An Exploration of the Distribution of Jiuling

VE414 Bayesian Analysis

Jingnan Gao Sizhe Zhou Zhujiang Gu

UM-SJTU Joint Institute

Contact Information:
Email: gjn0310@sjtu.edu.cn
sizhezhou@sjtu.edu.cn
guzhujiang@sjtu.edu.cn

1 Data Visualization

Data visualization is first performed to demonstrates the observation of Tayes since Bayesian frame is mainly based on data.Both route and the observed number are presented. Note that the weighted number here equals Close Tayes + 0.2Far Tayes.

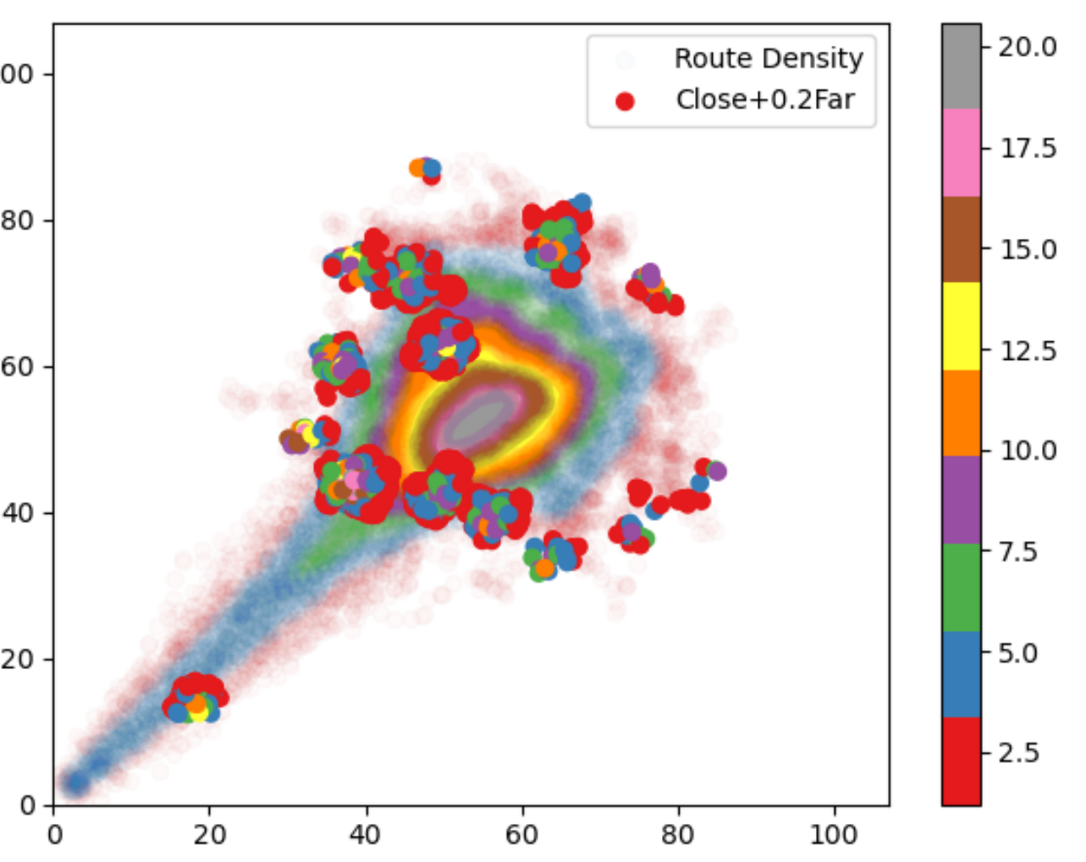


Figure 1: Route and Number of Jiulings (Weighted)

2 Estimation on Number of Jiuling

Prior: The distribution of Tayes from one Jiuling follows an uniform distribution $U(-3, 3)$, and more tayes means closer to Jiuling.
Transformation:By adding more points in a circle of radius 3 after setting a data point as the center according to the number of Jiuling that are classified as "Close" and 3 mile for "Far" Jiuling. This method keeps the generality since we could observe the same data within the range and density could be represented by numbers of scatters, then we utilized 3 methods to find the number of Jiuling by clustering the fruits observed and treat each cluster as one Jiuling. The following is the method we applied and the results.

