Journal of Wine Economics (2024), 1–25
doi:10.1017/jwe.2024.27

CAMBRIDGE
UNIVERSITY PRESS

https://doi.org/10.1017/jwe.2024.27

ARTICLE

Wine ratings and commercial reality

Gianni De Nicoló

Carey Business School, Johns Hopkins University, Baltimore, MD, USA
Email: gdenico1@jhu.edu

Abstract

Is the quality of a 91-point wine significantly different from that of an 89-point wine? Which wines are underpriced relative to their evaluation of quality? This paper addresses these questions by constructing a novel wine rating system based on scores assigned by a panel of wine experts to a set of wines. Wines are classified in ranked disjoint quality equivalence classes using measures of statistically significant and commercially relevant score differences. The rating system is applied to the "Judgment of Paris" wine competition, to data of Bordeaux *en-primeur* expert scores and prices, and to expert scores and price categories of a large database of Italian wines. The proposed wine rating system provides an informative assessment of wine quality for producers and consumers and a flexible rating methodology for commercial applications.

Keywords: wine ratings; ANOVA; quantile regressions; finite mixture models; hedonic price regressions

JEL classifications Q13; L15; C14

> *"Count what is countable, measure what is measurable, and what is not measurable, make it measurable." Galileo Galilei (1564–1642).*

I. Introduction

Wine quality ratings based on numerical scores assigned by a panel of wine experts to a set of wines are ubiquitously reported in wine magazines and wine guides, are prominent on the shelves of wine stores, and are used to grant quality "medals" to wines in wine competitions. These ratings are important for producers, offering them market visibility and directing their production decisions about wine style, as well as for consumers, orienting their purchasing decisions. The assessment of the value and the role of these ratings for producers and consumers, and their relationship with wine prices, are important ongoing areas of research in wine economics (Storchmann, 2012).

Wine ratings pose two important and commercially relevant questions. Is a 90-point wine significantly different from an 89-point wine? As observed by Gergaud et al. (2021), rating boundaries determining differences in wine quality ratings can

https://doi.org/10.1017/jwe.2024.27

be critical for wine marketing. Currently, a wine with a score equal to or greater than the 90-point threshold is marketed as a high-quality wine deserving special exposure in the press and on the wine store shelves. And, which wines are underpriced relative to their assessment of quality? The classifications of wine "bargains" or "top performers" in the popular press reviewed by Miller et al. (2015), and the results of how wine ratings can affect prices for wine rated above certain thresholds illustrated by Carlson et al. (2023), can significantly affect consumers' perceptions, with relevant impact on sales. This paper addresses these questions by constructing a wine rating *system* that aims at providing a statistical and commercially relevant basis for the determination and comparisons of wine ratings and their relationship with prices.

Rating methodologies used in many areas, such as finance (see e.g. FitchRatings, 2022), environmental standards (see e.g. Morgan Stanley Capital International, 2022), and quality ratings of consumer goods (see e.g. www.consumereports.org), are based on a mapping of statistics of measurable characteristics of the objects rated—the probability of defaults of debt securities, the environmental impact of specific gas emissions, the records of repairs of consumer goods—into *ranked disjoint quality equivalence classes*, which are typically labeled with alphabetically ordered letters or visual indicators of ranks. Although rating methodologies may be very similar, ratings of specific items issued by different rating organizations may differ depending on the type of databases used, as well as on the weight assigned to the set of quality factors whose aggregation determines the overall rating of an item. The proposed rating system adapts a standard rating methodology to the multifaceted dimensions of wine quality as captured by tasting protocols in professional evaluations, as reflected in the "wine scorecards" used in various settings reviewed by Jackson (2017). In essence, a wine score by a wine expert is a mapping of his/her quantitatively ordered sensory evaluation and weighting of wine quality factors onto a numerical rating scale. Similarly to the ratings produced in different fields, the rating of the set of wines by a panel of experts will depend on the size of the wine sample, the number of experts, and how experts' scores are aggregated.

Reported wine evaluations differ depending on the implicit or explicit wine scorecard that experts use, which determines the design of a rating system. In an *incomplete information* rating system, a single wine score is recorded with no information on the evaluation of the quality factors underlying that score. In this case, the evaluation of each of the quality factors underlying the wine score is not observable.[1] In a *complete information* rating system, quality factors are scored individually and aggregated into a final score according to predetermined weights. The wine scorecard used by the International Organization of Vine and Wine (OIV, 2022) for sponsored wine competitions is an example of this format: the card requires experts to score 10 quality factors. A complete information rating system based on data about the "why(s)" a particular score

[1] Wine publications often report scores of experts that are supplemented by tasting notes. Typically, longer tasting notes are associated with higher scores. However, the specific information content of these notes about the assessment of specific quality factors is difficult to ascertain on a comparative basis, as most of these notes are focused on sensory descriptors. For a review of the information content of wine-tasting notes and an evaluation of the price impact of descriptors, see Capehart (2021). On the potential role of tasting notes in wine marketing aided by AI technology, see Carlson et al. (2023).

https://doi.org/10.1017/jwe.2024.27

is assigned to a wine is undoubtedly more informative than an incomplete information setup. A scoring template that requires an expert to explicitly assess quality factors numerically is generally considered a standard of professional tasting procedures by enologists. However, most statistical evaluations of wine tastings in the literature have been carried out using an incomplete information setup due to data availability. As the data of our applications are all incomplete information databases, in this paper, we focus on an incomplete wine rating system.

