The Soul of Wine

Methods Primer

A Guide to the Statistical and Network Methods

Jure Skarabot · May 2026

Companion documents: Summary · Technical Appendix · Methods Primer · Data Appendix · Region Reference (Identity, Terroir)

Introduction

The Region Affinities tool runs two independent classification pipelines over 59 wine regions and tests whether they agree. The identity pipeline takes six expert-scored cultural dimensions, standardises them, and clusters the regions into six groups. The terroir pipeline takes factual text descriptions of each region's geography and climate, converts them into numerical features, compresses those features, and clusters the regions into seven groups. The two pipelines never share data. A formal independence test then asks whether the two cluster solutions agree more than chance would predict.

This primer explains the methods used at each stage, in plain language. Each section explains the intuition, gives the formal definition, and describes the algorithm. The treatment assumes comfort with basic statistics (mean, variance, correlation) but not prior exposure to clustering or text vectorisation. For full mathematical detail, see the Technical Appendix.

1. TF-IDF Vectorisation

What it does. TF-IDF (Term Frequency–Inverse Document Frequency) converts a collection of text documents into a matrix of numbers. Each row is a document; each column is a word or short phrase; each cell records how distinctively that word characterises that document. Words that appear often in one document but rarely across the corpus get high scores. Words that appear in nearly every document — "the", "and", "wine" — get low scores because they fail to distinguish documents.

How it works. For each (term, document) pair, TF-IDF multiplies two quantities. The term frequency measures how often the term appears in that document; the version used here applies sublinear scaling, $1 + \log f_{t,d}$, which dampens the impact of very frequent terms. The inverse document frequency measures how rare the term is across the corpus, computed as $\log(N/N_t)$, where N is the corpus size and N_t is the number of documents containing the term. The product gives high weight to terms that are frequent in a specific document but rare in the corpus overall.

In Region Affinities. The Layer 2 terroir narratives for the 59 regions are TF-IDF vectorised over a vocabulary of unigrams and bigrams (single words and two-word phrases), capped at 800

features. The result is a 59×800 matrix. Distinctive terms emerge naturally: "volcanic basalt" for Etna and Santorini, "slate slopes" for the Mosel, "alluvial gravels" for Bordeaux's Left Bank, "schist" for the Douro and parts of the Languedoc.

Why TF-IDF rather than raw word counts? Raw word counts let common words dominate. The word "wine" appears in every region's terroir description, so it provides no information about which regions are similar. TF-IDF down-weights such ubiquitous terms and up-weights distinctive ones — which is exactly what is needed for clustering by terroir vocabulary.

2. Standardisation

What it does. Standardisation puts all features on a common scale. For each feature, it subtracts the mean and divides by the standard deviation, producing values with zero mean and unit variance. This step matters because clustering algorithms measure distances, and distance computations are biased toward features with larger numerical ranges.

Why both pipelines need it. The D-scores already share a common scale (-2 to $+2$), but they differ in how they vary across regions. Standardisation ensures each of the six dimensions contributes equally regardless of which dimension happens to span a wider empirical range. The TF-IDF features have very different scales — some terms appear in many documents, others in only one — and standardisation neutralises this disparity before PCA.

3. Principal Component Analysis (PCA)

What it does. PCA finds new axes (the principal components) that capture as much variance in the data as possible, in order. The first component is the direction along which the data varies most. The second is orthogonal to the first and captures the next-most variance. And so on.

Why use it. Two reasons. First, dimensionality reduction: a 59×800 TF-IDF matrix is too sparse and too noisy for clustering directly. Reducing to 10 components retains the structural patterns while discarding much of the noise. Second, decorrelation: even when no dimensionality reduction is needed (as for the six D-scores), PCA produces orthogonal coordinates, which makes K-means more numerically stable.

Mechanics. For a centred data matrix X, the principal components are the eigenvectors of the covariance matrix $\Sigma = X^\top X/(n-1)$. The eigenvalues give the variance captured by each component. The cumulative ratio

$$\rho_m = \frac{\lambda_1 + \cdots + \lambda_m}{\lambda_1 + \cdots + \lambda_p}$$

tells how much of the total variance the first m components retain.

In Region Affinities. The identity pipeline retains all 6 components ($\rho_6 = 100\%$), so PCA acts as an orthonormal rotation rather than a reduction. The terroir pipeline reduces 800 TF-IDF features to 10 components, capturing 32.2% of the variance. The retained variance may seem modest, but for high-dimensional sparse text data, 32% across 10 components is substantial — the discarded variance is largely per-term noise that would not contribute meaningfully to cluster structure.

