KATEGORİK VERİLER İÇİN KÜMELEME ANALİZİ

By AAY BE

WORD COUNT 2067
TIME SUBMITTED 01-SEP-2021 01:23PM
PAPER ID 76166073

KATEGORİK VERİLER İÇİN KÜMELEME ANALİZİ

Barış ERGÜL[1], Arzu ALTIN YAVUZ[2]

[1] Eskişehir Osmangazi Üniversitesi Fen Edebiyat Fakültesi İstatistik Bölümü, Eskişehir/ TÜRKİYE,

Orcid No: 0000-0002-1811-5143

[2] Eskişehir Osmangazi Üniversitesi Fen Edebiyat Fakültesi İstatistik Bölümü, Eskişehir/ TÜRKİYE

Orcid No: 0000-0002-3277-740X

Özet

Amaç: *Kümeleme Analizi, veri matrisinde yer alan ve doğal gruplamaları kesin olarak bilinmeyen birimleri, birbiri ile benzer olan alt kümelere ayırmaya yardımcı olan yöntemler topluluğudur. Kümeleme analizi; birimleri veya değişkenler arası benzerlik ya da farklılıklara dayalı olarak hesaplanan bazı ölçülerden yararlanarak homojen gruplara bölmek amacıyla kullanılır. Kümeleme Analizi; ekonomiden psikolojiye, tıptan ziraat bilimine ve biyolojiye kadar birçok bilim dalında sınıflandırma yapmak amacıyla kullanılmaktadır. Bu çalışmanın amacı, hayvan türlerini birbirlerine olan benzerliklerine ve farklılıklarına göre, kümeleme analizi kullanılarak sınıflandırmak ve sınıflandırma performanslarını karşılaştırmaktır.*

Yöntem: *Kümeleme Analizinde, veri setinde yer alan değişkenlerin ölçme düzeyleri, sıralayıcı ve sınıflayıcı olduğunda, klasik kümeleme teknikleri işlemez duruma gelir. Bu durumda, farklı benzerlik ölçütleri kullanılarak analizlere devam edilmesi gereği ortaya çıkmaktadır. Kategorik verilerde kullanılan benzerlik ölçütlerinden birisi de Jaccard benzerlik ölçüsüdür. Bu benzerlik ölçütü kullanılarak, kümeleme analizi yapılmaktadır. Diğer bir kümeleme tekniği ise, K-modes tekniğidir. Bu teknik, K-Ortalamalar tekniğinin bir uzantısı olarak karşımıza çıkmaktadır. Bu teknikte, ortalamalar yerine mod kullanılmaktadır. Kategorik verilerin kümeleme teknikleri içinde sağlam teknikler de vardır. Bunlardan en sık kullanılanı ROCK (RObust Clustering using linKs) tekniğidir. ROCK tekniği, aynı kümeden gözlemler için bağlantıların toplamı maksimize edilmesini ve farklı kümelerdeki gözlemler için bağlantıların toplamlarının en aza indirgenmesini dikkate alan fonksiyonun maksimize edilmesini sağlar.*

Bulgular: *İlk aşamada, dünya üzerinde yaşayan ve türleri bilinen 20 adet hayvan 6 değişken bakımından (sıcak kanlı olma durumu, uçabilme durumları, omurgalı olma durumları, neslinin tükenme durumu, gruplar halinde yaşama durumları, saç durumu) oluşturulan kategorik veri setinden yola çıkılmıştır. Veri setinde ilgili özelliğe sahip olmayan türe 1, sahip olan türe 2 kodu verilmiştir. Sonraki aşamada, uygun küme sayısına çeşitli ölçütler göz önüne alınarak karar verilmiştir. Sonraki aşamalarda ilgili kategorik veri setinden oluşan hayvan veri seti için Jaccard, K-modes ve ROCK kümeleme teknikleri uygulanmıştır. Her bir hayvanın temsil ettiği kümelere ait sınıflar belirlenmeye çalışılmıştır.*

Sonuç: *Hayvan türlerine ait sınıflandırma performansları çeşitli ölçütler bakımından karşılaştırılmıştır.*