2.1 EM (Expectation-Maximum) Method

EM algorithm is utilized in this method by initializing several source points estimated after reviewing the visualization of the observations. Two parallel kind of estimation are considered in this method, which are shown as follows. Note that intersection here means if two Jiuling are close to each other, the region between them would also have a high probability with numerous fruits, but we don't consider that there is a Jiuling over there.

Prior: The distribution of Jiuling follows a Gaussian-Mixture Model.

1. No intersection. The validation by randomly splitting the data into 10 parts and recursively pick 9 of them, the results are shown below and the number of Jiuling by this method is 13.

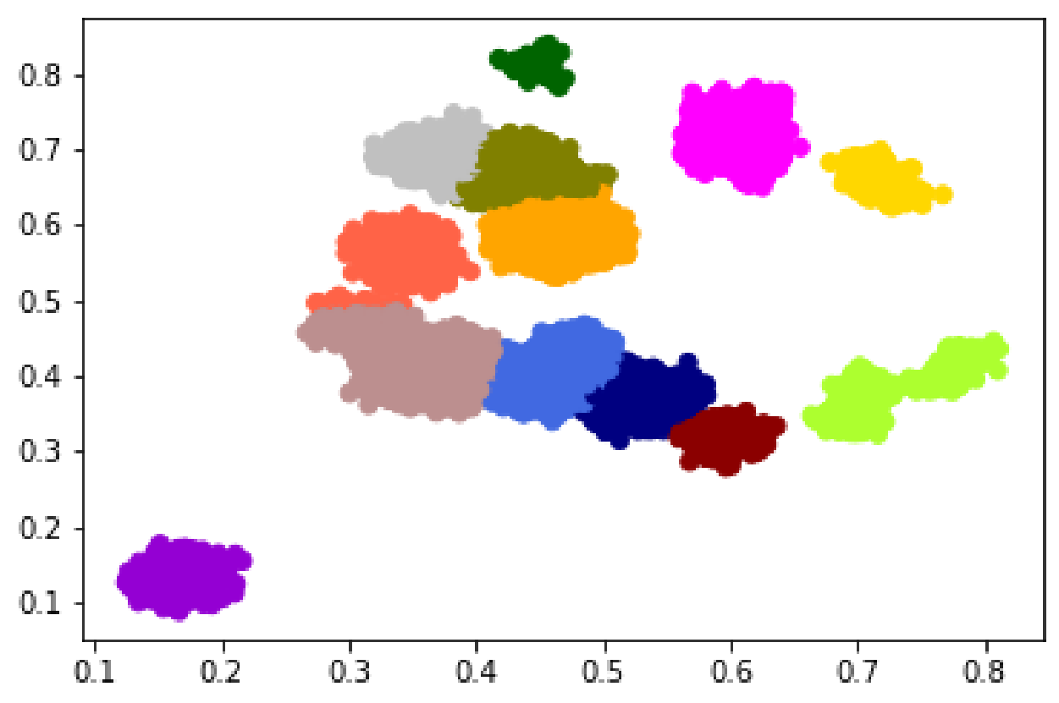


Figure 2: EM - No Intersection

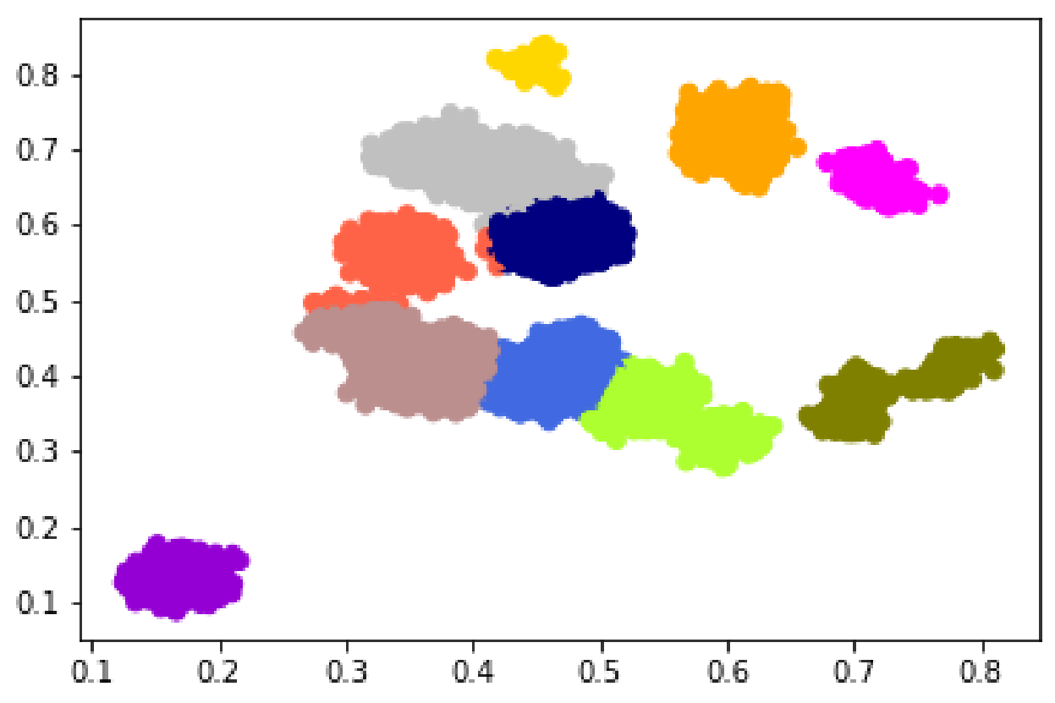
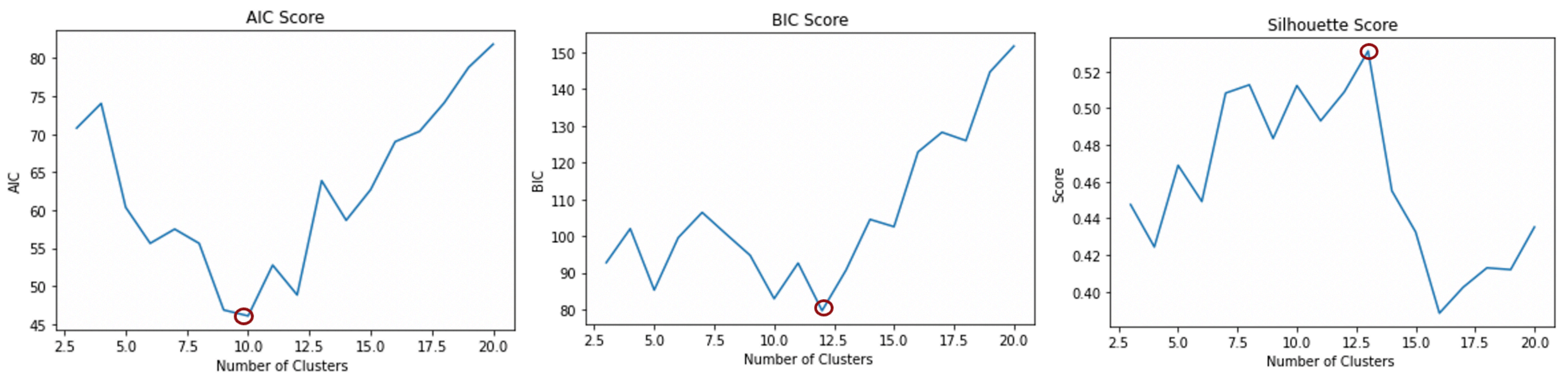


Figure 3: EM - Intersection

2. Intersection. For this trial, we keep updating the σ matrix if the "influence" of that initial points are larger than a threshold. Note that the "influence" of an initial points is determined as the mean value of "Weighted number" within 3 mile from that point and the threshold is determined as the median value of all the "influence". The validation by randomly splitting the data into 10 parts and recursively pick 9 of them, the results are shown above and the number of Jiuling by this method is 11.



AIC vs. Cluster Number BIC vs. Cluster Number Silhouette Score vs. Cluster Number

