Hyperparameter Selection And Cluster Derivation In OPTICS

In our study, the selection of OPTICS clustering hyperparameters and the determination of the final clusters follow the steps below.

Hyperparameter Selection

We estimate *min samples* from the knee point of the sorted k-NN distance curve (k-distance plot) to obtain density-based candidates (Yin et al. 2023). Let $p_q(k)$ denotes the qth percentile of the k-NN distances over all samples. Using the maximum chord-distance criterion, knee points for $q \in \{90, 95, 97.5\}$ occur at $k = 30, 43$, and 33, respectively (see Figure 1). We set *max eps* sufficiently large to capture all relevant density structure.

To obtain a robust *min samples*, we evaluate the three knee point derived candidates $k \in \{30, 33, 43\}$ by conducting a local sensitivity analysis over the widely used range $\xi \in [0.05, 0.10]$ (Ankerst et al. 1999). We further assess stability by perturbing *min samples* with multipliers {0.8, 1, 1.3, 1.5} and measuring the consistency of cluster assignments between the resulting clusterings (von Luxburg 2010).

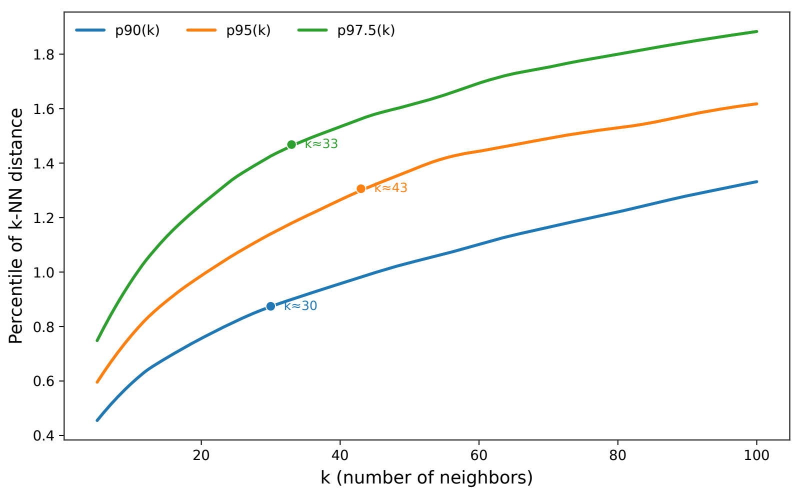
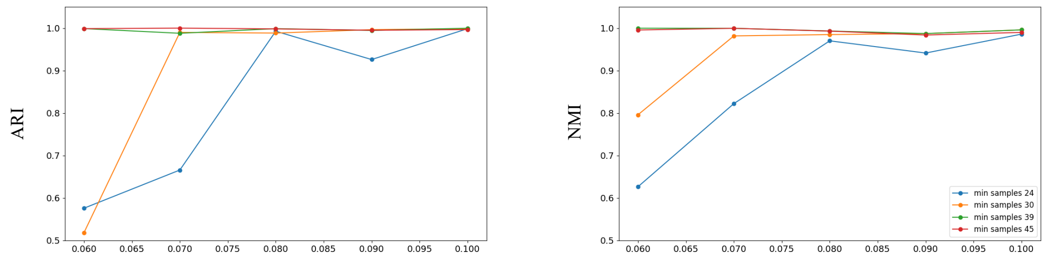


Figure 1: Knee-point detection on the k-th nearest neighbor distance percentiles.

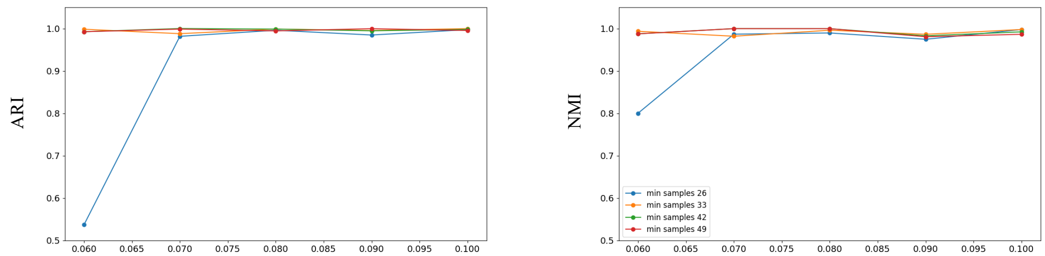
Analysis Of The ξ Parameter

Consistent with the principle that clustering should be robust to minor parameter variations (Lange et al. 2004), agreement between clusterings from adjacent parameter settings is quantified by the Adjusted Rand Index (ARI) (Vinh, Epps, and Bailey 2009) and Normalized Mutual Information (NMI) (Strehl and Ghosh 2002). High ARI and NMI values indicate strong consistency.