The proposed rating system builds on standard methods used in food science,[2] originally applied to wine evaluations by the pioneering contribution of Amerine and Roessler, (1983). In the context of these standard methods, we introduce the following novel assumptions regarding the standardization of expert scores, the definition and partitioning of data in ranked disjoint quality rating classes, and the identification of the price–quality rating component.

A. Scores standardization

Experts deliver a wine score in a specific numerical range. For instance, Wine Spectator magazine uses a 70–100 range, Decanter magazine uses a 50–100 range, and www.jancisrobinson.com uses the 12–20 range. Wine scores may differ according to experts' different experiences, sensory capacities, and, most importantly, different weights assigned to the perceived wine quality factors that are not observable. This is reflected in different ways the same rating scale is used by each expert, as observed by Cardebat and Paroissien (2015). The heterogeneity of views of a panel of experts who use implicitly the same wine scorecard is informative in providing an evaluation of wine quality from different perspectives. The differences in wine evaluations by Robert Parker and Jancis Robinson often mentioned in the literature illustrate this point.

To ensure the comparability of expert scores, the aggregation of their scores requires a standardization that reflects this heterogeneity while preserving each expert ranking of a wine. In the proposed rating system, rank-preserving standardization is simply implemented using the location and scale of the distribution of wine scores of each expert, measuring their evaluation with a *Z-score*.

B. Rating classes

Wines are classified and ranked in *disjoint quality equivalence classes* using the mean of standardized expert scores. Based on a standard analysis of variance (ANOVA), we compute a measure of *Minimum Significant Difference* (MSD) at a standard level of confidence for a set of rated wines. *Indivisible MSD units* provide a "numeraire," or "currency," to convert a standardized wine score in units of MSD, called the QV of a wine. The QV is computed as the integer ratio of a wine standardized score to the MSD. QVs automatically place wines in ranked disjoint equivalent quality classes, since wines with the same assessed quality have the same QV. The ratings of a set of wines are then delivered as a set of alphabetically ordered rating classes (e.g. A, B, C, ..., and so on). As detailed in our applications, the design of the system allows users to calibrate the number of rating classes to desired commercial objectives.

[2] An overview of the methods in food science is in Lawless and Heymann (2010). A review of the methods applied to wine are reviewed in Jackson (2020) and Lesschaeve and Noble (2022).

https://doi.org/10.1017/jwe.2024.27

Computing QVs is feasible when the scores of each wine by a set of experts are available. Yet, in many wine publications and wine guides, only a single score of a wine is reported with no information about the scores of individual experts. In this case, the construction of rating classes can be implemented by the estimation of a finite mixture model (FMM) using a given distribution of wine scores. The identification of ranked disjoint equivalent quality classes is based on the estimated posterior predicted probabilities that a wine score belongs to a particular rating class. These predicted posterior probabilities are the "Quality Value" counterparts derived from the entire distribution of wine scores.

C. Price–quality rating component

Given the availability of individual prices for a set of rated wines, the rating system can be used to identify wine underpricing within each rating class. The relationship between price and quality conditional on a wine being rated in a given rating class is obtained by estimating hedonic price *quantile* regressions (Koenker, 2005), where price is a function of the identified rating classes and other controls. These quantile regressions are estimated for a quantile lower than the median, whose level is chosen according to the desired stringency of the criterion defining "underpricing." An underpriced wine in each rating class is a wine whose price is lower than the predicted price at the chosen quantile.

The inclusion of underpricing in the rating system can be also implemented when individual wine prices are not available, but wines are classified in price ranges, or price "points": in this case, the identification of underpriced wines is simply obtained from the joint empirical distribution of wines' price ranges and rating classes using empirical quantiles within each rating class.

Therefore, the rating system can include underpricing information by expanding rating classes into subcategories. Similarly to the determination of rating classes, the design of these subcategories allows users to calibrate the degree of underpricing according to desired commercial objectives, as detailed in our applications.

