THE SOUL OF WINE

Technical Appendix

Pipeline Parameters, D-Score Matrix, SME Change Log, Terroir Cluster Membership

The Soul of Wine Research Project · April 2026

Model Summary

Metric	Identity Clustering	Terroir Clustering
Method	D-scores → StandardScaler → PCA → K-means	Layer 2 TF-IDF → PCA → K-means
Input	6 expert-scored dimensions (–2 to +2)	6-field terroir profiles, TF-IDF vectorised
k	6	7 (fully model-derived)
Silhouette	0.3029	0.2457
PCA Variance	100.0%	32.2% (10 components)
Canonical Match	57/59	N/A (model-derived)

Independence Test

Test	Value	Interpretation
Adjusted Rand Index	0.023	Near-zero: clustering solutions share no structure beyond chance
Chi-square	38.776	Test statistic for 6×7 contingency table
p-value	0.131	Fails to reject independence (p > 0.05)
Conclusion	—	Terroir and identity classifications are statistically independent

1. Pipeline Parameters

1.1 Three-Stage Additive Design

The pipeline operates in three additive stages, each adding one input to test whether it improves clustering. This nested design follows the logic of additive model comparison.

Stage	Inputs	k=6 Silhouette	Question Answered
1	Layer 1 TF-IDF only	0.1138	What structure does the text reveal on its own?
2	Layer 1 TF-IDF + D-scores (α=0.7)	0.1170	Does structured expert judgment add signal beyond text?
3 ✓	**D-scores only (preferred)**	**0.2976**	**Does simplest structured input outperform text? Yes (+0.180)**

1.2 Technical Parameters

Parameter	Value
TF-IDF Vectoriser	ngram_range=(1,2), max_features=800, sublinear_tf=True. Separate vectoriser per layer.
Normalisation	StandardScaler applied independently to each input before combination.
PCA (Identity)	Reduce to 6 components from 6 D-score dimensions (100% variance retained).
PCA (Terroir)	Reduce to 10 components from 800 TF-IDF features (32.2% variance).
K-means (Identity)	k=6, n_init=20, random_state=42.
K-means (Terroir)	k=7, n_init=20, random_state=42. Fully model-derived.
Evaluation	Silhouette score (internal validity), ARI (independence test), chi-square.
D-score Scale	Integer –2 to +2. Six dimensions scored per region from Layer 1 descriptions.

1.3 Comparison to Pass 3b

Metric	Pass 3b	Pass 5
Regions	56	**59**
Best silhouette	0.2319 (k=9)	**0.3121 (k=8)**
k=6 silhouette	0.2011	**0.2976**
k=6 ARI	0.2546	**0.2668**
Method	70/30 TF-IDF/D-scores	**D-scores only**

2. D-Score Matrix — All 59 Regions

Scores: +2 (strong positive pole) to –2 (strong negative pole). 0 = balanced/neutral. Colour coding: deep green (+2), light green (+1), grey (0), amber (–1), red (–2). Sorted by cluster, then by region number within each cluster.

D1 *Interiority↔Exteriority* · ***D2*** *Struggle↔Ease* · ***D3*** *Tradition↔Reinvention* · ***D4*** *Individual↔Collective* · ***D5*** *Urgency↔Timelessness* · ***D6*** *Earthly↔Transcendent*

Old World Interior · 10 regions

#	Region	Country	Metaphor	D1	D2	D3	D4	D5	D6
9	**Burgundy**	France	*Devotion*	+2	+1	+2	+1	-1	-1
25	**Campania**	Italy	*Memory*	+1	0	+1	0	-1	-1
26	**Etna**	Italy	*Awakening*	+1	+2	0	0	+1	-1
37	**Galicia**	Spain	*Longing (Morriña)*	+1	+1	+1	+1	0	-1
17	**Mosel**	Germany	*Poetry*	+2	+2	+2	+1	-2	-2
14	**Northern Rhône**	France	*Solitude*	+2	+2	+2	+2	-2	-1
29	**Piedmont**	Italy	*Philosophy*	+2	+1	+2	+1	-2	-1
20	**Rheingau**	Germany	*Nobility*	+1	0	+2	+1	-2	0
56	**Santa Cruz Mtns**	USA	*Obsession*	+2	+1	0	+2	0	0
23	**Tokaj**	Hungary	*Melancholy*	+2	+1	+2	+1	-2	-1

