

**ÇUKUROVA ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ**

DOKTORA TEZİ

Özlem AKAY

PANEL VERİLERDE KÜMELEME ANALİZİ

İSTATİSTİK ANABİLİM DALI

ADANA-2018

**ÇUKUROVA ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ**

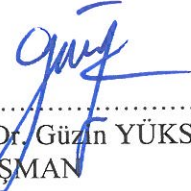
PANEL VERİLERDE KÜMELEME ANALİZİ

Özlem AKAY

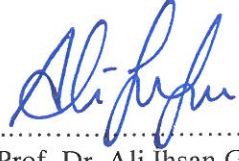
DOKTORA TEZİ

İSTATİSTİK ANABİLİM DALI

Bu tez 27/06/2018 Tarihinde Aşağıdaki Jüri Üyeleri Tarafından
Oybirliği/Oyçokluğu İle Kabul Edilmiştir.


Prof. Dr. Güzin YÜKSEL
DANIŞMAN


Prof. Dr. Hamza EROL
ÜYE


Prof. Dr. Ali İhsan GENÇ
ÜYE


Doç. Dr. Gülsen KIRAL
ÜYE


Doç. Dr. Nazif ÇALIŞ
ÜYE

Bu tez Enstitümüz İstatistik Anabilim Dalında hazırlanmıştır.
Kod No:

**Prof. Dr. Mustafa GÖK
Enstitü Müdürü**

Not: Bu tezde kullanılan özgün ve başka kaynaktan yapılan bildirişlerin, çizelge ve fotoğrafların kaynak gösterilmeden kullanımı, 5846 sayılı Fikir ve Sanat Eserleri Kanunundaki hükümlere tabidir.

ÖZ

DOKTORA TEZİ

PANEL VERİLERDE KÜMELEME ANALİZİ

Özlem AKAY

**ÇUKUROVA ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ
İSTATİSTİK**

Danışman : Prof. Dr. Güzin YÜKSEL
Yıl: 2018, Sayfa: 131
Jüri : Prof. Dr. Güzin YÜKSEL
: Prof. Dr. Hamza EROL
: Prof. Dr. Ali İhsan GENÇ
: Doç. Dr. Gülsen KIRAL
: Doç. Dr. Nazif ÇALIŞ

Panel veri yöntemi son yıllarda oldukça yaygın hale gelmiştir. Bunun nedeni, panel veri setinin, hem yatay kesit hem de zaman olmak üzere iki boyuta sahip olmasıdır. Panel veri zaman boyunca bireyler, firmalar, ülkeler vb. ile ilgili olduklarından bu birimlerde heterojenlik söz konusu olabilmektedir. Ancak panel veri modelinde ilgilenilen eğim parametresinin birimlere göre homojen olduğu varsayımı yapılmaktadır. Eğim parametrelerinin heterojenliğinin göz ardı edilmesi hatalı sonuçlara neden olacaktır.

Bu çalışmada, panel veri setinin birimlere göre heterojen olduğu dikkate alınarak panel veriye kümeleme analizi uygulanmıştır. Karma değişkenli panel veriyi kümelemek için yeni bir uzaklık ölçümü tanımlanmıştır. Gower, k-prototip uzaklıkları ve önerilen yeni uzaklık ölçümü toplamsal hiyerarşik kümeleme algoritmalarına entegre edilerek kümeleme analizi yapılmıştır. Ayrıca k-prototip algoritması ile de kümeleme yapılarak sonuçlar incelenmiştir. Kümeleme analizi ile birbirine benzer birimler aynı kümelerde gruplanmıştır. Elde edilen gruplara panel veri analizi uygulanmış ve parametre tahminleri yapılmıştır.

Anahtar Kelimeler: Panel veri, Kümeleme analizi, Uzaklık ölçümleri, Karma değişkenli panel veri

ABSTRACT

PhD THESIS

CLUSTERING ANALYSIS OF PANEL DATA

Özlem AKAY

ÇUKUROVA UNIVERSITY
INSTITUTE OF NATURAL AND APPLIED SCIENCES
DEPARTMENT OF STATISTICS

Supervisor : Prof. Dr. Güzin YÜKSEL
Yıl: 2018, Pages: 131
Jury : Prof. Dr. Güzin YÜKSEL
: Prof. Dr. Hamza EROL
: Prof. Dr. Ali İhsan GENÇ
: Assoc. Prof. Dr. Gülsen KIRAL
: Assoc. Prof. Dr. Nazif ÇALIŞ

The panel data method has become quite common in recent years because of the two dimensional nature of the panel data set; i.e. cross section and time section. Over the dimension of time, panel data are related to individuals, firms, countries and so forth. Thus, heterogeneity may occur in these units. However, it is assumed that the slope parameter of interest in the panel data model is homogeneous with respect to the units. Ignoring the heterogeneity of the slope parameters will cause incorrect results.

In this study, panel data cluster analysis was applied considering that the panel data set is heterogeneous according to the units. A new distance measurement has been defined to cluster the mixed variable panel data. The cluster analysis was performed by integrating Gower, k-prototypes distances and the proposed new distance measurement into the agglomerative hierarchical clustering algorithms. In addition, clustering has been done with the k-prototype algorithm and the results have been examined. Similar units are grouped in the same clusters by clustering analysis. Panel data analysis was applied to the obtained groups and parameter estimations were made.

Key Words: Panel Data, Cluster Analysis, Distance Measures, Mixed variable panel data

GENİŞLETİLMİŞ ÖZET

Panel veri, yatay kesit birimlerinin farklı zaman noktaları için tekrarlı gözlemlerini içerir. Bu tip veriler için boylamsal (longitudinal) veri ve tekrarlı ölçümler ifadeleri de kullanılmaktadır. Panel veri zaman boyunca bireyler, firmalar, ülkeler vb. ile ilgili olduklarından bu birimlerde heterojenlik söz konusu olabilmektedir. Ancak panel veri modelinde ilgilenilen eğim parametresinin birimlere göre homojen olduğu varsayılır. Eğim parametresi heterojenliğinin göz ardı edilmesi hem teorik hem de istatistiksel olarak geçerli tahminlerin yapılamamasına ve hatalı sonuçlara neden olacaktır. Bu nedenle panel veri setinin birimlere göre heterojen olduğu dikkate alınarak panel veriye kümeleme analizi uygulanır. Kümeleme analizi, farklı birimleri farklı kümelerde, benzer birimleri aynı kümede bir araya getirerek veri seti içerisindeki gizli kümeleri ortaya çıkaran ve bu kümelerin özelliklerini belirleyerek kitle hakkında tahminlerin yapılmasına olanak sağlayan yöntemlerden oluşan çok değişkenli istatistiksel bir analizdir. Kümeleme analizi ile küme içindeki birimler homojen kümeler arası birimlerde heterojen olacaktır.

Bu çalışmada, ilk olarak panel veri, panel veri modelleri ve panel veri modellerindeki eğim heterojenliğinden bahsedilmiştir. Panel verinin genel varsayımları verilerek varsayımlardan sapma durumunda, parametre taminleri için kullanılacak uygun tahmin yöntemleri açıklanmıştır. Sonra, kümeleme analizi ve toplamsal hiyerarşik kümeleme algoritmaları anlatılmıştır. Toplamsal hiyerarşik kümeleme algoritmalarında kullanılacak uzaklık ölçümleri verilmiştir. Nümerik, ikili ve karma (nümerik ve katagorik) değişkenli veri setlerinde kullanılacak uzaklık ölçümleri açıklanmıştır. Küme sayısı belirlemede kullanılan yöntemler ve küme geçerlilikleri yöntemleri verilmiştir. Daha sonra, panel veride kümeleme analizinin amacı ve uygulanma şekli anlatılmıştır. Ayrıca karma değişkenli panel verilerde kümeleme analizi yapılabilmesi için yeni bir uzaklık ölçümü tanımlanmıştır. Tanımlanan uzaklık ölçümünde, ilk olarak üç boyutlu panel verinin

iki boyutu çevrilerek tek boyuta dönüştürülür. Tek boyuttaki veri nümerik değişkenler ve ikili değişkenler olmak üzere iki alt boyuta ayrılır. Nümerik değişkenler $[0, 1]$ aralığında normalleştirilerek uzaklık ölçümü hesaplanır. Aynı şekilde ikili verilerde basit eşleştirme yöntemiyle uzaklık ölçümü elde edilir. Elde edilen uzaklık ölçümleri ağırlıklandırılarak birleştirilir vetoplamsal hiyerarşik kümeleme yöntemleriyle kullanılarak birimler kümelenir. Karma değişkenli verilerde kullanılan Gower, k-prototip uzaklıkları da karma değişkenli panel verileri kümelemek için uyarlanıp toplamsal hiyerarşik kümeleme yöntemleriyle kullanılmıştır. Ayrıca k-prototip algoritması karma değişkenli panel veriyi kümelemek için kullanılmıştır. Uygulama için 1993-2005 yılları arasında 25 ülkenin bulunduğu Turizm gelirleri karma değişkenli panel veri setine Gower, k-prototip ve önerilen uzaklık kullanılarak toplamsal hiyerarşik kümeleme yöntemleri ve k-prototip algoritması yöntemi ile kümeleme analizi yapılmıştır. Toplamsal hiyerarşik yöntemlerde uzaklık ölçümlerinin performanslarını kofenetik korelasyon katsayı değerleri ile karşılaştırılmıştır. Veri seti için uygun küme sayısı belirlendikten sonra uzaklık ölçümlerine göre kümeleme yöntemlerinin geçerlilik değerleri verilmiştir. Tüm uzaklıklar için küme sayının ortak olduğu Ward yöntemi ile kümeleme analizi yapılmıştır. Üç uzaklık içinde ikinci kümedeki birimler aynı olmasına rağmen birinci ve üçüncü kümedeki birimler değişiklik göstermiştir. Üç uzaklık kullanılarak elde edilen her bir küme için panel veri modelleri oluşturularak parametre tahmini yapılmıştır. Kullanılacak olan parametre tahmin yöntemini belirlemek için her bir kümeye yatay kesit bağımlılığı, birim kök testi, değişen varyans ve otokorelasyon testleri yapılmıştır. Yapılan testler sonucunda her üç uzaklık kullanılarak elde edilen kümelerin hepsinde yatay kesit bağımlılığı, birinci farkı alınan değişkenlerin durağan olduğu ve aynı zamanda değişen varyans ve otokorelasyon olduğu saptanmıştır. Bu nedenle parametre tahmini için modelde yatay kesit bağımlılığı, değişen varyans ve otokorelasyon varlığında dirençli tahmin edici olan Driscoll ve Kraay standart hatalarla sabit etkiler tahmini kullanılarak parametreler tahmin edilmiştir.

Sonuç olarak panel veri analizi yapmadan önce panel veriye kümeleme analizi uygulanmasının birbirine benzer birimleri aynı kümede toplayacağından daha doğru parametre tahminleri elde etmede yardımcı olacağı sonucuna varılmıştır.





TEŞEKKÜR

Bu çalışmanın gerçekleştirilmesinde, değerli bilgilerini benimle paylaşan, her sorun yaşadığımda yanına çekinmeden gidebildiğim, güler yüzünü ve samimiyetini benden esirgemeyen kıymetli danışman hocam Prof. Dr. Güzin YÜKSEL'e, çalışmaya olan katkıları ve yapıcı eleştirileri için Tez İzleme Komitesi üyelerim olan Sayın Prof. Dr. Hamza EROL ve Sayın Doç. Dr. Gülsen KIRAL hocalarıma teşekkürü bir borç bilir ve şükranlarımı sunarım.

Ayrıca bugüne kadar manevi desteğini hiç esirgemeyen her zaman yanımda olan aileme en içten sevgilerimi ve teşekkürlerimi sunarım.

İÇİNDEKİLER

SAYFA

ÖZ.....	I
ABSTRACT	II
GENİŞLETİLMİŞ ÖZET	III
TEŞEKKÜR.....	VII
İÇİNDEKİLER.....	VIII
ÇİZELGELER DİZİNİ.....	XII
ŞEKİLLER DİZİNİ	XVI
KISALTMALAR	XX
1. GİRİŞ.....	1
2. ÖNCEKİ ÇALIŞMALAR.....	5
3. PANEL VERİ VE PANEL VERİ MODELLERİ	11
3.1. Klasik Etki Modeli	13
3.2. Sabit Etki Modelleri	14
3.2.1. Birim Etkisine Sahip Sabit Etki Modeli	14
3.2.2. Zaman Etkisine Sahip Sabit Etki Modeli.....	15
3.2.3. Hem Birim hem de Zaman Etkisine Sahip Sabit Etki Modeli (İki Yönlü Sabit Etki Modeli).....	16
3.3. Rasgele Etki Modelleri	16
3.4. Heterojenlik	17
3.5. Panel Veri Modellerinde Değişen Varyans Testi	19
3.5.1. Değiştirilmiş Wald Testi.....	20
3.5.2. Levene, Brown ve Forsythe Testi.....	21
3.6. Panel Veri Modellerinde Otokorelasyon Testi	21
3.6.1. Baltagi Wu LBI Testi	22
3.6.2. Değiştirilmiş Durbin Watson Testi.....	23
3.7. Panel Veri Modellerinde Yatay Kesit Bağımlılığı Testi.....	24
3.7.1. Breusch Pagan Lagrange Çarpanı (LM).....	24

3.7.2. Friedman Testi.....	25
3.7.3. Frees Q Testi.....	26
3.8. Panel Birim Kök Testleri.....	27
3.9. Eş Bütünleşme Testleri.....	30
3.10. Varsayımlardan Sapmalara Dirençli Tahmin Ediciler.....	34
3.10.1. Parks-Kmenta Tahmincisi	36
3.10.2. Beck-Katz Tahmincisi	38
3.10.3. Driscoll ve Kraay Tahmincisi.....	39
4. KÜMELEME YÖNTEMLERİ VE ALGORİTMALAR	41
4.1. Hiyerarşik Kümeleme.....	42
4.1.1. Nümerik Değişkenler için Kullanılan Uzaklık Ölçüleri	43
4.1.2. İki Sonuçlu (Binary) Değişkenler için Kullanılan Uzaklık Ölçüleri .	44
4.1.3. Karma Değişkenler için Kullanılan Uzaklık Ölçüleri	46
4.1.3.1. k-prototip Uzaklığı	47
4.1.3.2. Optimal Ağırlık Uzaklığı.....	47
4.1.3.3. Goodall Uzaklığı.....	48
4.1.3.4. Gower Benzerlik Katsayısı	49
4.1.4. Matris Güncelleme Algoritmaları.....	50
4.2. Hiyerarşik Olmayan Kümeleme Algoritmaları	60
4.2.1. k-ortalamlar Algoritması.....	60
4.2.2. k-modlar Algoritması	62
4.2.3. k-prototip Algoritması	64
4.3. Küme Sayısı Tahmini	65
4.4. Küme Geçerlilikleri	66
4.4.1. Kofenetik Korelasyon Katsayısı.....	67
4.4.2. Bağlılılık	68
4.4.3. Silhouette Genişliği	68
4.4.4. Dunn İndeks.....	69
5. PANEL VERİLERDE KÜMELEME ANALİZİ	65

5.1. Tek Değişkenli Panel Veri Kümeleme Analizi	67
5.2. Çok Değişkenli Panel Veri Kümeleme Analizi	68
5.3. Karma (Nümerik ve İkili) Değişkenli Panel Veri için Kümeleme Analizi..	75
5.3.1.Önerilen Uzaklık Ölçümü.....	78
6. UYGULAMA.....	79
7. SONUÇLAR VE ÖNERİLER	113
KAYNAKLAR.....	117
ÖZGEÇMİŞ.....	131





ÇİZELGELER DİZİNİ

SAYFA

Çizelge 3.1. Panel veri çizelgesi.....	11
Çizelge 4. 1. İki sonuçlu p sayıda değişken içeren bir gözlem çiftinin (r,k) çapraz çizelgesi	45
Çizelge 4.2. İki sonuçlu p değişken içeren verinin gözlem çiftleri arasındaki uzaklık için kullanılan bazı eşleştirme tipli uzaklık ölçüleri.	46
Çizelge 4.3. Matris güncelleme algoritmaları için katsayı değerleri	53
Çizelge 5.1. Çok değişkenli panel veri yapısı	69
Çizelge 5.2. Panel veri seti	73
Çizelge 5.3. Yıllara göre ters çevrilmiş panel veri seti.....	73
Çizelge 6.1. Turizm gelirleri verisi için üç farklı uzaklık ölçümü kullanılarak hiyerarşik kümeleme algoritmalarına göre elde edilen kofenetik koralasyon katsayı değerleri.....	89
Çizelge 6.2. Turizm Geliri verisi için Gower uzaklığı kullanılarak hiyerarşik kümeleme algoritmalarına göre elde edilen Calinski-Harzbasz (1974) değerleri	89
Çizelge 6.3. Turizm Geliri verisi için Gower uzaklığı kullanılarak hiyerarşik kümeleme algoritmalarına göre elde edilen Duda-Hart (1973) değerleri.....	90
Çizelge 6.4. Turizm Geliri verisi için k-prototip uzaklığı kullanılarak hiyerarşik kümeleme algoritmalarına göre elde edilen Calinski-Harzbasz (1974) değerleri	90
Çizelge 6.5. Turizm Geliri verisi için k-prototip uzaklığı kullanılarak hiyerarşik kümeleme algoritmalarına göre elde edilen Duda-Hart (1973) değerleri.....	90
Çizelge 6.6. Turizm Geliri verisi için önerilen uzaklık kullanılarak hiyerarşik kümeleme algoritmalarına göre elde edilen Calinski-Harzbasz (1974) değerleri	91

Çizelge 6.7. Turizm Geliri verisi için önerilen uzaklık kullanılarak hiyerarşik kümeleme algoritmalarına göre elde edilen Duda-Hart (1973) değerleri.....	91
Çizelge 6.8. Turizm Geliri verisi için Gower uzaklığı kullanılarak hiyerarşik kümeleme algoritmalarına göre elde edilen geçerlilik değerleri	92
Çizelge 6.9. Turizm Geliri verisi için k-prototip uzaklığı kullanılarak hiyerarşik kümeleme algoritmalarına göre elde edilen geçerlilik değerleri.....	92
Çizelge 6.10. Turizm Geliri verisi için önerilen uzaklık kullanılarak hiyerarşik kümeleme algoritmalarına göre elde edilen geçerlilik değerleri	93
Çizelge 6.11. Gower uzaklığı ile elde edilen Küme 1'deki birimler için yatay kesit testleri	100
Çizelge 6.12. Önerilen uzaklık ile elde edilen Küme 1'deki birimler için yatay kesit testleri	100
Çizelge 6.13. Gower ve önerilen uzaklık ile elde edilen Küme 2'deki birimler için yatay kesit testleri.....	101
Çizelge 6.14. Gower uzaklığı ile elde edilen Küme 3'deki birimler için yatay kesit testleri	101
Çizelge 6.15. Önerilen uzaklık ile elde edilen Küme 3'deki birimler için yatay kesit testleri	101
Çizelge 6.16. Gower uzaklığı ile elde edilen Küme 1'deki birimler için birim kök testleri	102
Çizelge 6.17. Önerilen uzaklık ile elde edilen Küme 1'deki birimler için birim kök testleri	102
Çizelge 6.18. Gower ve önerilen uzaklık ile elde edilen Küme 2'deki birimler için birim kök testleri	103
Çizelge 6.19. Gower uzaklığı ile elde edilen Küme 3'deki birimler için birim kök testleri	103

Çizelge 6.20. Önerilen uzaklık ile elde edilen Küme 3'deki birimler için birim kök testleri	103
Çizelge 6.21. Gower uzaklığı ile elde edilen kümeler için değişen varyans ve otokorelasyon testi.....	104
Çizelge 6.22. Önerilen uzaklık ile elde edilen kümeler için değişen varyans ve otokorelasyon testi.....	105
Çizelge 6.23. Gower uzaklığı ile elde edilen Küme 1 için parametre tahminleri	105
Çizelge 6.24. Önerilen uzaklık ile elde edilen Küme 1 için parametre tahminleri	106
Çizelge 6.25. Gower ve önerilen uzaklık ile elde edilen Küme 2 için parametre tahminleri.....	107
Çizelge 6.26. Gower uzaklığı ile elde edilen Küme 3 için parametre tahminleri	107
Çizelge 6.27. Önerilen uzaklık ile elde edilen Küme 3 için parametre tahminleri	108



ŞEKİLLER DİZİNİ

SAYFA

Şekil 3.1.	Homojen eğim ve heterojen sabit durumuna örnekler (a,b,c).....	18
Şekil 3.2.	Eğim ve sabitlerin heterojen olması durumuna örnekler (a,b).....	19
Şekil 4.1.	Sınıflama tipleri ağacı.....	41
Şekil 4.2.	Tek bağlantı ve tam bağlantı kümeleme için matris güncelleme örneği	52
Şekil 4.3.	Matris güncellenme algoritmaları için dendrogram örnekleri.....	56
Şekil 5.1.	Yatay kesit veri örneği (Lu ve Huang, 2011).....	66
Şekil 5.2.	Panel veri kümeleme örneği için elde edilen dendrogram.....	75
Şekil 6.1.	Gower uzaklığı kullanılarak ortalama bağlantı yöntemi ile elde edilen dendrogram	80
Şekil 6.2.	Gower uzaklığı kullanılarak tam bağlantı yöntemi ile elde edilen dendrogram	80
Şekil 6.3.	Gower uzaklığı kullanılarak Ward yöntemi ile elde edilen dendrogram	81
Şekil 6.4.	Gower uzaklığı kullanılarak tek bağlantı yöntemi ile elde edilen dendrogram	81
Şekil 6.5.	Gower uzaklığı kullanılarak ağırlıklı ortalama bağlantı yöntemi ile elde edilen dendrogram.....	82
Şekil 6.6.	k-prototip uzaklığı kullanılarak ortalama bağlantı yöntemi ile elde edilen dendrogram.....	82
Şekil 6.7.	k-prototip uzaklığı kullanılarak tam bağlantı yöntemi ile elde edilen dendrogram	83
Şekil 6.8.	k-prototip uzaklığı kullanılarak Ward yöntemi ile elde edilen dendrogram	83
Şekil 6.9.	k-prototip uzaklığı kullanılarak tek bağlantı yöntemi ile elde edilen dendrogram	84

Şekil 6.10. k-prototip uzaklığı kullanılarak ağırlıklı ortalama bağlantı yöntemi ile elde edilen dendrogram.....	84
Şekil 6.11. Önerilen uzaklık kullanılarak ortalama bağlantı yöntemi ile elde edilen dendrogram	85
Şekil 6.12. Önerilen uzaklık kullanılarak tam bağlantı yöntemi ile elde edilen dendrogram	85
Şekil 6.13. Önerilen uzaklık kullanılarak Ward yöntemi ile elde edilen dendrogram	86
Şekil 6.14. Önerilen uzaklık kullanılarak tek bağlantı yöntemi ile elde edilen dendrogram	86
Şekil 6.15. Önerilen uzaklık kullanılarak ağırlıklı ortalama bağlantı yöntemi ile elde edilen dendrogram.....	87
Şekil 6.16. $\lambda = 0.1$ için küme sayısına göre küme içi kareler toplamı	94
Şekil 6.17. $\lambda = 0.5$ için küme sayısına göre küme içi kareler toplamı	94
Şekil 6.18. $\lambda = 0.7$ için küme sayısına göre küme içi kareler toplamı	95
Şekil 6.19. $\lambda = 1$ için küme sayısına göre küme içi kareler toplamı	95
Şekil 6.20. Gower uzaklığı kullanılarak elde edilen birinci küme için birimlere göre Turizm Gelirleri çizgi grafiği	97
Şekil 6.21. Önerilen uzaklık kullanılarak elde edilen birinci küme için birimlere göre Turizm Gelirleri çizgi grafiği	98
Şekil 6.22. Gower ve önerilen uzaklık kullanılarak elde edilen ikinci küme için birimlere göre Turizm Gelirleri çizgi grafiği	98
Şekil 6.23. Gower uzaklığı kullanılarak elde edilen üçüncü küme için birimlere göre Turizm Gelirleri çizgi grafiği	99
Şekil 6.24. Önerilen uzaklık kullanılarak elde edilen üçüncü küme için birimlere göre Turizm Gelirleri çizgi grafiği	99
Şekil 6.25. Gower uzaklığı kullanılarak elde edilen birinci küme için birimlere göre Turizm Gelirlerinin uydurulmuş doğrusuyla birlikte güven aralıkları.....	109

Şekil 6.26. Önerilen uzaklık kullanılarak elde edilen birinci küme için birimlere göre Turizm Gelirlerinin uydurulmuş doğrusuyla birlikte güven aralıkları.....	109
Şekil 6.27. Önerilen uzaklık ve Gower uzaklığı kullanılarak elde edilen üçüncü küme için birimlere göre Turizm Gelirlerinin uydurulmuş doğrusuyla birlikte güven aralıkları	110
Şekil 6.28. Gower uzaklığı kullanılarak elde edilen üçüncü küme için birimlere göre Turizm Gelirlerinin uydurulmuş doğrusuyla birlikte güven aralıkları.....	110
Şekil 6.29. Önerilen uzaklık kullanılarak elde edilen üçüncü küme için birimlere göre Turizm Gelirlerinin uydurulmuş doğrusuyla birlikte güven aralıkları.....	111



KISALTMALAR

ADF	: Augmented Dickey Fuller
ALM	: Adjusted Lagrange Multiplier
CADF	: Cross Section ADF
CH	: Calinski-Harzbasz
CPCC	: Cophenetic Correlation Coefficient
DH	: Duda-Hart
FE	: Fixed Effects
FGLS	: Feasible Generalized Least Square
GLS	: Generalized Least Square
IPS	: Im, Pesaran, Shin
LBI	: Locally Best Invariant
LLC	: Levin, Lin, Chu
LM	: Lagrange Multiplier
MI	: The Mutual Information
MSE	: Mean Square Error
OCIL	: Object Cluster Similarity
OECD	: Organisation for Economic Co-operation and Development
POLS	: Pooled OLS
PCSE	: Panel Corrected Standard Error
PMG	: Pair Group Method
RE	: Random Effects
SAHN	: Sequential, Agglomerative, Hierarchical, Nonoverlapping
UFT	: Unsupervised Feature Transformation
UPGMA	: Unweighted Pair Group Method With Averaging
UPGMC	: Unweighted Pair Group Centroid
WPGMA	: Weighted Pair Group Method with Arithmetic Mean

WPGMC : Weighted Pair Group Centroid Method



1. GİRİŞ

Panel veri ülkeler, firmalar, hane halkları vb. kesit gözlemlerin belli bir zaman dönemi içinde bir araya getirilmesidir. Panel veri setinde hem yatay kesit hem de zaman serisi boyutu bulunmaktadır. Bu nedenle panel veri ile yapılan analizler yatay kesit ve zaman serilerinin özelliklerini içerir. Önceden panel veri tarım ve astronomi alanında uygulanırken bilgisayar bilimleri ve istatistiksel yöntemler teorisinin gelişmesiyle birlikte ekonomi, sosyoloji, işletme ve psikoloji gibi sosyal bilimlerde de kullanımı yaygınlaşmıştır (Zheng ve ark, 2009).

Panel very modeli, Eşitlik (1.1)'de tanımlanmıştır.

$$y_{it} = \alpha_{it} + X'_{it}\beta_{it} + u_{it} \quad i=1,...,N \quad t=1,...,T \quad (1.1)$$

burada i , hanehalkları, bireyler, firmalar, ülkeler vb. gösterirken, t zamanı göstermektedir. Yani i yatay kesit boyutunu gösterirken t zaman serisi boyutunu göstermektedir. Modelde α_{it} sabittir, β_{it} , $k \times 1$ boyutlu parametre vektörü, X'_{it} , t zamanında i . birim için açıklayıcı değişkendir. y_{it} , t zamanında i . birimin bağımlı değişkenini göstermektedir. Panel veri modellerinde parametrelerin, her dönem ve her birim için değer almasına izin verilmektedir. Eğer eğim parametreleri birime ve zamana göre değişmiyorsa ($\beta_{it} = \beta$) birimler homojendir.

Panel veri karışık olaylar için daha doğru analizleri mümkün kılmaktadır. Panel veri analizlerinde en önemli konu, farklı birimlerin havuzlanabilirliğidir (Lu ve Huang, 2011). Ampirik çalışmaların çoğunda, panel veri modelinde ilgilenilen eğim parametresinin birimlere göre homojen olduğu varsayılır (Pesaran ve Yamagata, 2005). Panel veri analizi yapmadan önce Chow ve Wald gibi bazı testlerle havuzlanabilirlik kontrol edilmelidir. Sonuç olarak bu test başarısızsa birimler arasındaki heterojenlikten dolayı hem teorik hem de istatistiksel olarak geçerli tahminler yapılamayacaktır (Lu ve Huang, 2011). Dolayısıyla

heterojenliğin göz ardı edilmesi hatalı sonuçlara neden olacaktır. Ekonomideki çoğu çalışma, parametre heterojenliğinin varlığını ortaya çıkarmıştır. Örneğin, Hineline (2008) büyüme model katsayılarının homojenliği hipotezini reddetmiştir. Alt bölgelere ayrılmış ülkelere yeterli sayıda regresyon uygulayarak katsayıların ve önemliliklerinin anlamlı şekilde farklı olduğunu göstermiştir. Son zamanlarda araştırmacılar, zaman serileri açısından panel veriyi ele alarak eğitim heterojenliğine değinmektedirler. Örneğin Pesaran ve Smith (1995) ve Pesaran ve ark (1999) ayrı zaman serisiden kümelenmiş veya havuzlanmış ve kümlenmiş katsayıları tahmin eden ortalama grup ve havuzlanmış ortalama grup tahmin edicileri önermişlerdir. Pesaran ve ark (1999) kısa zaman periyodu için ortalama grup tahmin edicisinin ciddi bir şekilde yanlış olduğunu göstermişlerdir (Bobková, 2014).

Ekonometrideki son gelişmelerde, birimleri farklı gruplara kümeleyerek verinin kısmi havuzlanabilirliği araştırılmaktadır. Böylece, aynı gruptaki birimler bağımsız değişkenlerin etkinliği açısından homojendir (Lu, ve Huang, 2011). Sarafidis ve Weber (2009) ve Sarafidis ve Weber (2015), Chang-Ching ve Ng (2012) ve Kapetanios (2006) farklı yaklaşımlar önermişlerdir. Çalışmalarında heterojen kümelerde model tahmini yaparak kısmi havuzlanabilirliği araştırmışlardır. Tahmin edilen modelden artık kareler toplamına veya bilgi kriterine göre birimleri kümelemişlerdir. Fakat bu kümeleme türü tahmin edilen modele ve tanımlanmasına dayanır. Tahmin edilen modeldeki yanlış tanımlama hataları, kümelerde de yanlış tanımlamalara neden olacaktır (Bobková, 2014).

Bu çalışmada, panel veri setinin birimlere göre heterojen olduğu dikkate alınarak panel veriye kümeleme analizi uygulanmıştır. Bu yöntem modele değil veriye dayalı olarak birimler arasında heterojenliği dikkate alır. Kümeleme analizi ile küme içi homojen ve kümeler arası heterojen birimler elde edilir. Panel verinin değişken yapısına göre uygun uzaklık yöntemleri kullanılarak hiyerarşik kümeleme yöntemleri ve k-prototip algoritması ile kümeleme analizi yapılmıştır. Ayrıca karma değişkenli panel veri noktaları arasında yeni bir uzaklık ölçümü tanımlanmıştır. Elde edilen her bir kümedekümelenmiş panel veri seti için

parametre tahmini yapılmıştır. İlk olarak temel ekonometrik varsayımlar olan yatay kesit bağımlılığı, değişen varyans ve otokorelasyon gibi problemlerin olup olmadığı test edilmiştir. Bu varsayımlardan bir veya birkaçının sağlanmaması, tahmin edilen parametrelerde etkinlik kaybına ve standart hataların yanlış tahmin edilmesine neden olacaktır. Bu nedenle bu problemlere karşı tutarlı standart hatalar üretebilen tahmin yöntemleri kullanılmıştır.

Çalışma yedi ana bölümden oluşmaktadır. Çalışmanın ilk bölümünde çalışmanın amacı açıklanmıştır. İkinci bölümde konu ile ilgili daha önceden yapılan çalışmalar incelenmiş, konunun zaman içerisinde gelişimi literatür özeti olarak verilmiştir. Üçüncü bölümde, panel veri, panel veri modelleri ve panel veri tahmin yöntemleri kapsamlı bir şekilde açıklanmıştır. Çalışmanın dördüncü bölümünde kümeleme analizi, yaygın kullanılan uzaklık ölçümleri ve kümeleme yöntemleri açıklanmıştır. Çalışmanın beşinci bölümünde panel verilerde kümeleme analizi açıklanmış ve karma değişkenli panel veriler için yeni bir uzaklık ölçümü önerilmiştir. Altıncı bölüm uygulama bölümüdür. Karma değişkenli panel veri setine kümeleme analizi uygulanmış ve her bir küme için parametre tahminleri yapılmıştır. Analizler Stata ve R paket programları yardımıyla yapılmıştır. Çalışmanın son bölümü elde edilen bilgilerin yorumlandığı ve değerlendirildiği sonuç bölümüdür.



2. ÖNCEKİ ÇALIŞMALAR

Çalışmanın bu bölümünde panel veri analizi, panel veri kümeleme analizi ve karma değişkenli veriler için önerilen uzaklık ölçümleri ile ilgili yapılan çalışmaların literatür özeti verilecektir.

Conway ve Kniesner (1992), erkek işgücü arzı fonksiyonu tahminin hassasiyeti için En Küçük Kareler, sabit etki ve rasgele etki gibi tahmin tekniklerine alternatif seçenekler incelemişlerdir.

Kim ve Maddala (1992), 1976 ve 1987 yılları arasındaki on iki yıllık süre boyunca ABD’de oldukça büyük firmaların hisse senedi hareketlerini panel veri kullanarak analiz etmişlerdir.

İslam, Fiebig ve Neade (2002), telekomünikasyon alanında eksik panel veri için tahmin modeli ele almışlardır. Kısa, hem kısa hem de eşit olmayan zaman serili ve içinde kayıp veri olmayan farklı bir havuzlama yöntemi önermişlerdir.

Mayda (2005), 1980 ve 1995 yılları arasında, 14 merkezi OECD ülkesi içinde meydana gelen göçün ekonomik ve ekonomik olmayan değişimini incelemiştir. Göç akışında kullanılan yıllık panel veri seti, hem zaman serisi hem de yatay kesit değişiminden faydalanmayı mümkün kılmıştır. Uluslararası standart göç modelini tahmin etmiştir.

Berrington ve ark (2006), yapısal eşitlik modellerinin iki farklı tipine, sabit ve rasgele etkiye, grafik zincirine, değişim skorunun pratik uygulamalarına ve teorisine giriş sağlamayı amaçlamıştır. Panel veri analizi yapmak için bu yöntemleri kullanmıştır.

Rao ve Sabavala (1981) müşterilerin homojen grupları için, panel veride hiyerarşik kümeleme analizine dayalı yöntem önermişlerdir.

Bonzo ve Hermosilla (2002), panel veride hiyerarşik olmayan kümeleme analizinin yapısal korumave karışık hesaplamalar gibi problemleri için, durum vektörünün örnekleme dağılımını, özellikle dağılım sürecinin eğimini, bozarak geliştirilmiş sezgisel bir yöntem önermişlerdir.

Mouchart ve Rombouts (2005), kısa zaman serili veya kayıp veri bulunması gibi birkaç eksikli panel veriye dayalı tahminlerde, verinin en iyi şekilde kullanımını amaçlayan modelleme stratejisi önermişlerdir. Bu fikir, genel panel veri modeli belirlemede kümeleme yaklaşımı içermektedir.

Kapetanios (2006) bilgi kriterine dayalı bir yöntem önermiştir. Bu yöntem serinin havuzlanmış seriye ayrılmasını mümkün kılar. Havuzlanmış seride altvektörlerin ortak parametre kanıtı varken havuzlanmamış serinin böyle bir kanıtı yoktur.

Mariel ve ark (2006), Almanya otomobil endüstrisindeki otomobillerin özelliklerini gösteren değişkenler arasında çoklu iç ilişki olması durumunda, otomobilleri büyük, orta ve küçük olarak sınıflandırarak otomobil satışını etkileyen faktörleri panel veri modeliyle incelemişlerdir.

Tunay ve Silpar (2006), hiyerarşik olmayan kümeleme analizi kullanarak sektörde yer alan bankaları gruplandırıp, panel veri yöntemiyle Türkiye’de faaliyet gösteren ticari bankaların karlılık performansını tahmin etmişlerdir.

Ren ve Shi (2009), çok değişkenli panel verinin zaman boyutuna sıralı kümeleme yöntemini uygulamışlardır.

Zheng ve ark (2009) panel veri analizine kümeleme yöntemi uygulamışlardır. Bu yöntem panel veriye sahip bağımlı değişkenin heterojenlik sorununu çözmektedir.

Juárez ve Steel (2010), panel veri modelinin normal dağılıma sahip olmadığı ve aynı zamanda modelde otoregresiflik olması durumunda kümeleme yöntemini incelemişlerdir.

Lu ve Huang (2011), 18 OECD ülkesinin kümelemesini yaparken küme sayısının belirlenmesinde k-ortalamalar küme algoritmasını kullanarak panel veri modeliyle parametre tahminlerini yapmışlardır.

Thompson (2011), birimler ve zaman gibi iki boyut arasındaki eş zamanlı korelasyona karşı dirençli standart hataları hesaplamanın çok kolay olduğunu

göstermiştir. Bu yöntemler, rezidülerin ya sadece birimlere göre yada sadece zamana göre korelasyonlu olduğu durumda geçerlidir.

Lombardo ve Falcane (2011), coğrafi bölge, ekonomik performans ve suçluluk arasındaki etkileşimi daha iyi anlamak için kısa çoklu zaman serilerine kümeleme analizi uygulamışlardır. Ayrıca potansiyel suç kontrol programı kurulması için yararlı, gelişme eğilimli bir dizi gizli özellikler belirlemişlerdir.

Nie ve ark (2012), panel veri kümelemede kullanılabilen yeni uzaklık ölçümü tanımlamışlardır. Tanımlanan bu uzaklık, ağırlıklı ve ağırlıksız hesaplanabilir.

Andreica ve Andreica (2014), Romanya endüstrisindeki işsizliği etkileyen ekonomik faaliyetleri kümeleyerek 2012-2013 yıllarındaki işsizliği panel veri modeliyle tahmin etmişlerdir.

Zheng ve Li (2014), nümerik özellikli çok değişkenli panel veriyi kümelemek için yeni uzaklık ve kareler toplamı fonsiyonu tanımlamışlardır.

Bobková (2014), daha basit ve veriye dayalı kümeleme tahmin yöntemi önermiştir. Açıklayıcı değişkenin zaman ortalamalarına dayalı olarak K-ortalamalar kümelemeye göre ülkeleri ayırmayı önermiştir.

Li ve ark (2015), panel veride hiyerarşik kümeleme belirlemek için gri bölgesel analiz kullanarak dizi toplamaya dayanan geliştirilmiş gri kümeleme analizi (AGRA) önermişlerdir.

Hou ve Ai (2015), panel verilerde kümeleme analizi, panel veri modelinin özellik testi ve parametre tahminleri üzerinde çalışmışlardır. Panel veriler üzerinde bir kümeleme analizi yaparak, Çin şehir sakinlerinin tüketim ve gelir arasındaki sekiz gelir düzeyinin 2007'den 2012'ye olan ilişkisini incelemişlerdir.

Stephenson ve ark (2017), kümelenmiş panel veri için var olan yöntemler zamana göre küme boyutunu sabit varsaydığından, kümelenme boyutlarının zaman içinde değişmesine izin veren kümelenmiş panel veri seti için geçerli parametre tahmini sağlayan, konu ağırlıklı genelleştirilmiş tahmin denklemleri şemasını sunmuşlardır. Simülasyon yoluyla mevcut yöntemlerin performansını konu

ağırlıklı yaklaşımla karşılaştırmışlardır. Konu ağırlıklı yaklaşım, tüm model parametreleri için göz ardı edilemez yanlılık gösteren, mükemmel kapsamlı tek yöntemdir.

Gowda ve Diday (1991), sembolik kümelene için sembolik nesnelerin “konumu”, “yayılımı” ve “içeriği” temelli yeni bir benzemezlik ölçüsü önermişlerdir. Benzemezlik ölçüsü, benzerlik ölçüsünün sadece başka bir yönü olmadığı için yenidir.

Ichino ve Yaguchi (1994), koordinat eksenlerinin sadece nicel özelliklerle değil, aynı zamanda niteliksel ve yapısal özelliklerle de ilişkili olduğu çok boyutlu özellik uzayında basit ve kullanışlı genelleştirilmiş Minkowski metriklerini sunmuşlardır.

Gibert ve Cortés (1997), Klass tarafından kullanılan metriklerin özelliklerini ve detaylarını sunmuşlardır. Klass, karşılıklı komşu algoritmasının versiyonunu uygulayan ve adapte edilen kötü yapılandırılmış alanların sınıflandırılmasına yönelik bir kümeleme sistemidir.

He ve ark (2002), nümerik vekategorik değişkenler içeren veri için yeni bir böl ve yönet tekniğini önermişlerdir. İlk olarak, orijinal karma veri kümesi, saf kategorik veri kümesi ve salt sayısal veri kümesi olmak üzere iki alt veri kümesine ayrılmıştır. Daha sonra, farklı türde veri kümeleri için tasarlanan mevcut iyi kurulmuş kümeleme algoritmaları, ilgili kümeleri elde etmek için kullanılmıştır.

