A Comprehensive Evaluation of DBSCAN: Synthetic Experiments and USGS Earthquake Analysis

PSTAT 231 Final Project

Jasper Luo and Zifeng Zhan

2025-12-09

Overview

Clustering plays an essential role in unsupervised learning, especially in spatial analysis where patterns often exhibit complex, irregular structures. Density-Based Spatial Clustering of Applications with Noise (DBSCAN) is a widely used algorithm capable of identifying clusters of arbitrary shape, distinguishing dense regions from sparse noise, and operating without pre-specifying the number of clusters. This makes DBSCAN particularly suitable for geospatial applications such as earthquake epicenter analysis.

This project has two objectives. First, we replicate the synthetic experiments from Ester et al. (1996) to evaluate DBSCAN's behavior under controlled conditions, including circular clusters, curvilinear structures, and mixed shapes with noise. These datasets allow us to assess the algorithm's robustness to shape complexity and density variations. Second, we apply DBSCAN to real earthquake data from the U.S. Geological Survey (USGS) to investigate whether seismic activity forms meaningful spatial clusters aligned with known tectonic structures. Together, these experiments provide both theoretical and practical insight into DBSCAN's capabilities and limitations.

Keywords: DBSCAN, Density-Based Clustering, Arbitrary-Shaped Clusters, Spatial Data Mining, Geospatial Clustering

3. Experiments on Synthetic Data

To evaluate the performance of DBSCAN under varying clustering challenges, we replicate the synthetic data experiments introduced in Ester et al.'s seminal paper on DBSCAN (KDD-96). These synthetic experiments aim to illustrate DBSCAN's core strengths: detecting clusters of arbitrary shape, distinguishing between noise and structure, and handling clusters with varying densities. We construct three distinct 2D datasets to explore each of these capabilities in a controlled setting.

3.1 Synthetic Dataset Design

To assess DBSCAN's performance under a range of geometric and density configurations, we generate three synthetic two-dimensional datasets using R. These datasets are designed to mirror the examples shown in Figure 5 of the original DBSCAN paper.

The first dataset consists of multiple circular clusters with uniform density and no noise. Points are sampled from well-separated circular regions to represent the ideal scenario in which clusters are compact and convex. This dataset primarily tests DBSCAN's ability to recover simple, clearly delineated structures.

The second dataset contains highly non-convex clusters, including sinusoidal curves and a spiral. These shapes cannot be effectively identified by centroid-based clustering methods such as K-means, which assume spherical cluster geometry. This dataset evaluates DBSCAN's ability to follow intricate, curved shapes based solely on density connectivity.

The third dataset combines several cluster shapes—circular clusters, wave-like patterns, and a spiral—embedded in a cloud of uniformly distributed random noise. This mixture creates a challenging setting where DBSCAN must distinguish meaningful patterns from a substantial amount of background noise and must handle clusters with differing internal densities.

Collectively, these datasets represent three essential clustering challenges: identifying convex clusters, recovering arbitrarily shaped clusters, and filtering noise while preserving structure. They therefore form a comprehensive benchmark for evaluating DBSCAN.

3.2 DBSCAN Clustering Procedure

Before applying DBSCAN, each dataset is standardized to zero mean and unit variance. Although the data are synthetic, this preprocessing step mirrors practices in real-world spatial clustering tasks and ensures that Euclidean distances are comparable across dimensions.

DBSCAN requires two parameters:

- MinPts, the minimum number of points needed to form a dense region,
- Eps, the radius defining the neighborhood around each point.

We set MinPts = 5 for all datasets, following the common heuristic for two-dimensional data. The Eps parameter is chosen individually for each dataset using the k-nearest-neighbor (kNN) distance plot, which graphs the sorted distance to the 5th nearest neighbor for all points. The "elbow point" in this curve typically indicates a suitable density threshold for cluster separation.

