna benzer şekilde, değişken kesinlik parametreli beta regresyonu için de tahmin ediciler en çok olabilirlik yöntemi ile bulunmaktadır. Regresyon modelleri için çeşitli artık türlerinden bahsedilebilmektedir [12]. Bunlardan ilki, ham yanıt değişkeni artıklarıdır. Bu artıklar, $z_i - \hat{\mu}_i$ şekildedir. Ancak bu tip artıklar, ancak sabit varyanslılık özelliği gösteren modeller için kullanılabilmektedir. Beta regresyon modeli için kullanılabilecek artıklar, Pearson artıklarıdır. Bu artıklar,

$$r_{p,i} = \frac{z_i - \hat{\mu}_i}{V(z_i)} \tag{5.14}$$

şeklindedir. Ferrari ve Cribati-Neto [11], 2004 yılındaki çalışmasında bu artıkları, "standartlaştırılmış basit artıklar" şeklinde tanımlamıştır. $V(z_i) = \hat{\mu}_i(1-\hat{\mu}_i)/(1+\hat{\phi}_i)$ eşitliği düşünüldüğünde, Pearson artıkları denkleminin paydasında yer alan varyans ifadesinin bulunmasında değişken varyanslılık özelliği gösteren beta regresyon modelinde yer alan iki farklı bağ fonksiyonunun tersi de yer almaktadır.

5.3. Sonlu Karma Beta Dağılım Modeli

Sonlu karma modeller, herhangi bir veri seti için, veri seti içerisinde farklı özelliklere sahip olduğu varsayımı yapılabilecek veriler için kullanılabilir. Bu veriler için herhangi bir sınıflandırma bilgisi olmaması gerekmektedir.

Bir karma modelde, uygulayıcı sadece verisetinde aynı özelliklere sahip verileri ayırmakla kalmayıp, sınıflanan bu verilerin dağılımsal özelliklerini ve her bir sınıfın içerisinde yer alan veri sayısını da öğrenmek istemektedir. Bu sayede, ayrılan her bir sınıfa ait istatistiksel bilgiye sahip olacaktır.

Bağımlı değişkenin beta dağılımına uygunluk sağlaması koşulunda bu bağımlı değişken için sonlu karma beta dağılım modeli uygulanabilir. Sınıflandırma işlemi bağımlı değişkenin ortalama ve kesinlik parametrelerindeki farklılaşma göz önüne alınarak uygulanır. Karma dağılımdaki her bir kümenin eleman sayısı ise bu farklılığa bağlı olarak değişkenlik gösterecektir [58]

K bileşenli bir karma dağılım modeli Eş. 5.15 şeklinde gösterilmektedir.

$$h(z|x,\psi) = \sum_{k=1}^{K} \pi_k f(z|x,(g_1^{-1}(x^T\beta_k),g_2^{-1}(y^T\gamma_k)) \tag{5.15}$$

Eş. 5.15'te h fonksiyonu karma dağılımın olasılık yoğunluk fonksiyonunu, f fonksiyonu ise ortalama ve kesinlik parametreleri ile tanımlanmış beta dağılımına ait bir olasılık yoğunluk fonksiyonunu göstermektedir. π ifadesi ise her bir bileşenin ağırlığını göstermektedir. Her bir ağırlık değerlerinin sıfırdan büyük ve tüm ağırlık değerleri toplamının

1 olması gerekmektedir. $\theta_k = \left(g_1^{-1}(x^T\beta_k), g_2^{-1}(y^T\gamma_k)\right)$ ifadesi beta fonksiyonuna ait parametreleri göstersin. Bu durumda $\psi = (\pi_1, \pi_2, \pi_3, ..., \pi_K, \theta_1, \theta_2, \theta_3, ..., \theta_K)$ ifadesi karma dağılımda yer alan tüm dağılımların bilgisini içermektedir.

Herhangi bir (x, z) gözleminin i. kümeye ait olması sonsal olasılığı Eş. 5.16 ile verilir.

$$P(i|x, z, \psi) = \frac{\pi_i f(z|x, \theta_i)}{\sum_{k=1}^{K} \pi_l f(z|x, \theta_l)} \tag{5.16}$$

Verisetindeki her bir gözlem değerinin kümelere ayrılması aşamasında sonsal olasılık fonksiyonları kullanılır. Gözlem değerinin, sonsal olasılık yoğunluk fonksiyonu değeri hangi kümede en yüksek değeri gösterirse o kümeye ataması yapılır.

5.4. Sonlu Karma Beta Dağılım Modelinde Parametre Tahmini

N adet gözlem değerine sahip bir örneklemin log-olabilirlik fonksiyonu Eş. 5.17'de gösterildiği şekildedir.

$$logL = \sum_{n=1}^{N} \log h(z_n|x_n, \psi) = \sum_{n=1}^{N} \log \left(\sum_{i=1}^{K} \pi(k) f(z_n|x_n, \theta_k) \right) \tag{5.17}$$

Beta regresyon modelinde, log-olabilirlik fonksiyonunun doğrudan maksimizasyonu yapılamamaktadır; çünkü log-olabilirlik fonksiyonda gerek olasılık değerleri gerekse ağırlık ifadesi bilinememektedir. Bu sorunun çözümü için Dempter ve diğerleri [40] tarafından iterasyon bazlı EM algoritması kullanılmaktadır.

EM algoritması aşamaları iki bölümde incelenebilmektedir. Bu bölümler sırasıyla tahmin ve maksimizasyon aşamalarıdır [45].

Tahmin Aşaması:

Bu aşamada her bir gözlem için sonsal olasılık fonksiyonları kulllanılarak sonsal sınıf olasılıkları bulunur.

$$\hat{p}_{nk} = P(k|x_n, z_n, \hat{\psi}) \tag{5.18}$$

Eş. 5.18'de bulunan sonsal sınıf olasılıkları bulunduktan sonra her bir kümeye ait ağırlıklar Eş. 5.19 kullanılarak elde edilir.

$$\pi_k = \frac{\sum_{n=1}^{N} \hat{p}_{nk}}{N} \tag{5.19}$$

Maksimizasyon Aşaması:

Eş. 5.20'de gösterilen şekilde, sonsal olasılıklar kullanılarak her bir bileşen için log-olabilirlik fonksiyonu maksimize edilir.

$$\max_{\theta_k} \sum_{n=1}^{N} \hat{p}_{nk} \log f(z_n|x_n, \theta_k) \tag{5.20}$$

Tahmin ve maksimizasyon aşamaları, olabilirlik fonksiyonu maksimize edildiği sürece devam ettirilir. Ancak süreç maksimum iterasyon sayısına ya da herhangi bir eşik değerine ulaştığında durdurulur.

Küme sayısının seçimi için çeşitli bilgi kriterİ analizleri kullanılabilir. Bu bilgi kriteri analizleri sırasıyla Akaike Bilgi Kriteri(AIC), Bayezyen Bilgi Kriteri(BIC) analizleridir.

6. SILHOUTTE (GÖLGE) METODU

Gölge metodu, 1986 yılında Rousseuw [59] tarafından geliştirilen bir grafiksel gösterim yöntemidir. Bu yöntem ile kümelere ayrılan verilerin hem küme içindeki uzaklıklarına hem de diğer kümeler ile uzaklıklarına bakılarak ortalama gölge genişliği tanımı geliştirilmiştir. Ortalama gölge genişliği, yapılan kümelemenin geçerliliği ile birlikte kümeleme için kullanılabilecek uygun küme sayısına ait değerlendirme yapma imkanı sunmaktadır.

Literatürde verilerin özelliklerini ayırt edilebilmesi için bir çok kümeleme algoritması geliştirilmiştir. Bu algoritmalar başta uzaklık bazlı ve dağılım bazlı olmak üzere ikiye ayrılırken uzaklık bazlı kümeleme algoritmalarının bazıları k-ortalamalar, k-ortancalar ya da k-ağırlık merkezi algoritmaları olabilmektedir. Dağılım bazlı kümeleme algoritmalarına örnek olarak sonlu karma dağılımlar gösterilebilmektedir. Dağılım bazlı kümeleme algoritmalarında veri hangi dağılıma uygunluk sağlıyorsa o dağılıma ait karma dağılım modelleri kullanılır.

6.1. Gölge Metodu

Bu yöntem içerisinde iki tip yakınlık kavramından bahsedilmektedir. Bunlar benzemezlik(dissimilarity) ve benzerlik(similarity) kavramlarıdır. Benzemezlik kavramı, herhangi iki verinin birbirine ne kadar uzak olduğu ölçmeye yarmaktadır. İkinci kavram olan benzerlik(similarity) kavramı ise herhangi iki noktanın birbirine ne kadar benzediğini ölçmektedir.

Kümeleme algoritması olarak herhangi uzaklık bazlı algoritma kullanıldığını varsayılsın. Herhangi n adet verinin kümelendiği süşünülsün. Bu durumda algoritmanın amacı, verisetindeki n adet noktanın birbirine en yakın olacak şekilde k adet kümeye ayrılması olacaktır. Kullanılan uzaklık bazlı kümeleme algoritmasında, aynı kümeye ait noktaların birbirine yakın, farklı kümelerdeki noktaların ise birbirinden uzak olması amaçlanmaktadır. Bu durumun test edilebilmesi, 6.1 ifadesinin minimize edilmesi ile mümkün olacaktır.

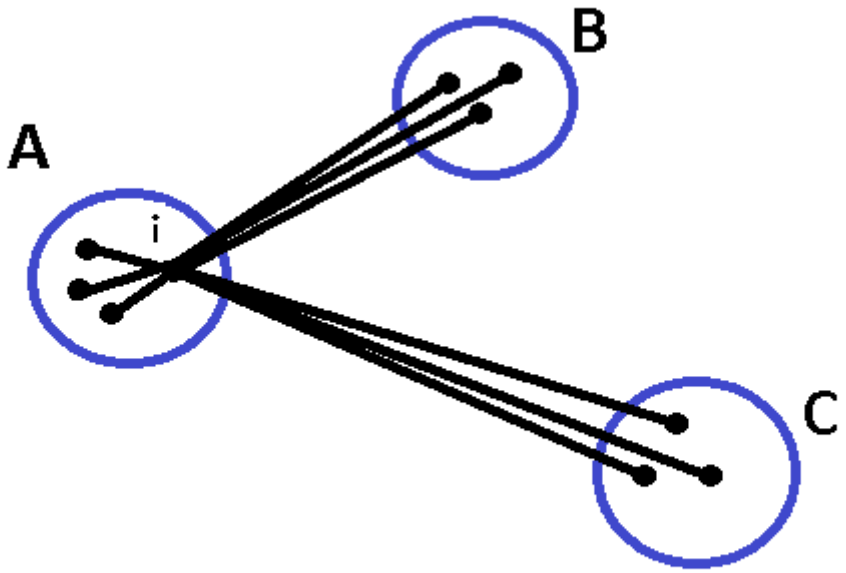
$$\frac{\sum_{i=1}^{n} d(i, m(i))}{n} \tag{6.1}$$

6.1 ifadesinin payında yer alan $d(i, (m(i))$ ifadesi herhangi bir i noktası ve bu i noktasına en yakın noktayı temsil eden $m(i)$ noktası arasındaki benzemezlik ölçüsünü göstermektedir.

Herhangi bir kümeleme algoritması kullanıldığında verinin k adet ayrımı yapılabilir ancak algoritma bu ayrımın kalitesi hakkında bilgi vermez. Kümeleme algoritmasının kalitesinin test edilmesi için öncelikle her bir nokta için gölgelerin belirlenmesi gerekir.

6.2. Gölgelerin Belirlenmesi

Gölgelerin belirlenmesi için iki duruma ihtiyaç vardır. İlk olarak kümeleme algoritması ile elde edilecek ayrışmaya, ikincisi olarak veriler arasında benzer olanların birleşimine ihtiyaç vardır.



Şekil 6.1: i noktası ve Kümelerin Gösterimi

Şekil 6.1'e göre, verisetinde A kümesine ataması gerçekleştirilmiş herhangi bir i noktası ele alınsın ve A kümesine ait tek nokta i noktası olmasın. Bu durumda $a(i)$, A kümesine ait tüm noktaların i'ye göre ortalama farklılığını göstersin. Ayrıca herhangi A kümesinden farklı bir B ve C kümeleri de mevcut olsun. Bu durumda $d(i, B)$ ve $d(i, C)$ sırasıyla i noktasının B ve C kümelere ait noktalara göre ortalama farklılığını göstermektedir. $b(i)$ değerinin $d(i, C)$ değerinin en düşük değeri olduğu düşünülsün. Bu

durumda B kümesi i noktası için komşu küme yani en iyi ikinci seçim olacaktır. Bu bakış açısı kullanılarak $s(i)$ indeksi Eş. 6.2'deki gibi elde edilmiştir.

$$s(i) = \left\{ \begin{array}{cc} 1 - a(i)/b(i) & \text{eğer } a(i) < b(i) \\ 0 & \text{eğer } a(i) = b(i) \\ b(i)/a(i) - 1 & \text{eğer } a(i) > b(i) \end{array} \right\} \tag{6.2}$$

Gölge metodu indeksi olan,$s(i)$ Eş. 6.2'de gösterilen şekilde parçalı fonksiyon gibi ele alınırken, aynı indeks Eş. 6.3'te verildiği gibi, tek bir eşitlik ile de gösterilebilmektedir. Eş. 6.3 sayesinde $s(i)$ indeksinin -1 ve 1 aralığında tanımlı olabileceği rahatlıkla görülebilmektedir.

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))} \tag{6.3}$$

Gölge metoduna ait $s(i)$ indeksini daha iyi anlamak için uç değerleri incelemek gerekmektedir. $s(i)$ değeri 1 değerine yakın olduğu durumda, $a(i)$ içerisindeki farklılığın, $b(i)$ farkının en küçüğünden de küçük olması beklenir. Bu durumda i noktasının iyi bir şekilde kümelendiği söylenebilir.

Diğer bir durum ele alınırsa, $s(i)$ değeri -1 değerine yakın olduğu durum içerisinde, $a(i)$ değeri $b(i)$ değerinden daha büyük olacaktır yani i noktası A kümesine değil B kümesine yakın olacaktır. Bu durumda i noktası, yanlış kümelenmiş olacaktır.

Son durum ise $s(i)$ değerinin 0 olması durumudur. Bu durumda i noktası $a(i)$ ve $b(i)$ noktası yaklaşık olarak eşit olacaktır. Bu durumda i noktasının kümelenmesinde kayıtsızlık mevcut olacaktır ve bu nokta A ya da B arasında rastgele kümelenecektir.

6.3. Gölge Metodunun Yorumlanması

Gölge değerleri hesaplandıktan sonra bu değerlerin grafik şekilde gösterimi yapılabilir. Gölge metodu, verisetindeki noktaların sahip oldukları kümeyi ne kadar iyi ya da kötü temsil ettiklerini açıklamaya yaramaktadır [60].

Gölge grafiğinin genişliği, $s(i)$ değerini göstermektedir. Gölge grafiğinin genişliğinin fazla olması, bu küme için büyük $s(i)$ değerlerinin varlığını simgelemektedir, bu durumda kullanılan kümeleme algoritasının etkin bir sonuç verdiği sonucuna varılabilir.

Bu sonuç, veriseti için kullanılan bu algoritmanın doğru sonuçlar vereceğine işaret etmektedir. Gölge grafiğinin diğer bir göstergesi de grafik boyudur. Bu yöntemde gölgenin boyu herhangi bir kümeye ait nokta sayısını göstermektedir. Eğer bir kümede nokta sayısı fazla ise, bu kümeye ait gölge grafiğinde de bu büyüklüğün görülmesi beklenir.

Genellikle gölge metodunun grafiksel gösteriminde kümeleme algoritmaları sonucunda elde edilen tüm kümeler için $s(i)$ değerleri elde edilir ve grafik her bir küme alt alta gösterilerek oluşturulur. Bu sayede hangi kümenin daha iyi hangi kümenin daha kötü ayrıştırıldığı kolayca görülebilir.