D. Applications and plan of the paper

The proposed rating system is applied to three examples of commercially important wine ratings: the 1976 "Judgment of Paris" wine competition, a sample of 2021 ratings of Bordeaux *en-primeur* wines, and a large database of ratings and price categories of Italian wines published online for subscribers by the National Association of Wine Tasters ONAV (Organizzazione Nazionale Assaggiatori Vino) in 2022. The dataset of the "Judgment of Paris" wine competition includes wine scores by the panel of experts, but prices are not available. The Bordeaux *en-primeur* dataset includes both expert scores and prices. The ONAV dataset includes aggregate scores by panels of experts, as well as price *ranges* of a large set of Italian wines, but scores of the experts composing the panels and individual wine prices are not available.

Three desirable properties characterize the proposed rating system, as illustrated in these applications. First, ratings are obtained by a standard statistical procedure that embeds an "economic" evaluation of the quality values (QVs) of wines, delivering ranked disjoint quality equivalent classes. Second, the system can easily incorporate sub-ratings related to a price–quality relationship that takes into account both wine characteristics and rating classes. Third, the flexibility of the system allows potential users to calibrate the parameters that define the set of rating classes and the incorporation of price–quality sub-ratings according to desired commercial objectives. Wine rating reports based on the proposed rating system may provide a more transparent and informative assessment of wine quality for producers and consumers than current methods.

The remainder of the paper is composed of five sections and an Appendix. Section 2 details the rating methodology. Sections 3–5 implement the rating system using the three datasets described above. Section 6 concludes. The Appendix reports additional data tables referenced in the text.

II. Methods

The tasting panel is composed of N experts indexed by $i \in \{1, 2, 3, \dots, N\}$ who evaluate M wines indexed by $j \in \{1, 2, 3, \dots, M\}$ on a numerical rating scale defined on the positive real line. The score assigned by expert i to wine j is denoted by X_{ij}.

A. Scores standardization

As noted, experts may use a rating scale differently according to the (unobservable) weights assigned to different wines' quality factors. For example, experts may assign different weights to sensory quality factors such as concentration, balance, persistence, or harmony, which are common factors requiring a specific evaluation in most wine scorecards.

To make experts' evaluations comparable, we standardize the raw scores of each expert with respect to the location and scale of his/her score distribution. To this end, we use a standard *Z-score*, given by $Z_{ij} = \left(X_{ij} - \mu_i\right)\sigma_i^{-1}$, where $\mu_i = M^{-1}\sum\limits_j X_{ij}$ is the mean and $\sigma_i = \sqrt{M^{-1}\sum\limits_j \left(X_{ij} - \mu_i\right)^2}$ is the standard deviation of expert i's wine evaluation. Under this standardization, the distribution of standardized scores of each expert has the same location and scale, i.e. a zero mean and a unit variance. Different evaluations of a wine by an expert will then reflect different quality evaluations relative to the expert's own set of weights assigned to the (unobservable) quality factors. Note that any standardization of experts' scores must be *rank-preserving* to consistently reflect their preference ordering. This condition is automatically satisfied for Z_{ij} since the Z-score is a linear function of the raw score X_{ij}.[3] To work with positive standardized scores, and with no change of any of the results that follow, a second standardization is implemented with respect to the location and scale parameters of the overall distribution of standardized scores of wines of the panel of experts, denoted by μ_P and

[3] The standardization based on a transformation on the entire cumulative distribution of each expert score relative to that of one expert proposed by Cardebat and Paroissien (2015), and applied by Gergaud et al. (2021) and the Global Wine Score website www.globalwinescore.com, is *not* rank-preserving due to the pervasive presence of sets of wines ranked with the same score (ties).

https://doi.org/10.1017/jwe.2024.27

σ_P respectively. This standardization can be useful to compare differences in scores of any expert relative to the overall distribution of the scores of the panel if so desired. Under this (double) standardization, the score of expert i of wine j, denoted by $Z(P)_{ij}$, satisfies

$$\frac{Z(P)_{ij} - \mu_P}{\sigma_P} = \frac{X_{ij} - \mu_i}{\sigma_i} \equiv Z_{ij} \;\Rightarrow Z(P)_{ij} = \mu_P + \sigma_P Z_{ij} \tag{1}$$

B. ANOVA

The score of a wine by a panel of experts is the *mean* of experts' standardized scores, given by

$$Z(P)_j = N^{-1} \sum_{i=1}^{N} Z(P)_{ij} \tag{2}$$

There has been a debate in the literature on whether ranks or averages are the most appropriate statistics to aggregate experts' wine scores. Quandt (2006) advocated ranks, although he recognized that there is a loss of information about perceived differences in the quality of wines, as ranks can be the same across experts, but the value of their evaluations can be very different. In their detailed analysis of voting and grading systems, Balinski and Laraki, (2007, 2010) discussed how different rules of aggregation of scores and rankings proposed in the wine literature may deliver different and often inconsistent results. They proposed to use the *median* to reflect "majority judgment." Specifically, they showed that the median is the correct statistic of an aggregation function of experts' scores that is consistent with a set of basic set of preference axioms. A key assumption in their framework is that experts share a "common language," which we associate with experts sharing a "wine scorecard." Note that if the distribution of experts' scores is approximately normal (hence, approximately symmetric), then the mean and the median are approximately equal, implying that the use of average scores is consistent with "majority judgment."