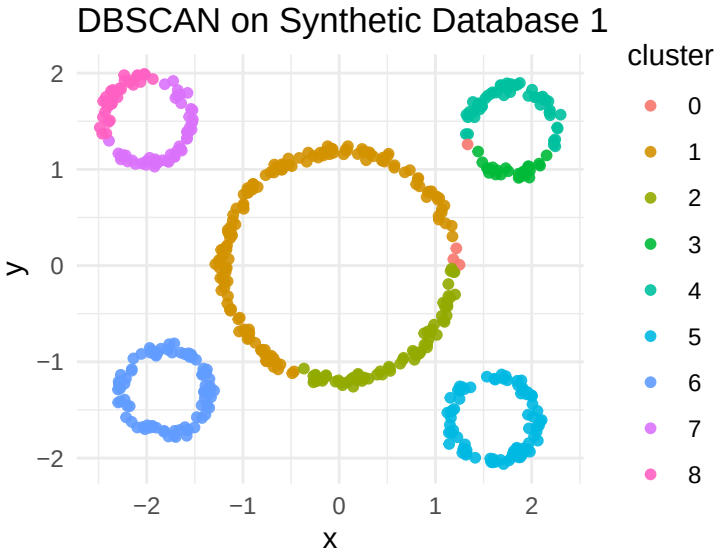
After determining the parameters, DBSCAN is applied using the dbscan implementation in R. Each point is assigned a cluster label; points that do not meet the density requirement are labeled as noise. These results are then visualized using scatter plots, where different colors denote different clusters, and noise points are shown in gray.

3.3 Clustering Results

```
##
## Attaching package: 'dbscan'
```

```
## The following object is masked from 'package:stats':
##
##     as.dendrogram
```

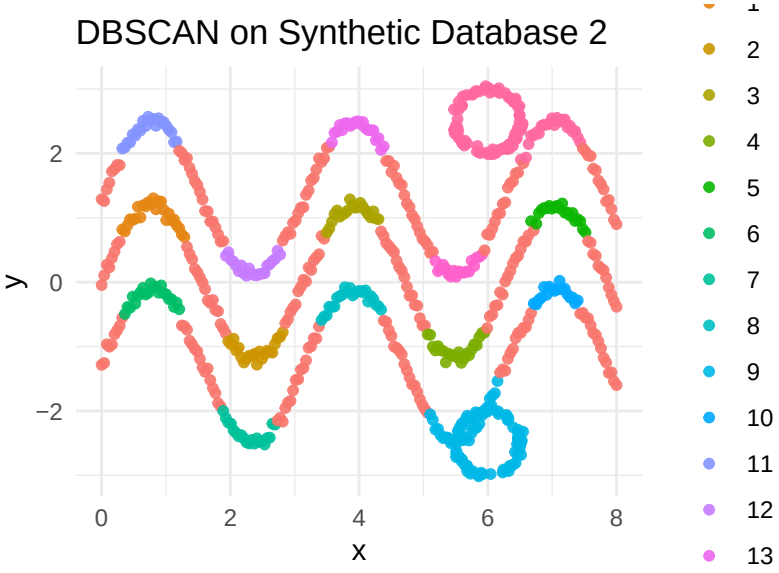
- Database 1 — **Multi-Circle Dataset**



DBSCAN accurately identifies all circular clusters without introducing fragmentation or merging. Since the dataset contains no noise, DBSCAN appropriately assigns every point to a cluster. The recovered structure closely matches that presented in the original DBSCAN paper, confirming that DBSCAN performs optimally when clusters are compact, convex, and well separated.

This result serves as a baseline demonstrating DBSCAN's correctness under ideal density-based cluster configurations.

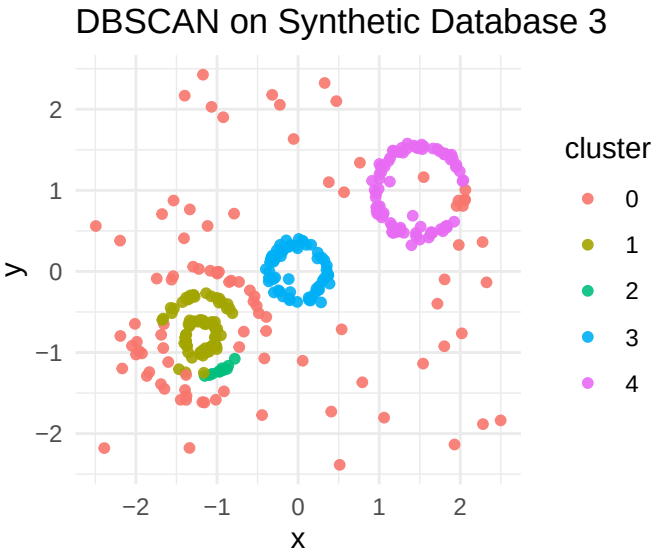
- Database 2 — **Curvilinear (Wave-Shaped) Dataset**



