BeCode.org/BXL-Bouman-2.22

Density Based Spatial Clustering of Applications with Noise (*DBSCAN*)

Saba Yahyaa

Dec. 2020

Introduction:

. DBSCAN Algorithm.

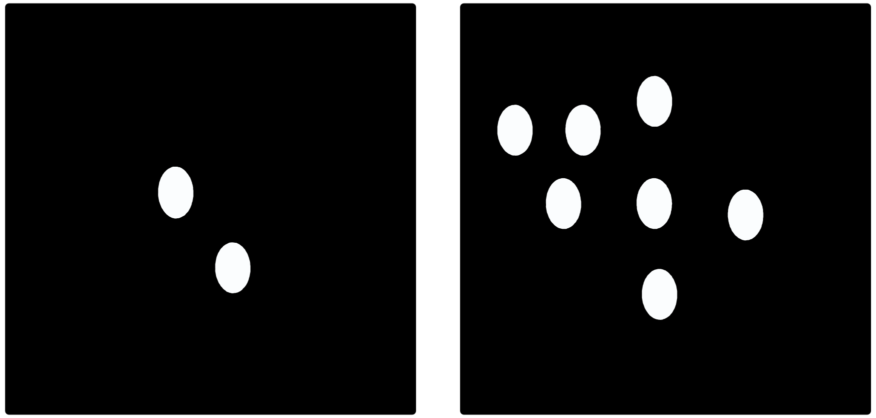
. Advantages and disadvantages.

. Hyper-parameters tuning.

. Comparison different clustering algorithms.

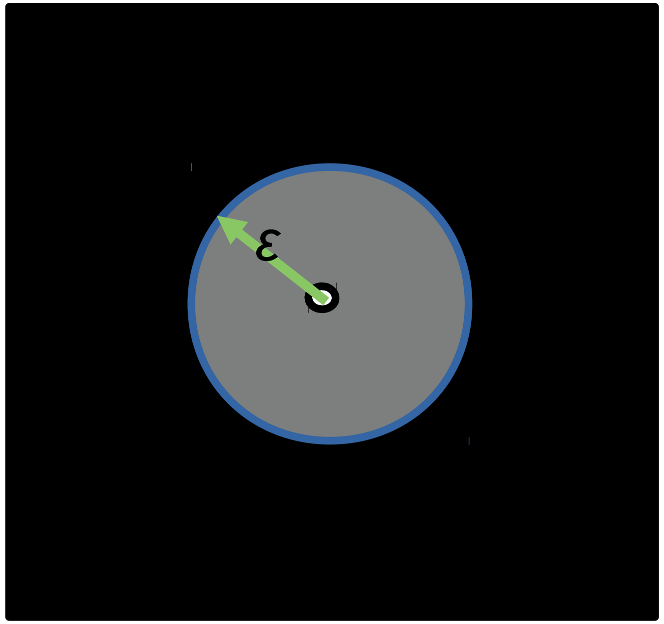
DBSCAN:

Density Based Spatial Clustering of Applications with Noise

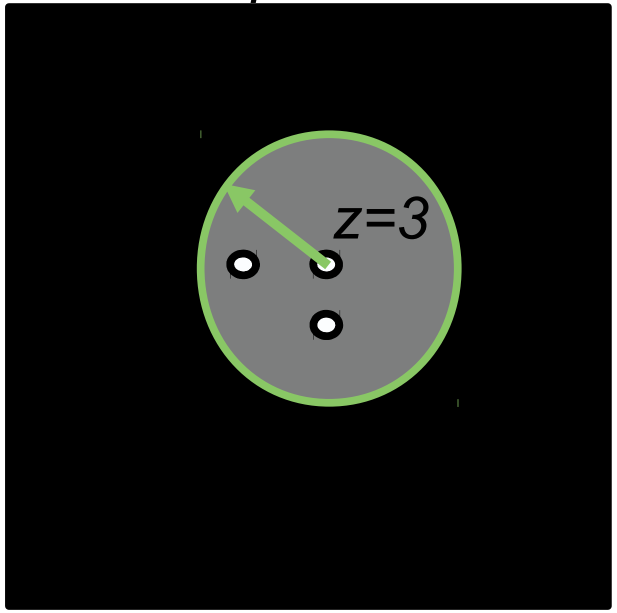


***DBSCAN* Parameters:**

Eps (ε): measure of neighborhood

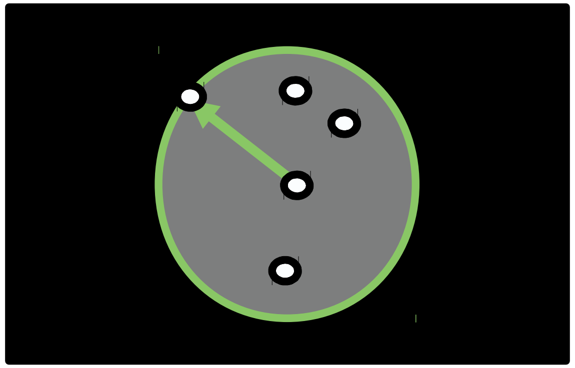


MinPts (z): min number of neighboring points inside the circle



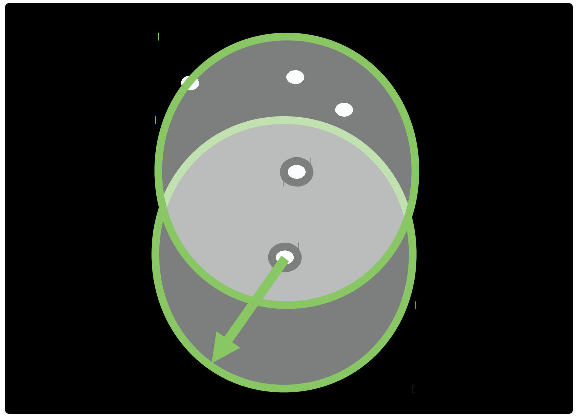
Classify each Point in the Data Set:

Core Point

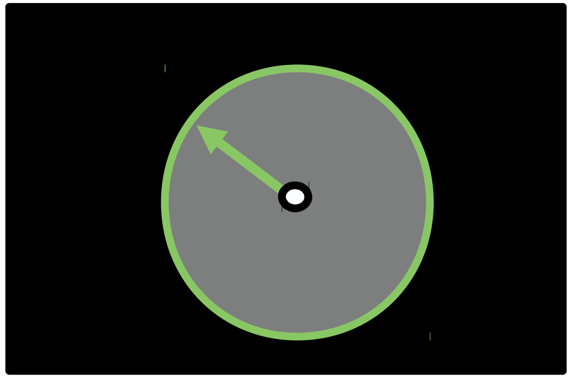


Z=5, ε=0.4

Border Point

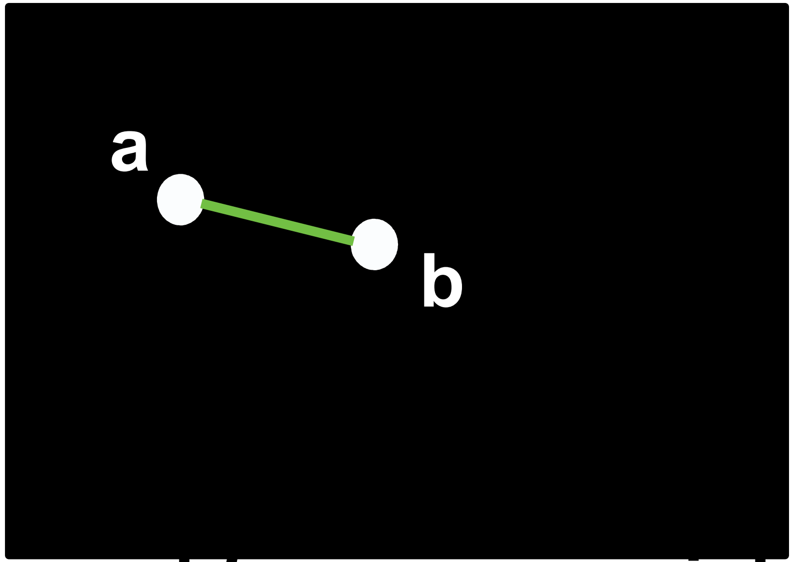


Noise Point



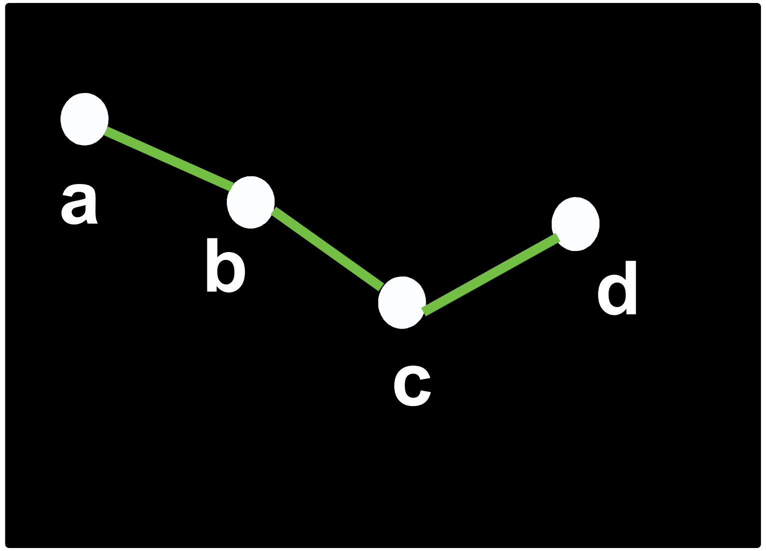
DBSCAN :

Density edge



***a* and *b* are core points**

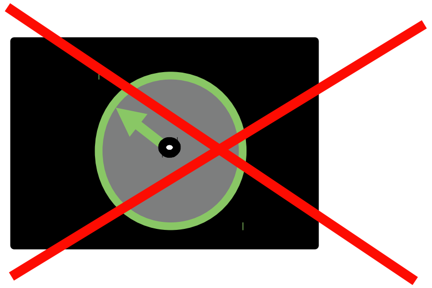
Density connected points



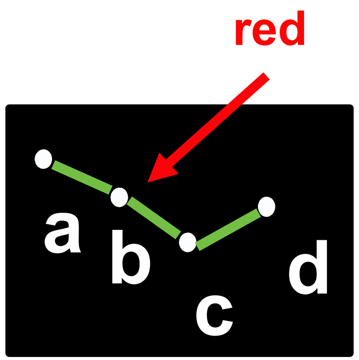
***a, b, c* and *d* are core points**

DBSCAN Algorithm:

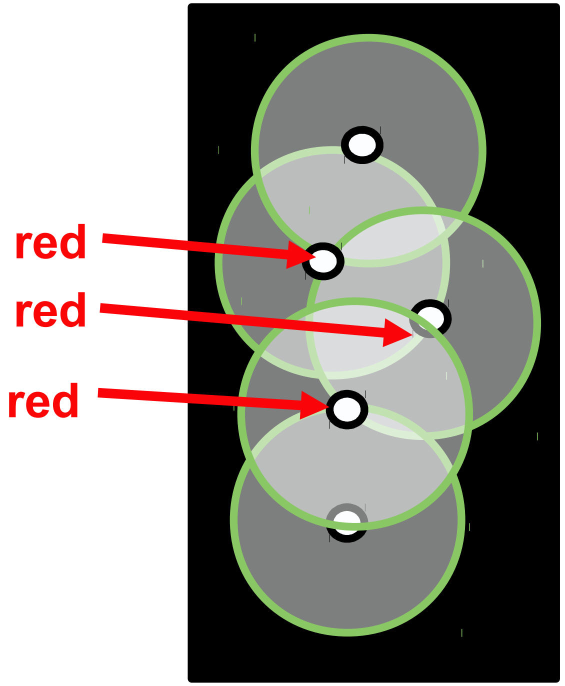
- **1. Classify** points
- 2. Discard **noise** points



- 3. Assign **cluster to a core point**



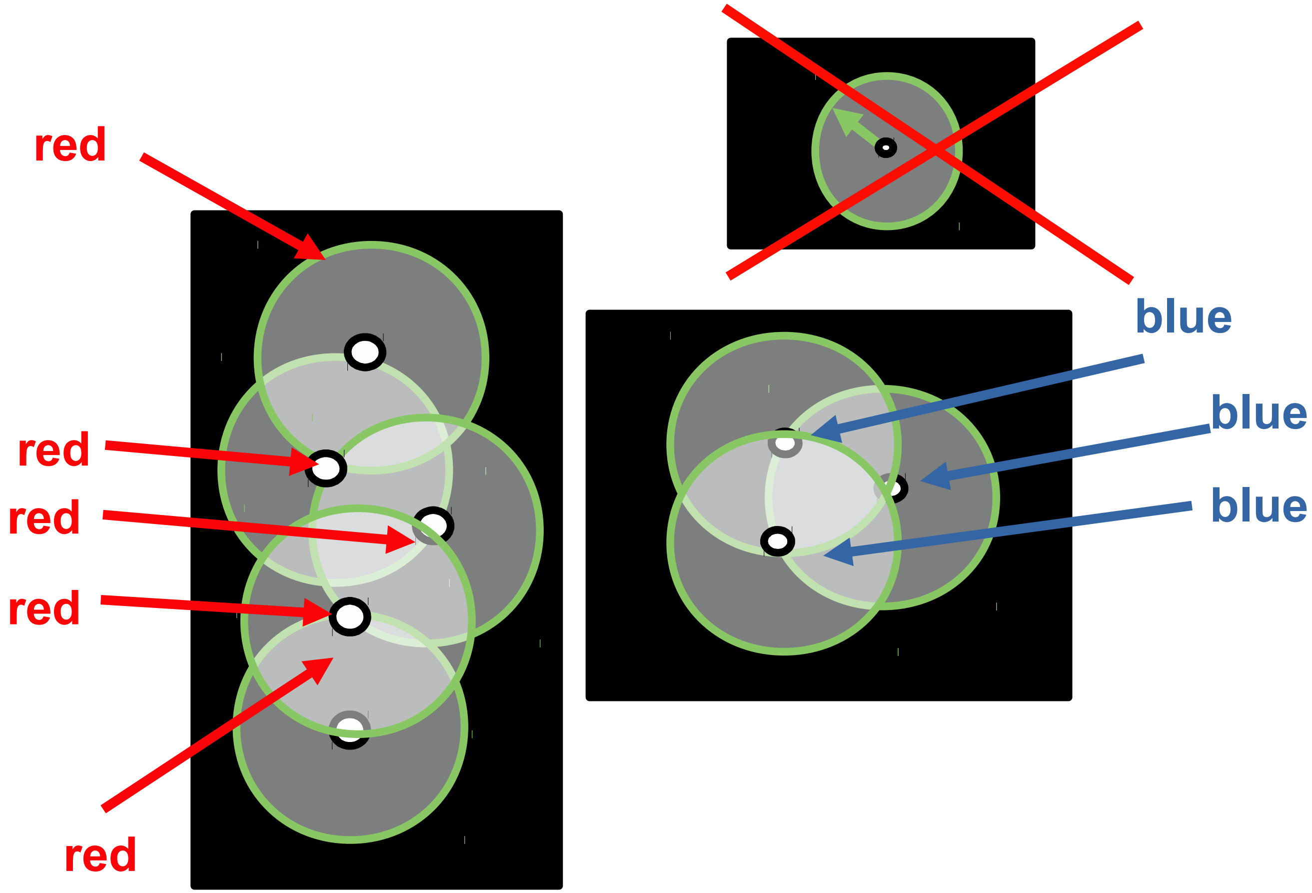
- 4. Color all the **density edge** connected point of a core point