Anahtar Kelimeler: *Hayvan, K-modes, ROCK, Jaccard, Kümeleme*

Cluster Analysis for Categorical Data

Abstract

Purpose: *The clustering analysis is that the natural groupings in the data matrix are a collection of methods that help separate unknown units. It is a clustering analysis that is used to divide some dimensions based on units or variables based on similarities or differences based on homogeneous groups. The clustering analysis is used in psychology from the economy to agricultural science and biology of medicine to graduate in many sciences. This study aims to classify and compare classification performance using clustering analysis according to their similarities and differences of animal species.*

Methods: *However, when the variables in the data set are measurement levels, ordinal and nominal, classical clustering techniques are not operating. In this case, it is necessary to continue analyzes using different similarity criteria. One of the similarity criteria used in categorical data is the measure of Jaccard similarity. By using this similarity measure, clustering analysis is performed. Another clustering technique is the K-Modes technique. This technique is an extension of the K-means technique as an extension. This technique uses modes instead of means. There are also robust techniques in clustering techniques of categorical data. The most commonly used is the Rock (Robust Clustering Using Links) technique. The Rock technique allows the total function to maximize the sum of the connections for observations and to maximize the functions that are considered to be minimized for the observations in different sets.*

Findings: *In the first stage that live on the world and the types of known 20 animals and, 6 variable (warm-blooded, can fly, vertebrate, endangered, conditions of living in groups, have hair) formed from the categorical data set. In the data set; Type 1 that do not have the relevant feature and Type 2 that the relevant feature. At the next stage, the appropriate cluster number is decided to consider a variety of criteria. In the next stages, Jaccard, K-Modes, and Rock clustering techniques were applied to the animal data set consisting of the categorical data set. Classification belonging to each animal represented is tried to be determined.*

Results: *The classification performances of animal species were compared in terms of various criteria.*

Keywords: *Animal, K-modes, ROCK, Jaccard, Cluster*

1.Giriş

Kümeleme analizi, doğal grupları kesin olarak bilinmeyen gözlemleri veya değişkenleri, birbiri ile benzer olan alt kümelere ayırma amacı olan yöntemler topluluğu olarak bilinmektedir. Başka bir ifadeyle, kümeleme analizi; gözlemler veya değişkenleri benzerlik ya da farklılıklarına göre hesaplanan bazı uzaklık ölçülerinden hareketle homojen gruplara bölmek amacıyla kullanılır (Johnson ve Wicherin, 1992).

Kümeleme analizinin uygulama aşamalarından ilki, veri matrisinin belirlenmesi aşamasıdır. Benzerlik matrisinin oluşturulması aşamasından sonra, uygun kümeleme yöntemi kullanılarak, gözlemlerin veya değişkenlerin uygun sayıda kümeye ayrılması aşaması gelmektedir. Elde edilen kümelerin yorumlanması aşaması da son aşama olarak belirlenmiştir (Tatlıdil, 1996).

Kümeleme Analizi; ekonomiden psikolojiye, tıptan ziraat bilimine ve biyolojiye kadar birçok bilim dalında sınıflandırma yapmak amacıyla kullanılmaktadır.

Kümeleme Analizinde en çok kullanılan kümeleme tekniklerinden birisi *k*-ortalamalar tekniğidir. Bu tekniğinin uygulanabilmesi için en önemli koşul, veri setindeki değişkenlerin en azından eşit aralıklı ölçek ile ölçülmesidir. Çünkü küme merkezleri oluşturulurken her bir iterasyon sonucu oluşan kümeler için değişkenlerin ortalamaları alınmaktadır. Kümeleme Analizinde, veri setinde yer alan değişkenlerin ölçme düzeyleri, sıralayıcı ve sınıflayıcı olduğunda, klasik kümeleme teknikleri işlemez duruma gelir. Bu durumda, farklı benzerlik ölçütleri kullanılarak analizlere devam edilmesi gereği ortaya çıkmaktadır.

Bu çalışmanın amacı, hayvan türlerini birbirlerine olan benzerliklerine ve farklılıklarına göre, kümeleme analizi kullanılarak sınıflandırmak ve sınıflandırma performanslarını karşılaştırmaktır. Hayvan türlerine ait veri setindeki değişkenler, sınıflayıcı ölçek ile ölçülmüş olduğundan, klasik kümeleme analizi kullanılmaz. İlgili veri seti için farklı benzerlik ölçütleri ve kümeleme analizi teknikleri geliştirilmiştir. Bu tekniklerin performanslarının karşılaştırılması ile istenen amaca ulaşılacaktır.