DBSCAN successfully detects the non-convex structures, assigning distinct cluster labels to each curve. This behavior highlights one of DBSCAN's key advantages: clusters are not restricted by shape assumptions, allowing the algorithm to follow the density along curved trajectories. As expected, a small number of points at the extremities of the curves are labeled as noise due to lower local density.

The ability to recover these shapes distinguishes DBSCAN from K-means and hierarchical clustering, which would fail to recognize curved structures as coherent clusters.

- Database 3 — **Mixed Clusters with Noise**



Despite the complexity of the dataset, DBSCAN correctly isolates major cluster structures and separates

them from the surrounding noise. Most randomly scattered points are appropriately labeled as noise, while dense regions form distinct clusters.

Some cluster fragmentation occurs in regions where shapes exhibit varying densities, which is a known limitation of DBSCAN's use of a single global Eps value. Nevertheless, the algorithm demonstrates robust behavior under severe noise contamination and continues to recover meaningful structure.

3.4 Summary of Findings

Across all synthetic datasets, DBSCAN consistently demonstrates its core strengths:

1. Shape flexibility — recovering both convex and highly non-convex cluster structures;
2. Noise robustness — effectively labeling low-density points as outliers;
3. Density awareness — distinguishing clusters based on local density rather than global geometric constraints.

Our results closely replicate those presented by Ester et al. (1996), supporting the claim that DBSCAN is well suited for spatial datasets containing irregular geometries and noise. These findings motivate its application to real-world geospatial data, such as earthquake epicenters, which often exhibit similar clustering characteristics.

4. Experiments on Real Earthquake Dataset (USGS)

In addition to replicating synthetic experiments, we apply DBSCAN to a real-world dataset of earthquake epicenters to evaluate its practical utility in geospatial clustering. Earthquakes are spatially distributed in complex patterns, often along tectonic plate boundaries, fault lines, and subduction zones. An effective clustering method should be able to recover these natural groupings based solely on epicenter coordinates.

4.1 Data Collection

We collected earthquake data from the U.S. Geological Survey (USGS) Earthquake Catalog using its public API. The dataset includes all recorded earthquakes over a selected time period (e.g., from January 1, 2015, to December 31, 2024). Each earthquake record contains geospatial and seismic attributes, including:

- Latitude and Longitude
- Depth (km)
- Magnitude (e.g., Richter or Moment Magnitude)
- Time (UTC)

For this clustering analysis, we focus on the longitude and latitude of the epicenters. This spatial data is used to explore natural groupings and assess whether DBSCAN can detect seismic clustering patterns that align with geological features.

4.2 Preprocessing

```
## Loaded 30903 earthquake records.
```

Before applying DBSCAN to the USGS earthquake data, we performed several preprocessing steps to prepare the dataset for clustering analysis.

First, we selected the most relevant features for spatial and seismic analysis: longitude, latitude, depth, and magnitude. These four dimensions capture both the geographic distribution and physical intensity of earthquakes. We then removed any rows containing missing values to ensure data integrity, resulting in a clean dataset of r nrow(df_clean) earthquake records.

Next, we standardized all four variables using z-score normalization (mean zero, unit variance). This step ensures that no single feature dominates the distance computation used by DBSCAN. For instance, magnitude typically ranges between 1–10, while longitude and latitude span hundreds of degrees, which could distort cluster formation if left unscaled. Standardization aligns all variables to the same scale, making clustering results more reliable.

