

OCA v1.75

Methods, Region Formation and Limitations

Optimization Cluster Analyzer

Reference based on implementation v1.75

1. Parameter-space representation

Stage	Implementation	Effect
Effective dimensions	Removes constant parameters, exact duplicate columns, numeric relationships with Pearson >= 0.999999, and excessively high-cardinality text dimensions.	Prevents uninformative or redundant dimensions from distorting distance calculations.
Numeric parameters	Computes Q05 and Q95. That robust range is used only when tails materially stretch the scale; otherwise minimum and maximum are used.	Reduces excessive tail influence without forcing robust clipping on every variable.
Categorical parameters	Bool, enum and string remain categorical. Their internal numeric encoding is not interpreted as ordinal distance.	Equal categories contribute 0 to Gower distance; different categories contribute 1.
Observed step	Median positive difference between adjacent unique observed values.	Supports neighborhood reconstruction and refine.set grid construction.

Numeric normalization	Observed-step inference														
$x^* = \text{clamp}(\frac{x-l}{H-l}, 0, 1)$	$\text{step} = \text{median}\{\Delta v_i : \Delta v_i > 0\}$														
<table><tr><th>Symbol</th><th>Meaning</th></tr><tr><td>x</td><td>Raw value.</td></tr><tr><td>L, H</td><td>Q05/Q95 when the robust range is materially narrower; otherwise observed min/max.</td></tr><tr><td>clamp</td><td>Restricts output to [0,1].</td></tr></table>	Symbol	Meaning	x	Raw value.	L, H	Q05/Q95 when the robust range is materially narrower; otherwise observed min/max.	clamp	Restricts output to [0,1].	<table><tr><th>Symbol</th><th>Meaning</th></tr><tr><td>Delta v_i</td><td>Adjacent positive difference between unique observed values.</td></tr><tr><td>step</td><td>Median of those positive differences.</td></tr></table>	Symbol	Meaning	Delta v_i	Adjacent positive difference between unique observed values.	step	Median of those positive differences.
Symbol	Meaning														
x	Raw value.														
L, H	Q05/Q95 when the robust range is materially narrower; otherwise observed min/max.														
clamp	Restricts output to [0,1].														
Symbol	Meaning														
Delta v_i	Adjacent positive difference between unique observed values.														
step	Median of those positive differences.														

Gower distance

$$d_G(A, B) = \frac{1}{m} \sum \delta_j$$
$$\delta_j^{num} = \min\left(1, \frac{\sum_{i=1}^m |a_i - b_i|}{R_j}\right)$$
$$\delta_j^{cat} = 0 \text{ if equal; } \delta_j^{cat} = 1 \text{ if different}$$

Symbol	Meaning
m	Number of comparable effective dimensions.
delta_j	Contribution of dimension j.
R_j	Numeric range used in dimension j.
Numeric	Absolute difference divided by range, capped at 1.
Categorical	0 if equal; 1 if different.

Step limitation: In genetic optimization, intermediate @opt levels may be absent. The inferred step can therefore exceed the original MT5 step.

2. Clustering methods and adaptive policy

Optimization type	Possible regime	Clustering path
Genetic	SPARSE or HYBRID. Never DENSE solely because many passes exist.	Quality-elite ensemble. K-Medoids + Gower is the canonical prototype; HDBSCAN adds non-forced density evidence when enough elite points exist; hierarchical clustering is used for smaller elite samples.
Slow / complete	SPARSE, HYBRID or DENSE, depending on observed grid coverage and evidence.	With a usable grid, Quality Plateau is used. Otherwise OCA falls back to K-Medoids + HDBSCAN, with hierarchical clustering or the sparse ensemble when required.