2.Yöntem

Kümeleme Analizinin en kritik konusu, uygun küme sayısı hakkında karar vermektir. Ancak günümüzde yayınlanan birçok makalede bu konuda kesin bir ölçüt bulunmamaktadır (Günay, 2008).

2.1.Uygun Küme Sayısı

Uygun küme sayısına, benzerlik matrisleri ve kümeleme tekniği göz önüne alınarak karar verilmelidir. Ancak son yıllarda, literatürde sıklıkla kullanılan 3 yöntem göze çarpmaktadır. Bunlar; Elbow (Dirsek), Silhouette (Siluet, Gölge) ve Gap (Uçurum) istatistiği yöntemleridir.
Elbow yönteminin temel amacı, küme içi toplam değişimin en aza indirilmesi sayesinde uygun küme sayısına karar vermektir. Bu yöntem, küme içi kareler toplamını küme sayısının fonksiyonu olarak tanımlar.
Silhouette yöntemi, farklı kümeler için ortalama bir siluet değeri hesaplar. Kümeler için çeşitli olası değerler üzerinden ortalama siluet değerini maksimize eden küme sayısı, uygun küme sayısıdır.
Gap yöntemi, farklı kümelerin farklı küme içi değişim değerleri için toplamı, verilerin H_0 hipotezi altında beklenen değerleri ile karşılaştırır. Optimum kümelerin sayısı, gap istatistiğini en üst düzeye çıkaran değer olacaktır (Kassambara, 2017).

2.2.Jaccard Uzaklık Matrisi

Uygun küme sayısına karar verildikten sonra, ilgili kümeleme analizi uygulanmaktadır. Ancak, bu çalışmada, kategorik verilerde kullanılan benzerlik ölçütlerinden birisi de Jaccard benzerlik ölçüsü olması nedeni ile kümeleme analizinde bu ölçüt kullanılarak da analiz yapılacak ve sonuçları tartışılacaktır.

Jaccard benzerlik katsayısı olarak da bilinen Jaccard endeksi, kümelerin benzerliğini karşılaştırmak için kullanılan bir istatistiktir. Kümeler arasındaki farklılıkları ölçen Jaccard benzerlik ölçüsü, Jaccard katsayısının tamamlayıcısıdır. Aşağıdaki formül yardımıyla hesaplanır (Shameem ve Ferdous, 2009).

$$Uzaklık\ J(A,B) = 1 - J(A,B) = \frac{|A \cup B| - |A \cap B|}{|A \cup B|} \quad (1)$$

Jaccard uzaklık matrisi oluşturulması ile aşamalı kümeleme analizlerinde ve k-ortalamalar tekniğinde rahatlıkla kullanılmaktadır.

2.3.k-Modes Tekniği

Kategorik verilerin kümelemesinde kullanılan diğer bir kümeleme tekniği ise, k-Modes tekniğidir. Bu teknik, k-Ortalamalar tekniğinin bir uzantısı olarak karşımıza çıkmaktadır. Bu teknikte, ortalamalar yerine mod kullanılmaktadır.

Yöntemin algoritması aşağıdaki gibidir: (Aranganayagi ve Thangavel, 2009).

- Başlangıç olarak *k* adet küme, modları kullanılarak seçilir.
- Adım (1) kullanarak, modu en yakın kümeye bir gözlem atayın. Her atamadan sonra kümenin modunu güncelleyin.
- Tüm gözlemleri ilgili kümeye atadıktan sonra, gözlemleri yeni mod değerini kullanarak tekrar elde edin ve kümeleri güncelleyin.
- Kümelerde değişiklik olmayana dek Adım (2) ve Adım (3) tekrar edin.

2.4.ROCK Tekniği

Kategorik verilerin kümeleme teknikleri içinde sağlam teknikler de vardır. Bunlardan en sık kullanılanı ROCK (RObust Clustering using linKs) tekniğidir. ROCK, linklerin kavramına dayanan sağlam bir hiyerarşik kümeleme algoritmasıdır. Ayrıca büyük veri kümelerine de uygulanabilir. ROCK, aralarında uzaklık olmayan veri noktalarını kullanır. Bu algoritmada, Küme Benzerliği, ortak olarak komşuları olan farklı kümelerdeki nokta sayısına dayanmaktadır. ROCK tekniği, aynı kümeden gözlemler için bağlantıların toplamının maksimize edilmesini ve farklı kümelerdeki gözlemler için bağlantıların toplamlarının en aza indirgenmesini dikkate alan fonksiyonun maksimize edilmesini sağlar (Rani ve Rohil, 2013; Elavarasi ve Akilandeswari, 2014).

