Supplemental Data for:

Puga G and Anderson K. 2023. Concentrations and similarities across countries in the mix of winegrape cultivars. Am J Enol Vitic 74:0740018.
DOI: 10.5344/ajev.2023.22067

Supplemental Table 1 Area, cultivar similarity index (CSI) relative to the world, cultivar concentration index (CCI) for each country in 2000 and 2016, and cluster classification based on both CSIs relative to the world and CCIs in 2016.

Country	Area 2000 (ha)	Area 2016 (ha)	CSI 2000	CSI 2016	CCI 2000	CCI 2016	Cluster 2016[a]
Algeria	30200	8300	0.42	0.45	18.5	23.3	Green
Argentina	197418	206342	0.28	0.39	7.6	9.1	Orange
Armenia	11206	14705	0.09	0.02	7.8	58.0	Blue
Australia	130602	132435	0.48	0.67	12.0	15.6	Green
Austria	48496	45439	0.10	0.14	16.2	13.8	Orange
Brazil	52840	33205	0.09	0.07	14.7	17.7	Orange
Bulgaria	95997	52974	0.34	0.58	10.5	11.5	Green
Cambodia		10		0.50		26.0	Green
Canada	8498	12600	0.39	0.42	5.9	7.5	Green
Chile	113966	145873	0.46	0.68	14.3	12.7	Green
China		178000		0.65		6.8	Green
Croatia	59448	11746	0.12	0.21	11.3	20.3	Orange
Cyprus	18282	5133	0.02	0.01	38.3	52.9	Blue
Czechia	11331	13600	0.16	0.24	7.7	7.8	Orange
Ethiopia		169		0.12		39.3	Blue
France	864846	814882	0.67	0.76	6.6	5.7	Green
Georgia	37419	48000	0.10	0.08	32.2	32.2	Blue
Germany	104233	94501	0.12	0.19	11.0	10.0	Orange
Greece	50915	50845	0.06	0.17	9.4	8.8	Orange
Hungary	86886	63881	0.22	0.32	2.5	4.3	Orange
India		2700		0.30		22.4	Orange
Israel	4851	5000	0.46	0.62	9.6	11.4	Green
Italy	636662	604551	0.37	0.46	3.2	3.3	Green
Japan		3869		0.14		9.0	Orange
Kazakhstan		6938		0.09		28.2	Blue
Korea, Rep.	5400	5400	0.01	0.00	31.7	31.7	Blue
Lebanon		4000		0.71		16.2	Green
Luxembourg	1348	1300	0.09	0.14	19.0	14.8	Orange
Mexico		5465		0.51		7.2	Green
Moldova	89844	82600	0.37	0.52	10.0	6.7	Green
Morocco	49600	17590	0.09	0.28	13.2	10.2	Orange
Myanmar		70		0.40		27.1	Orange
New Zealand	9942	35463	0.36	0.31	16.5	37.3	Blue
North Macedonia		24777		0.15		22.6	Orange
Norway		13		0.00		45.0	Blue
Peru		3831		0.04		20.4	Orange
Portugal	205003	182649	0.09	0.28	2.0	4.0	Orange
Romania	222173	182762	0.33	0.42	1.7	2.0	Green
Russia	56332	50794	0.17	0.55	6.6	8.0	Green
Serbia	68999	22014	0.14	0.61	28.3	4.0	Green
Slovakia	15580	7748	0.18	0.16	12.6	9.8	Orange
Slovenia	23472	15989	0.29	0.38	3.8	5.1	Orange
South Africa	93656	95775	0.34	0.53	10.6	9.7	Green
Spain	1181806	883558	0.70	0.58	13.5	11.8	Green
Switzerland	15042	14793	0.14	0.25	24.4	16.6	Orange

Continued on next page.

Supplemental Data for:
Puga G and Anderson K. 2023. Concentrations and similarities across countries in the mix of winegrape cultivars. Am J Enol Vitic 74:0740018.
DOI: 10.5344/ajev.2023.22067

Supplemental Table 1 *continued* Area, cultivar similarity index (CSI) relative to the world, cultivar concentration index (CCI) for each country in 2000 and 2016, and cluster classification based on both CSIs relative to the world and CCIs in 2016.