Hsu ve ark (2007), kategorik değerler arasındaki benzerliği ifade etmeyi kolaylaştıran ve aynı zamanda sayısal ve kategorik değerlerin uzaklık ölçümünü birleştiren bir uzaklık gösterim şeması "uzaklık hiyerarşisi" önermişlerdir.

Ahmad ve Dey (2007), değerlerin birlikte ortaya çıkmasına dayalı yeni maliyet fonksiyonu ve uzaklık ölçümü önermişlerdir. Ölçümler ayrıca kümelene sürecine yönelik bir değişkenin önemini de dikkate almaktadır.

Gonçalves ve ark (2008), genetik uzaklıkları tahmin etmek için çeşitli algoritmaları karşılaştırmış ve kesikli ve nümerik değişkenlerin ayırma ve ortak analizlerine dayanarak, bu koleksiyonun 40 domates kalıtımının genetik farklılığını

nicelleştirmişlerdir. Kesikli ve nümerik değişkenler arasındaki farklar ve birliktelik analizleri Mahalanobis, Cole Rodgers ve Gower uzaklıklarına dayalı olarak hesaplanmıştır.

Shih ve ark (2010), karma yani kategorik ve nümerik veri türleri üzerindeki kümeleri bulmak için yeni bir iki aşamalı kümeleme yöntemi önermişlerdir. Bu yaklaşımda, kategorik değişkenler için, ortak oluşum (co-occurrence) fikrine dayanarak aralarındaki benzerlik veya ilişkiler bulunur. Daha sonra tüm kategorik değişkenler, bu yapılandırılmış ilişkilere dayalı olarak nümerik değişkenlere dönüştürülür. Son olarak, tüm kategorik veriler sayısal hale getirildiği için, mevcut kümeleme algoritmaları veri kümesine uygulanabilir

Lim ve ark (2012), kümelenme sırasında bilgi kaybını en aza indirirken, veri kümelerinin karma öznitelik türüyle (sayısal, kategorik) kümelenmesini destekleyen bir kümelenme çerçevesi önermişlerdir.

Reddy ve Kavitha (2012), karma sayısal ve kategorik özellikleri olan veriler için iyi çalışan benzerlik ağırlığı ve filtre yöntemi paradigmasına dayalı bir kümeleme algoritması sunmuşlardır. Sayısal verilerin sadece sınırlandırılmasının üstesinden gelmek ve kümelerin daha iyi karakterizasyonunu sağlamak için küme merkezinin değiştirilmiş bir tanımını önermişlerdir.

Soundaryadevi ve Jayashree (2014), nesne kümeleme benzerliğine (OCIL) dayanan yinelemeli bir kümeleme algoritmasının, karma değişkenli öğrenci veri setine uygulanabilirliğini araştırmışlardır.

Wei ve ark (2015), sayısal olmayan özellikleri bilgi kaybı olmaksızın sayısal özelliklere dönüştürebilen, karşılıklı bilgi (MI) tabanlı eğitimsiz özellik dönüşümünü (UFT), heterojen veri kümeleme için konvansiyonel k-ortalamlar algoritması ile değerlendirmişlerdir.



3. PANEL VERİ VE PANEL VERİ MODELLERİ

Panel veri aynı birimin zaman içerisinde tekrarlı gözlemlerinden oluşan veri seti olarak tanımlanabilir (Cameron ve Trivedi, 2005; Berrington ve ark, 2006). Yatay kesit veriler ile zaman serilerinin birleşmesi sonucu oluşan panel veri hem yatay kesit hem de zaman serilerinin özelliklerini içermektedir (Gujarati, 2004).

Boylamsal (longitudinal) veri olarak da bilinen panel veri modern ekonometrik uygulamalarda, ekonomik verilerin zaman serisi ve yatay kesitsel değişimini bir arada kullanarak son yıllarda oldukça popüler olmuştur.

Kesitte yer alan gözlemlerin yıllar itibarıyla tekrarı söz konusu olduğu için panel veri analizinin temelinde tekrarlı varyans analizi ile varyans analizi modellerinin bulunduğu söylenebilir.

Günümüzde ekonomi alanında yıllara göre farklı ülkelerin iktisadi göstergeleri veya sağlık alanında aynı türden farklı ilaçların zaman içerisinde vücutta meydana getirdiği değişimler incelendiğinde kullanılacak veri seti panel veri yapısına uymaktadır (Verbeke ve Molenberghs, 2000). Panel veri setine ait Çizelge genel olarak aşağıdaki biçimde gösterilir.

Çizelge3.1. Panel veri çizelgesi

i	t	Y_{it}	X_{kit}		
1	1	Y_{11}	X_{111}		X_{q11}
.	.	.	.	...	.
1	T	Y_{1T}	X_{11T}		X_{q1T}
.	.	.	.	...	.
.	.	.	.	...	.
.	.	.	.	...	.
N	1	Y_{N1}	X_{1N1}		X_{qN1}
.	.	.	.	...	.
N	T	Y_{NT}	X_{1NT}		X_{qNT}

Çizelgede, $i=1,...,N$ yatay kesit boyutudur. Yani, i . bireye ait gözlem değerini gösterir. $t=1,...,T$ zaman boyutudur ve t . zamandaki gözlem değeridir. $k=1,...,q$ bağımsız değişkenlerin indisidir ve q adet açıklayıcı değişkenin varlığını belirtir. Y_{it} bağımlı değişken X_{kit} bağımsız değişkenlerdir.

Yatay kesitte yer alan her bir birim için tüm zamanlara ait gözlemlerin sayısı aynı ise dengeli panel veri, değilse dengesiz panel veri olarak adlandırılmaktadır. Baltagi (2005) ve Gujarati (2004) panel veri yönteminin üstünlüklerini şu şekilde sıralamaktadırlar:

- Panel veri, zaman boyunca bireyler, firmalar, ülkeler vd. ile ilgili olduklarından bu birimlerde bir heterojenliğin var olması kesin gibidir. Panel veri tahmin teknikleri, açık bir şekilde bu tür heterojenlikleri kesite özgü bazı değişkenlere izin vererek hesaba katabilmektedir.
- Panel veri yöntemi kesit ve zaman serisi gözlemlerini birleştirdiğinden daha fazla veri, daha çok çeşitlilik, değişkenler arasında daha az bağıllık, daha fazla serbestlik derecesi ve daha fazla etkinlik vardır (Baltagi, 2005).
- Çoklu doğrusal bağlantı daha çok zaman serisi ile yapılan çalışmalarda görülebilmektedir. Zaman serileri yerine panel veri kullanıldığında iki boyut söz konusu olduğundan bağımsız değişkenlerin aralarında kuvvetli ilişki olma olasılığı daha az olacaktır. Bu nedenle çoklu doğrusal bağlantı olma olasılığında azalacaktır (Güriş, 2015).
- Kısa zaman serisi ve/veya yetersiz kesit gözleminin var olduğu durumlarda da ekonometrik analiz yapılmasına olanak sağlar.
- Panel veri, tekrarlı yatay kesit gözlemlerini incelediğinden değişim dinamiklerini incelemede daha uygundur.
- Panel veri, sadece yatay kesit veya sadece zaman serisi verilerinde gözlemlenemeyen etkileri daha iyi ölçebilir ve tespit edebilir (Gujarati, 2004).

Sabit, eğim katsayısı ve hata terimi hakkında yapılan varsayımlara bağlı olarak panel veri regresyonu değişik şekillerde modellenenebilir. Bunlar,

1. Eğim katsayılarının ve sabit katsayının birime ve zamana göre farklılık gösterdiği modellerdir.
2. Eğim parametrelerinin ve sabit katsayının birimden birime farklılık gösterdiği modeldir.
3. Eğim katsayılarının birime ve zamana göre meydana gelen farklılıklardan etkilenmediği ve sabit katsayının birim ve zamana göre değişmediği model klasik model olarak adlandırılmaktadır.
4. Eğim parametrelerinin sabit olduğu ve sabit katsayının birimlere göre farklılık gösterdiği model tek yönlü sabit etki modeli olarak bilinmektedir.
5. Eğim katsayılarının birime ve zamana göre meydana gelen farklılıklardan etkilenmediği ancak sabit katsayının birim ve zamana göre değiştiği model iki yönlü sabit etki modelidir.

Panel veri modelleri sabit ve eğim katsayılarının sabit olduğu klasik model, sabit ve/veya eğim katsayılarının değişken olduğu sabit ve rasgele etki modelleri olmak üzere üç şekilde ele alınır.

3.1. Klasik Etki Modeli

Klasik modelde, hem sabit hem de eğim parametrelerinin birimlere ve zamana göre değişmediği yani bütün gözlemlerin homojen olduğu varsayılmaktadır. Klasik etki panel veri modeli,

$$Y_{it} = \alpha + X'_{it}\beta + u_{it} \quad i=1,...,N \quad t=1,...,T \quad k=1,...,q \quad (3.1)$$

şeklinde (Cameron ve Trivedi, 2005). Modelde $u_{it} \sim iid(0, \sigma_u^2)$ olduğu varsayılır. Her birim için gözlemler korelasyonsuz ve birime ve zamana karşı

hatalar homoskedastiktir. Bu modelde Havuzlanmış En Küçük Kareler tahmin edicisi ile parametre tahmini yapılır.

3.2. Sabit Etki Modelleri

Sabit katsayısının birimlere, zamana veya hem birimlere hem de zamana göre değiştiğinin varsayıldığı modellerdir. Sabit etki modelleri birim etkisine sahip sabit etki modeli, zaman etkisine sahip sabit etki modeli olmak üzere tek yönlü sabit etki modelleridir.

Hem birim hem de zaman etkinse sahip sabit etki modeli de iki yönlü sabit etki modeli olarak adlandırılır.

3.2.1. Birim Etkisine Sahip Sabit Etki Modeli

Bu modelde eğim parametrelerinin hem birime hem de zamana göre değişmediği ancak sabit katsayısının birime göre farklılık gösterdiği varsayılır.

$$Y_{it} = \alpha_i + X'_{it}\beta + u_{it} \quad i=1,\dots,N \quad t=1,\dots,T \quad k=1,\dots,q \quad (3.2)$$

α_i , her birime göre farklı olan sabit parametredir. β_k , eğim parametresi ve $u_{it} \sim iid(0, \sigma_u^2)$ dağılımına sahip hata terimidir. u_{it} 'nin X_{kit} ile korelasyonsuz olduğu ve sıfır ortalama ve σ_u^2 varyansı ile bağımsız özdeş dağılan rasgele bir değişken olduğu varsayımı yapılmaktadır. u_{it} (hata terimi), göz ardı edilmiş değişkenlerin hem birimlere hem de zamana özgü etkilerini ortaya koymaktadır.

Sabit etkiyi modele sokmanın yöntemi her bir yatay kesit için bir gösterge değişkenin kullanılmasıdır. i . birimin her bir gözlemi için 1 diğer birimler için 0 değerini alan gösterge değişkenler kullanılarak bireylerin gözlenemeyen etkilerinin modele katılmasını sağlayan bu yönteme “Gösterge Değişkenli En Küçük Kareler” tahmincisi denir.

Birimler arasındaki farklılık modele açıklayıcı değişken olarak dahil edilen gösterge değişkenler yardımıyla ele alınabilir.

$$Y_{it} = \alpha_1 D_{1t} + \alpha_2 D_{2t} + \dots + \alpha_N D_{NT} + \beta_1 X_{1it} + \beta_2 X_{2it} + \dots + \beta_k X_{kit} + u_{it} \\ i=1,\dots,N \quad t=1,\dots,T \quad k=1,\dots,q \quad (3.3)$$

D , $NT \times N$ boyutlu birim gösterge değişken matrisidir ve $D_N = I_N \otimes I_T$ olarak tanımlanır (Stock and Watson, 2007).

3.2.2. Zaman Etkisine Sahip Sabit Etki Modeli

Modelde eğim parametrelerinin hem birime hem de zamana göre değişmediği ancak sabit katsayısının zamana göre farklılık gösterdiği varsayılır.

$$Y_{it} = \lambda_t + X'_{it}\beta + u_{it} \quad i=1,\dots,N \quad t=1,\dots,T \quad k=1,\dots,q \quad (3.4)$$

λ_t , zamana göre farklılık gösteren sabit parametredir.

Zaman etkisine sahip sabit etki modelinde zaman dönemleri arasında farklılık yaratmak amacıyla modele T sayıda zaman gösterge değişkeni eklenir.

$$Y_{it} = \lambda_1 D_{i1} + \lambda_2 D_{i2} + \dots + \lambda_N D_{iT} + \beta_1 X_{1it} + \beta_2 X_{2it} + \dots + \beta_k X_{kit} + u_{it} \\ i=1,\dots,N \quad t=1,\dots,T \quad k=1,\dots,q \quad (3.5)$$

D , $NT \times T$ boyutlu zaman gösterge değişken matrisidir ve $D_T = I_N \otimes I_T$ olarak tanımlanır (Stock ve Watson, 2007).

3.2.3. Hem Birim hem de Zaman Etkisine Sahip Sabit Etki Modeli (İki Yönlü Sabit Etki Modeli)

Hem birim hem de zamana göre meydana gelen farklılıklar birim gösterge değişkeni D_N ve zaman gösterge değişkeni D_T ile modele dahil edilir. Ancak her iki değişkenin de modelde yer alması belirlenme problemine yol açar. $[D_N D_T]$ matrisinin sütunları tam doğrusal bağlantılı olduğu için N sayıda α ve T sayıda λ belirlenemez (Matyas ve Sevestre, 1996). Tahminin gerçekleştirilebilmesi için biri zamana diğeri de birimlere ilişkin iki değişken ihmal edilir (Fomby ve ark, 1984).

Sabit katsayısı değişken modellerin temel varsayımı, modelde yer alamayan ve gözlenmeyen değişkenlerin etkilerinin modelde, sabit terim ya da hata terimi ile ifade edilmesidir. Sabit etki modelleri, Gösterge Değişkenli En Küçük Kareler, Kovaryans (grup içi) ve Genelleştirilmiş En Küçük Kareler yöntemleri ile tahmin edilebilir.

3.3. Rasgele Etki Modelleri

Rasgele etki modellerinde birime, zamana, hem birim hem de zamana göre meydana gelen değişiklikler rasgele değişken olarak düşünülmektedir ve modele hata teriminin bir bileşeni olarak dahil edilmektedir (Pazarlıoğlu ve Gürler, 2007). Bu nedenle modele hata bileşenler modeli de denilmektedir. Sabit etkiler modelinde tahmin edilmesi gereken çok sayıda parametre varken, eğer model rasgele etki modeli içeriyorsa tahmin edilecek parametre sayısı azalmakta ve serbestlik derecesi kaybı önlenebilmektedir.

$$Y_{it} = \alpha + X'_{it}\beta + u_{it} \quad i=1,...,N \quad t=1,...,T \quad k=1,...,q \quad (3.6)$$

olarak ifade edilen modelde hata terimi u_{it} , hata bileşenlerine göre üç şekilde tanımlanabilir.

$$u_{it} = \mu_i + v_{it} \quad (3.7)$$

$$u_{it} = \lambda_t + v_{it} \quad (3.8)$$

$$u_{it} = \mu_i + \lambda_t + v_{it} \quad (3.9)$$

Eşitlikteki $\mu_i \sim iid(0, \sigma_u^2)$ gözlenemeyen birim etkisini, $\lambda_t \sim iid(0, \sigma_\lambda^2)$ gözlenemeyen zaman etkisini ve $v_{it} \sim iid(0, \sigma_v^2)$ birleşik hata birimini göstermektedir. (3.7) ve (3.8) eşitliği tek yönlü rasgele etki modelini gösterirken, (3.9) eşitliği de iki yönlü rasgele etki (hem birim hem de zaman etkisine sahip rasgele etki) modelini göstermektedir. Tüm i ve t 'ler için X_{it} 'ler μ_i , λ_t , ve v_{it} 'den bağımsızdır (Baltagi, 2005).

3.4. Heterojenlik

Heterojenlik çoğu panel veri setinin yaygın bir özelliğidir (Frees, 2004). Panel veri analizinde kullanılan bireyler, firmalar, şehirler gibi birimler genelde heterojendir. Heterojenliğin neredeyse tamamıyla göz ardı edilmesi, hatalı sonuçlara neden olmaktadır.

Belirli bir model ile çalışırken, birimlere özgü önemli etkenlerin önemsenmeyip ihmal edilmesi sonucu, bütün gözlemlerin her birim ve her zaman dönemi için aynı parametrik dağılımdan türediği varsayımı gerçekçi olmamaktadır. Bireysel özelliklerden kaynaklanan farklı kitle parametrelerinin (sabit ya da eğim parametreleri) bu durumuna “parametre heterojenliği” denir.

$$Y_{it} = \alpha_i + \beta_i' X_{kit} + u_{it} \quad i=1, \dots, N \quad t=1, \dots, T \quad (3.10)$$

Eşitlik (3.10)'de, parametrelerin zamandan bağımsız ancak birimden birime değiştiği varsayılmıştır. Bu model yerine parametre heterojenliğinin ihmal

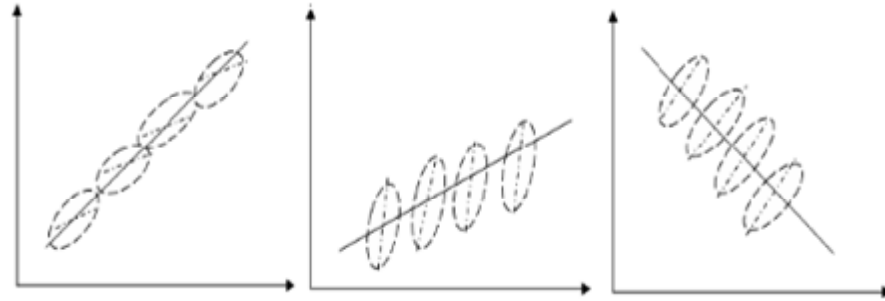
edildiği (3.11) eşitliğindeki modelin tahmin edilmesi ile ortaya çıkabilecek olası sonuçlar Şekil 3.4.1(a,b,c)'de gösterilmiştir.

$$Y_{it} = \alpha + \beta' X_{kit} + u_{it} \quad i=1,\dots,N \quad t=1,\dots,T \quad (3.11)$$

Bu sonuçların yorumunu iki şekilde yapmak mümkündür. Gerçekte sabit terimlerin heterojen ve eğimlerin homojen olması, ele alınan ilk durumdur. İkinci durumda ise, hem sabit hem de eğim parametrelerinde heterojenlik söz konusudur.

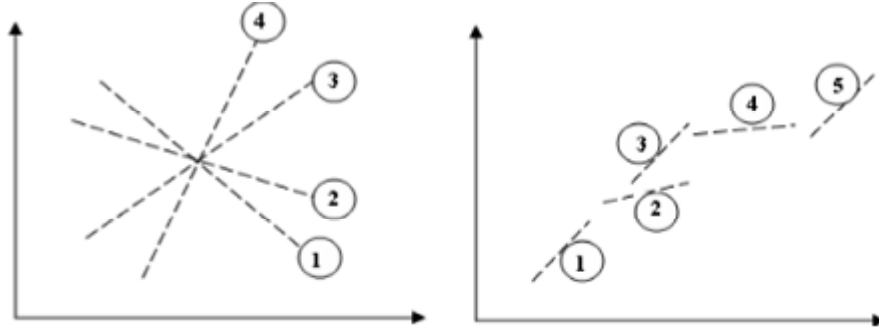
$\alpha_i \neq \alpha_j$ ve $\beta_i = \beta_j$, $i \neq j$ olması durumunda;

Şekil 3.4.1'de görülen kesikli elipsler her birim için ayrı ayrı zamansal saçılımları, kesikli düz çizgiler bireysel doğrusal regresyonları ve sürekli düz çizgiler ise birleştirilmiş regresyon doğrusunu göstermektedir. Burada dikkati çeken, bireysel eğimlerin çok farklı yönlerde karşımıza çıkabileceğidir. Bu durumda kesinlikle (3.13) eşitliği kullanılmalıdır.



Şekil 3.1. Homojen eğim ve heterojen sabit durumuna örnekler (a,b,c)

$\alpha_i \neq \alpha_j$ ve $\beta_i \neq \beta_j$, $i \neq j$ olması durumuna ilişkin olası sonuçlar Şekil 3.4.2 (a,b)'de gösterilmiştir.



Şekil 3.2. Eğim ve sabitlerin heterojen olması durumuna örnekler (a,b)

Burada serpilme diyagramları çizilmemiştir. Yuvarlak çizgiler içerisindeki sayılar, ilgili regresyon doğrusunun ait olduğu birim numarasını göstermektedir. Burada her bir birim için parametrelerdeki heterojenliği yok sayan, tüm $N \times T$ sayıda gözlemi birleştirerek uygulanan regresyon, gerçeği yansıtmayacaktır. Aslında bu tarz bir yaklaşımda, temsilci regresyon çizgisinin eğrisel olması beklenir (Hsiao, 2003).

Panel verilerde genellikle verilerin birleşimini gösteren havuzlanmış en küçük kareler (POLS), sabit etki modeli (FE) ve rassal etki modeli (RE) kullanılmaktadır. Her üç model yapısı da temelde yatay kesit bağımlılığı, otokorelasyon ve değişen varyans problemlerinin olmadığı varsayımlarına dayanmaktadır. Panel veri modelinde bu varsayımlardan bir ya da birkaçının sağlanmaması tahmin edilen parametrelerde etkinlik kaybına ve standart hataların yanlış tahmin edilmesine neden olmaktadır. Dolayısıyla bu varsayımların geçerli olup olmadığının test edilmesi gerekmektedir.

3.5. Panel Veri Modellerinde Değişen Varyans Testi

Panel veri modelleri ile ilgili varsayımlardan biri birimler arası hata terimlerinin varyanslarının özdeş yani sabit olduğudur. Fakat uygulamadaki birçok panel veri çalışması farklı yapılarda yatay kesit birimleri içermektedir. Değişen varyans sorunu varsa ve göz ardı edilerek tahmin yapılırsa katsayılar tutarlı olmaya

devam edecek ancak bu tahminler etkin olmayacaktır. Ayrıca tahmin sonucunda elde edilen standart hatalar sapmalı olacaktır. Sabit etki modelinde değişen varyans sorunu doğrudan panel veri modelinin hata teriminden kaynaklanır. Birimlere göre değişen varyans söz konusudur. Sabit etkili panel veri modellerinde değişen varyans sorunun olup olmadığını incelenmesinde Wald testi kullanılmaktadır. Rassal etki modellerinde ise değişen varyans sorunu, hata bileşenlerinin birinin ya da her ikisinin birden varyansının panel birimler için değişmesinden kaynaklanır. Rassal etki modelinin hata terimlerinin, eş varyans varsayımına uyup uymadığını test etmek için kullanılabilecek testler Levene (1960), Brown ve Forsythe (1974) testleridir.

3.5.1. Değiştirilmiş Wald Testi

Hata süreci, yatay kesit birimleri içinde homoskedastik olabilir, ancak varyansı birimden birime farklılık gösterebilir. Bu, grupsal heteroskedasite (değişen varyans) olarak bilinen bir durumdur. Değiştirilmiş Wald istatistiği, sabit etki panel veri modelinin artıklarında değişen varyans olup olmadığını “varyanslar birimlere göre homoskedastiktir” ($H_0 = \sigma_i^2 = \sigma^2 \ i=1,..., N$) yokluk hipotezini sınamak için kullanılmaktadır. i . yatay kesit biriminin hata varyansının tahmin edicisi $\hat{\sigma}_i^2 = T_i^{-1} \sum_{t=1}^{T_i} e_{it}^2$ ’dir ve $\hat{\sigma}_i^2$ ’nin varyansının tahmini,

$$V_i = T_i^{-1}(T_i - 1)^{-1} \sum_{t=1}^{T_i} (e_{it}^2 - \hat{\sigma}_i^2)$$

olarak tanımlanır. Değiştirilmiş Wald test istatistiği,

$$W = \sum_{i=1}^N \frac{(\hat{\sigma}_i^2 - \hat{\sigma}^2)}{V_i} \quad (3.12)$$

şeklindedir ve N serbestlik dereceli χ^2 dağılımına uyacak şekilde tanımlanır (Baum, 2001).

3.5.2. Levene, Brown ve Forsythe Testi

Levene (1960), normal dağılım varsayımının gerçekleşmediği durumdada dirençli bir heteroskedasite testi önermiştir. X_{ij} 'i. gruptaki X 'in j . gözlemi olsun ve $Z_{ij} = |X_{ij} - \bar{X}_i|$ olmak üzere Levene testi istatistiği,

$$W_0 = \frac{\sum_i n_i (\bar{Z}_i - \bar{Z})^2 / (g-1)}{\sum_i \sum_j (Z_{ij} - \bar{Z})^2 / \sum_i (n_i - 1)} \quad (3.13)$$

şeklinde hesaplanmaktadır. n_i , gözlem sayısı ve g , birim sayısıdır. $\bar{Z}_i = \sum \frac{Z_{ij}}{n_i}$ ve $Z = \sum \sum \frac{Z_{ij}}{n_i}$ 'dir. W_0 kritik değeri, $(g-1)$ ve $\sum_i (n_i - 1)$ serbestlik dereceli Snedecor F tablosundan elde edilir. Brown ve Forsythe (1974) tarafından önerilen iki test istatistiğinin ilkinde (W_{50}) $\bar{X}_i$ yerine X_{ij} 'nin i . birim medyanı, ikincisinde (W_{10}) $\bar{X}_i$ yerine X_{ij} 'nini. birim %10 kırılmış ortalaması yer almaktadır (Datong ve ark, 2017).

3.6. Panel Veri Modellerinde Otokorelasyon Testi

Otokorelasyon sorununun göz ardı edilmesi durumunda yapılacak tahminlerde katsayı sonuçları hala tutarlı olacak ancak etkin olmayacaktır. Bununla birlikte elde edilen standart hatalarda sapmalı olacaktır. Bhargava, Franzini ve Narendranathan DW-d testi ve Baltagi-Wu LBI testleri hem sabit hem de rassal etki modellerinde kullanılabilir. Baltagi ve Wu (1999), yerel en iyi değişmez (LBI) test istatistiğini geliştirmişlerdir. Test dengesiz panel veri modelleri için de kullanılabilir. Rassal etki modellerinde, hata bileşenini oluşturan büyüklüklerde korelasyon çok sık rastlanan bir problemdir. Tahmin sürecinin ardından rasgele etki modelinde otokorelasyon durumunun sınanması

gerekmektedir. Otokorelasyon olması durumunda, tahmin ediciler etkinliğini yitirmekte ve standart hatalar sapmasızlık özelliğini kaybetmektedir. Rassal etkili panel veri modellerinde otokorelasyon olmaması varsayımının sağlanıp sağlanamaması konusunda uygulanabilecek testler Lagrange Çarpanı (LM) testi ve Düzeltilmiş Lagrange Çarpanı (ALM) testleridir.

3.6.1. Baltagi Wu LBI Testi

Baltagi ve Wu (1999)'un yerel en iyi değişmez (LBI) testini, $H_0: \rho = 0$ yokluk hipotezine karşı $H_a: \rho > 0$ ve ya $H_a: \rho < 0$ alternatif hipotezi için önermişlerdir. Panel veri modeli,

$$y_{it} = X_{it}\beta + u_{it}, \quad u_{it} = v_{it} + \mu_i \quad (3.14)$$

şeklinde iken; matris formu ile aşağıdaki gibi yazılır.

$$y = X\beta + \text{diag}(l_{n_i})\mu + v \quad (3.15)$$

Burada l_{n_i} , n_i boyutunun birim vektörüdür. $v_{it} = \rho v_{it-1} + \varepsilon_{it}$ olmak üzere,

$$\Sigma_v(\rho) = E(vv') = \sigma_\varepsilon^2 \text{diag}(V_i) = \sigma_\varepsilon^2 \Omega_v(\rho) \quad (3.16)$$

eşitliği yazılır. n_i boyutunun ortagonal matrisi O_{n_i} olmak üzere, $O_{n_i} = (\frac{l_{n_i}}{\sqrt{n_i}}, B_i)'$ dir. Burada $B_i, B_i' l_{n_i} = 0, \quad B_i' B_i = l_{n_i-1}$ ve $B_i B_i' = l_{n_i} - \bar{J}_{n_i}$ özelliklerini sağlayan $n_i(n_i - 1)$ boyutlu bir matristir. J_{n_i} , n_i boyutunun birim matrisi olmak üzere $\bar{J}_{n_i} = J_{n_i}/n_i$ eşitliği vardır. Yukarıdaki eşitlik $\text{diag}(B_i')$ ile çarpılırsa,

$$\text{diag}(B_i')y = X\beta + \text{diag}(B_i')v \quad (3.17)$$

elde edilmektedir. Birim etki, $B_i'1_{n_i} = 0$ olması nedeniyle, modelden düşmüştür. $\tilde{y} = \text{diag}(B_i')y$, $\tilde{X} = \text{diag}(B_i')X$ ve $\tilde{v} = \text{diag}(B_i')v$ alınırsa dönüştürülmüş model, kısaca aşağıdaki gibi yazılır.

$$\tilde{y} = \tilde{X}\beta + \tilde{v} \quad (3.18)$$

Yapılan işlemler grup içi dönüşümü, yukarıdaki model ise grup içi tahminciyi ifade etmektedir. LBI testi için d istatistiği,

$$d = \frac{z'A_0z}{z'z} \quad (3.19)$$

şeklindedir. Burada, $A_0 = \frac{\partial \Omega_{ii}^{-1}(\rho)}{\partial \rho} \Big|_{\rho=0} = -\frac{\partial \Omega_{ii}(\rho)}{\partial \rho} \Big|_{\rho=0}$ ve $z = \tilde{P}_{\tilde{X}}\tilde{Y}$ 'dir (Baltagi ve Wu, 1999).

3.6.2. Değiştirilmiş Durbin Watson Testi

Bhargava ve ark (1982), panel veri modelleri için, yokluk hipotezinin hata teriminde otokorelasyon olmadığını ($H_0: \rho = 0$), alternatif hipotezde ise birinci mertebe otokorelasyon olduğunu ($H_0: |\rho| < 1$) ortaya koyan bir test geliştirmişlerdir. Test istatistiği panel veriye uyarlanmış bir Durbin Watson istatistiğinin hesaplanması ile elde edilmiştir.

$$DW = \frac{\sum_{i=1}^N \sum_{t=2}^T (\hat{v}_{it} - \hat{v}_{it-1})^2}{\sum_{i=1}^N \sum_{t=1}^T \hat{v}_{it}^2} \quad (3.20)$$

Burada $\hat{v}_{it} = y_{it} - x'_{it}\hat{\beta} - \hat{v}_i$ 'dir. $\hat{\beta}$ ve $\hat{v}_i$ sırasıyla β ve v 'nin en küçük kareler gösterge değişkeni (veya grup içi) tahminleridir (Baltagi, 2005: Born ve Breitung, 2016).

Bu iki test için söz konusu model öncelikle AR(1) kalıntıları kullanılarak sabit etkiler varsayımı ile tahmin edilmektedir. Her iki otokorelasyon test istatistiğinin değeri 2'den küçük ise birinci dereceden otokorelasyonun olmadığını belirten yokluk hipotezi reddedilmektedir (Arı ve Özcan, 2012).

3.7. Panel Veri Modellerinde Yatay Kesit Bağımlılığı Testi

Panel veri modeli analizlerinde seriye belli bir şok geldiğinde panel veride yer alan tüm yatay kesit birimlerinin ilgili şoktan aynı derecede etkilenip etkilenmediği araştırılmalıdır. Bu çalışma yatay kesit bağımlılığı olarak bilinmektedir. Kalıntılar arasında sıfır korelasyon beklenmemekte, mevcut korelasyonun anlamlılığı istatistiki testlerle araştırılmaktadır. Sabit etkili panel veri modellerinde yatay kesit bağımlılığının araştırılmasında Breusch ve Pagan (1980) tarafından önerilen Lagrange Çarpanı (LM) testi kullanılır. Bu test panel veri modelinde $T > N$ durumunda sabit etkili modelin yatay kesit bağımlılığını araştırmak için uygundur. Pesaran (2004) hem sabit hem de rassal etkili modellerde kullanılabilecek CD_{LM} olarak adlandırılan test istatistiğini türetmiştir. Bu test $N > T$ durumunda kullanılmaktadır. Panel veri modelinde yatay kesit bağımlılığının araştırılmasında hem sabit hem de rasgele etki modellerinde kullanılabilecek bir başka test ise Friedman (1937) R testidir. Panel boyutu için $T < N$ olması durumunda kullanılabilecek testlerden biridir. Frees Q testi ise Friedman test istatistiğinin hesaplanmasında kullanılan sıra korelasyon değerinin kare büyüklüklerine dayanan bir testdir (Güriş, 2015).

3.7.1. Breusch Pagan Lagrange Çarpanı (LM)

Breusch ve Pagan (1980) tarafından önerilen LM istatistiğien yaygın bilinen yatay kesit bağımlılığı testidir. Bu test N sabit T sonsuzken geçerlidir ve LM istatistiği;

$$LM = T \sum_{i=1}^{N-1} \sum_{j=i+1}^N \hat{p}_{ij}^2 \quad (3.21)$$

olarak ifade edilir. Burada $\hat{p}_{ij}$ artıkların ikili korelasyonlarının örnek tahminidir.

$$\hat{p}_{ij} = \hat{p}_{ji} = \frac{\sum_{t=1}^T \hat{u}_{it} \hat{u}_{jt}}{\sqrt{(\sum_{t=1}^T \hat{u}_{it}^2)(\sum_{t=1}^T \hat{u}_{jt}^2)}} \quad (3.22)$$

burada $\hat{u}_{it}$, u_{it} 'nin tahminidir. LM, ilgili yokluk hipotezi altında $N(N-1)/2$ serbestlik dereceli asimtotik χ^2 dağılımına sahiptir (De Hoyos ve Sarafidis, 2006). Bu tür testlerin temel fikri, yokluk hipotezi altında sınırlı modelden elde edilen parametre tahminlerini skor vektörü ile değiştirmek ve sıfır vektörünün sıfıra yeterince yakın olup olmadığını kontrol etmektir. Boşluk altında test istatistiği bireysel spesifik EKK regresyonlarının kalıntılarına dayanabileceğini ortaya çıkarmaktadır (Sarafidis ve Wansbeek, 2010). Bununla birlikte, bu test, uygulamalarda sıklıkla karşılaşılan bir durum olan N büyük ve T sonlu olduğunda, önemli ölçüde boyut bozulmaları gösterecektir. Çünkü LM istatistiği, sonlu T için doğru bir şekilde merkezlenemez ve N büyük iken yanlılık muhtemelen artar (De Hoyos ve Sarafidis, 2006).

3.7.2. Friedman Testi

Friedman (1937) uygun büyüklükte N ve sabit T durumunda, Spearman'ın rank korelasyon katsayısına dayalı parametrik olmayan bir test önermiştir (Sarafidis ve Wansbeek, 2010). Katsayı, düzenli ürün-moment korelasyon katsayısı olarak, yani Spearman'ın rank korelasyon katsayısının ranklardan hesaplanması dışında, hesaplanan değişkenlik oranı bakımından düşünülebilir. Özellikle $\{r_{i,1}, \dots, r_{i,T}\}$, $\{u_{i,1}, \dots, u_{i,T}\}$ 'nin rankı olarak tanımlanırsa, ortalama rank $(T + \frac{1}{2})$ 'dir ve Spearman'ın rank korelasyon katsayısı,

$$r_{ij} = r_{ji} = \frac{\sum_{t=1}^T \{r_{i,t-(T+\frac{1}{2})}\} \{r_{j,t-(T+\frac{1}{2})}\}}{\sum_{t=1}^T \{r_{i,t-(T+1/2)}\}^2} \quad (3.23)$$

şeklindedir. Friedman'ın istatistiği ortalama Spearman'nın katsayısına dayanır ve

$$R_{ave} = \frac{2}{N(N-1)} \sum_{i=1}^{N-1} \sum_{j=i+1}^N \hat{r}_{ij} \quad (3.24)$$

olarak hesaplanır (De Hoyos ve Sarafidis, 2006).

3.7.3. Frees Q Testi

Frees (1995, 2004) rank korelasyon katsayısının kareleri toplamına dayalı bir istatistik

$$R_{ave}^2 = \frac{2}{N(N-1)} \sum_{i=1}^{N-1} \sum_{j=i+1}^N \hat{r}_{ij} \quad (3.25)$$

olarak önermiştir.

Frees (1995) $FRE = N[R_{ave}^2 - (T-1)^{-1}]$ 'in T sabit N sonsuza giderken asimtotik Q dağılımlı olduğunu göstermiştir (Sarafidis ve Wansbeek, 2010).

$$Q = a(T)\{x_{1,T-1}^2 - (T-1)\} + b(T)\{x_{2,\frac{T(T-3)}{2}}^2 - \frac{T(T-3)}{2}\} \quad (3.26)$$

Burada, $x_{1,T-1}^2$ ve $x_{2,\frac{T(T-3)}{2}}^2$, sırasıyla $(t-1)$ ve $T(T-3)/2$ serbestlik dereceli bağımsız χ^2 rasgele değişkenleridir. $a(T) = 4(T+2)/\{5(T-1)^2(T+1)\}$ ve $b(T) = 2(5T+6)/\{5T(T-1)(T+1)\}$ 'dir. (De Hoyos ve Sarafidis, 2006). R_{ave}^2 , $(T-1)^{-1} + Q_q/N$ 'den büyük ise burada Q_q , Q dağılımının uygun kuantilidir (Sarafidis ve Wansbeek, 2010).

Q dağılımı T 'nin boyutuna bağlıdır ve iki χ^2 dağılımlı rasgele değişkenin

(ağırlıklandırılmış) toplamıdır. Bu nedenle, uygun kuantillerin hesaplanması sorunlu olabilir. T 'nin küçük olmadığı durumlarda Q 'nun varyansını hesaplayarak Q dağılımına normallik yaklaşımı önerilmiştir. Aşağıdaki sonuç kullanılmıştır.

$$\frac{FRE}{\sqrt{Var(Q)}} \approx N(0,1) \quad (3.27)$$

burada,

$$Var(Q) = \frac{32}{25} \frac{(T+2)^2}{(T-1)^3(T+1)^2} + \frac{4}{5} \frac{(5T+6)^2(T-3)}{T(T-1)^2(T+1)^2}$$

şeklindedir (De Hoyos ve Sarafidis, 2006).

3.8. Panel Birim Kök Testleri

Panel verilerde hem birimlere hem de zamana ait bilgiler analize katılabilmektedir. Panel verideki zaman etkisinin incelenmesiyle panel veriyi yaratan sürecin durağan olup olmadığı belirlenebilecektir (Güriş, 2015). Ortalaması ve varyansı zaman içinde değişmeyen ve iki dönem arasındaki varyansı bu ortak varyansın hesaplandığı döneme değilde yalnızca iki dönem arasındaki uzaklığa bağlı olan olasılıklı bir süreç için durağandır denir (Gujarati, 2004). Bir serinin durağan olması için seriyi yaratan sürecin,

$$\begin{aligned} E(Y_t) &= \mu \\ Var(Y_t) &= \sigma^2 \end{aligned}$$

şeklinde sabit bir ortalama ve varyansa sahip olması ve serinin kovaryansının zamana değil

$$cov(Y_t, Y_{t+s}) = cov(Y_t, Y_{t-s}) = \gamma_s$$

biçiminde zamanlar arası farka (s) bağlı olarak sabit olması gerekir. Bu üç özelliği taşıyan seriler zayıf durağan olarak adlandırılır. Tanımlanan özelliklere ek olarak incelenen seri normal dağılım özelliği gösteriyorsa serinin güçlü durağan olduğundan söz edilir (Güriş, 2015). Eğer seri durağansa çoğu tahmin edicinin limit dağılımı T sonsuza giderken normale yaklaşacaktır. Eğer durağan değilse ya da birim kök içeriyorsa T sonsuza giderken standart tahmin ediciler standart olmayan dağılıma sahip olacaklardır (Hsiao, 2003)

Panel birim kök testleri yatay kesit bağımlılığına bağlı olarak birinci nesil ve ikinci nesil birim kök testleri olarak ayrılmaktadır. Birinci nesil birim kök testleri, yatay kesit bağımsızlığı durumunda uygulanan testlerdir. Dickey ve Fuller (1979) ve Genişletilmiş Dickey Fuller (ADF) test yaklaşımları üzerine oturtulmuşlardır. Levin, Lin ve Chu (2002) testi, Harris ve Tzavalis (1999) testi, Breitung (2000) testi, Im, Pesaran ve Shin (1997) testi, Maddala ve Wu (1999) testi, Hadri (2000) testi, Choi (2001) testi, Levin Lin ve Chu (2002) testi, ve Im, Pesaran ve Shin (2003) birinci nesil kök testleridir. Im, Pesaran ve Shin (1997, 2003) testi, Maddala ve Wu (1999) testi ve Choi (2001) testi otoregrasif ρ 'nun birimden birime değiştiğini varsayarken diğerleri ρ 'nun değişmediğini varsayar. İkinci nesil birim kök testleri ise yatay kesit bağımlılığı durumunda kullanılan testlerdir. Bai ve Ng (2004), Moon ve Perron (2004), Phillips ve Sul (2003) ve Pesaran (2004) tarafından üretilen testler yatay kesit bağımlılığının düşük boyutlu bir faktör modeli şeklinde modellendiği birim kök testleridir. O'Connell (1998), Maddala ve Wu (1999), Taylor ve Sarno (1998) ve Chang (2002, 2004) türetilen testler kalıntıların kovaryans matrisine çok az sayıda kısıtlama konulduğu ya da hiç kısıtlama konulmadığı birim kök testleridir. Breitung ve Das (2005) ve Sul (2006) tarafından önerilen ikinci nesil birim kök testleri de vardır (Tatoğlu, 2013b). Bu bölümde Pesaran (2004) tarafından önerilen Yatay Kesit Genelleştirilmiş Dickey Fuller (CADF) testi açıklanacaktır.