The reliability of tests of mean differences using F-tests based on ANOVA rests on the assumptions of equality of variances across experts' scores and approximate normality of the distribution of the relevant regression errors. Equality of variance across experts' scores is guaranteed by the *Z-score* standardization of Equation 1. The normality assumption needs to be tested. We test the normality of the residuals associated with the ANOVA regression using Bera et al. (2016) test, which exhibits good power for small samples and is detailed in our first application. If normality is not rejected, we compute the relevant F-test from the ANOVA at a 5% significance level. If normality is rejected, we use a *robust* ANOVA, implemented by computing a modified F-test using trimmed means and winsorized variances. In this case, the assessment of significant mean differences is based on the Yean statistics, as detailed in Wilcox (2022).

C. Rating classes as disjoint quality equivalent classes

A standard measure of statistical difference of the mean score of a pair of wines is given by the MSD at a given statistical significance level, typically chosen to be equal to 5%. The MSD is determined by the distribution of the test statistics under the null

https://doi.org/10.1017/jwe.2024.27

hypothesis of no difference in means. A wine rating system partitions the set of wines evaluated by a set of experts into *disjoint* equivalent quality classes. If class A is labeled as superior to class B, all wines in class A have scores statistically and significantly higher than the scores of all wines in class B.

Let's illustrate the role of the MSD in our rating system with simple examples. If wine 1's score is significantly higher than wine 2's score, then $Z(P)_1 - Z(P)_2\ MSD$. A wine equivalence class can be defined as the label assigned to the set of wines whose scores are not statistically significantly different. Yet, the computation and use of the MSD is necessary, but not sufficient, to fulfill the commercial need to classify wines in ranked disjoint equivalent quality classes, since wines with close numerical scores may not be assigned to such classes.

The common practical solution is to classify wines in different commercially significant categories based on partitions of raw scores in numerical quality categories treated as *absolute* ranks of quality.[4] For example, 90-point wines are classified strictly better than 89-point wines. If the value of the MSD is greater than 1, a 90-point wine is not significantly different from an 89-point wine. As long as some scores are viewed as a threshold of quality, such as a 90-point score, whether or not a 90-point score or an 89-point score represents significantly different quality levels may be highly relevant commercially.

More complications arise when we consider multiple mean comparisons. For example, let the scores of wines A, B, and C be 91, 90, and 89 respectively. If the MSD is less than 1, these wines are in disjoint quality equivalent classes. Using preference ordering notation, $A \succ B \succ C$. If the MSD is greater than 2, then these wines are in the same equivalence class, that is, $A \sim B \sim C$. If the MSD is 1.5, however, wine A is better than C ($A \succ C$) but equivalent to B ($A \sim B$), while wine B is equivalent to C ($B \sim C$). We can "separate" A from C, but we are unable to place B in one or the other class. Several examples of this situation are reported in Amerine and Roessler (1983), and arise in all our datasets as well. From a statistical viewpoint these comparisons are perfectly reasonable and informative, but they are not useful commercially. Paraphrasing the rating categories of some wine competitions, any wine receiving a "gold" medal should be classified as *strictly* better than any wine receiving a "silver" medal. In other words, quality equivalent classes must be *disjoint*.

Statistical significance and commercial relevance can be reconciled as follows. Consider an estimated MSD as an *indivisible unit of account* of quality. The QV of a wine can be defined by

$$QV\left(Z(P)_j\right) = int\left(\frac{Z(P)_j}{MSD}\right) \tag{3}$$

where the *int* operator truncates any fraction, transforming the value of wine in an integer number. In other words, $QV\left(Z(P)_j\right)$ simply transforms the original standardized score of wine j into indivisible MSD units. These units can be viewed as the

[4] Typical examples of these classifications are the rating tables used by magazines such as Wine Spectator or Decanter, where quality categories are associated with number ranges of a given scale anchored by descriptions indicating progressively higher quality as number ranges increase.