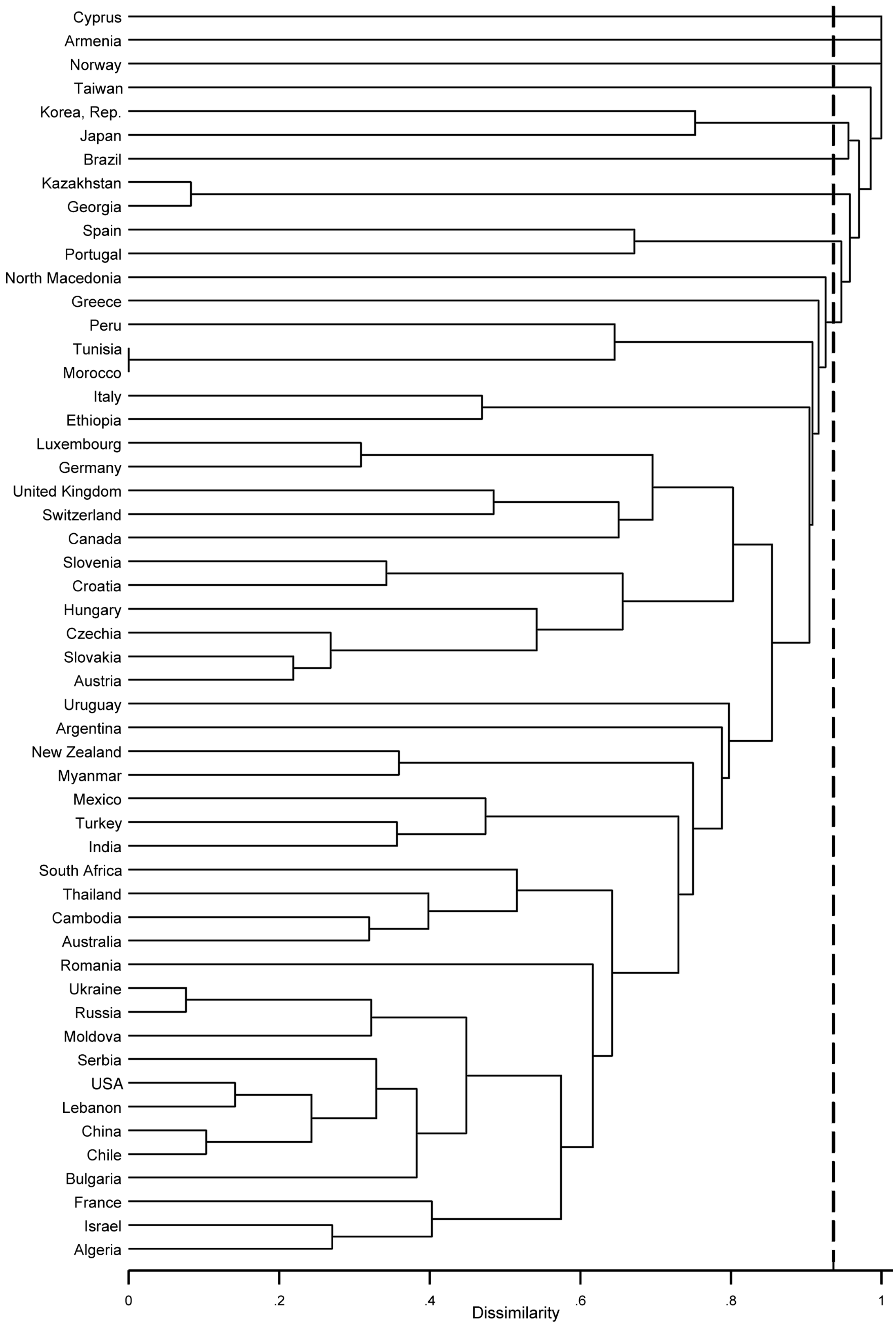
Country	Area 2000 (ha)	Area 2016 (ha)	CSI 2000	CSI 2016	CCI 2000	CCI 2016	Cluster 2016[a]
Taiwan	2833	149	0.01	0.00	39.6	51.3	Blue
Thailand		208		0.27		21.8	Orange
Tunisia	16836	3400	0.31	0.28	22.6	10.2	Orange
Turkey		13704		0.24		8.7	Orange
Ukraine		25166		0.46		14.7	Green
United Kingdom	873	1839	0.07	0.32	7.0	19.3	Orange
USA	175693	239632	0.48	0.72	9.2	8.9	Green
Uruguay	8880	6743	0.17	0.33	20.3	13.2	Orange
WORLD average[b]	90706	84587		0.33		17.28	
WORLD total[b]	4807408	4483130				2.23	
Correlation with area[b]			0.64	0.44	-0.25	-0.30	