DBSCAN Algorithm:

- **5. Repeat** steps *3*, *4* for uncolored core point

- **6.** Color **border** point according to nearest core point



DBSCAN Advantages and Disadvantages:

- **Advantage:**

  1. not sensitive to noise.

  2. can **find** the non-linearity separable cluster.

- **Disadvantage:**

  1. choosing ε and *z* is **difficult.**

  2. cannot cluster data sets well with **large differences** in densities.

  3. does not perform well with **large number** of features.

MinPts (z) and Eps (ε) Tuning on chipotle_stores data:

- Eps, ε using Euclidean distance:

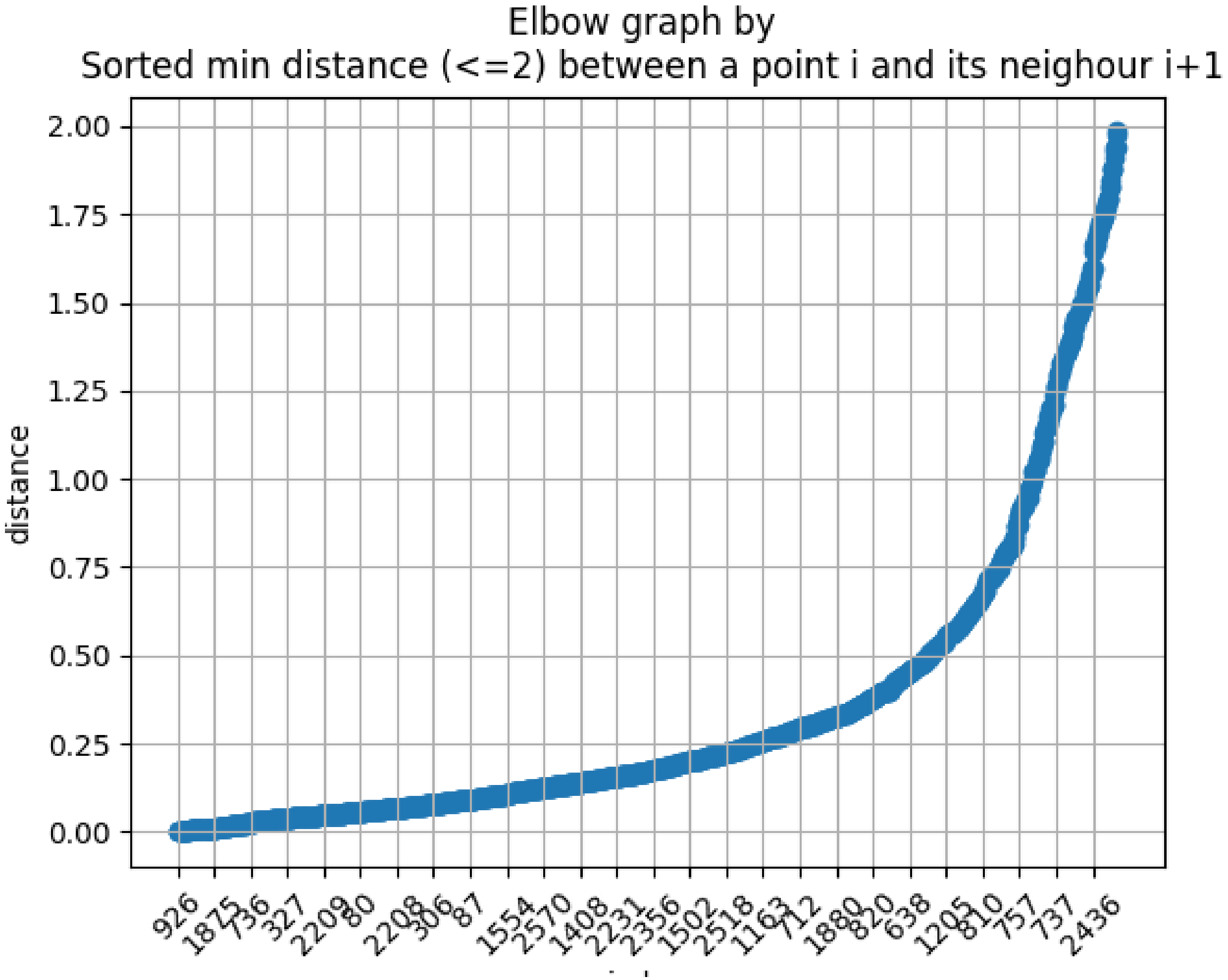
  1. sort the data,

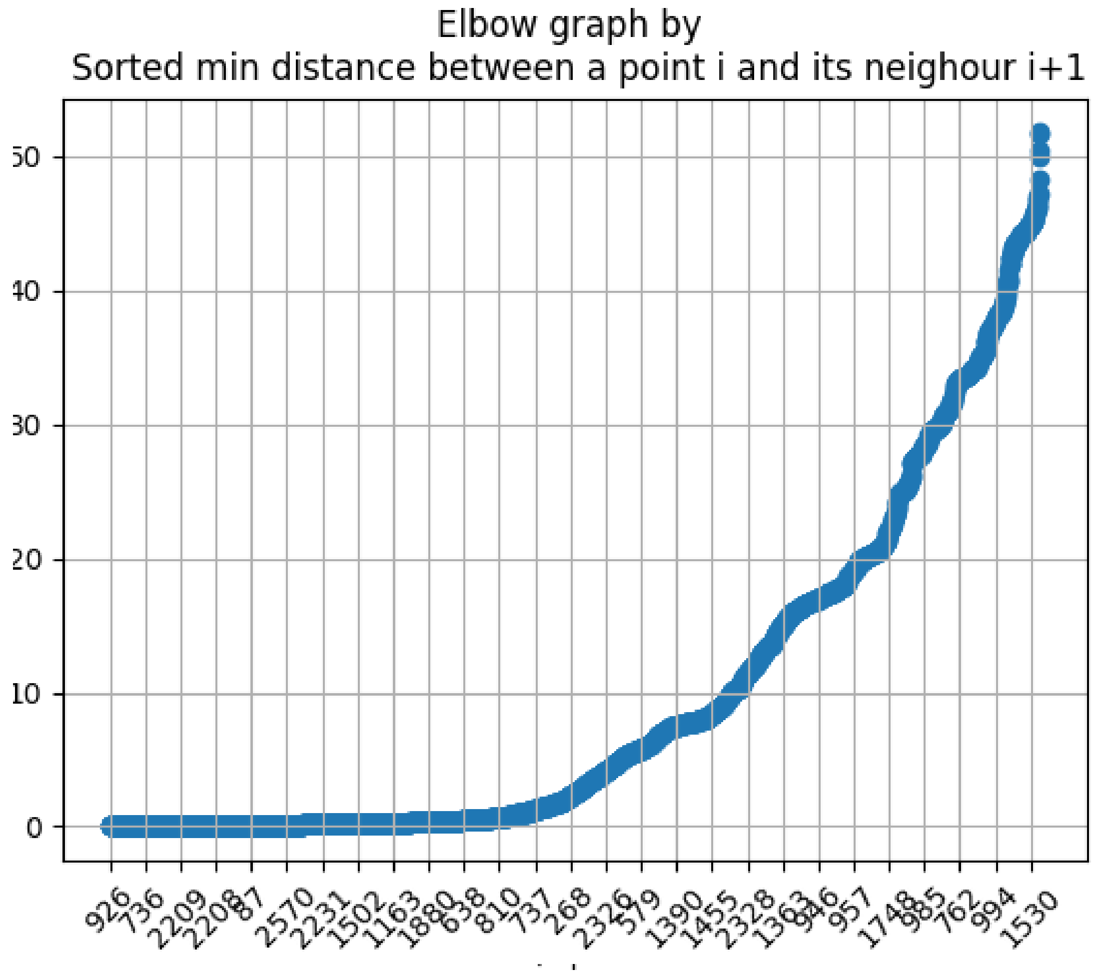
  2. find the **distance** among its neighbors,

  3. find the **minimum distance** between them,

  4. and **plot** the minimum distance.

MinPts (z) and Eps (ε) Tuning on chipotle_stores data:





MinPts (z) and Eps (ε) Tuning on chipotle_stores data:

- **Fix** ***z=12*** **for** ***ε=[0.1, 0.2,……, 2.4, 2.5]*****, find Silhouette score for each** ***ε*****:**

  for *ε=0.1*, silhouette=*-0.472*
  for ε=0.1, silhouette=*0.22*
  ….
  for *ε=0.4*, silhouette=*0.42*
  for *ε=0.5*, silhouette=*0.472*
  for *ε=0.6*, silhouette=*0.379*

MinPts (z) and Eps (ε) Tuning on chipotle_stores data:

- **Fix *ε=0.5* for *z=[5, 6,……., 49, 50]*, find Silhouette score for each *z*:**

for z=*7*, silhouette=*0.388*
for z=8, silhouette=*0.49*
for z=9, silhouette=0.484
….
for z=*49*, silhouette=*0.114*
for z=*50*, silhouette=*0.111*

Comparison k-mean, DBSCAN, OPTICS, and Hierarchical Agglomerrative:

---- Hierarchical (n_clusters = 5, affinity = **'euclidean'**, linkage =**'ward'**) ------
The average of dissimilarity for samples using Hierarchical is 92.498

------------------------------ Optics (min_samples=8) ---
The average of dissimilarity for samples using OPTICS is 1394.807

------------------------ DBSCAN (eps=0.5, min_samples=8) ------------------------------
The average of dissimilarity for samples using DBSCAN is 786.326