https://doi.org/10.1017/jwe.2024.27

"currency" employed in valuing a specific set of wines rated by a specific set of experts. The computation of the QVs automatically delivers the ranking of wines in disjoint quality equivalent classes and the relevant rating distribution. A standard measure of the MSD is obtained by computing the relevant F-tests from the ANOVA and the associated *Fisher Least Significant Difference* (FLSD) at a 5% significance level. We treat the FLSD as a benchmark, since it is the most liberal test of pairwise comparisons, being based on a Type I error rate that assumes individual pairwise comparisons.[5] If normality is rejected, we use a *robust* ANOVA, implemented on trimmed means and winsorized variances, using the Yean statistics to obtain the corresponding *robust* FLSD. Having obtained the FLSD, we parameterize the MSD as $MSD\left(k\right) = kFLSD$, where $k \geqslant 1$. The MSD is thus calibrated as a multiple of the FLSD. By varying k, we can determine a desired number of ranked disjoint quality equivalent classes depending on commercial objectives. Therefore, a (calibrated) wine QV is computed as

$$QV\left(Z(P)_j, k\right) = int\left(\frac{Z(P)_j}{kFLSD}\right) \tag{4}$$

The value $k = 1$ is the benchmark ($MSD\left(1\right) = FLSD$), since it delivers the maximum number of disjoint quality equivalent classes. If some $k1$ is chosen, the QV of a wine is expressed in terms of indivisible *multiples* of FLSD units. In this case, the number of ranked disjoint equivalent quality classes is typically reduced as k is increased.

In sum, given the FLSD and a choice of k, the computation of wines' QVs determines the rating classes of the wine sample, which can be denoted by alphabetically ordered capital letters or other descriptors conveying a scale of different quality levels.

D. Rating underpriced wines

Núñez et al., (2024) review the extensive literature on hedonic linear regressions, which have been widely used in assessing the wine price–quality relationship. Few applications, such as Amédée-Manesme et al. (2020) and Castriota et al. (2022), have also used hedonic quantile regressions with the aim of identifying possible nonlinearities between wine price, wine characteristics, and expert scores.

We use hedonic quantile regressions to identify underpriced wines within each rating class as follows. Suppose the rating classes of a set of wines are labeled in decreasing quality order as $\{A, B, C, D, \ldots\}$. Denote with P_j the price of wine j, with Y_j a vector of wine characteristics, and with I_R a set of indicator functions classifying wines in

[5] Pairwise comparisons of means following statistically significant F-tests are used to detect which particular means in a group are significantly different. When multiple independent tests are conducted, each test has an inherent Type I error rate α, but the overall *family-wise* Type I error rate accounting for all the $\left(n - 1\right)/2$ comparisons is equal to $1 - \left(1 - \alpha\right)^n$, where n is the number of comparisons. The Fisher LSD is the most liberal, as it does not control for the family-wise error rate. In applications, the Fisher LSD is considered "protected" from underestimation of the Type I error rate by an ANOVA F-test resulting in a very small p-value. For a review of multiple pairwise mean comparisons following ANOVA, see Sauder and DeMars (2019).

each rating class $R \in \{A, B, C, D, ..\}$. We estimate the following hedonic price *quantile* regression:

$$P_j = \alpha_q + Y_j\beta_q + \sum_R \gamma_q^R I_R + I_j^q \tag{5}$$

where the value of q is set to a value strictly lower than the median ($q = 0.5$). The quantile regression estimates the predicted quantile q of the price of wine j in rating class R. Denote with $\widehat{P}_j^q\,(R)$ the predicted quantile q of the price of wine j in rating class R. We identify an underpriced wine as a wine for which the observed price is lower than the estimated price at quantile q, i.e. $P_j\widehat{P}_j^q\,(R)$. The rating of this wine is then set equal to $R+$. In other words, for each identified rating class, we define a sub-rating class marked by a "+," which includes wines that are underpriced within their quality level.

To sum up, the choice of q determines the magnitude of estimated underpricing, with lower levels of $q = 0.5$ indexing higher degrees of underpricing. For reporting purposes, the choice of q will ultimately be determined by commercial considerations.

III. The "Judgment of Paris" wine competition

The 1976 Paris Wine Tasting was organized by Steven Spurrier, owner of a wine shop, and Patricia Gallagher, a manager of a wine school. In this event, nine French experts evaluated in blind tastings a set of top-quality white and red wines from France and California. In each of the tastings, a California wine ranked first. Taber (2005) illustrates the mechanics of the tasting and vividly describes the significant marketing impact of the competition in the international wine world: for the first time, New World wines were ranked as superior to top-quality Bordeaux and Burgundy wines by French experts. Experts used a (0–20) numerical scale, assigning separate points to four quality factors: eye, nose, mouth, and harmony. Unfortunately, the scores of each quality factor have not been made available.