Dağılım için tek faktör yapısı ve başlangıç değeri verildiğinde Y_{it} , N yatay kesit birimli ve T zaman serisi gözlemlili panel veri için Pesaran (2007) basit

dinamik heterojen doğrusal model kullanır. Model,

$$Y_{it} = (1 - \delta_i)\mu_i + \delta_i Y_{i,t-1} + u_{it} \quad i = 1, \dots, N \quad t = 1, \dots, T \quad (3.28)$$

$$u_{it} = \lambda_i f_t + e_{i,t}$$

olarak verilir ve burada f_t , gözlenmeyen genel etkiler ve $e_{i,t}$, birime özel hatalardır. Bu denklemler birlikte yazılırsa,

$$\Delta Y_{it} = \alpha_i - (1 - \delta_i)Y_{i,t-1} + \lambda_i f_t + e_{it} \quad (3.29)$$

Burada $\alpha_i = (1 - \delta_i)\mu_i$ ve $\Delta Y_{i,t} = Y_{i,t} - Y_{i,t-1}$ 'dir. $\delta_i = 0$ yokluk hipotezi (tüm yatay kesit birimleri (i) için) $\delta_i < 1$ $i = 1, \dots, N_1$ ve $\delta_i = 0$ $i = N_1 + 1, \dots, N_1$ alternatif hipotezlerine karşı test edilir. Pesaran (2007) birim kök içeren yokluk hipotezi için aşağıdaki yatay kesit genelleştirilmiş DF (CADF) regresyonunda b_i tahminini OLS'nin t -oranına bağlı olarak test etmeyi önermiştir.

$$\Delta Y_{i,t} = \alpha_i - b_i Y_{i,t-1} + c_i \bar{Y}_{t-1} + d_i \Delta \bar{Y}_t + \epsilon_{i,t} \quad (3.30)$$

burada $\bar{Y}_t = \frac{1}{N} \sum_{i=1}^N Y_{i,t}$, $\Delta \bar{Y}_t = \frac{1}{N} \sum_{i=1}^N \Delta Y_{i,t}$ ve $\epsilon_{i,t}$ ise regresyon hatasını gösterir.

Im, Pesaran ve Shin (2003) testine benzer olarak, Pesaran (2007) IPS testinin yatay kesit düzeltilmiş versiyonunu önermiştir.

$$CIPS = \frac{1}{N} \sum_{i=1}^N CADF_i \quad (3.31)$$

burada $CADF_i$, Eşitlik (3.30) CADF regresyonunda b_i 'nin t oranı tarafından verilen i . yatay kesit birimi için yatay kesit düzeltilmiş Dickey-Fuler istatistiğidir (Gengenbach ve ark, 2008).

3.9. Eş Bütünleşme Testleri

Yapılan birim kök testleri ile serinin durağan olmadığı sonucuna varılırsa, durağanlığı sağlamak için seriler üzerinde fark alma gibi işlemler yapılır. Ancak yapılan fark alma işlemi uzun dönem ilişkileri yani trendlerinde ortadan kalkmasına neden olabilmektedir. Bu nedenle fark alma işlemi yapılmadan önce eşbütünleşme analizi yapılmaktadır. Böylece ele alınan değişkenler arasında eşbütünleşme varsa bu durumda durağanlaştırma işlemine gerek kalmaz.

Panellerdeki eşbütünleşme testleri, literatürde iki farklı yönde ilerlemiştir. Birincisi, eşbütünleşmenin olmamasını yokluk hipoteziolarak alır ve Engle ve Granger (1987) statik regresyonundan üretilen artıkları kullanır. Pedroni (1995, 1997a) tarafından yapılan çalışmalar ve Kao (1999)'nun ilgili makaleleri bu yöneme dayanır. İkincisi, eşbütünleşmenin varlığını yokluk hipotezi olarak alan testlerdir. McCoskey ve Kao (1998) tarafından önerilen testin temelini oluşturmaktadır. Ayrıca Harris ve Inder (1994), Shin (1994), Leybourne ve McCabe (1994) ve Kwiatowski ve ark. (1992) testleride artık temelli testlerdir (Banerjee, 1999).

Pedroni eşbütünleşme testi gibi artık terimlere dayalı eşbütünleşme testlerinde, değişkenlerin düzey değerleri için, tahmin edilen uzun dönem katsayıları ile birinci farklar kullanılarak tahmin edilen kısa dönem hata düzeltme katsayılarının birbirine eşit olması gerekmektedir. Bu ise testlerin gücünü düşürmekte ve böylece değişkenler arasında eşbütünleşme ilişkisi olsa bile, bunun sahte biçimde reddedilmesine yol açmaktadır.

Westerlund (2007) Pedroni testlerinin eksikliğini gidermek için, hata düzeltme modeline dayalı dört panel eşbütünleşme testi geliştirmiştir. Bu testlerin ikisi grup ortalama istatistikleri, diğer ikisi ise panel istatistikleri olarak adlandırılmaktadır. Panel istatistikleri, panelin yatay kesit boyutundaki hata düzeltme ile ilgili bilginin bir araya getirilmesi ile oluşmaktadır. Grup ortalama istatistikleri ise bu bilgileri kullanmamaktadır. Testler arasındaki ayrım, alternatif hipotez testlerine dayanmaktadır. Westerlund (2007)'ın geliştirdiği bu test, paneli

oluşturan serilerin aynı derecede ve birinci farkta I(1) durağan olduğu varsayımına dayanmaktadır.

Westerlund (2007) test istatistikleri aşağıdaki süreç izlenerek hesaplanmaktadır.

$$\Delta y_{it} = \delta_i' d_t + \alpha_i (y_{it-1} - \beta_i' x_{it-1}) + \sum_{j=1}^{p_i} \alpha_{ij} \Delta y_{it-j} + \sum_{j=0}^{p_i} \gamma_{ij} \Delta x_{it-j} + e_{it}$$

Burada $d_t = (1, t)'$, $\delta_i' = (\delta_{1i}, \delta_{2i})'$ olarak tanımlanan parametre vektörü ile deterministik bileşenleri içermektedir. Test istatistiklerinin hesaplanabilmesi için asıl problem hata düzeltme parametresi α_i 'nin nasıl tahmin edileceğidir. β_i 'nin bilindiği varsayımı uygun bir yaklaşım olmayacağı için yukarıdaki denklemi yeniden yazmak gerekir.

$$\Delta y_{it} = \delta_i' d_t + \alpha_i y_{it-1} + \lambda_i' x_{it-1} + \sum_{j=1}^{p_i} \alpha_{ij} \Delta y_{it-j} + \sum_{j=0}^{p_i} \gamma_{ij} \Delta x_{it-j} + e_{it}$$

Burada $\lambda_i = \alpha_i \beta_i$ 'dir. Westerlund (2007) test istatistiklerinin dördüde α_i 'nin tahmin edilerek t-istatistiğinin tahmin edilmesine bağlıdır.

Westerlund (2007) grup ortalama istatistiklerinin hesaplanması üç aşamalı süreçle oluşmaktadır.

1. Her birim için yukarıdaki denklem tahmin edilir. Tahmin edilen model şu şekilde gösterilir.

$$\Delta y_{it} = \hat{\delta}_i' d_t + \hat{\alpha}_i y_{it-1} + \hat{\lambda}_i' x_{it-1} + \sum_{j=1}^{p_i} \hat{\alpha}_{ij} \Delta y_{it-j} + \sum_{j=0}^{p_i} \hat{\gamma}_{ij} \Delta x_{it-j} + \hat{e}_{it}$$

Burada gecikme sayısını gösteren p_i , Akaike bilgi kriteri gibi bilgi kriteri veya $\hat{\alpha}_{ij}$ ve $\hat{\gamma}_{ij}$ 'nin son gecikmelerinin anlamlı olup olmamasına göre karar verilir ve gecikme sayıları birim boyunca değişebilir. Alternatif olarak T 'nin sabit bir fonksiyonu olarak da alınabilir.

2. $\alpha_i(1) = 1 - \sum_{j=1}^{p_i} \alpha_{ij}$ 'nin tahminini içermektedir. Tahmin süreci için normalde parametrik yöntemler kullanılır ve $\tilde{\alpha}_i(1) = 1 - \sum_{j=1}^{p_i} \hat{\alpha}_{ij}$ kullanılarak $\alpha_i(1)$ tahmin edilir. Ancak, $\tilde{\alpha}_i(1)$ 'i temel alan testlerin özellikle p_i büyük olduğunda otoregresif parametrelerin tahminine özgü belirsizlikten ötürü küçük örneklemelerde zayıf performans gösterdiği bilinmektedir. Bu nedenle temel çekirdek (kernel) tahmincisi alternatif bir yöntem olarak kullanılabilir.

$$\hat{\omega}_{yi}^2 = \frac{1}{T-1} \sum_{j=-M_i}^{M_i} \left(1 - \frac{j}{M_i+1}\right) \sum_{t=j+1}^T \Delta y_{it-j} \quad (3.32)$$

Burada M_i , çekirdekte tahmin edilecek kovaryans sayısını belirleyen bant genişliği parametresidir. Δy_{it} 'nin uzun dönem varyansı $\hat{\omega}_{yi}^2$, $\hat{\omega}_{ui}^2/\alpha_i(1)^2$ ile sıfır hipotezi altında kolaylıkla değerlendirilir. Burada da $\hat{\omega}_{ui}^2$ karma (birleşik) hata terimi $u_{it} = \gamma_i(L)'v_{it} + e_{it}$ 'nin uzun dönem varyansını göstermektedir. Bu ifade de $\alpha_i(1)$ alternatif olarak $\hat{\omega}_{ui}^2/\hat{\omega}_{yi}^2$ kullanılarak da tahmin edilebileceği önerilmektedir. $\hat{\omega}_{ui}^2$ yukarıdaki gibi Δy_{it} 'nin $\hat{u}_{it} = \sum_{j=0}^{p_i} \hat{\gamma}_{ij} \Delta x_{it-j} + \hat{e}_{it}$ ile yer değiştirilerek çekirdek (kernel) tahminin kullanılması ile bulunabilir. Burada $\hat{\gamma}_{ij}$ ve $\hat{e}_{it}$ ikinci denklemden elde edilir. $\alpha_i(1)$ 'in yarı parametrik çekirdek (kernel) tahmin edicisi sonuçları da $\tilde{\alpha}_i(1)$ olarak tanımlanır.

3. Grup ortalama test istatistikleri Eşitlik (3.33) ve Eşitlik (3.34)'deki gibi hesaplanır.

$$G_\tau = \frac{1}{N} \sum_{i=1}^N \frac{\hat{\alpha}_i}{SE(\hat{\alpha}_i)} \quad (3.33)$$

$$G_\alpha = \frac{1}{N} \sum_{i=1}^N \frac{T \hat{\alpha}_i}{\hat{\alpha}_i(1)} \quad (3.34)$$

Burada $SE(\hat{\alpha}_i)$, $\hat{\alpha}_i$ 'nin standart hatasıdır. G_α için asimtotik olarak bir problem olmasada özellikle gecikme sayısı büyük olduğunda sıfır hipotezinin red edilmesi daha sık olmaktadır. Bu durumda T yerine herbir gözlem sayısının ortalamasını yazmak küçük örneklerde daha iyi performans üretmektedir.

Panel istatistikleri birinci denklemdeki modelden farklı olarak her yatay kesit birim için değişime izin verdiğinden üç adımlı bir süreç ile testlerin oluşumu özetlenebilir.

1. Grup ortalama istatistikleri gibi gecikme sayısı p_i belirlenir. Her birim için ilk denklem tahmin edilir ve hatalar bulunur. Tahmin edilen hatalar,

$$\Delta \tilde{y}_{it} = \Delta y_{it} - \hat{\delta}'_i d_t - \hat{\lambda}'_i x_{it-1} - \sum_{j=1}^{p_i} \hat{\alpha}_{ij} \Delta y_{it-j} - \sum_{j=0}^{p_i} \hat{\gamma}_{ij} \Delta x_{it-j} \quad (3.35)$$

ve

$$\tilde{y}_{it-1} = y_{it-1} - \hat{\delta}'_i d_t - \hat{\lambda}'_i x_{it-1} - \sum_{j=1}^{p_i} \hat{\alpha}_{ij} \Delta y_{it-j} - \sum_{j=0}^{p_i} \hat{\gamma}_{ij} \Delta x_{it-j} \quad (3.36)$$

şeklindedir.

2. Ortak (genel) hata düzeltme parametresini α ve standart hatasını tahmin edebilmek için $\Delta \tilde{y}_{it}$ ve y_{it-1} kullanılır.

$$\hat{\alpha} = \left(\sum_{i=1}^N \sum_{t=2}^T \tilde{y}_{it-1}^2 \right)^{-1} \sum_{i=1}^N \sum_{t=2}^T \frac{1}{\hat{\alpha}_{i(1)}} \tilde{y}_{it-1} \Delta \tilde{y}_{it} \quad (3.37)$$

$\hat{\alpha}$ 'nın standart hatası,

$$SE(\hat{\alpha}) = \left((\hat{S}_N^2)^{-1} \sum_{i=1}^N \sum_{t=2}^T \tilde{y}_{it-1}^2 \right)^{-1/2} \quad (3.38)$$

olarak tanımlanır. Burada $\hat{S}_N^2 = \frac{1}{N} \sum_{i=1}^N \hat{S}_i^2$ ve eğer $\hat{\sigma}_i$ ikinci denklemden elde edilen regresyonun standart hatası olarak alınırsa, anakütle için tanımlanan $\frac{\sigma_i}{\alpha_i(1)}$ 'nin tutarlı tahmini $\hat{S}_i = \frac{\hat{\sigma}_i}{\hat{\alpha}_i(1)}$ olarak belirlenir.

3. Panel test istatistikleri

$$P_\tau = \frac{\hat{\alpha}}{SE(\hat{\alpha})} \quad (3.39)$$

ve

$$P_\tau = T \hat{\alpha} \quad (3.40)$$

olarak hesaplanır. Burada G_α olduğu gibi P_α 'da küçük örneklemelerde T yerine her birimin gözlem sayısının ortalaması yazmak daha iyi performans üretmektedir (Westerlund, 2007).

3.10. Varsayımlardan Sapmalara Dirençli Tahmin Ediciler

Değişen varyans, otokorelasyon ve yatay kesit bağımlılığı olması durumunda hata teriminin varyans kovaryans matrisi birim matrise eşit değildir. Bu durum, büyük örneklerle çalışıldığı zaman tutarsızlığa neden olmaz, fakat etkinliği etkilemektedir. Bir başka ifade ile, varyanslar ve dolayısıyla standart hataların, t ve F istatistiklerinin, R^2 'nin ve güven aralıklarının geçerliliği etkilenmektedir. Bu nedenlerle modelde değişen varyans, otokorelasyon veya yatay kesit bağımlılığından en az biri varsa, ya parametre tahminlerine dokunmadan standart hatalar düzeltilmeli ya da uygun yöntemlerle tahminler yapılmalıdır.

Dirençli standart hatalar için ilk çalışmalar Huber (1967), Eicker (1967) ve White (1980) tarafından yapılmıştır. Yalnızca değişen varyans olduğu durumlarda Huber, Eicker ve White tahmincisini önermişlerdir. Huber, Eicker ve White tahmincilerinden sonra çalışmalar Arellano (1987), Froot (1989) ve Rogers (1993)

tarafından geliştirilmiş ve kalıntıların bağımsız dağılımlı olması varsayımının esnekleştiği durumda da tahminler yapılmıştır. Panel veri modellerinde birim içerisinde korelasyonlu, birimler arasında korelasyonsuz olduğu durumda dirençli standart hatalar üretilmiştir. Wooldridge (2002), Huber, Eicker ve White tahmincilerini varyans kovaryans matrisinin bilindiği ve köşegensel olmadığı varsayımı altında, bir başka ifade ile sadece değişen varyans ve otokorelasyonun olduğu durumlar için geliştirmiştir ve Tam Dirençli Asimptotik Varyans Tahmindisi olarak adlandırılmaktadır. Bu tahmincede izin verilen otokorelasyon türü, hata terimlerinin birimler içinde korelasyonlu olması, fakat birimler arasında korelasyonsuz olmasıdır. Newey ve West (1987, 1994) tarafından önerilen başka bir yaklaşımda değişen varyans ve otokorelasyon varlığında tutarlı tahminciler üretilmektedir. Newey-West'in genelleştirilmiş momentler temelli kovaryans matrisi, White'in tahmincisinin genişletilmiş bir halidir ve otokorelasyon için gecikme uzunluğu sıfır seçildiğinde (otokorelasyon olmadığında), Newey-West tahmincisi White tahmincisine eşittir. Panel veri modellerinde, birimler arası (uzamsal) korelasyonu dikkate alan ilk çalışmalardan biri Anselin (1988) tarafından yapılmıştır. Panel veri modellerinde, değişen varyans yanında dönemsel ve uzamsal korelasyonu dikkate alan ilk çalışma Parks (1967) tarafından yapılmıştır. Parks, daha sonra Kmenta (1986) tarafından tanıtılan esnek genelleştirilmiş en küçük kareler yöntemi temelli bir algoritma önermiştir. Parks-Kmenta'nın yönteminden elde edilen varyanslar aşağıya doğru sapmalıdır, ayrıca birim sayısının zaman boyutundan büyük olduğu durumlarda kullanılması uygun değildir. Parks-Kmenta yönteminin bu problemlerinden dolayı Beck ve Katz (1995), 'panel düzeltilmiş standart hataları' (PCSE) yöntemini önermiştir. Beck ve Katz (1995), birimler arası korelasyonu düzeltmek için kullanılan ve büyük T asimptotik temelli standart hatalar üreten bu yöntemin küçük panellerde de kullanımının uygun olduğunu göstermiştir. Bununla birlikte, PCSE tahmincisinin küçük örnek özellikleri yatay kesit boyutu N 'nin zaman boyutu T 'den büyük olduğu durumlarda zayıftır. Zaman boyutu T 'nin büyük olduğu düşünüldüğünde,

Driscoll ve Kraay (1998) standart parametrik olmayan serisi kovaryans matris tahmincilerinin uzamsal ve dönemsel korelasyonun tüm genel formları için dirençli olabilecek şekilde geliştirilebileceğini göstermiştir. Driscoll ve Kraay'ın metodolojisi, yatay kesit ortalamaları serisi için Newey-West türü yapmaktadır. Bu şekilde düzeltilmiş standart hata tahminleri, yatay kesit boyutu N 'den bağımsız olarak, N sonsuz bile olsa, kovaryans matris tahmincilerinin tutarlılığını garantilemektedir. Böylece Driscoll ve Kraay'ın yaklaşımı, özellikle mikro ekonometrik panellerde karşılaşılan yatay kesit boyutun büyüklüğü durumunda zayıf olan sadece T olduğu durumda tutarlı kovaryans matris tahmincileri üreten Parks-Kmenta ya da PCSE yaklaşımlarına alternatif olarak üretilmiştir. Bu tahminci, büyük T ve N durumunda bile değişen varyans varlığında tutarlı, uzamsal ve dönemsel korelasyonun genel formlarında dirençli standart hatalar üretmektedir (Tatoğlu, 2013a).

3.10.1. Parks-Kmenta Tahmincisi

İlk olarak Parks (1967) tarafından önerilen ve Kmenta (1986) tarafından tartışılan Park- Kmenta yöntemi, genelleştirilmiş en küçük kareler (GLS) uygulaması kullanılır. Bu yöntem için regresyon eşitliği,

$$y_{it} = \beta_1 + \sum_{k=2}^k \beta_k x_{kit} + e_{it} \quad (3.41)$$

şeklindedir. Eşitlikte tek kesim ve zaman ve birim noktalarına karşı eğim katsayısı vardır. Fakat Parks-Kmenta yöntemine göre, bu eşitlik GLS ile tahmin edilebilmelidir. Çünkü bu yöntem süreci, klasik regresyon modelinden, regresyon dağılımının davranışına ilişkin daha az kısıtlayıcı varsayımlara ve dolayısıyla klasik regresyon modele göre varyans-kovaryans matrisi Ω 'ya dayanmaktadır. Bu nedenle, GLS tahmini zaman serileri ve yatay kesit gözlemleriyle özel ilgilidir. β parametre tahminleri problem ile ilgili eşitlik,

$$(x'\Omega^{-1}x)^{-1}x'\Omega^{-1}y \quad (3.42)$$

şeklindedir. Bu tahmin, hataların varyans kovaryans matrisi Ω 'nın bilindiği varsayımına dayanır. Fakat varyans kovaryans matrisinin bilinmediği durumlardan dolayı GLS kullanılmayıp “Esnek” genelleştirilmiş en küçük kareler (FGLS) kullanılır. Varyans kovaryans matrisinin tahmini kullanıldığından esnektir. Ω 'nın tutarlı tahmin edicisi $\hat{\Omega}$ bulunur ve formülde Ω yerine $\hat{\Omega}$ yazılarak β tahmini elde edilir.

Parks–Kmenta yöntemi hataların değişen varyans, yatay kesit bağımlılığı ve otokorelasyon varsayımlarını içerir. Bu varsayımları özellikleri,

$$\begin{aligned} E(e_{it}^2) &= \sigma_{ii} \\ E(e_{it}e_{jt}) &= \sigma_{ij} \\ e_{it} &= \rho_i e_{it-1} + v_{it} \end{aligned}$$

olarak ifade edilir. Diğer bir deyişle bu yaklaşım AR(1) (birinci mertebeden otoregresif model) adı verilen otokorelasyon, yatay kesit bağımlılığı ve değişen varyans için sırasıyla model belirleyerek hata komplikasyonları ile ilgilenir. Burada ρ_i birinci mertebeden otoregresiflik katsayısıdır. ρ_i ve hataların varyans kovaryans matrisinin elemanları σ^2 'nin tutarlı tahminlerinin bulunması gerekmektedir. Bu amaç için, Parks–Kmenta yöntemi iki ardışık FGLS dönüşümü içerir. İlkinde, hataların otokorelasyonu yok edilir sonra hataların yatay kesit bağımlılığı yok edilir. Eşitlik (3.42), EKK ile tahmin edilir. Bu tahminden elde edilen artıklar hataların birim özel otokorelasyonunu tahmin etmek için kullanılır. Daha sonra hataların yatay kesit bağımlılığını tahmin etmek için kullanılır ve veri komplikasyonlardan arındırılmış hatalarla EKK tahmini yapmak için tekrar dönüştürülür. ρ_i ve σ^2 'nin tutarlı tahminleri elde edilerek Ω 'nın elemanlarının tutarlı tahminlerinin türetilmesi görevi tamamlanır. Böylece Ω yerine $\hat{\Omega}$ yazılarak

standart hataların ve istenen katsayıların tahminleri elde edilir (Podesta, 2000).

3.10.2. Beck-Katz Tahmincisi

Beck ve Katz(1995), Parks-Kmenta yöntemini revize etmişlerdir. GLS panel veri için optimal özellikteyken gerçekte uygulanan FGLS aynı değildir. Bunun nedeni, FGLS hata süreçlerinin tahmininde kullanılmasına rağmen, standart hatalar için FGLS hataların varyans kovaryans matrisinin bilindiği varsayımını sağlamaması ve tahmin edilememesidir. Beck ve Katz $N > T$ durumunda standart hatalardaki güven aşırılığından dolayı Parks-Kmenta yönteminin uygun olmadığını göstermiştir.

Beck ve Katz EKK parametre tahmin edicisini tutmayı önerir fakat değişen varyans ve hataların yatay kesit bağımlılığını dikkate alan panel düzeltilmiş standart hatalar (PCSEs) ile EKK standart hatalarını yer değiştirir. Fakat hataların otokorelasyonu PCSEs hesaplanmadan önce yok edilmelidir. Otokorelasyon, geçikmeli bağımlı değişken bağımsız değişken olarak ele alınıp modellenabilir veya Parks-Kmenta'nın önerdiği yöntem ile otoregresiflik için model tahmin edilerek düzeltilir. Aşağıdaki eşitlik ile geliştirilebilir.

$$y_{it} = \beta_1 + \beta_2 y_{it-1} + \sum_{k=3}^k \beta_k x_{kit} + e_{it} \quad (3.43)$$

Eşitliğin sağ tarafına bağımlı değişkenin gecikmeli değerinin birleştirilmesi bağımlı değişkende basıklık veya sürekliliğin açık bir tahmini üretir. Burada y_{it-1} , bağımlı değişkenin ilk gecikmesidir ve β_2 eğim katsayısıdır. Dinamiklik açıklanırken EKK ile model parametreleri, değişen varyans ve hataların yatay kesit bağımlılığını dikkate almak için PCSEs ile standart hatalar tahmin edilir. EKK tahminlerinin örnek geçerliliği için doğru formül aşağıda ifade edilen terimin kökleri ile verilir.

$$Cov(\beta) = (x'x)^{-1}(x'\Omega x)(x'x)^{-1} \quad (3.44)$$

Eşitliğin orta terimi panel veri için korelasyon içerir. Artıkların değişen varyanslı ve yatay kesit bağımlılığı koşulu altında, hataların varyans kovaryans matrisi Ω , $N \times N$ 'lik yatay kesit kovaryans matrisli $NT \times NT$ boyutlu blok köşegen matristir. Eşitlik (3.44)'i tahmin etmek için Θ 'nın tutarlı tahminine ihtiyaç vardır. Eşitlik (3.42)'ün EKK tahmini tutarlı olduğundan Θ 'nın tutarlı tahminini üretmek için EKK artıkları kullanılır (Podesta, 2000).

3.10.3. Driscoll ve Kraay Tahmincisi

Driscoll ve Kraay (1998) zaman boyutunun (T) büyük olduğu durumlarda, standart parametrik olmayan zaman serisi kovaryans matris tahmin edicilerinin uzamsal ve dönemsel korelasyonun tüm genel formları için dirençli olabilecek şekilde geliştirilebileceğini göstermiştir.

Driscoll ve Kraay tahmincisi yatay kesit boyutun büyüklüğü durumunda zayıf olan, sadece büyük T olduğu durumda tutarlı kovaryans matris tahmincileri üreten Parks-Kmenta ya da PCSE yaklaşımlarına alternatif olarak türetilmiştir. Bu tahminci, büyük T ve N durumunda bile değişen varyans varlığında tutarlı, uzamsal ve dönemsel korelasyonun genel formlarında dirençli standart hatalar üretmektedir. Hata teriminin değişen varyans, otokorelasyonlu ve yatay kesit bağımlılığı olduğu varsayımları altında, parametreler havuzlanmış EKK yöntemi ile tutarlı bir şekilde tahmin edilebilmektedir. Parametre tahminlerinin Driscoll ve Kraay standart hataları ise, asimptotik (dirençli) kovaryans matrisinin diagonal elemanlarının karekökleri yardımıyla elde edilmektedir.

$$V(\hat{\beta}) = (x'x)^{-1} \hat{S}_T (x'x)^{-1} \quad (3.45)$$

Burada $\hat{S}_T = \hat{\Omega}_0 + \sum_{j=1}^{m(T)} w(j, m) [\hat{\Omega}_j + \hat{\Omega}_j^2]$ dur. $m(T)$, otokorelasyon için

gecikme uzunluğunu ifade etmektedir. $w(j, m(T)) = 1 - j / (m(T) + 1)$ olarak ifade edilen Barlett ağırlıkları, $\hat{S}_T$ 'nin pozitif tanımlı olmasını sağlamak ve örnek otokovaryans fonksiyonunda yüksek mertebeden gecikmelerin düşük ağırlıklar almasına imkan sağlamaktadır. $(K+1) \times (K+1)$ boyutlu $\hat{\Omega}_j$ matrisi ise $\hat{\Omega}_j = \sum_{t=j+1}^T h_t(\hat{\beta}) h_{t-j}(\hat{\beta})'$ 'dir. Burada $h_t(\hat{\beta}) = \sum_{i=1}^{N(t)} h_{it}(\hat{\beta})$ eşitliği vardır. Her bir birim için t moment koşullarının karesi $h_{it}(\hat{\beta})$, farklı T 'lere sahip N 'ler için hesaplanmaktadır. Bu küçük düzeltme ile Driscoll ve Kraay kovaryans matris tahmincisi, dengesiz panel veri modellerinde de kullanılabilir.

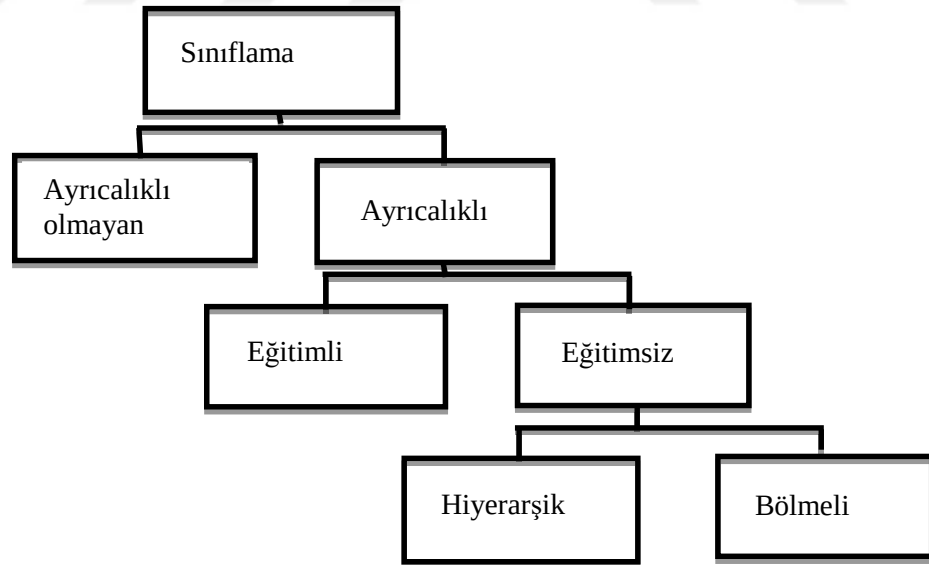
Yatay kesit ortalamalarına dayanan bu yaklaşımla, standart hata tahminleri, birimlerin yatay kesit boyutu N 'ye bağlı olmaksızın tutarlıdır (Tatoğlu, 2013a).

4. KÜMELEME YÖNTEMLERİ VE ALGORİTMALAR

Kümeleme, birimleri gruplara ayıran bir sınıflama tipidir. En basit tanımıyla benzer özellik gösteren veri elemanlarının kendi aralarında gruplara ayrılmasıdır (Demiralay ve Çamurcu, 2005). Birimler arasındaki ilişki, birimlere karşılık gelen satır ve sütunlardaki yakınlık/uzaklık matrisinde verilir. d -boyutlu metrik uzaydaki noktalar karakterize edilirse, nokta çiftleri arasındaki yakınlıklar uzaklık olabilir. Kümeleme analizi, birimleri değişkenler arası benzerlik ya da uzaklıklara dayalı olarak hesaplanan bazı ölçülerden yararlanarak homojen gruplar oluşturmaya çalışır (Martinez ve Martinez, 2005).

Kümeleme mühendislik, bilgisayar bilimleri, yaşam ve tıp bilimleri, astronomi ve dünya bilimleri, sosyal bilimler ve ekonomi gibi geniş bir uygulama alanına sahiptir (Xu ve Wunsch, 2009).

Kümeleme sınıflamanın özel bir türüdür. Şekil 4.1’de kümeleme ve sınıflama arasındaki ilişki verilmiştir.



Şekil 4.1. Sınıflama tipleri ağacı

- a. Ayrıcalıklı olmayana karşı ayrıcalıklı sınıflama: Ayrıcalıklı sınıflama, birimlerin bölmeli bir grubudur. Her bir birim kesinlikle bir alt gruba ya da kümeye aittir. Ayrıcalıklı olmayan kümeleme de ise, birimler birkaç sınıfa atanır. Örneğin insanları yaş veya cinsiyete göre gruplamak ayrıcalıklı iken kişi eşzamanlı birkaç rahatsızlığa sahip ise kategorik olarak gruplamak ayrıcalıklı olmayan sınıflamadır.
- b. Eğitilmiş karşı eğitimsiz: Eğitimsiz sınıflama, sınıflama yapmak için sadece yakınlık matrisini kullanır. Eğitilmiş sınıflama ise yakınlık matrisinin yanı sıra birimlerin kategorik etiketlerini de kullanır.
- c. Bölmeliye karşı hiyerarşik: Ayrıcalıklı, eğitimsiz sınıflama verinin yapısına göre hiyerarşik ve bölmeli olarak tekrar ikiye ayrılır. Hiyerarşik sınıflama bölümlerin iç içe geçmiş dizisi iken bölmeli sınıflama tek bölümdür. Bu yüzden hiyerarşik sınıflama bölmeli sınıflamanın özel bir durumudur.

Ayrıcalıklı, eğitimsiz, bölmeli sınıflama için kümeleme terimi (hiyerarşik olmayan ya da bölmeli kümeleme) ve ayrıcalıklı, eğitimsiz, hiyerarşik sınıflama için hiyerarşik kümeleme terimi kullanılır (Jain ve Dubes, 1998).

4.1. Hiyerarşik Kümeleme

Hiyerarşik küme algoritmaları hiyerarşik kümeleme yapabilmek için özel adımlardır. Ardışık bir işlem süreci içerirler. Hiyerarşik kümeleme yöntemlerinin, araştırmacının incelediği veri setinde kaç küme bulunduğunu başlangıçta bilmediği durumlarda çok uygun bir yöntem olduğunu ve yöntemin araştırmacıya incelediği veri setinde daha önce gözlenmemiş ilişkileri ve prensipleri keşfetme olanağı vermesi nedeniyle faydalı olduğu belirtilmektedir (Anderberg, 1973).

Kümeleme analizi yöntemlerinde en temel adım birimler arasında benzerliğin bulunmasında kullanılan uzaklık ölçümleridir. Uzaklıklar veya benzerlikler kümeleme analizinde önemli rol oynar. Her kümeleme algoritması, veri

noktaları arasındaki benzerlik (yakınlık) veya benzemezlik (uzaklık) indeksine dayalıdır. Uzaklık ölçümleri anlamlı değilse, anlamlı bir kümeleme analizi mümkün değildir. Veri kümeleme literatüründe benzerlik ölçümleri, benzerlik katsayıları, benzemezlik ölçümleri veya uzaklıklar iki noktanın veya iki kümenin benzerliğini yada benzemezliğini en iyi şekilde tanımlamalıdır. Veri nokta çiftleri arasında benzerlik veya benzemezlik ölçümü yoksa anlamlı bir kümeleme analizi mümkün değildir. Veri setlerindeki birim sayısı, göz önüne alınacak değişken sayısı, değişkenler arasındaki ilişki yapısı, küme sayısı, ölçek türü gibi birçok özellik farklılık gösterebileceğinden herhangi bir kümeleme yönteminin her zaman herhangi bir uzaklık türü ile en iyi sonucu vereceği şeklinde bir genelleme yapılamaz (Erişoğlu, 2011).

Algoritmalarda birimlerin birbirine uzaklıklarını hesaplamak için kullanılan uzaklık ölçümleri verinin tipine (nümerik, ikili ve karma değişkenli) göre farklılık göstermektedir.

4.1.1. Nümerik Değişkenler için Kullanılan Uzaklık Ölçüleri

Algoritmalarda nümerik değişkenlere sahip her bir birimin birbirine uzaklığını hesaplamak için kullanılan formüller aşağıda açıklanmıştır.

Öklit Uzaklığı ve Öklit Uzaklığının Karesi: Öklit uzaklığı ve öklit uzaklığının karesi formüllerinde standartlaştırılmamış (işlenmemiş) verilerle hesaplama yapılır. Öklit uzaklıkları kümeleme analizinde sıra dışı olabilecek yeni nesnelerin eklenmesinden etkilenmezler. Ancak boyutlar arasındaki ölçek farklılıkları öklit uzaklıklarını önemli ölçüde etkilemektedir. Öklit uzaklık formülü en yaygın olarak kullanılan uzaklık hesaplama formülüdür. Öklit ve Öklit uzaklığının karesinin formülleri sırasıyla Eşitlik (4.1) ve Eşitlik (4.2)'de verilmiştir.

$$dis(x, y) = \{\sum_i (x_i - y_i)^2\}^{1/2} \quad (4.1)$$

$$dis(x, y) = \sum_i (x_i - y_i)^2 \quad (4.2)$$

Manhattan Uzaklık Ölçümü: Manhattan uzaklığı boyutlar arasındaki ortalama farka eşittir. Manhattan uzaklığı,

$$dis(x, y) = \sum_i |x_i - y_i| \quad (4.3)$$

eşitliği ile verilir.

Chebychev Uzaklığı: İki nesne arasındaki mutlak maksimum uzaklığa eşittir ve

$$dis(x, y) = \max |x_i - y_i| \quad (4.4)$$

olarak ifade edilir.

$dis(x, y)$: x ve y noktaları arasındaki uzaklık, hatta parametredir. x , y ise aralarındaki uzaklık hesaplanan nesneleri uzayda temsil eden noktalardır (Kaufman ve Rousseeuw, 1990; Demiralay ve Çamurcu, 2005).

4.1.2. İki Sonuçlu (Binary) Değişkenler için Kullanılan Uzaklık Ölçüleri

İkili değişken yalnızca iki mümkün sonuca sahiptir. Örneğin, insanları kümelerken birkaç ikili değişken kullanılabilir: erkek/kadın, sigara içen/içmeyen, evet/hayır gibi. Veri matrisinde böyle değişkenler sıklıkla 0 ve ya 1 ile kodlanır. f ikili değişken r . birim ya $X_{rf} = 0$ ya da $X_{rf} = 1$ olacaktır. Özellikte var olanlar sıklıkla 1 değerini alırken yokluk 0 ile gösterilir. İkili değişkenler bazen aralık ölçümü olarak hesaplanabilir, yani Öklid ve Manhattan uzaklığı uygulanabilir. Bu bazen iyi sonuçlar vermesine rağmen ikili veri için tasarlanan özel yaklaşımlar vardır.

İkili verilerden benzemezlik (uzaklık) matrisi hesaplamak ve kümeleme algoritmalarından birini uygulamak mümkündür. r ve k birim arasındaki benzerlik $s(r,k)$ veya benzemezlik $d(r,k)$ hesaplanırken 2x2'lik olasılık Çizelgesi aşağıdaki gibidir.

Çizelge 4. 1. İki sonuçlu p sayıda değişken içeren bir gözlem çiftinin (r,k) çapraz çizelgesi

		k . birim		
		1	0	
r . birim	1	a	b	a+b
	0	c	d	c+d
		a+c	b+d	p

Burada a , iki birim için 1'e eşit değişken sayısıdır. Benzer olarak b , $X_{if} = 1$ ve $X_{kf} = 0$ olan f değişkenlerinin sayısıdır. Açık olarak, $a+b+c+d=p$ toplam değişken sayısıdır. Basit eşleştirme uzaklık ölçümü için benzerlik ve benzemezlik (uzaklık) ölçümü sırasıyla aşağıdaki gibi tanımlanmıştır.

$$s(r, k) = \frac{a+d}{a+b+c+d} \quad (4.5)$$

$$d(r, k) = \frac{b+c}{a+b+c+d} \quad (4.6)$$

$d(r, k) = 1 - s(r, k)$ olarak da tanımlanabilir (Kaufman ve Rousseeuw, 2005).

İki sonuçlu p değişkenden oluşan veriler için birçok eşleştirme tipli benzerlik ölçüsü kullanılmakla birlikte en yaygın olarak kullanılanların bir listesi Çizelge 4.1.2.2 ile verilmiştir. İki sonuçlu p değişkenden oluşan gözlemler için bu kadar çok eşleştirme tipli benzerlik ölçüsünün kullanılmasının nedeni 0-0 eşleştirmesinin nasıl yorumlandığı ile ilgilidir. Bazı durumlarda 0-0 eşleştirmesi, 1-1 eşleştirmesi

ile eşdeğer olabilir. Örneğin cinsiyetin erkek için 1 değeri ve kadın için 0 değeri ile kodlanması veya tam tersi şeklinde kodlanması arasında bir fark yoktur. Bazı durumlarda ise 0–0 eşleştirmesi bir gözlem üzerinde herhangi bir özelliğin bulunup bulunmadığı ile ilgili olabilir. Örneğin bir böceğin bir özelliği kanat olup olmaması ise kanatlı olmasının 1, olmamasının 0 ile kodlanması gibi (Servi, 2009). İki sonuçlu p değişken içeren verinin gözlem çiftleri arasındaki uzaklık için kullanılan bazı eşleştirme tipli benzerlik ölçüleri Çizelge 4.1.2.2’de verilmiştir.

Çizelge 4.2. İki sonuçlu p değişken içeren verinin gözlem çiftleri arasındaki uzaklık için kullanılan bazı eşleştirme tipli uzaklık ölçüleri. (Kaufman ve Rousseeuw, 2005)

Ölçüm	$d(i,j)$
Eşleştirme Katsayısı (Zubin 1938)	$\frac{b + c}{a + b + c + d}$
Rogers ve Tanimoto (1960)	$\frac{2(b + c)}{(a + d) + 2(b + c)}$
Sokal ve Sneath (1963)	$\frac{b + c}{2(a + d) + (b + c)}$
Jaccard Katsayısı (1908)	$\frac{b + c}{a + b + c}$
Dice (1945)	$\frac{b + c}{2a + b + c}$
Sokal ve Sneath (1963)	$\frac{2(b + c)}{a + 2(b + c)}$

4.1.3. Karma Değişkenler için Kullanılan Uzaklık Ölçüleri

Gerçek hayatta karma değişkenli veri grupları yaygındır. Karma veri tipleri ve gruplarının geleneksel küme yaklaşımı; bir tipteki değişkeni diğerine çevirerek yeniden sunulan uzaklığı kullanmaktır. İlk yol ikili veriyi basit eşleştirme veya ikili kodlarla nümerik veriye dönüştürmektir. Diğer ise nümerik değişkenleri kategorik veriye dönüştürmektir. Fakat bu yaklaşımlar bilgi kaybına neden olmaktadır. Dönüşüm sürecindeki bilgi kaybı, kümelemede kayda değer bir yanlışlık içerir. Bu

problemin üstesinden gelmek için karma veri gruplarında bilgi kaybı olmaksızın etkili çalışan bir uzaklık ölçümü önerilir.