Outward Ease · 5 regions

#	Region	Country	Metaphor	D1	D2	D3	D4	D5	D6
7	**Beaujolais**	France	*Joy*	-1	-1	-1	-1	+1	+1
48	**Marlborough**	NZ	*Assertion*	-2	0	-1	-1	+2	+1
15	**Provence**	France	*Pleasure*	-2	-2	0	-1	+1	+1
33	**Veneto**	Italy	*Commerce*	-2	0	-1	-1	+1	+1
58	**Walla Walla**	USA	*Community*	+1	-1	-1	-2	+1	+1

New World Reinvention · 12 regions

#	Region	Country	Metaphor	D1	D2	D3	D4	D5	D6
36	**Catalonia**	Spain	*Identity*	-1	+1	-2	+1	+2	+1
46	**Central Otago**	NZ	*Adventure*	-1	+1	-1	+1	+1	0
51	**Columbia Valley**	USA	*Determination*	-1	+1	-1	+1	+1	+2
52	**Finger Lakes**	USA	*Conviction*	-1	+1	-1	+1	+1	0
40	**Mendoza**	Argentina	*Reinvention*	-1	0	-2	0	+1	+1
53	**Napa Valley**	USA	*Ambition*	-2	0	-1	-1	+2	+1
54	**Paso Robles**	USA	*Independence*	-1	0	-2	+1	+1	+1
41	**Patagonia**	Argentina	*Extremity*	0	+2	-2	+1	+2	0
31	**Sicily**	Italy	*Resurrection*	-1	+1	-2	+1	+1	+1
49	**Stellenbosch**	S. Africa	*Aspiration*	-1	0	-1	+1	+1	+1
50	**Swartland**	S. Africa	*Rebellion*	-1	+1	-2	+1	+2	+1
57	**Sonoma**	USA	*Authenticity*	-1	0	-1	+1	0	+1

Old World Exterior · 7 regions

#	Region	Country	Metaphor	D1	D2	D3	D4	D5	D6
16	**Baden**	Germany	*Warmth*	-1	-1	+1	-1	-1	+1
8	**Bordeaux**	France	*Business*	-2	-1	+2	-2	0	+1
10	**Champagne**	France	*Society*	-2	-1	+2	-2	0	-1

#	Region	Country	Metaphor	D1	D2	D3	D4	D5	D6
11	**Châteauneuf**	France	*Family*	-1	0	+2	-2	-1	+1
5	**Dalmatian Coast**	Croatia	*Tranquility*	+2	-2	+1	-1	-2	+2
19	**Pfalz**	Germany	*Generosity*	-1	-2	+1	-1	-1	+1
39	**Rioja**	Spain	*Patience*	-1	0	+2	-1	-1	+1

Note: Old World Exterior is the most internally diverse cluster. Bordeaux and Champagne define the extreme end (D1=–2, D4=–2, D3=+2), while Baden, Pfalz, and Rioja express the same institutional character more moderately. Dalmatian Coast is the cluster's most distinctive member — its D1=+2 (strongly interior) runs counter to the group's outward tendency, but its deep timelessness and tradition place it here. The cluster holds together through shared institutional, outward-facing character, not identical D-score profiles.