Method	Role in OCA	Internal support conditions
K-Medoids + Gower	Prototype partition whose center is always an observed pass.	Silhouette ≥ 0.17 ; Davies-Bouldin ≤ 3 ; stability ≥ 0.45 ; validity ≥ 0.43 ; adequate minimum cluster size.
HDBSCAN analogue	Non-forced density evidence using sparse kNN, mutual reachability, MST and a condensed hierarchy.	For ≥ 2 clusters: noise $\leq 90\%$; silhouette ≥ 0.18 ; DB ≤ 3 ; sparse DBCV ≥ 0.15 ; hierarchy persistence ≥ 0.20 .
Hierarchical single-link	Fallback for smaller samples; cuts the largest MST edges.	Silhouette ≥ 0.18 ; DB ≤ 3 ; cut stability ≥ 0.04 ; validity ≥ 0.42 ; safe limit 600 rows.
Mixed K-Means / Gower	Retained for diagnostics and compatibility.	Forced partition. It does not establish final structure by itself in the quality-first path.

Composite validity used by forced partitions

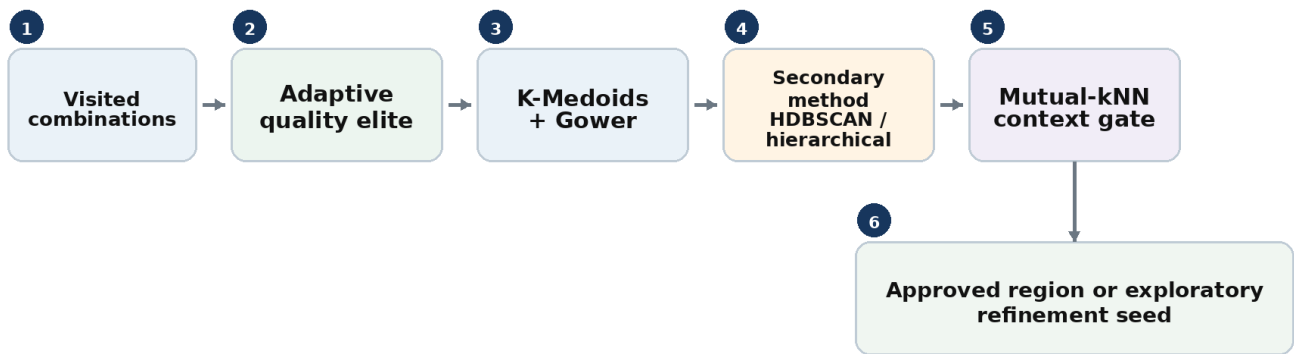
$$sep = \text{clamp}\left(\frac{silhouette + 1}{2}, 0, 1\right), \quad comp = \frac{1}{1 + DB}, \quad bal = \text{clamp}(5f_{min}, 0, 1)$$

$$Validity = \text{clamp}(0.45 sep + 0.25 comp + 0.15 bal + 0.15 stability, 0, 1)$$

Symbol	Meaning
sep	Separation term derived from silhouette and mapped to [0,1].
comp	Compactness term: $1 / (1 + \text{Davies-Bouldin})$.
f_min	Fraction of the sample in the smallest cluster.
bal	Balance term derived from $5 \times f_{min}$, capped at 1.
stability	Method-specific stability estimate in [0,1].
Validity	Weighted composite score clipped to [0,1].

3. Region formation in genetic optimization

GENETIC OPTIMIZATION - REGION FORMATION



Component	Rule
Adaptive elite	SPARSE uses approximately 20% of the sample; HYBRID approximately 15%, subject to dimension-dependent minimum and maximum elite sizes.
Prototype	K-Medoids + Gower is always enabled inside the sparse elite ensemble.
Secondary method	HDBSCAN if elite size $\geq \max(30, 4D)$. Otherwise hierarchical clustering when elite size $\geq \max(16, 2D+8)$ and ≤ 600 .
Context	mutual-kNN, adaptive radius, reciprocal support and same-region fraction.
Minimum same-region fraction	0.60 in SPARSE; 0.65 in HYBRID.