Gölge metodunun bir diğer avantajı da belirli bir kümeleme algoritması için en uygun küme sayısını belirleyerek modelin iyileştirilmesini sağlamaktadır. En uygun küme sayısı, verisetindeki her bir nokta için en uygun olan komşu küme ve sahip olunan küme arasındaki uzaklık baz alınarak hesaplanmaktadır. Elde edilebilecek en büyük ortalama gölge büyüklüğü değerine sahip küme sayısı, model için en uygun küme sayısı olacaktır [61].

7. UYGULAMA

7.1. Tarım Sigortaları Havuzu (TARSİM)

7.1.1 Kuruluş Amacı

Türkiye benzeri tarım bazlı ekonomilerde, tarım gerek yaşamsal faaliyetlerin devamında gerekse ekonomik ve sosyal sürekliliğin sağlanmasında ve gelişmesinde çok önemli bir yere sahiptir. Bu nedenle üretimin ekonomik olarak korunması ve çeşitli meteorolojik olaylardan etkilenmeden devam etmesi ya da mümkün olduğunca az etkilenmesi için tarımsal ürünlerin sigortalanması devlet tarafından desteklenmektedir. Bu amaçla üreticiler Çiftçi Kayıt Sistemi'ne kaydedilerek sisteme dahil edilir. Üreticinin kişisel bilgileri ve sahip olduğu tarımsal arazinin özellikleri, yetiştirilen ürünler hakkında bilgiler sürekli olarak toplanmaktadır. Böylece sistem kendi içerisinde güncelliğini korurken, yeni üretici kayıtları ile yeni üretim bölgeleri hakkında bilgi alabilmektedir.

Tarımsal üretimi tehdit eden meteorolojik risklerin teminat altına alınması ve tarımsal üretimin güvence altına alınması için 14 Haziran 2005 tarihinde 5363 sayılı "Tarım Sigortaları Kanunu" çıkartılmıştır. Kanun kapsamına alınan meteoreolojik riskler ile ilgili olarak yapılacak sigorta sözleşmelerinde standardın sağlanması, riskin en iyi koşullarda transferi için uygun ortam oluşturulması, oluşacak hasarlarda tazminatın tek merkezden ödenmesi ve tarım sigortalarının geliştirilmesi, yaygınlaştırılması hedeflerine yönelik Tarım Sigortaları Havuzu(TARSİM) kurulmuştur [62].

TARSİM'in kuruluş amaçları; sırasıyla, sel, kuraklık benzeri katastrofik riskler için sigorta teminatı sağlamak, reasürans şirketlerinin tarımsal ürünlerin sigortası kapsamında katılımını sağlayarak reasürans teminatını ve kapasitesini artırmak, sigorta primlerinin tek fiyattan belirlenmesini sağlayarak fiyat farklarından oluşacak adaletsizliği ortadan kaldırmak, üreticilere sigorta bilinci aşılamak, devlet destekli üretimin üreticiye daha anlamlı ulaşmasını sağlamak gösterilebilir.

7.1.2 Faaliyet Alanları

2005 yılından bugüne Tarım Sigortaları Havuzu (TARSİM) tarımsal ürünlerin sigortalanması, tarımsal üretimin dengeli bir şekilde devam etmesi amacıyla faaliyet vermektedir. Bu amaçla, TARSİM, bitkisel ürün, sera, ilçe bazlı kuraklık verim sigortası, büyükbaş ve küçükbaş hayvan hayat sigortası, kümes hayvanları, su ürünleri hayat sigortası ve son olarak arıcılık (arılı kovan) sigortası alanlarında çalışmalarını sürdürmektedir.

Devlet destekli bitkisel ürün sigortalarında temel olarak don, dolu, fırtına, hortum, yangın, deprem, heyelan, sel ve su baskını gibi doğal meteorolojik olayların bitkisel ürünlerde neden olduğu miktar kaybı, yine dolu olayı için yaş meyve, yaş sebze ve kesme çiçeklerde neden olduğu kalite kaybı teminat kapsamına alınmaktadır. [63].

7.1.3 Muafiyet ve Müşterek Sigorta Tanımları

Tarım Sigortaları Havuzu (TARSİM), tarife ve genel şartlarında prim hesaplanırken iki sigorta tanımından faydalanır. Bunlar sırasıyla muafiyet ve müşterek sigorta kavramlarıdır. Muafiyet kavramı poliçede yazılı önceden belirlenmiş bir miktara kadar olan hasarların sigortacı tarafından ödenmemesi ve/veya belirlenmiş o miktardan daha yüksek hasarların o miktar tenzil edildikten sonra ödenmesi olarak tanımlanabilmektedir [62]. Müşterek sigorta ise teminat kapsamındaki risklerin gerçekleşmesi sonucunda meydana gelen bir hasarın belli bir bölümünü veya belli bir yüzdesini sigortalının üzerinde tutması anlamına gelmektedir [62].

Muafiyet ve müşterek sigorta kavramları sayesinde üretici, sigortalılık süresince sigortacı gibi davranarak bilinçli bir şekilde üretim faaliyetinde bulunmaktadır. Herhangi bir hasarın gerçekleşmesi durumunda üreticinin de hasara katılımı beklenir. Örnek olarak, Çizelge 7.1'de yer alan dolu sigortasına ait oranlar kullanılarak ödenecek tazminat tutarı,gerçek hasar tutarının muafiyet ve müşterek sigorta tutarlarından çıkartılması ile bulunmaktadır. Sigortalıya ödenen tazminat muafiyet ve müşterek sigorta oranları kullanılarak manipüle edildiğinden, ödenen tazminat tutarı ile ortaya çıkan gerçek hasar arasında farklılık olmaktadır.

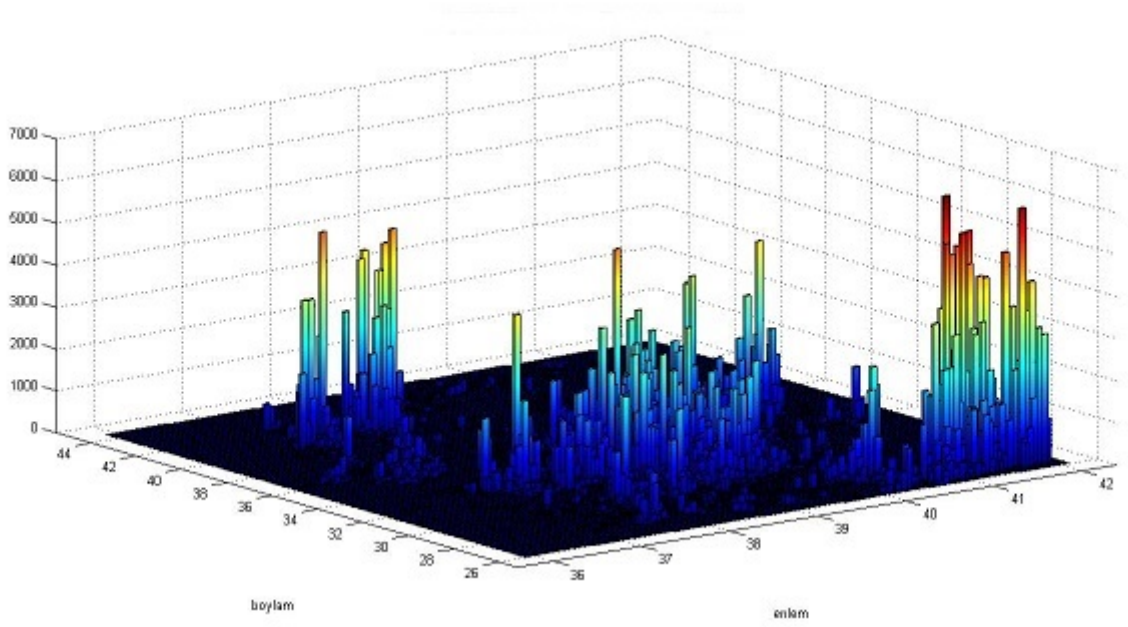
Çizelge 7.1: Dolu Sigortası için Muafiyet ve Müşterek Sigorta Oranı

Teminat	Ürünler	Toplam Sigorta Bedeli Üzerinden Muafiyet Oranı (%)	Sigortalının Üzerinde Kalan Müşterek Sigorta Oranı (%)
Dolu	Tüm Ürünler	10	0

TARSİM bünyesinde teminat altına alınmış hasar bölgelerinin risk sınıflandırması çalışmasında gerçekleşen hasar ile farklı değerler sunan ödenen tazminat değişkeni yerine herhangi bir şekilde değişikliğe uğramayan hasar gerçekleşme oranı ve hasar gerçekleşme olasılığı değişkenleri kullanılmıştır.

7.2. TARSİM Konumsal Kümeleme Uygulaması

Türkiye'de 2010-2014 yılları arasında buğday ürünü bitkisel ürün sigortası kapsamında dolu riskine bağlı olarak teminat verilmiş konumların dağılımı Şekil 7.1'te görülmektedir. Bu çalışmada, 5 yıl içerisinde en az 1 kez TARSİM sisteminde yer alan konumlar ele alınmıştır. Konumsal kümeleme çalışması kapsamında kullanılan toplam konum sayısı 1.314.114 adettir. Şekil 7.1'te görüleceği üzere, buğday ürününe ait bitkisel ürün sigortası kapsamında teminat verilen konumlar arasında bölgesel farklılıklar görülmektedir.



Şekil 7.1: Buğday Dolu Sigortası Poliçesine Sahip Konumlar

Buğday ürününe ait dolu riskine karşı sigortalanan konumlar incelendiğinde, Şekil 7.1'e göre konumların temelde üç ana parçaya ayrılabileceği görülmektedir. Bu üç bölge, sırasıyla Marmara Bölgesi ve çevresi, Orta Anadolu Bölgesi ve Güneydoğu, Doğu Anadolu bölgeleri şeklindedir. Ancak bu bölgelerin de kendi içerisinde teminat altına alınmış konum sayılarında farklılık gösterdiği görülmektedir. Örneğin, sadece Orta Anadolu Bölgesine bakıldığında, tek bir tepe değerinin olmadığı, konum sayısının birden fazla bölgede yoğunluk gösterdiği görülmektedir.

TARSİM kapsamında tüm Türkiye'de buğday ürünü baz alınarak dolu riskine karşı sigortalanmış poliçelere ait konumların yıllara göre dağılımı Çizelge 7.2'de verilmiştir.

Buğday ürünü için gerçekleştirilen hasar kümeleme çalışmasında 2010-2014 yılları arası, 5 yıllık veri göz önüne alındığında hasarlı konum sayılarının yıllara göre farklılık gösterdiği anlaşılmaktadır. Bu kümeleme çalışmasında 14.415 adet sigortalanmış konum

Çizelge 7.2: Yıllara Göre Buğday Dolu Hasarlı Konum Sayıları

Yıl	Konum Sayısı	Hasarlı Konum Sayısı
2010	5650	732
2011	6297	707
2012	9304	487
2013	10100	673
2014	10165	1030

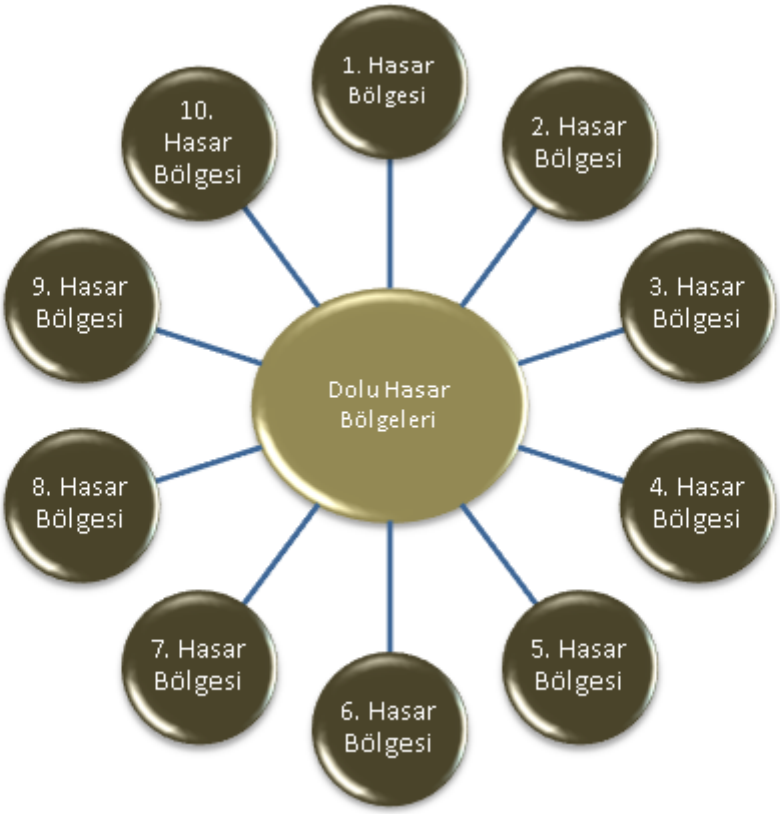
incelenmiştir.

Bu konumsal farklılıkların incelenmesinde, gerek uzaklık bazlı kümeleme algoritması olan küresel k-ortalamalar gerekse dağılım bazlı kümeleme algoritması olan sonlu karma von-Mises Fisher dağılımı modellleri gibi konumsal kümeleme algoritmaları kullanmak gerekmektedir.

Konumların incelenebilmesi için 5 yıllık süreçin herhangi bir diliminde bitkisel ürün sigortası kapsamında o konuma teminat verilmiş olması gerekmektedir. Belirlenen konumlar için hem uzaklık hem de dağılım bazlı kümeleme çalışması yapılmıştır. Bu konumlar için konumsal uzaklıklar ele alınarak küresel k-ortalamalar algoritması, konumların açısal değerlerinin dağılım özelliği göstermesi varsayımı kullanılarak dağılım bazlı konumsal kümeleme algoritması olan sonlu Karma von-Mises Fisher dağılımı algoritması çalıştırılmıştır. Birbirine yakın olan hasar bölgelerinin, coğrafik ve meteorolojik benzerlikleri göz önüne alınarak risk sınıflandırması yapılmıştır.

