MerQur
Veri Bilimi ve Yöntemleri Dergisi
Journal of Data Science and Methods

Cilt 1, Sayı 1
2026
ISSN: — / e-ISSN: —

MerQur'da Kümeleme ve Boyut İndirgeme: K-Means'ten UMAP'a Denetimsiz Öğrenme

Clustering and Dimensionality Reduction in MerQur: Unsupervised Learning from K-Means to UMAP

Ömer K. Örücü [1]

[1] Süleyman Demirel Üniversitesi, Mimarlık Fakültesi, Peyzaj Mimarlığı Bölümü, Isparta/Türkiye ORCID: 0000-0002-2162-7553 E-posta: omerorucu@sdu.edu.tr · Resmi site: https://www.sekizgenacademy.com/journals/index.php/merqur/tr/index

Yazışmadan sorumlu yazar: Ömer K. Örücü (omerorucu@sdu.edu.tr)

Tür: Davetli Editöryal Sunum / Invited Editorial Showcase **Geliş:** 2026-05-17 · **Kabul:** 2026-05-17 · **Yayım:** 2026-05-17 **DOI:** — *(ISSN başvurusu sonrası eklenecek)*

Öz

Denetimsiz öğrenme yöntemleri — kümeleme ve boyut indirgeme — etiketsiz verilerden yapı çıkarmanın iki ana yoludur. Kümeleme benzer gözlemleri gruplara ayırırken, boyut indirgeme yüksek boyutlu verinin az boyutlu (genellikle 2D veya 3D) bir temsile dönüştürülmesini sağlar. Bu çalışmada **MerQur** masaüstü yazılımının Kümeleme kategorisinde sunulan **7 analiz** ayrıntılı olarak tanıtılmıştır: K-Means Kümeleme, Hiyerarşik Kümeleme, DBSCAN, Temel Bileşenler Analizi (PCA), t-Distributed Stochastic Neighbour Embedding (t-SNE), Multidimensional Scaling (MDS) ve Uniform Manifold Approximation and Projection (UMAP). Her analiz için (i) yöntemin temeli ve denetimsiz öğrenmedeki yeri, (ii) hiperparametreler ve seçim stratejileri (K seçimi: elbow / silhouette / gap statistic; DBSCAN için ε ve min_samples; PCA için bileşen sayısı; UMAP için komşu sayısı), (iii) MerQur'daki form alanları ve seçenekler, (iv) raporlanan istatistikler ve görselleştirme çıktıları (silhouette grafiği, dendrogram, biplot, 2D embedding), ve (v) tipik bir araştırma sorusu için yorumlama önerisi sunulmuştur. Geometrik mesafe-tabanlı yöntemler (K-Means, Hiyerarşik) ile yoğunluk-tabanlı yaklaşımlar (DBSCAN) arasındaki ayrım; doğrusal (PCA, MDS) ile doğrusal olmayan boyut indirgemeler (t-SNE, UMAP) arasındaki tasarım farklılıkları tartışılmıştır. Sonuç olarak MerQur'un Kümeleme kategorisi, klasik istatistiksel boyut indirgemeden modern manifold öğrenmesine, geometrik gruplama yöntemlerinden yoğunluk-tabanlı aykırı tespiti edebilen kümeleme algoritmalarına uzanan kapsamı tek bir grafik arayüzde toplamaktadır.

Anahtar Kelimeler: kümeleme, K-Means, DBSCAN, hiyerarşik, PCA, t-SNE, UMAP, MDS, denetimsiz öğrenme, MerQur

Abstract

Unsupervised learning methods — clustering and dimensionality reduction — are the two main ways of extracting structure from unlabelled data. Clustering groups similar observations, while dimensionality reduction transforms high-dimensional data into a low-dimensional (typically 2D or 3D) representation. This study introduces in detail the **7 analyses** offered in MerQur's Clustering category: K-Means Clustering, Hierarchical Clustering, DBSCAN, Principal Component Analysis (PCA), t-Distributed Stochastic Neighbour Embedding (t-SNE), Multidimensional Scaling (MDS), and Uniform Manifold Approximation and Projection (UMAP). For each: (i) the basis of the method and its place in unsupervised learning, (ii) hyperparameters and selection strategies, (iii) form fields and options in MerQur, (iv) reported statistics and visualisation outputs, and (v) interpretation guidance for a typical research question. The distinction between geometric distance-based methods (K-Means, Hierarchical) and density-based approaches (DBSCAN); the design differences between linear (PCA, MDS) and non-linear dimensionality reductions (t-SNE, UMAP) are discussed. Overall, MerQur's Clustering category brings together within a single graphical interface a spectrum spanning classical statistical dimensionality reduction, modern manifold learning, geometric grouping methods, and density-based clustering algorithms capable of outlier detection.