Several statistical analyses of the results of this competition have been carried out, although most studies have erroneously included the scores of Steven Spurrier and Patricia Gallagher that were not included in the total count of wine scores (see Taber (2005), p. 202). Ashenfelter and Quandt (1999) and Quandt (2006) analyzed the results of the red wine competition converting scores into ranks, and discussed various methods of aggregations of expert scores and an evaluation of the rank correlations among experts as a measure of "consensus." Cicchetti (2006) focused on the degree of agreement of experts in both the red and white wine competitions, pointing out the possibility of different outcomes if the panel was split according to some measure of experts' "consistency." Hulkower (2009) applied Borda method of ranking, while Balinski and Laraki (2013) used their proposed "majority judgment" method: both studies reported results different from those originally publicized since French red wines were found to rank first. More recently, Gergaud et al. (2021) have reviewed how the rankings of red wines would have changed using different ranking procedures. To the best of our knowledge, virtually all of the analyses of this competition have not examined systematically whether differences in total wine scores or ranks are statistically significant. What would have been the results of this competition if our proposed

https://doi.org/10.1017/jwe.2024.27

Table 1. Judgment of Paris: Wines ranked by standardized mean score.

White Wines	Mean		Red Wines	Mean	
a (U.S.)	15.29		A (U.S.)	13.98	
b (FR)	14.64		B (F)	13.94	
c (U.S.)	13.57		C (F)	13.86	
h (FR)	11.78		D (F)	13.81	
d (U.S.)	11.76		E (U.S.)	12.21	
e (U.S.)	10.94		F (F)	10.90	
f (FR)	10.88		I (U.S.)	10.26	
I (U.S.)	10.07		G (U.S.)	9.99	
g (FR)	10.06		H (U.S.)	9.90	
j (U.S.)	4.56		J (U.S.)	9.26	
	Average	Rank sum		Average	Rank sum
U.S. wines	11.03	5.50		10.93	6.67
French wines	11.84	5.50		13.13	3.75

rating system had been applied using the scores of the nine French judges? And what would have been the event's commercial impact?

The Appendix reports basic information and statistics of this wine competition. Table A.1 lists white and red wines, which include famous Burgundy and Bordeaux wines respectively. Table A.2 lists the experts, generally considered the apex of professional wine expertise in France at that time. Tables A.3 (white wines) and A.4 (red wines) report the original scores, the standardized scores according to Equation 1, statistics for each wine and each expert, and aggregate scores both including and excluding the scores of the experts whose tally was not included in the final scores.[6]

Table 1 summarizes the ranking of white and red wines by the standardized mean scores.

Note that a U.S. wine is first in the rank in both the white and red wine groups. As measured by the average of scores and rank sums by country provenance, U.S. white wines performed similarly to French white wines: this was a notable feat, given the top quality of French Burgundy wines. However, in the red wine category, U.S. wines performed worse than French wines on average, except the first red California wine. However, the difference between the standardized mean score of the first-ranked U.S. red wine and the second-ranked French red wine is minuscule.

Recall that experts arrived at their total wine score by rating wines according to four quality factors. An indirect gauge of how differences in quality assessment among experts might have affected their final score can be obtained by estimating a simple factor model. As shown in Table 2, the results of a standard estimation of common factors used by experts indicate that about 91% and 88% of variations of scores for

[6] Perhaps unsurprisingly, a comparison of the aggregate scores for both white and red wines reveals fairly different values of means and standard deviations for most wine scores when experts number 4 and 8 are not included in the total count.

https://doi.org/10.1017/jwe.2024.27

Table 2. Judgment of Paris: Factor analysis.

White wines				
Factor	Eigenvalue	Difference	Proportion	Cumulative
Factor1	5.50137	3.748	0.6113	0.6113
Factor2	1.75337	0.77336	0.1948	0.8061
Factor3	0.98001	0.57931	0.1089	0.915
Red wines				
Factor	Eigenvalue	Difference	Proportion	Cumulative
Factor1	4.41576	2.00116	0.4906	0.4906
Factor2	2.41459	1.35419	0.2683	0.7589
Factor3	1.06041	0.44555	0.1178	0.8768