Sparse-region admission

Condition	Threshold
Region size	$n \geq 4$
Context-supported members	$\geq \max(3, \text{ceil}(0.10 \times n))$
Mean normalized quality	≥ 0.62
Mean local quality	≥ 0.55
Mean local dispersion	≤ 0.24

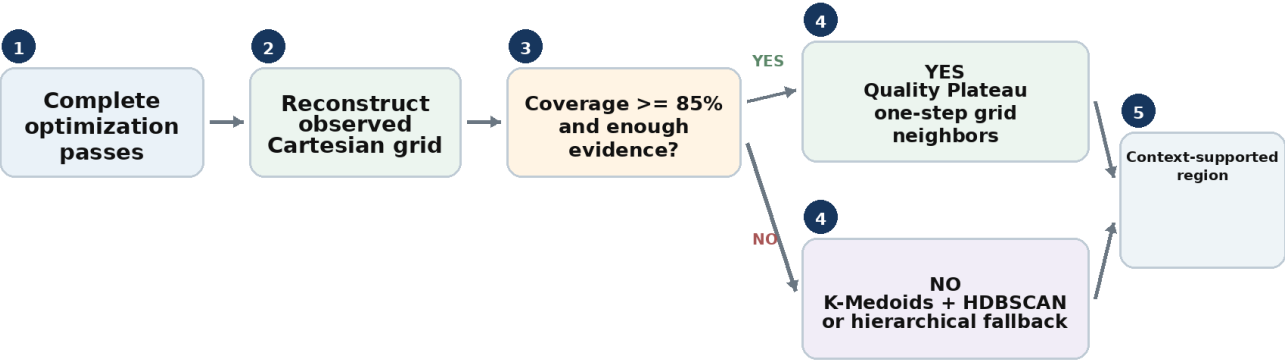
Sparse-region score

$$Score_{sparse} = 0.45 \bar{q} + 0.35 \bar{q}_{local} + 0.20(1 - \min(1, \bar{d}/0.30))$$

Symbol	Meaning
$\bar{q}$	Mean normalized quality of the region.
$\bar{q}_{local}$	Mean local quality of the region.
$\bar{d}$	Mean local dispersion.
$Score_{sparse}$	Ordering score applied only after the admission gate.

4. Region formation in slow / complete optimization

SLOW / COMPLETE OPTIMIZATION - REGION FORMATION



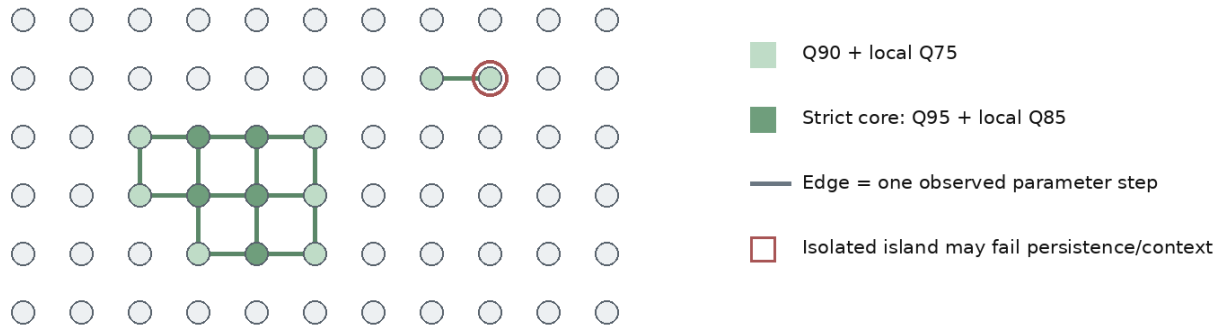
Observed-grid coverage

$$Grid\ coverage = \frac{N_{unique\ parameter\ combinations}}{\prod_j Cardinality_j}$$

Symbol	Meaning
N_unique	Number of distinct effective parameter vectors observed in the report.
Cardinality_j	Number of distinct observed values in effective dimension j.
Grid coverage	Observed combinations divided by the Cartesian product reconstructed from observed cardinalities.

Quality Plateau requirement	Threshold
Optimization type	SLOW_COMPLETE
Observed-grid coverage	>= 0.85
Unique combinations	N >= max(64, 12D)
Topology	At least one effective dimension must support coherent one-step adjacency.