Keywords: clustering, K-Means, DBSCAN, hierarchical, PCA, t-SNE, UMAP, MDS, unsupervised learning, MerQur

1. Giriş

Denetimsiz öğrenme (unsupervised learning), gözlemlere eşlik eden etiket bilgisi olmadığı durumlarda veriden yapısal örüntü çıkarma çalışmasıdır. İki ana koldan oluşur: **kümeleme** — benzer gözlemleri aynı kümeye atama — ve **boyut indirgeme** — yüksek boyutlu veriyi düşük boyutlu bir temsile dönüştürme. Bu iki kol sıklıkla birlikte kullanılır: önce boyut indirgenir, sonra düşük-boyutlu uzayda kümeleme yapılır (Hastie ve ark., 2009).

Klasik istatistiğin denetimsiz yöntemleri (PCA, hiyerarşik kümeleme, MDS) 20. yüzyılın ilk yarısında geliştirildi. 1990'larda DBSCAN gibi yoğunluk-tabanlı algoritmalar mekânsal veri madenciliğinde ön plana çıktı. 2008'de **t-SNE** ve 2018'de **UMAP**, doğrusal olmayan manifold öğrenmesinin modern temsilcileri olarak yüksek-boyutlu veri görselleştirmesinde standart araç haline geldi.

Bu çalışmanın amacı, **MerQur** masaüstü yazılımının Kümeleme kategorisinde yer alan 7 analizi sistemli olarak tanıtmaktır. Kategori adı "Kümeleme" olsa da, boyut indirgeme yöntemleri (PCA, t-SNE, MDS, UMAP) de aynı kategoride bulunur — çünkü denetimsiz öğrenme akışında bu iki tür yöntem birbirini tamamlar.

2. K-Means Kümeleme

2.1 Yöntem

K-Means, K adet kümeyi temsil eden centroid'ler etrafında, gözlemleri en yakın centroid'e atayarak çalışan iteratif bir algoritmadır (MacQueen, 1967). Amacı küme-içi varyansı (within-cluster sum of squares, WCSS) minimize etmektir:

$$\min \sum_{i=1}^{K} \sum_{x \in C_i} \| x - \mu_i \|^2$$

K seçimi: elbow method (WCSS — K grafiği), silhouette skoru, gap statistic. Önemli sınırlılık: K önceden belirlenir; küresel kümelere varsayar; aykırı değerlere duyarlıdır.

2.2 MerQur'da uygulama

İstatistik → Kümeleme → K-Means Kümeleme

. Form alanları:

- **Özellikler:** sayısal değişken seti
- **K (küme sayısı):** sayı veya otomatik (elbow + silhouette)
- **Başlatma:** k-means++ (varsayılan), random
- **Standardizasyon:** dahil/hariç (varsayılan dahil — değişkenler farklı ölçekteyse şart)
- **n_init:** 10 (farklı başlangıçlar)

Çıktıda her gözlemin küme atanması, küme centroid'leri, küme-içi SS, silhouette skoru (her gözlem + ortalama), elbow grafiği, **küme dağılımı 2D scatter** (PCA boyutlarında).

2.3 Uygulama örneği

Müşteri segmentasyonu: 5 özellik (yaş, gelir, alışveriş sıklığı, sepet ortalama, sadakat skoru). Elbow ve silhouette K = 4 önerir; silhouette ortalaması = 0.46 (orta-iyi). Kümeler: gençler-yüksek harcamalı, yaşlılar-düşük frekanslı, vd.

3. Hiyerarşik Kümeleme

3.1 Yöntem

Hiyerarşik kümeleme, gözlemler arasındaki **mesafe matrisini** kullanarak bir **dendrogram** üretir. İki yaklaşım vardır:

- **Agglomerative (alttan-yukarı):** her gözlem ayrı kümeyle başlar, en yakın iki küme birleştirilir, tüm gözlemler tek kümede toplanana kadar devam eder
- **Divisive (yukarıdan-aşağı):** tüm gözlemler tek kümede başlar, en heterojen küme bölünür

Birleştirme kriterleri (linkage): single (en yakın çift), complete (en uzak çift), average, Ward (varyans artışını minimize eden) — Ward en popüler.

K seçimi dendrogram üzerinde **kesim çizgisi** ile yapılır.