4. K-Means Clustering

What it does. K-means partitions n points into k disjoint clusters by alternating two steps until the cluster boundaries stabilise. (1) Each point is assigned to the cluster whose centre is closest. (2) Each cluster's centre is recomputed as the mean of its assigned points. The objective minimised is the sum of squared distances from each point to its cluster centre.

Lloyd's algorithm. The classic implementation. Starting from k initial centroids:

1. Assign each point to its nearest centroid.
2. Recompute each centroid as the mean of its assigned points.
3. Repeat until no points change cluster.

The initialisation problem. K-means with random initial centroids can converge to local minima — partitions that are locally stable but globally suboptimal. Two standard countermeasures are used: *k-means++ seeding* chooses initial centroids that are spread far apart, dramatically improving convergence quality; and n_{init} *random restarts* run the algorithm n_{init} times from different seedings and return the best result. The pipeline uses k-means++ with $n_{\text{init}} = 20$.

Choosing k. The cluster count k is not learnt by the algorithm; it must be specified. The choice is justified by a combination of silhouette scores (which prefer clusters that are internally cohesive and well-separated), interpretability (whether each cluster has a clear character), and stability (whether perturbations of the algorithm produce comparable results). For Region Affinities, $k_I = 6$ is the silhouette optimum and the smallest k at which all conceptually distinct identity types emerge as separate clusters. $k_T = 7$ produces the most geographically interpretable terroir partition.

In Region Affinities. K-means is applied independently to the PCA-reduced identity matrix ($k_I = 6$) and the PCA-reduced terroir matrix ($k_T = 7$). Each pipeline is run 20 times with different random initialisations; the best partition (lowest within-cluster sum of squares) is retained.

5. Silhouette Score

What it does. The silhouette score measures how good a clustering solution is, on a scale from -1 to $+1$. For each point, the score compares how close it is to other points in its own cluster versus points in the next-nearest cluster. A score near $+1$ means the point is firmly inside its cluster and far from any neighbouring cluster. A score near 0 means the point is on a cluster boundary. A score near -1 means the point is closer to a neighbouring cluster than its own — a likely misassignment.

Formal definition. For each point i, $a(i)$ is the average distance from i to other points in its own cluster, and $b(i)$ is the average distance from i to points in the nearest neighbouring cluster. The silhouette is

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}.$$

The overall silhouette is the mean of $s(i)$ over all points.

Interpreting the value. For tabular data with well-separated clusters, silhouettes above 0.5 are excellent. For social-science data with inherently overlapping categories, silhouettes between 0.2 and 0.4 are the typical range for meaningful structure. Identity scores 0.3029 and terroir 0.2457 — both in the meaningful-structure range, with identity stronger than terroir, consistent

with D-scores being a deliberately designed bipolar encoding versus terroir being derived from open-ended text.

6. The Adjusted Rand Index (ARI)

What it does. The ARI measures how much two clustering solutions agree on the same set of objects. It examines all possible pairs of objects and asks: do the two solutions agree on whether this pair belongs together or apart? Crucially, it adjusts for the fact that some agreement happens by chance — the more clusters there are, the more often two random partitions will accidentally agree.

The intuition. Take 59 wine regions. There are $\binom{59}{2} = 1{,}711$ possible pairs. For each pair, the identity partition either places them in the same cluster or different clusters; the terroir partition does the same. Each pair therefore falls into one of four categories: *both same* (in the same cluster in both solutions), *both different* (in different clusters in both solutions), or one of two mixed categories. The Rand Index is the fraction of pairs in the two "agree" categories. The Adjusted Rand Index corrects for the agreement that would occur if the two partitions were random.

Properties. ARI $= 1$ if the two partitions are identical (modulo cluster relabelling). ARI $= 0$ if they agree no more than two random partitions of the same sizes would. ARI < 0 if they actively disagree (less agreement than chance). ARI is symmetric: $\text{ARI}(P, Q) = \text{ARI}(Q, P)$. ARI is meaningful when the two partitions have different numbers of clusters — the comparison does not depend on label matching.

In Region Affinities. The ARI between the identity and terroir partitions is 0.023 — effectively zero. The two classification systems agree no more than random partitions would.

7. Chi-Squared Test of Independence

What it does. The chi-squared test of independence is a classical statistical test. It asks: in a two-way contingency table (rows: identity clusters; columns: terroir clusters; cells: counts of regions in each combination), is there evidence that the row variable and the column variable are statistically dependent?