QUALITY PLATEAU - ONE-STEP CONNECTIVITY



Component selection	Rule
Plateau elite	Quality $\geq$ Q90; local quality $\geq$ Q75; local dispersion ≤ 0.25 ; minimum one-step context satisfied.
Strict core	Quality $\geq$ Q95 and local quality $\geq$ Q85.
Minimum component size	$\max(5, \min(64, \text{ceil}(\sqrt{N}/4)))$.
Persistence	Strict-core count $\geq \max(3, \text{ceil}(0.08 \times n))$ and strict fraction $\geq 8\%$.
Final context	Contextual coverage $\geq 55\%$ and stable fraction $\geq 55\%$.

Quality Plateau validity and stability

$$\text{Validity} = \text{clamp}(0.45 \bar{q} + 0.40 \bar{q}_{\text{local}} + 0.15 C, 0, 1)$$

$$\text{Stability} = \text{clamp}(0.50 \min(1, P/0.30) + 0.50 F_{\text{stable}}, 0, 1)$$

Symbol	Meaning
$\bar{q}$	Mean normalized quality of accepted components.
$\bar{q}_{\text{local}}$	Mean local quality of accepted components.
C	Contextual coverage.
P	Fraction of strict-core points inside accepted components.
F_{stable}	Fraction of covered points satisfying stricter local-stability conditions.
Validity / Stability	Both scores are clipped to [0,1].

5. Context gate, cluster admission and representative

Per-pass context gate

Test	Rule
Isolated peak	Rejected.
Reciprocal neighbors	Must meet the region-specific minimum support.
Same-region neighbors	Must meet the region-specific minimum support.
Same-region fraction	Must meet the regime floor: 0.60 / 0.65 / 0.70.
Local quality	Must meet the adaptive floor; fallback is 0.50.
Local dispersion	≤ 0.22 .

Regional support: For regions larger than two members, the base support is $\max(2, \min(4, \text{ceil}(\log_2(n))))$; the final requirement also respects the neighborhood analyzer's minimum support and the region's actual size.

Standard cluster admission

Condition	Threshold
Members	≥ 4
Context-supported members	$\geq \max(3, \text{ceil}(0.10 \times n))$
Mean dispersion	≤ 0.24
Mean boundary score	≤ 0.52
Mean normalized quality	≥ 0.35

Approved-cluster ordering score

$$Score_{cluster} = 0.45 \bar{q} + 0.30(1 - \min(1, \bar{d}/0.30)) + 0.25(1 - \bar{b})$$

Symbol	Meaning
$\bar{q}$	Mean normalized quality.
$\bar{d}$	Mean local dispersion.
$\bar{b}$	Mean boundary score; lower means more interior.
$Score_{cluster}$	Ordering score after all admission thresholds are met.

5.1 Candidate and representative selection

Candidate quality

$$Q_{cand} = 0.32P + 0.25L + 0.16D + 0.17I + 0.10(1 - \min(1, d/0.30)) - 0.25Iso$$

Symbol	Meaning
P	Normalized performance quality.
L	Local neighborhood quality.
D	Normalized neighborhood density.
I	Interior term: 1 - boundary score.
d	Local dispersion.
Iso	1 for an isolated peak; otherwise 0.

Primary representative priority

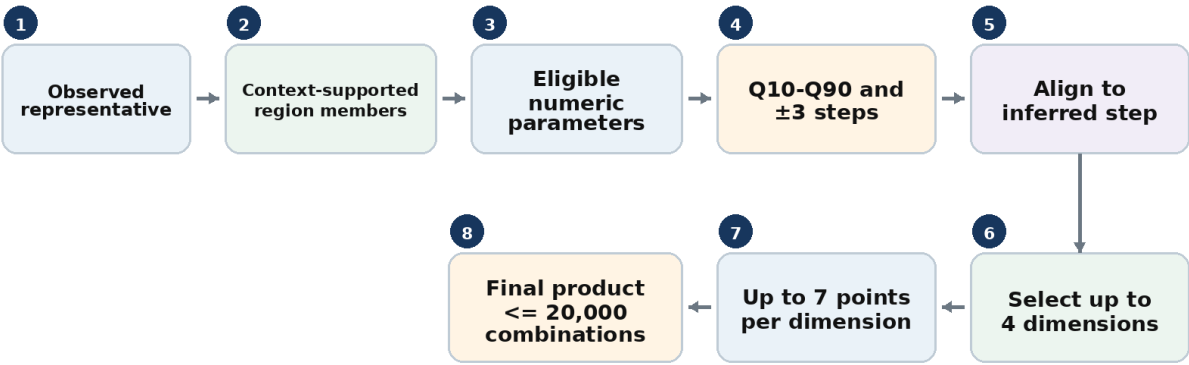
$$Priority = 0.72 Q_{cand} + 0.18(1 - \min(1, C_{medoid})) + 0.10 S_{local}$$