3.Uygulama

İlk aşamada, dünya üzerinde yaşayan ve türleri bilinen 20 adet hayvan 6 değişken bakımından (sıcak kanlı olma durumu, uçabilme durumları, omurgalı olma durumları, neslinin tükenme durumu, gruplar halinde yaşama durumları, saç durumu) oluşturulan kategorik veri setinden yola çıkılmıştır. Veri setinde ilgili özelliğe sahip olmayan türe 1, sahip olan türe 0 kodu verilmiştir.

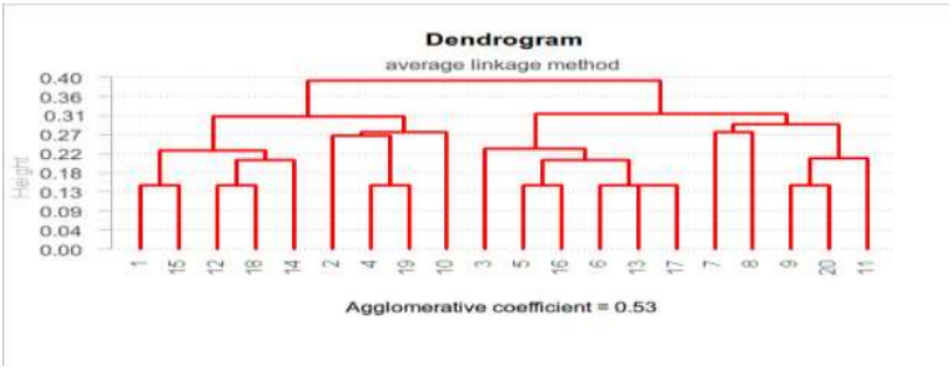
Sonraki aşamada uygun küme sayısına karar verilir. Bunun için R paket programının Hierarchical Cluster Analysis of Nominal Data paketinden yararlanılmıştır. Paket üzerindeki komut çalıştırıldığı zaman çeşitli kriterler bakımından uygun küme sayısı gösterilmektedir.

Optimal number of clusters based on the evaluation criteria:

	PSFM	PSFE	BIC	AIC	BK	SI
1	2	2	1	2	2	2

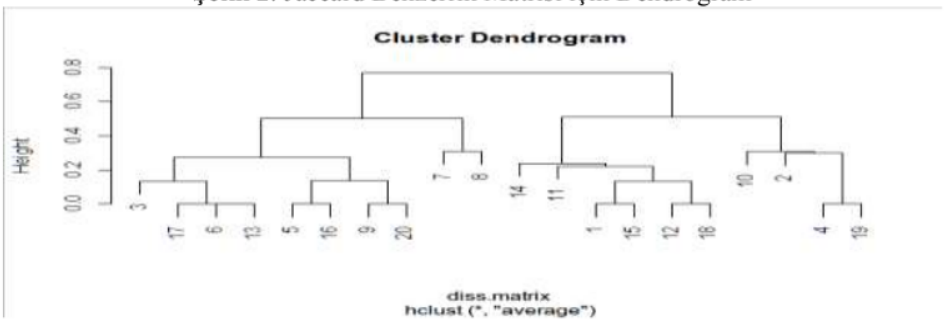
Hayvan türleri için uygun küme sayısının, Silhouette SI kriterine göre 2 olduğuna karar verilmiştir. 9 hayvan türü bir kümede, diğer geri kalan 11 hayvan türü de bir kümede olacak şekilde kümelenmişlerdir. Uygun küme sayısına ait dendrogram Şekil 1'de gösterilmiştir.