3.2 MerQur'da uygulama

İstatistik → Kümeleme → Hiyerarşik Kümeleme

. Form alanları:

- **Özellikler:** sayısal değişken seti
- **Mesafe metriği:** Euclidean (varsayılan), Manhattan, Cosine
- **Linkage:** Ward (varsayılan), single, complete, average
- **Kesim seçimi:** K kümede veya yükseklik eşiği

Çıktıda dendrogram, küme atamaları, kümeler arası mesafe matrisi, silhouette analizi.

3.3 Uygulama örneği

20 değişkenli psikometrik veri. Ward linkage + Euclidean, dendrogram 3 doğal küme önerir. Silhouette = 0.42, kümeler psikolojik tipler olarak yorumlanır.

4. DBSCAN

4.1 Yöntem

Density-Based Spatial Clustering of Applications with Noise (DBSCAN), bir gözlemin **yoğunluk komşuluğunda** yeterli sayıda komşusu varsa onu küme çekirdeği olarak işaretler; çekirdek-bağlantılı gözlemleri aynı kümeye dahil eder; izole gözlemleri **gürültü** (noise/outlier) olarak işaretler (Ester ve ark., 1996). İki temel parametre: ε (komşuluk yarıçapı) ve *min_samples* (çekirdek koşulu için min komşu sayısı).

Avantajları: K önceden belirtilmez; küresel olmayan küme şekilleri (uzun, eğri) yakalanır; aykırı değerler doğal olarak ayrılır. Sınırlılığı: farklı yoğunluklu kümelere zayıf.

4.2 MerQur'da uygulama

İstatistik → Kümeleme → DBSCAN

. Form alanları:

- **Özellikler:** sayısal değişken seti
- **ε (eps):** sayı veya k-distance grafiğinden öneri
- **min_samples:** genelde 2 × özellik sayısı

Çıktıda küme atamaları (-1 = gürültü), küme sayısı (otomatik bulunan), her küme için silhouette, k-distance grafiği (ε seçim yardımı), 2D scatter.

4.3 Uygulama örneği

Mekânsal nokta dağılımı (örn. ağaç koordinatları). DBSCAN 3 küme + 18 aykırı nokta bulur; K-Means'in yapamayacağı eğri küme şekillerini başarıyla yakalar.

5. Principal Component Analysis (PCA)

5.1 Yöntem

PCA, korelasyonlu çok sayıda değişkeni birbirinden bağımsız (ortogonal) **temel bileşenlere** indirger. Her bileşen, açıklanan varyans büyüklüğüne göre sıralanır. İlk birkaç bileşen genellikle toplam varyansın büyük bir kısmını kapsar.

Yorumlama: her bileşen, orijinal değişkenlerin **doğrusal kombinasyonudur**; yükler (loadings) hangi değişkenin bileşene ne kadar katkı verdiğini gösterir.

5.2 MerQur'da uygulama

İstatistik → Kümeleme → PCA

. Form alanları:

- **Özellikler:** sayısal değişken seti
- **Standardizasyon:** dahil (varsayılan — farklı ölçekler için şart)
- **Bileşen sayısı:** sayı / varyans eşiği / Kaiser kriteri / scree

Çıktıda özdeğerler, varyans yüzdeleri (kümülatif), bileşen yükleri, bireysel skorlar, **biplot** (gözlemler + değişken vektörleri), **scree grafiği**.

5.3 Uygulama örneği

19 biyoklimatik değişken (WorldClim). İlk 3 bileşen %78 varyansı açıklıyor. PC1 sıcaklık, PC2 yağış, PC3 mevsimsellik olarak yorumlanır. Bu 3 PC sonraki niş modellemede kullanılır.

6. t-SNE

6.1 Yöntem

t-Distributed Stochastic Neighbour Embedding, yüksek-boyutlu komşuluk olasılık dağılımını düşük-boyutlu (genellikle 2D) bir uzaya **KL diverjansını minimize ederek** yansıtır (van der Maaten & Hinton, 2008). Yerel yapıyı korur — yakın komşular yakın kalır; global mesafeler tam korunmaz.

Önemli hiperparametre: **perplexity** (komşuluk büyüklüğü) — tipik 5–50 arası. Yorum: t-SNE haritaları görsel cluster keşfi için ideal ama mesafeler nicel yorumlanamaz.

6.2 MerQur'da uygulama

İstatistik → Kümeleme → t-SNE

. Form alanları:

- **Özellikler:** sayısal değişken seti
- **Perplexity:** 30 (varsayılan)
- **Çıktı boyutu:** 2 (varsayılan), 3
- **n_iter:** 1000 (varsayılan)
- **Random state:** sayı (tekrar üretilebilirlik için)

Çıktıda 2D/3D scatter (etiket varsa renkli), KL diverjans kalıntı değeri.