Bazı araştırmacılar nümerik ve nominal uzaklıkları dengelemeye çalışmışlardır. Huang (1998) k-prototip algoritmasında kategorik uzaklık için ağırlık vektörü tanımlamıştır. Modha ve Spangler (2003) nümerik ve nominal uzaklığı dengeleyen optimal ağırlığı önermişlerdir. Goodall (1966) tarafından önerilen Goodall uzaklığı, kategorik ve nümerik uzaklığı ortak bir taslakta birleştirmiştir. Ahmad ve Dey (2007) farklı değişkenlerde değerlerin birliktelik bağlantısını kullanarak eşdizimlilik uzaklığını tanımlamışlardır. k-prototip, optimal ağırlık, Goodall uzaklıkları ve Gower benzerlik katsayısı aşağıda verilmiştir.

4.1.3.1. k-prototip Uzaklığı

Karma veri grubu kümelenirken, Huang (1998), nümerik ve nominal değişkenlerin uzaklığını dengelemek için kullanıcıya özel γ ağırlığını kullanmıştır ve k-prototip algoritmasında bu ölçümü uygulamıştır. Nümerik uzaklık öklid uzaklığının karesidir ve kategorik uzaklık basit eşleştirmedir. Tüm nümerik değişkenler $[0,1]$ aralığında normalleştirilir. Küçük γ değeri kümelemede nümerik değişkenlerin baskın olduğunu, büyük γ değeri kümelemede kategorik değişkenlerin baskın olduğunu gösterir. Huang (1998), ağırlığın $[0.5, 1.4]$ aralığında olmasını önermiştir. x ve y ikili birimlerin arasındaki toplam uzaklık Eşitlik (4.7)'deki gibi tanımlanır.

$$d(x, y) = \sum_{i \in A_r} \left\{ \frac{a_i(x) - a_i(y)}{\max(V_{a_i}) - \min(V_{a_i})} \right\}^2 + \gamma \sum_{i \in A_c} D^i(x, y) \quad (4.7)$$

4.1.3.2. Optimal Ağırlık Uzaklığı

Modha ve Spangler (2003) dengeli nümerik ve nominal uzaklıkları daha iyi dengelemek için optimal yöntemleri uygulamışlardır. Kullanıcıya özel ağırlık

yerine küme sayısı verildiğinde kümeler içi ve kümeler arası ağırlıklandırılmış uzaklıkların oranını minimize etmek için optimal ağırlığı araştırmışlardır. Nümerik özellikler ortalama ve standart sapmaya bağlı olarak standartlaştırılmıştır ve uzaklık öklid uzaklığının karesi alınarak bulunmuştur. Her bir kategorik değer $1-v$ (v , değişken değerlerinin sayısıdır) kodu kullanılarak ikili vektör olarak sunulmuştur ve uzaklık kosinüs uzaklığı alınarak bulunmuştur. Optimal ağırlık uzaklığı iki veri tipinin ağırlıklandırılmış uzaklığı ile birleştirilmiştir. Optimal ağırlık uzaklığı,

$$d(x, y) = (1 - w) \sum_{i \in A_r} \frac{(a_i(x) - a_i(y))^2}{\text{var}(V_{a_i})} + w \sum_{i \in A_c} D^i(x, y) \quad (4.8)$$

olarak ifade edilmektedir. Optimal ağırlığı elde etmek için, küme sayısı önceden bilinmelidir. Bu minimizasyon probleminin amaç fonksiyonu lineer olmadığından, optimal bir çözüm bulmak zordur. Modha ve Spangleren iyi çözümünü bulmak için $[0, 1]$ aralığında çok sayıda değer kullanarak objektif değeri hesaplamışlardır (Yang, 2012).

4.1.3.3. Goodall Uzaklığı

Goodall (1966), ikili birimin ortak olmayan değişken değerinin diğer ikili çiftlerin kendi aralarında sahip oldukları ortak değerden daha yakın olduğu fikrine dayalı benzerlik indeksi önermiştir. Örneğin, alabalık ve levrek aynı ölçülere sahiptir fakat yunus ve alabalık omurgalıdır. Çoğu hayvan ölçüden ziyade omurgaya sahip olduğundan alabalık yunustan daha çok levreğe yakındır.

Li ve Biswas(2002) bu uzaklığı toplamsal hiyerarşik kümeleme yöntemlerinde kullanmışlardır. Aynı olmayan nominal değerler için uzaklık $D^i(x, y) = 1$, $a_i(x) \neq a_i(y)$, $a_i \in A_c$ 'dir. Örneğin, meyve sepetinde sarı ve kırmızı arasındaki uzaklık $D^{renk}(\text{muz}, \text{çilek}) = 1$ 'dir. Fakat $D^{renk}(\text{muz}, \text{çilek})$ ise oldukça karmaşıktır. Aynı nominal değerli ikililer için uzaklık, eşit veya

oldukça benzer ikili çiftin seçilme olasılıklarının toplamıdır. Yani düşük veya eşit sıklığa sahip olmasıdır. $s = a_i(x) = a_i(y)$, $a_i \in A_c$ iken formül aşağıdaki gibidir.

$$D^i(x, y) = \sum_{r \in V_{a_i}, \text{freq}(r) \leq \text{freq}(s)} \frac{\text{freq}(r)(\text{freq}(r)-1)}{n(n-1)} \quad (4.9)$$

Aynı nümerik değerler arasındaki uzaklık Eşitlik (4.9) kullanılarak nominal bileşim olarak hesaplanır.

İki farklı nümerik değer arasındaki uzaklığı hesaplamak için, alan (domain) nümerik özelliğin benzersiz (tek) değerleri ile ardışık parçalara bölünür ve her aralıkta frekansı sayılır. $s = a_i(x)$, $t = a_i(y)$, $a_i(x) \neq a_i(y)$, $a_i \in A_r$ verilsin nümerik değişken A_i 'deki uzaklık aşağıdaki gibi tanımlanır.

$$D^i(x, y) = \sum_{r \in V_{a_i}} \frac{\text{freq}(r)(\text{freq}(r)-1)}{n(n-1)} + \sum_{|m-l| < |t-s|} \left\{ \frac{|m-l|=|t-s|, \text{freq}(|m-l|) \leq \text{freq}(|t-s|)}{n(n-1)} \right\} \frac{2\text{freq}(l)(\text{freq}(m)-1)}{n(n-1)}$$

Her bir değişken için uzaklık hesaplanırken, ki-kare değerine karşılık gelen χ^2 dönüşümü kullanılır. Bu değerlerin toplamı değişken sayısının 2 katı olan serbestlik dereceli χ^2 dağılımına sahiptir. Bu toplamın olasılığı ikili birimlerin Goodall uzaklığıdır (Yang, 2012).

4.1.3.4. Gower Benzerlik Katsayısı

Gerçek veri gruplarında, hem nümerik hem de kategorik özelliklerin aynı zamanda görülmesi oldukça yaygındır. Yani veri karma durumdadır. Bu tip veriler için Gower (1971) tarafından benzerlik ölçüsü önerilmiştir. x_i ve x_j iki boyutlu karma veri birimleri için benzerlik ölçümü;

$$S(x_i, x_j) = \frac{\sum_{l=1}^d \delta_{ijl} S_{ijl}}{\sum_{l=1}^d \delta_{ijl}} \quad (4.10)$$

şeklindedir. Burada S_{ijl} , l . özellik için iki birim arasındaki benzerliği gösterir. δ_{ijl} : iki birim ölçümünde kayıp veri olup olmamasına dayalı olarak 0-1 katsayısıdır. $D(x_i, y_i) = 1 - S(x_i, y_i)$ eşitliği kullanılarak uzaklık ölçüsü elde edilebilir. Kesikli değişken için (ikili değişken içeren) benzerlik;

$$S_{ijl} = \begin{cases} 1 & x_{il} = x_{jl} \\ 0 & x_{il} \neq x_{jl} \end{cases} \quad (4.11)$$

olarak ifade edilir. Nümerik değişkenler için S_{ijl} ,

$$S_{ijl} = 1 - \frac{|x_{il} - x_{jl}|}{R_l} \quad (4.12)$$

şeklindedir. Burada R_l tüm birimlerin l . değişkenin aralığıdır ve $R_l = \max x_{ml} - \min x_{ml}$ olarak hesaplanır (Xu ve Wunsch, 2009).

Veri normalleştirme, ön süreç işlemlerinden biridir, burada değişkenlerin $[-1,1]$ veya $[0,1]$ gibi özel aralığa düşmesi sağlanır. Kümelemeden önce normalleştirme, değişken ölçümlerindeki veya büyüklüklerindeki farklılıklara hassas olan Öklid gibi metrik uzaklık için özellikle gereklidir. Gerçek uygulamalarda değişken değerlerinin aralıklarından dolayı, bir değişken diğerini çok etkilemiş olabilir. Normalleştirme maaş gibi daha geniş aralıklı değişkenlerin yaş gibi daha küçük aralıklı değişkenlere daha ağır basmasını önler. Amaç bu değişkenlerin değişkenliğini ve boyut veya büyüklüğünü eşitlemektir.

4.1.4. Matris Güncelleme Algoritmaları

Tek Bağlantı ve Tam Bağlantı kümeleme için matris güncelleme algoritmaları geliştirilmiştir. Bu yaklaşım King (1967) tarafından önerilmiş ve bu yöntemi formülleştiren Johnson (1967) tarafından popüler hale getirilmiştir. Bu

algoritma, eski küme yeni kümeyle birleşince uzaklık matrisindeki satır ve sütunları silen toplamsal bir tasarımıdır.

$\mathcal{D} = [d(i, j)]$ $n \times n$ boyutlu uzaklık matrisidir. Kümelere $0, 1, \dots, (n-1)$ sıra numaraları atanır ve $L(k)$ k . kümenin düzeyidir. m sıra numaralı küme (m) ile gösterilir ve (r) ve (s) kümeleri arasındaki uzaklık $d[(r), (s)]$ ile gösterilir. Tek ve Tam Bağlantı kümeleme için Johnson Algoritması adımları aşağıdaki gibi verilmiştir.

Adım 1. $L(0)=0$ düzeyinde ve sıra numarası $m=0$ olan ayrı kümelerle başlanır.

Adım 2. Şu anki kümedeki en az uzaklığa sahip olan küme çifti bulunur $\{(r), (s)\}$ çifti

$$d[(r), (s)] = \min\{d[(i), (j)]\}$$

burada minimum şu anki kümedeki tüm küme çiftlerini kapsar.

Adım 3. Sıra numarası artımı: $m \leftarrow m+1$, (r) ve (s) kümeleri bir sonraki m kümesinde birleştirilir. $L(m) = d[(i), (j)]$ kümenin düzeyidir.

Adım 4. (r) ve (s) kümelerine karşılık gelen satır ve sütunlar silinerek uzaklık matrisi $\mathcal{D}$ güncellenir ve yeni oluşturulan kümeyle karşılık gelen satır ve sütunlar eklenir. Yeni küme arasındaki uzaklık $d(r, s)$ olarak gösterilir ve eski küme (k) aşağıdaki gibi tanımlanır.

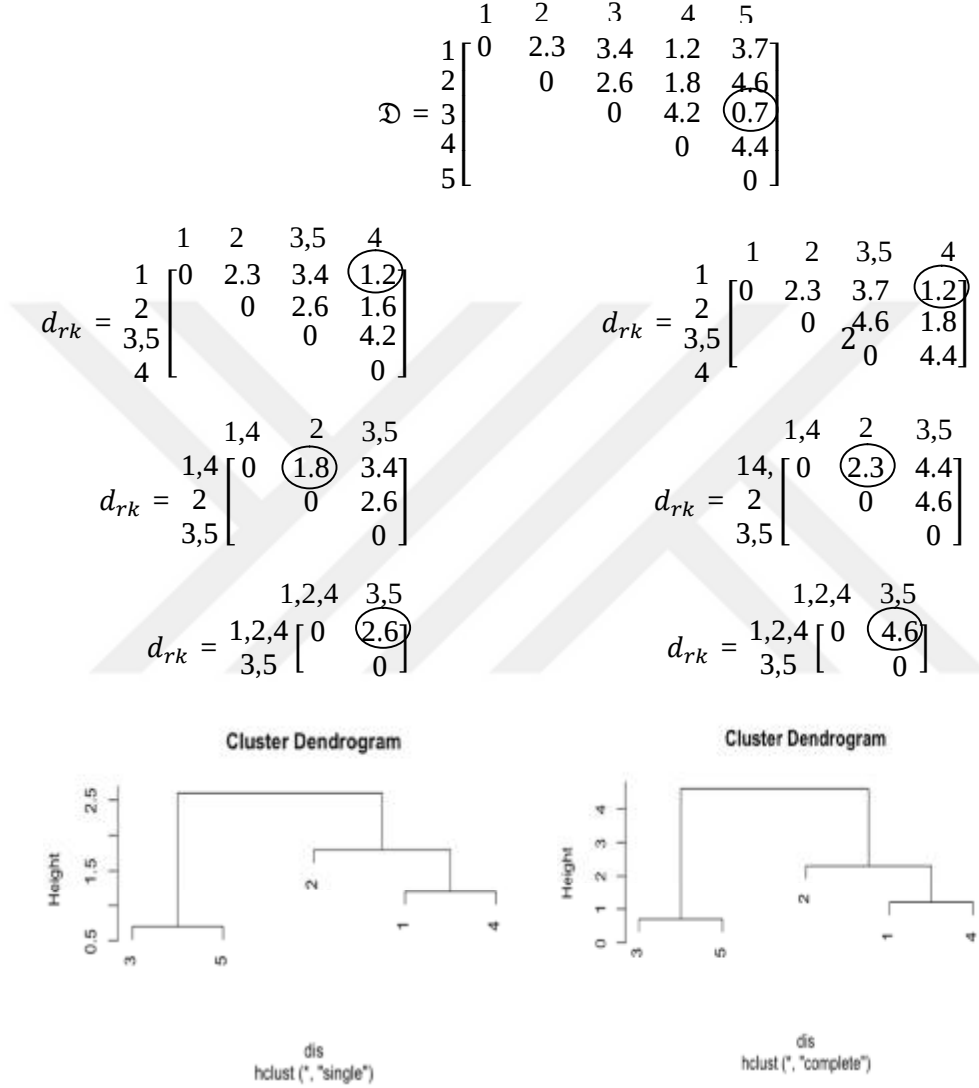
Tek bağlantı ve tam bağlantı yöntemleri için sırasıyla Eşitlik (4.13) ve Eşitlik (4.14)'teki gibi ifade edilir.

$$d[(k), (r, s)] = \min\{d[(k), (r)], d[(k), (s)]\} \quad (4.13)$$

Tam bağlantı yöntemi için;

$$d[(k), (r, s)] = \max\{d[(k), (r)], d[(k), (s)]\} \quad (4.14)$$

Adım 5. Tüm elemanlar tek bir küme oluşturunca durulur aksi taktirde Adım 2'ye dönülür.



Şekil 4.2. Tek bağlantı ve tam bağlantı kümeleme için matris güncelleme örneği

Tek bağlantı ve Tam bağlantı algoritmaları, $\{(k), (s)\}$ ve $\{(k), (r)\}$ ikilileri arasındaki uzaklığın sırasıyla minimumunu ve maksimumunu kullanır. Diğer kümeleme yöntemleri için uzaklıkların farklı kombinasyonları özelleştirilerek

tanımlanır. Algoritmadaki adım 4 için n_k elemanlı var olan k kümesi ve yeni oluşturulan (r,s) kümesi arasındaki uzaklık için uzaklık matrisinin nasıl güncelleneceğine dair formül tanımlanır.

Hiyerarşik kümeleme yöntemlerinin en genel formülü aşağıdaki gibidir.

$$d[(k), (r, s)] = \alpha_1 d[(k), (r)] + \alpha_s d[(k), (s)] + \beta d[(r), (s)] + \gamma |d - [(k), (r)] - d[(k), (s)]|$$

En yaygın algoritmalar için parametre değerleri Çizelge 4.1.2.1’de gösterilmiştir.

Çizelge 4.3. Matris güncelleme algoritmaları için katsayı değerleri

Kümeleme Yöntemleri	α_r	α_s	β	γ
Tek Bağlantı	1/2	1/2	0	-1/2
Tam Bağlantı	1/2	1/2	0	1/2
UPGMA(grup ortalaması)	$\frac{n_r}{n_r + n_s}$	$\frac{n_r}{n_r + n_s}$	0	0
WPGMA(ağırlıklandırılmış ortalama)	1/2	1/2	0	0
UPGMC(ağırlıklandırılmış imiş merkezi)	$\frac{n_r}{n_r + n_s}$	$\frac{n_r}{n_r + n_s}$	$\frac{-n_r n_s}{(n_r + n_s)^2}$	0
WPGMC(ağırlıklandırılmış merkezi)	1/2	1/2	-1/4	0
Ward's Yöntemi	$\frac{n_r + n_k}{n_r + n_s + n_k}$	$\frac{n_s + n_k}{n_r + n_s + n_k}$	$\frac{-n_k}{n_r + n_s + n_k}$	0

PGM, ikili grup yöntemini ifade eder. U ve W sırasıyla ağırlıklandırılmamış ve ağırlıklandırılmış olduğunu gösterir. Ağırlıklandırılmamış yöntemde kümedeki her birime eşit davranılır. Ağırlıklandırılmış yöntemde tüm

kümeler aynı ağırlıklandırılır. Böylece küçük kümedeki birimler büyük kümelerdeki birimlerden daha fazla ağırlıklandırılırlar. A, aritmetik ortalama ve C'de merkez anlamındadır. UPGMA, aritmetik ortalama kullanan ağırlıklandırılmamış ikili grup yöntemini, UPGMC, merkezi yöntemi ve WPGMC, medyan yöntemini ifade etmektedir. Tek ve tam bağlantı yöntemlerinin sapan değerlerinden sakınmak için aritmetik ortalama kullanılır. İki kümedeki en yakın ikili birimlerin alınması tek bağlantı yöntemi, var olan ve olası küme arasındaki uzaklığı ölçerken en uzak çiftin alınması tam bağlantı yöntemidir.

UPGMA ve WPGMA yöntemleri uzaklığın aritmetik ortalamasını kullanır. Aritmetik ortalamanın basit geometrik yorumu yoktur. Tersine, birimler d -boyutlu uzayda örüntü olarak gösterildiğinden, UPGMC ve WPGMC yöntemleri doğrudan geometrik yorum içerir. Merkezi yöntemler merkezler arası uzaklıkla iki küme arasındaki uzaklığı inceler. UPGMC yöntemi her bir kümedeki tüm örneklerden merkezi hesaplayarak uzaklığı ölçer. WPGMC yöntemi, yeni küme oluşturmak için iki kümenin merkezinden merkezler hesaplar. UPGMA, her bir kümenin boyutunu ele alarak her bir örüntünün katkısını eşit derecede ağırlıklandırır oysaki WPGMC küçük kümelerdeki örüntüyü büyük kümelerdeki örüntüden daha fazla ağırlıklandırır. Merkezi yöntemler yalnızca birimler örüntü olarak gösterildiğinde ve uzaklık ölçümü öklit uzaklığının karesiyken kullanılır. Merkezi yöntemler ve diğer SAHN algoritmaları arasındaki en önemli fark monotonluktur.

Örnek 1. Aşağıda üç boyutlu uzayda tanımlanan altı birimli bir vektör verilmiştir.

$$\begin{aligned} x_1 &= (1 \ 2 \ 2)^T & x_4 &= (3 \ 4 \ 3)^T \\ x_2 &= (2 \ 1 \ 2)^T & x_5 &= (0 \ 3.5 \ 3.5)^T \\ x_3 &= (0 \ 1 \ 3)^T & x_6 &= (2 \ 2.5 \ 2.5)^T \end{aligned}$$

$$X = \begin{bmatrix} 1 & 2 & 0 & 3 & 0 & 2 \\ 2 & 1 & 1 & 4 & 3.5 & 2.5 \\ 2 & 2 & 3 & 3 & 3.5 & 2.5 \end{bmatrix}$$

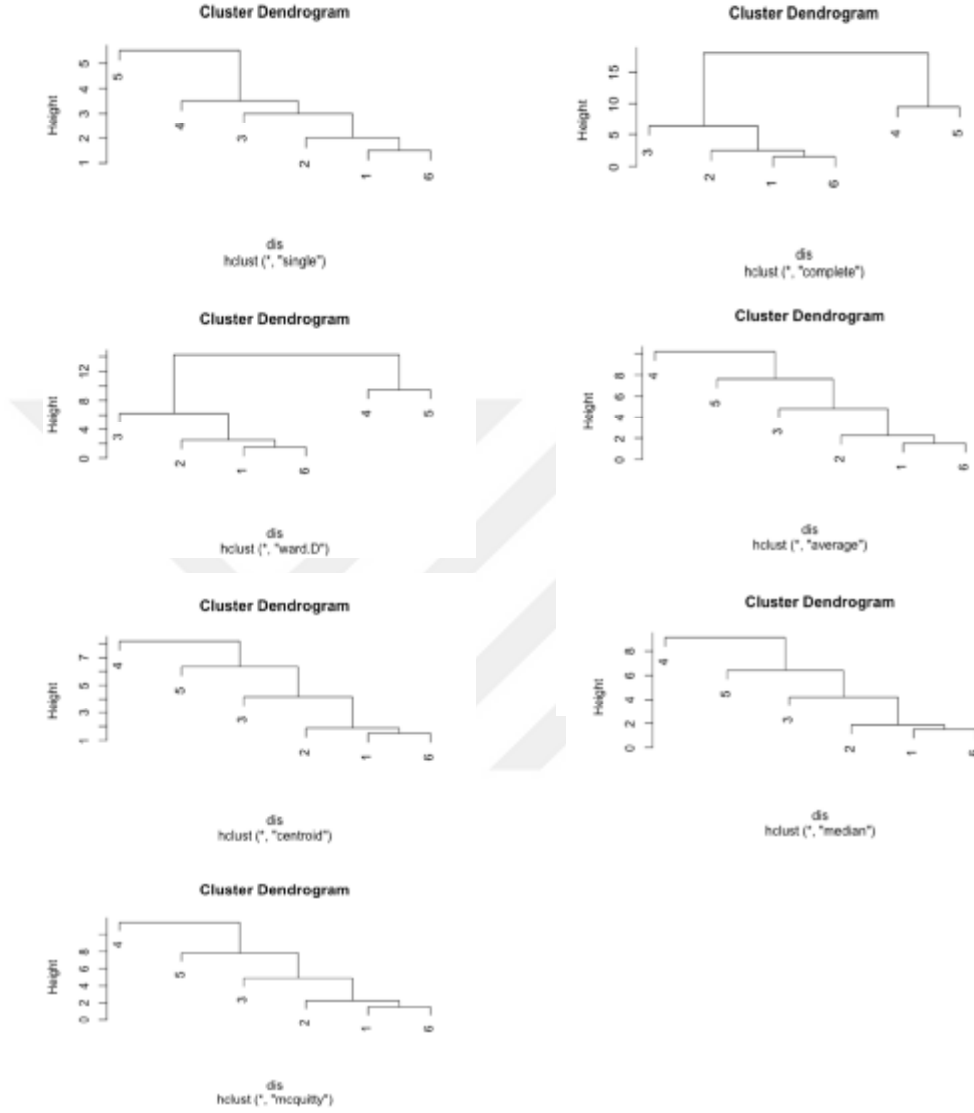
Eşitlik (4.2) kullanılarak her ikili çiftin uzaklıkları elde edilir. Aşağıda öklit uzaklığının karesi ile $d(x_1, x_2)$ ikili çifti için uzaklık elde edilmiştir.

$$d(x_1, x_2) = (1 - 2)^2 + (2 - 1)^2 + (2 - 2)^2 = 2$$

Tüm ikili çiftlere Eşitlik (4.2) uygulanarak uzaklık matrisi aşağıdaki gibi elde edilir.

$$\mathfrak{D} = \begin{matrix} & \begin{matrix} 1 & 2 & 3 & 4 & 5 & 6 \end{matrix} \\ \begin{matrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \end{matrix} & \begin{bmatrix} 0 & 2 & 3 & 9 & 5.5 & 1.5 \\ & 0 & 5 & 11 & 12.5 & 2.5 \\ & & 0 & 18 & 6.5 & 6.5 \\ & & & 0 & 9.5 & 3.5 \\ & & & & 0 & 6 \\ & & & & & 0 \end{bmatrix} \end{matrix}$$

(x_3, x_5) ve (x_3, x_6) örnek çiftleri arasındaki uzaklık bağlantısı bir dendogramda belirsizliğe neden olmaz.



Şekil 4.3. Matris güncellenme algoritmaları için dendrogram örnekleri

Ward yöntemi, iki aritmetik ortalama yöntemi (UPGMA, WPGMA) ve iki merkezi yöntem (UPGMC, WPGMC) için hazırlanan dendrogramlar aynı topolojiye sahiptir ancak düzeyleri farklıdır. x_4 sapan değerdir çünkü kümeye en son dahil edilmiştir. Tam bağlantı dendrogramında (x_4, x_5) kümelenmiştir. (x_1, x_2, x_6)

kümesi tüm dendogramlarda güçlü bir kümedir. Ward yöntemi minimum varyans yöntemi olarak da adlandırılır. Ward yöntemi verinin örneği olarak görüldüğü durumlar için tasarlanmıştır. Böylece d -boyutlu uzayda örneği ile başlanır.

x_{ij}^k , i . örneğin j . özelliği olsun. Burada k . kümedeki i örneği 1'den n_k 'ya ve j 'de 1'den d 'ye değer alır. $[m_1^{(k)}, \dots, m_d^{(k)}]$ olarak gösterilen k küme ağırlık merkezi, k . kümesindeki küme merkezi ya da n_k örneğinin ortalamasıdır.

$$m_j^{(k)} = \frac{1}{n_k} \sum_{i=1}^{n_k} x_{ij}^k$$

k küme için hata kareleri, k kümesindeki tüm örneğin ağırlık merkezine uzaklıklarının karesi toplamıdır.

$$e_k^2 = \sum_{i=1}^{n_k} \sum_{j=1}^d [x_{ij}^k - m_j^{(k)}]^2$$

K tane kümenin hata kareleri, bireysel kümelerin hata kareler toplamıdır.

$$E_k^2 = \sum_{k=1}^K e_k^2$$

Ward yöntemi ΔE_{pq}^2 'yi minimize eden küme çiftlerini birleştirir. p ve q kümelerinin t kümesi ile birleşmesiyle oluşan sonraki küme, E_k^2 'de değişikliğe neden olacaktır. Bu üç küme hariç tutulursa tüm kümeler için hata kareler aynı kalır.

$$\Delta E_{pq}^2 = e_t^2 - e_p^2 - e_q^2$$

Birkaç cebirsel işlemden sonra sadece merkeze bağlı hata karelerdeki değişim bulunur.

$$\Delta E_{pq}^2 = \frac{n_p n_q}{n_p + n_q} \sum_{j=1}^d [m_j^{(p)} - m_j^{(q)}]^2 \quad (4.15)$$

Bağlantı için seçilen p ve q kümeleri bu miktarı minimize eden kümelerdir. Küme sayısı azaldıkça hata kareler artar. Fakat Ward yöntemindeki artış mümkün olduğu kadar küçüktür. Oluşan yeni küme uzaklığı;

$$d[(r), (t)] = \frac{n_r + n_p}{n_r + n_t} d[(r), (p)] + \frac{n_r + n_q}{n_r + n_t} d[(r), (q)] - \frac{n_r}{n_r + n_t} d[(p), (q)] \quad (4.16)$$

ile hesaplanır (Jain ve Dubes, 1998).

Örnek 1 genel matris güncelleme yöntemlerinden biri olan Ward yöntemi ile kümelendirilsin. $d(x_1, x_6) = 1.5$ minimum değere sahip olduğundan 1 ve 6. küme birleştirilir. Uzaklık matrisi Eşitlik (4.16)'daki formül kullanılarak tekrar oluşturulur.

$$\begin{aligned} r=2, p=1, q=6 \text{ için;} \quad d[2, (1,6)] &= \frac{2}{3}x2 + \frac{2}{3}x2.5 - \frac{1}{3}x1.5 = 2.5 \\ r=3, p=1, q=6 \text{ için;} \quad d[3, (1,6)] &= \frac{2}{3}x3 + \frac{2}{3}x6.5 - \frac{1}{3}x1.5 = 5.8 \\ r=4, p=1, q=6 \text{ için;} \quad d[4, (1,6)] &= \frac{2}{3}x9 + \frac{2}{3}x3.5 - \frac{1}{3}x1.5 = 7.8 \\ r=2, p=1, q=6 \text{ için;} \quad d[5, (1,6)] &= \frac{2}{3}x5.5 + \frac{2}{3}x6 - \frac{1}{3}x1.5 = 7.1 \end{aligned}$$

$$\mathfrak{D} = \begin{matrix} & \begin{matrix} (1,6) & 2 & 3 & 4 & 5 \end{matrix} \\ \begin{matrix} (1,6) \\ 2 \\ 3 \\ 4 \\ 5 \end{matrix} & \begin{bmatrix} 0 & \textcircled{2.5} & 5.8 & 7.8 & 7.16 \\ & 0 & 5 & 11 & 12.5 \\ & & 0 & 18 & 6.5 \\ & & & 0 & 9.5 \\ & & & & 0 \end{bmatrix} \end{matrix}$$

$d(x_2, (x_1, x_6)) = 2.5$ minimum değerine sahip olduğundan dolayı 1 ve 6'yı birleştiren kümeye 2'de dahil edilerek yukarıdaki işlemler tek bir küme elde edilinceye kadar işlem devam ettirilir.

$$d[3, (1,6,2)] = \frac{3}{4}x5.8 + \frac{2}{4}x5 - \frac{1}{4}x2.5 = 6.22$$

$$d[4, (1,6,2)] = \frac{3}{4}x7.8 + \frac{2}{4}x11 - \frac{1}{4}x2.5 = 10.7$$

$$d[5, (1,6,2)] = \frac{3}{4}x7.16 + \frac{2}{4}x12.5 - \frac{1}{4}x2.5 = 10.9$$

$$\mathfrak{D} = \begin{array}{c} ((1,6),2) \\ 3 \\ 4 \\ 5 \end{array} \begin{array}{c} ((1,6),2) \\ 3 \\ 4 \\ 5 \end{array} \begin{array}{c} 3 \\ 4 \\ 5 \end{array} \begin{array}{c} 4 \\ 5 \end{array} \begin{array}{c} 5 \end{array} \begin{bmatrix} 0 & 6.22 & 10.7 & 10.9 \\ & 0 & 18 & 6.5 \\ & & 0 & 9.5 \\ & & & 0 \end{bmatrix}$$

$d(x_3, (x_1, x_6, x_2)) = 6.22$ minimum değerine sahip olduğundan 1, 6, ve 2'nin birleştiği kümeye 3'te eklenir.

$$d[4, (1,6,2,3)] = \frac{4}{5}x10.7 + \frac{2}{5}x18 - \frac{1}{5}x6.22 = 14.51$$

$$d[5, (1,6,2,3)] = \frac{4}{5}x10.9 + \frac{2}{5}x6.5 - \frac{1}{5}x6.22 = 10.0$$

$$\mathfrak{D} = \begin{array}{c} (1,6,2,3) \\ 4 \\ 5 \end{array} \begin{array}{c} (1,6,2,3) \\ 4 \\ 5 \end{array} \begin{array}{c} 4 \\ 5 \end{array} \begin{array}{c} 5 \end{array} \begin{bmatrix} 0 & 14.51 & 10.07 \\ & 0 & 9.5 \\ & & 0 \end{bmatrix}$$

$d((x_4, x_5), (x_1, x_6, x_2, x_3)) = 9.5$ minimum değerine sahip olduğundan 1, 6, 2 ve 3'ün birleştiği kümeye 4 ve 5 eklenir.

$$d[(4,5)(1,6,2,3)] = \frac{5}{6} \times 10.7 + \frac{3}{6} \times 6.5 - \frac{2}{6} \times 6.22 = 10.09$$

$$\mathfrak{D} = \begin{bmatrix} 0 & 10.09 \\ & 0 \end{bmatrix}$$

Tek küme altında toplanır ve işlemler sonlandırılır (Jain ve Dubes, 1988).

4.2. Hiyerarşik Olmayan Kümeleme Algoritmaları

Küme sayısının önceden bilindiği durumlarda araştırmacı küme sayısına göre çözümler üretir. Küme sayısı konusunda ön bilgi varsa veya araştırmacı anlamlı olacak küme sayısına karar vermiş ise bu durumda, çok uzun zaman alan hiyerarşik yöntemler yerine hiyerarşik olmayan yöntemler kullanılmaktadır (Anderberg, 1973).

Hiyerarşik olmayan teknikler ya gruplardaki birimlerin ilk parçalarından ya da kümelerin çekirdeklerinden oluşan çekirdek noktaların ilk setiyle işe başlar. İlk gruplaşmalar için seçim serbesttir. Başlangıç yollarından biri birimler arasındaki çekirdek noktaların rasgele seçilmesi veya ilk gruplardaki birimlerin rasgele parçalanmasıdır (Johnson ve Wichern, 1998).

Bu bölümde ele k-ortalamlar, k-modlar ve K prototip algoritmaları ele alınacaktır.

4.2.1. k-ortalamlar Algoritması

En yaygın kullanılan kümeleme algoritmalarından birisi olan k-ortalamlar bölmeli yada hiyerarşik olmayan kümeleme yöntemi olarak sınıflandırılır. $k (\leq n)$ tam sayı ve X nümerik eleman seti verilsin, K ortalamlar algoritması, X adet birimden oluşan veri setini, grup içi hata kareler toplamını minimize edecek şekilde giriş parametresi olarak verilen k adet kümeye bölmeyi amaçlar. Süreç sıklıkla matematiksel programlama problemi olarak P ile aşağıdaki gibi ifade edilir.

$$P(W, Q) = \sum_{l=1}^k \sum_{i=1}^n w_{i,l} d(X_i, Q_l) \quad (4.17)$$

$$\sum_{l=1}^n w_{i,l} = 1, \quad 1 \leq i \leq n$$

$$w_{i,l} \in \{0,1\}, \quad 1 \leq i \leq n, \quad 1 \leq l \leq k$$

Burada W , $n \times k$ boyutlu ayırma matrisidir. $Q = \{Q_1, Q_2, \dots, Q_k\}$ aynı gruptaki birimlerin setidir ve $d(., .)$ iki birim arasındaki öklid uzaklığının karesidir. $X = \{X_1, X_2, \dots, X_n\}$, n birimli settir ve $X_i = [x_{i,1}, x_{i,2}, \dots, x_{i,1}]$ 'dir. $1 \leq j \leq m$ için $x_{i,j} = x_{k,j}$ ise $X_i = X_k$ 'dir. Fakat bu durum gerçek dünya veritabanında X_i ve X_k 'nin aynı elemanlar olduğu anlamına gelmez. İki elemanın $A_1, A_2, \dots, A_m$ değişkenleri için eşit değere sahip olduğu anlamına gelir.

Aşağıdaki iki problemi iteratif olarak çözmek P problemine çözüm getirebilir.

1. Problem P_1 : $Q = \hat{Q}$ alınır ve indirgenmiş $P(W, \hat{Q})$ çözülür.
2. Problem P_2 : $W = \hat{W}$ alınır ve indirgenmiş $P(\hat{W}, Q)$ çözülür.

Problem P_1 aşağıdaki gibi çözülür.

$$w_{i,l} = 1 \text{ eğer } d(X_i, Q_l) \leq d(X_i, Q_t), \quad 1 \leq t \leq k \text{ için}$$

$$w_{i,t} = 0 \quad t \neq l \text{ için}$$

Ve problem P_2 de aşağıdaki gibi çözülür.

$$q_{l,j} = \frac{\sum_{i=1}^n w_{i,l} x_{i,j}}{\sum_{i=1}^n w_{i,l}} \quad 1 \leq l \leq k \text{ ve } 1 \leq j \leq m \text{ için} \quad (4.18)$$

P problemini çözmek için temel algoritma aşağıdaki gibidir.

1. Başlangıç Q^0 seçilir ve W^0 elde etmek için $P(W, Q^0)$ çözülür. $t=0$ olarak alınır.
2. $\hat{W} = W^t$ olsun ve Q^{t+1} elde etmek için $P(\hat{W}, Q)$ çözülür. $P(\hat{W}, Q^t) = P(\hat{W}, Q^{t+1})$ ise sonuç $\hat{W}, Q^t$ 'dir ve durulur. Değilse adım 3'e geçilir.
3. $\hat{Q} = Q^{t+1}$ olsun ve W^{t+1} elde etmek için $P(W, \hat{Q})$ çözülür. $P(W^t, \hat{Q}) = P(W^{t+1}, \hat{Q})$ ise sonuç $W^t, \hat{Q}$ 'dir ve durulur. Değilse $t=t+1$ alınır ve adım 2'ye gidilir.

k-ortalamlar algoritmasının önemli özellikleri vardır. Bunlar;

1. Büyük veri setleri için etkindir.
2. Sıklıkla, yerel en iyide sona erer.
3. Yalnızca nümerik değerler için kullanılır.
4. Kümeler konveks yapıya sahiptir (Huang, 1998).

4.2.2. k-modlar Algoritması

k-modlar algoritması Huang (1997) tarafından geliştirilmiştir. k-modlar kümeleme yöntemi, büyük kategorik verilerin kümelenmesinde kullanılan ilk algoritmalarından biridir. k-ortalamlar yönteminin kategorik özellikli birimleri kümelememesinin nedeni benzemezlik ölçümüdür ve P_2 probleminin çözümünde kullanılan yöntemdir. k-ortalamalara aşağıdaki değişimler uygulanarak bu olumsuzluklar ortadan kaldırılır.

1. Kategorik özellikli birimler için basit eşleştirme benzemezlik ölçümü kullanılır.
2. Küme ortalamlarının yerini modlar alır.
3. P_2 problemini çözmek ve model kurmak için sıklık temelli yöntem kullanılır.

X ve Y , m adet kategorik değişkenle tanımlanan iki kategorik birim olsun. X ve Y arasındaki benzemezlik (uzaklık) ölçümü iki birimin değişken kategorilerine karşılık gelen toplam uyumsuzluk olarak tanımlanabilir. İki birim arasındaki daha küçük uyumsuzluklar daha büyük benzerliktir. Bu ölçüm genellikle basit eşleştirmedir. Genel olarak,

$$d_1(X, Y) = \sum_{j=1}^m \delta(x_j, y_j) \quad (4.19)$$

şeklindedir. Burada,

$$\delta(x_j, y_j) = \begin{cases} 0 & (x_j = y_j) \\ 1 & (x_j \neq y_j) \end{cases}$$

olarak ifade edilir. $X, A_1, A_2, \dots, A_m$ kategorik değişkenler tarafından tanımlanan kategorik birimlerin seti olsun.

$X = \{X_1, X_2, \dots, X_n\}$ 'nin modu $Q = [q_1, q_2, \dots, q_m]$ vektörüdür.

$$D(X, Q) = \sum_{i=1}^n d_1(X_i, Q) \quad (4.20)$$

Burada Q , X 'in elemanı olması zorunlu değildir.

A_j değişkeninde k . kategori $c_{k,j}$ 'ye sahip birimlerinin sayısı $n_{c_{k,j}}$ olsun.

$$f_r(A_j = c_{k,j} \setminus X) = \frac{n_{c_{k,j}}}{n} X \text{te } c_{k,j} \text{ kategorisinin nispi sıklığıdır.}$$

Kategorik birimler için benzemezlik ölçümü olarak Eşitlik (4.19) kullanılır. Maliyet fonksiyonu;

$$P(W, Q) = \sum_{l=1}^k \sum_{i=1}^n \sum_{j=1}^m w_{i,l} \delta(x_{i,j}, q_{l,j}) \quad (4.21)$$

Burada $w_{i,l} \in W$ ve $Q_1 = [q_{1,1}, q_{1,2}, \dots, q_{1,m}] \in Q$ 'dur.

P_1 problemini çözmek için basit eşleştirme benzemezlik ölçümü ve P_2 problemini çözmek için ortalama yerine mod kullanarak K ortalamalar modifiye edilir ve amaç fonksiyonu minimize edilir.

Algoritmanın temelinde yeni W ve Q elde edildiğinde P 'nin hesaplanması gereklidir. Daha etkin hesaplama yapılabilmesi için aşağıdaki algoritma kullanılır.

1. Her bir küme için k başlangıç modu seçilir.
2. Birim, Eşitlik (4.18)'e göre en yakın moda sahip kümeye atanır. Her bir atamadan sonra kümenin modu güncellenir.
3. Birimler kümelere atandıktan sonra şu anki modlara göre birimlerin benzemezlikleri tekrar test edilir. Birim, şu anki küme yerine en yakın modlu kümeye ait olacak şekilde bulunursa, birim o kümeye yeniden atanır ve her iki kümenin modları güncellenir.
4. Kümelerdeki birimler sabit olana kadar adım 3 tekrarlanır (Huang, 1998).