Against the Odds · 10 regions

#	Region	Country	Metaphor	D1	D2	D3	D4	D5	D6
42	**Barossa Valley**	Australia	*Fortitude*	+1	+2	+1	0	0	+1
34	**Douro**	Portugal	*Endurance*	+1	+2	+2	+1	-1	+1
43	**Hunter Valley**	Australia	*Defiance*	+1	+2	+1	+1	-1	+1
12	**Jura**	France	*Eccentricity*	+2	+1	+2	+1	0	+1
28	**Liguria**	Italy	*Intimacy*	+2	+1	+1	+1	-1	+1
21	**Macedonia**	Greece	*Austerity*	+2	+2	+2	+1	0	+1
38	**Ribera del Duero**	Spain	*Severity*	+1	+2	+1	+1	0	+2
22	**Santorini**	Greece	*Survival*	+2	+2	+1	+1	+1	+1
30	**Sardinia**	Italy	*Stubbornness*	+2	+1	+2	-1	-1	+2
4	**Wagram**	Austria	*Earth*	+2	0	+2	0	0	+1

The Moderates · 15 regions

#	Region	Country	Metaphor	D1	D2	D3	D4	D5	D6
6	**Alsace**	France	*Duality*	0	0	+1	0	0	0
24	**Alto Adige**	Italy	*Precision*	0	0	+1	0	0	+1
27	**Friuli-VG**	Italy	*Dialogue*	-1	0	0	0	0	0
35	**Goriška Brda**	Slovenia	*Fortune*	-1	0	0	0	0	+1
47	**Hawke's Bay**	NZ	*Confidence*	0	0	0	0	-1	+1
1	**Kamptal**	Austria	*Discipline*	+1	0	+1	+1	0	0
13	**Loire**	France	*Sentimentality*	-1	0	0	0	0	0
45	**Maipo Valley**	Chile	*Pride*	-1	0	+1	0	-1	+1
44	**Margaret River**	Australia	*Composure*	0	0	0	0	-1	0
18	**Nahe**	Germany	*Subtlety*	+1	0	+1	+1	-1	0
55	**Santa Barbara**	USA	*Serendipity*	-1	0	0	0	0	+1
2	**Steiermark**	Austria	*Clarity*	+1	0	0	+1	+1	+1
32	**Tuscany**	Italy	*Art*	-1	-1	0	+1	0	0
3	**Wachau**	Austria	*Monumentality*	0	0	+1	+1	0	+1
59	**Willamette Valley**	USA	*Idealism*	-1	+1	-1	+1	0	-1

3. SME Change Log

16 score and metaphor changes across 13 regions, applied after SME review of the Pass 5 D-score matrix. All changes are final and incorporated into canonical documents. Three intentional cluster shifts resulted: Etna and Santa Cruz Mountains → Old World Interior; Nahe and Wachau → The Moderates; Sonoma → New World Reinvention.

Region	Metaphor	Change	Rationale
Bordeaux	*Business*	D3: +1 → +2	Deeper institutional tradition anchor; separates from Outward Ease
Champagne	*Society*	D3: +1 → +2	Matches Bordeaux reasoning; tradition as dominant institutional force
Etna	*Awakening*	D1: −1 → +1, D2: +1 → +2, D3: −1 → 0	Volcanic introspection is interior; extreme conditions justify +2 struggle; D3 balanced. Cluster shift: → Old World Interior
Goriška Brda	*Fortune*	Metaphor: Serendipity → Fortune, D3: −1 → 0	Resolves duplicate metaphor with Santa Barbara; accidental fortune, not reinvention
Rioja	*Patience*	Metaphor: Memory → Patience	Resolves duplicate with Campania; Patience captures extended aging philosophy
Santa Barbara	*Serendipity*	D3: −1 → 0	Serendipity is discovery, not disruption; balanced on tradition/reinvention axis
Santa Cruz Mtns	*Obsession*	D2: +2 → +1	Obsessive private vision, not extreme endurance. Cluster shift: → Old World Interior
Sonoma	*Authenticity*	D1: 0 → −1, D3: 0 → −1	Deliberate informality masking serious intent; moderately outward-facing. Cluster shift: → New World Reinvention
Stellenbosch	*Aspiration*	D4: 0 → +1	Individual professional ambition; anchors in New World Reinvention
Tuscany	*Art*	Metaphor: Heritage → Art, D2: 0 → −1, D3: +1 → 0	Art better captures Renaissance cultural frame; moderate ease and balanced tradition
Wachau	*Monumentality*	D1: +1 → 0, D2: +1 → 0	Balanced, composed grandeur — not extreme struggle. Cluster shift: → The Moderates
Wagram	*Earth*	D1: +1 → +2, D3: +1 → +2, D6: +2 → +1	Deeply earthly and grounded; strong custodial tradition of the loess landscape
Walla Walla	*Community*	D1: 0 → +1	Place-anchored community identity; inward-facing despite collaborative outreach