6.3 Uygulama örneği

500 hastanın 30 biyomarker profili. t-SNE 2D haritada 4 belirgin küme — sonradan klinik tanı kategorileriyle örtüşür.

7. Multidimensional Scaling (MDS)

7.1 Yöntem

MDS, gözlemler arası verilen bir mesafe matrisini, düşük boyutlu bir uzayda **mesafeleri en iyi koruyarak** yerleştirir. **Klasik (metric) MDS** Öklid mesafelerini, **non-metric MDS (NMDS)** sıralı mesafe ilişkilerini

korur — NMDS özellikle ekoloji ve psikometride yaygındır.

7.2 MerQur'da uygulama

İstatistik → Kümeleme → Multidimensional Scaling (MDS)

. Form alanları:

- **Mesafe matrisi:** verilen veya hesaplanacak özellikler
- **Tür:** Classical / Non-metric
- **Boyut sayısı:** 2 (varsayılan), 3
- **Stress hedefi:** otomatik

Çıktıda 2D/3D koordinatlar, Shepard diyagramı (stress diagnosis), stress değeri (yorum: < 0.05 mükemmel, 0.1 iyi, 0.2 kabul edilebilir).

7.3 Uygulama örneği

Bitki toplulukları arası Bray-Curtis dissimilarity matrisi. NMDS 2D, stress = 0.14 (iyi). Topluluk türleri görsel olarak belirgin gruplara ayrışır.

8. UMAP

8.1 Yöntem

Uniform Manifold Approximation and Projection, t-SNE'ye benzer ama matematiksel olarak topolojik manifold öğrenmesine dayanır (McInnes ve ark., 2018). Avantajları: t-SNE'den hızlı, hem yerel hem global yapıyı görece daha iyi korur, daha büyük veri setlerine ölçeklenir.

Önemli hiperparametreler: $n\_neighbors$ (yerel-global denge — 5 yerel, 50 global), min_dist (kümelerin ne kadar sıkı olacağı).

8.2 MerQur'da uygulama

İstatistik → Kümeleme → UMAP

. Form alanları:

- **Özellikler:** sayısal değişken seti
- **n_neighbors:** 15 (varsayılan)
- **min_dist:** 0.1 (varsayılan)
- **Çıktı boyutu:** 2 (varsayılan), 3
- **Mesafe metriği:** Euclidean (varsayılan), Manhattan, Cosine

Çıktıda 2D/3D scatter, varsa kategori etiketiyle renkli.

8.3 Uygulama örneği

10.000 hastanın 50 özelliği. UMAP 2D haritada 6 küme; t-SNE'ye göre 5x daha hızlı (45 saniye vs 4 dakika). Kümeler hastalık alt-tiplerine karşılık geliyor.

9. Yöntem Seçim Rehberi

Tablo 1 ana karar eksenlerini özetler.

Tablo 1. Denetimsiz yöntem seçim rehberi.

Senaryo	Önerilen yöntem
Önceden bilinen K + küresel kümeler + hız önemli	K-Means
Hiyerarşik küme yapısı + dendrogram istenir	Hiyerarşik (Ward)
Bilinmeyen K + aykırı tespit + küresel olmayan şekil	DBSCAN
Doğrusal indirgeme + yorumlanabilir bileşenler	PCA
Verilen mesafe matrisi + 2D görselleştirme	MDS (klasik / NMDS)
Görsel keşif + yerel yapı + küçük-orta veri	t-SNE
Büyük veri + global + yerel yapı + hız	UMAP

10. Karşılaştırmalı Değerlendirme

Tablo 2, MerQur'un Kümeleme kategorisindeki 7 analizin diğer GUI tabanlı açık erişimli alternatiflerle karşılaştırmasını özetler.

Tablo 2. Kümeleme ailesinin MerQur, JASP, jamovi ve PSPP'deki desteği.

Analiz	MerQur	JASP	jamovi	PSPP
K-Means	✓	+ (modül)	+ (modül)	✓
Hiyerarşik	✓	+	+	✓
DBSCAN	✓	–	–	–
PCA	✓	✓	✓	✓
t-SNE	✓	–	–	–
MDS	✓	–	–	–
UMAP	✓	–	–	–

MerQur'un belirgin avantajları: (i) DBSCAN, t-SNE, MDS ve UMAP'in tam destekli olması — bu yöntemler diğer açık erişim GUI'lerde yok, (ii) PCA biplot ve scree grafiklerinin otomatik üretilmesi, (iii) hiyerarşik kümelemede tüm linkage yöntemleri ve dendrogram kesim seçenekleri.