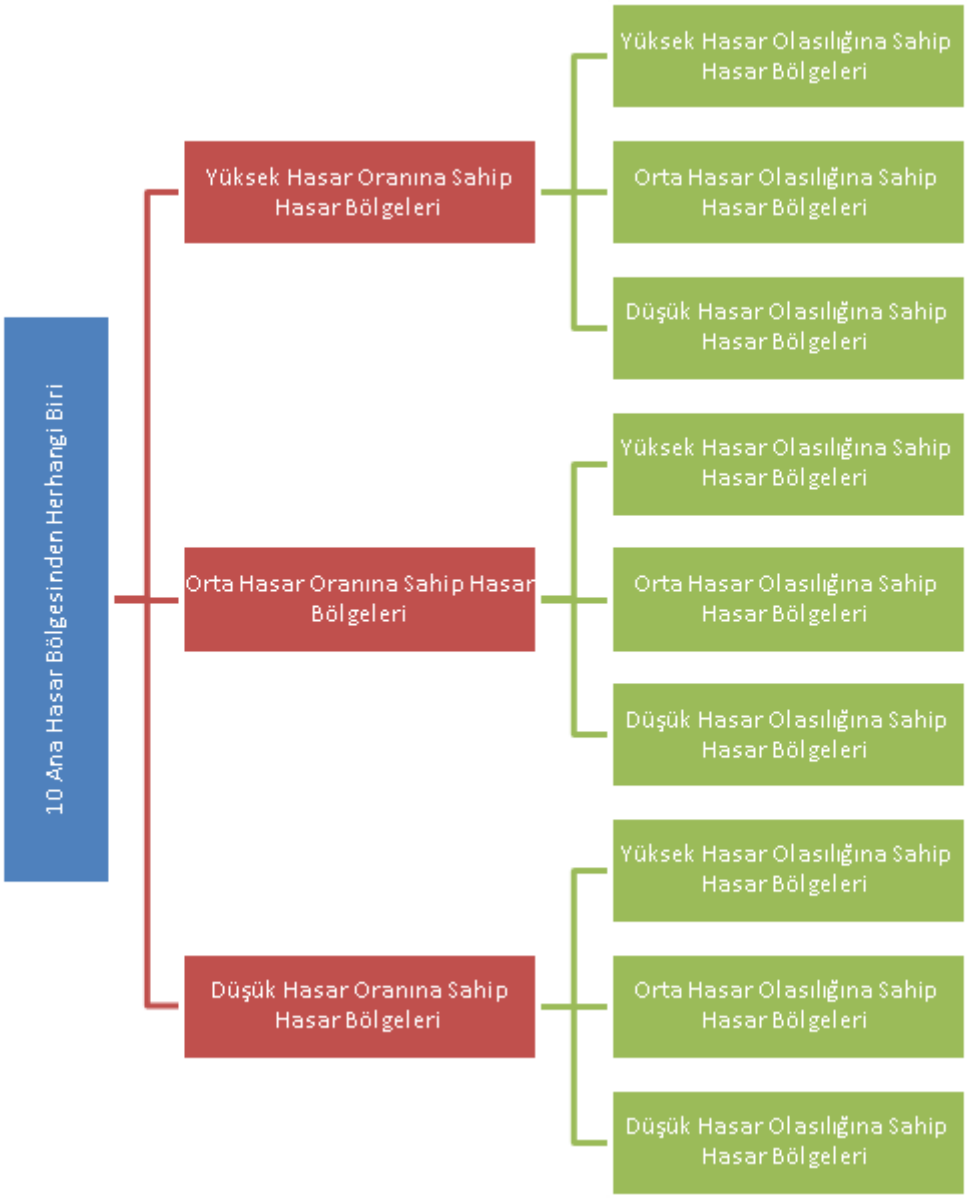
Hasar bölgeleri için öncelikle dağılım bazlı kümeleme algoritması olan sonlu karma von-Mises Fisher kümeleme algoritması kullanılarak Bayezyen bilgi kriteri(BIC) ve log-olabilirlik değerleri incelenmiştir. Ardından hasar bölgelerinin konumsal bilgileri kullanılarak bu kez küresel k-ortalamalar algoritması çalıştırılmıştır. Kümeleme algoritmasının uygunluğunun test edilebilmesi için gölge metodundan yararlanılmıştır. Bu sayede kümeleme algoritmasının kalitesi ölçülmüştür.

Şekil 7.2'de poliçelendirilmiş konumların enlem boylam bilgileri kullanılarak konumsal kümeleme sonucunda yapılan ayrım gösterilmiştir. Konumsal kümeleme çalışmasının ardından her bir küme için öncelikle sonlu Karma Beta dağılımı modeli içerisinde ortalama hasar gerçekleşme oranları bağımlı değişken kabul edilerek, konumların düşük, orta ve yüksek hasar gerçekleşme oranlarına sahip bölgeler olmak üzere kümeleme çalışması yapılmıştır.



Şekil 7.2: Dolu Hasar Bölgeleri Şeması

Ortalama hasar gerçekleşme oranları değişkenine göre yapılan kümeleme çalışmasının ardından ayrı ayrı elde edilen konumsal özelliklere ve hasar gerçekleşme oranları ortalamasına sahip konumlar için, Şekil 7.3'de görüleceği üzere bu kez ortalama hasar gerçekleşme olasılıkları değişkeni kullanılarak hasar gerçekleşme olasılıklarının dağılım özelliklerine göre de düşük, orta ve yüksek olasılığa sahip hasar bölgeleri olarak sınıflandırılması amaçlanmıştır.



Şekil 7.3: Düşük,Orta, Yüksek Hasar Ortalaması ve Olasılığı Ayrılış Şeması

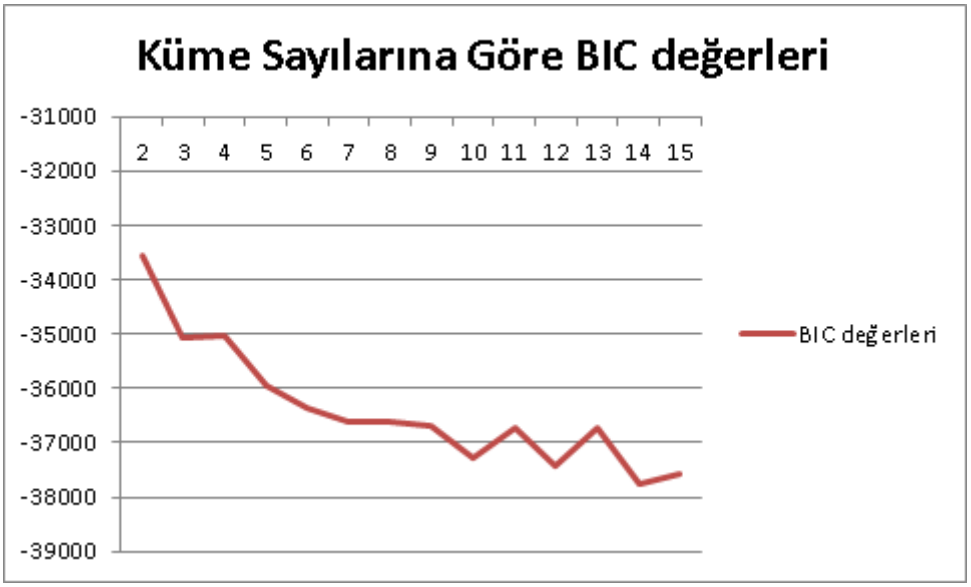
7.3. TARSİM Konumsal Kümeleme Uygulaması Sonuçları

Uygulamanın ilk aşamasında, dolu riskine karşı teminat vermiş buday üretimi yapılan konumlar için sonlu karma von-Mises Fisher dağılımı modeli kullanılmıştır. Bu model farklı küme sayıları için uygulanıp Log-olabilirlik ve BIC değerleri elde edilmiştir. Dağılım bazlı kümeleme çalışmalarında Log-olabilirlik ve BIC değerleri incelenerek optimal küme sayısı belirlenebilmektedir.

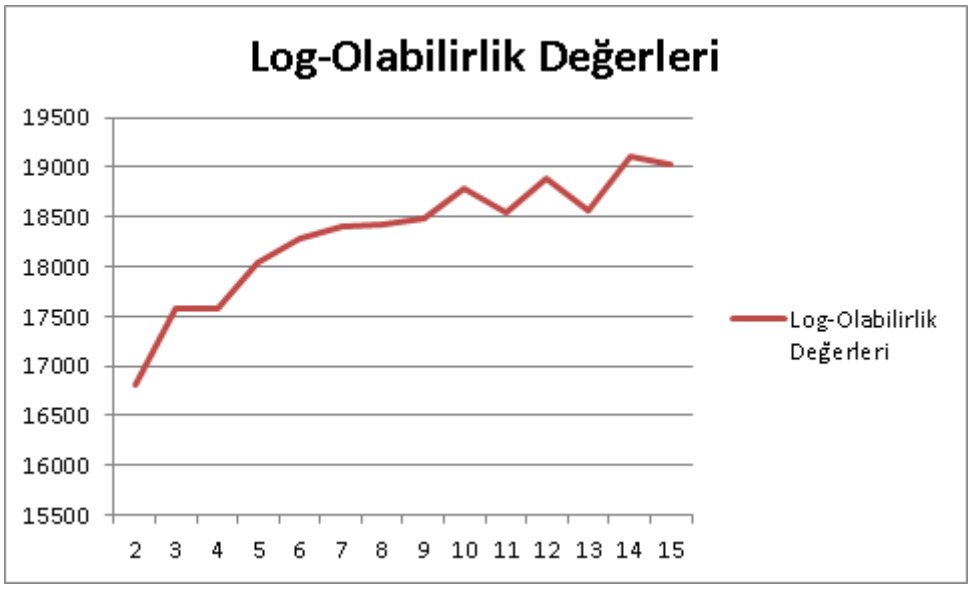
Şekil 7.4 ve Şekil 7.5'de görüleceği üzere, grafiklerde sırasıyla yer alan Bayezyen Bilgi Kriterei(BIC) ve log-olabilirlik değerleri incelendiğinde, bu konumsal kümeleme ça-

lışması için 10 ve 12 kümenin hem en düşük BIC hem de en yüksek Log-olabilirlik değerlerinden birini verdiği görülebilmektedir.

Şekil 7.5'e göre, Log-olabilirlik değerindeki ilk tepe değeri küme sayısı 10 olduğunda gerçekleşirken, ikinci tepe değeri ise küme sayısı 12 olduğunda elde edilmektedir. Bu durumda 10 ve 12 küme sayısının, diğer küme sayılarına göre modeli daha iyi açıklayabilme imkanı olabilecektir. 10 ve 12 küme sayılarına ait log-olabilirlik değerleri incelendiğinde iki değer arasındaki mutlak farkın az olduğu görülmektedir. Kümeleme çalışmalarında, mevcut farklılıkları mümkün olan en az küme sayısı ile açıklamak amaçlandığınından, kullanılan bu dağılım bazlı kümeleme çalışması için optimal küme sayısının 10 olabileceği düşünülmüştür.

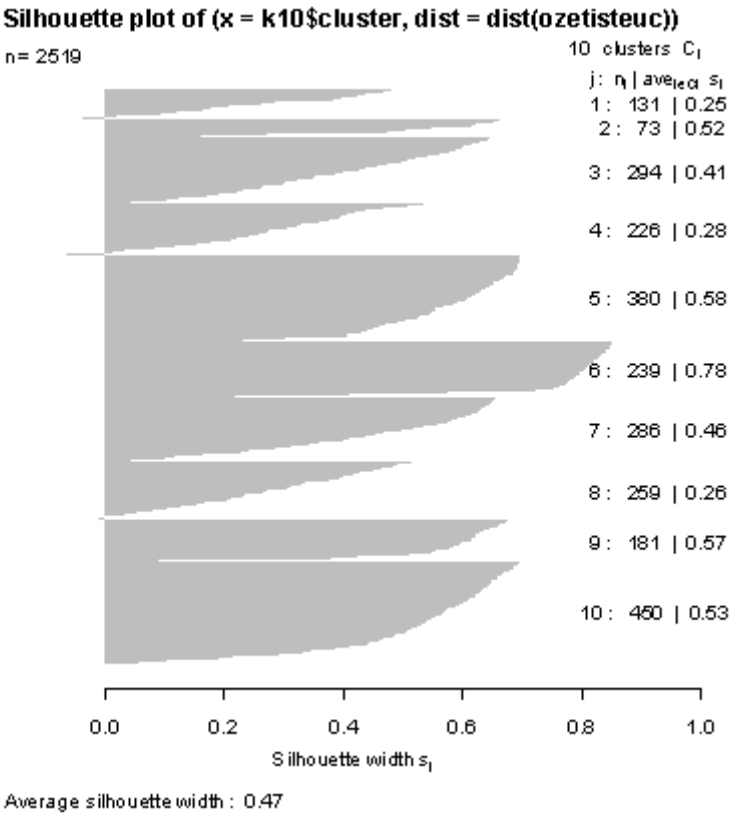


Şekil 7.4: Küme Sayısına Göre BIC değerleri



Şekil 7.5: Küme Sayısına Göre Log-olabilirlik Değerleri

Dağılım bazlı kümeleme çalışması için optimal küme sayısı elde edildikten sonra aynı küme sayısı küresel k-ortalamalar algoritması için veri kabul edilerek çalıştırılmıştır. Şekil 7.6'ya göre, hasar bölgeleri, uzaklık bazlı kümeleme algoriması olan küresel k-ortalamalar yöntemi kullanılarak 10 adet kümeye ayrıldığında ortalama gölge genişliği 0.47 değerini vermektedir. Bu sonuç, buğday ürünü için teminat altına alınan konumlar için küresel k-ortalamalar algoritmasının anlamlı bir kümeleme sonucu ortaya koyduğunu göstermektedir.



Şekil 7.6: Gölge Metodu Grafiksel Gösterimi

Hem dağılım bazlı hem de uzaklık bazlı kümeleme yöntemleri için uygun küme sayıları belirlenirken yöntemlerin kümeleme kaliteleri test edilerek, dolu riskine karşı teminat verilen konumlar için iki farklı kümeleme sonucu elde edilmiştir.

7.3.1 Sonlu Karma von-Mises Fisher Dağılımı Yöntemi ile Kümeleme

Sonlu karma von-Mises Fisher dağılımı yönteminde, konumların hasar geçmişine bakılmaksızın dağılım bazlı konumsal kümeleme çalışması yapılmıştır. Hasar geçmişi olan bölgeler için sonlu karma von-Mises Fisher dağılımı kümeleme sonucu Şekil 7.7'de, 5 yıl içerisinde herhangi bir şekilde riske maruz kalmamış yani hasarsız bölgeler ise Şekil 7.8'de gösterilmiştir.

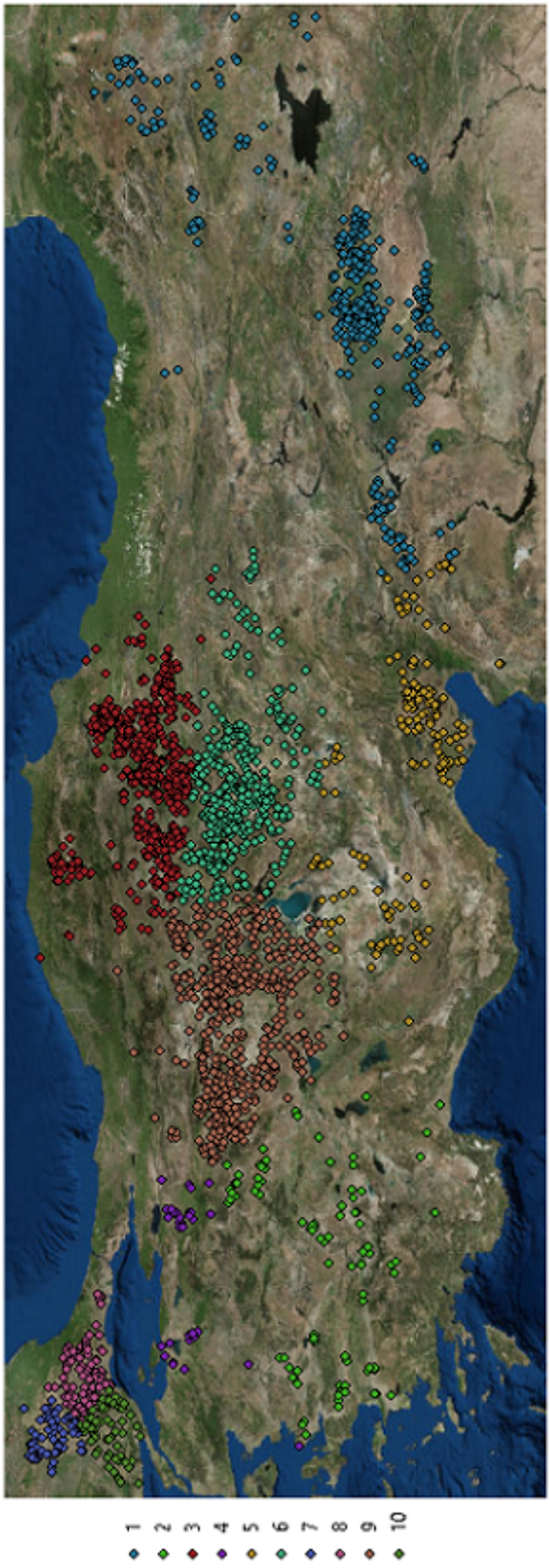
Kümeleme sonucu elde edilen kümeler için, Rivest'in [64] 1986 yılında ortaya koyduğu von-Mises Fisher uyum iyiliği testi yapıldığında, bu kümelerin von-Mises Fisher dağılımına uygunluk gösterdiği görülmüştür. Ek 1'de sınıflandırılan konumlar için elde edilen parametreler ve uyum iyiliği test sonuçları (p değerleri) verilmiştir.

