CLUSTER STABILITY IN ELECTORAL DATA (2000–2016)

Formal Analysis using Hierarchical Ward Clustering and AWE Criterion

Problem Setup

Given $X \in \mathbb{R}^{3142\times 5}$, where each row corresponds to a U.S. county and each column contains the Democratic vote share in the elections $\{2000, 2004, 2008, 2012, 2016\}$, perform the following:

1. Compute the pairwise **Euclidean distance matrix** $D \in \mathbb{R}^{3142\times 3142}$.
2. Apply Ward's hierarchical clustering to D.
3. For $k = 2$ to 6, compute:

$$\mathrm{AWE}_k = 2\log B_k, \quad \text{where} \quad B_k = \frac{P(m_k)}{P(m_1)}$$

   assuming: $\log B_k = 25(k-1) - 2k$.
4. Identify optimal k via $\Delta_k = \mathrm{AWE}_k - \mathrm{AWE}_{k-1}$.

MATHEMATICAL DERIVATION

Given:

$$\log B_k = 25(k-1) - 2k \quad \Rightarrow \quad \mathrm{AWE}_k = 2\log B_k = 2(23k - 25) = 46k - 50$$

2whitegray!10

k	AWE_k	$\Delta_k = \mathrm{AWE}_k - \mathrm{AWE}_{k-1}$
1	0	–
2	42	42
3	88	46
4	134	46
5	180	46
6	226	46

Conclusion: Maximum Δ_k occurs first at $k = 3$; hence, $k^* = 3$ is chosen via the parsimony principle.

INTERPRETATION AND POLITICAL RELEVANCE

1. THREE PERSISTENT BLOCS

The value $k^* = 3$ supports a stable trichotomy:

$$(i)\text{Democratic urban clusters}, \quad (ii)\text{Republican rural zones}, \quad (iii)\text{Heterogeneous swing regions}$$

These reflect spatial and ideological cohesion across time.

2. TEMPORAL STABILITY SIGNAL

For $k > 3$, Δ_k remains constant:

$$\Delta_k = 46 \quad \text{for all } k = 3, 4, 5, 6$$

This plateau in AWE_k implies no significant informational gain, evidencing that the blocs are temporally resilient over five election cycles.

3. STRATEGIC TARGETING

The third (swing) cluster likely holds:

Highest campaign leverage and policy reactivity.

Campaigns should prioritize this region, whereas realignment of entrenched urban/rural groups would require structural shocks of substantial magnitude.

SUMMARY INSIGHT

Ward clustering with AWE analysis identifies three enduring U.S. electoral blocs. Their high internal cohesion and unchanged Δ_k beyond $k = 3$ affirm long-term bloc persistence (2000–2016). These results align strongly with spatially observable voting behavior and provide a principled basis for both political analysis and campaign strategy.

DISCRIMINANT SHAPE & OUTLIER DETECTION

Model-Based Clustering in U.S. County Demographics

Problem Setup

Let matrix $X \in \mathbb{R}^{1000\times 4}$ represent data for 1,000 U.S. counties with features:

- x_1: Percent of adults with a college degree
- x_2: Median household income (USD)
- x_3: Urbanization index ($0 =$ fully rural, $1 =$ fully urban)
- x_4: Democratic vote share (most recent election)

Model-based clustering assuming Gaussian components yields eigenvalue ratios:

$$\begin{aligned}
&\text{Cluster 1:} \quad \lambda_1/\lambda_4 = 5.3\\
&\text{Cluster 2:} \quad \lambda_1/\lambda_4 = 1.2\\
&\text{Cluster 3:} \quad \lambda_1/\lambda_4 = 18.6
\end{aligned}$$

A new county has feature vector:

$$\mathbf{x}_{\text{new}} = \begin{bmatrix} 0.22 \\ 44{,}000 \\ 0.65 \\ 0.52 \end{bmatrix} \quad \text{assigned to Cluster 2, whose statistics are:} \quad \boldsymbol{\mu}_2 = \begin{bmatrix} 0.25 \\ 48{,}000 \\ 0.62 \\ 0.55 \end{bmatrix}$$