-------------------- k-mean (n_clusters=15, init=**'k-means++'**) ---------------------
The average of dissimilarity for samples using K-mean is 303.265

Comparison k-mean, DBSCAN, :

of clusters=15

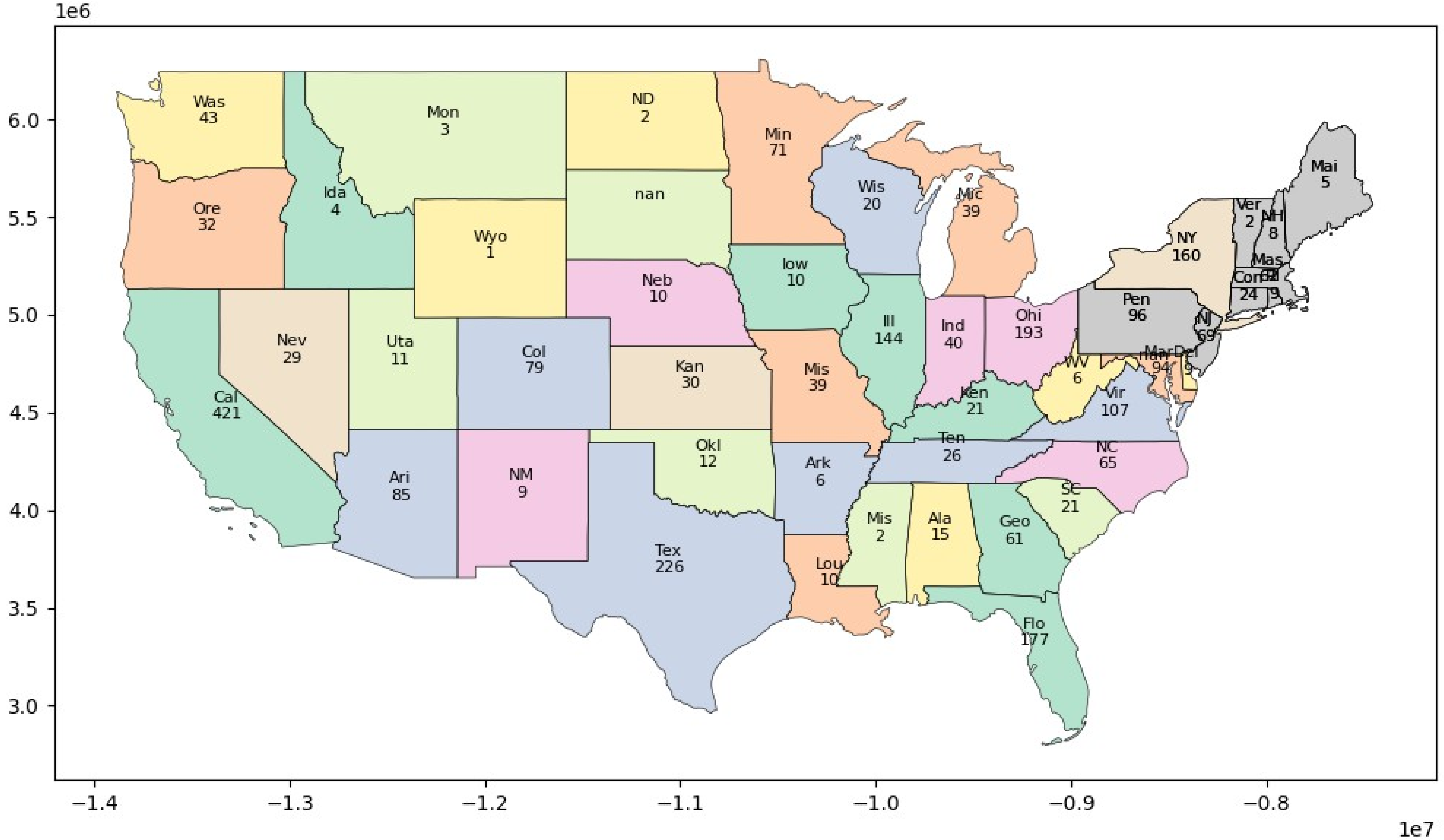
------------------------- DBSCAN (eps=0.5, min_samples=48) -------------------------------