Bu kümeleme çalışmasının sonucunda elde edilen kümeleme bilgisi, hasar geçmişi olan

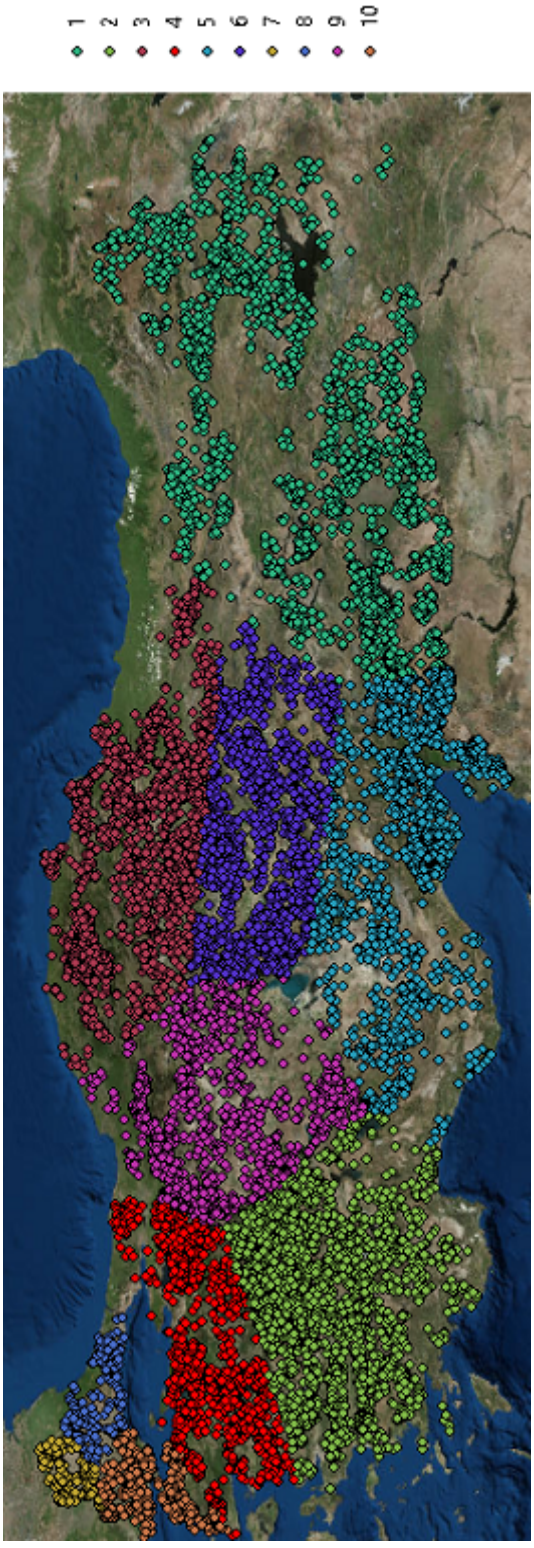
bölgeler için, bir sonraki aşama olan hasar gerçekleşme oranları kümeleme çalışmasında kullanılmaktadır. Şekil 7.9'a göre, Hasar gerçekleşme oranları kümeleme çalışması sonucunda 29 adet farklı risk grupları elde edilir. Bu çalışmanın ardından her bir hasar bölgesi, o risk grubuna ait konumların hasar gerçekleşme olasılıklarına göre tekrar ayrıştırılarak 54 farklı hasar bölgesi elde edilir.

Kümeleme çalışmasının ikinci ve üçüncü aşamalarında kullanılan sonlu karma Beta dağılımı modeli için her bir sınıflama aşamasında en küçük Akaike Bilgi Kriteri değeri elde edilecek şekilde kümeleme algoritması çalıştırılmıştır. Ek 2'de sonlu karma Beta dağılımı ve sonlu karma Beta dağılım modellerinde optimal küme sayısı kullanılarak yapılan her bir dallanma sonucunda elde edilen Akaike Bilgi Kriteri değerleri yer almaktadır. Ek 4'te ise uyum iyiliği testleri sonucu, her bir risk sınıfı için hasar gerçekleşme oranı ve olasılığı değişkenlerinin Beta dağılımına uyumlu olduğu gösterilmiştir.

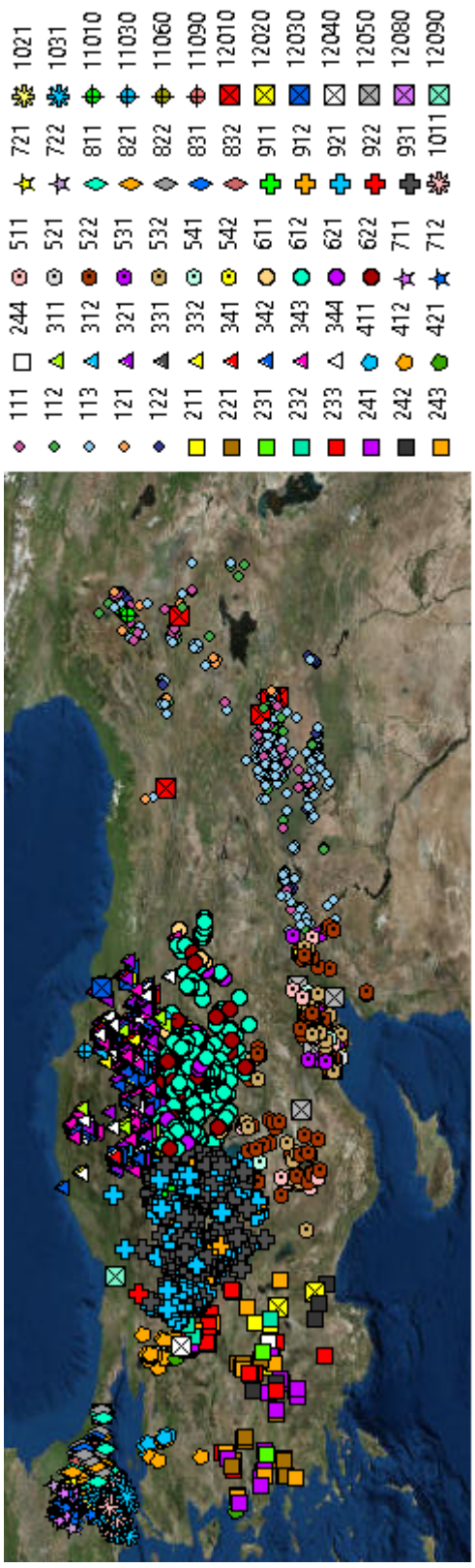
Konumsal özellikleri sonlu karma von-Mises Fisher dağılımı, sigorta ve hasar geçmişi sonlu karma Beta dağılımı modeli kullanılarak ayrıştırılan hasar bölgeleri için elde edilen hasar gerçekleşme oranı ve hasar gerçekleşme olasılığı tanımlayıcı istatistikleri Çizelge 7.3'te verilmiştir.



Şekil 7.7: von-Mises Fisher Dağılımı Kümeleme Algoritması-Hasarlı Bölgeler Haritası



Şekil 7.8: von-Mises Fisher Dağılımı Kümeleme Algoritması Sonucu-Hasarsız Bölgeler Haritası



Şekil 7.9: von-Mises Fisher Dağılımı ve Sonlu Karma Beta Dağılımı Modeli ile Hasarlı Bölgeler Haritası

Çizelge 7.3: Sonlu Karma von-Mises Fisher Algoritası ve Sonlu Karma Beta Dağılımı Modeli Sonucu Kümelerin Hasar Gerçekleşme Oranı ve Olasılığı Tanımlayıcı İstatistikleri

Sonlu Karma von-Mises Fisher Algoritması				
Kume	Hasar Oranı Ortalaması (Yüzde)	Hasar Ortalaması Standard Sapması (Yüzde)	Hasar Olasılığı Ortalaması	Hasar Olasılığı Standart Sapması
111	32.9440	12.2325	0.2659	0.0491
112	34.3891	12.9643	0.5401	0.1534
113	30.4919	12.7239	0.0737	0.0531
121	77.1523	12.0784	0.3355	0.1988
122	74.1666	10.0118	0.0366	0.0250
211	77.6667	12.0554	0.0199	0.0188
221	8.3769	3.3795	0.0419	0.0404
231	34.5416	3.1230	0.1569	0.0043
232	53.8373	5.3260	0.2637	0.0435
233	39.4327	9.8881	0.0315	0.0305
241	20.6440	4.0471	0.0893	0.0225
242	19.9928	2.9630	0.2929	0.1303
243	18.8871	4.0530	0.0208	0.0065
244	20.0000	7.0710	0.0008	0.0004
311	79.9151	8.1424	0.0453	0.0215
312	72.9915	5.3973	0.1917	0.1096
321	49.4277	8.4214	0.1522	0.1358
331	5.7833	3.7604	0.0150	0.0054
332	7.9159	3.2958	0.0867	0.0605
341	25.9432	4.6994	0.5328	0.1236
342	25.6293	6.3823	0.0883	0.0220
343	22.8140	5.6978	0.0282	0.0125
344	24.9633	5.6133	0.2175	0.0660
411	24.5317	4.4760	0.0093	0.0026
412	25.8419	12.4598	0.1222	0.1056
421	65.5833	5.5132	0.1435	0.1175
511	62.7857	13.2566	0.1035	0.1119
521	19.5047	9.3239	0.6477	0.1615
522	17.8917	4.8916	0.0815	0.0825
531	34.7922	5.5646	0.4637	0.1540
532	34.4218	5.9461	0.0717	0.0583
541	5.0000	2.5298	0.0232	0.0134
542	6.6333	1.4651	0.1568	0.0602
611	31.4375	16.8191	0.4682	0.1822
612	29.0927	13.0364	0.0600	0.0577
621	73.0222	13.7419	0.2729	0.1397
622	71.418	12.1031	0.0518	0.0332
711	19.0367	6.6729	0.0391	0.0274
712	15.3391	8.2664	0.0055	0.0030
721	51.6775	8.2054	0.1249	0.0297
722	40.3549	9.5391	0.0327	0.0244
811	6.2311	2.5759	0.0114	0.0120
821	16.4985	3.9247	0.0086	0.0039
822	18.5483	4.3564	0.0591	0.0550
831	40.3344	14.5076	0.1245	0.0439
832	46.6182	6.9411	0.0118	0.0078
911	70.5000	0.7071	0.1882	0.0166
912	83.1488	8.3835	0.0260	0.0198
921	42.4030	8.6914	0.0685	0.0571
922	45.1926	8.0880	0.4177	0.1434
931	20.2630	6.6880	0.0578	0.0818
1011	5.2646	1.2036	0.0198	0.0239
1021	44.9032	7.8525	0.0543	0.0363
1031	18.0364	7.1790	0.0394	0.0532
11010	100.0000	0.0000	0.1469	0.1082
11030	100.0000	0.0000	0.1153	0.1449
11060	100.0000	0.0000	0.0484	0.0631
11090	100.0000	0.0000	0.0394	0.0299
12010	22.2733	13.7476	1.0000	0.0000
12020	15.9000	4.2225	1.0000	0.0000
12030	37.4833	21.8923	1.0000	0.0000
12040	26.8333	16.7348	1.0000	0.0000
12050	34.8000	24.7022	1.0000	0.0000
12080	6.0000	0.0000	1.0000	0.0000
12090	41.0000	0.0000	1.0000	0.0000

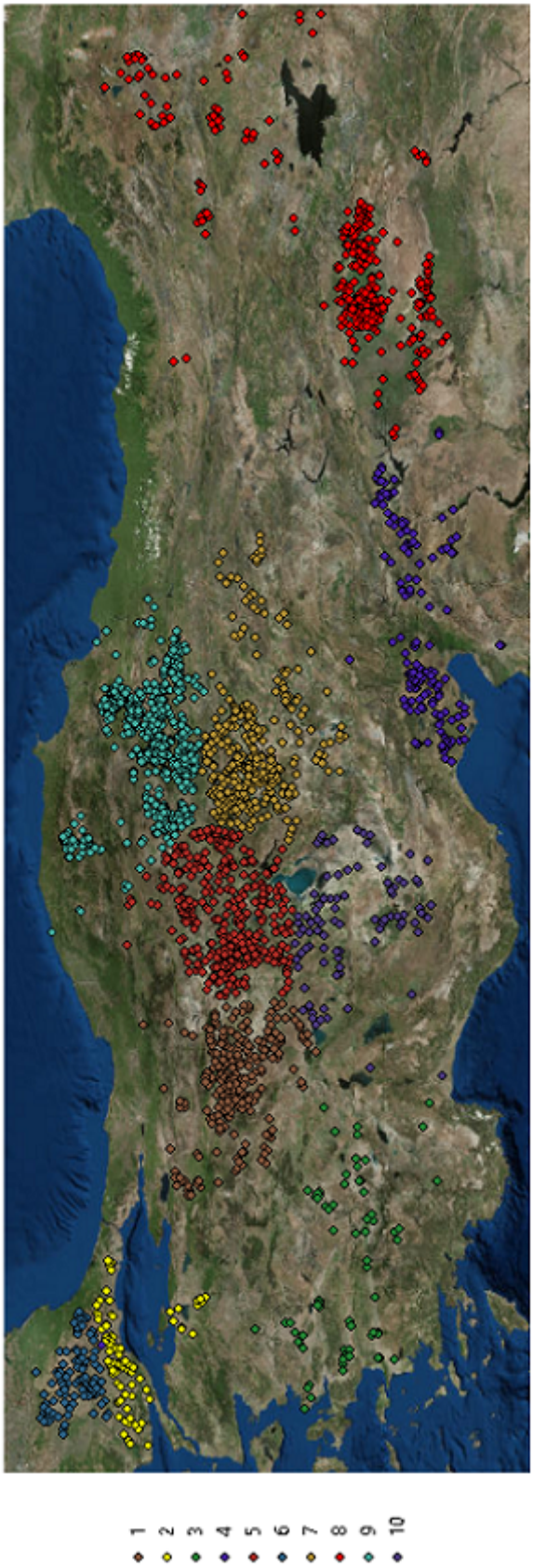
7.3.2 Küresel K-ortalamalar Kümeleme Yöntemi ile Kümeleme

Bu kümeleme çalışmasında ise, dağılım bazlı kümeleme çalışmasında ele alınan konumsal veri kullanılarak, buğday üretilen konumların 10 adet kümeye ayrılması başarılmıştır. Ardından, küresel k-ortalamalar kümeleme çalışmasının sonucunda elde edilen bilgi, aynı sonlu Karma von-Mises Fisher yönteminde olduğu gibi bir sonraki aşama olan hasar gerçekleşme oranları kümeleme çalışmasında kullanılmaktadır.