[a]The green, orange, and blue clusters include countries with: high CSIs relative to the world and low CCIs, low CSIs relative to the world and low CCIs, and low CSIs relative to the world and high CCIs, respectively.
[b]WORLD average is a simple average (i.e., not area weighted). Both WORLD average and WORLD total refer to the 53 countries in the data set, which include the 20 countries with the largest winegrape growing areas. Correlation with area is the correlation coefficient between the indices (either CSIs or CCIs) and the areas in the respective year, based on the information for all the countries for which there is area data for that year.

DOI: 10.5344/ajev.2023.22067

Supplemental Table 2 and **Supplemental Table 3** are available online at ajevonline.org.

Supplemental Data for:
Puga G and Anderson K. 2023. Concentrations and similarities across countries in the mix of winegrape cultivars. Am J Enol Vitic 74:0740018.
DOI: 10.5344/ajev.2023.22067



Supplemental Figure 1 Dendrogram based on winegrape cultivar similarity indices (CSIs), 2016. Dissimilarity = 1 - CSI. The clustering method is average linkage. The dashed vertical line shows a nine-cluster classification based on CSIs between countries.

DOI: 10.5344/ajev.2023.22067

Supplemental Data:

Concentrations and Similarities across Countries in the Mix of Winegrape Cultivars

More detailed explanation of the indices used in this study

Cultivar similarity index. For analyzing similarities in the mix of winegrape cultivars between two countries, we use the cultivar similarity index (CSI). This index was first introduced by Anderson (2010) and it is also known as the regional similarity index. The CSI for countries *i* and *j* takes the form:

$$CSI_{ij} = \frac{\sum_{c=1}^{C} f_{i,c} f_{j,c}}{\left(\sum_{c=1}^{C} f_{i,c}^2\right)^{1/2} \left(\sum_{c=1}^{C} f_{j,c}^2\right)^{1/2}}.$$

This equation uses data on the bearing area for all cultivars available in both countries. $f_{i,c}$ ($f_{j,c}$) is the area of cultivar c in country i (j), as a proportion of the total winegrape bearing area in that country. We also use this equation for calculating the CSI between each country and the world as a whole with data for 2000 and 2016.

The CSI ranges between 0 and 1, and it is higher when the mix of winegrape cultivars between two countries is more similar. An index of 0 represents a completely different mix of winegrape cultivars, while an index of 1 means that both countries have exactly the same cultivars and the same proportional area for each of those cultivars.

Cultivar concentration index. For analyzing concentrations in the mix of winegrape cultivars, we use a novel index that we call the cultivar concentration index (CCI), given by:

$$CCI_i = 100\left(\sum_{c=1}^{C} f_{i,c}^2\right).$$

The same formula has been used in other disciplines: the Herfindahl-Hirschman index for analyzing concentration in economics, the Simpson index in ecology (Simpson 1949), the Hunter-Gaston index in microbiology (Hunter and Gaston 1988), and the effective number of parties index in politics (Laakso and Taagepera 1979).

The interpretation of the CCI is that if two different vineyard blocks are randomly chosen anywhere in a country, the probability in percentage of those vineyard blocks having the same cultivar is equal to the value of the index. We computed this index for each country and for the world as a whole with data for 2000 and 2016.

Comment on the implication of minor not-reported cultivars on the indices computation. Not all cultivars are reported in Anderson and Nelgen (2020a, 2020b). Some countries provide a list of 'other' cultivars that are not separately identified. These 'other' cultivars, which may differ in number between 2000 and 2016, are not accounted for in the indices' computation. Additionally, new cultivars are occasionally discovered and reported (Pastore et al. 2020).