Şekil 1. Uygun Küme Sayısı için Dendrogram



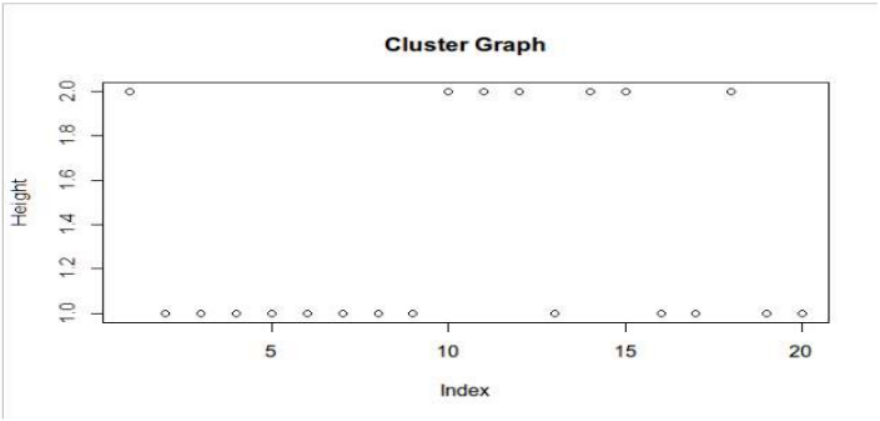
Uygun küme sayısına karar verildikten sonra kategorik veriler için kullanılan Jaccard benzerlik matrisini kullanarak hayvan türleri için kümeleme analizi yapılmıştır. Kümelme analizi için elde edilen dendrogram Şekil 2'de gösterildiği gibidir. 10 hayvan türü bir kümede, diğer geri kalan 10 hayvan türü de bir kümede olacak şekilde kümelenmişlerdir.

Şekil 2. Jaccard Benzerlik Matrisi için Dendrogram



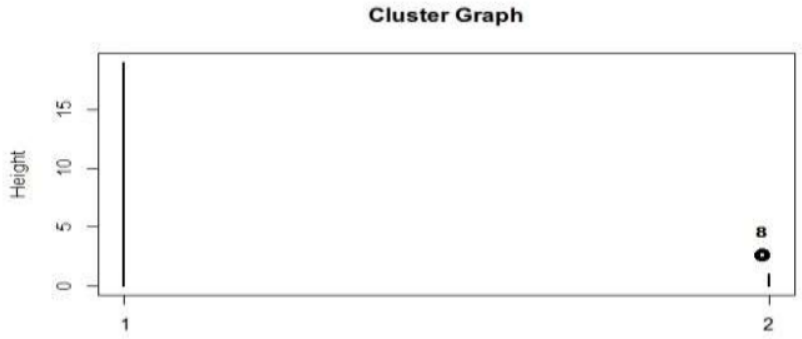
Sonraki aşamada, k-modes tekniği uygulanmış ve kümelere atanan hayvanlar Şekil 3'te gösterilmiştir. 7 hayvan türü bir kümede, diğer geri kalan 13 hayvan türü de bir kümede olacak şekilde kümelenmişlerdir.

Şekil 3. K-Modes için Hayvanların Kümelenmesi



Son olarak ROCK tekniği uygulanmış ve kümelere atanan hayvanlar Şekil 4'te gösterilmiştir. 1 hayvan türü bir kümede, diğer geri kalan 19 hayvan türü de bir kümede olacak şekilde kümelenmişlerdir.

Şekil 4. ROCK için Hayvanların Kümelenmesi



Sonraki aşamada ise, hayvanların 2 kümeye atanmaları için doğru sınıflandırma oranları ve R^2 değerleri her bir teknik için, verilerin kategorik olması sebebiyle lojistik regresyon analizi ile bulunmuştur. İlgili tablo, Tablo 1'de gösterildiği gibidir. Doğru Sınıflama Oranları ve R^2 değerleri birlikte değerlendirildiğinde, hayvan türleri verileri için en iyi tekniğin Jaccard benzerlik matrisi kullanılarak yapılan analiz olduğu görülmektedir.

Tablo 1. Teknikler için Doğru Sınıflandırma Oranları

Teknik	D.S.O	Cox-Snell R^2
Jaccard	**99.0**	**0.750**
k-Modes	99.0	0.726
ROCK	99.0	0.328

4.Sonuç ve Öneriler

Dünya üzerinde yaşayan ve türleri bilinen 20 adet hayvan 6 değişken bakımından oluşturulan kategorik veri setinden yola çıkılmıştır. İlgili veri seti için uygun küme sayısının 2 olması gerektiği kararlaştırılmış ve analizler 2 küme üzerinden yürütülmüştür.