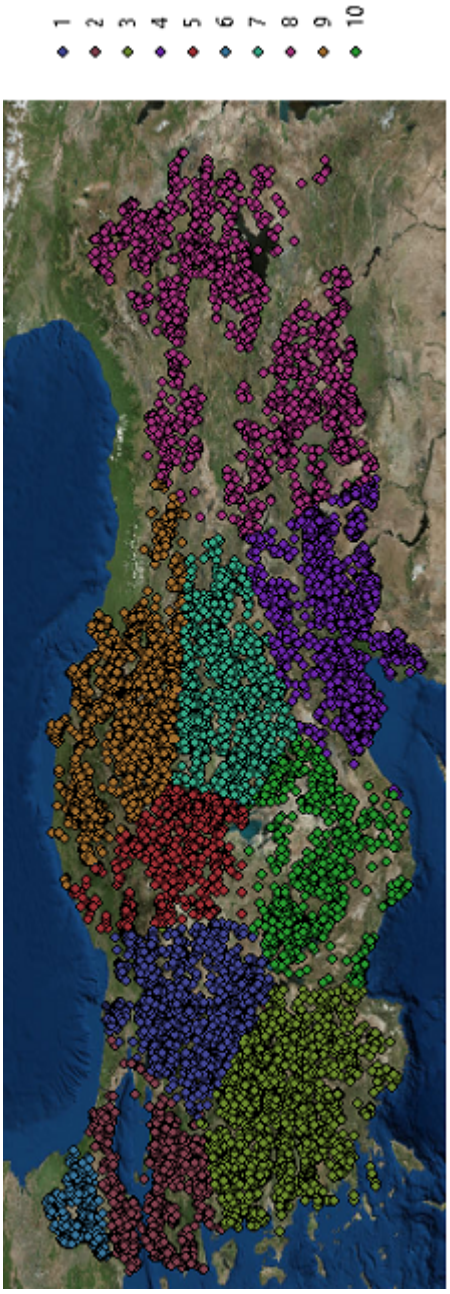
Hasar geçmişi olan bölgeler için küresel k-ortalamalar kümeleme çalışmasının sonucu Şekil 7.10'da, 5 yıl içerisinde herhangi bir şekilde riske maruz kalmamış yani hasarsız bölgeler ise Şekil 7.11'de gösterilmiştir. Bu kümeleme çalışması sonucunda hasar gerçekleşme oranları açıklayıcı değişken olarak ele alındığında, Şekil 7.11'de 24 adet farklı hasar bölgesi elde edilir. Bu çalışmanın ardından her bir hasar bölgesi, o bölgeye ait konumların hasar gerçekleşme olasılıklarına göre tekrar ayrıştırılarak 48 farklı hasar bölgesi elde edilir.

Kümeleme çalışmasının ikinci ve üçüncü aşamalarında, hasar gerçekleşme oranı ve olasılığı değişkenlerinin incelemesinde kullanılan sonlu karma Beta dağılımı modeli için her bir sınıflama aşamasında en küçük Akaike Bilgi Kriteri değeri elde edilecek şekilde kümeleme algoritması çalıştırılmıştır. Ek 2'de küresel k-ortalamalar algoritması ve sonlu karma Beta dağılım modellerinde optimal küme sayısı kullanılarak yapılan her bir dallanma sonucunda elde edilen Akaike Bilgi Kriteri değerleri yer almaktadır. Ek 3'te ise uyum iyiliği testleri sonucu, her bir risk sınıfı için hasar gerçekleşme oranı ve olasılığı değişkenlerinin Beta dağılımına uyumlu olduğu gösterilmiştir.

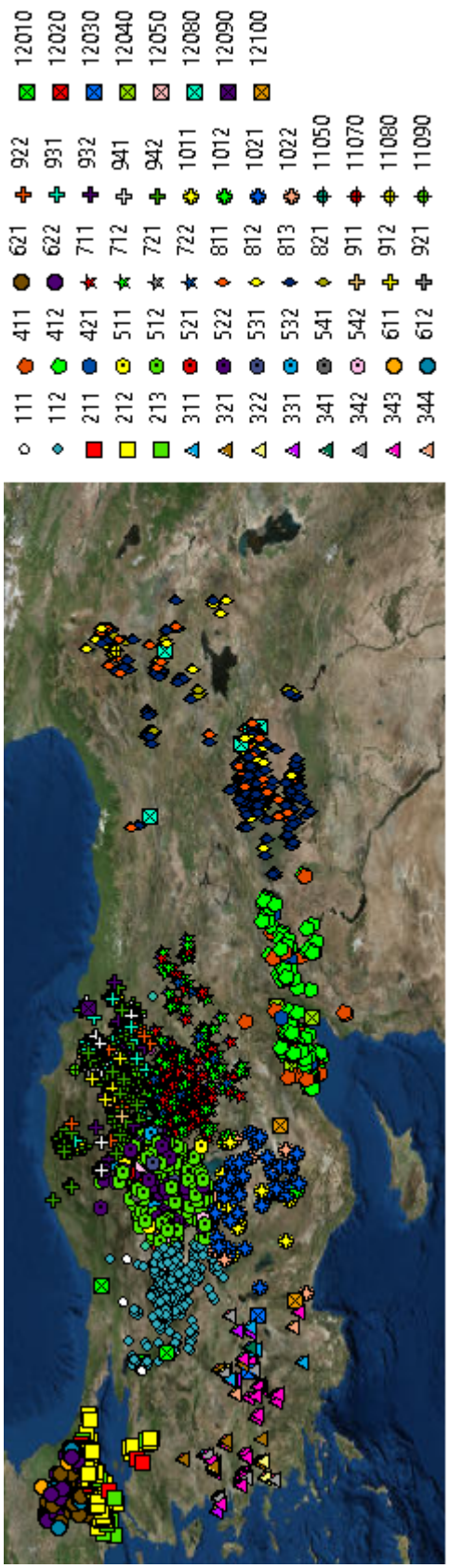
Küresel k-ortalamalar algoritması ve sonlu karma Beta dağılımı modelleri kullanılarak ayrıştırılan hasar bölgeleri için elde edilen hasar gerçekleşme oranı ve olasılığı tanımlayıcı istatistikleri Çizelge 7.4'te verilmiştir.



Şekil 7.10: Konumsal K-ortalamalar Kümeleme Algoritması Sonucu-Hasarlı Bölgeler Haritası



Şekil 7.11: Konumsal K-ortalamalar Kümeleme Algoritması Sonucu-Hasarsız Bölgeler Haritası



Şekil 7.12: Konumsal K-ortalamalar Kümeleme Algoritması ve Sonlu Karma Beta Dağılımı Modeli Sonucu-Hasarlı Bölgeler Haritası

Çizelge 7.4: Konumsal K-ortalamalar Algoritası ve Sonlu Karma Beta Dağılımı Modeli Sonucu Kümelerin Hasar Gerçekleşme Oranı ve Olasılığı Tanımlayıcı İstatistikleri

Konumsal K-ortalamalar Algoritması				
Kume	Hasar Oranı Ortalaması (Yüzde)	Hasar Oran Standart Sapması(Yüzde)	Hasar Olasılığı Ortalaması	Hasar Olasılığı Standart Sapması
111	35.9882	16.6115	0.4992	0.1896
112	32.0040	17.3985	0.0644	0.0607
211	24.4648	5.0751	0.2103	0.0527
212	14.6964	9.3648	0.0124	0.0095
213	20.3620	9.6674	0.0715	0.0216
311	49.5523	3.8532	0.0562	0.0978
321	7.2578	3.2563	0.0816	0.0706
322	6.9333	3.0858	0.0054	0.0028
331	74.2500	12.1484	0.0194	0.0156
341	20.0000	7.0711	0.0008	0.0004
342	18.0250	3.0504	0.0209	0.0055
343	22.8681	6.6086	0.1009	0.0509
344	18.1666	2.7537	0.4407	0.0577
411	26.7592	11.9080	0.4228	0.1657
412	22.4080	11.1609	0.0832	0.0598
421	75.7547	10.4581	0.0924	0.1131
511	25.3957	5.9363	0.3821	0.1467
512	24.8263	6.43044	0.0577	0.0491
521	49.0145	13.7974	0.4188	0.1713
522	37.3361	16.1315	0.0616	0.0526
531	80.4629	9.9234	0.0235	0.0179
532	76.3246	8.1062	0.1575	0.0356
541	4.5000	0.7071	0.0889	0.0400
542	4.2857	1.2999	0.0201	0.0149
611	41.0005	6.6183	0.0168	0.0109
612	49.6999	10.9442	0.1063	0.0345
621	19.7877	5.3718	0.0272	0.0351
622	19.7877	5.3718	0.0272	0.0351
711	29.5768	14.7139	0.0152	0.0079
712	34.6075	15.7182	0.1253	0.1063
721	83.6731	12.4206	0.4036	0.1157
722	78.1832	9.3609	0.0871	0.0562
811	35.3308	14.2820	0.2817	0.0426
812	37.9023	14.8170	0.5434	0.1540
813	32.6302	14.2084	0.0797	0.0595
821	86.1153	7.4979	0.2319	0.2445
911	5.7833	3.7604	0.0151	0.0054
912	8.2479	3.3975	0.0864	0.0582
921	68.8128	6.0538	0.3223	0.1974
922	77.4502	9.0562	0.0579	0.0312
931	48.2420	7.5951	0.0777	0.0539
932	48.8932	8.2029	0.2871	0.0568
941	25.9579	5.0728	0.3445	0.1343
942	24.1868	6.0588	0.0809	0.0556
1011	53.4873	14.7162	0.0174	0.0135
1012	49.0353	13.8344	0.1314	0.0352
1021	20.9573	6.3684	0.0157	0.0113
1022	23.1030	7.5579	0.0833	0.0488
11050	100.0000	0.0000	0.1425	0.1400
11070	100.0000	0.0000	0.0328	0.0485
11080	100.0000	0.0000	0.1469	0.1082
11090	100.0000	0.0000	0.0608	0.0907
12010	31.5555	14.3849	1.0000	0.0000
12020	6.0000	0.0000	1.0000	0.0000
12030	17.7500	3.8891	1.0000	0.0000
12040	27.2500	20.8226	1.0000	0.0000
12050	54.9167	0.0000	1.0000	0.0000
12080	22.2733	13.7476	1.0000	0.0000
12090	33.1250	22.6362	1.0000	0.0000
12100	38.6000	37.3352	1.0000	0.0000

7.4. Kümeleme Çalışması Sonucunda Buğday Ürünü İçin Prim Hesabı

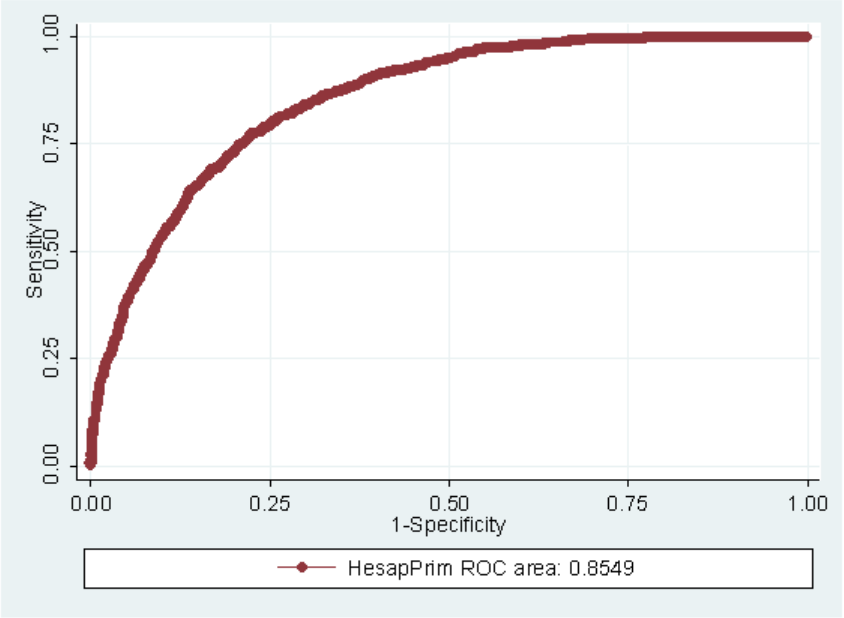
Buğday ürünü için dolu riskine bağlı net prim hesabı için, öncelikle, TARSİM bünyesinde yer alan 1.314.114 adet poliçe için sahip oldukları köy kodlarına göre filtreleme işlemi yapılmıştır. Filtreleme işleminin ardından elde edilen 14.415 farklı konum için ortalama hasar gerçekleşme oranı, ortalama hasar gerçekleşme olasılığı, ortalama sigorta bedeli değişkenleri elde edilmiştir.

Konumlara ait enlem boylam verileri kullanılarak, hem sonlu karma von-Mises Fisher dağılımı hem de küresel k-ortalamalar algoritması ile konumsal kümeleme çalışmaları yapılmıştır. Belirlenen kümeler için sonlu karma beta dağılımı kullanılarak hem hasar gerçekleşme oranı hem de hasar gerçekleşme olasılığı değişkenlerine göre derecelendirme yapılmış ve her bir küme için ortalama hasar gerçekleşme oranı ve ortalama hasar gerçekleşme olasılığı değerleri tahmin edilmiştir.

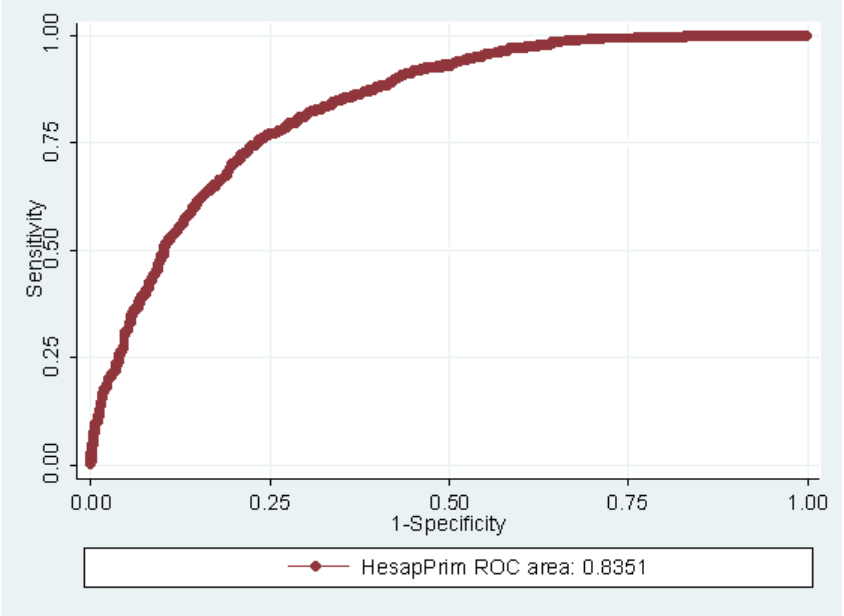
Hasar geçmişi bulunan konumlar için, buğday ürünü için risk primi hesaplamasında, o konuma ait ortalama sigorta bedeli, hasar gerçekleşme oranı ve hasar gerçekleşme olasılığı çarpılarak net aktüeryal risk primi hesaplanmıştır [30].

Her iki kümeleme yöntemi için ortak kabul edilen küme sayıları kullanılarak, bu konumlar için sonlu karma von-Mises Fisher dağılımı ile hesaplanan primler, küresel k-ortalamalar kümeleme algoritması sonucu hesaplanan primler ile karşılaştırılmıştır. Düşük riskli bölgeler için en düşük prim seviyeleri elde edilirken, yüksek risli bölgeler için en yüksek prim seviyeleri tahmini elde edilmiştir.

Çalışmanın sonucunda; farklı konumsal özelliklere ve hasar geçmişine sahip konumlar için primlerin farklı olması gereği açıklanmıştır.