4.2.3. k-prototip Algoritması

Karma tipteki birimleri kümelemek için kullanılan k-prototip algoritması içine k-ortalamalar ve k-modlar algoritmalarını entegre etmek kolaydır. Gerçek dünya veri tabanlarında sıkça karşılaşılan birimler karma tip birimler oldukları için k-prototip algoritması pratikte daha yararlıdır.

$A_1^r, A_2^r, \dots, A_p^r, A_{p+1}^c, \dots, A_m^c$ değişkenleri ile tanımlanan iki karma tip birim X ve Y olmak üzere, bu birimler arasındaki benzemezlik Eşitlik (4.22) ile ölçülür.

$$d(X, Y) = \sum_{j=1}^p (x_j - y_j)^2 + \gamma \sum_{j=p+1}^m \delta(x_j, y_j) \quad (4.22)$$

Burada birinci toplam, nümerik özellikli değişkenler için öklid uzaklığının karesidir. İkinci toplam, kategorik özellikli değişkenler için basit eşleştirme uzaklık

(benzemezlik) ölçümüdür. Karma veri tipleri için Eşitlik (4.22) kullanılarak amaç fonksiyonu (Eşitlik (4.17)) aşağıdaki gibi geliştirilir.

$$P(W, Q) = \sum_{l=1}^k \left(\sum_{i=1}^n w_{i,l} \sum_{j=1}^p (x_{i,j} - q_{l,j})^2 + \gamma \sum_{i=1}^n w_{i,l} \sum_{j=p+1}^m \delta(x_{i,j}, q_{j,l}) \right) \quad (4.23)$$

$$P_l^r = \sum_{i=1}^n w_{i,l} \sum_{j=1}^p (x_{i,j} - q_{l,j})^2 \quad (4.24)$$

$$P_l^c = \sum_{i=1}^n w_{i,l} \sum_{j=p+1}^m \delta(x_{i,j}, q_{j,l}) \quad (4.25)$$

olsun. Eşitlik yeniden yazılırsa,

$$P(W, Q) = \sum_{l=1}^k (P_l^r + P_l^c)$$

olur. Hem P_l^r hem de P_l^c negatif olmadığından, $P(W, Q)$ 'yu minimize etmek $1 \leq l \leq k$ için P_l^r ve P_l^c 'yi minimize etmeye eşittir.

Yerel en iyi Q^*, W^{**} bulmak için $d(.,.)$ hariç başka bir değişiklik olmadığından aynı k-ortalamalar algoritması kullanılır. $\hat{Q}$ verildiğinde W 'yi hesaplamak için k-ortalamalar algoritmasındaki aynı yol ile Eşitlik (4.23) kullanılır. $\hat{W}$ verildiğinde $1 \leq l \leq k$ için P_l^r ve P_l^c 'yi minimize ederek Q bulunur. Eşitlik (4.18) ile $q_{l,j}$ hesaplanırsa P_l^r minimize olur. k-modlardan $p + 1 \leq j \leq m$ için $q_{l,j}$ seçilerek P_l^c minimize edilir (Huang, 1998).

4.3. Küme Sayısı Tahmini

Kümeleme analizinde uygun küme sayısının bulunması önemli bir sorundur. Bazı araştırmacılar, literatürde bu zorluğun üstesinden gelmek için yeni yöntemler önermişlerdir. En seçkin yöntemlerden ikisi özetle tanıtılırsa;

1. Calinski ve Harabasz (1974) İndeks: Calinski-Harabasz indeksi Eşitlik (4.26) ile tanımlanır.

$$CH(K) = \frac{\frac{B(K)}{(K-1)}}{\frac{W(K)}{N-K}} \quad (4.26)$$

burada N , veri noktalarının sayısıdır. Küme sayısı Eşitlik (4.26)'te verilen fonksiyon maksimize edilerek seçilir. $B(K)$, kümeler arası kareler toplamını ve $W(K)$, kümeler içi kareler toplamını vermektedir.

2. Duda ve Hart (1973) İndeks: Duda-Hart Je (2) / Je (1) durdurma kural indeksi değeri tam anlamıyla Je (2)'nin Je (1)'e oranıdır. Je (1), bölünecek grup içi hataların kareleri toplamıdır. Je (2) ise sonuçta ortaya çıkan iki alt gruptaki hataların kareleri toplamıdır.

4.4. Küme Geçerlilikleri

Kümeleme analizinde kullanılan farklı algoritmalarla veya aynı algoritmada alınan farklı parametre başlangıç değeri ile aynı verinin kümelenmesinde farklı birçok kümeleme yapısı elde edilebilir. Aynı veri kümesinden elde edilen farklı kümeleme gruplarından hangisinin geçerli olup olmadığı kümeleme geçerliliği olarak adlandırılır (Gordon 1998: Halkidi ve ark, 2002a, Halkidi ve ark, 2002b). Kümeleme analizinin sonuçlarının kalitesini değerlendirmek için küme geçerliliği yöntemlerine ihtiyaç vardır (Bolshakova ve Azuaje, 2003). Kümeleme algoritmaların tipine bağlı olarak dışsal, içsel ve oransal olmak üzere üç farklı tipte küme geçerlilik ölçütü vardır. Dışsal kriterlerde, veri hakkında sahip olunan ön bilgi ile kümeleme algoritması sonunda elde edilen kümeleme yapısı karşılaştırılır; kümeleme sonunda elde edilen gözlemlerin küme etkileri ile daha önceden bilgi sahibi olunan gözlemlerin kategori etiketleri karşılaştırılır. İçsel kriterlerde, veri seti ile kümeleme yapısı arasındaki uyumun

belirlenmesinde sadece veri setindeki doğal yapıyı ve nicel değerleri göz önünde bulundurarak kümeleme sonuçları değerlendirilir. İçsel kriterlerin çoğu kümeler içi kareler toplamını yada kümeler arası kareler toplamını temel alarak değerlendirme yapmaktadır (Theodoridis ve Koutroumbas, 2006). Oransal kriterler ise farklı kümeleme algoritmaları ya da aynı kümeleme algoritmasından farklı parametre değeri ile elde edilen kümeleme sonuçlarını karşılaştırır. İyi bir kümeleme ile aynı kümede yer alan birimler arasındaki benzerliğin olabildiğince fazla, farklı kümelerde yer alan birimler arasındaki benzerliğin ise olabildiğince az olması amaçlanmaktadır. Küme geçerliliği teknikleri küme içi ve kümeler arası değişimlerin farklı açılardan değerlendirilmesi ile birbirinden ayrılmaktadır (Bolshakova ve Azuaje, 2003). Kümeleme analizi ile ilgili çalışmalarda daha çok içsel kriterlerin küme geçerlilik ölçütü olarak kullanıldığı görülmektedir. Bunun sebebi olarak, içsel kriterlerin dışsal kriterlere göre doğal ve anlamlı küme yapıları oluşturmada daha başarılı sonuçlar vermesi gösterilebilir (Aydın ve Seven, 2015). Bu çalışmada içsel ve oransal geçerlilik ölçütleri kullanılacaktır.

4.4.1. Kofenetik Korelasyon Katsayısı

Kofenetik korelasyon katsayısı (CPCC) hiyerarşik kümeleme yapılarının geçerliliğinde kullanılan bir indekstir. X 'in uzaklık matrisi $P = \{p_{ij}\}$ verildiğinde, CPCC aynı kümede gruplanan iki veri noktasının uzaklık seviyesini kaydeden elemanlardan oluşan kofenetik matris $Q = \{q_{ij}\}$ ile P arasındaki benzerliğin derecesini ölçer. μ_p ve μ_q sırasıyla P ve Q 'nun ortalamaları olmak üzere,

$$\mu_p = \frac{1}{M} \sum_{i=1}^{N-1} \sum_{j=i+1}^N p_{ij} \quad (4.27)$$

$$\mu_q = \frac{1}{M} \sum_{i=1}^{N-1} \sum_{j=i+1}^N q_{ij} \quad (4.28)$$

şeklindedir. Burada $M=N(N-1)/2$ 'dir ve CPCC aşağıdaki gibi tanımlanır.

$$CPCC = \frac{\frac{1}{M} \sum_{i=1}^{N-1} \sum_{j=i+1}^N p_{ij} q_{ij} - \mu_p \mu_q}{\sqrt{(\frac{1}{M} \sum_{i=1}^{N-1} \sum_{j=i+1}^N p_{ij}^2 - \mu_p^2)(\frac{1}{M} \sum_{i=1}^{N-1} \sum_{j=i+1}^N q_{ij}^2 - \mu_q^2)}} \quad (4.29)$$

Kofenetik korelasyon katsayısı $[-1,1]$ aralığında değer alır ve 1'e ne kadar yakın değer alırsa kümeleme de o kadar başarılıdır (Xu ve Wunsch, 2009).

4.4.2. Bağlanırlık

N veri setindeki toplam gözlem sayısını (satırları), M 'de toplam sütun sayısını gösterir. $nn_{i(j)}$, i . gözlemin en yakın komşusu j . olarak tanımlanır ve eğer i ile j aynı kümedeyse sıfır, değilse $\frac{1}{j}$ değerini alır. Belirli bir kümeleme bölümü için K ayrı kümedeki N gözlemin $\mathcal{C} = \{C_1, C_2, \dots, C_K\}$ 'nin birbirinden uzak kümelere bölünmesi durumunda, bağlanırlık eşitlikteki gibi tanımlanır.

$$Conn(\mathcal{C}) = \sum_{i=1}^N \sum_{j=1}^L x_{i,nn_{i(j)}} \quad (4.30)$$

Burada L , kullanılan en yakın komşulukların sayısını veren parametredir. Bağlanabilirlik sıfır ve ∞ aralığındadır ve minimum değer almalıdır (Brock ve ark, 2008).

4.4.3. Silhouette Genişliği

Rousseeuw (1987), her bir birimin kendi kümesine uygunluğunu tanımlayacak bir Silhouette indeksi önermiştir. Silhouette genişliği her bir gözlemin Silhouette değerinin ortalamasıdır. Silhouette değeri kümelere atanan gözlemlerin güvenilirlik derecesini ölçer. İyi kümelenmiş gözlemler 1'e yakın değerler alırken zayıf kümelenmiş gözlemler -1'e yakın değerler alırlar. Negatif değerler yalnızca bir birim en uygun kümesine atanmadığında ortaya çıkar. i . gözlem için Silhouette değeri,

$$S(i) = \frac{b_i - a_i}{\max(b_i, a_i)} \quad (4.31)$$

Olarak ifade edilir. Burada a_i , i . birimin kendi kümesindeki tüm noktalara olan ortalama uzaklığını (benzerliğini) ve b_i , i . birimin diğer kümelerdeki tüm noktalara olan ortalama uzaklığını gösterir.

$$b_i = \min_{C_k \in \mathcal{C} \setminus C(i)} \sum_{j \in C_k} \frac{\text{dist}(i, j)}{n(C_k)} \quad (4.32)$$

Burada $C(i)$, i . gözlemi içeren kümedir. $\text{dist}(i, j)$ ise i ve j gözlemleri arasındaki uzaklıktır (Öklid, Manhattan vb.). $n(C)$, C kümesinin niceliğidir. Silhouette genişliği $[-1, 1]$ aralığında değer alır ve maksimum olmalıdır (Rousseeuw 1987: Brock ve ark, 2008).

4.4.4. Dunn İndeks

Dunn indeks (Dunn, 1974), yüksek yoğunluklu iyi ayrımlı kümelerin belirlenmesinde oldukça etkin bir kriterdir. Sırasıyla C_i ve C_j kümelerinde yer alan x ve y nokta çiftleri arasındaki minimum uzaklık olarak C_i ve C_j kümeleri arasındaki uzaklık $D(C_i, C_j)$,

$$D(C_i, C_j) = \min_{x \in C_i, y \in C_j} D(x, y) \quad (4.33)$$

olarak tanımlanır ve ikiyelemanı arasındaki maksimum uzaklık olarak C_i kümesinin çapı $\text{çap}(C_i)$ 'dir.

$$\text{çap}(C_i) = \max_{x, y \in C_i} D(x, y) \quad (4.34)$$

Dunn indeks tekrar yapılandırılırsa,

$$Du(K) = \min_{i=1,\dots,K} \left(\min_{j=i+1,\dots,K} \left(\frac{D(C_i, C_j)}{\max_{l=1,\dots,K} \text{çap}(C_l)} \right) \right) \quad (4.35)$$

şeklinde ifade edilir. Dunn indeks değeri daha büyük olan kümeleme, geçerli kümeleme olarak kabul edilir (Xu ve Wunsch, 2009).



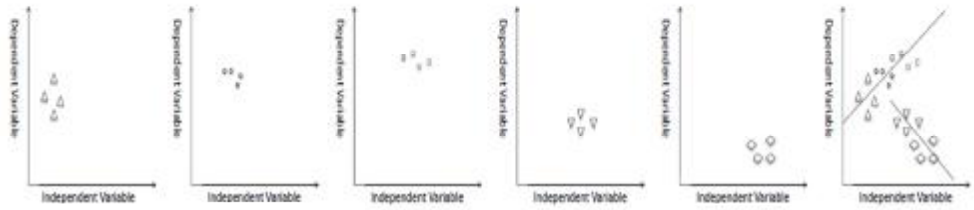
5. PANEL VERİLERDE KÜMELEME ANALİZİ

Panel veri zaman ve yatay kesit içeren iki boyutlu veridir (Zheng ve ark, 2009). Panel veriyle model kurmak, zaman serisi verisi veya yatay kesit verileriyle model kurmaktan çok daha avantajlıdır. Özellikle panel veri gözlemlenmemiş bireysel etkileri veya aynı zamandaki zaman etkilerini tahmin edebilir. Bu nedenle panel veri modeli daha etkilidir (Hou ve Ai, 2015).

Panel veride önemli konulardan biri verideki farklı birimlerin havuzlanabilirliğidir. Bu alandaki klasik lineer regresyon modellerinde, regresyondaki parametrelerin farklı birimlere karşı homojen olduğu düşünülürse farklı birimler hep birlikte havuzlanabilir olacaktır. Fakat genellikle yüksek heterojenlikten dolayı birimler havuzlanamayabilir. Son zamanlardaki bazı çalışmalar farklı gruptaki birimleri kümeleyerek kısmi havuzlanabilirliği araştırmışlardır (Lu ve Huang, 2011).

Panel veri grubu $k \in [1, p]$, $i \in [1, n]$ ve $t \in [1, T]$ 'li $\{Y_{it}, X_{kit}\}$ formuna sahiptir. X_{kit} , çok boyutlu bağımsız değişken vektörü ve Y_{it} bağımlı değişkendir. Panel veri grubunun her bir birimi için, farklı lineer fonksiyona veri noktaları uydurularak Y bağımlı değişkeni ve X bağımsız değişkeni arasındaki ilişki modellenabilir. Bu modeller Bölüm 3'te ayrıntılı bir biçimde ele alınmıştır. Eğer model Eşitlik (3.1)'deki gibi klasik etki modeli ise En Küçük Kareler Tahmin edicisi (EKK) α ve β parametre değerlerinin tahmini için kullanılan klasik yöntem olduğu, Bölüm 3'te belirtilmiştir. Gerçek dünyada panel veri analizinin zorluğu her bir birimdeki mevcut veri noktalarının genellikle çok sınırlı olmasıdır. Sonuç olarak, küçük örnekten tahmin edilen α ve β parametre değerleri güvenilir değildir. Örnek olarak Şekil (5.1a)'da sadece dört veri noktalı bağımlı ve bağımsız değişkenler arasında lineer ilişki önemli değildir. Bu durumda EKK tahmin edicisi yanılsamalı olabilir. Fakat beş değişken içeren panel veri havuzlanabilirse (Şekil (5.1a)-(5.1e)'deki) Şekil (5.1f)'de gösterildiği gibi iki lineer örüntü tanımlamak kolaydır. Bu küçük örnek kesin ve güvenilir tahmin elde etmek için farklı değişkenlerin

havuzlanabilirliğinin gerekliliğini öne sürer ve aynı zamanda bu örnek değişkenler heterojense panel veri kümelemenin önemli olduğunu gösterir. Şekil (5.1f)'de aynı doğruya uydurulan değişkenlerin kümelendiği ve homojen olduğu düşünülmektedir (Lu ve Huang, 2011).



(a)Değişken 1 (b)Değişken 2 (c)Değişken 3 (d)Değişken 4 (e)Değişken 5 (f)Kesit değişken
Şekil 5.1.Yatay kesit veri örneği (Lu ve Huang, 2011)

Regresyon modeli hazırlamak bağımlı değişkenin kümeleme sonuçları ve bağımsız değişkenin kümeleme sonuçlarını karşılaştırmak ve heterojen veri sorununu çözmek için regresyon analizinden önce panel veri kümelemesi yapılır (Zheng ve ark, 2009).

Panel verideki tüm bireysel birimlerin, ortak niceliklerini tahmin ederken esneklik kazanmak amacıyla bazı özelliklere göre gruplar oluşturulur. Bu gruplama, eğer birimlerin varsayılan ortak özellikleri oldukça farklıysa, sonuçları güçlü bir şekilde etkiler. Böyle durumlarda paneldeki birimleri kümelemek faydalı olabilir. Panel veri modellerinde birimler arasındaki parametre heterojenliğini gidermek için kümeleme analizi uygulanır.

Kümeleme analizinin, doğal sınıflamaları hakkında açık bilgi bulunmayan durumlarda, kitleye ilişkin tahminlerin yapılmasında yararlanılan bir yöntem olduğu Bölüm 4'te ele alınmıştır.

5.1. Tek Değişkenli Panel Veri Kümeleme Analizi

Kümeleme, bir sınıftaki birimin diğerleriyle en yakın benzerliğe ve her bir sınıf arasındaki birimlerin en az benzerliğe sahip olduğu k sınıftaki p boyutlu veriyi sınıflandırır. Panel veri kümelenirken genel kümeleme adımlarını uygulanabilir fakat panel veri özelliğinin farkında olunması gerekir (Zheng ve ark, 2009).

Tek değişkenli panel veri kümeleme iki yol ile elde edilebilir. İlk yol, tek değişkenli panel verinin zaman boyutunu yatay kesit verinin gösterge boyutuna çevirmektir. Bu iki veri türü benzer istatistik tanımına sahiptir ve birim uzaklığıyla ilgili algoritmalar kümeleme sürecinde hemen hemen aynıdır. Bu nedenle tek değişkenli panel veri kümeleme yatay kesit veri kümelemesi yöntemiyle yapılabilir ve ilgili yazılımlarla doğrudan hesaplanabilir. Diğer ise tek boyutlu örneklemedeki sıralı kümedir. Bu yol tek değişkenli panel verilerinin mekansal boyutunu, sıralı örneklemin gösterge boyutuna dönüştürmektir. Tek boyutlu örnek üzerinde sıralı kümenin hesaplanmasını tamamlamak için bazı profesyonel yazılımlar vardır (Ren ve Shi, 2009).

Tek değişkenli panel veri setinin aşağıdaki veri matrisinden oluştuğu varsayılın.

$$X = \begin{bmatrix} X_{1t_1} & X_{1t_2} & \cdots & X_{1t_p} \\ X_{2t_1} & X_{2t_2} & \cdots & X_{2t_p} \\ \vdots & \vdots & \ddots & \vdots \\ X_{nt_1} & X_{nt_2} & \cdots & X_{nt_p} \end{bmatrix}$$

n birim numarası, t_j zaman değişkenidir ve X_{it_j} , t_j zamandaki i . birimin veri değeridir.

Tek değişkenli panel veri kümeleme adımları aşağıdaki gibi verilmiştir.

1. Yatay kesit panel veri, aynı zaman yatay kesit gözlemlerindeki birimlerin sayısıdır. Veri dönüşümü gerekliyse, standartlaştırma veya normalleştirme dönüşümü devam etmelidir. Standartlaştırma dönüşüm eşitliği;

$$X_{it_j} = \frac{X_{it_j} - \bar{X}_{t_j}}{S_{t_j}} \quad i = 1, \dots, n; \quad t_j = t_1, \dots, t_p \quad (5.1)$$

$$S_{t_j} = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (X_{it_j} - \bar{X}_{t_j})^2} \quad (5.2)$$

olarak ifade edilmektedir. Normalleştirme dönüşüm eşitliği Eşitlik (5.3)'de verilmiştir.

$$X_{it_j} = \frac{X_{it_j} - \min_{1 \leq i \leq n} \{X_{it_j}\}}{\max_{1 \leq i \leq n} \{X_{it_j}\} - \min_{1 \leq i \leq n} \{X_{it_j}\}} \quad i = 1, \dots, n; \quad t_j = t_1, \dots, t_p \quad (5.3)$$

2. Kümeleme istatistiği hesaplanır. Öklid uzaklığı en yaygın kullanılan uzaklıktır. Panel veriye dayalı eşitlik;
- 3.

$$d_{is} = \sqrt{\sum_{j=1}^p (X_{it_j} - X_{st_j})^2} \quad i, s = 1, \dots, n \quad (5.4)$$

şeklindedir.

4. En yaygın kümeleme yöntemleri, gruplar arası bağlantı, gruplar içi bağlantı, tek bağlantı (en yakın komşuluk), tam bağlantı (en uzak komşuluk), merkezi kümeleme, medyan kümeleme ve Ward yöntemleridir. Önceki hesaplama sonuçlarına göre panel veri kümelemede en uygun yöntem seçilir (Zheng ve ark, 2009).

5.2. Çok Değişkenli Panel Veri Kümeleme Analizi

Çok değişkenli panel veri yapısı tek değişkenli panel veriden daha karmaşıktır. Üç boyutlu yapıda veya matris formunda açıklanarak gösterilebilir. Çizelge 5.2.1'de gösterildiği gibi, iki sınıf taşınarak tek boyuta devredilir.

Toplamdaki birim sayısının N olduğunu, her birimin özelliğinin p değişkeni ile gösterildiğini $(X_1, X_2, \dots, X_j, X_p)$ ve uzunluğun T olduğu varsayalım. Böylece $X_{ij}(t)$ t zamanındaki i . birimin j . değişken değeridir.

Çizelge 5.1. Çok değişkenli panel veri yapısı (Zheng ve Li, 2014)

Zaman	1	...	T
İndeks	$X_1 \dots X_j \dots X_p$	...	$X_1 \dots X_j \dots X_p$
Birim			
1	$X_{11}(1) \dots X_{1j}(1) \dots X_{1p}(1)$	...	$X_{11}(T) \dots X_{1j}(T) \dots X_{1p}(T)$
	$X_{21}(1) \dots X_{2j}(1) \dots X_{2p}(1)$	...	$X_{21}(T) \dots X_{2j}(T) \dots X_{2p}(T)$
2	...	...	...
...	$X_{i1}(1) \dots X_{ij}(1) \dots X_{ip}(1)$	...	$X_{i1}(T) \dots X_{ij}(T) \dots X_{ip}(T)$
i	...	...	...
...	$X_{N1}(1) \dots X_{Nj}(1) \dots X_{Np}(1)$	...	$X_{N1}(T) \dots X_{Nj}(T) \dots X_{Np}(T)$
N			

Çok değişkenli panel verinin bazı istatistikleri verilsin.

$i \in [1, n]$, $j \in [1, p]$ ve $t \in [1, T]$ olmak üzere bu istatistikler kümeleme analizinde kullanılmaktadır.

t . zamandaki j . değişkenin ortalaması;

$$\bar{X}_j(t) = \frac{1}{N} \sum_{i=1}^N X_{ij}(t) \quad (5.5)$$

j . değişkenin ortalaması;

$$\bar{X}_j = \frac{1}{T} \sum_{t=1}^T \sum_{i=1}^N X_{ij}(t) \quad (5.6)$$

t . zamandaki j . değişkenin varyansı;

$$Var_{X_j(t)} = \frac{1}{N} \sum_{i=1}^N [X_{ij}(t) - \bar{X}_j(t)]^2 \quad (5.7)$$

j . değişkenin varyansı;

$$Var_{X_j} = \frac{1}{T} \frac{1}{N-1} \sum_{i=1}^N [X_{ij}(t) - \bar{X}_j(t)]^2 \quad (5.8)$$

olarak ifade edilir.

Çok değişkenli panel veri kümeleme analizi oldukça karmaşıktır. Şimdilik, doğrudan kullanılabilecek ilgili bir yazılım yoktur, bu panel verinin çok değişkenli istatistiksel araştırmalarda seyrek kullanılmasının önemli bir nedenidir. Problemi çözme isteği çok keskin değilse bozulma fikrine uyum sağlanabilir. Zaman boyutundaki her değişkenin ortalaması alınır ve özel bir zamandaki durum olarak özetlenir. Böylece zaman boyutu yok edilir ve panel veri bozularak kesit verisi olur. Fakat bu bozulma en az iki kusura sahiptir. İlki, birçok bilgi kaybına neden olur. Çünkü ortalama değer sadece birimin ortalama değişimini gösterir ve sapma derecesi gibi diğer dağılım özelliklerini göstermez. İkincisi, resesif hipotez vardır yani tüm birimlerin aynı değişkeni aynı yönde değişir. Aksi halde, doğru olmayan veya yanlış sonuç ortaya çıkar. Fakat problemi çözme isteği kesinse, uzaklık fonksiyonunu ve kümeleme algoritmasını yeniden yapılandırmaya ihtiyaç vardır. Bilinen verilerin temelinde, birimlerin veya değişkenlerin yakınlığı gözlemlenir. Bazı kurallara göre, diğerlerinden daha büyük yakınlığa sahip örnekler veya değişkenler bir kümede toplanır ve daha büyük yakınlıktaki diğerleri diğer kümede birleştirilir. Bu süreçte, aynı kümedeki farkların küçük ve kümeler arasındaki farkın daha büyük olmasını garanti etmek gerekir. Bu süreç aynı yöntemle devam eder. En sonunda, birimler veya değişkenler birkaç kümeye bölünür. İki ana sorun vardır. İlki, örneklerin yakınlığı alınırken ne istatistiği kullanılacağı diğeri, hiyerarşik kümeleme yöntemlerinden hangi yöntemin uyum sağlayacağı ya da türlerin yakınlığını belirlemek için hangi kuralın uygunluğudur.

Karışık veri grubundan türlerin basit yapısını vermek için, yakınlık ölçüsü gereklidir. Bu birimler toplandığında “yakınlık” bazı uzunluklar tarafından tanımlanabilir. Bütünlükteki birim r ve birim k arasındaki uzaklık d_{rk} olarak gösterilir ve d_{rk} aşağıdaki gibi bazı koşulları sağlamalıdır.

1. $d_{rk} \geq 0$ eğer $X_{rj}(t) = X_{kj}(t)$ ise $d_{rk} = 0$ 'dır.
2. $d_{kr} = d_{rk}$ tüm $X_{rj}(t), X_{kj}(t)$
3. $d_{rk} \leq d_{rs} + d_{ks}$ tüm $X_{rj}(t), X_{kj}(t), X_{sj}(t)$

Benzer uzaklık fonksiyonları, Blok uzaklık, Öklid uzaklığı, Minkowski uzaklığı, Chebychev uzaklığı, Mahalanobis uzaklığını içerir. Zaman boyutu eklenen çok değişkenli panel veri için uzaklık fonksiyonu kesit verisi uzaklık fonksiyonundan farklıdır. Zamansal ve uzaysal Öklid uzaklığı olarak yeniden adlandırılır. Birim r ve birim k arasındaki zamansal ve uzaysal Öklid uzaklığı d_{rk} olarak gösterilir ve aşağıdaki gibi tanımlanır.

$$d_{rk} = \{\sum_{t=1}^T \sum_{j=1}^p [X_{rj}(t) - X_{kj}(t)]^2\}^{1/2} \quad (5.9)$$

İkililer arasındaki tüm birimlerin uzaklığı uzaklık matrisini oluşturur. Bu matris simetriktir ve köşegen elemanları sıfırdır. Alt üçgensel matris olarak aşağıdaki gibi gösterilir.

$$d_{rk} = \begin{bmatrix} 0 & & & \\ d_{21} & 0 & & \\ \vdots & \vdots & \ddots & \\ d_{N1} & d_{N1} & \dots & 0 \end{bmatrix}$$

Tanımlanan uzaklık ölçümü hiyerarşik kümeleme yöntemleri ile (Tek bağlantı (en yakın uzaklık) yöntemi, Tam bağlantı (en uzak uzaklık) yöntemi,

merkezi kümeleme yöntemi, medyan kümeleme yöntemi ve sapmaların kareleri (Wards) yöntemi) kullanılarak kümeleme analizi yapılabilir. Bununla birlikte, k -ortalamlar kümeleme, fuzzy kümeleme, iki aşamalı kümeleme yöntemleri gibi bir çok çeşit içeren zengin sınıflama yöntemleri ile de kümeleme yapılabilir. Zheng ve Li (2014) çalışmalarında, çok değişkenli panel verinin kümeleme analizini açıklamak için yeni bir uzaklık yöntemi önermişler ve Ward yöntemini kullanmışlardır. Uzaklık fonksiyonu olarak aynı olan, çok değişkenli panel verinin sapmasının kareleri toplamı fonksiyonu, kesit verisi fonksiyonundan farklıdır. g grubunun karesi toplamı S_g olarak gösterilir. S_g fonksiyonu aşağıdaki gibi tanımlanır.

$$S_g = \sum_{t=1}^T \sum_{j=1}^p \sum_{i \in i^g} [X_{ij}(t) - \bar{X}_j^g(t)]^2 \quad (5.10)$$

i^g , g grubundaki tüm sayı dizisinin grubudur. $\bar{X}_j^g(t)$ g grubundaki tüm birimlerin t zamandaki j . değişkeninin ortalamasını açıklar. Sayısal olarak Eşitlik (5.5)'te görülebilir.

Ward, eklenen kareler toplamının iki grup tek gruba birleştirildiğinde r grubundan k grubuna uzaklığın ölçebildiğini düşünür. Eklenen kareler toplamı en küçük olan iki grup birleştirilecektir. Eğer eklenen kareler toplamı eşitse ilgili birimler sırasıyla aynı zamanda birleştirilmelidir. r grubu ve k grubu birleştirilirken birbirleri arasındaki uzaklık aşağıdaki fonksiyon ile hesaplanır (k grubunda iki birim olduğu varsayılın. Eğer daha fazla birim varsa, k grubunun iki alt birimle birleştirildiğini düşünebilir ve yenileme devam eder).

$$D_{rk} = \Delta S_{rk} = \frac{n_r + n_p}{n_r + n_k} S_{rp} + \frac{n_r + n_q}{n_r + n_k} S_{rq} - \frac{n_r}{n_r + n_k} S_{pq} \quad (5.11)$$

n_r , r grubundaki birim sayısını açıklar. S_{rp} , r grubu ve alt grup p yeni bir grupta birleşeceği zaman kareler toplamını açıklar (Zheng ve Li, 2014).

İşlemlerin daha kolay anlaşılması açısından Çizelge 5.2.2.'de küçük bir panel veri seti verilmiştir. Veri seti, 3 birime sahip 2 değişkenli ve iki zaman boyutlu bir panel veridir. Çok değişkenli panel veri kümeleme analizi örnek üzerinde gösterilmiştir.

Çizelge 5.2. Panel veri seti

			Değişkenler	
Numara	Birim	Zaman	X_1	X_2
1	Adana	2015	3	2
1	Adana	2016	8	6
2	Ankara	2015	5	1
2	Ankara	2016	7	3
3	İstanbul	2015	4	3
3	İstanbul	2016	1	9

Çizelge 5.3. Yıllara göre ters çevrilmiş panel veri seti

Zaman		2015		2016	
Numara	Birim	X_1	X_2	X_1	X_2
1	Adana	3	2	8	6
2	Ankara	5	1	7	3
3	İstanbul	4	3	1	9

Eşitlik (5.9) kullanılarak her ikili çiftin uzaklıkları elde edilir. Aşağıda öklid uzaklığının karesi ile d_{12} ikilisi için uzaklık elde edilmiştir.

$$d_{12} = \{[X_{11}(2015) - X_{21}(2015)]^2 + [X_{12}(2015) - X_{22}(2015)]^2 + [X_{11}(2016) - X_{21}(2016)]^2 + [X_{12}(2016) - X_{22}(2016)]^2\}^{1/2}$$

$$d_{12} = \{[3 - 5]^2 + [2 - 1]^2 + [8 - 7]^2 + [6 - 3]^2\}^{1/2} = \sqrt{15}$$

Tüm ikili çiftlere Eşitlik (5.9) uygulanarak uzaklık matrisi aşağıdaki gibi elde edilir.

$$d_{rk} = \frac{1}{2} \begin{bmatrix} 1 & 2 & 3 \\ 0 & \sqrt{15} & \sqrt{51} \\ & 0 & \sqrt{78} \\ & & 0 \end{bmatrix}$$

Zheng ve Li (2014)'nin önerdiği uzaklık fonksiyonu Ward yöntemi ile kullanılarak çok değişkenli panel veri kümeleme analizi aşağıdaki gibi yapılır.

Eşitlik (5.10) kullanılarak her ikili çiftin uzaklıkları elde edilir. Panel veri sapmaları kareler toplamı fonksiyonu ile S_{12} ikili çifti için uzaklık elde edilmiştir.

$$S_{12} = \{[3 - 4]^2 + [5 - 4]^2 + [2 - 1.5]^2 + [1 - 1.5]^2 + [8 - 7.5]^2 + [7 - 7.5]^2 + [6 - 4.5]^2 + [3 - 4.5]^2\} = 7.5$$

Tüm ikili çiftlere bu eşitlik uygulanarak uzaklık matrisi aşağıdaki gibi elde edilir.

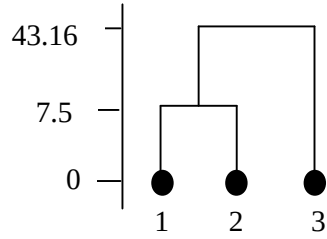
$$S_{rk} = \frac{1}{2} \begin{bmatrix} 1 & 2 & 3 \\ 0 & 7.5 & 30 \\ & 0 & 38.5 \\ & & 0 \end{bmatrix}$$

$S_{12} = 7.5$ minimum değere sahip olduğundan 1. ve 2. küme birleştirilerek uzaklık matrisi tekrar oluşturulur. Eşitlik (5.11) kullanılarak kümeleme yapılır.

$$\Delta S_{(3,(1,2))} = \frac{2}{3} 30 + \frac{2}{3} 38.5 - \frac{1}{3} (7.5) = 43.16$$

$$\Delta S_{rk} = \begin{matrix} & (1,2) & 3 \\ (1,2) & 0 & 43.16 \\ 3 & & 0 \end{matrix}$$

Birimler tek kümede toplandığı için işlem sonlandırılır. Kümelemenin dendrogramı aşağıdaki gibi elde edilir.



Şekil 5.2. Panel veri kümeleme örneği için elde edilen dendrogram

5.3. Karma (Nümerik ve İkili) Değişkenli Panel Veri için Kümeleme Analizi

Geleneksel kümeleme tekniklerinin ya sadece nümerik ya da sadece kategorik özellikli olan tek tip veri gruplarına odaklandığı Bölüm 4'te bahsedilmiştir. Gerçek hayatta panel veri grupları oluşturulurken zaman değişmezi ya da zamana göre değişen kategorik verilerle de oluşturulabilmektedir. Bu veri grupları hem nümerik hem de kategorik özelliklere sahip olduğundan kümelemede bazı zorluklara neden olmaktadır. Daha doğru ve daha anlamlı kümeleme için karma veri grupları birlikte düşünülmelidir. Bu nedenle karma değişkenli panel veri gruplarını kümeleyebilmek için uygun uzaklık ölçümleri ya da algoritmalar tasarlanmalıdır.

Karma değişkenli panel veriyi kümeleyebilmek için karma verileri kümelemede kullanılan Gower uzaklığı ve k-prototip uzaklığı karma değişkenli panel veri için düzenlenmiştir. Ayrıca Zheng ve Li (2014)'nin önerdiği uzaklık yöntemi karma değişkenli panel veri için geliştirilmiştir. Ek olarak k-prototip algoritmasında karma değişkenli panel veriyi kümelemek için uyarlanmıştır.

Gower benzerlik ölçümü karma değişkenli panel veri için Eşitlik (5.12)'deki gibi düzenlenir.

$$S(X_{ij}(t), X_{sj}(t)) = \frac{\sum_{t=1}^T \sum_{j=1}^p \delta_{isj}(t) S_{isj}(t)}{\sum_{t=1}^T \sum_{j=1}^p \delta_{isj}(t)} \quad (5.12)$$

Burada $S_{isj}(t)$, t zamanında j . özellik için iki birim arasındaki benzerliği gösterir. $\delta_{isj}(t)$, iki birim ölçümünde kayıp veri olup olmamasına dayalı olarak 0-1 katsayıdır. $D(X_{ij}(t), X_{sj}(t)) = 1 - S(X_{ij}(t), X_{sj}(t))$ basit eşitliği kullanılarak uzaklık ölçüsü elde edilebilir. Kesikli değişken için (ikili değişken içeren) benzerlik;

$$S_{isj}(t) = \begin{cases} 1 & X_{ij}(t) = X_{sj}(t) \\ 0 & X_{ij}(t) \neq X_{sj}(t) \end{cases} \quad (5.13)$$

şeklinde. Nümerik değişkenler için S_{isj} ,

$$S_{isj}(t) = 1 - \frac{|X_{ij}(t) - X_{sj}(t)|}{R_j(t)} \quad (5.14)$$

olarak ifade edilir. Burada $R_j(t)$ tüm birimlerin t zamanında j . değişkenin aralığıdır ve $R_j(t) = \max X_j(t) - \min X_j(t)$ olarak hesaplanır.

Eşitlik (4.7)'de verilen k -prototip uzaklığı karma değişkenli panel veri için aşağıdaki gibi düzenlenir.

$$d(i, s) = \sum_{i, s \in A_r} \sum_{t=1}^T \sum_{p=1}^P \left\{ \frac{X_{ij}(t) - X_{sj}(t)}{\max(X_j(t)) - \min(X_j(t))} \right\}^2 + \gamma \sum_{i, s \in A_c} \sum_{t=1}^T \sum_{j=p+1}^m \delta(X_{ij}(t), X_{sj}(t))$$

Eşitliğin ilk terimi nümerik değişkenler için zamansal ve uzaysal öklid uzaklığının karesidir ve tüm nümerik değişkenler $[0,1]$ aralığında normalleştirilir. Eşitliğin ikinci terimi ikili değişkenler için basit eşleştirme uzaklık ölçümüdür

(Eşitlik (4.6)) ve $[0.5, 1.4]$ aralığında değer alan γ değeri ile ağırlıklandırılır. Her biri p değişkenli iki gözlem çifti arasındaki eşleştirme tipli benzerlik ölçüsü bulunurken Çizelge 4.1.2.1’de verilen çaprazlık çizelgesinden faydalanılır. Buna göre, çaprazlık çizelgesinde her bir gözlem çiftinin karşılıklı değişkenlerinin aynı anda 1 ($X_{ij}(t) = X_{sj}(t) = 1$) değerini alanların sayısı a , 0 ($(X_{ij}(t) = X_{sj}(t) = 0)$) değeri alanların sayısı d , değişkenleri biri 1 diğeri 0 ($X_{ij}(t) = 0, X_{sj}(t) = 1$) alanların sayısı b , tam tersi durumda ($X_{ij}(t) = 1, X_{sj}(t) = 0$) c ile ifade edilmiştir.

k-prototip algoritması için Eşitlik (4.23) karma değişkenli panel veri için aşağıdaki gibi düzenlenir.

$$P(W, Q) = \sum_{l=1}^k \left(\sum_{i=1}^n w_{il} \sum_{t=1}^T \sum_{j=1}^p (X_{ij}(t) - q_{lj}(t))^2 + \gamma \sum_{i=1}^n w_{il} \sum_{t=1}^T \sum_{j=p+1}^m \delta(X_{ij}(t), q_{jl}(t)) \right)$$

Burada,

$$q_{lj}(t) = \frac{\sum_{i=1}^n w_{il} X_{ij}(t)}{\sum_{i=1}^n w_{il}} \quad 1 \leq l \leq k \text{ ve } 1 \leq j \leq p \quad (5.15)$$

olarak ifade edilir.

$$P_l^r = \sum_{i=1}^n w_{il} \sum_{t=1}^T \sum_{j=1}^p (X_{ij}(t) - q_{lj}(t))^2 \quad (5.16)$$

$$P_l^c = \sum_{i=1}^n w_{il} \sum_{t=1}^T \sum_{j=p+1}^m \delta(X_{ij}(t), q_{jl}(t)) \quad (5.17)$$

eşitlik yeniden yazılırsa,

$$P(W, Q) = \sum_{l=1}^k (P_l^r + P_l^c) \quad (5.18)$$

elde edilir. $P(W, Q)$ 'yu minimize etmek için k-prototip algoritması uygulanır.

5.3.1.Önerilen Uzaklık Ölçümü

Kriter fonksiyonu olarak, birimler arasındaki uzaklık ölçümü kümeleme sürecindeki öncelikli adımlardan biridir. Birimler arasındaki uzaklık ölçümünü bulmak için bilinen birçok yöntem vardır. Fakat geleneksel kümeleme algoritmalarının çoğu sadece nümerik ya da kategorik veriler için tasarlanmıştır.

Zheng ve Li (2014) çok değişkenli nümerik panel veri grubunu kümelemek için önerdiği uzaklık ölçümü karma (nümerik ve ikili) değişkenli panel veri kümeleme için geliştirilmiştir.

Panel veri Çizelge 5.2.1'de gösterildiği gibi tek boyuta çevrildikten sonra, önerilen uzaklık ölçümü için karma veri grubunu sadece nümerik ve sadece ikili değişkenler olmak üzere iki alt kümeye ayırır. İkili veri grubu basit eşleştirme ve nümerik veri grubu veriyi $[0,1]$ aralığında normalleştiren uzaklık ölçümleri kullanılarak birimler arasındaki uzaklık hesaplanır. Elde edilen uzaklıklar birleştirilir ve toplamsal hiyerarşik kümeleme algoritmaları kullanılarak kümeleme yapılır.