Covariance matrix: $\Sigma_2 = \text{diag}(0.01,\ 2500^2,\ 0.01,\ 0.01)$

MAHALANOBIS DISTANCE CALCULATION

$$\mathbf{d} = \mathbf{x}_{\text{new}} - \boldsymbol{\mu}_2 = \begin{bmatrix} -0.03 \\ -4000 \\ 0.03 \\ -0.03 \end{bmatrix}, \quad \Sigma_2 = \text{diag}(0.01,\ 2500^2,\ 0.01,\ 0.01)$$

Because Σ_2 is diagonal:

$$\text{MD}^2 = \sum_{i=1}^{4} \frac{d_i^2}{\sigma_i^2} = \frac{0.03^2}{0.01} + \frac{(-4000)^2}{2500^2} + \frac{0.03^2}{0.01} + \frac{0.03^2}{0.01} = 0.09 + 2.56 + 0.09 + 0.09 = 2.83$$

$$\boxed{\text{MD} = \sqrt{2.83} \approx 1.68}$$

OUTLIER DETERMINATION (THRESHOLD: 3.5)

Since $1.68 < 3.5$, this county lies **well within** the typical spread of Cluster 2 and is **not flagged** as an outlier.

CLUSTER SHAPE INTERPRETATION

Cluster 2's eigenvalue ratio $\lambda_1/\lambda_4 = 1.2$ implies *near-isotropy*, i.e., almost equal variance across all principal directions. This results in a **spherical cluster**, sharply contrasting with elongated ellipsoids seen in Clusters 1 (5.3) and 3 (18.6).
Combined with the low Mahalanobis distance, this geometry supports the conclusion that the new county's demographic profile is **statistically coherent** with Cluster 2's internal shape.

SPATIAL CLUSTER CONTIGUITY IN STATEWIDE PATTERNS

Analyzing Geopolitical Coherence in Regional k-Means Voting Blocs

Problem Setup

Suppose a U.S. state with 100 counties is studied. You are given:

- Matrix $X \in \mathbb{R}^{100\times 3}$ with columns:
  1. Voter turnout
  2. Democratic vote share
  3. Republican vote share
- Adjacency matrix $A \in \{0,1\}^{100\times 100}$ where $A_{ij} = 1$ iff counties i and j share a border.

You perform **k-means clustering** on X for $k = 2, 3, 4, 5$. Define the **Cluster Contiguity Ratio** (CCR_k) as:

$$\mathrm{CCR}_k = \frac{\text{Number of adjacent pairs in the same cluster}}{\text{Total number of adjacent pairs}}$$

Assume the following empirical values:

k	CCR_k
2	0.94
3	0.78
4	0.61
5	0.43

SPATIAL COHERENCE VERDICT

Maximum spatial coherence: $\mathrm{CCR}_2 = 0.94$ is highest, implying the $k = 2$ clustering preserves almost all adjacency pairs—yielding the strongest **geographic cohesion**.

INTERPRETIVE RATIONALE

SPATIAL AUTOCORRELATION

High Moran-I statistics in electoral behavior indicate that adjacent counties vote similarly. Thus, the 94% adjacency retention at $k = 2$ signals alignment with latent spatial dependence structures.

VOTING REGIONALISM

A binary partition ($k = 2$) often captures:

Urban/metro Democratic core vs. Rural Republican periphery

This split mirrors enduring macro-regional cleavages, making the $k = 2$ configuration both politically and geographically legible.

COMPACTNESS VS INTERPRETABILITY

k	Geometric Compactness	Spatial Contiguity	Interpretability
2	Lowest (broad centroids)	Highest (0.94)	Coarse—collapses swing regions
3	Moderate	Good (0.78)	Adds suburban/swing class; interpretable
4–5	Highest	Fragmented (≤ 0.61)	Scattered clusters; cartographically unstable

Trade-off:

- Increasing k:
  - Improves **within-cluster homogeneity**, aiding modeling accuracy
  - Reduces **spatial contiguity**, disrupting geographic storytelling

In applied political geography, the smallest k that preserves distinct, interpretable blocs is often favored. Here, $k = 3$ offers a strong compromise—retaining over 75% spatial cohesion while isolating a pivotal suburban swing belt. Nevertheless, the most geographically coherent solution remains unambiguously $k = 2$.