Şekil 7.13: Sonlu Karma von-Mises Fisher Dağılımı Kümeleme Algoritması Sonucu-Hasar Olasılığı ROC Analizi

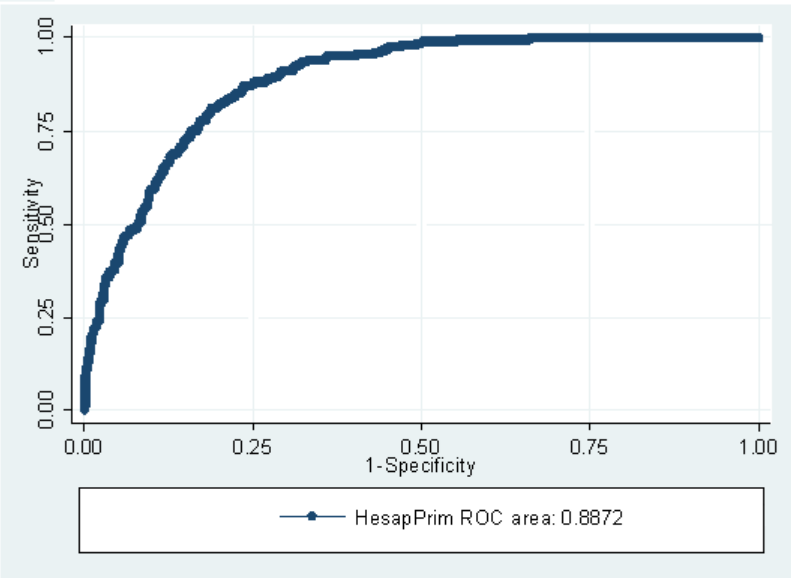


Şekil 7.14: Küresel K-ortalamalar Kümeleme Algoritması Sonucu-Hasar Olasılığı ROC Analizi

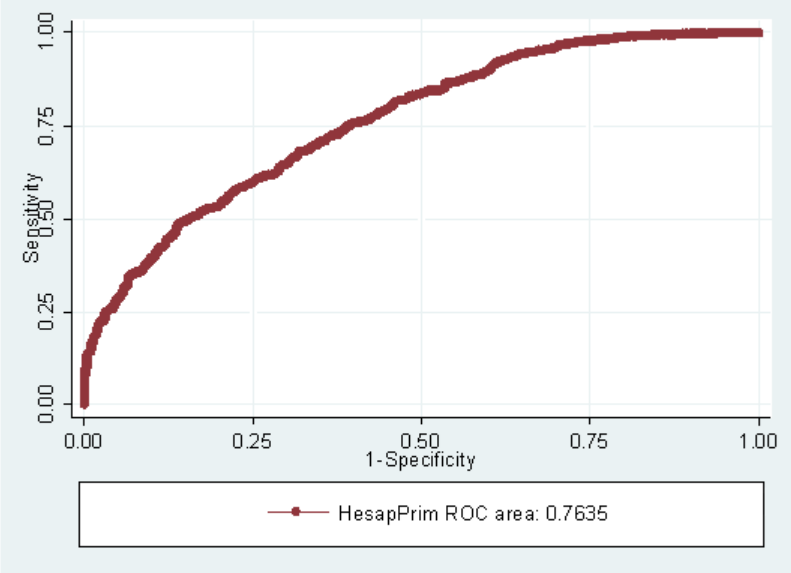
Uygun kümeleme algoritmasının seçilmesi için, her iki algoritma ile hesaplanan primlerin, hasar gerçekleşme oranı ve olasılığı değerlerinde gerçekleşen değişimi açıklanabilirliği karşılaştırılmalıdır. Şekil 7.13'e göre, sonlu karma von-Mises Fisher dağılımı ile hesaplanan primlerin, hasar gerçekleşme olasılığındaki değişimin yüzde 85'inin açıkla-

nabildiği görülmektedir. Uzaklık bazlı kümeleme algoritması ile hesaplanan primler ele alındığında ise, Şekil 7.14'e göre, primlerin hasar gerçekleşme olasılığındaki değişimin yüzde 83'lük kısmını açıklayabildiği görülmektedir.

Sonlu Karma von-Mises Fisher dağılımına ait kümeleme çalışması ve konumsal k-ortalamalar yöntemi ile elde edilen primlerin, ortalama hasar gerçekleşme oranındaki değişimi ne kadar açıklayabileceğine bakıldığında ise, dağılım bazlı kümeleme sonucunda elde edilen primlerin uzaklık bazlı kümeleme algoritması ile hesaplanan prime göre değişimi daha iyi açıkladığı sonucuna varılmıştır. Şekil 7.15'e göre dağılım bazlı kümeleme algoritması ile hesaplanan primler yüzde 88.7 ve Şekil 7.16'ya göre k-ortalamalar algoritması ile hesaplanan primler ise yüzde 76.35 oranında hasar oranındaki değişkenliği açıklayabildiği görülmektedir.

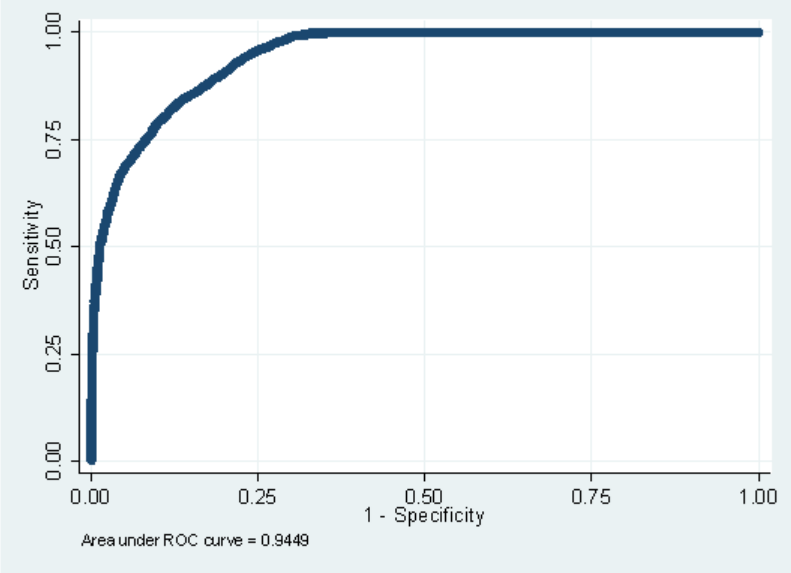


Şekil 7.15: von-Mises Fisher Kümeleme Algoritması Sonucu-Hasar Oranı ROC Analizi

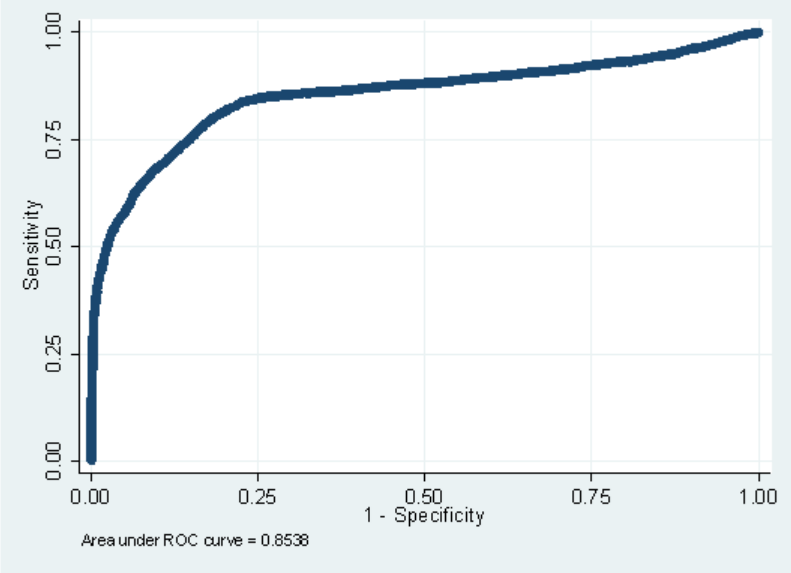


Şekil 7.16: Küresel K-ortalamalar Kümeleme Algoritması Sonucu-Hasar Oranı ROC Analizi

Ayrıca Şekil 7.17'e göre, hasar geçmişi olan bölgeler için sonlu karma von-Mises Fisher dağılımı kümeleme yöntemi elde edilen primler, ortalama sigorta bedelindeki gerçekleşen değişimin yüzde 94.5'lik kısmını açıklamaktadır.



Şekil 7.17: Sonlu Karma von Mises Fisher Kümeleme Algoritması Sonucu-Sigorta Teminatı ROC Analizi



Şekil 7.18: Küresel K-ortalamalar Kümeleme Algoritması Sonucu-Sigorta Teminatı ROC Analizi

Şekil 7.18'e göre, küresel k-ortalamalar algoritması kullanılarak hesaplanan aktüeryal primler, ortalama sigorta teminatındaki değişiminin yüzde 85'ini açıklayabilmektedir.

Bu sonuçlar ışığında, her iki yöntemin de ortalama hasar gerçekleşme oranı, olasılığı ve sigorta bedelinde gerçekleşen değişimi anlamlı bir şekilde açıkladığı sonucuna varılmıştır. Bu sonuç, düşük ve/veya yüksek değerlere sahip ortalama hasar gerçekleşme oranı ve olasılığı değişkenlerinin doğru bir şekilde ayrıldığını ve hesaplanan primlerin bu değişimi doğru bir şekilde ölçebildiğini göstermektedir. Her iki kümeleme çalışması karşılaştırıldığında, dağılım bazlı sonlu karma von-Mises Fisher dağılımının uzaklık bazlı konumsal k-ortalamalar algoritmasına göre hem hasar gerçekleşme oranı hem de hasar gerçekleşme olasılığı değerlerinde gerçekleşen değişimi daha iyi ölçebildiği sonucuna varılmıştır.

7.5. Farklı Küme Sayıları ile Hesaplanan Primlerin Karşılaştırılması

Uygulama kapsamında, kümeleme algoritmalarında kullanılacak optimal küme sayısı, hem silhoutte(gölge) yöntemiyle hem de sonlu karma von-Mises Fisher dağılımına ait log-olabilirlik ve BIC değerleri vasıtasıyla belirlenmişti. Bu analizlerin sonucunda ise 10 adet kümenin uygun olabileceği düşünülmüştü. Ancak bilindiği üzere, her ne kadar

optimal küme sayısı belirlenip hasar gerçekleşme oranı ve olasılığı değişkenleri ayrı ayrı incelenip, farklı kümeleme algoritmaları için açıklanabilirlik analizi yapılsa da, bazı koşullarda optimalliğin ötesini düşünerek hareket etmek gerekir.

Özellikle, tarım sigortalarının, gerek devlet politikaları gerekse ekonomik göstergelerden kolayca etkilenecek bir piyasa olduğu düşünülürse, her zaman optimallik düşüncesi ile ekonomik faaliyetleri sürdürmek mümkün olmayacaktır.

Bu bakış açısıyla, öncelikle konumsal kümeleme algoritmasında kullanılan küme sayısının değişimi incelenecek. Ardından bu durumdan etkilenen hasar bölgeleri için tekrar hasar gerçekleşme oranı ve hasar gerçekleşme olasılığı değerlerine göre kümeleme yapılacaktır. Sonucunda ise belirlenen küme sayısına göre risklilik ölçütünde yer alan hasar sayısı ve şiddeti tanımlarına denk gelen hasar gerçekleşme olasılığı ve hasar gerçekleşme oranı değerleri kullanılarak performans değerlendirmesi yapılacaktır.

Çalışmanın bu bölümünde, küme sayısı 8 ve 12 olacak şekilde farklı küme sayılarına göre hesaplanan aktüeryal risk priminin, hasar gerçekleşme oranı ve hasar gerçekleşme olasılığını ne kadar açıklayabildiği test edilmiştir.

Test edilecek küme sayısının seçiminde dikkat edilmesi gereken çeşitli durumlar söz konusudur. Örneğin, hasar bölgeleri için az sayıda küme seçilmesi durumunda, hasar gerçekleşme oranı ve hasar gerçekleşme olasılığı incelenirken, alt kümelerde yığılmalar olacaktır. Düşük, orta ve yüksek olarak derecelendirilen değişkenler, üç dereceden daha fazla sınıfa ayrılabilecektir ve bu da her bir hasar bölgesi için dallanmayı artıracaktır. Bu durumda, hasar bölgeleri, konumsal özelliklerine göre tam olarak ayrılmadığı için hatalı sonuçlar ortaya çıkabilecektir. Diğer taraftan, hasar bölgeleri için küme sayısı fazla seçildiğinde ise, hasar gerçekleşme oranı ve hasar gerçekleşme olasılığı değişkenlerinde gerçekleşen değişim tam olarak ölçülemeyecektir. Bu nedenle test edilecek küme sayıları, optimal küme sayısına yakın değerler seçilmiştir.

Çizelge 7.5: Farklı Küme Sayıları için ROC Alan Analizi

Küme Sayısı	Yöntem	Hasar Oranı	Hasar Olasılığı	Sigorta Bedeli
8	Sonlu Karma von-Mises Fisher Dağılımı	0.6654	0.8673	0.8763
	Küresel k-ortalamalar Algoritması	0.6138	0.7487	0.7893
10	Sonlu Karma von-Mises Fisher Dağılımı	0.8872	0.8549	0.9450
	Küresel k-ortalamalar Algoritması	0.7635	0.8351	0.8540
12	Sonlu Karma von-Mises Fisher Dağılımı	0.7484	0.8274	0.7830
	Küresel k-ortalamalar Algoritması	0.6063	0.7936	0.7698

Çizelge 7.5'e göre, test edilen küme sayısı 8 ve 12 olarak seçildiğinde, her iki küme sayısına göre de, sonlu karma von-Mises Fisher dağılımı kullanarak hesaplanan aktüeryal risk primi, küresel k-ortalamalar algoritması kullanarak yapılan aktüeryal risk primine göre hasar gerçekleşme oranında, hasar gerçekleşme olasılığında ve sigorta bedelinde gerçekleşen değişimi daha iyi açıklamaktadır.

Farklı küme sayıları ve optimal olarak seçilen 10 küme sayısı kullanılarak yapılan kümeleme çalışmalarına göre, optimal olarak 10 küme seçilerek yapılan prim hesabının diğer küme sayılarına göre yapılan hesaba göre hem hasar gerçekleşme oranında hem hasar gerçekleşme olasılığında hem de sigorta bedelinde gerçekleşen değişimi daha iyi açıkladığı görülmektedir.

8. SONUÇLAR VE ÖNERİLER

Riskliliği en doğru şekilde sınıflayan yöntemlerle adil risk primi hesabı yapılması hem sigortacı hem sigortalı açısından oldukça önemlidir.

Sigortacılık sektöründe, riskliliğin ölçülebilmesi için kullanılan istatistiksel yöntemlerden en yaygın olanı kümeleme çalışmalarıdır. Kümeleme çalışması, diğer tüm etkenler sabit tutulduğunda, bir etkendeki değişkenliğin sınıflandırılmasına imkan sağlamaktadır.

Bu tez çalışmasında, Türkiye'de buğday üretimi yapılan alanlarda hasar verisi kullanılarak dolu riski için adil prim hesaplanmıştır. Öncelikle bu bölgeler için konumsal özellikleri dikkate alınarak; kümeleme çalışması yapılmış, ardından ise sigorta geçmişi ele alınarak risk kümelerine göre sınıflama yapılmıştır. Bu çalışmanın ardından risk kümelerine ayrılan konumlar için net prim ilkesi kullanılarak, aktüeryal risk primi hesabı yapılmıştır.

Tarımsal zararları oluşturan risk türleri, içerisinde en sık tekrarlayanı ve en çok hasara sebep olanı hava durumuna bağlı olaylardır. Meteorolojik özellikler dikkate alınarak yapılacak olan kümeleme çalışması, her bölgenin kendine özgü olan riske katkı tipini ve büyüklüğünü bölgeler arasında ayrıştırarak, her bölgeye farklı bir sabit etki belirler. Maalesef meteorolojik istasyonlardan gelen sağlıksız ve yetersiz veriler ile yapılan endeksleme çalışmalarından doğan baz riskinden dolayı istasyon verisi bazlı kümeleme çalışmalarından çıkan sonuçlar şüphelidir. Bu sebeple bu tez çalışmasında meteorolojik değişkenleri kullanmak yerine konum bazlı hasar verisinden yararlanılmıştır. Bu tez çalışmasında TARSİM'den alınan 2010-2014 yılları arası 5 yıllık veri kullanılmıştır.