İki farklı veri grubunu ölçmek için formül aşağıdaki gibi tanımlanır.

$V=1(1:K)$, X_k değişkenlerinin indeksi olsun.

$$D^2(i, s) = \sum_{k \in V} \delta_k^2(i, s) \quad (5.19)$$

olarak ifade edilir. Burada $\delta_k^2(i, s)$, X_k değişkeni ile ilgili i ve s birimleri arasındaki uzaklığı ölçer. Genel olarak V , $Q = \{k \in V: X_k \text{ nümerik}\}$ ve $C = \{k \in V: X_k \text{ ikili}\}$ olarak bölünür. Böylece $V = Q \cup C$ 'dir ve X_k 'nın nümerik ve ikili değişken olmasına göre Eşitlik (5.20) genelleştirilebilir. X_k değişkeninin tipine göre δ_k^2 hesaplanabilir.

$$\delta_k^2(i, s) = \begin{cases} \frac{[X_{ij}(t) - \bar{X}_j^g(t)]^2}{R_j^2(t)} & k \in Q \\ d^2(i, s) & k \in C \end{cases}$$

Farklı veri tiplerine göre uzaklık ölçümleri bulunduktan sonra (i, s) ikilisi için uzaklık ölçümü aşağıdaki gibi tanımlanır.

$$D^2(i, s) = w_1 \sum_{t=1}^T \sum_{j=1}^p \sum_{i \in i^g} \frac{[X_{ij}(t) - \bar{X}_j^g(t)]^2}{R_j^2(t)} + w_2 \sum_{t=1}^T \sum_{j=p+1}^m d^2(i, s) \quad (5.20)$$

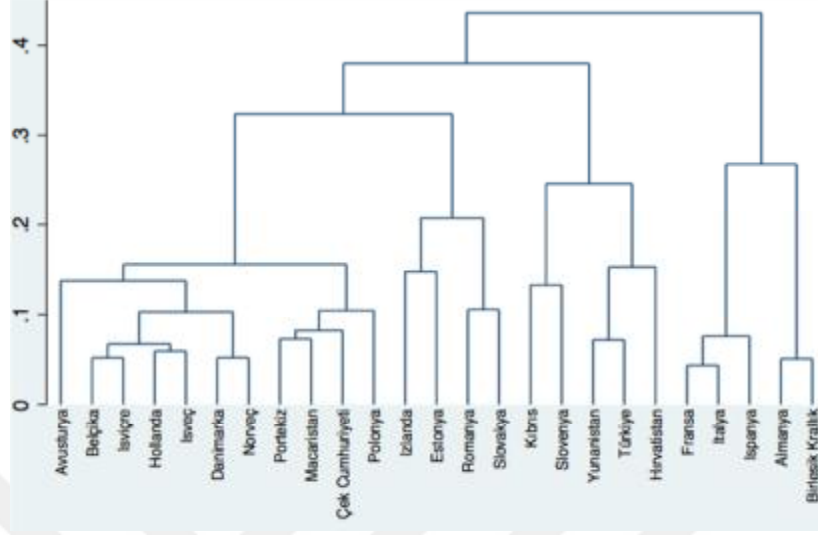
burada i^g g kümesindeki tüm dizi numaralarını gösterir, $\bar{X}_j^g(t)$ g kümesindeki tüm birimlerin t . zamandaki j . değişkenin ortalamasını gösterir. $R_j(t)$ tüm birimlerin t . zamandaki j . değişkenin aralık değeridir ve $R_j(t) = \max X_j(t) - \min X_j(t)$ olarak hesaplanır. $d^2(i, s)$ ikili veri için basit eşleştirme uzaklık ölçümüdür ve karma değişkenli panel veriyi kümelemek için uyarlanan k-prototip uzaklığında ifade edilmiştir. Ağırlıklar oransal olarak belirlenir ve değerleri, $w_1 = Q/V$ ve $w_2 = C/V$ olarak önerilir.



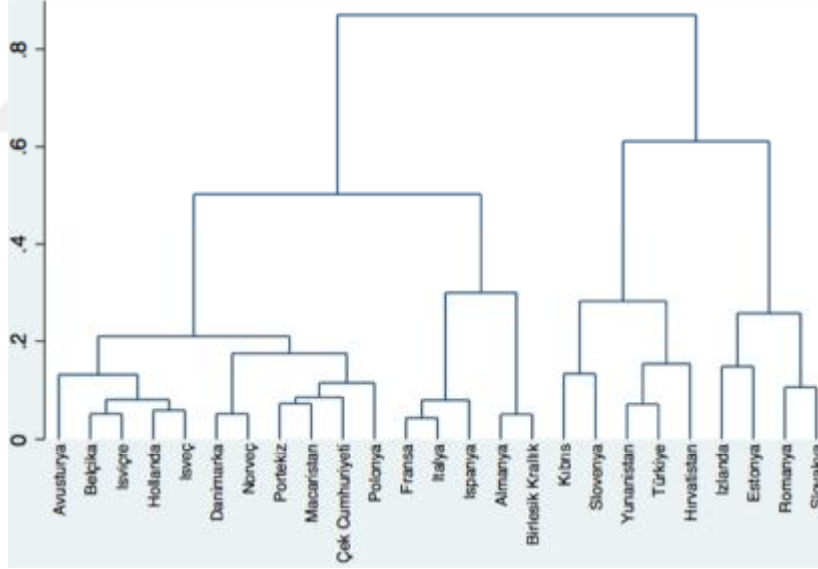
6. UYGULAMA

Bu çalışmada, karma değişkenli panel veri seti üzerinde hiyerarşik kümeleme yöntemlerinde seçilen uzaklık ölçümlerinin ve k-prototip algoritmasının kümelemeye etkisi araştırılacaktır. Bu amaç doğrultusunda, Gower, k-prototip ve önerilen uzaklık ölçümleri hiyerarşik kümeleme algoritmalarına entegre edilip karma değişkenli panel veri setinde kullanılarak sonuçlar incelenmiştir. Ayrıca karma değişkenli panel veri setine k-prototip algoritması uygulanarak sonuçlar değerlendirilmiştir. Karma değişkenli panel veri kümelemeye örnek olması açısından, 1993-2005 yılları arasında 25 ülkeye ait turizm gelirleri (TG), ülkelere gelen turist (GT) sayısı, turizm harcamaları (TH), ikili değişken olarak Akdeniz’e kıyısı olan ülkeler için 1, diğerleri için 0 değerini alan gösterge (D1) değişkeni, internet kullanıcı (IN) sayısı, yatak sayısı (YS) değişkenlerine sahip karma değişkenli panel veri seti ele alınmıştır. Veri seti ESDS International (Economic and Social Data Service) aracılığı ile elde edilmiştir. Veri setine Wald testi uygulanarak eğim parametresi heterojenliği kontrol edilmiştir (wald testi: 71.50, p:0.000). Wald testine göre eğim parametrelerinin heterojen olduğu görülmüştür. Kümeleme analizi için panel veri seti Stata programına girilip verilen zaman boyutu yatay kesit gösterge boyutuna çevrilerek işlemler yapılmıştır.

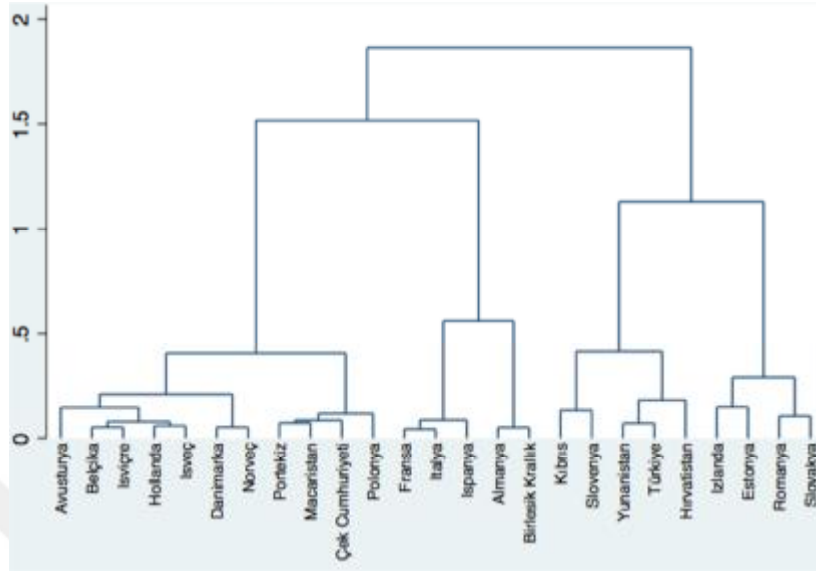
Gower, k-prototip ve önerilen uzaklık ölçümleri kullanılarak, hiyerarşik kümeleme algoritmaları ile bu veri setine kümeleme analizi yapılmıştır. Farklı uzaklıklara göre algoritma yöntemlerinin birimleri nasıl kümelediği dendogramlarda açıkça görülmektedir. Şekil 6.1-6.5’te Gower uzaklığı, Şekil 6.6-6.10’da k-prototip ve Şekil 6.11-6.15’te önerilen uzaklık ölçümleri kullanılarak yapılan toplamsal hiyerarşik kümeleme yöntemlerine ait dendogramlar verilmiştir. Bu algoritmalar için yapılan kümeleme sonuçlarında küme sayısı Calinski-Harabasz (CH) ve Duda Hart (DH) indekslerine göre belirlenmiştir. Kümelerin geçerlilik indeksleri olan Bağlılılık, Silhouette ve Dunn indeks değerlerine göre her bir uzaklığın algoritmalarındaki etkinliği karşılaştırılmıştır.



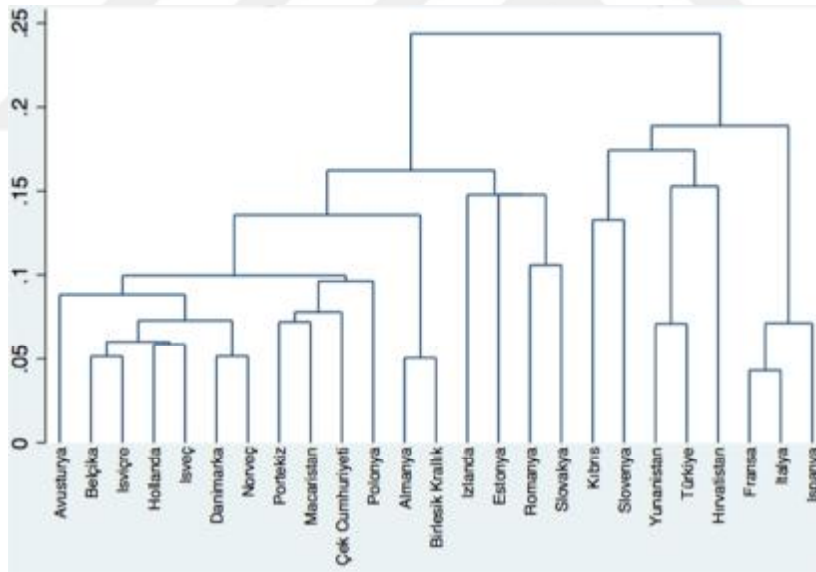
Şekil 6.1. Gower uzaklığı kullanılarak ortalama bağlantı yöntemi ile elde edilen dendrogram



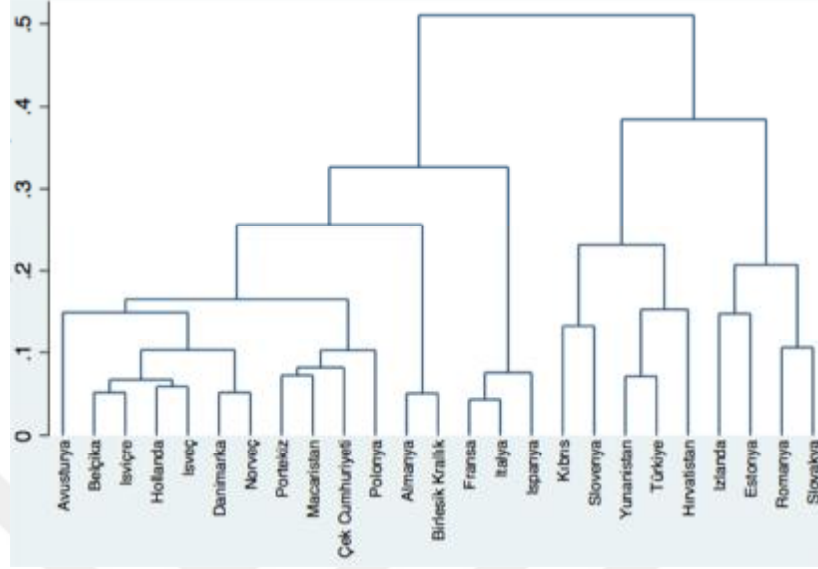
Şekil 6.2. Gower uzaklığı kullanılarak tam bağlantı yöntemi ile elde edilen dendrogram



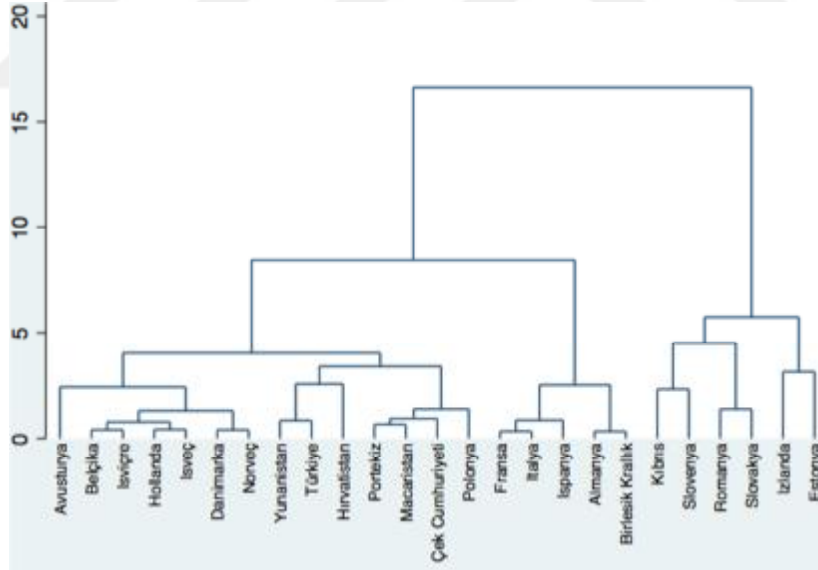
Şekil 6.3. Gower uzaklığı kullanılarak Ward yöntemi ile elde edilen dendrogram



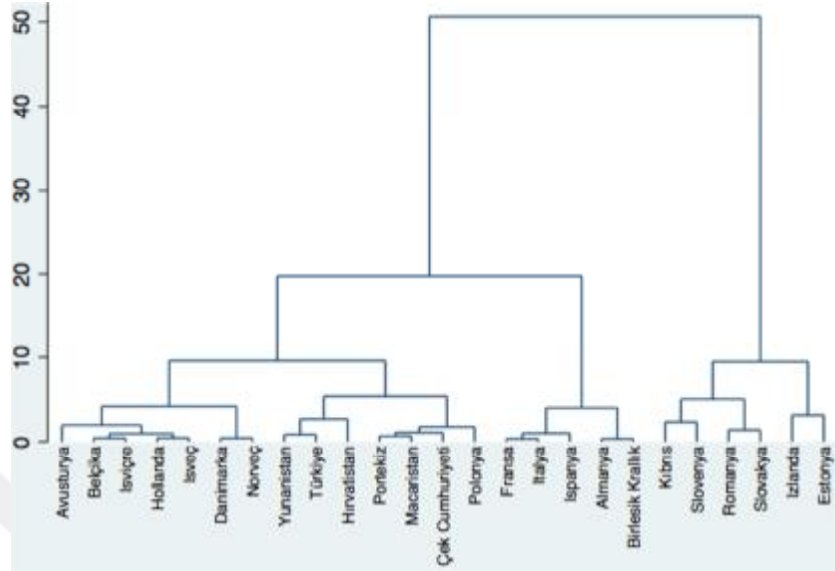
Şekil 6.4. Gower uzaklığı kullanılarak tek bağlantı yöntemi ile elde edilen dendrogram



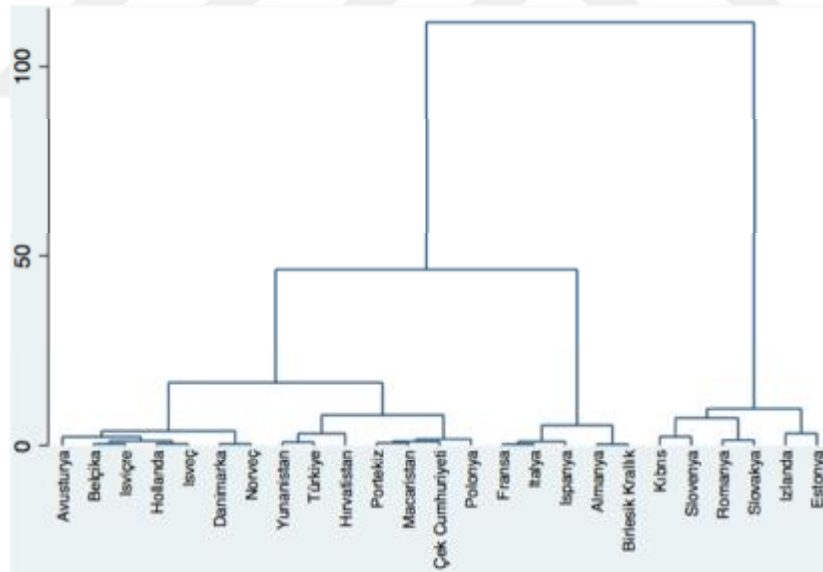
Şekil 6.5. Gower uzaklığı kullanılarak ağırlıklı ortalama bağlantı yöntemi ile elde edilen dendrogram



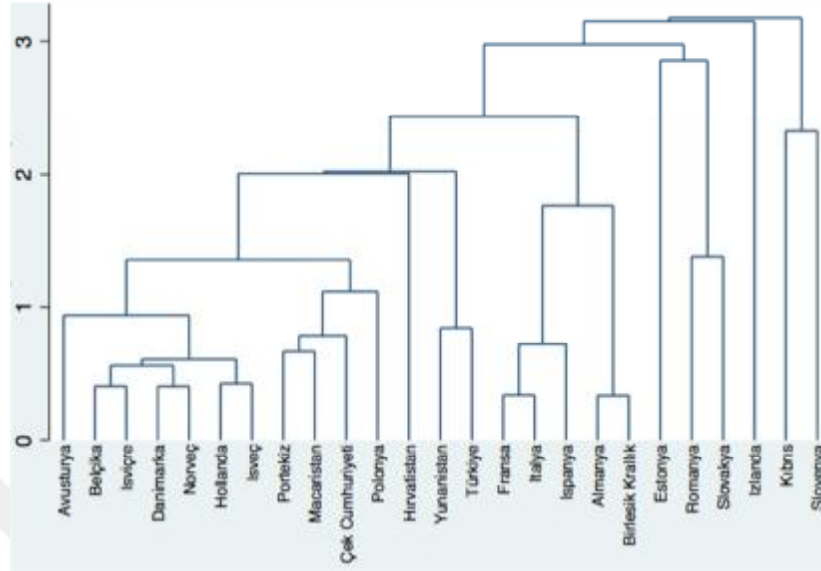
Şekil 6.6. k-prototip uzaklığı kullanılarak ortalama bağlantı yöntemi ile elde edilen dendrogram



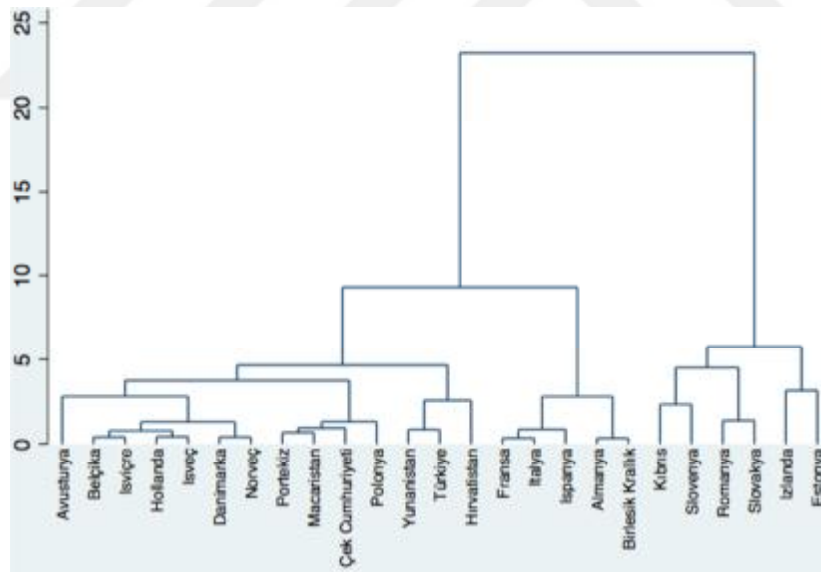
Şekil 6.7. k-prototip uzaklığı kullanılarak tam bağlantı yöntemi ile elde edilen dendrogram



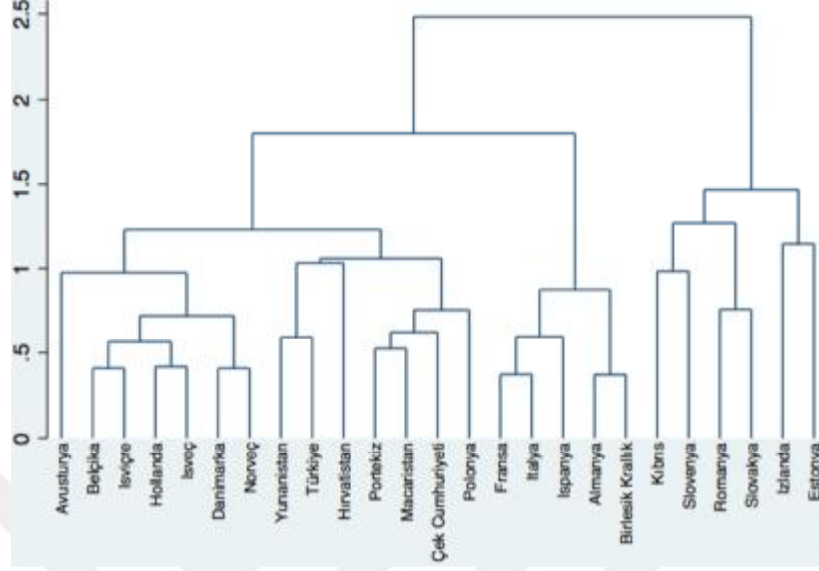
Şekil 6.8. k-prototip uzaklığı kullanılarak Ward yöntemi ile elde edilen dendrogram



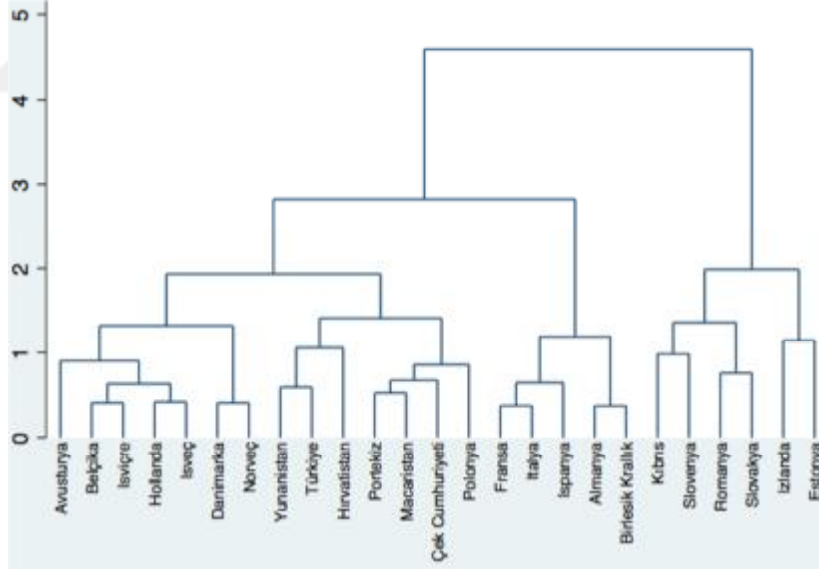
Şekil 6.9. k-prototip uzaklığı kullanılarak tek bağlantı yöntemi ile elde edilen dendrogram



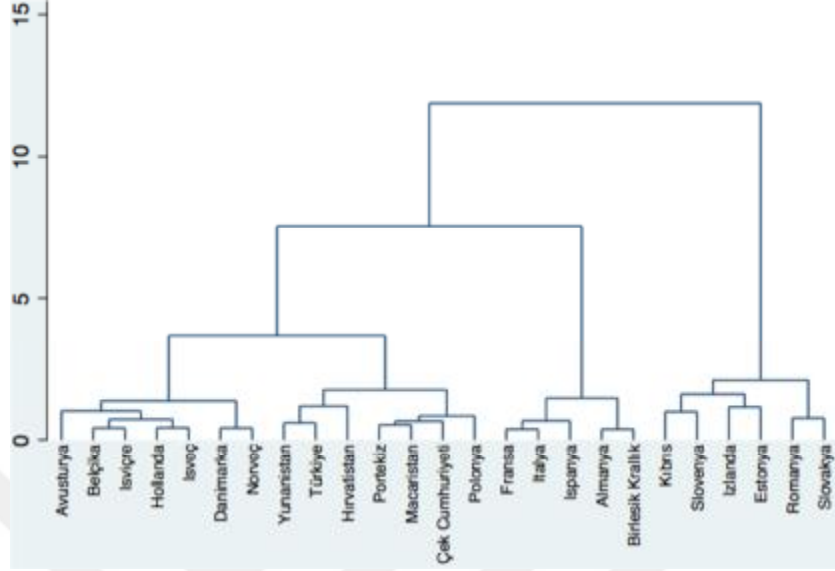
Şekil 6.10. k-prototip uzaklığı kullanılarak ağırlıklı ortalama bağlantı yöntemi ile elde edilen dendrogram



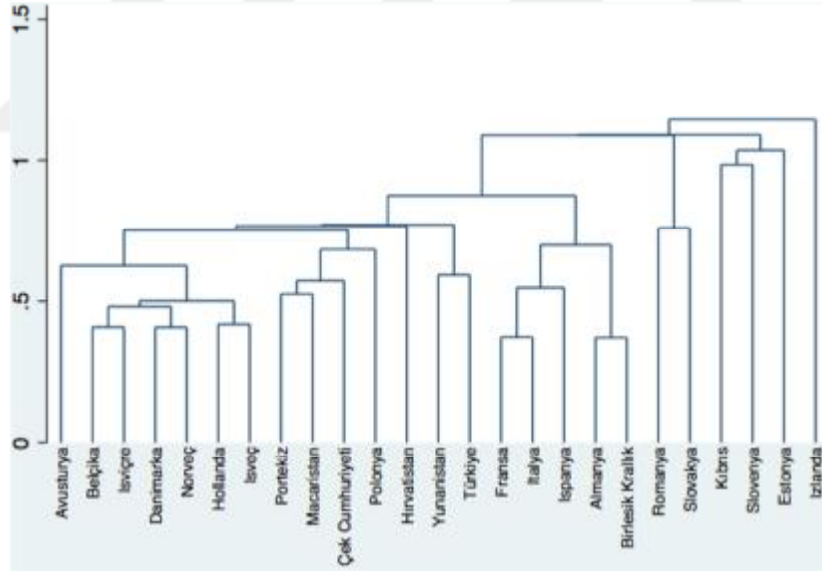
Şekil 6.11. Önerilen uzaklık kullanılarak ortalama bağlantı yöntemi ile elde edilen dendrogram



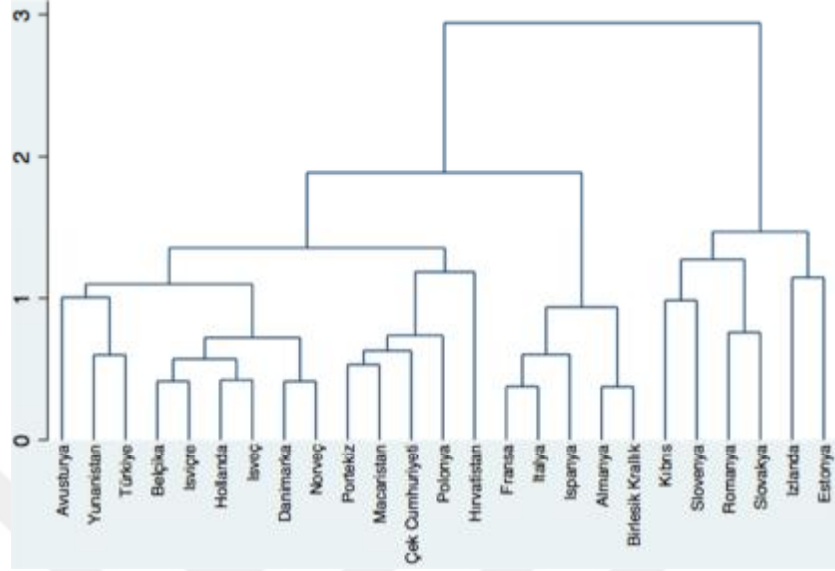
Şekil 6.12. Önerilen uzaklık kullanılarak tam bağlantı yöntemi ile elde edilen dendrogram



Şekil 6.13. Önerilen uzaklık kullanılarak Ward yöntemi ile elde edilen dendrogram



Şekil 6.14. Önerilen uzaklık kullanılarak tek bağlantı yöntemi ile elde edilen dendrogram



Şekil 6.15. Önerilen uzaklık kullanılarak ağırlıklı ortalama bağlantı yöntemi ile elde edilen dendrogram

Farklı uzaklıklar kullanılarak her bir yöntem için sırasıyla dendogramlar incelenirse birimlerin birbirine bağlanma uzaklıklarının kullanılan uzaklık matrisine bağlı olarak farklı olduğu açıkça görülmektedir. k-prototip ve önerilen uzaklık kullanılarak ortalama bağlantı yöntemiyle elde edilen dendogramda birimlerin kümelenme sırası aynıdır. Gower uzaklığı kullanıldığında ise kümelenmede farklılıklar oluşmuştur. Yunanistan, Türkiye ve Hırvatistan'ın oluşturduğu küme Kıbrıs ve Slovakya'nın oluşturduğu küme ile bağlanmıştır. k-prototip ve önerilen uzaklık kullanılarak tam bağlantı yöntemiyle elde edilen dendogramda yine birimler aynı sıra ile kümelenmişlerdir. Gower uzaklığı kullanıldığında ise Yunanistan, Türkiye ve Hırvatistan'ın oluşturduğu küme, Kıbrıs ve Slovenya kümesi ile birleşmiştir. Diğer iki dendogramda ise Kıbrıs ve Slovenya'nın bulunduğu küme Romanya ve Slovakya'nın bulunduğu küme ile birleşmiştir.

Ward yöntemiyle elde edilen dendogramdan devam edilecek olursa, k-prototip ve önerilen uzaklıkla elde edilen ortalama ve tam bağlantı dendogramları

birimlerin kümelenişi açısından aynıdır. Gower uzaklığı kullanılarak elde edilen dendogram da Gower uzaklığı kullanılarak elde edilen tam bağlantı dendogramı ile aynıdır. Üç farklı uzaklık kullanılarak Ward yöntemiyle elde edilen dendogramlarda birimlerin birleşmesinde farklılıklar vardır. k-prototip uzaklığının kullanıldığı dendogramda Kıbrıs ve Slovenya'nın, Romanya ve Slovakya'nın bulunduğu küme ile birleşirken, önerilen uzaklık kullanıldığında İzlanda ve Estonya'nın bulunduğu küme ile birleşmiştir. Gower uzaklığı kullanıldığında ise, Yunanistan, Türkiye ve Hırvatistan'ın bulunduğu küme ile birleşmiştir. Tek bağlantı yöntemi ile elde edilen dendograma gelinirse, k-prototip uzaklığı kullanıldığında Kıbrıs ve Slovenya'nın bulunduğu küme diğer tüm birimlerin oluşturduğu küme ile birleşmiştir. Önerilen uzaklık kullanıldığında Kıbrıs ve Slovenya, Estonya ile bağlanmıştır. Gower uzaklığı kullanıldığında ise Yunanistan, Türkiye ve Hırvatistan'ın oluşturduğu küme ile birleşmiştir. Dendogramı elde edilen son yöntem ağırlıklı ortalama bağlantı yöntemidir. k-prototip uzaklığı kullanıldığında Yunanistan ve Türkiye ve Hırvatistan'ın oluşturduğu küme Avusturya, Belçika, İsviçre, Hollanda, İsveç, Danimarka, Norveç, Portekiz, Macaristan, Çek Cumhuriyeti ve Polonya'nın oluşturduğu küme ile kümelenirken önerilen uzaklık kullanıldığında Türkiye ve Yunanistan, Avusturya ile kümelenmiştir. Gower uzaklığı kullanıldığında da Kıbrıs ve Slovenya'nın oluşturduğu küme ile kümelenmiştir.

Üç farklı uzaklık, beş farklı hiyerarşik kümeleme yönteminde kullanılarak dendogramlar elde edilmiş ve birimlerin kümelenme yapıları ayrıntılı bir şekilde incelenmiştir.

Uzaklıkların kümeleme yöntemlerindeki performansını belirlemek için her bir kümeleme yönteminde üç farklı uzaklık kullanılarak kofenetik korelasyon katsayıları elde edilmiştir. Elde edilen değerler Çizelge 6.1'de verilmiştir.

Çizelge 6.1. Turizm gelirleri verisi için üç farklı uzaklık ölçümü kullanılarak hiyerarşik kümeleme algoritmalarına göre elde edilen kofenetik koralasyon katsayı değerleri

		Gower	K-prototip ($\gamma=1$)	Önerilen
Turizm Gelirleri	Ortalama Bağ.	0.7327	0.6548	0.7642
	Tam Bağ.	0.6621	0.6509	0.7567
	Ward yönt.	0.6750	0.6514	0.7518
	Merkezi bağ.	0.7030	0.6535	0.6997
	Tek bağ.	0.6922	0.5736	0.71.83
	WPGMA	0.6745	0.6518	0.7592
	Medyan bağ.	0.6988	0.6497	0.6864

Çizelge 6.1’de görüldüğü gibi kofenetik korelasyon katsayısı Gower uzaklığı kullanılan merkezi ve medyan bağlantı yöntemlerinde daha yüksek iken önerilen uzaklık kullanılan diğer yöntemlerde daha iyi sonuçlar vermiştir. k-prototip kullanıldığında ise, yöntemlerin kofenetik korelasyon değerlerinin küçük olduğu görülmüştür. Kofenetik korelasyon katsayısı 1’e ne kadar yakınsa kümeleme o kadar başarılı olacaktır.

Veriye uygun doğru küme sayısını belirlemek için farklı uzaklıklar kullanılarak kümeleme yöntemleri 2’den 7’ye kadar kümelenerek Calinski-Harabasz ve Duda Hart indeks değerleri hesaplanmıştır. Değerler Çizelge 6.2-6.7’de verilmiştir.

Çizelge 6.2. Turizm Geliri verisi için Gower uzaklığı kullanılarak hiyerarşik kümeleme algoritmalarına göre elde edilen Calinski-Harabasz (1974) değerleri

Gower	Küme Sayısı	2	3	4	5	6	7
Calinski- Harabasz (1974)	Ortalama Bağ.	14.18	8.55	15.68	12.24	13.01	13.86
	Tam Bağ.	18.13	11.91	15.68	12.24	13.01	13.86
	Ward yönt.	18.13	17.37	15.68	12.24	13.01	15.69
	Merkezi bağ.	14.18	7.33	5.87	12.24	13.01	13.31
	Tek bağ.	1.52	4.19	3.60	10.26	8.30	8.30
	WPGMA	18.13	11.91	10.88	12.24	13.01	13.86
	Medyan bağ.	11.29	11.91	10.88	10.26	13.01	13.31

Çizelge 6.3. Turizm Geliri verisi için Gower uzaklığı kullanılarak hiyerarşik kümeleme algoritmalarına göre elde edilen Duda-Hart (1973) değerleri

Gower		2	3	4	5	6	7
	Ortalama Bağ.	0.905	0.426	0.313	0.310	0.446	0.207
Duda-Hart	Tam Bağ.	0.673	0.459	0.313	0.310	0.315	0.696
(1973)	Ward yönt.	0.459	0.673	0.313	0.446	0.315	0.557
Je(2)/Je(1)	Merkezi bağ.	0.313	0.905	0.426	0.310	0.400	0.217
index	Tek bağ.	0.317	0.310	0.467	0.207	0.400	0.217
	WPGMA	0.673	0.720	0.542	0.310	0.315	0.446
	Medyan bağ.	0.682	0.720	0.310	0.542	0.400	0.217

Çizelge 6.4. Turizm Geliri verisi için k-prototip uzaklığı kullanılarak hiyerarşik kümeleme algoritmalarına göre elde edilen Calinski-Harabasz (1974) değerleri

k-prototip		2	3	4	5	6	7
	Ortalama Bağ.	22.82	24.72	19.21	15.81	20.01	19.90
	Tam Bağ.	22.82	24.72	23.80	21.82	20.28	19.90
Calinski-	Ward yönt.	22.82	24.72	23.80	21.82	20.28	19.90
Harabasz	Merkezi bağ.	22.82	24.72	19.21	15.81	13.15	17.90
(1974)	Tek bağ.	3.50	4.98	8.90	6.96	13.15	11.12
	WPGMA	22.82	24.72	19.21	17.38	15.53	19.90
	Medyan bağ.	22.82	24.72	19.21	15.81	13.15	13.62

Çizelge 6.5. Turizm Geliri verisi için k-prototip uzaklığı kullanılarak hiyerarşik kümeleme algoritmalarına göre elde edilen Duda-Hart (1973) değerleri

K-prototip		2	3	4	5	6	7
	Ortalama Bağ.	0.515	0.597	0.398	0.550	0.537	0.000
Duda-Hart	Tam Bağ.	0.515	0.550	0.597	0.537	0.398	0.557
(1973)	Ward yönt.	0.515	0.550	0.597	0.537	0.398	0.313
Je(2)/Je(1)	Merkezi bağ.	0.515	0.597	0.398	0.000	0.550	0.537
index	Tek bağ.	0.790	0.633	0.217	0.515	0.000	0.880
	WPGMA	0.515	0.597	0.756	0.398	0.446	0.000
	Medyan bağ.	0.515	0.597	0.000	0.756	0.446	0.313

Çizelge 6.6. Turizm Geliri verisi için önerilen uzaklık kullanılarak hiyerarşik kümeleme algoritmalarına göre elde edilen Calinski-Harzbasz (1974) değerleri

Önerilen		2	3	4	5	6	7
Calinski-Harabasz (1974)	Ortalama Bağ.	22.82	24.72	19.21	15.81	20.01	17.90
	Tam Bağ.	22.82	24.72	19.21	21.82	20.28	19.90
	Ward yönt.	22.82	24.72	23.80	21.85	20.31	19.90
	Merkezi bağ.	4.46	11.05	18.95	14.92	13.15	11.12
	Tek bağ.	4.46	8.20	9.04	9.96	5.55	11.12
	WPGMA	22.82	24.72	19.21	20.57	18.72	16.93
	Medyan bağ.	22.82	13.18	9.80	7.57	6.44	5.35

Çizelge 6.7. Turizm Geliri verisi için önerilen uzaklık kullanılarak hiyerarşik kümeleme algoritmalarına göre elde edilen Duda-Hart (1973) değerleri

K-prototip		2	3	4	5	6	7
Duda-Hart (1973) Je(2)/Je(1) index	Ortalama Bağ.	0.515	0.597	0.398	0.550	0.000	0.537
	Tam Bağ.	0.515	0.597	0.550	0.537	0.398	0.557
	Ward yönt.	0.515	0.550	0.595	0.537	0.510	0.313
	Merkezi bağ.	0.595	0.514	0.703	0.398	0.000	0.822
	Tek bağ.	0.683	0.748	0.390	0.000	0.515	0.880
	WPGMA	0.515	0.597	0.601	0.398	0.468	0.000
	Medyan bağ.	0.882	0.624	0.703	0.398	0.000	0.494

Çizelge 6.2'ye bakıldığında Gower uzaklığı kullanılan çoğu kümeleme yöntemlerinin CH değerleri küme sayısını 2 olarak önerirken (tam bağlantı, Ward yöntemi, merkezi ve ağırlıklı ortalama bağlantı) Çizelge 6.3'deki DH değerleri küme sayısını 3 olarak önermektedir (ward yöntemi, merkezi, ağırlıklı ortalama ve medyan bağlantı).

Çizelge 6.4'de k-prototip uzaklığı kullanılan kümeleme yöntemlerinin CH değerleri tek bağlantı yöntemi hariç küme sayısını 3 olarak önermiştir. Çizelge 6.5'deki DH değerleri ise ortalama ve merkezi bağlantı yöntemi için küme sayısını 3 olarak önerirken tam bağlantı, ward yöntemi ve ağırlıklı ortalama yöntemleri için 4 olarak belirlemektedir.

Çizelge 6.6’da ise önerilen uzaklık kullanılarak elde edilen kümeleme yöntemlerinin CH değerleri genel olarak 3 küme etrafında yoğunlaşmıştır. Çizelge 6.7’deki DH değerleri ortalama ve tam bağlantı için 3, Ward yöntemi ve ağırlıklı ortalama için küme sayısını 4 olarak belirlemektedir. Genel olarak bakıldığında ise CH ve DH indekslerinin maksimum değerleri 3 küme etrafında yoğunlaştığı görülmektedir. Bu nedenle bu veri seti için optimal küme sayısı 3 olarak belirlenmiştir.