How it works. If the two variables are independent, the expected count in cell (i, j) is $e_{ij} = (\text{row total})_i \cdot (\text{col total})_j / n$. The chi-squared statistic measures how far the observed counts n_{ij} deviate from these expected counts:

$$\chi^2 = \sum_{i,j} \frac{(n_{ij} - e_{ij})^2}{e_{ij}}.$$

Under the null hypothesis of independence, this statistic follows a chi-squared distribution with $(\text{rows} - 1)(\text{cols} - 1)$ degrees of freedom. Large values produce small p-values and reject the null. Small values produce large p-values, consistent with independence.

A practical caveat. The chi-squared distribution is an asymptotic approximation that requires expected counts of at least 5 per cell. With 59 regions distributed across a 6×7 table, many cells have expected counts well below 5, weakening the test's accuracy. The ARI is therefore the primary independence measure; the chi-squared test is reported as a corroborating diagnostic.

In Region Affinities. $\chi^2 = 38.776$ with df $= 30$ and $p = 0.131$. The null of independence is not rejected at $\alpha = 0.05$.

8. The D-Score System

What it does. The D-scores are six expert-scored dimensions that capture the cultural identity of each wine region. Unlike the terroir classification, which is fully algorithmic, the D-scores involve structured human judgement: a subject matter expert reads each region's identity narrative (Layer 1) and scores it on six bipolar axes, each ranging from -2 to $+2$.

The dimensions.

- D_1 Interiority $\leftrightarrow$ Exteriority — inward-facing and place-defined ($+2$) vs. outward-facing and market-oriented (-2).
- D_2 Struggle $\leftrightarrow$ Ease — defined by difficulty and endurance ($+2$) vs. pleasure and comfort (-2).
- D_3 Tradition $\leftrightarrow$ Reinvention — anchored in deep custodial tradition ($+2$) vs. defined by disruption and reinvention (-2).
- D_4 Individual $\leftrightarrow$ Collective — driven by solitary vision ($+2$) vs. communal, cooperative character (-2).
- D_5 Urgency $\leftrightarrow$ Timelessness — urgent and present-focused ($+2$) vs. timeless and eternal (-2).
- D_6 Earthly $\leftrightarrow$ Transcendent — grounded and earthly ($+2$) vs. reaching toward something spiritual (-2).

Why expert scoring rather than survey instruments? Cross-cultural psychology (Hofstede, GLOBE, Ronen–Shenkar) typically derives identity dimensions from large-scale attitudinal surveys. Wine regions cannot easily be polled in this manner — a region is not a population of respondents. The D-score system substitutes expert reading of a structured identity narrative for survey collection, accepting the trade-off of single-rater subjectivity in exchange for tractability across a moderately sized region set. The structured narrative format constrains the rater's interpretation; the bipolar dimensional framing constrains the dimensions of judgement.

9. The Pipeline in Summary

The methods form two parallel pipelines that produce comparable cluster outputs.

Identity pipeline. 6-dimensional D-score matrix $\rightarrow$ standardisation $\rightarrow$ PCA (6 components, 100% variance) $\rightarrow$ K-means ($k = 6$) $\rightarrow$ silhouette score 0.3029.

Terroir pipeline. 6-field text per region $\rightarrow$ TF-IDF vectorisation (800 features) $\rightarrow$ standardisation $\rightarrow$ PCA (10 components, 32.2% variance) $\rightarrow$ K-means ($k = 7$) $\rightarrow$ silhouette score 0.2457.

Independence test. The two cluster assignments form a 6×7 contingency table. The Adjusted Rand Index is 0.023; chi-squared is 38.776 with df $= 30$, $p = 0.131$. Both measures fail to find evidence that the partitions are related. Terroir and cultural identity are independent classification systems.

References

- Arthur, D. & Vassilvitskii, S. (2007). k-means++: The advantages of careful seeding. In *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, 1027–1035.
- Hartigan, J. A. & Wong, M. A. (1979). A K-Means Clustering Algorithm. *Applied Statistics*, 28(1), 100–108.
- Hubert, L. & Arabie, P. (1985). Comparing partitions. *Journal of Classification*, 2(1), 193–218.
- Jolliffe, I. T. (2002). *Principal Component Analysis* (2nd ed.). Springer.
- Pearson, K. (1900). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine*, 50(302), 157–175.
- Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20, 53–65.
- Salton, G. & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, 24(5), 513–523.