Dağılım bazlı ve uzaklık bazlı olarak iki farklı konumsal kümeleme yöntemi kullanılmıştır. Konumsal kümeleme yöntemleri karşılaştırıldığında, dağılım bazlı yönsel, uzaklık bazlı kümeleme yöntemine göre riskin sınıflandırılması daha iyi sonuçlar vermektedir. Sonuç olarak, her iki yöntemle hesaplanan primler karşılaştırıldığında, dağılım bazlı kümeleme algoritması kullanılarak hesaplanan primin, uzaklık bazlı kümeleme algoritması kullanılarak hesaplanan prime göre, hasar gerçekleşme oranında ve olasılığındaki değişimi daha iyi yansıttığı sonucuna varılmıştır. Ayrıca konumsal kümeleme algoritmaları kullanılırken, belirlenen optimal küme sayısı sınandığında, dağılım bazlı kümeleme algoritmasının daha da iyi sonuçlar verdiği görülmüştür.

Bu çalışmanın katkısı, hasar geçmişi olan tarımsal bölgeler için konumsal kümeleme

çalışması yapılarak, bu bölgeler için risk gruplarına göre adil primin alınabileceğini göstermektir. Gelecekte buğday üretimi yapılabilecek herhangi tarımsal üretim bölgesi için, hangi risk grubuna ait olduğu bilindiğinde, o risk grubuna ait hesaplanan ortalama hasar gerçekleşme oranı ve olasılık değişkenleri bilinecektir. Bu sayede, o bölgeye ait olarak ürünün birim fiyatı ve miktarı düşünülerek bildirilen sigorta bedeli bilgisi ile aktüeryal risk primi hesabı yapılabilecektir. Bu modelde, düşük hasar oranı ve olasılığı özelliği gösteren risk grupları için düşük prim, yüksek hasar oranı ve olasılığı özelliği gösteren risk grupları için yüksek prim alınması gerektiği sonucu elde edilmiştir.

Çalışmanın uygulama kısmında yer alan TARSİM veri setinde, incelenen köykodları için yazılı il, ilçe, köy bilgileri birleştirilerek, adresten koordinat bulan web bazlı internet siteleri kullanılırken, çıktı olarak enlem ve boylam bulunup, her bir çıktı değerinin uygunluğu test edilerek kullanılmıştır. İlerideki çalışmalarda adreslerden enlem ve boylam bulunması konusunda daha işlevsel bir çözüm üretilebilir ya da yazılım bazlı yardımlar alınarak bu sorunun üstesinden gelinebilir.

Gelecekte yapılacak çalışmalarda, farklı tarımsal ürün sigortaları ve farklı risk türleri için de konumsal kümeleme çalışmaları yapılarak aktüeryal risk primi hesaplanabilir. Ayrıca hasar bölgelerine ait tarihsel veri kullanılarak, Türkiye kapsamında ya da bölgesel uzay-zaman (spatio-temporal) analizleri yapılabilir.

KAYNAKLAR

[1] Sümer, G. and Polat, Y. Dünyada Tarım Sigortaları Uygulamaları ve TARSİM. *Gazi Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi*, 18(1):236–263, **2016**.

[2] Iturrioz, R. Agricultural insurance. Technical report, The World Bank, **2009**.

[3] Halcrow, H.G. Actuarial structures for crop insurance. *Journal of Farm Economics*, 31(3):418–443, **1949**.

[4] Skees, J.R. and Reed, M.R. Rate making for farm-level crop insurance: Implications for adverse selection. *American Journal of Agricultural Economics*, 68(3):653–659, **1986**.

[5] Salton, G. and McGill, M.J. *Introduction to Modern Information Retrieval*. McGraw-Hill, Inc., New York, NY, USA, **1986**.

[6] Rasmussen, E. Information retrieval. chapter Clustering Algorithms, pages 419–442. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, **1992**.

[7] Dhillon, I.S. and Modha, D.S. Concept decompositions for large sparse text data using clustering. *Machine Learning*, 42(1-2):143–175, **2001**.

[8] Maitra, R. and Ramler, I.P. A k-mean-directions Algorithm for Fast Clustering of Data on the Sphere. *Journal of Computational and Graphical Statistics*, 19(2):377–396, **2010**.

[9] Banerjee, A., Dhillon, I., Ghosh, J., and Sra, S. Generative model-based clustering of directional data. *ACM SIGKDD international conference on Knowledge discovery and data mining*, page 19, **2003**.

[10] Banerjee, A. and Dhillon, I. Clustering on the unit hypersphere using von Mises-Fisher distributions. *Journal of Machine Learning Research*, 6:1345–1382, **2005**.

[11] Ferrari, S.L.P. and Cribari-Neto, F. Beta Regression for Modelling Rates and Proportions. *Journal of Applied Statistics*, 31(7):799–815, **2004**.

[12] Espinheira, P.L., Ferrari, S.L.P., and Cribari-Neto, F. On Beta Regression Residuals. *Journal of Applied Statistics*, 35(4):407–419, **2008**.

[13] Miranda, M.J. Area-yield crop insurance reconsidered. *American Journal of Agricultural Economics*, 73(2):233–242, **1991**.

[14] Goodwin, B.K. Premium rate determination in the federal crop insurance program: what do averages have to say about risk? *Journal of Agricultural and Resource Economics*, pages 382–395, **1994**.

[15] T.O. Knight, and Coble, K.H. Survey of us multiple peril crop insurance literature since 1980. *Review of Agricultural Economics*, 19(1):128–156, **1997**.

[16] Skees, J.R., Black, J.R., and Barnett, B.J. Designing and rating an area yield crop insurance contract. *American journal of agricultural economics*, 79(2):430–438, **1997**.

[17] Goodwin, B.K. and Ker, A.P. Nonparametric estimation of crop yield distributions: implications for rating group-risk crop insurance contracts. *American Journal of Agricultural Economics*, 80(1):139–153, **1998**.

[18] Goodwin, B.K. Problems with market insurance in agriculture. *American Journal of Agricultural Economics*, 83(3):643–649, **2001**.

[19] Goodwin, B.K. and Mahul, O. Risk modeling concepts relating to the design and rating of agricultural insurance contracts. **2004**.

[20] Nelson, C.H. The influence of distributional assumptions on the calculation of crop insurance premia. *North Central Journal of Agricultural Economics*, 12(1):71–78, **1990**.

[21] Turvey, C.G. Weather derivatives for specific event risks in agriculture. *Review of Agricultural Economics*, 23(2):333–351, **2001**.

[22] Farzaneh, M., M.S.Allahyari,, Damalas, C.A., and Seidavi, A. Crop insurance as a risk management tool in agriculture: The case of silk farmers in northern iran. *Land Use Policy*, 64:225–232, **2017**.

[23] Lou, W. and Sun, S. Design of agricultural insurance policy for tea tree freezing damage in zhejiang province, china. *Theoretical and Applied Climatology*, pages 1–16, **2013**.

[24] Zhao, Y., Chai, Z., Delgado, M.S., and Preckel, P.V. A test on adverse selection of farmers in crop insurance: Results from inner mongolia, china. *Journal of Integrative Agriculture*, 16(2):478–485, **2017**.

[25] Zhang, X., Yin, W., Wang, J., Ye, T., Zhao, J., and Wang, J. Crop insurance premium ratemaking based on survey data: a case study from dingxing county, china. *International Journal of Disaster Risk Science*, 6(3):207, **2015**.

[26] Murray, K., Farrin,and A.G. The effect of index insurance on returns to farm inputs: Exploring alternatives to zambia's fertilizer subsidy program. In *Selected paper for Agricultural and Applied Economics Association meeting, Minneapolis*, pages 27–29, **2014**.

[27] Ahmed, O. and Serra, T. Economic analysis of the introduction of agricultural revenue insurance contracts in spain using statistical copulas. *Agricultural Economics*, 46(1):69–79, **2015**.

[28] Tack, J., Coble, k.H., and Barnett, B. Warming temperatures will likely induce higher premium rates and government outlays for the us crop insurance program. **2017**.

[29] Ghimire, Y.N., Timsina, K.P., Kandel, G., Devkota, D., Thapamagar, D.B., Gautam, S., and Sharma, B. Agricultural insurance issues and factors affecting its adoption: A case of banana insurance in nepal. *Evaluation*, 11(1), **2016**.

[30] Şahin, Ş., Karabey, U., Karageyik, B.B., Nevruz, E., and Yıldırak, K. Türkiye'de buğday bitkisel Ürün sigortası için aktüeryal prim hesabı. *Turkish Journal of Agricultural Economics*, 22(2), **2016**.

[31] Binici, T. and Zulauf, C.R. Determining wheat crop insurance premium based on area yield insurance scheme in konya province, turkey. *Journal of Applied Sciences*, 6:1148–1152, **2006**.

[32] Ömer Ozan Evkaya. *Modelling weather index based drought insurance for provınces in the Central Anatolıa regıon*. PhD thesis, Orta Doğu Teknik Üniversitesi, **2012**.

[33] Anonim,Paralel Meridyen Enlem Boylam. `http://www.turkcebilgi.org/cografya/genel-cografya/paralel-meridyen-enlem-boylam-32074.html`. (Haziran,**2017**).

[34] Peker, K.Ö. *Dairesel Veriler ve Ardışık Testlerde Kullanımı*. PhD thesis, Anadolu Üniversitesi,Fen Bilimleri Enstitüsü,İstatistik Anabilim Dalı, **2002**.

[35] Johnson, R.A. and Wichern, D.W. *Applied Multivariate Statistical Analysis*. **2007**.

[36] Peker, K.Ö. and Bacanlı, S. Dairesel verilere uygulanan tanımlayıcı İstatistiksel yöntemler ve meteorolojik bir uygulama. **2004**.

[37] Jammalamadaka, S.R. and SenGupta, A. *Topics in Circular Statistics*, volume 6. **2001**.

[38] Fisher, N.I., Lewis, T., and Embleton, B. J. J. Statistical Analysis of Spherical Data. page 343, **1987**.

[39] Mardia, K.V. and Jupp, P.E. *Directional statistics*, volume 494. John Wiley & Sons, **2009**.

[40] Dempster, A.P., Laird, N.M., and Rubin, D.B. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society Series B Methodological*, 39(1):1–38, **1977**.

[41] Sra, S. A short note on parameter approximation for von Mises-Fisher distributions: and a fast implementation of I s (x). *Computational Statistics*, 27(1):177–190, **2012**.

[42] Hornik, K. and Grün, B. movMF: An R Package for Fitting Mixtures of von Mises-Fisher Distributions. *Journal of Statistical Software*, 58(10):1–31, **2014**.

[43] Mardia, K.V. and Jupp, P.E. Directional Statistics. **1999**.

[44] Tanabe, A., Fukumizu, K., Oba, S., Takenouchi, T., and Ishii, S. Parameter estimation for von Mises-Fisher distributions. *Computational Statistics*, 22(1):145–157, **2007**.

[45] Neal, R.M. and Hinton, G.E. A view of the em algorithm that justifies incremental, sparse, and other variants. In *Learning in graphical models*, pages 355–368. Springer, **1998**.

[46] Kearns, M., Mansour, Y., and Ng, A.Y. An Information-Theoretic Analysis of Hard and Soft Assignment Methods for Clustering. *Proc. of Conference on Uncertainty in Artificial Intelligence*, pages 282–293, **1997**.

[47] Salah, A., Rogovschi, N., and Nadif, M. Model-based Co-clustering for High Dimensional Sparse Data. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, pages 866–874, **2016**.

[48] Buchta, C., Kober, M., Feinerer, I., and Hornik, K. Spherical k-means clustering. *Journal of Statistical Software*, 50(10):1–22, **2012**.

[49] Frakes, W.B. and Baeza-Yates, R. Information retrieval: data structures and algorithms. **1992**.

[50] Dhillon, I.S., Guan, Y., and Kogan, J. Iterative clustering of high dimensional text data augmented by local search. In *Data Mining, 2002. ICDM 2003. Proceedings. 2002 IEEE International Conference on*, pages 131–138. IEEE, **2002**.

[51] Dhillon, I.S., Fan, J., and Guan, Y. Efficient clustering of very large document collections. *Data mining for scientific and engineering applications*, 2:357–381, **2001**.

[52] McCullagh, P. Generalized linear models. *European Journal of Operational Research*, 16(3):285–292, **1984**.

[53] Ohlsson, E. and Johansson, B. Non-life insurance pricing. *Non-Life Insurance Pricing with Generalized Linear Models*, pages 1–14, **2010**.

[54] Zeileis, A., Cribari-Neto, F., Grün, B., Kosmidis, I., Simas, A.B., and Rocha, A.V. Package 'betareg'. **2016**.

[55] Smithson, M. and Verkuilen, J. A better lemon squeezer? Maximum-likelihood regression with beta-distributed dependent variables. *Psychological methods*, 11(1):54, **2006**.

[56] Simas, A.B., Barreto-Souza, W., and Rocha, A.V. Improved estimators for a general class of beta regression models. *Computational Statistics & Data Analysis*, 54(2):348–366, **2010**.

[57] Grün, B., Kosmidis, I., and Zeileis, A. Extended beta regression in r: shaken, stirred, mixed, and partitioned. Technical report, Working Papers in Economics and Statistics, **2011**.

[58] Grün, B. and Leisch, F. Flexmix version 2: finite mixtures with concomitant variables and varying and constant parameters. *Journal of Statistical Software*, 28(4):1–35, **2008**.

[59] Rousseeuw, P.J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65, **1987**.

[60] Chen, G., Jaradat, S.A., Banerjee, N., Tanaka, T.S., Ko, M.S.H., and Zhang, M.Q. Evaluation and comparison of clustering algorithms in analyzing es cell gene expression data. *Statistica Sinica*, pages 241–262, **2002**.

[61] Wang, K., Wang, B., and Peng, L. Cvap: Validation for cluster analyses. *Data Science Journal*, 8:88–93, **2009**.

[62] TARSİM. *Devlet Destekli Bitkisel Ürün Sigortası Genel Şartlar*, **2016**.

[63] TARSİM. *Devlet Destekli Bitkisel Ürün Sigortası Tarife ve Talimatlar*, **2016**.

[64] Rivest, L.P. Modified Kent's statistics for testing goodness of fit for the Fisher distribution in small concentrated samples. *Statistics & probability letters*, 4(1):1–4, **1986**.