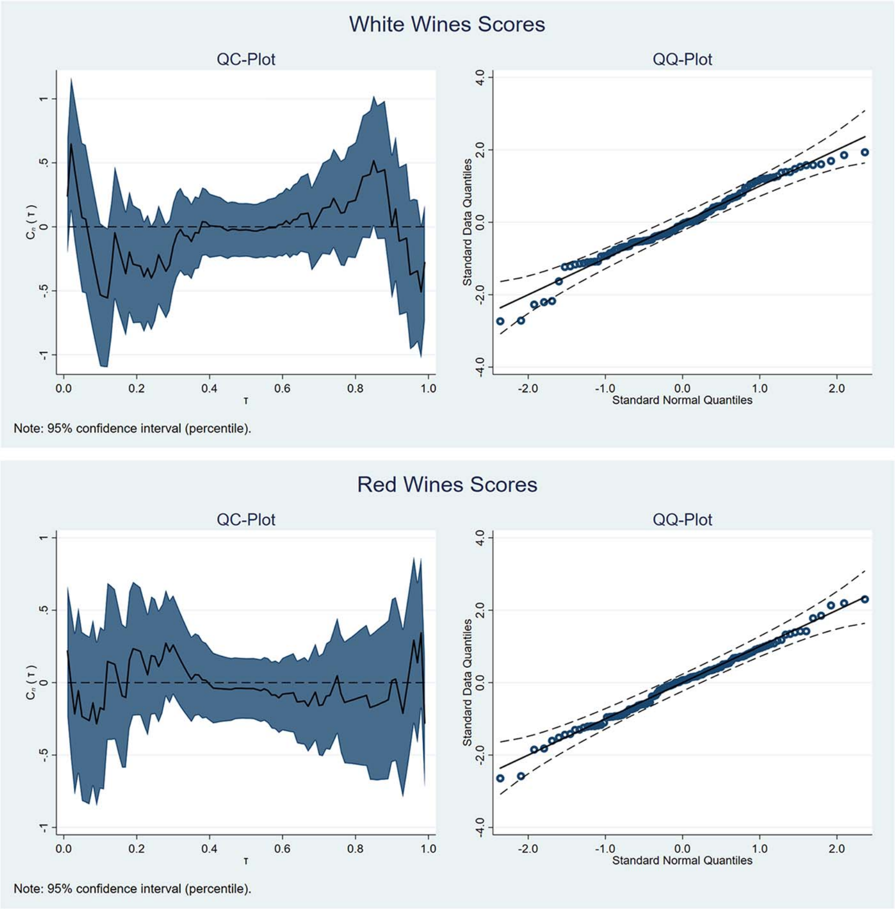
white and red wines respectively is spanned by three common factors. This suggests that experts likely assigned different weights to at least three of the four quality factors composing the rating scale. The results of this simple factor analysis appear consistent with a three-level partition of quality factors.[7]

As previously mentioned, the use of standard F-tests to compute MSDs rests on the assumption of normality of the conditional distribution of scores across experts. As a formal test, we use the statistical procedure introduced by Bera et al. (2016). They show that normality can be assessed based on the asymptotic Quantile-Covariance (QC) function, defined as the ratio of the expected quantile loss function over the density function evaluated at each quantile. Bera et al. (2016) show that the QC function is constant *if and only if* the underlying distribution is normal, and show that this property can be tested using standard Kolmogorov-type statistics. Graphically, a QC plot would exhibit approximate normality if it is close to a horizontal line. The results of the test can be also represented by QQ plots inclusive of confidence bands. As shown in Figure 1, the 95% confidence bands of the QC include the horizontal line, and the QQ plots indicate that all score observations are inside the relevant 95% confidence band: thus, the null of normality is not rejected.

Table 3 reports ANOVA tables for two estimates of the MSD: the benchmark FLSD, and the more conservative Fisher-Hayter (Hayter, 1986) MSD measure, denoted by FHLSD, whose test of mean differences is based on the Type I error rate associated with *all* 45 $(N(N-1)/2)$ pairwise comparisons of the $N = 10$ wines in the sample. As expected, the FHLSD is notably greater than the FLSD, indicating that a more stringent criterion used to identify differences among standardized mean scores will generally result in a smaller number of rating classes. Table 4 reports the QVs and the ratings according to the FLSD and the FHLSD. The QV of each wine is computed

[7] The “eye” part of the evaluation might have been very similar across experts since this quality factor is typically assessed to identify potential faults and some features of the typology of a wine, which was known to the judges. Moreover, the “eye” part typically receives the lowest weight in most professional wine quality evaluation protocols. The remaining three quality factors are those where most of the evaluations might have differed.

https://doi.org/10.1017/jwe.2024.27



White Wines Scores

QC-Plot

QQ-Plot

Note: 95% confidence interval (percentile).

Red Wines Scores

QC-Plot

QQ-Plot

Note: 95% confidence interval (percentile).

Figure 1. Judgment of Paris: QC and QQ plots.

according to Equation 2, using as denominator the FLSD and the FHLSD respectively. The resulting rating classes are labeled with alphabetically ordered capitalized letters..

Let's compare the results under the FLSD and FHLSD "price systems." Consider columns (1) and (2) of the white wines panel: the first two wines, one U.S. and one French wine, are worth six FLSDs and are placed in the A rating class; the third wine, a U.S. wine, is worth five FLSDs and is placed in the B rating class; all wines ranked from forth to nine (three U.S. and three French wines) are worth three FLSDs and are placed in the C rating class; the 10th-ranked U.S. wine is worth only one FLSD, and it is placed in the F rating class. According to the FLSD "currency," white wines are ranked in the six classes (A, B, C, D, E, F), where the D and E classes are empty. When we use a more conservative criterion of significant difference, such as the FHLSD, then the number of rating classes shrinks: as shown in columns (3) and (4) of the white wines panel, QVs are smaller and the rating classes are reduced to three, (A, B, C), and within each class the number of U.S. and French wines is the same.

Table 3. Judgment of Paris: ANOVA.