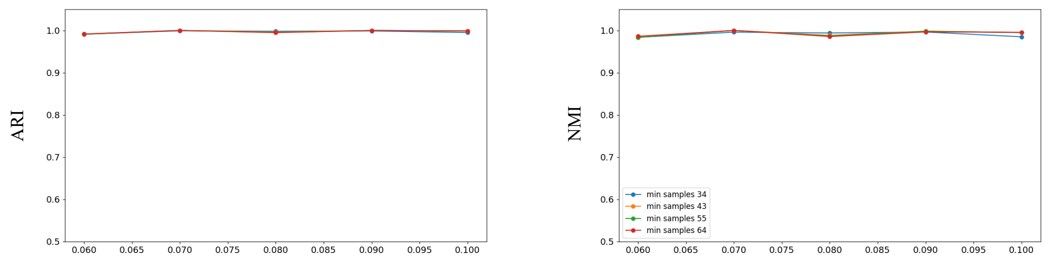
As shown in Figure 2, compared to $k = 30$ and $k = 33$, setting *min samples* $= 43$ yields consistently high ARI and NMI values, all above 0.98, at and around this setting, indicating highly stable performance. Moreover, the metric curves for $k = 30$ and $k = 33$ reveal that *min samples* within $[40, 49]$ also forms a local stability plateau with similarly high ARI and NMI values. Together, these results support the selection of *min samples* $= 43$ as a robust choice.



(a) Metric curves vs. ξ (k=30)



(b) Metric curves vs. ξ (k=33)



(c) Metric curves vs. ξ (k=43)

Figure 2: Comprehensive analysis of clustering performance across different parameters.

Final Clustering Result

With $min\ samples = 43$ selected, we analyze the effect of ξ on cluster count and clustering quality using metrics including noise rate, DBI, and Silhouette score (Figure 3). Over the interval $\xi \in [0.05, 0.09]$, the solution consistently produces 14 clusters, forming a stability plateau that reflects a stable and reliable clustering of the data. Throughout this range, the Silhouette score remains high, while both the noise rate and DBI remain consistently low, further supporting the robustness of this parameter region. Based on these metrics, we therefore select $\xi = 0.07$ as the operating point within this stability plateau.

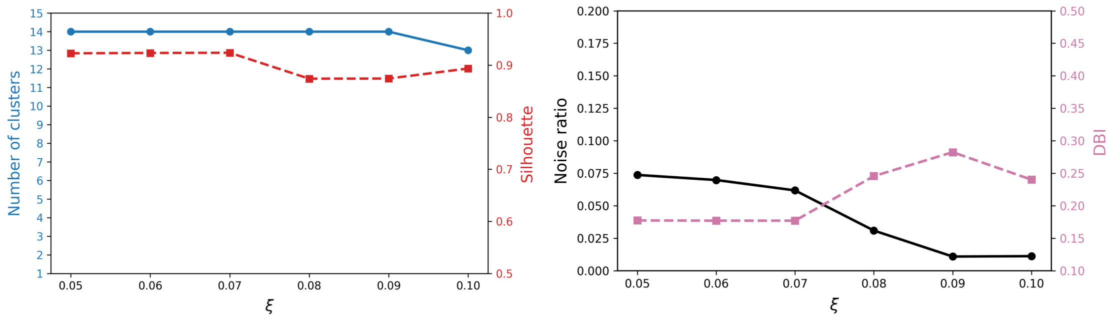


Figure 3: Cluster counts and quality metrics as functions of ξ with $min\ simples = 43$.

The parameter set ($min\ samples = 43$, $\xi = 0.07$) produces a reachability plot with well-separated valleys that correspond to 14 distinct clusters (Figure 4). This clustering is adopted as the final partition.

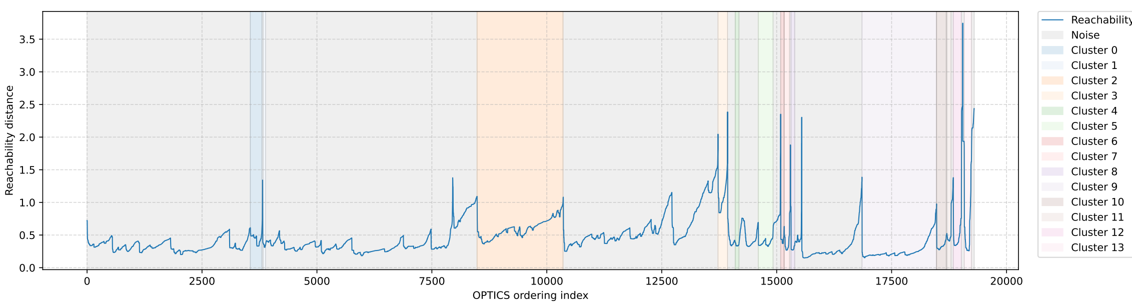


Figure 4: Reachability plot derived from OPTICS under $\xi = 0.07$.

References

Ankerst M, Breunig MM, Kriegel HP, Sander J, 1999 *Optics: Ordering points to identify the clustering structure. Proceedings of the 1999 ACM SIGMOD international conference on Management of data*, 49–60.

Lange T, Roth V, Braun ML, Buhmann JM, 2004 *Stability-based validation of clustering solutions. Neural computation* 16(6):1299–1323.

Strehl A, Ghosh J, 2002 *Cluster ensembles—a knowledge reuse framework for combining multiple partitions. Journal of machine learning research* 3(Dec):583–617.

Vinh N, Epps J, Bailey J, 2009 *Information theoretic measures for clusterings comparison: Variants. Properties, normalization and correction for chance* 18.

von Luxburg U, 2010 *Clustering stability: an overview. Foundations and Trends in Machine Learning* 2(3):235–274.

Yin L, Hu H, Li K, Zheng G, Qu Y, Chen H, 2023 *Improvement of dbscan algorithm based on k-dist graph for adaptive determining parameters. Electronics* 12(15):3213.