4. Terroir Cluster Membership (k=7, Fully Model-Derived)

Terroir clusters are produced entirely by the pipeline: Layer 2 TF-IDF → PCA (10 components) → K-means (k=7). Cluster count, membership, and naming are determined by the algorithm with no manual override. Top TF-IDF terms indicate the vocabulary features that most distinguish each cluster.

Iberian Continental (6)

Douro, Catalonia, Galicia, Ribera del Duero, Rioja, Mendoza

Southern Hemisphere & International (12)

Alto Adige, Patagonia, Barossa Valley, Hunter Valley, Margaret River, Maipo Valley, Central Otago, Hawke's Bay, Marlborough, Stellenbosch, Swartland, Napa Valley

American West Coast (7)

Columbia Valley, Paso Robles, Santa Barbara, Santa Cruz Mountains, Sonoma, Walla Walla, Willamette Valley

French Viticultural (10)

Alsace, Beaujolais, Bordeaux, Burgundy, Champagne, Châteauneuf-du-Pape, Jura, Loire, Northern Rhône, Provence

Germanic Rhine (5)

Baden, Mosel, Nahe, Pfalz, Rheingau

Austrian Danube (3)

Kamptal, Wachau, Wagram

Mediterranean & Volcanic (16)

Steiermark, Dalmatian Coast, Macedonia, Santorini, Tokaj, Campania, Etna, Friuli-Venezia Giulia, Liguria, Piedmont, Sardinia, Sicily, Tuscany, Veneto, Goriška Brda, Finger Lakes

5. D-Score Dimension Definitions

Each region is scored on six dimensions using integer values from –2 to +2. Scores are assigned by reading the Layer 1 identity description independently, without reference to previous pass scores or to Layer 2 terroir data.

D1 — Interiority ↔ Exteriority

+2: Interior, place-defined, inward-looking. The region's identity is inseparable from its specific place.

–2: Outward-facing, social, market-oriented. The region projects outward and builds institutional identity.

D2 — Struggle ↔ Ease

+2: Extreme difficulty, endurance, physical hardship. The region's character is forged through adversity.

–2: Pleasure, comfort, ease. The region's character is associated with generosity and accessibility.

D3 — Tradition ↔ Reinvention

+2: Deep custodial tradition. The region is defined by continuity, preservation, and inherited practice.

–2: Radical reinvention, disruption. The region is defined by deliberate construction of new identity.

D4 — Individual ↔ Collective

+2: Strongly individual, solitary. Individual winemaker vision dominates regional identity.

–2: Strongly collective, communal. Regional identity is a shared, institutional enterprise.

D5 — Urgency ↔ Timelessness

+2: Urgent, present-focused. The region's energy is contemporary, forward-looking, proving itself now.

–2: Timeless, eternal. The region’s identity exists outside contemporary urgency.

D6 — Earthly ↔ Transcendent

+2: Deeply grounded, earthly. Identity is rooted in physical soil, landscape, and material reality.

–2: Transcendent, spiritual. Identity reaches beyond the material toward the philosophical or sublime.

6. Reproducibility

The complete analysis pipeline is available at analysis/pipeline.py in the project repository. Running the script reproduces all clustering solutions, PCA coordinates, and statistical tests. Required packages: scikit-learn, numpy, scipy, pandas. The pipeline reads Layer 2 descriptions from CANONICAL 8 (source document) and D-scores from the embedded REGIONS array. All random states are fixed (random_state=42) for deterministic output.

Pipeline outputs (to data/ directory): identity_pca.json (PCA coordinates for identity scatter plot), terroir_pca.json (PCA coordinates for terroir scatter plot), terroir_clusters.json (terroir cluster assignments), pipeline_report.txt (full report).