Ek 1: 10 Küme için Sonlu Karma von-Mises Fisher Dağılımı Kümeleme Çalışması Parametre Sonuçları

Küme Sayısı	Theta	x	y	z	Kappa	fishkent test p-değeri
1	0.1870	0.7827601	0.4713487	0.4063460	2440.032	0.511
2	0.1508	0.7854139	0.5401032	0.3023467	7361.261	0.546
3	0.1670	0.7578859	0.5332089	0.3758951	7135.873	0.530
4	0.0846	0.7670537	0.5634131	0.3069108	8426.419	0.496
5	0.1182	0.7961348	0.4946601	0.3485410	5378.812	0.510
6	0.1194	0.7746361	0.5162052	0.3653369	8428.317	0.547
7	0.0131	0.7469053	0.5930282	0.3007490	96838.750	0.612
8	0.0182	0.7513593	0.5834065	0.3083765	49402.620	0.532
9	0.1130	0.7702573	0.5417944	0.3363964	7887.499	0.511
10	0.0292	0.7585522	0.5818313	0.2933783	30091.500	0.484

Ek 2: Hasar Gerçekleşme Oranı ve Olasılığı Değerleri için AIC Değerleri

Sonlu Karma von-Mises Fisher-10Kume				Küresel k-ortalamalar-10Kume			
1. Katman Küme Adı	1. Katman AIC Değeri	2. Katman Küme Adı	2. Katman AIC Değeri	1. Katman Küme Adı	1. Katman AIC Değeri	2. Katman Küme Adı	2. Katman AIC Değeri
10.1	-352.6793	10.1.1	-528.7032	10.1	-291.7139	10.1.1	-579.4496
		10.1.2	-38.0358	10.2	-85.44071	10.2.1	-262.4051
10.2	-83.4620	10.2.1	-51.1047	10.3	-548.9324	10.3.1	-170.2417
		10.2.2	-7.1626			10.3.2	-83.2915
		10.2.3	-57.6790			10.3.3	-807.9889
		10.2.4	-157.9614			10.3.4	-254.6772
10.3	-549.5134	10.3.1	-89.3448	10.4	-528.9043	10.4.1	-829.5966
		10.3.2	-251.9071			10.4.2	-228.2012
		10.3.3	-175.5417	10.5	-232.7585	10.5.1	-726.3368
		10.3.4	-836.7276			10.5.2	-262.3127
10.4	-32.5701	10.4.1	-80.9629			10.5.3	-35.4049
		10.4.2	-2.0176			10.5.4	-27.1857
10.5	-226.7589	10.5.1	-27.2412	10.6	-194.9488	10.6.1	-130.0408
		10.5.2	-16.8818			10.6.2	-1044.1660
		10.5.3	-303.5007	10.7	-77.3741	10.7.1	-43.0905
		10.5.4	-43.1103			10.7.2	-304.9316
10.6	-194.9487	10.6.1	-1044.1660	10.8	-147.5542	10.8.1	-251.0974
		10.6.2	-130.0359			10.8.2	-125.6981
10.7	-77.3429	10.7.1	-272.3190	10.9	-513.7670	10.9.1	-65.9231
		10.7.2	-69.2550			10.9.2	-525.0572
10.8	-144.4057	10.8.1	-114.3714			10.9.3	-647.2911
		10.8.2	-82.1888			10.9.4	-483.4996
		10.8.3	-180.1965	10.10	-122.8807	10.10.1	-94.0629
10.9	-517.9456	10.9.1	-452.8704			10.10.2	-349.1516
		10.9.2	-1156.715				
		10.9.3	-62.0563				
10.10	-118.1937	10.10.1	-11.44536				
		10.10.2	-72.6145				
		10.10.3	-223.4660				

Ek 3: Küresel K-Ortalamalar Kümeleme Sonucu-Beta Uyum İyiliği Testi Sonuçları

Küresel k-ortalamalar Yöntemi-Beta dağılımı Uyum İyiliği Test Sonuçları							
Küme	Değişken	Kolmogorov Smirnov Test İstatistiği Değeri	Anderson Darling Test İstatistiği Değeri	Küme	Değişken	Kolmogorov Smirnov Test İstatistiği Değeri	Anderson Darling Test İstatistiği Değeri
111	Olasılık	0.1571	0.9261	611	Olasılık	0.1541	2.1901
	Ortalama	0.1992	2.1682		Ortalama	0.2230	0.9037
112	Olasılık	0.0943	2.3552	612	Olasılık	0.2562	2.0109
	Ortalama	0.0423	0.3007		Ortalama	0.2909	1.0210
211	Olasılık	0.1585	1.3962	621	Olasılık	0.1029	0.9211
	Ortalama	0.1867	2.6369		Ortalama	0.0972	0.5601
212	Olasılık	0.1525	1.4363	622	Olasılık	0.1935	2.7023
	Ortalama	0.0596	0.1885		Ortalama	0.0762	0.3942
213	Olasılık	0.1102	0.6105	711	Olasılık	0.0503	1.2181
	Ortalama	0.1400	1.5514		Ortalama	0.0957	1.1080
311	Olasılık	0.4480	1.1650	712	Olasılık	0.0778	1.5227
	Ortalama	0.1609	1.2649		Ortalama	0.0544	0.7469
321	Olasılık	0.2620	1.7970	721	Olasılık	0.2148	2.0345
	Ortalama	0.1599	1.6803		Ortalama	0.2381	0.7284
322	Olasılık	0.2130	0.5810	722	Olasılık	0.1875	1.7772
	Ortalama	0.2751	1.4485		Ortalama	0.0945	0.7480
331	Olasılık	0.2762	0.6701	811	Olasılık	0.1027	1.3806
	Ortalama	0.3211	1.5598		Ortalama	0.0742	0.2076
341	Olasılık	0.3511	1.1020	812	Olasılık	0.1625	2.2264
	Ortalama	0.5183	1.7656		Ortalama	0.0673	0.1797
342	Olasılık	0.1507	0.9351	813	Olasılık	0.0481	1.501
	Ortalama	0.1889	0.5893		Ortalama	0.0519	0.6798
343	Olasılık	0.2131	0.5810	821	Olasılık	0.2377	1.5636
	Ortalama	0.1787	2.9918		Ortalama	0.2307	1.3780
344	Olasılık	0.3189	1.4540	911	Olasılık	0.1047	0.7987
	Ortalama	0.3173	1.3844		Ortalama	0.1889	1.0847
411	Olasılık	0.1147	0.9053	912	Olasılık	0.1761	1.9283
	Ortalama	0.1091	1.2081		Ortalama	0.1071	2.2488
412	Olasılık	0.0515	0.5283	921	Olasılık	0.2244	1.2603
	Ortalama	0.0459	0.3516		Ortalama	0.2055	4.0195
421	Olasılık	0.2873	1.1055	931	Olasılık	0.07001	4.2137
	Ortalama	0.1972	1.6117		Ortalama	0.0706	1.1098
511	Olasılık	0.1412	1.0798	932	Olasılık	0.1212	1.3999
	Ortalama	0.2808	1.1969		Ortalama	0.0771	0.5352
512	Olasılık	0.1046	2.2919	941	Olasılık	0.1664	6.2420
	Ortalama	0.0796	1.5719		Ortalama	0.0917	0.3529
521	Olasılık	0.1721	0.7028	942	Olasılık	0.0604	2.6362
	Ortalama	0.2397	1.2764		Ortalama	0.0381	0.4538
522	Olasılık	0.1371	2.9654	1011	Olasılık	0.1511	0.9891
	Ortalama	0.2911	1.7185		Ortalama	0.1731	1.0542
531	Olasılık	0.1318	0.7041	1012	Olasılık	0.1666	1.516
	Ortalama	0.2561	2.0363		Ortalama	0.2522	0.4885
532	Olasılık	0.5121	1.6521	1021	Olasılık	0.0732	2.1682
	Ortalama	0.7115	2.9836		Ortalama	0.0884	0.3688
541	Olasılık	0.5791	1.0109	1022	Olasılık	0.2391	2.5306
	Ortalama	0.6594	1.9848		Ortalama	0.1436	1.7962
542	Olasılık	0.2597	2.5857				
	Ortalama	0.2613	2.7798				

Ek 4: Sonlu Karma von Mises Fisher Dağılımı İçin Beta Dağılımı Uyum İyiliği Testi Sonuçları

Sonlu Karma von Mises-Fisher dağılımı -Beta Dağılımı Uyum İyiliği Test Sonuçları							
Küme	Değişken	Kolmogorov Smirnov Test İstatistiği Değeri	Anderson Darling Test İstatistiği Değeri	Küme	Değişken	Kolmogorov Smirnov Test İstatistiği Değeri	Anderson Darling Test İstatistiği Değeri
111	Olasılık	0.0754	0.7440	611	Olasılık	0.1862	1.3224
	Ortalama	0.0670	0.3327		Ortalama	0.1576	0.4809
112	Olasılık	0.0845	0.4392	612	Olasılık	0.0794	2.3652
	Ortalama	0.0726	0.2412		Ortalama	0.0679	2.0187
113	Olasılık	0.0409	0.4715	621	Olasılık	0.1492	1.0741
	Ortalama	0.0667	1.0931		Ortalama	0.0637	0.5059
121	Olasılık	0.1137	1.3305	622	Olasılık	0.1003	1.1224
	Ortalama	0.1276	0.9928		Ortalama	0.0845	2.1990
122	Olasılık	0.1111	1.1187	711	Olasılık	0.1549	1.0369
	Ortalama	0.1417	2.2132		Ortalama	0.1863	1.3224
211	Olasılık	0.0901	0.8977	712	Olasılık	0.1854	2.0466
	Ortalama	0.0955	1.0448		Ortalama	0.1087	1.7244
221	Olasılık	0.1446	1.9679	721	Olasılık	0.3067	1.6645
	Ortalama	0.0982	1.6023		Ortalama	0.2217	1.1461
231	Olasılık	0.2635	0.7540	722	Olasılık	0.1936	0.9427
	Ortalama	0.2636	0.7541		Ortalama	0.1924	1.1704
232	Olasılık	0.2499	2.3729	811	Olasılık	0.1311	1.1209
	Ortalama	0.2499	1.3812		Ortalama	0.1732	2.1598
233	Olasılık	0.1090	0.6451	821	Olasılık	0.1205	1.3964
	Ortalama	0.1269	2.4773		Ortalama	0.1414	1.6472
241	Olasılık	0.1570	1.4513	822	Olasılık	0.1931	1.2243
	Ortalama	0.1018	0.9929		Ortalama	0.0891	0.8752
242	Olasılık	0.2150	1.3419	831	Olasılık	0.1591	2.0205
	Ortalama	0.2026	2.5162		Ortalama	0.1860	1.0684
243	Olasılık	0.1244	0.9858	832	Olasılık	0.1999	1.4781
	Ortalama	0.2540	1.8608		Ortalama	0.2683	2.0294
244	Olasılık	0.3511	1.1020	911	Olasılık	0.2635	0.7540
	Ortalama	0.3183	1.7656		Ortalama	0.2635	0.7540
311	Olasılık	0.1388	0.2805	912	Olasılık	0.1498	1.1809
	Ortalama	0.1443	2.3750		Ortalama	0.1649	1.9729
312	Olasılık	0.1803	0.8612	921	Olasılık	0.1071	1.4232
	Ortalama	0.1538	1.6469		Ortalama	0.0314	0.3093
321	Olasılık	0.1058	2.2708	922	Olasılık	0.1591	0.8149
	Ortalama	0.0440	0.6152		Ortalama	0.2028	1.3444
331	Olasılık	0.1047	0.7987	931	Olasılık	0.0281	0.2792
	Ortalama	0.1889	1.0847		Ortalama	0.0524	0.4181
332	Olasılık	0.1802	1.3695	1011	Olasılık	0.2799	1.3531
	Ortalama	0.1113	1.3329		Ortalama	0.2686	1.6019
341	Olasılık	0.1022	1.1607	1021	Olasılık	0.2539	2.8408
	Ortalama	0.0617	0.6924		Ortalama	0.2618	0.8373
342	Olasılık	0.1123	1.1656	1031	Olasılık	0.1467	1.6776
	Ortalama	0.0662	2.3339		Ortalama	0.0713	0.6143
343	Olasılık	0.0619	0.6131				
	Ortalama	0.0491	0.3706				
344	Olasılık	0.0610	1.4106				
	Ortalama	0.0412	0.1859				
411	Olasılık	0.2440	1.5265				
	Ortalama	0.2440	1.5265				
412	Olasılık	0.1533	1.4551				
	Ortalama	0.1126	1.1415				
421	Olasılık	0.3686	1.0705				
	Ortalama	0.3078	1.3129				
511	Olasılık	0.1620	0.9375				
	Ortalama	0.1706	1.2102				
521	Olasılık	0.2052	2.3535				
	Ortalama	0.2440	1.5265				
522	Olasılık	0.1119	1.4021				
	Ortalama	0.0698	0.4810				
531	Olasılık	0.1315	1.4961				
	Ortalama	0.1622	1.1337				
532	Olasılık	0.1284	1.7748				
	Ortalama	0.1119	2.4021				
541	Olasılık	0.2194	1.3428				
	Ortalama	0.3320	2.0341				
542	Olasılık	0.3299	2.3456				
	Ortalama	0.2027	1.3770				

ÖZGEÇMİŞ

Kimlik Bilgileri

Adı-Soyadı : İsmail GÜR
Doğum Yeri : ANKARA
Medeni Hali : Bekar
E-posta : ismail.gur@hacettepe.edu.tr
Adresi : Çiçekli Mah. Boztepe Sok. 25/3 İncirli/ANKARA

Eğitim:

Lise : 2006-2010 TED Ankara Koleji Özel Lisesi
Lisans : 2010-2014 Hacettepe Üniversitesi Aktüerya Bilimleri Bölümü
Yandal-Lisans : 2010-2014 Hacettepe Üniversitesi İktisat Bölümü
Yüksek Lisans : 2014-2017 Hacettepe Üniversitesi Aktüerya Bilimleri Bölümü

Yabancı Dil ve Düzeyi

İngilizce : YDS 80 (Eylül 2014)

İş Deneyimi

2014 Eylül-2015 Şubat :Fırat Üniversitesi Aktüerya Bilimleri Bölümü-
Araştırma Görevlisi

2015 Şubat-.... :Hacettepe Üniversitesi Aktüerya Bilimleri Bölümü
Araştırma Görevlisi

Deneyim Alanları
Hayatdışı Sigortaları Matematiği
Konumsal Kümeleme Çalışmaları
Veri Madenciliği

Tezden Üretilmiş Projeler ve Bütçesi

Tezden Üretilmiş Yayınlar

Tezden Üretilmiş Tebliğ ve/veya Poster Sunumu ile Katıldığı Toplantılar

-Gür İ.., Yıldırak K.,Lazoğlu Ç., Frost Risk Premium Calculation Using Spatial Clustering. 3nd International Researchers, Statisticians And Young Statisticians Congress(IRSYSC 2017),24-26 Mayıs 2017, Konya,Türkiye

-Lazoğlu Ç., Gür İ., Estimation of Earthquake Probabilities With Non-Parametric Methods In Semi-Markov Model, 2nd International Conference on Computational Mathematics and Engineering Sciences(CMES-2017),20-22 Mayıs 2017,İstanbul,Türkiye

-Gür İ., Yıldırak K., Dolu Sebepli Kayısı Hasar Tahmin Modeli: Dairesel Regresyon Uygulaması, Yöneylem Arastırması ve Endüstri Mühendisligi 36. Ulusal Kongresi,13-15 Temmuz 2016,İzmir,Türkiye

HACETTEPE ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ
YÜKSEK LİSANS/DOKTORA TEZ ÇALIŞMASI ORJİNALLİK RAPORU

HACETTEPE ÜNİVERSİTESİ
FEN BİLİMLER ENSTİTÜSÜ
AKTÜERYA BİLİMLERİ BÖLÜMÜ ANABİLİM DALI BAŞKANLIĞI'NA

Tarih: 12./07/2017

Tez Başlığı / Konusu: TARIM SİGORTALARINDA KONUMSAL KÜMELEME ÜZERİNE BİR ÇALIŞMA

Yukarıda başlığı/konusu gösterilen tez çalışmamın a) Kapak sayfası, b) Giriş, c) Ana bölümler d) Sonuç kısımlarından oluşan toplam 82 sayfalık kısmına ilişkin, 11/07/2017 tarihinde şahsım/tez danışmanım tarafından *Turnitin* adlı intihal tespit programından aşağıda belirtilen filtrelemeler uygulanarak alınmış olan orijinallik raporuna göre, tezimin benzerlik oranı % 4 'tür.

Uygulanan filtrelemeler:

1- Kaynakça hariç
2- Alıntılar hariç
3- 5 kelimeden daha az örtüşme içeren metin kısımları hariç

Hacettepe Üniversitesi Fen Bilimleri Enstitüsü Tez Çalışması Orjinallik Raporu Alınması ve Kullanılması Uygulama Esasları'nı inceledim ve bu Uygulama Esasları'nda belirtilen azami benzerlik oranlarına göre tez çalışmamın herhangi bir intihal içermediğini; aksinin tespit edileceği muhtemel durumda doğabilecek her türlü hukuki sorumluluğu kabul ettiğimi ve yukarıda vermiş olduğum bilgilerin doğru olduğunu beyan ederim.

Gereğini saygılarımla arz ederim.

Tarih ve İmza

12.07.2017

Adı Soyadı: İSMAİL GÜR
Öğrenci No: N14124550
Anabilim Dalı: AKTÜERYA BİLİMLERİ BÖLÜMÜ
Programı: AKTÜERYA BİLİMLERİ BÖLÜMÜ
Statüsü: ☑ Y.Lisans ☐ Doktora ☐ Bütünleşik Dr.

DANIŞMAN ONAYI

UYGUNDUR.

DOÇ. DR. ŞAHAP KASIRGA YILDIRAK

(Unvan, Ad Soyad, İmza)