Symbol	Meaning
Q_cand	Candidate-quality score.
C_medoid	Mean Gower medoid cost inside the region, capped at 1 in the priority term.
S_local	Same-region local support fraction.
Priority	Final ranking used to select one observed representative per approved cluster.

Multiple clusters: Each approved cluster can export exactly one primary observed representative. Final export is not truncated by a user quota.

6. refine.set construction

refine.set - LOCAL GRID CONSTRUCTION



Local refinement range

$$[start, stop] = [Q_{10}, Q_{90}] \cap [current - 3 \text{ step}, current + 3 \text{ step}]$$

Symbol	Meaning
Q10, Q90	10th and 90th percentiles of context-supported same-region values.
current	Parameter value of the observed representative.
step	Inferred observed step.
start / stop	Intersection of the percentile interval with the +/-3-step interval, then step-aligned.

Rule	Implementation
Eligible parameters	Only int/double parameters that are non-constant, non-redundant and have step >0.
Local sample	Up to 2,000 region members are scanned; only context-supported observations contribute values.
Minimum evidence	>=5 valid contextual observations and >=3 distinct values.
Range	Q10-Q90, containing the current value and capped by current +/-3 x step.
Alignment	Start and stop are aligned to the inferred step and clipped to observed min/max.
Points per dimension	Maximum 7.
Active dimensions	Maximum 4.
Grid product	Maximum 20,000 combinations.
Unselected parameters	Fixed at the representative value.

Dimension-selection score

$$Score_j = \min(U_j, 32) \cdot \min\left(1, \frac{N_j}{\max(5, N_{region})}\right)$$

Symbol	Meaning
U_j	Distinct contextual values observed for parameter j , capped at 32.
N_j	Valid context-supported observations for parameter j .
N_{region}	Number of members in the region.
$Score_j$	Priority used to choose which eligible dimensions are expanded first.

7. Limitations

Limitation	Practical consequence
Genetic sample is selective	Only visited combinations are available to the analyzer; unvisited regions are not represented.
Inferred step	May differ from the original @opt step when intermediate levels are absent from the report.
Reconstructed coverage	Observed unique combinations divided by the observed-cardinality product does not prove coverage of the original MT5 theoretical grid.
Heuristic thresholds	ARI, quality, dispersion, boundary and coverage limits are v1.75 engineering policies, not universal significance levels.
Duplicates	Exact parameter duplicates do not weight the global geometry, but statistics after label expansion can again reflect multiplicity.
Refinement	Adds fitting inside the same in-sample data.
Repeated genetic runs	Measure operational repeatability; they do not establish statistically independent searches.
Statistical significance	OCA does not provide p-values, confidence intervals or formal significance tests for future performance.
Multiple testing / data snooping	No formal multiplicity correction is applied to the large number of parameter configurations examined.

Computational limits and approximations

Component	Limit / behavior
Text file	256 MiB per file.
Imported table	<2,000,000 rows and <=10,000,000 cells.
Feature matrix	<=30,000,000 cells; no full $N \times N$ distance matrix is stored.
K-Medoids	Direct path up to 2,500 rows; above that, deterministic landmarks are used (approximately 650-1,000 depending on D).
HDBSCAN	Exact pair-evaluation budget of 2,000,000 pairs; above that, deterministic landmarks are used, capped at 1,000.
Hierarchical	Safe limit of 600 rows.
Primary representative	Scans at most 400 members per cluster.
refine.set	Scans at most 2,000 contextual members per parameter.