Küme sayısının 3 olduğu göz önüne alınarak birimlerin doğru kümelendiğine ilişkin üç farklı uzaklığın kullanıldığı kümeleme yöntemlerine ait küme geçerlilikleri olan bağlantılık silhouette ve dunn indeks değerleri Çizelge 6.8-6.10’da verilmiştir.

Çizelge 6.8. Turizm Geliri verisi için Gower uzaklığı kullanılarak hiyerarşik kümeleme algoritmalarına göre elde edilen geçerlilik değerleri

		Bağlanırlık	Silhouette	Dunn indeks
Gower	Ortalama Bağ.	12.1444	0.42	0.2447
	Tam Bağ.	12.1206	0.37	0.3117
	Ward yönt.	14.7718	0.38	0.2223
	Merkezi bağ.	10.1599	0.28	0.2223
	Tek bağ.	10.5511	0.39	0.2576
	WPGMA	12.12063	0.37	0.3217
	Medyan bağ.	12.12063	0.37	0.3217

Çizelge 6.9. Turizm Geliri verisi için k-prototip uzaklığı kullanılarak hiyerarşik kümeleme algoritmalarına göre elde edilen geçerlilik değerleri

		Bağlanırlık	Silhouette	Dunn indeks
K-prototip ($\gamma = 1$)	Ortalama Bağ.	11.0071	0.59	0.2539
	Tam Bağ.	11.0071	0.59	0.2539
	Ward yönt.	11.0071	0.59	0.2539
	Merkezi bağ.	11.0071	0.59	0.2539
	Tek bağ.	8.8297	0.59	0.0757
	WPGMA	11.0071	0.59	0.2539
	Medyan bağ.	11.0071	0.59	0.2539

Çizelge 6.10. Turizm Geliri verisi için önerilen uzaklık kullanılarak hiyerarşik kümeleme algoritmalarına göre elde edilen geçerlilik değerleri

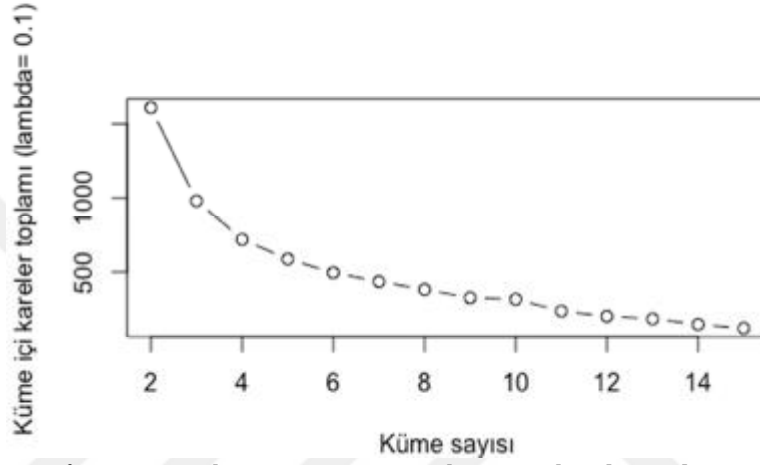
		Bağlanırlık	Silhouette	Dunn indeksi
Önerilen	Ortalama Bağ.	10.7714	0.41	0.4400
	Tam Bağ.	10.7714	0.41	0.4400
	Ward yönt.	10.7714	0.41	0.4400
	Merkezi bağ.	8.6861	0.26	0.2840
	Tek bağ.	9.3087	0.30	0.3261
	WPGMA	10.7711	0.41	0.3196
	Medyan bağ.	9.1849	0.15	0.4400

Çizelge 6.8-6.10'a bakıldığında önerilen uzaklık kullanılarak elde edilen hiyerarşik yöntemler bağlanırlık ve dunn indeks değerlerine göre en iyi performansı vermektedir. Silhouette değerine göre ise en yüksek değer k-prototip uzaklığı kullanılarak elde edilmiştir. Ortalama bağlantı algoritmasına göre önerilen uzaklık kullanıldığında bağlanırlık değeri 10.7714 iken Gower uzaklığı kullanıldığında 12.1444 ve k-prototip kullanıldığında ise 11.0071 dir. Önerilen uzaklık için dunn indeks değeri 0.4400 iken Gower uzaklığı kullanıldığında 0.2447 ve k-prototip kullanıldığında ise 0.2539'dur. Silhouette değeri ise, önerilen uzaklık kullanıldığında 0.41, Gower uzaklığı kullanıldığında 0.42 ve k-prototip kullanıldığında ise 0.59 değerini almaktadır. Diğer yöntemlerde aynı şekilde yorumlanabilir. Tekrar hatırlatılacak olursa bağlanırlık minimum, silhouette ve dunn indeks maksimum olmalıdır. Geçerlilik ölçümlerine göre genel olarak önerilen uzaklığın iyi performans gösterdiği söylenebilir.

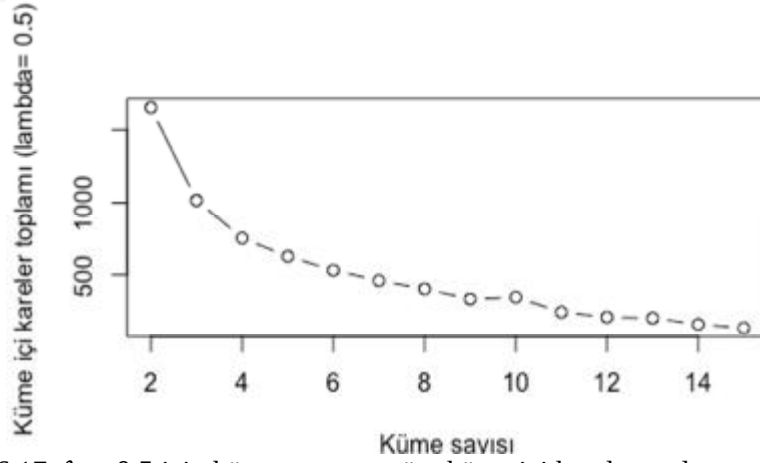
Karma değişkenli panel veri setine üç farklı uzaklık kullanılarak hiyerarşik yöntemlerle kümeleme analizi uygulanmış ve sonuçlar değerlendirilmiştir. Şimdi, k-prototip algoritması kullanılarak karma değişkenli panel veri seti kümelensin. Veri setine farklı λ değerleri için k-prototip algoritmasıyla kümeleme analizi uygulanmıştır.

k-prototip algoritması ile sırasıyla $\lambda = 0.1, 0.5, 0.7, 1$ değerleri için elde edilen küme sayısına göre küme içi kareler toplamına ait grafik Şekil 6.16-

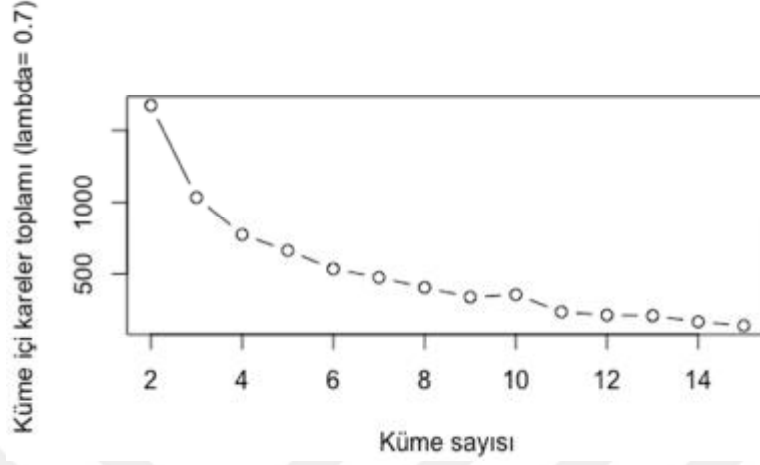
6.19’da verilmiştir. Birbirine benzer birimler aynı kümede iken birbirine benzemeyenlerde farklı kümelerde bulunacaktır. Kümeler içi kareler toplamı yardımıyla her bir küme için, kümedeki her bir noktanın küme merkezine olan uzaklığının karesinin toplamı alınır. Kümeler içi kareler toplamı hangi küme sayısında düşük çıkarsa uygun küme sayısını o’dur.



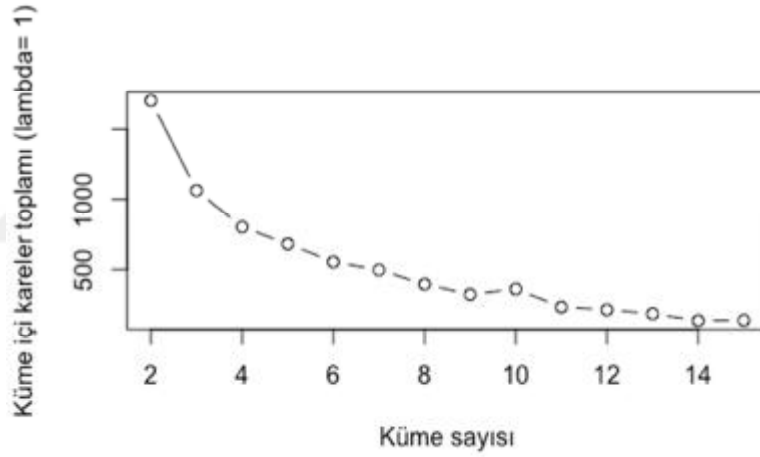
Şekil 6.16. $\lambda = 0.1$ için küme sayısına göre küme içi kareler toplamı



Şekil 6.17. $\lambda = 0.5$ için küme sayısına göre küme içi kareler toplamı



Şekil 6.18. $\lambda = 0.7$ için küme sayısına göre küme içi kareler toplamı



Şekil 6.19. $\lambda = 1$ için küme sayısına göre küme içi kareler toplamı

Şekil 6.16-6.19'da görüldüğü gibi λ 'nın aldığı değere göre optimum küme sayısı değişmemektedir. Genel olarak tüm λ değerlerinde küme sayısı 4 iken küme içi kareler toplamının birden azaldığı görülmektedir. Hiyerarşik kümeleme ile küme sayısı 3 olarak belirlenmiştir. Kümelemedeki temel amaç birbirine benzeyen birimleri sınırlı bir sayıda kümede toplayabilmektir. k-prototip algoritması için $\lambda = 1$ değeri alınarak karma değişkenli panel veri seti için küme sayısı 4 olarak belirlenmiştir. Kümelerdeki birimler;

Küme 1: Avusturya, Belçika, Danimarka, Hollanda, Norveç, İsveç, İsviçre

Küme 2: Yunanistan, Portekiz, Türkiye, Hırvatistan, Çek Cumhuriyeti, Macaristan, Polonya

Küme 3: Kıbrıs, İzlanda, Estonya, Romanya, Slovakya, Slovenya

Küme 4: Fransa, Almanya, İtalya, İspanya, İngiltere olarak elde edilir.

Turizim gelirleri karma değişkenli panel veri seti için Gower, k-prototip ve önerilen uzaklık ölçümleri kullanılarak hiyerarşik kümeleme algoritmaları ile kümeleme analizi yapılmıştır. Üç uzaklık yöntemi içinde küme sayısının 3 olarak belirlendiği ortak algoritmalar Ward yöntemi ve ağırlıklı ortalama yöntemidir. 3 küme için dendogramlara bakıldığında, Gower uzaklığı kullanılan tek bağlantı hariç diğer algoritmalarda kümedeki birimlerin aynı ancak birleşim noktaları ve yerlerinde değişiklik olduğu görülmektedir. Bu nedenle uygun kümeleme yöntemi Ward yöntemi olarak belirlenir. k-prototip ve önerilen uzaklık kullanıldığında Ward yöntemi için kümedeki birimlerin aynı, Gower uzaklığı kullanıldığında ise Türkiye, Yunanistan ve Hırvatistan'ın farklı kümede olduğu görülmektedir. Buna göre kümelerdeki birimler,

Küme 1: Avusturya, Belçika, İsviçre, Hollanda, İsveç, Danimarka Norveç, Portekiz, Macaristan, Çek Cumhuriyeti, Polonya.

Küme 2: Fransa, İtalya, İspanya, Almanya, Birleşik Krallık.

Küme 3: Kıbrıs, Slovenya Yunanistan, Türkiye, Hırvatistan, İzlanda, Estonya, Romanya, Slovakya,

olarak elde edilmiştir. Benzer şekilde önerilen ve k-prototip uzaklıkları kullanılarak elde edilen kümelerdeki birimler;

Küme 1: Avusturya, Belçika, İsviçre, Hollanda, İsveç, Danimarka, Norveç, Yunanistan, Türkiye, Hırvatistan, Portekiz, Macaristan, Çek Cumhuriyeti, Polonya.

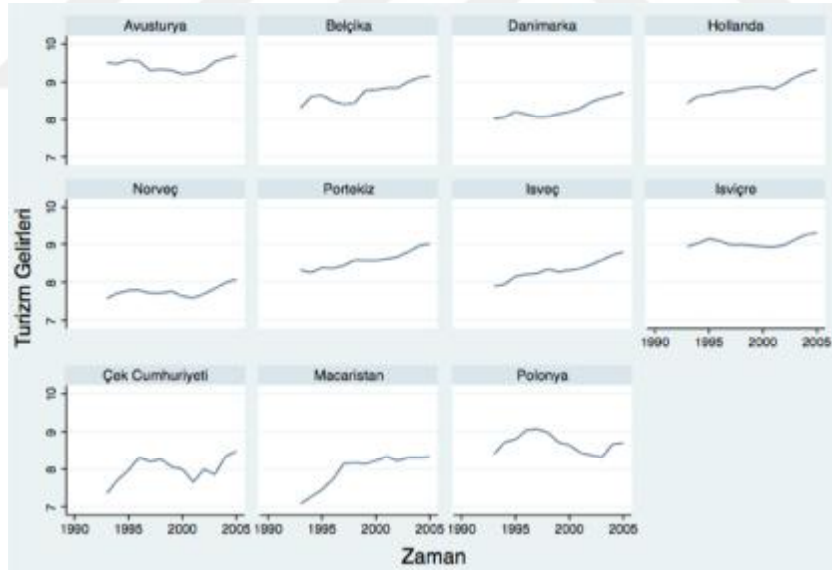
Küme 2: Fransa, İtalya, İspanya, Almanya, Birleşik Krallık.

Küme 3: Kıbrıs, Slovenya, Romanya, Slovakya, İzlanda, Estonya, şeklindedir.

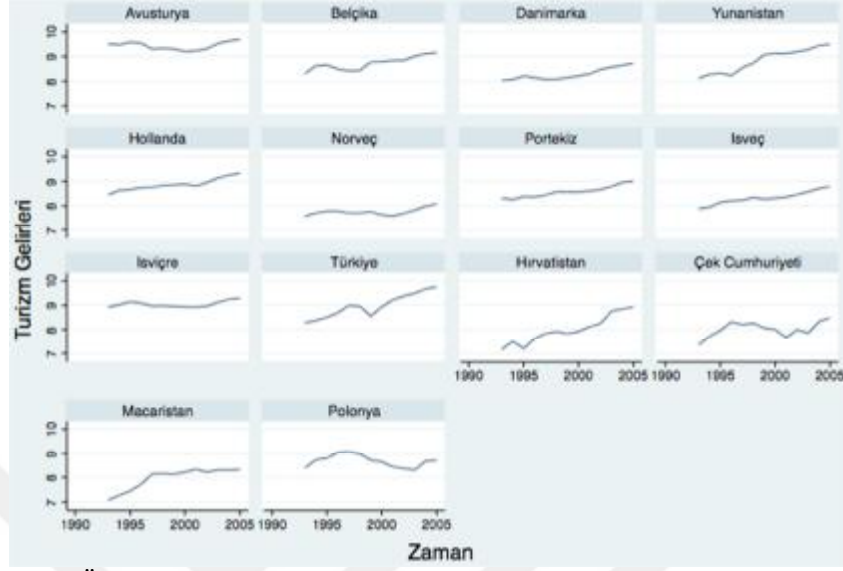
Her iki durumda da ikinci kümelerdeki birimler aynı olmasına rağmen birinci ve üçüncü kümelerdeki bazı birimler farklılık göstermektedir. Şekil 6.3 ve Şekil 6.13’de ki dendogramlardan açıkça görülmektedir. Gower uzaklığına göre Yunanistan, Türkiye ve Hırvatistan üçüncü kümede iken önerilen ve k-prototip uzaklıklarına göre ise bu ülkeler birinci kümede yer almaktadırlar.

Farklı uzaklık ölçümleri kullanılarak toplamsal hiyerarşik kümeleme yöntemleri ile karma değişkenli panel veriye kümeleme analizi yapılmıştır. Böylece küme içi homojen kümeler arası heterojen birimler elde edilmiştir. Elde edilen kümelere panel veri analizi yapılarak farklı uzaklık yöntemleri ile kümelenen panel veri seti için parametre tahmin sonuçları incelenmiştir.

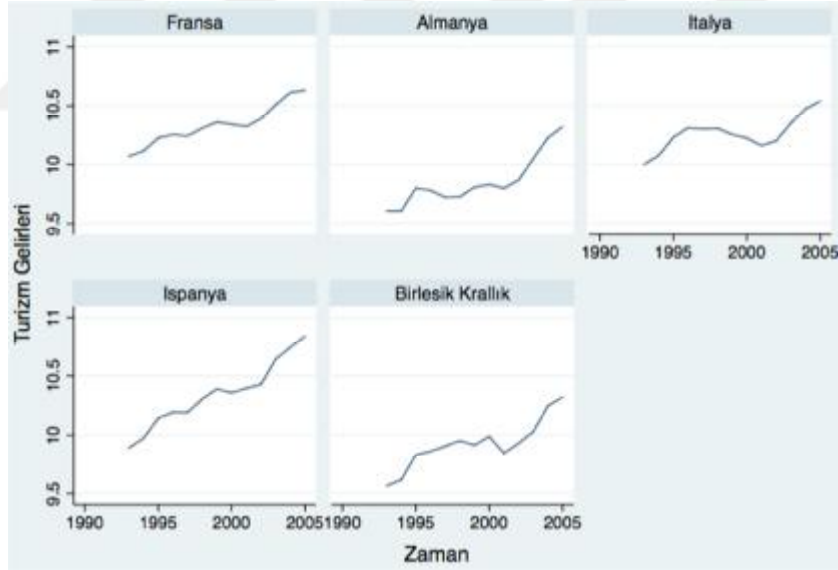
Şekil 6.20-6.24’te Turizm Gelirleri bağımlı değişkeninin her bir kümedeki birimler için çizgi grafiği gösterilmiştir. Kümeler arası birimlerin yıllara göre farklı dalgalanma gösterdikleri açıkça görülmektedir.



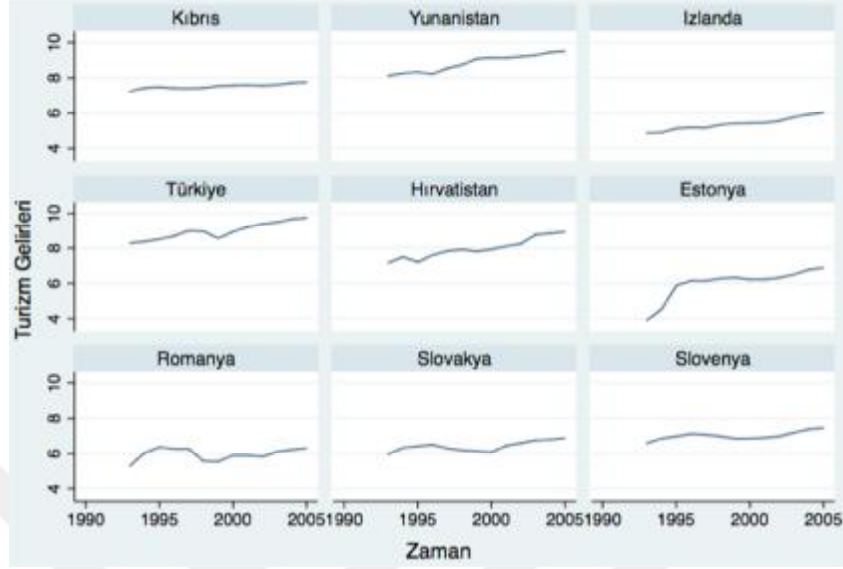
Şekil 6.20. Gower uzaklığı kullanılarak elde edilen birinci küme için birimlere göre Turizm Gelirleri çizgi grafiği



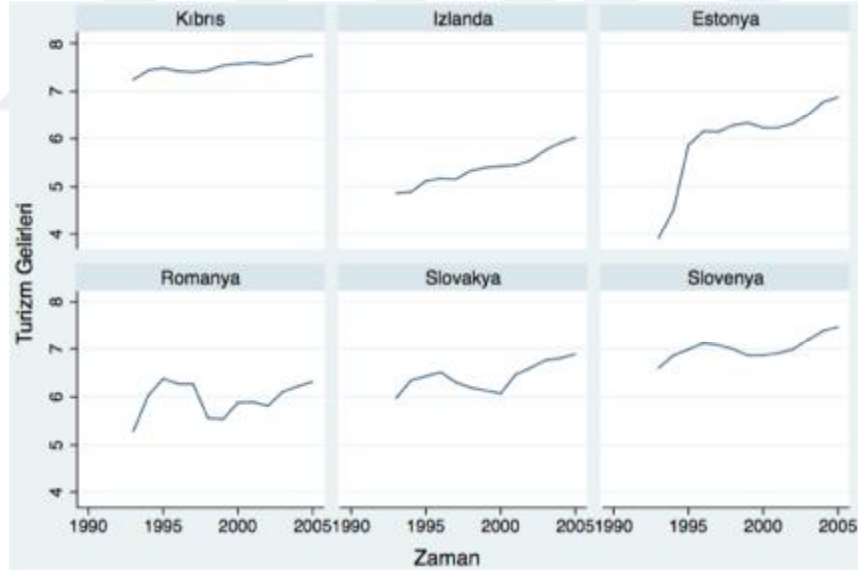
Şekil 6.21. Önerilen uzaklık kullanılarak elde edilen birinci küme için birimlere göre Turizm Gelirleri çizgi grafiği



Şekil 6.22. Gower ve önerilen uzaklık kullanılarak elde edilen ikinci küme için birimlere göre Turizm Gelirleri çizgi grafiği



Şekil 6.23. Gower uzaklığı kullanılarak elde edilen üçüncü küme için birimlere göre Turizm Gelirleri çizgi grafiği



Şekil 6.24. Önerilen uzaklık kullanılarak elde edilen üçüncü küme için birimlere göre Turizm Gelirleri çizgi grafiği

Turizm Gelirleri verisine kümeleme analizi uygulandıktan sonra her bir küme için panel veri analizi yapılır. Parametre tahmini için kullanılacak yöntemi

belirlemek için öncelikle her bir küme için yatay kesit bağımlılığı, birim kök testi, değişen varyans ve otokorelasyon testleri yapılır.

Panel veri regresyon modelleri, birimler arasında yatay kesit bağımsızlığı varsayımına dayanmakta olup, Baltagi (2005) panel veri analizinde uzun dönemde serilerde rastlanacak yatay kesit bağımlılığının standart sabit etki (FE) ve rassal etki (RE) tahminlerinin tutarlı ancak etkin olmamalarına ve tahmin edilen standart hataların sapmalı olmasına neden olduğunu belirtmiştir. Bu nedenle panel veri analizlerinde yatay kesit bağımlılığını test eden birkaç farklı test kullanılmaktadır. Çalışmada bu testlerden birimler arasında korelasyon varlığına bakan Pesaran (2004) testi, Spearman'ın rank korelasyon katsayısını kullanarak ortalama değer hesaplayan ve parametrik olmayan Friedman (1937) testi ve rank korelasyon katsayılarının karelerinin toplamına dayanan Frees (1995) testi kullanılmıştır. İki farklı uzaklık kullanılarak elde edilen tüm kümeler için yapılan testlerin sonuçları Çizelge 6.11-6.15'te verilmiştir.

Çizelge 6.11. Gower uzaklığı ile elde edilen Küme 1'deki birimler için yatay kesit testleri

Test	İstatistik değeri	Anlamlılık değeri
Pesaran	6.026	0.0000
Friedman	30.042	0.0008
Frees	1.748	Q dağılımı kritik değeri 0.10 için 0.1984 0.05 için 0.2620 0.01 için 0.3901

Çizelge 6.12. Önerilen uzaklık ile elde edilen Küme 1'deki birimler için yatay kesit testleri

Test	İstatistik değeri	Anlamlılık değeri
Pesaran	8.334	0.0000
Friedman	43.752	0.0000
Frees	2.336	Q dağılımı kritik değeri 0.10 için 0.1984 0.05 için 0.2620 0.01 için 0.3901

Çizelge 6.13. Gower ve önerilen uzaklık ile elde edilen Küme 2'deki birimler için yatay kesit testleri

Test	İstatistik değeri	Anlamlılık değeri
Pesaran	6.996	0.0000
Friedman	36.791	0.0000
Frees	1.280	Q dağılımı kritik değeri 0.10 için 0.1984 0.05 için 0.2620 0.01 için 0.3901

Çizelge 6.14. Gower uzaklığı ile elde edilen Küme 3'deki birimler için yatay kesit testleri

Test	İstatistik değeri	Anlamlılık değeri
Pesaran	2.139	0.0324
Friedman	22.623	0.0039
Frees	0.942	Q dağılımı kritik değeri 0.10 için 0.1984 0.05 için 0.2620 0.01 için 0.3901

Çizelge 6.15. Önerilen uzaklık ile elde edilen Küme 3'deki birimler için yatay kesit testleri

Test	İstatistik değeri	Anlamlılık değeri
Pesaran	0.079	0.9368
Friedman	14.484	0.0128
Frees	0.947	Q dağılımı kritik değeri 0.10 için 0.1984 0.05 için 0.2620 0.01 için 0.3901

Tüm kümeler için yapılan her üç testte de değişkenler arasında yatay kesit bağımlılığı tespit edilmiştir. Çalışmanın bundan sonraki kısmında uygulanacak panel birim kök testi için yatay kesit bağımlılığı test sonuçları oldukça önem taşımaktadır. Bunun nedeni birinci nesil panel birim kök testlerinin, yatay kesit birimlerinin birbirinden bağımsız olduğu varsayımı altında oluşturulmasıdır. Paneli oluşturan yatay kesit birimlerinin, seriye gelen bir şok karşısında birbirinden hiçbir zaman etkilenmediğini varsaymak gerçekçi olmayacağı için, daha sağlıklı sonuçlar

elde edebilmek adına yatay kesit bağımlılığını dikkate alan ikinci nesil birim kök testlerinin uygulanması gereklidir. Yatay kesit birimlerine ait hata terimlerinin faktör yapılarının modellenmesine dayanan ikinci nesil panel birim kök testlerine Choi (2002), Phillips ve Sul (2003), Bai ve Ng (2004), Moon ve Peron (2004) testleri örnek verilebilir. Bu çalışmada ise yatay kesit bağımlılığı tespit edilmiş olduğundan, Pesaran (2007)'ın hata terimlerinin faktör yapılarının tahminlenmesi yerine, uygulama kolaylığı sağlayan yatay kesit bağımlılığına dirençli olarak geliştirdiği genişletilmiş Dickey-Fuller (Cross-Sectionally Augmented Dickey-Fuller (CADF)) panel birim kök testi uygulanmış ve serilerin durağanlığı incelenmiştir. Gower ve önerilen uzaklık ölçümleri ile elde edilen kümeler için birim kök test sonuçları Çizelge 6.16-6.20'de verilmiştir.

Çizelge 6.16. Gower uzaklığı ile elde edilen Küme 1'deki birimler için birim kök testleri

Değişken	Düzy		Birinci Fark	
	Z[t-bar]	Anlamlılık değeri	Z[t-bar]	Anlamlılık değeri
TG	-1.491	0.767	-3.522	0.000
GT	-1.728	0.503	-2.784	0.001
YS	-1.735	0.494	-3.053	0.000
TH	-1.185	0.951	-3.081	0.000
IN	-2.114	0.122	-3.794	0.000
D1	2.610	1.000	2.610	1.000

Çizelge 6.17. Önerilen uzaklık ile elde edilen Küme 1'deki birimler için birim kök testleri

Değişken	Düzy		Birinci Fark	
	Z[t-bar]	Anlamlılık değeri	Z[t-bar]	Anlamlılık değeri
TG	-1.792	0.415	-3.321	0.000
GT	-1.836	0.359	-2.563	0.002
YS	-1.486	0.799	-2.738	0.000
TH	-1.690	0.555	-3.084	0.000
IN	-1.843	0.350	-3.714	0.000
D1	2.610	1.000	2.610	1.000

Çizelge 6.18. Gower ve önerilen uzaklık ile elde edilen Küme 2’deki birimler için birim kök testleri

Değişken	Düzy		Birinci Fark	
	Z[t-bar]	Anlamlılık değeri	Z[t-bar]	Anlamlılık değeri
TG	-1.377	0.753	-3.794	0.000
GT	-1.622	0.572	-2.867	0.009
YS	-1.687	0.519	-2.712	0.020
TH	-2.329	0.102	-2.879	0.008
IN	-2.742	0.017	-3.643	0.000
D1	2.610	1.000	2.610	1.000

Çizelge 6.19. Gower uzaklığı ile elde edilen Küme 3’deki birimler için birim kök testleri

Değişken	Düzy		Birinci Fark	
	Z[t-bar]	Anlamlılık değeri	Z[t-bar]	Anlamlılık değeri
TG	-2.072	0.159	-3.281	0.000
GT	-1.919	0.283	-2.083	0.152
YS	-2.277	0.059	-2.780	0.002
TH	-2.383	0.032	-2.923	0.000
IN	-2.466	0.019	-3.966	0.000
D1	2.610	1.000	2.610	1.000

Çizelge 6.20. Önerilen uzaklık ile elde edilen Küme 3’deki birimler için birim kök testleri

Değişken	Düzy		Birinci Fark	
	Z[t-bar]	Anlamlılık değeri	Z[t-bar]	Anlamlılık değeri
TG	-1.748	0.466	-2.078	0.002
GT	-1.916	0.321	-2.078	0.204
YS	-2.389	0.063	-2.750	0.010
TH	-1.964	0.284	-3.230	0.000
IN	-2.355	0.074	-3.788	0.000
D1	2.610	1.000	2.610	1.000

Çizelge 6.16-6.20’de yer alan sonuçlar, tüm kümelerdeki değişkenlerin serilerinde birim kök olduğunu göstermiştir. Bu nedenle birinci farkları alınarak serilerin CADF test sonuçlarına da bakılmıştır. Birinci farkları alınan

değişkenlerden Küme 3'te ülkeye gelen turist (GT) sayısı değişkeni ve ülkelerin akdenize kıyısı olup olmadığını gösteren ikili değişken (D1) serisi hariç diğer değişkenler durağan hale gelmektedir.

Değişkenlere ait serileri birim kök içerdiği durumlarda, bu serilerin doğrusal bileşimleri durağan olabilmekte ve seriler uzun dönemde ilişkili olabilmektedir. Sistemi etkileyen kalıcı şoklara rağmen, değişkenler arasında uzun dönemde bir denge ilişkisinin de test edilmesi gerekmektedir. Panel Birim Kök Testleri gibi panel eşbütünleşme testlerini kullanmanın zaman serisinde eşbütünleşme uygulamasına göre avantajı, testlerin gücünü artırmasıdır.

Panel eşbütünleşme testi Westerlund (2008) testi ile yapılacaktır. Bu testi, diğer eşbütünleşme testlerinden ayıran temel özellik, diğer bütün testler değişkenler arasındaki eşbütünleşik ilişkiyi belirlerken, modeldeki değişkenlerin tamamının durağan olmaması gerekliliği ön şartına dayanırken, Westerlund (2008) testi modeldeki değişkenlerden en az birinin durağan olmasının yeteceği ön şartı ile çalışmaktadır. Her iki uzaklık kullanılarak elde edilen tüm kümeler yapıları eşbütünleşme testi Westerlund (2008) test sonuçlarına göre %5 anlamlılık düzeyinde (aynı zamanda %1 ve %10 düzeylerinde de) seriler arasında eşbütünleşik ilişkinin olmadığı sonucuna varılmıştır.

Modelde var olabilecek diğer istatistiksel sorunlar değişen varyans ve otokorelasyon sorunlarıdır. Değişen varyans ve Otokorelasyon testleri sonuçları Çizelge 6.21 ve Çizelge 6.22'te verilmiştir.

Çizelge 6.21. Gower uzaklığı ile elde edilen kümeler için değişen varyans ve otokorelasyon testi

	Küme 1	Küme 2	Küme 3
Ki-kare	6084.65	509.13	421.17
p	0.000	0.000	0.000
Durbin –Watson	0.5432	0.7105	0.6035
Baltagi-Wu LBI	0.8558	1.0437	0.8711

Çizelge 6.22. Önerilen uzaklık ile elde edilen kümeler için değişen varyans ve otokorelasyon testi

	Küme 1	Küme 2	Küme 3
Ki-kare	181.06	509.13	641.26
p	0.000	0.000	0.000
Durbin –Watson	0.4664	0.7105	0.6271
Baltagi-Wu LBI	0.7351	1.0437	0.9186

Çizelge 6.21-6.22’de görüldüğü gibi iki farklı uzaklık kullanılarak elde edilen tüm kümelerde $p=0.000$ olduğundan sabit varyansı (hemoskedasite) ifade eden H_0 hipotezi reddedilmektedir. Modelde değişen varyans vardır. Durbin–Watson ve Baltagi-Wu LBI testleri içinde değerler 2’den küçük olduğundan modelde otokorelasyon varlığından söz edilir.

Yapılan tüm testler sonucunda tüm kümelerde yatay kesit bağımlılığı, değişen varyans ve otokorelasyon olduğu görülmüştür. Bu nedenle parametre tahmini için modelde yatay kesit bağımlılığı, değişen varyans ve otokorelasyon varlığında dirençli tahmin edici olan Driscoll ve Kraay tahmincisi kullanılmıştır. Yapılan F testine göre Driscoll ve Kraay standart hatalarla havuzlanmış en küçük kareler tahmin edicisi yerine Driscoll ve Kraay standart hatalarla sabit etkiler tahmin edicisi kullanılmasının daha uygun olduğu görülmüştür. Sonuçlar Çizelge 6.23-6.28’de verilmiştir.

Çizelge 6.23. Gower uzaklığı ile elde edilen Küme 1 için parametre tahminleri

TG	Katsayı	Standart Hata	p-değeri	95% Güven Aralıkları	
GT	0.400	0.188	0.055	-0.009	0.810
YS	0.147	0.144	0.327	-0.166	0.461
TH	0.212	0.125	0.117	-0.061	0.486
IN	0.049	0.023	0.053	-0.0008	0.100
D1	0	Omitted			
sabit	1.805	1.678	0.303	-1.851	5.462
Düz. R^2	0.408				

Çizelge 6.24. Önerilen uzaklık ile elde edilen Küme 1 için parametre tahminleri

TG	Katsayı	Standart Hata	p-değeri	95% Güven Aralıkları	
GT	0.766	0.098	0.000	0.531	0.961
YS	0.137	0.109	0.233	-0.100	0.375
TH	0.213	0.113	0.084	-0.033	0.460
IN	0.057	0.020	0.019	0.011	0.102
D1	0	Omitted			
sabit	-1.443	0.598	0.158	-3.531	0.644
Düz. R^2	0.640				
MSE	0.665				

Çizelge 6.23 ve Çizelge 6.24'te her iki uzaklık yönteminden elde edilen Küme 1 için Driscoll ve Kraay standart hatalarla sabit etkiler tahmin sonuçları yer almaktadır. Gower uzaklığı ile elde edilen Küme 1 için parametre tahminlerine bakıldığında modeldeki hiçbir değişkenin modele katkı sağlamadığı ve ülkelerin Akdeniz'e kıyısı olup olmadığını gösteren iki değişkeninde modelden düşürüldüğü ve düzeltilmiş R^2 değerinin %40 olduğu görülmektedir. Önerilen uzaklık ile elde edilen Küme 1 için parametre tahminlerinde ise Gelen Turist (GT) sayısı ve Internet Kullanıcı (IN) sayısı değişkenlerinin modele anlamlı katkı sağladığı, düzeltilmiş R^2 değerinin %64 ve MSE değerinin %66 olduğu görülmektedir. Önerilen uzaklık kullanılarak elde edilen Küme 1'in parametre tahminleri için %95 güven aralıklarının Gower uzaklığı kullanılarak elde edilen Küme 1'in parametre tahminleri için %95 güven aralıklarından daha dar olduğu da söylenebilir.

Çizelge 6.25. Gower ve önerilen uzaklık ile elde edilen Küme 2 için parametre tahminleri

TG	Katsayı	Standart Hata	p-değeri	95% Güven Aralıkları	
GT	0.648	0.102	0.000	0.425	0.871
YS	-0.033	0.135	0.809	-0.329	0.262
TH	0.538	0.082	0.000	0.358	0.719
IN	-0.007	0.017	0.669	-0.045	0.305
D1	0	Omitted			
sabit	-1.845	1.095	0.118	-4.233	0.541
Düz. R^2	0.872				
MSE	0.303				

Çizelge 6.25'te her iki uzaklık yönteminden elde edilen Küme 2 için Driscoll ve Kraay standart hatalarla sabit etkiler tahmin sonuçları yer almaktadır. İki uzaklıkta da elde edilen Küme 2'deki birimler aynı olduğundan parametre tahminleri de aynı sonucu vermiştir. Küme 2 için yapılan parametre tahminlerinde Gelen Turist (GT) sayısı ve Turizm Harcamalarının (TH) modele anlamlı bir katkı sağladığı ancak Yatak Sayısı (YS) ve İnternet Kullanıcı (IN) sayısı değişkenlerinin modele anlamlı bir katkı sağlamadığı görülmektedir. Ülkelerin Akdeniz'e kıyısı olup olmadığını gösteren (D1) ikili değişkeni modelden düşürülmüştür. Düzeltilmiş R^2 değeri %87 ve MSE değeri %30'dur.

Çizelge 6.26. Gower uzaklığı ile elde edilen Küme 3 için parametre tahminleri

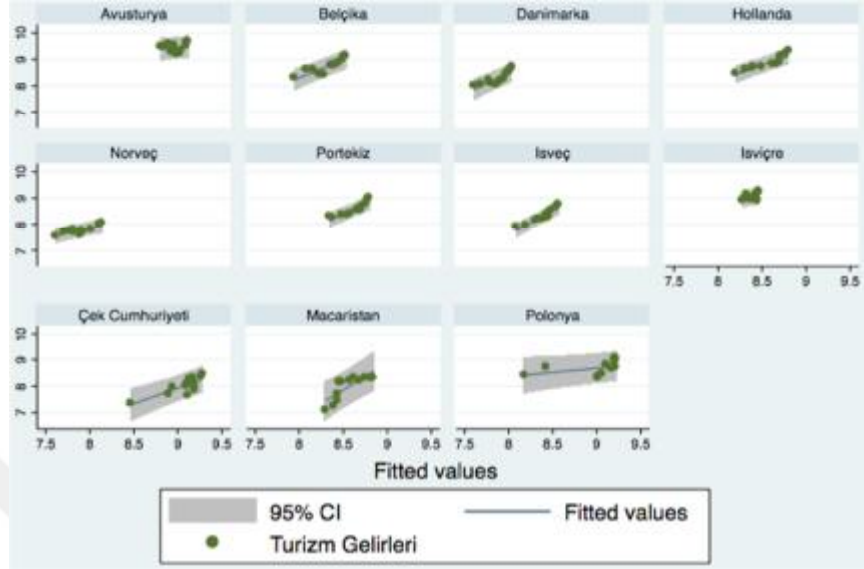
TG	Katsayı	Standart Hata	p-değeri	95% Güven Aralıkları	
GT	0.648	0.126	0.000	0.410	0.959
YS	0.094	0.106	0.391	-0.136	0.325
TH	0.562	0.150	0.003	0.235	0.889
IN	-0.009	0.021	0.659	-0.055	0.036
D1	0	Omitted			
sabit	-2.503	0.751	0.006	-4.141	-0.864
Düz. R^2	0.765				
MSE	0.767				

Çizelge 6.27. Önerilen uzaklık ile elde edilen Küme 3 için parametre tahminleri

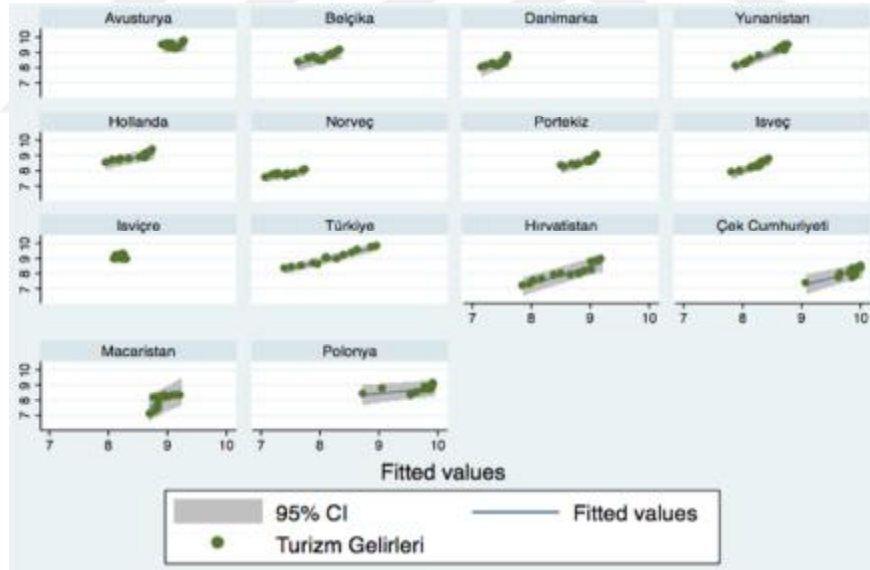
TG	Katsayı	Standart Hata	p-değeri	95% Güven Aralıkları	
GT	0.329	0.096	0.005	0.118	0.540
YS	0.154	0.092	0.121	-0.047	0.355
TH	0.656	0.161	0.002	0.303	1.008
IN	-0.031	0.021	0.168	-0.077	0.014
D1	0	Omitted			
sabit	-0.449	0.876	0.618	-2.359	1.461
Düz. R^2	0.736				
MSE	0.617				

Çizelge 6.26 ve Çizelge 6.27’de her iki uzaklık yönteminden elde edilen Küme 3 için Driscoll ve Kraay standart hatalarla sabit etkiler tahmin sonuçları yer almaktadır. Gower uzaklığı ile elde edilen Küme 3 için parametre tahminlerine bakıldığında Gelen Turist (GT) sayısı ve Turizm Harcamaları (TH) modele katkı sağlarken Yatak Sayısı (YS) ve İnternet Kullanıcı (IN) sayısının modele anlamlı bir katkısının olmadığı görülmektedir. Ülkelerin Akdeniz’e kıyısı olup olmadığını gösteren iki değişkeninden modelden düşürülmüştür. Düzeltilmiş R^2 değeri %76 ve MSE değeri %77’dir. Önerilen uzaklık ile elde edilen Küme 3 için parametre tahminlerinde Yatak Sayısı (YS) ve İnternet Kullanıcı (IN) sayısı değişkenlerin modele anlamlı katkı sağlamazken Gelen Turist (GT) sayısı ve Turizm Harcamaları (TH) modele anlamlı bir katkı sağlamaktadır. Ülkelerin Akdeniz’e kıyısı olup olmadığını gösteren iki değişkeni (D1) modelden düşürülmüştür. Düzeltilmiş R^2 değeri %74 ve MSE değeri %62’dir. Önerilen uzaklık kullanılarak elde edilen Küme 3’ün parametre tahminleri için %95 güven aralıklarının Gower uzaklığı kullanılarak elde edilen Küme 3’ün parametre tahminleri için %95 güven aralıklarından daha dar olduğu da söylenebilir.

