The Soul of Wine

Methods Primer

A plain-language guide to the statistical and computational methods used in this study. No prior knowledge of statistics or computer science is assumed.

Natural Language Processing (NLP)

Natural language processing is a branch of computer science that teaches machines to read and interpret human text. In this study, we use NLP to convert written descriptions of wine regions into numerical data that a computer can analyse. The goal is to find patterns in how regions are described — patterns that might be invisible to a human reader working through 59 profiles one at a time.

TF-IDF: Turning Words into Numbers

TF-IDF stands for Term Frequency–Inverse Document Frequency. It is a way of measuring how important a word is to a particular document within a collection. The intuition is straightforward: if a word appears frequently in one region's terroir profile but rarely across other profiles, it is probably distinctive to that region. The word "slate," for example, appears heavily in descriptions of Mosel and Rheingau but almost nowhere else — so TF-IDF gives it a high score for those regions.

The "TF" part counts how often a word appears in a single document. The "IDF" part penalises words that appear in many documents — common words like "wine" or "region" get low scores because they do not help distinguish one region from another. The product of TF and IDF gives each word a score for each region. The result is a table where each row is a region and each column is a word, with the cells containing importance scores. This table is the input to the clustering algorithm.

Principal Component Analysis (PCA)

When we convert text into TF-IDF scores, we end up with hundreds of columns — one for each word and word-pair in the vocabulary. Many of these columns carry similar information. PCA is a technique that compresses this high-dimensional data into a smaller number of dimensions while preserving as much of the meaningful variation as possible.

Think of it as finding the "essential axes" along which regions differ most. If many words tend to co-occur (say, "granite," "schist," and "steep" all appear together), PCA combines them into a single dimension that captures that pattern. In this study, we reduce the TF-IDF data to 10 principal components for terroir clustering, and use all 6 dimensions (the D-scores) directly for identity clustering.

PCA also enables the scatter plots in the Cluster Map visualisation. Each region is plotted using its first two principal components as coordinates — the two axes that capture the most variation. Regions that are close together in the plot are similar in their overall profile.

K-Means Clustering

K-means is an algorithm that groups data points into a specified number of clusters (k). The algorithm works iteratively: it starts by randomly placing k "centres" in the data space, assigns each region to its nearest centre, then moves each centre to the average position of its assigned regions. This process repeats until the centres stop moving. The result is k groups of regions, each defined by proximity in the data space.

The choice of k — how many clusters — is a modelling decision. For identity clustering, we use k=6, producing six identity types. For terroir clustering, we tested k=3 through k=10 and selected

k=7 because it produced the most interpretable clusters while maintaining strong statistical independence from the identity solution. The terroir clustering is fully algorithmic: no human judgment is involved in assigning regions to terroir clusters.

Measuring Cluster Quality

Silhouette Score

The silhouette score measures how well-defined the clusters are. For each region, it compares how similar the region is to others in its own cluster versus how similar it is to regions in the nearest neighbouring cluster. Scores range from -1 to +1. A score near +1 means the region fits snugly in its cluster; near 0 means it sits on the boundary between clusters; near -1 means it may be in the wrong cluster. The overall silhouette score is the average across all 59 regions. In social science data, scores above 0.2 indicate meaningful structure.

Testing Independence

Adjusted Rand Index (ARI)

The central question of this study is whether terroir clustering and identity clustering agree. The Adjusted Rand Index is a standard measure of agreement between two groupings of the same objects. It looks at every possible pair of regions and asks: do the two clustering solutions agree on whether this pair belongs together or apart?

ARI = 1.0 means the two solutions are identical. ARI = 0.0 means they agree no more than random chance would predict. ARI < 0 means they agree less than chance — they are actively disagreeing. In this study, the ARI between terroir clusters and identity clusters is approximately zero, meaning the two classification systems share no structure whatsoever.

Chi-Square Test

As a second check, we use the chi-square test of independence. This is a standard statistical test that asks: if I know which terroir cluster a region belongs to, does that help me predict which identity cluster it belongs to? The test produces a p-value — a number between 0 and 1 that represents the probability of seeing the observed data if the two systems were truly independent. A high p-value (above 0.05 by convention) means we have no evidence that the systems are related. In this study, the p-value is well above 0.05, confirming independence.

The D-Score System

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 judgment — a subject matter expert reads each region's identity narrative and scores it on six bipolar axes, each ranging from -2 to +2:

D1: Interiority vs. Exteriority — Is the region's identity inward-facing and place-defined (+2), or outward-facing and market-oriented (-2)?
D2: Struggle vs. Ease — Is the region defined by difficulty and endurance (+2), or by pleasure and comfort (-2)?
D3: Tradition vs. Reinvention — Is the region anchored in deep custodial tradition (+2), or

defined by disruption and reinvention (-2)?
D4: Individual vs. Collective — Is identity driven by solitary vision (+2), or by communal, cooperative character (-2)?
D5: Urgency vs. Timelessness — Does the region feel urgent and present-focused (+2), or timeless and eternal (-2)?
D6: Earthly vs. Transcendent — Is the region grounded and earthly (+2), or reaching toward something spiritual or transcendent (-2)?

Putting It All Together

The study runs two parallel pipelines. The identity pipeline takes expert D-scores, standardises them, and clusters regions into 6 groups using k-means. The terroir pipeline takes factual text descriptions, converts them to numerical features using TF-IDF, compresses them with PCA, and clusters regions into 7 groups using k-means. The two pipelines never share data. Finally, we measure whether their outputs agree — and find that they do not. Terroir and cultural identity are independent classification systems. The map is not the soul.