Sonraki aşamalarda, kategorik veriler için oluşturulan veri setine ilgili kümeleme analizleri uygulanmış ve hayvan türlerinin kümelenmesi sağlanmıştır. Jaccard benzerlik matrisi kullanılarak uygulanan kümeleme analizi sonucuna göre, 10 hayvan türü bir kümede, diğer geri kalan 10 hayvan türü de bir kümede olacak şekilde kümelenmişlerdir. K-Modes tekniği kullanılarak uygulanan kümeleme analizi sonucuna göre, 7 hayvan türü bir kümede, diğer geri kalan 13 hayvan türü de bir kümede olacak şekilde kümelendikleri tespit edilmiştir. ROCK tekniği kullanılarak, 1 hayvan türü bir kümede, diğer geri kalan 19 hayvan türü de bir kümede olacak şekilde kümelendikleri görülmüştür. Sonrasında, ilgili tekniklerin doğru sınıflama oranları ve R^2 değerleri hesaplanmış ve iki kriter göz önüne alınarak, Jaccard benzerlik matrisi kullanılarak yapılan kümeleme analizinin hayvan türlerini sınıflamada en iyi teknik olduğu sonucuna ulaşılmıştır.

Kaynakça

ARANGANAYAGI, S. ve THANGAVEL, K., (2009), Improved K-Modes for Categorical Clustering Using Weighted Dissimilarity Measure, World Academy of Science, Engineering and Technology International Journal of Computer and Information Engineering, 3 (3), 729-735.

ELAVARASI, S.A. ve AKILANDESWARI, J., (2014), Survey on Clustering Algorithm and Similarity Measure for Categorical Data, Journal on Soft Computing, 4 (2), 715-722.

GÜNAY, A.C., (2008), Kümeleme Analizinde Küme Sayısının Belirlenmesi Üzerine Bir Çalışma, Ankara Üniversitesi Fen Bilimleri Enstitüsü, Yüksek Lisans Tezi.

JOHNSON, R. ve WICHERN, D., (1992), Applied Multivariate Statistical Analysis, 3.th ed., Prentice Hall, USA.

KASSAMBARA, A., (2017), Practical Guide to Cluster Analysis in R, STHDA.

RANI, Y. ve ROHIL, H. (2013), A Study of Hierarchical Clustering Algorithm, International Journal of Information and Computation Technology, 3 (10), 1115-1122.

SHAMEEM, M. ve FERDOUS, R., (2009), An efficient K-Means Algorithm integrated with Jaccard Distance Measure for Document Clustering, IEEE.

TATLIDİL, H., (1996), Uygulamalı Çok Değişkenli İstatistiksel Analiz, Cem Ofset Ltd. Şti., Ankara.

KATEGORİK VERİLER İÇİN KÜMELEME ANALİZİ

ORIGINALITY REPORT

16%
SIMILARITY INDEX

PRIMARY SOURCES

1	dergipark.org.tr Internet	131 words — 6%
2	iibf.ogu.edu.tr Internet	42 words — 2%
3	app.trdizin.gov.tr Internet	22 words — 1%
4	Partha Sarathi Bishnu, Vandana Bhattacherjee. "Software cost estimation based on modified K-Modes clustering Algorithm", Natural Computing, 2015 Crossref	21 words — 1%
5	dspace.lboro.ac.uk Internet	21 words — 1%
6	etheses.whiterose.ac.uk Internet	21 words — 1%
7	ikee.lib.auth.gr Internet	18 words — 1%
8	www.itobiad.com Internet	13 words — 1%
9	juti.if.its.ac.id Internet	12 words — 1%

10	scindeks.ceon.rs Internet	11 words — 1%
11	Kyoungok Kim. "A weighted k-modes clustering using new weighting method based on within-cluster and between-cluster impurity measures", Journal of Intelligent & Fuzzy Systems, 2017 Crossref	9 words — < 1%
12	YAŞAR, Ercan and YAŞAR, Mine. "Küresel Servet Eşitsizliği ve Çokuluslu Bir Sınıflama", Melih Topaloğlu, 2017. Publications	8 words — < 1%
13	dspace.gazi.edu.tr Internet	8 words — < 1%

EXCLUDE QUOTES	OFF	EXCLUDE MATCHES	OFF
EXCLUDE BIBLIOGRAPHY	OFF		