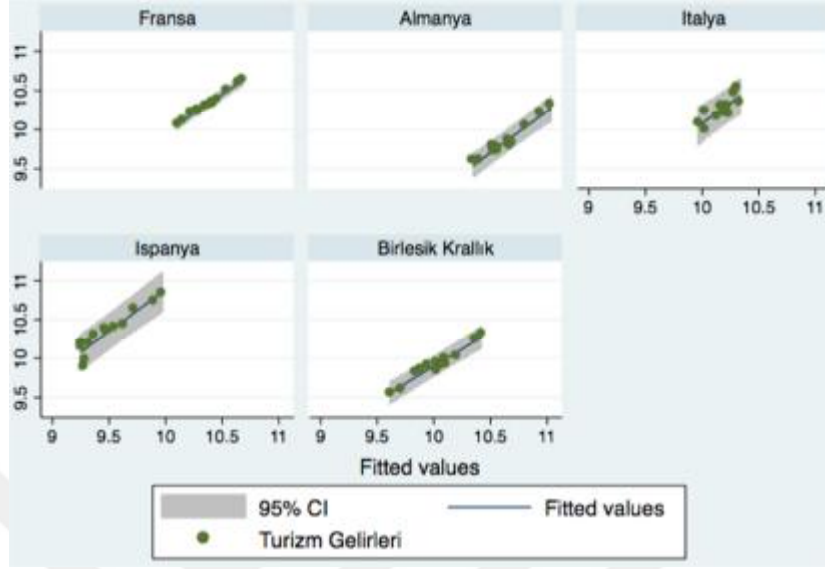
Şekil 6.25-6.29’de Gower ve önerilen uzaklık ölçümleri ile elde edilen kümeler için Turizm gelirleri değişkeninin uydurulmuş doğru ve güven aralıkları grafiği yer almaktadır.



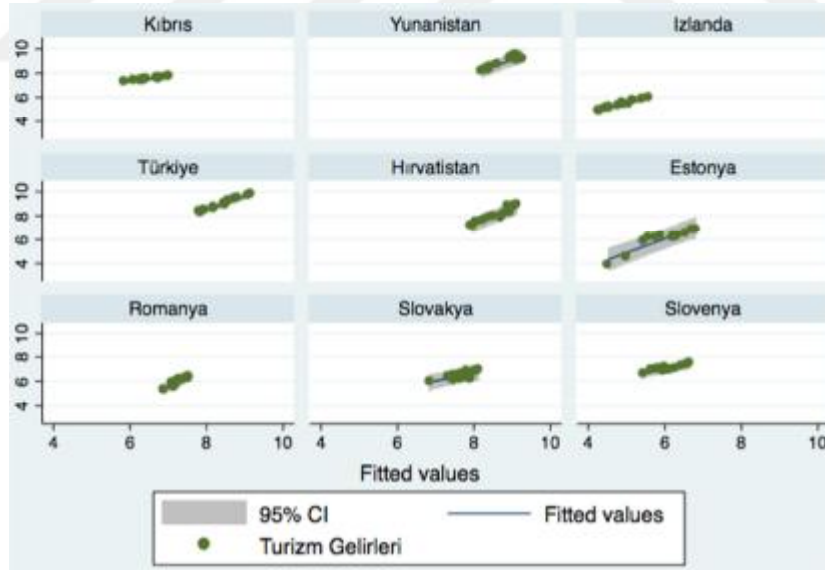
Şekil 6.25. Gower uzaklığı kullanılarak elde edilen birinci küme için birimlere göre Turizm Gelirlerinin uydurulmuş doğrusuyla birlikte güven aralıkları



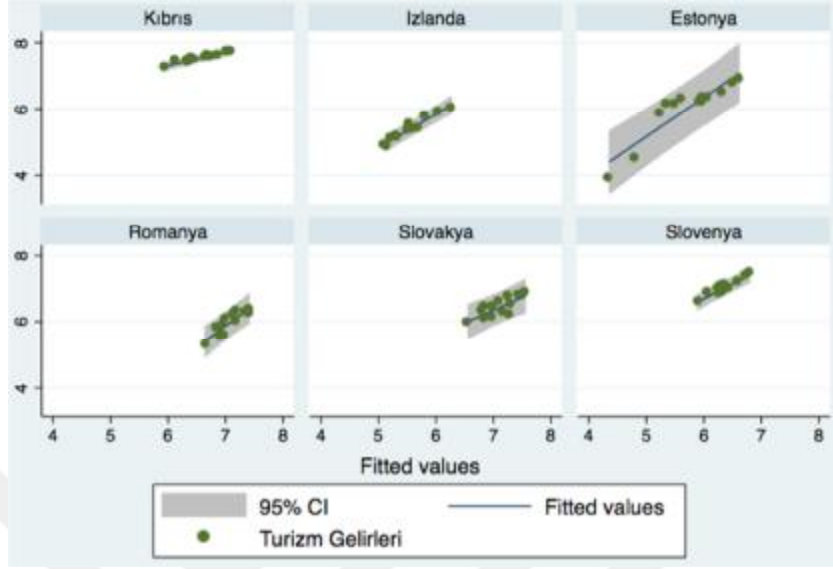
Şekil 6.26. Önerilen uzaklık kullanılarak elde edilen birinci küme için birimlere göre Turizm Gelirlerinin uydurulmuş doğrusuyla birlikte güven aralıkları



Şekil 6.27. Önerilen uzaklık ve Gower uzaklığı kullanılarak elde edilen üçüncü küme için birimlere göre Turizm Gelirlerinin uydurulmuş doğrusuyla birlikte güven aralıkları



Şekil 6.28. Gower uzaklığı kullanılarak elde edilen üçüncü küme için birimlere göre Turizm Gelirlerinin uydurulmuş doğrusuyla birlikte güven aralıkları



Şekil 6.29. Önerilen uzaklık kullanılarak elde edilen üçüncü küme için birimlere göre Turizm Gelirlerinin uydurulmuş doğrusuyla birlikte güven aralıkları

Şekil 6.25-6.29'dan görüldüğü gibi önerilen uzaklık ölçümü ile elde edilen kümelere uygulanan panel veri analizi sonuçlarına göre güven aralıklarının daha dar olduğu görülmektedir.



7. SONUÇLAR VE ÖNERİLER

Panel veri, zaman serisi ve yatay kesit verilerinin birleşiminden oluşmaktadır. Yatay kesit ve zaman serisi modellerinin iki aşamalı olarak kullanıldığı çalışmalar, söz konusu aşamaların panel veri yöntemiyle birleştirilmesiyle verilerin analizinde etkinlik sağlanabilmektedir. Panel veri modelleri son yıllarda bilgisayar paket programlarının sağladığı destek ile daha geniş kullanım alanları bulabilmektedir. Karışık olaylar için daha doğru analizleri mümkün kılan panel veri analizlerinde en önemli konu farklı birimlerin havuzlanabilirliğidir. Genel olarak, panel veri modelinde ilgilenilen eğim parametresinin birimlere göre homojen olduğu varsayımı yapılmaktadır. Ancak ekonomideki çoğu çalışma, parametre heterojenliğinin varlığını ortaya çıkarmıştır. Parametre heterojenliğinin göz ardı edilmesi ile birimler arasındaki heterojenlikten dolayı hem teorik hem de istatistiksel olarak geçerli tahminler yapılamayacaktır. Bu nedenle panel veri setinin birimlere göre heterojen olduğu dikkate alınarak panel veriye kümeleme analizi uygulanmıştır. Kümeleme analizi ile benzer özellik gösteren birimler kendi aralarında gruplara ayrılır ve küme içi homojen ve kümeler arası heterojen birimler elde edilir.

Çalışmada, karma değişkenli verileri kümelemede kullanılan Gower ve k-prototip uzaklık ölçümleri karma değişkenli panel veri setini kümelemek için uyarlanmıştır ve ayrıca yeni bir uzaklık ölçümü tanımlanmıştır. Bu uzaklık ölçümünün uygulanmasında ilk olarak üç boyutlu panel verinin iki boyutu çevrilerek tek boyuta dönüştürülür. Tek boyuttaki veri nümerik değişkenler ve ikili değişkenler olmak üzere iki alt boyuta ayrılır. Nümerik değişkenler $[0,1]$ aralığında normalleştirilerek uzaklık ölçümü hesaplanır. Aynı şekilde ikili verilerde basit eşleştirme yöntemiyle uzaklık ölçümü elde edilir. Elde edilen uzaklık ölçümleri birleştirilip toplamsal hiyerarşik kümeleme yöntemleri kullanılarak birimler kümelenir. Ayrıca karma değişkenli verileri kümelemek için kullanılan k-prototip algoritması ile de karma değişkenli panel veri kümelenmiştir.

Uygulamada 1993-2005 yılları arasında 25 ülkenin bulunduğu Turizm Gelirleri karma değişkenli panel veri seti ele alınmıştır. Bu veri seti, Gower, k-prototip ve önerilen uzaklık ölçümleri kullanılarak toplamsal hiyerarşik kümeleme yöntemleri ve k-prototip algoritması ile kümelendirilmiştir ve sonuçlar incelenmiştir. Toplamsal hiyerarşik kümeleme yöntemleri için CH ve DH indeks sonuçlarına göre uygun küme sayısı 3 olarak belirlenmiştir. Uzaklık ölçümlerinin performansını karşılaştırmak için kofenetik korelasyon, bağlantırlık, silhouette ve Dunn indeks sonuçları incelenmiştir. Kofenetik korelasyon sonuçlarına göre, önerilen uzaklığın kullanıldığı toplamsal hiyerarşik kümeleme yöntemlerinin çoğunluğunda iyi performans verdiği görülmüştür. Bağlantırlık ve Dunn indeks değerlerine göre de, önerilen uzaklığın kullanıldığı toplamsal hiyerarşik kümeleme yöntemlerinin çoğunluğunda iyi performans vermiştir. Silhouette indeksi sonuçlarında ise önerilen uzaklık, k-prototip uzaklığından kötü ancak Gower uzaklığından daha iyi sonuç vermiştir. Veri setine farklı uzaklık yöntemleri kullanılarak toplamsal hiyerarşik kümeleme yöntemleri ile kümeleme analizi yapıldıktan sonra k-prototip algoritması kullanılarak da kümeleme analizi yapılmıştır. Veri seti dört farklı λ değerine göre k-prototip algoritması ile kümelendirilmiştir. Her bir λ değeri için küme sayısına göre küme içi kareler toplamı hesaplanmıştır. Bu dört λ değeri içinde uygun küme sayısı 4 olarak bulunmuştur.

Elde edilen kümelerin panel veri analizine geçmeden önce hangi yöntem ile elde edilen kümelerin ele alınacağı önemli bir sorundur. Çalışmada, uzaklıkların performansı açısından toplamsal hiyerarşik yöntemlere göre seçim yapılmıştır. Buna göre toplamsal hiyerarşik yöntemlerden, Ward yöntemi üç uzaklıktada CH ve DH indeks sonuçlarına göre 3 küme önerdiğinden veri seti için uygun yöntem olarak belirlenmiştir. Üç uzaklık içinde ikinci kümedeki birimler aynı olmasına rağmen birinci ve üçüncü kümedeki birimler değişiklik göstermiştir. k-prototip ve önerilen uzaklık ölçümlerinin kullanıldığı Ward yöntemi ile elde edilen kümelerdeki birimler aynı olduğundan çalışma önerilen ve Gower uzaklığı üzerinden devam etmiştir. Bu iki uzaklık kullanılarak elde edilen her bir küme için

panel veri analizi ile parametre tahminleri yapılmıştır. Kullanılacak parametre tahmin yöntemini belirlemek için her bir kümeye yatay kesit bağımlılığı, birim kök testi, değişen varyans ve otokorelasyon testleri yapılmıştır. Yapılan testler sonucunda her iki uzaklık kullanılarak elde edilen kümelerin hepsinde yatay kesit bağımlılığı, birinci farkı alınan değişkenlerin durağan olduğu ve aynı zamanda değişen varyans ve otokorelasyon olduğu saptanmıştır. Bu nedenle parametre tahmini için modelde yatay kesit bağımlılığı, değişen varyans ve otokorelasyon varlığında dirençli tahmin edici olan Driscoll ve Kraay standart hatalarla sabit etkiler tahmin kullanılmıştır. Elde edilen parametre tahminlerinin güven aralıkları sonuçlarına göre önerilen uzaklıkla elde edilen kümelere uygulanan panel veri analizinin daha iyi sonuçlar verdiğini söylenebilir. Ayrıca her kümede modele katkı sağlayan değişkenlerin farklı olduğu görülmektedir. Bu nedenle panel veri analizi yapılmadan önce kümeleme analizi yapılması birbirine benzer birimleri bir araya getireceğinden daha doğru panel veri modelleri elde edilecektir.



KAYNAKLAR

- Ahmad, A., and Dey, L., 2007. A K-mean Clustering Algorithm for Mixed Numeric and Categorical Data. *Data & Knowledge Engineering*, 63(2):503-527.
- Anderberg, M. R., 1973. *Cluster analysis for applications*. Academic Press, New York, 359s.
- Andreica, M. E., and Andreica, M., 2014. Forecasts of Romanian industry employment using simulations and panel data models. *Romanian Journal of Economic Forecasting*, 17(2):130-140.
- Anselin, L., 1988. Lagrange Multiplier Test Diagnostics for Spatial Dependence and Spatial Heterogeneity. *Geographical Analysis*, 20(1):1-17.
- Arellano, M., 1987. Practitioners' Corner: Computing Robust Standard Errors for Within-groups Estimators. *Oxford bulletin of Economics and Statistics*, 49(4):431-434.
- Arı, A., ve Özcan, B., 2012. Hollanda Hastalığı: Gelismekte Olan Ülkeler Üzerine Bir Uygulama. *Sosyoekonomi*, 2:153.
- Aydın, N., ve Seven, A., 2015. İl Nüfus ve Vatandaşlık müdürlüklerinin İş Yoğunluğuna göre Hibrid Kümeleme ile Sınıflandırılması. *Yönetim ve Ekonomi Araştırmaları Dergisi*, 13(2):181-201.
- Bai, J., and Ng, S., 2004. A Panic Attack on Unit Roots and Cointegration. *Econometrica*, 72(4):1127-1177.
- Baltagi, B., 2005. *Econometric Analysis of Panel Data*. John Wiley & Sons, England, 302s.
- Baltagi, B. H., and Wu, P. X., 1999. Unequally Spaced Panel Data Regressions with AR (1) Disturbances. *Econometric Theory*, 15(6):814-823.
- Banerjee, A., 1999. Panel Data Unit Roots and Cointegration: an Overview. *Oxford Bulletin of Economics and Statistics*, 61(1):607-629.

- Baum, C. F., 2001. Residual Diagnostics for Cross-Section Time Series Regression models. *The Stata Journal*, 1(1):101-104.
- Beck, N., and Katz, J. N. (1995). What to do (and not to do) with Time-Series Cross-Section Data. *American political science review*, 89(3):634-647.
- Berrington, A., Smith, P., and Sturgis, P., 2006. An Overview of Methods for the Analysis of Panel Data. ESRC National Centre for Research Methods Briefing Paper, 57s.
- Bhargava, A., Franzini, L., and Narendranathan, W., 1982. Serial correlation and the Fixed Effects Model. *The Review of Economic Studies*, 49(4):533-549.
- Bobková, B., 2014. On Estimation of Gravity Equation: A Cluster Analysis (No. 37/2014). IES Working Paper.
- Bolshakova, N., and Azuaje, F., 2003. Cluster Validation Techniques for Genome Expression Data. *Signal processing*, 83(4):825-833.
- Bonzo, D. C., and Hermosilla, A. Y., 2002. Clustering Panel Data via Perturbed Adaptive Simulated Annealing and Genetic Algorithms. *Advances in Complex Systems*, 5(04):339-360.
- Born, B., and Breitung, J., 2016. Testing for Serial Correlation in Fixed-Effects Panel Data Models. *Econometric Reviews*, 35(7):1290-1316.
- Breitung, J., 2000. The local power of some unit root tests for panel data. In *Nonstationary panels, panel cointegration, and dynamic panels*, Advances in Econometrics, JAI Press, Amsterdam 161-177.
- Breitung, J., and Das, S., 2005. Panel Unit Root Tests under Cross-Sectional Dependence. *Statistica Neerlandica*, 59(4):414-433.
- Breusch, T. S., and Pagan, A. R., 1980. The Lagrange Multiplier Test and its Applications to Model Specification in Econometrics. *The Review of Economic Studies*, 47(1):239-253.
- Brock, G. Pihur, V., Datta, S. and Datta, S., 2008. clValid: An R package for cluster validation. *Journal of Statistical Software*, 25(4):1-22.

- Brown, M. B., and Forsythe, A. B. (1974). Robust Tests for the Equality of Variances. *Journal of the American Statistical Association*, 69(346):364-367.
- Calinski, T., and J. Harabasz. 1974. A Dendrite Method for Cluster Analysis. *Communications in Statistics* 3:1-27.
- Cameron, A. C., and Trivedi, P. K., 2005. *Microeconometrics: methods and applications*. Cambridge university press, New York, 1058s.
- Chang, Y., 2002. Nonlinear IV Unit Root Tests in Panels with Cross-Sectional Dependency. *Journal of Econometrics*, 110(2):261-292.
- Chang, Y., 2004. Bootstrap unit root tests in panels with cross-sectional dependency. *Journal of Econometrics*, 120(2):263-293.
- Choi, I., 2001. Unit Root Tests for Panel Data. *Journal of international money and Finance*, 20(2):249-272.
- Conway, K. S., and Kniesner, T. J., 1992. Estimating Labour Supply Disequilibrium with Fixed-Effects Random-Coefficients regression. *Applied Economics*, 24(7):781-789.
- Datong, G. M., Choji, N. M., and Nyam, I. A., 2017. Testing the Poolability of a Set of Panel Data. *International Journal of Innovative Mathematics, Statistics & Energy Policies*, 5(1):20-23
- De Hoyos, R. E., and Sarafidis, V., 2006. Testing for cross-sectional dependence in panel-data models. *Stata Journal*, 6(4):482.
- Demiralay, M., Çamurcu, A. Y., 2005. Cure, Agnes ve K- Means Algoritmalarındaki Kümeleme Yeteneklerinin Karşılaştırılması. *İstanbul Ticaret Üniversitesi Fen Bilimleri Dergisi*, 8:1-18.
- Dice, L. R., 1945. Measures of the amount of ecologic association between species. *Ecology*, 26(3): 297-302.
- Dickey, D. A., and Fuller, W. A., 1979. Distribution of the Estimators for Autoregressive Time Series with a Unit Root. *Journal of the American statistical association*, 74(366a):427-431.

- Driscoll, J. C., and Kraay, A. C., 1998. Consistent Covariance Matrix Estimation with Spatially Dependent Panel Data. *Review of economics and statistics*, 80(4):549-560.
- Duda, R. O., and Hart, P. E., 1973. *Pattern classification and scene analysis*. A Wiley-Interscience Publication, New York, 1973.
- Dunn J. C., 1974. Well Separated Clusters and Fuzzy Partitions. *Journal on Cybernetics*, 4:95–104.
- Eicker, F., 1967. Limit Theorems for Regressions with Unequal and Dependent Errors. In *Proceedings of the fifth Berkeley Symposium on Mathematical Statistics and Probability*, 1(1):59-82).
- Engle, R. F., and Granger, C. W., 1987. Co-integration and Error Correction: Representation, Estimation, and Testing. *Econometrica: journal of the Econometric Society*, 55(2):251-276.
- Erişoğlu, M., 2011. Uzaklık Ölçülerinin Kümeleme Analizine olan Etkilerinin İncelenmesi ve Geliştirilmesi. Çukurova Üniversitesi Doktora Tezi, Adana, 199s.
- Fomby, T. B., Hill, R. C., and Johnson, S. R., 1984. *Advanced econometric methods*. Springer Science & Business Media, 624s.
- Frees, E. W., 1995. Assessing Cross-sectional Correlation in Panel Data. *Journal of econometrics*, 69(2):393-414.
- Frees, E. W., 2004. *Longitudinal and Panel Data: Analysis and Applications for the Social Sciences*. Cambridge University Press, 497s.
- Friedman, M., 1937 . The Use of Ranks to Avoid the Assumption of Normality Implicit in the Analysis of variance. *Journal of the american statistical association*, 32(200):675-701.
- Froot, K. A., 1989. Consistent Covariance Matrix Estimation with Cross-Sectional dependence and Heteroskedasticity in Financial Data. *Journal of Financial and Quantitative Analysis*, 24(3):333-355.

- Gengenbach, C., Palm, F. C., and Urbain, J. P., 2009. Panel Unit Root Tests in the Presence of Cross-Sectional Dependencies: Comparison and Implications for Modelling. *Econometric Reviews*, 29(2):111-145.
- Gibert, K., and Cortés, U., 1997. Weighting Quantitative and Qualitative Variables in Clustering Methods. *Mathware & soft computing*. 4(3):251-266.
- Gonçalves, L.S.A, Rodrigues, R., Junior, A.T.A, Karasawa, M., and Sudre, C.P., 2008. Comparison of Multivariate Statistical Algorithms to Cluster Tomato Heirloom Accessions. *Genetics and Molecular Research*, 7(4):1289-1297.
- Goodall, D. W., 1966. A New Similarity Index Based on Probability. *Biometrics*, 22(4):882-907.
- Gordon, A. D., 1998. Cluster validation. In *Data science, classification, and related methods*, Springer, Tokyo, 22-39.
- Gowda, K. C., and Diday, E., 1991. Symbolic Clustering Using a New Dissimilarity Measure. *Pattern Recognition*, 24(6):567-578.
- Gower, J.C., 1971. A General Coefficient of Similarity and Some of its Properties. *Biometrics*, 27:857-874.
- Gujarati, D., 2004. *Basic Econometrics*. United States Military Academy, West Point, 1024s.
- Güriş, S., 2015. *Stata ile panel veri modelleri*. Der yayınevi, İstanbul, 318p.
- Hadri, K., 2000. Testing for Stationarity in Heterogeneous Panel Data. *The Econometrics Journal*, 3(2):148-161.
- Halkidi, M., Batistakis, Y., ve Vazirgiannis, M., 2002a. Cluster validity methods: part I. *ACM Sigmod Record*, 31(2):40-45.
- Halkidi, M., Batistakis, Y., ve Vazirgiannis, M., 2002b. Clustering validity checking methods: part II. *ACM Sigmod Record*, 31(3):19-27.
- Harris, D., and Inder, B., 1994. A Test of the Null Hypothesis of Cointegration: In *Nonstationary Time Series Analysis and Cointegration*. Hargreaves, Colin P., Ed., Oxford University Press, New York.

- Harris, R. D., and Tzavalis, E., 1999. Inference for Unit Roots in Dynamic Panels where the Time Dimension is Fixed. *Journal of econometrics*, 91(2):201-226.
- Hsiao, C., 2003. *Analysis of Panel Data*. Cambridge University Press, United Kingdom, 366s.
- Hou, F., and Ai, K., 2015. The Empirical Research of Relationship between Consumption and Income for Chinese Urban Residents. *Open Journal of Applied Sciences*, 5:251-258.
- He, Z., Xu, X., and Deng, S., 2002. Clustering Mixed Numeric and Categorical Data: A Cluster Ensemble Approach. Department of Computer Science And Engineering. Harbin Institute of Technology.
- Hineline, D. R., 2008. Parameter Heterogeneity in Growth Regressions. *Economics Letters*, 101(2):126-129.
- Hsu, C., Chen, C., and Su, Y., 2007. Hierarchical Clustering of Mixed Data Based on Distance Hirarchy. *Information Sciences an International Journal*, 177:4474-4492.
- Huang, Z., 1997. Clustering Large Data Sets with Mixed Numeric and Categorical Values. *Proceedings of the First Pacific Asia Knowledge Discovery and Data Mining Conference*, Singapore. 21-34.
- Huang, Z., 1998. Extensions to the K-Modes Algorithm for Clustering Large Data Sets with Categorical Values. *Data Mining and Knowledge Discovery*, 2(3):283-304.
- Huber, P. J., 1967. The behavior of maximum likelihood estimates under nonstandard conditions. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, 1(1):221-233.
- Ichino, M., and Yaguchi, H., 1994. Generalized Minkowski Metrics for Mixed Feature-Type Data Analysis. *IEEE Transactions on Systems, Man, and Cybernetics*, 24(4):698-708.

- Im, K. S., Pesaran, M. H., and Shin, Y., 1997. Testing for Unit Roots in Heterogeneous Panels. Department of Applied Economics, University of Cambridge, Working Paper, 9526.
- Im, K. S., Pesaran, M. H., and Shin, Y., 2003. Testing for unit roots in heterogeneous panels. *Journal of econometrics*, 115(1):53-74.
- Islam, T., Fiebig, D. G., and Meade, N., 2002. Modelling Multinational Telecommunications Demand with Limited Data. *International Journal of Forecasting*, 18(4):605-624.
- Jaccard, P., 1908. Nouvelles recherches sur la distribution florale. *Bull. Soc. Vaud. Sci. Nat.*, 44: 223-270.
- Jain, A. K., and Dubes, R. C., 1998. Algorithms for Clustering Data. New Jersey 07632: Prentice Hall, Englewood Cliffs, New Jersey, 320p.
- Johnson, S. C., 1967. Hierarchical Clustering Schemes. *Psychometrika*, 32(3):241-254.
- Johnson, R., and Wichern, D., 1998. Applied multivariate statistical methods. Prentice Hall, New Jersey, 828s.
- Juárez, M. A., and Steel, M. F., 2010. Model-Based Clustering of Non-Gaussian Panel Data Based on Skew-t Distributions. *Journal of Business & Economic Statistics*, 28(1):52-66.
- Kao, C., 1999. Spurious regression and residual-based tests for cointegration in panel data. *Journal of econometrics*, 90(1):1-44.
- Kapetanios, G., 2006. Cluster Analysis of Panel Data Sets Using Non-Standard Optimisation of Information Criteria. *Journal of Economic Dynamics and Control*, 30(8):1389-1408.
- Kaufman, L., and Rousseeuw, P. J., 1990. Findings Groups in Data: an Introduction to Cluster Analysis, John Willey and Sons, New Jersey, 342s.
- Kim, B. S., and Maddala, G. S., 1992. Estimation and Specification Analysis of Models of Dividend Behavior Based on Censored Panel Data. In *Panel Data Analysis*, 111-124.

- King, B., 1967. Step-wise Clustering Procedures. *Journal of the American Statistical Association*, 62(317):86-101.
- Kmenta, J., 1986. *Elements of econometrics*. Macmillan, New York, 786s.
- Konstantakis, K. N., Papageorgiou, T., Michaelides, P. G., and Tsionas, E.G., 2015. Economic Fluctuations and Fiscal Policy in Europe: A Political Business Cycles Approach Using Panel and Clustering. *Open Econ Rev.* 26: 971-998.
- Kwiatkowski, D., Phillips, P. C., Schmidt, P., and Shin, Y., 1992. Testing the Null Hypothesis of Stationarity Against the Alternative of a Unit Root: How Sure are we that Economic Time Series have a Unit Root?. *Journal of econometrics*, 54(1-3): 159-178.
- Levene, H. (1960). Robust tests for equality of variance. In 'Contributions to Probability and Statistics'. (Eds I Olkin, SG Ghurye, W Hoeffding, WG Madow, HB Mann), 278–292.
- Levin, A., Lin, C. F., and Chu, C. S. J., 2002. Unit Root Tests in Panel Data: Asymptotic and Finite-Sample Properties. *Journal of econometrics*, 108(1):1-24.
- Leybourne, S. J., and McCabe, B. P., 1994. A Consistent Test for a Unit Root. *Journal of Business & Economic Statistics*, 12(2):157-166.
- Li, C., and Biswas, G., 2002. Unsupervised Learning with Mixed Numeric and Nominal Data. *IEEE Transactions on Knowledge and Data Engineering*, 14(4):673-690.
- Li, X., Hipel, K. W., and Dang, Y., 2015. An improved Grey Relational Analysis Approach for Panel Data Clustering. *Expert Systems with Applications*, 42(23):9105-9116.
- Lim, J., Jun, J., Kim, S., and McLeod, D., 2012. A Framework for Clustering Mixed Attribute Type Datasets. *Proceeding of the fourth International Conference on Emerging Databases (EDB12)*, 92-100.

- Lin, C. C., ve Ng, S. (2012). Estimation of panel data models with parameter heterogeneity when group membership is unknown. *Journal of Econometric Methods*, 1(1):42-55.
- Lombardo, R., and Falcone, M., 2011. Crime and Economic Performance. A cluster analysis of panel data on Italy's NUTS 3 regions. Università della Calabria, Dipartimento di Economia, Statistica e Finanza (Ex Dipartimento di Economia e Statistica), Working Paper, 1-33.
- Lu, H., and Huang, S., 2011. Clustering Panel Data. SIAM International Workshop on Data Mining held in conjunction with the 2011 SIAM International Conference on Data Mining. 1-10.
- Maddala, G. S., ve Wu, S., 1999. A comparative study of unit root tests with panel data and a new simple test. *Oxford Bulletin of Economics and statistics*, 61(S1): 631-652.
- Mariel, P., López, C., and Fernández, K., 2006. Sales Advertising Relationship: an Application of Panel Data from the German Automobile Industry. *Prague Economic Papers*, 1(1):29-43.
- Martinez, W. L., Martinez A. R., 2005. *Exploratory Data Analysis with Matlab*, CRC Press, London, 405s.
- Matyas, L., and Sevestre, P., 1996. *The Econometrics of Panel Data. A Handbook of the Theory with Applications*. Springer, Berlin, 977s.
- Mayda, A. M., 2005. International migration: a panel data analysis of economic and non-economic determinants.
- McCoskey, S., and Kao, C., 1998. A Residual-Based Test of the Null of Cointegration in Panel Data. *Econometric reviews*, 17(1):57-84.
- Modha, D. S., and Spangler, W. S., 2003. Feature Weighting in K-means Clustering. *Machine learning*, 52(3):217-237.
- Moon, H. R., and Perron, B., 2004. Testing for a unit root in panels with dynamic factors. *Journal of econometrics*, 122(1):81-126.

- Mouchart, M., and Rombouts, J. V., 2005. Clustered Panel Data Models: An Efficient Approach for Nowcasting from Poor Data. *International Journal of Forecasting*, 21(3):577-594.
- Newey, W. K., and West, K. D., 1987. A Simple, Positive Semi-definite and Autocorrelation Consistent Covariance Matrix. *Econometrica*, 55(3): 703-708.
- Newey, W. K., and West, K. D., 1994. Automatic Lag Selection in Covariance Matrix Estimation. *The Review of Economic Studies*, 61(4):631-653.
- Nie, G., Chen, Y., Zhang, L., and Guo, Y., 2012. Credit Card Customer Analysis Based on Panel Data Clustering. *Procedia Computer Science*, 1(1):2489-2497.
- O'Connell, S., 1998. Improving Quality in Locum General Practice. *BMJ*, 317(7150):S2a-7150.
- Parks, R. W., 1967. Efficient Estimation of a System of Regression Equations when Disturbances are both Serially and Contemporaneously Correlated. *Journal of the American Statistical Association*, 62(318):500-509.
- Pazarlıoğlu, M. V., ve Gürler, Ö. K., 2007. Telekomünikasyon yatırımları ve ekonomik büyüme: panel veri yaklaşımı. *Finans Politik ve Ekonomik Yorumlar*, 44(508):35-43.
- Pedroni, P., 1995. Panel Cointegration: Asymptotic and Finite Sample Properties of Pooled Time Series Tests with an Application to the PPP Hypothesis. *Indiana University Working Papers in Economics*, 95-013.
- Pedroni, P., 1997a. Panel Cointegration: Asymptotic and Finite Sample Properties of Pooled Time Series Tests with an Application to the PPP Hypothesis, New Results. *Indiana University Working Paper in Economics*.
- Pedroni, P., 1997b, On the Role of Cross Sectional Dependency in Dynamic Panel Unit Root and Panel Cointegration Exchange Rate Studies. *Indiana University Working Paper in Economics*.

- Pesaran, M. H., 2004. General diagnostic tests for cross section dependence in panels.
- Pesaran, M. H., 2007. A simple panel unit root test in the presence of cross-section dependence. *Journal of applied econometrics*, 22(2):265-312.
- Pesaran, M. H., and Smith, R., 1995. Estimating long-run relationships from dynamic heterogeneous panels. *Journal of econometrics*, 68(1):79-113.
- Pesaran, M. H., Shin, Y., and Smith, R. P., 1999. Pooled Mean Group Estimation of Dynamic Heterogeneous Panels. *Journal of the American Statistical Association*, 94(446): 621-634.
- Pesaran, M. H., and Yamagata, T., 2005. Testing Slope Homogeneity in Large Panels. University of Southern California IEPR Working Paper, 1-19.
- Phillips, P. C., and Sul, D., 2003. Dynamic Panel Estimation and Homogeneity Testing under Cross Section Dependence. *The Econometrics Journal*, 6(1):217-259.
- Podesta, F., 2000. Recent Developments in Quantitative Comparative methodology: The case of pooled time series cross-section analysis. *DSS papers Soc*, 3(2):5-44.
- Rao, V. R., and Sabavala, D. J., 1981. Inference of hierarchical choice processes from panel data. *Journal of Consumer Research*, 8(1):85-96.
- Reddy, M. V. J., and Kavitha, B., 2012. Clustering the Mixed Numerical and Categorical Dataset Using Similarity Weight and Filter Method. *International Journal of Database Theory and Application*, 5:121-133.
- Ren, J., and Shi, S., 2009. Multivariable Panel Data Ordinal Clustering and Its Application in Competitive Strategy Identification of Appliance-wiring Listed Companies. *International Conference on Management Science & Engineering*, 16:14-16.
- Rogers, W. H, 1993. Regression Standard Errors in Clustered Samples. *Stata Technical Bulletin*, 3(13):19-23.

- Rogers, D. J., and Tanimoto, T. T., 1960. A computer program for classifying plants. *Science*, 132(3434): 1115-1118.
- Rousseeuw, P. J., 1987. Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis. *Journal of Computational and Applied Mathematics*, 20: 53–65.
- Sarafidis, V and Wansbeek, T., 2010. Cross-sectional Dependence in Panel Data Analysis. The University of Sydney MPRA Paper, No: 20367, 1-43.
- Sarafidis, V., and Wansbeek, T., 2012. Cross-sectional Dependence in Panel Data Analysis. *Econometric Reviews*, 31(5):483-531.
- Sarafidis, V., and Weber, N., 2009. To pool or not to pool: a partially heterogeneous framework. *Munich Personal RePEc Archive*, 1-36.
- Sarafidis, V., and Weber, N., 2015. A partially heterogeneous framework for analyzing panel data. *Oxford Bulletin of Economics and Statistics*, 77(2):274-296.
- Servi, T., 2009. Çok Değişkenli Karma Dağılım Modeline Dayalı Kümeleme Analizi, Çukurova Üniversitesi Doktora Tezi, Adana, 242s.
- Shih, M., Jheng, J., and Lai, L., 2010. A Two-Step Method for Clustering Mixed Categorical and Numeric Data. *Tamkang Journal of Science and Engineering*. 13:11-19.
- Shin, Y., 1994. A Residual-Based Test of the Null of Cointegration Against the Alternative of No Cointegration'. *Econometric Theory*, 19:91-115.
- Sokal, R. R., and Sneath, P. H. A., 1963. Principles of numerical taxonomy. W. H. Freeman and Company, San Francisco and London I, 963s.
- Soundaryadevi, M., and Jayashree, L., 2014. Clustering of Data with Mixed Attributes based on Unified Similarity Metric. *International Journal of Innovative Research in Computer and Communication Engineering*, 2:1865-1870.

- Stephenson, M., Ali, R. A., and Darlington, G. A., 2017. A Simple Approach to Analyzing Clustered Longitudinal Data. *Communications in Statistics-Simulation and Computation*, 46(5):3553-3562.
- Stock, J. H., and Watson, M. W., 2007. *Introduction to econometrics*. Pearson Addison Wesley, Boston , 796s.
- Sul, D., 2006. "Panel Unit Root Tests under Cross Sectional Dependence with Recursive Mean Adjustment. *Economics Letters*, 105(1):123-126.
- Tatoğlu, F. Y., 2013a. *Panel Veri Ekonometrisi*. Beta Basım, İstanbul, 306s.
- Tatoğlu, F. Y., 2013b. *İleri Panel Veri Analizi*. Beta Basım, İstanbul, 290s.
- Taylor, M. P., and Sarno, L., 1998. The Behavior of Real Exchange Rates during the Post-Bretton Woods period. *Journal of international Economics*, 46(2):281-312.
- Theodoridis, S., and Koutroumbas, K., 2006. Clustering: Basic Concepts. *Pattern Recognition*, 483-516.
- Thompson, S. B., 2011. Simple Formulas for Standard Errors that Cluster by both Firm and Time. *Journal of financial Economics*, 99(1):1-10.
- Tunay, K. B., ve Silpar, A. M., 2006. *Türk Ticari Bankacılık Sektöründe Karlılığa Dayalı Performans Analizi-I. Türkiye Bankalar Birliği Araştırma Tebliği Serisi*,1.
- Verbeke, G. M., and Molenberghs, G. G., 2000. *Linear Mixed Models for Longitudinal Data*. Springer, New York, 579s.
- Wei, M., Chow, T., and Chan, R., 2015. Clustering Heterogeneous Data with k-Means by Mutual Information-Based Unsupervised Feature Transformation. *Entropy*, 17:1535-1548.
- Westerlund, J., 2007. Testing for Error Correction in Panel Data. *Oxford Bulletin of Economics and statistics*, 69(6):709-748.
- White, H., 1980. A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity. *Econometrica: Journal of the Econometric Society*, 48(4):817-838.

- Wooldridge, J. M., 2002. *Econometric Analysis of Cross Section and Panel Data*. The MIT Press, London, 735s.
- Xu, R., and Wunsch, D. C., 2009. *Clustering*, John Wiley&Sons,:IEEE Press,New Jersey, 358s.
- Yang, R., 2012. *A Hierarchical Clustering and Validity Index for Mixed Data*. Iowa State University Graduate Theses and Dissertations, Ames, 66s.
- Zheng, B., and Li, S. 2014. Multivariable Panel Data Cluster Analysis and Its Application. *Computer Modeling &New Technologies*, 18:553-557.
- Zheng, T., Zhu, D., Wang, X., and Yu, B., 2009. Panel Data Clustering and Its Application to Discount Rate of B Stock in China. 2009 Second International Conference on Information and Computing Science. 163-166.
- Zubin, J., 1938. A technique for measuring like-mindedness. *The Journal of Abnormal and Social Psychology*, 33(4), 508-516.

ÖZGEÇMİŞ

Özlem AKAY 1986 yılında Burdur'un Gölhisar ilçesinde doğmuştur. 2004 yılında Selçuk Üniversitesi İstatistik Bölümü'nde başladığı lisans eğitiminden 2008 yılında mezun olmuştur. Aynı yıl Selçuk Üniversitesi İstatistik Anabilim Dalında başladığı yüksek lisans eğitimini 2011 yılında tamamlamıştır. 2013 yılında Çukurova Üniversitesi İstatistik Anabilim Dalında doktora programına başlamıştır. 2011 yılından itibaren Çukurova Üniversitesi İstatistik Bölümü'nde araştırma görevlisi olarak çalışmaktadır.