The av of dissimilarity for samples using DBSCAN is 308.058

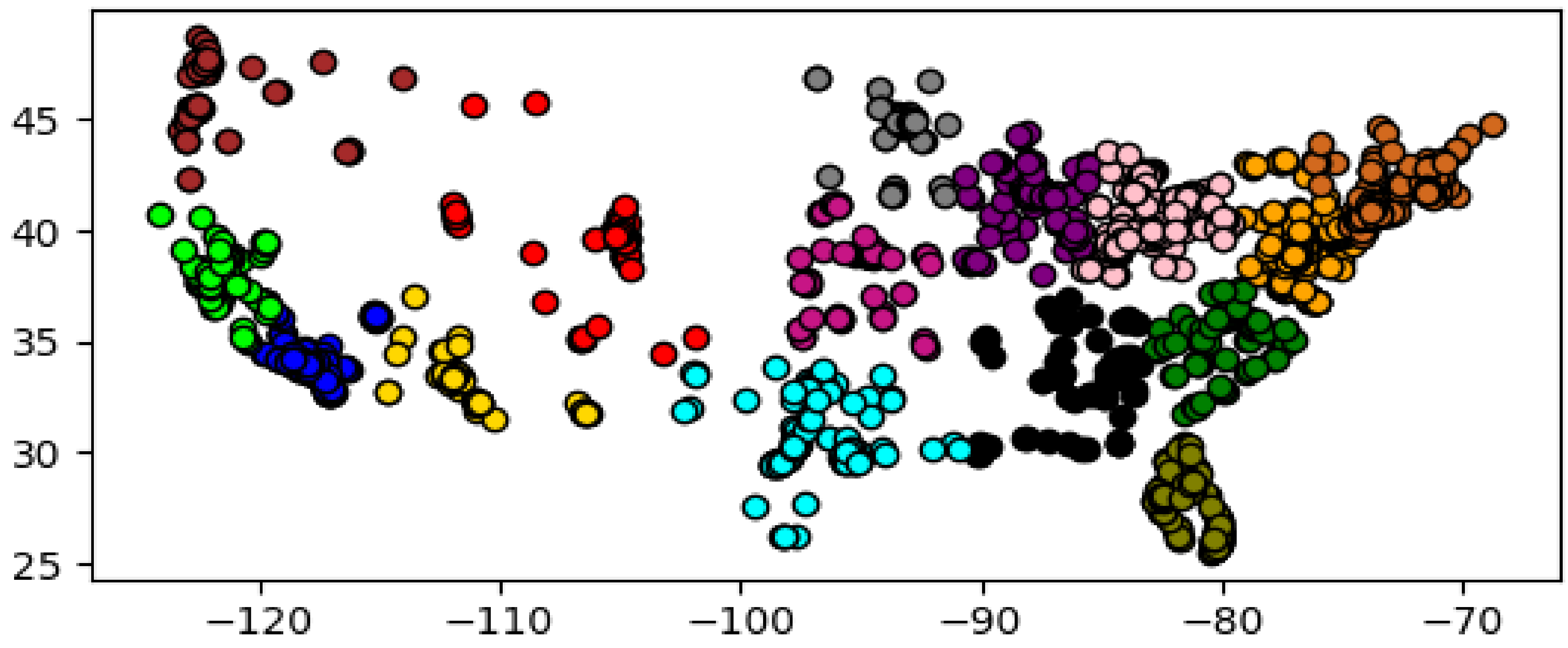
------------- k-mean (n_clusters=15, init=**'k-means++'**) ---------

The av of dissimilarity for samples using K-mean is 303.265

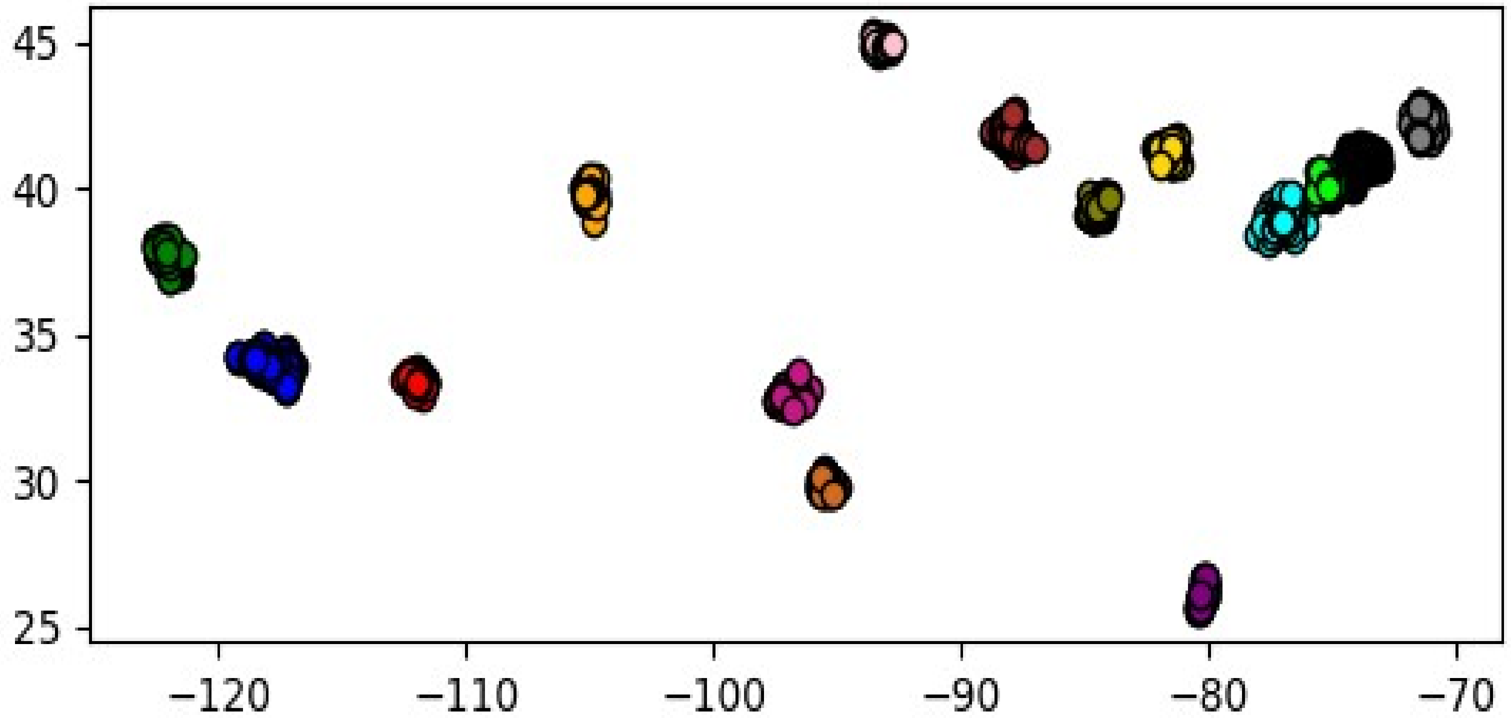
USA map with Chipotle Data:



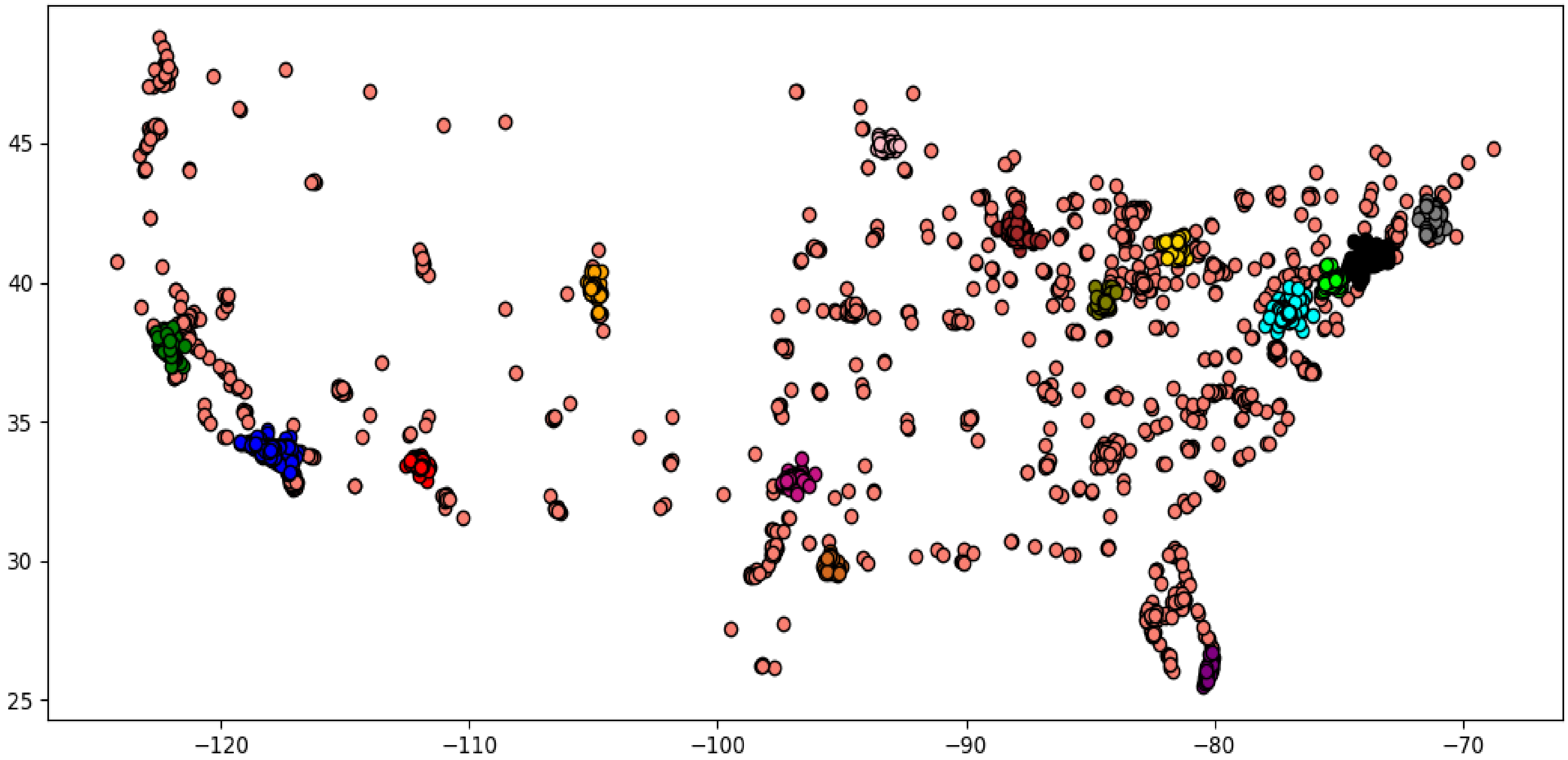
USA, k-means with n_clusters=15:



USA, DBSCAN with n_clusters=15



USA, DBSCAN with n_clusters=15 and Noise:



Thanks