Figure 4: Three metrics for EM under different cluster numbers

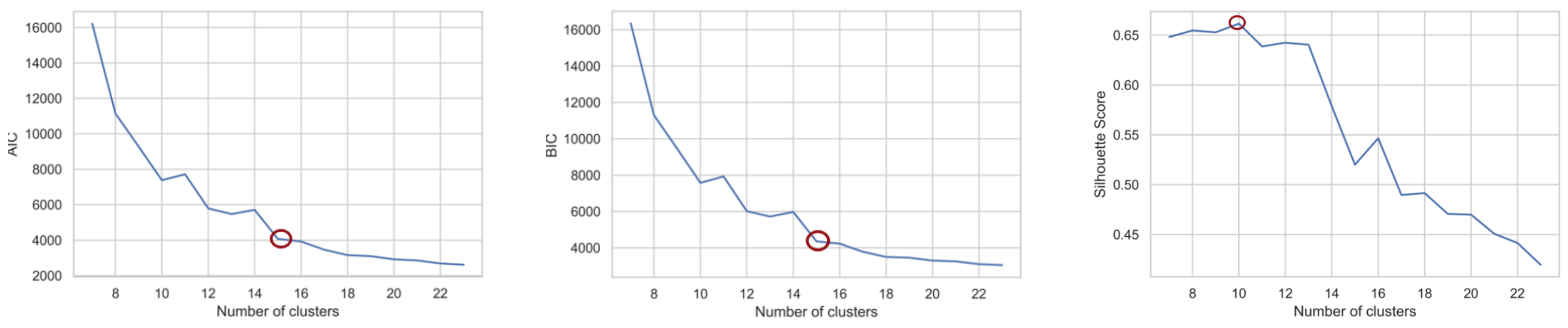
2.2 Bisecting K-means

As a comparison with EM to cluster the Tayes, we introduce bisecting K-means, which is improved from traditional K-means algorithm to alleviate converging to local extremum. The traditional K-means can be derived from EM (GMM) algorithm as follows. Hard cluster is applied in the E-step. Namely,

$$
\begin{aligned}
P(z=c|x_i,\theta_{old}) &= \lim_{\sigma\to 0}\frac{\pi_c\cdot exp(-\frac{1}{2\sigma}||x_i-\mu_c||_2^2)}{\sum_{j=1}^{c}\pi_j\cdot exp(-\frac{1}{2\sigma}||x_i-\mu_j||_2^2)}\\
&=\begin{cases}1 & \text{for smallest } ||x_i-\mu_c||_2^2\\ 0 & \text{otherwise}\end{cases}
\end{aligned}
$$

In the M-step, the new centroids are recomputed. The bisecting K-means is a hierarchical variant of K-means where the starting point is one whole cluster and we keep splitting the one cluster such that the SSE (sum of squared error) reduction is maximized until the predefined number of clusters is reached.

To determine the number of clusters (Jiuling), three metrics AIC, BIC, and Silhouette coefficient are applied on data which yields the result in Fig. 5. By using Elbow method, the number of Jiuling is estimated to be 15 by AIC and BIC. While by maximum of Silhouette score, estimated number is 10. The sample cluster result of 10 and 15 predefined clusters are shown by Fig. 8.



AIC vs. Cluster Number BIC vs. Cluster Number Silhouette Score vs. Cluster Number

Figure 5: Three metrics for bisecting K-means under different cluster numbers

Note that bisecting K-means is based on spherical and similar clusters to ensure correct approximation to the optimum. This means it may not perform well on correlated clusters and clusters of various sizes and shapes, unlike GMM which introduces variances and covariances to accomodate this.

2.3 DBSCAN (Density-Based Spatial Clustering of Applications with Noise)

As another comparison to EM clustering model, DBSCAN is based on density which means no specific number of clusters is required and it's not prone to noise. The main algorithm divides data into 3 kinds: core points (point with at least minPts points are within distance of ϵ), reachable points (non-core points that fall into some core point's ϵ neighborhood), and noise points (non-core points and non-reachable points). The implementation is basically clustering reachable points and removing noise points.

To find the optimal parameters (minPts and ϵ), Silhouette score is introduced for grid search. Due to the limit of computer memory and speed, the range of grid search is hence limited ($\epsilon \geq 4$) is undoable for 12G RAM. Basically, a coarse grid search, with minPts ranging from 3 to 9 with step as 1 and ϵ ranging from 0 to 4 with step as 1, is applied first. And then a finer grid search, with minPts ranging from 3 to 5 with step as 1 and ϵ ranging from 1 to 3 with step as 0.25, is applied.

Part of the grid searching result is shown by Fig 6 and Fig. 7. The optimum Silhouette score suggest that 7 clusters is the best while 8 ∼14 except for 11 clusters have similar performance. The sample cluster result is shown by Fig. 9.