Sınırlılıklar: (i) Gaussian Mixture Models (GMM) panel olarak yer almaz — yumuşak kümeleme için K-Means alternatif, (ii) Spectral clustering ve Affinity Propagation panel olarak yoktur, (iii) Self-Organizing Maps (SOM) yoktur.

11. Sonuç

Bu çalışma, MerQur masaüstü yazılımının Kümeleme kategorisinde sunulan 7 analizi sistemli olarak tanıtmıştır. K-Means'in basit ve hızlı kümeleme yaklaşımından DBSCAN'in aykırı tespit yetisine, PCA'nın klasik doğrusal boyut indirgemesinden UMAP'ın modern manifold öğrenmesine uzanan kapsam, denetimsiz öğrenmenin geniş bir yelpazesini tek arayüzde sunar. Özellikle DBSCAN, t-SNE ve UMAP'ın doğrudan panel olarak yer alması MerQur'u diğer açık erişimli GUI istatistik yazılımlarından belirgin biçimde ayrıştırır. Sonraki davetli editöryal sunumlarda **İleri Düzey** kategorisi (VARCOMP, GAM, robust/quantile regresyon, Bayesian, mediation, path, diskriminant analizi vd.) ayrıntılı olarak incelenecektir.

Beyanlar

Etik Kurul Onayı: Bu çalışma insan ya da hayvan denek içermediğinden etik kurul onayı gerektirmemiştir.

Çıkar Çatışması: Yazar, MerQur yazılımının geliştiricisidir.

Finansman: Spesifik bir dış fon alınmamıştır.

Veri ve Kod Erişim Beyanı: Bu derleme orijinal araştırma verisi içermez. MerQur yazılımı https://merqur.sdu.edu.tr adresinden ücretsiz indirilebilir.

Yapay Zekâ Kullanımı: Bu makalenin yazımı sırasında üretken yapay zekâ (Claude, Anthropic) dil ve yapı düzeltmesi amacıyla destekleyici olarak kullanılmıştır.

Yazar Katkı Beyanı (CRediT): Ömer K. Örücü — Kavramsallaştırma, Yöntem, Yazılım, Doğrulama, Yazma (orijinal taslak), Yazma (gözden geçirme & düzenleme).

Kaynakça

Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In

Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining (KDD-96) (pp. 226–231). AAAI Press.

Halkidi, M., Batistakis, Y., & Vazirgiannis, M. (2001). On clustering validation techniques.

Journal of Intelligent Information Systems , 17, 107–145. https://doi.org/10.1023/A:1012801612483

Hastie, T., Tibshirani, R., & Friedman, J. (2009).

The elements of statistical learning (2nd ed.). Springer.

Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: A review and recent developments.

Philosophical Transactions of the Royal Society A , 374(2065), 20150202. https://doi.org/10.1098/rsta.2015.0202

Kaufman, L., & Rousseeuw, P. J. (1990).

Finding groups in data: An introduction to cluster analysis

. Wiley.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In

Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability

(Vol. 1, pp. 281–297). University of California Press.

McInnes, L., Healy, J., & Melville, J. (2018). UMAP: Uniform manifold approximation and projection for dimension reduction.

Journal of Open Source Software

, 3(29), 861. https://doi.org/10.21105/joss.00861

Pearson, K. (1901). On lines and planes of closest fit to systems of points in space.

Philosophical Magazine

, 2(11), 559–572. https://doi.org/10.1080/14786440109462720

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. https://doi.org/10.1016/0377-0427(87)90125-7

Schubert, E., Sander, J., Ester, M., Kriegel, H. P., & Xu, X. (2017). DBSCAN revisited, revisited: Why and how you should (still) use DBSCAN.

ACM Transactions on Database Systems

, 42(3), 1–21. https://doi.org/10.1145/3068335

Tibshirani, R., Walther, G., & Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic.

Journal of the Royal Statistical Society: Series B

, 63(2), 411–423. https://doi.org/10.1111/1467-9868.00293

Torgerson, W. S. (1952). Multidimensional scaling: I. Theory and method.

Psychometrika

, 17(4), 401–419. https://doi.org/10.1007/BF02288916

van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE.

Journal of Machine Learning Research

, 9, 2579–2605.

Ward, J. H. (1963). Hierarchical grouping to optimize an objective function.

Journal of the American Statistical Association

, 58(301), 236–244. https://doi.org/10.1080/01621459.1963.10500845