How important is this limitation? The data in Anderson and Nelgen (2020a, 2020b) is reported at a great level of detail. While more cultivars have been reported in 2016, most of these newly reported cultivars are minor in terms of area coverage. The formulas for both the CSI and the CCI give little weight to the least-planted cultivars. For illustrating this, we calculated the CCI for the world in 2016 using only the top 150 cultivars in terms of area, i.e., less than 10% of all the cultivars in that year. At two decimal points, the CCI is the same (2.23) whether we use the top 150 cultivars or all cultivars. Therefore, we argue that this first limitation is not quantitatively important in our study, but it may be more relevant in other studies in which the number and relevance of reported cultivars changes considerably across countries or regions, or between time periods.

More detailed explanation of the cluster analysis methods used in this study

Hierarchical cluster analysis based on CSIs. We compute a matrix of CSIs for all the countries for which there are winegrape growing area data in 2016. We then transform this matrix into a dissimilarity matrix in which the dissimilarity index between two countries *i* and *j* is 1 - CSI_{ij}. With this dissimilarity matrix, we clustered the countries using an average-linkage hierarchical clustering method.

Hierarchical clustering starts with all countries assigned to N separate groups, each group containing one country. The two countries with the highest CSI (lowest dissimilarity index) are merged into one group, leading to N - 1 groups. The closest two groups are then merged so that the total number of groups becomes N - 2. This process continues until all countries are merged into one single group of size N. Average-linkage clustering determines the closest two groups based on the average dissimilarity between countries in the two groups, and gives equal weight to each country.

We use this cluster method to classify the countries based on their CSIs in 2016. Theoretically, it is possible to classify the countries into up to N clusters by choosing a dissimilarity level as the threshold. For choosing the number of clusters, we rely on the Calinski-Harabasz and the Duba-Hart stopping rules. The Calinski-Harabasz stopping rule provides a pseudo F-index. Higher pseudo-F indices indicate more distinct clustering. The Duba-Hart stopping rule gives a Je(2)/Je(1) index and a pseudo-T^2 value. Higher Je(2)/Je(1) and lower pseudo-T^2 values point out more distinct cluster solutions.

K-means cluster analysis based on the CSIs *relative to the world and the* CCIs. We use both the CSIs relative to the world and the CCIs to cluster the countries using a k-means method with data from 2016. We have used the Minkowski distance metric with argument 2 (Euclidean distance) for comparing the observations (countries) based on two variables (CSI relative to the world and CCI). The process starts with all countries randomly assigned to the (k) number of groups. The mean for each group is calculated based on the Euclidean distance between countries, and each country is re-assigned to the group with the closest mean. This process repeats until no country changes groups. Since the Duba-Hart stopping rule only applies to hierarchical clustering methods and k-means is a partition method, we rely on the Calinski-Harabasz stopping rule to choose the (k) number of clusters.

Stata codes for the cluster analysis performed in this study

Code for hierarchical cluster analysis based on CSIs. We use Stata 17 to perform this analysis. Its corresponding code is as follows:

```
**# CSIs 2016 dendrogram and classification

*This code is to perform an average linkage cluster analysis of all countries using data from 2016 to create Supplemental Figure 1.


***Perform cluster analysis:

clustermat averagelinkage D, add name(alink)

*D is the dissimilarity matrix (using the data provided in the Supplemental Data).


***Use stopping rules to determine the appropriate number of clusters:

cluster stop alink, rule(calinski) groups(3/30) variables(v*)

cluster stop alink, rule(duda) groups(3/30) variables(v*)

*The Calinski-Harabasz stopping rule provides a pseudo F-index. Higher pseudo-F indices indicate more distinct clustering.

*The Duba-Hart stopping rule gives a Je(2)/Je(1) index and a pseudo-T2 value. Higher Je(2)/Je(1) and lower pseudo-T2 values point out more distinct cluster solutions.

*Therefore, the results of these stopping rules suggest that nine cluster is the most distinct cluster solution.
```

***Generate nine clusters:

cluster generate g9 = group(9)

***Order and sort observations in the data set:

order g9, after(Countryofplanting)

sort g9 Country

***Generate graph (same as Supplemental Figure 1):

cluster dendrogram alink, horizontal labels(Countryofplanting) ylabel(, angle(0) labsize(*.45)) xtitle("Dissimilarity", size(11.5pt)) xlabel(, labsize(*.4)) color(black) lwidth(vvthin) graphregion(color(white)) xsize(8.25) ysize(11.75) xline(.936, lwidth(1pt) lcolor(black) lpattern(dash)) title("")

Code for k-means cluster analysis based on the CSIs relative to the world and the CCIs. We use Stata 17 to perform this analysis. Its corresponding code is as follows:

**# K-means cluster analysis using 2016 CSIs and CCI

*This code is to perform a k-means cluster analysis of all countries using data for 2016, and to create Supplemental Figure 2.