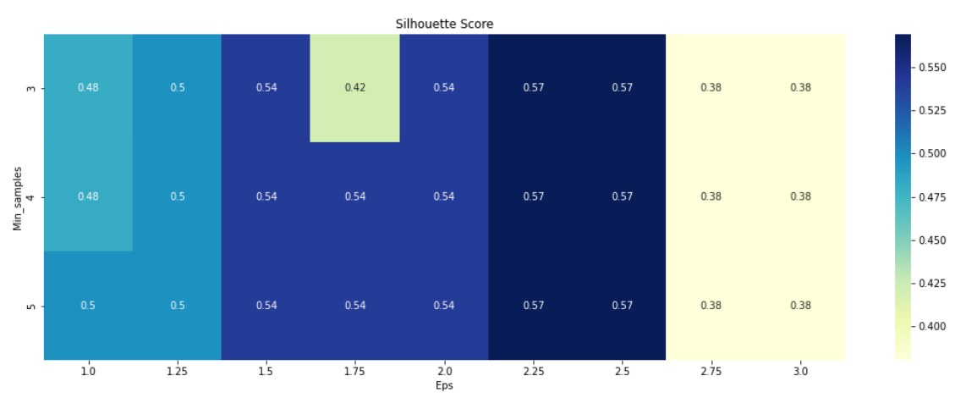


Figure 6: Silhouette Score on DBSCAN's Parameters

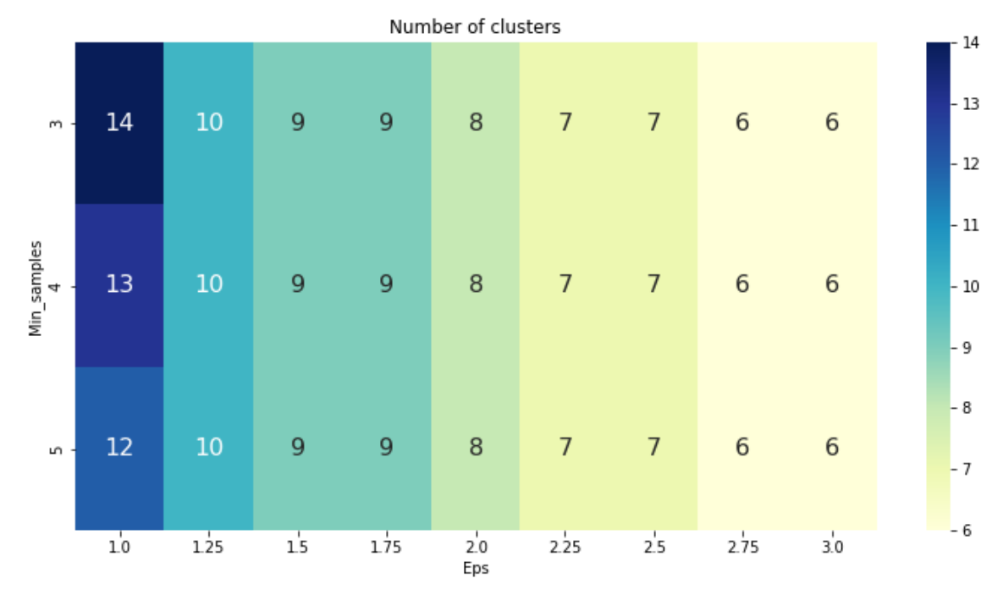


Figure 7: Cluster Number on DBSCAN's Parameters

Also note that DBSCAN tends to merge overlapping clusters (see the dense Tayes cluster together in the left bottom of Forbidden Forest from Fig. 9) and may yields unreasonable results under this context (see the purple data points spread over forest from Fig. 9). It also has inability when dealing with clusters of large density differences due to its inflexible setting of parameters.