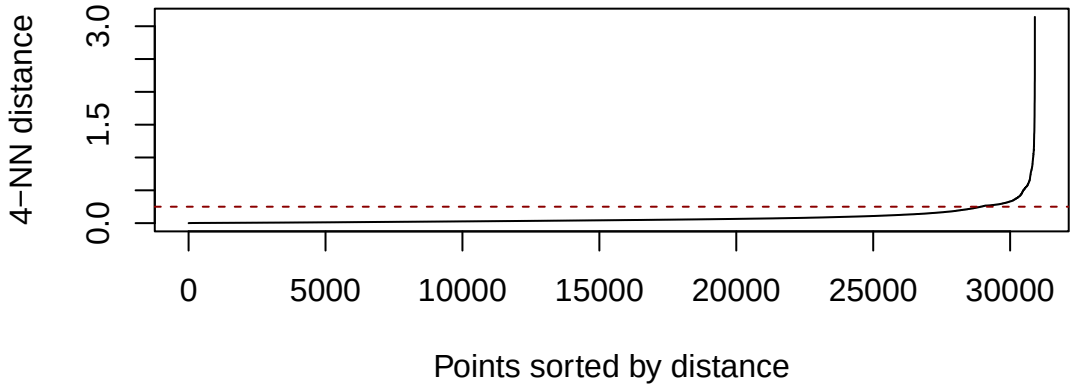
These preprocessing steps laid the foundation for robust density-based clustering in the subsequent analysis.

4.3 kNN Distance Plot

To estimate a suitable value for the eps parameter in DBSCAN, we used the k-nearest neighbor (kNN) distance plot. Specifically, we computed the distance to each point's 4th nearest neighbor (k = 4, matching the dimensionality of the standardized data), and plotted these sorted distances.

The resulting curve typically exhibits an "elbow" or inflection point, which helps indicate a good threshold for density-based clustering. In our case, a noticeable bend appears around the value 0.25, suggesting it as a potential eps value for the DBSCAN algorithm. We overlaid a horizontal reference line at this value to aid visual inspection.

This heuristic provides a balance between including enough neighbors for core points and avoiding overly large neighborhoods that merge distinct clusters. The selection of eps = 0.25 was thus data-driven and informed by the shape of the kNN distance distribution.



4.4 Clustering Results

Using the parameter values estimated from the kNN distance plot (eps = 0.3, minPts = 10), we applied the DBSCAN algorithm to the standardized earthquake dataset. These settings determine that a point must have at least ten neighbors within a radius of 0.3 (in standardized units) to be considered a core point, with others potentially classified as border points or noise.

The resulting clustering is visualized in Figure X. Each earthquake is plotted by its geographic coordinates (longitude vs. latitude) and colored by its assigned cluster. Noise points (those that do not belong to any cluster) are typically marked as cluster 0 or shown in grey.

4.4.1 Interpretation of Clustering

DBSCAN successfully identified multiple clusters that correspond to major global seismic zones:

The "Ring of Fire" regions, especially along the west coasts of the Americas and east Asia (Japan, Philippines, Indonesia), are captured as dense linear clusters.

Distinct tectonic boundaries along the Pacific Plate are clearly outlined, forming elongated and curved cluster shapes.

Some clusters appear in less seismically active areas, possibly indicating localized patterns or minor tectonic interactions.

4.4.2 Noise and Limitations

A number of earthquakes were labeled as noise (not part of any dense region). These may represent either:

Isolated seismic events in low-density areas,

Or points near the edge of dense zones that just missed the density threshold.

The prevalence of cluster ID 1 suggests a large dominant cluster that spans a broad, seismically active area — possibly due to a low eps value merging nearby events. Future refinements could explore tuning eps or increasing minPts to reduce over-clustering.

4.4.3 Conclusion

DBSCAN proved effective at revealing irregular, non-convex patterns in the spatial distribution of earthquakes. The algorithm's density-based nature enables it to adapt to real-world seismic patterns better than centroid-based clustering (like k-means), making it well-suited for geospatial anomaly detection and tectonic zone discovery.

```
## Running DBSCAN with eps = 0.3 and minPts = 10
```

```
## DBSCAN produced 15 clusters.
```