***Standardize variables

generate CSI2016z = std(CSI2016)

generate CCI2016z = std(CCI2016)

*The new variables are the standardized CSIs and CCIs (using the data provided in the Supplemental Data).

***Perform cluster analysis:

cluster kmeans CSI2016z CCI2016z, k(2) name(CS2) s(kr(1234))

cluster kmeans CSI2016z CCI2016z, k(3) name(CS3) s(kr(1234))

cluster kmeans CSI2016z CCI2016z, k(4) name(CS4) s(kr(1234))

cluster kmeans CSI2016z CCI2016z, k(5) name(CS5) s(kr(1234))

cluster kmeans CSI2016z CCI2016z, k(6) name(CS6) s(kr(1234))

cluster kmeans CSI2016z CCI2016z, k(7) name(CS7) s(kr(1234))

cluster kmeans CSI2016z CCI2016z, k(8) name(CS8) s(kr(1234))

cluster kmeans CSI2016z CCI2016z, k(9) name(CS9) s(kr(1234))

cluster kmeans CSI2016z CCI2016z, k(10) name(CS10) s(kr(1234))

cluster kmeans CSI2016z CCI2016z, k(11) name(CS11) s(kr(1234))

cluster kmeans CSI2016z CCI2016z, k(12) name(CS12) s(kr(1234))

*1234 is a seed for replicability.

***Use stopping rule to determine appropiate number of clusters:

cluster stop CS2

cluster stop CS3

cluster stop CS4

cluster stop CS5

cluster stop CS6

cluster stop CS7

cluster stop CS8

cluster stop CS9

cluster stop CS10

cluster stop CS11

cluster stop CS12

*The Calinski-Harabasz stopping rule provides a pseudo F-index. Higher pseudo-F indices indicate more distinct clustering.

*Therefore, the results of these stopping rules suggest that three clusters is the most distinct cluster solution.

***Order and sort observations in the data set:

order Country CS3

sort CS3 Area

***Generate graph (same as Supplemental Figure 2):

graph twoway (scatter CSI2016 CCI2016 if CS3==1, color(orange_red) msize(medlarge) msymbol(square)) (scatter CSI2016 CCI2016 if CS3==2, color(green) msize(medlarge)) (scatter CSI2016 CCI2016 if CS3==3, color(blue) msize(medlarge) msymbol(triangle)), aspect(0.425) graphregion(color(white)) legend(off) xtitle("CCI 2016", size(11.5pt)) ytitle("CSI 2016", size(11.5pt))

References

Anderson K. 2010. Varietal intensities and similarities of the world's wine regions. J Wine Econ 5:270-309. DOI: 10.1017/S193143610000095X

Anderson K and Nelgen S. 2020a. Database of Regional, National, and Global Winegrape Bearing Areas by Variety, 2000, 2010, and 2016. Wine Economics Research Centre, University of Adelaide, Adelaide, Australia.

Anderson K and Nelgen S. 2020b. Which Winegrape Varieties are Grown Where? A Global Empirical Picture (Revised Edition). University of Adelaide Press, Adelaide, Australia.

Hunter PR and Gaston MA. 1988. Numerical index of the discriminatory ability of typing systems: An application of Simpson's index of diversity. J Clin Microbiol 26:2465-2466. DOI: 10.1128/jcm.26.11.2465-2466.1988

Laakso M and Taagepera R. 1979. "Effective" number of parties: A measure with application to West Europe. Comp Polit Stud 12:3-27. DOI: 10.1177/001041407901200101

Pastore C, Fontana M, Raimondi S, Ruffa P, Filippetti I and Schneider A. 2020. Genetic characterization of grapevine varieties from Emilia-Romagna (Northern Italy) discloses unexplored genetic resources. Am J Enol Vitic 71:334-343. DOI: 10.5344/ajev.2020.19076

Simpson EH. 1949. Measurement of diversity. Nature 163:688-688. DOI: 10.1038/163688a0