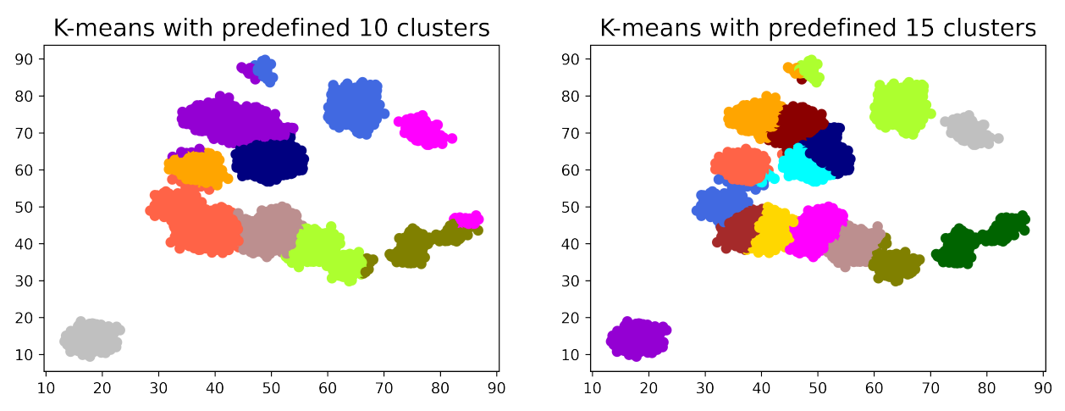


Figure 8: Cluster Result of Bisecting K-means

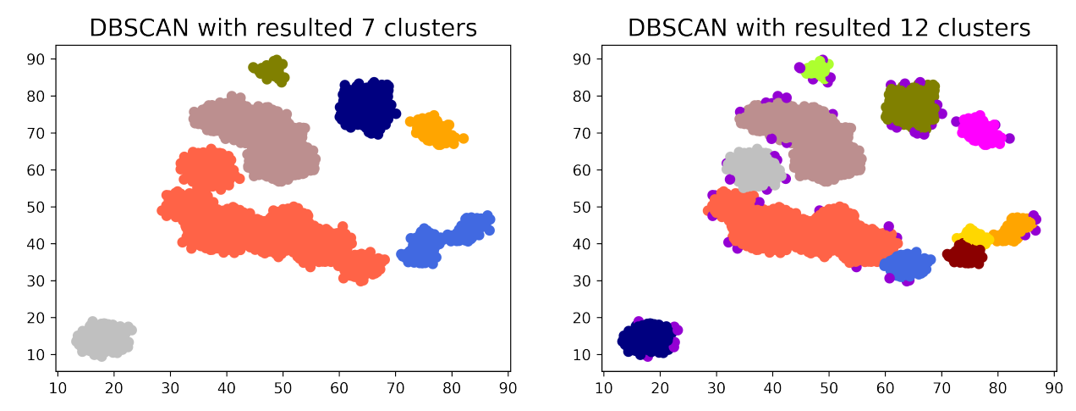


Figure 9: Cluster Result of DBSCAN

2.4 Validation on Number of Jiuling

Note that both the EM and the bisecting K-means algorithms contain the same parameter which is the predefined number of clusters. So it's necessary to conduct a sensitivity validation on number of clusters to see which predefined number of clusters are the most suitable.

3 Thoughts on Moving Jiuling

When the Jiulings are mobilizable, the primitive idea to estimate the number of Jiuling is still by clustering. But since Jiulings can move, the newly fallen Tayes rather than accumulated Tayes are the actual target to be clustered. Based on this, the first step is to divide the whole Forbidden Forrest into certain regions to tolerate the error of spotting the Tayes. And then obtain the distribution of newly found Tayes in each region by difference between data in consecutive timestamps. The ideal case is that the newly fallen Tayes entirely come from the temporarily static (or moves by little compared to the size of each region) Jiulings during those two timestamps. Then through clustering, the number of Jiulings can be estimated. So the timestamps between which Jiulings are static or move just by little compared to reasonable region sizes that can be defined are needed to conquer the main task.

4 Conclusion

This project is to estimate the number and the location of Jiuling, the mysterious and invisible plant in the Forbidden Forest. Data visualization is first performed to the positions of Tayes, and then three different methods, including Expectation-Maximum, Bisecting K-means and DBSCAN, are utilized to figure out the estimated number and locations assumed that Jiuling is static. Taking the results of multiply models into account, there are estimated 11 Jiuling trees in the forest. The overall number of trees in the entire forest is estimated to be 55 under the assumption of uniform distribution. Further steps and processing for moving Jiuling as future work are proposed as well.

References

[1] Timothy L Bailey, Charles Elkan, et al. Fitting a mixture model by expectation maximization to discover motifs in bipolymers. 1994.

[2] T.K. Moon. The expectation-maximization algorithm. IEEE Signal Processing Magazine, 13(6):47–60, 1996.