WHITE WINES	Partial SS	*df*	MS	*F*	Prob *F*	FLSD	FHLSD
Model	733.06531	17	43.121489	6.03	0	2.392	5.523
Experts	3.22E−29	8	4.02E−30	0	1		
Wines	733.06531	9	81.451701	11.38	0		
Residual	515.13186	72	7.1546092				
Total	1248.1972	89	14.024687				
RED WINES	Partial SS	*df*	MS	*F*	Prob *F*	FLSD	FHLSD
Model	308.9548	17	18.173812	2.11	0.0153	2.626	6.063
Experts	7.93E−13	8	9.92E−14	0	1		
Wines	308.9548	9	34.328312	3.98	0.0004		
Residual	620.80727	72	8.6223232				
Total	929.76207	89	10.446765				

The results for red wines differ markedly from those of the white wines. Under the FLSD three classes are obtained, (A, B, C): four wines are rated A, with only one U.S. wine in the A class; two wines are rated B, one U.S. and one French; the remaining four wines are rated C and are all U.S. wines. Under the FHLSD, rating classes shrink to two, (A, B): five wines are rated A, with two U.S. wines in the list; the remaining five wines are rated B, with only one French wine in the list. Note that the use of different "price systems" can be useful to assess how sensitive is the placement of certain wines on the boundaries of rating classes to changes in rating classes. We explore the usefulness of variations in "currency" denominations for commercial purposes in the next application. Overall, French red wines performed better than the U.S. red wines, although the performance of the two highest-ranked U.S. red wines was as good as that of the French red wines in the two highest rating classes under the benchmark FLSD.

Summing up, the application of our proposed rating system to this important wine competition would have delivered a more balanced assessment of the quality of the wines involved. California white wines were on average comparable to French white wines, and some of them were in a higher rating class than some white French wines. By contrast, French red wines were in higher rating classes than California red wines, except the Stags' Leap Winery Cabernet, whose ranking determined its rise to fame and the main marketing punch of this wine competition. Interestingly, while the evidence for California white wines being better or equivalent to French wines was compelling, the marketing galore was mostly focused on red wines.

IV. Bordeaux *en-primeur*

The commercial importance of pricing for Bordeaux wines in primary and secondary markets is stressed and analyzed by Masset et al. (2023), who review the literature on the "efficiency" of Bordeaux wine pricing. The Bordoverview.com website contains

https://doi.org/10.1017/jwe.2024.27

https://doi.org/10.1017/jwe.2024.27

Table 4. Judgment of Paris: QVs and ratings.

		QV	Rating	QV	Rating
	Standardized	FLSD	FHLSD	FHLSD	FHLSD
Wines	Score	(1)	(2)	(3)	(4)
WHITE WINES					
a (U.S.)	15.29	6	A	2	A
b (FR)	14.64	6	A	2	A
c (U.S.)	13.57	5	B	2	A
h (FR)	11.78	4	C	2	A
d (U.S.)	11.76	4	C	2	A
e (U.S.)	10.94	4	C	1	B
f (FR)	10.88	4	C	1	B
i (U.S.)	10.07	4	C	1	B
g (FR)	10.06	4	C	1	B
j (U.S.)	4.56	1	F	0	C
RED WINES					
A (U.S.)	13.98	5	A	2	A
B (FR)	13.94	5	A	2	A
C (FR)	13.86	5	A	2	A
D (FR)	13.81	5	A	2	A
E (U.S.)	12.21	4	B	2	A
F (FR)	10.90	4	B	1	B
I (U.S.)	10.26	3	C	1	B
G (U.S.)	9.99	3	C	1	B
H (U.S.)	9.90	3	C	1	B
J (U.S.)	9.26	3	C	1	B

ratings and prices for a large set of Bordeaux *en-primeur* wines. The sample of Left Bank red wines in 2021 includes 170 wines rated by five experts: William Kelley for Wine Advocate (WA), Jeff Leve (JL), Jane Anson for Decanter Magazine (JA), Chris Kissack (CK), and Jancis Robinson (JR). All experts rated wines on a 75–100 scale except JR, who used a 10–20 scale, which we converted into a 75–100 scale by a simple linear transformation. Table A.4 in the Appendix reports the distribution of the rated 170 wines by Appellation d'Origine Contrôlée (AOC) and classification.[8]

Table 5 reports statistics of the scores of each wine by the experts, the relevant Spearman rank correlation matrix of experts' evaluations, and the results of a standard estimation of common factors and factor loadings. Note that not all experts rated each

[8] Classifications are: the 1855 Médoc Grand Cru Classé Classification (1st–5th GCC), with a total of 61 wines in five subcategories; the Médoc Cru Bourgeois Classification (CB), with a total of 56 wines; and the Graves classification (CC Graves), with 11 wines.

https://doi.org/10.1017/jwe.2024.27

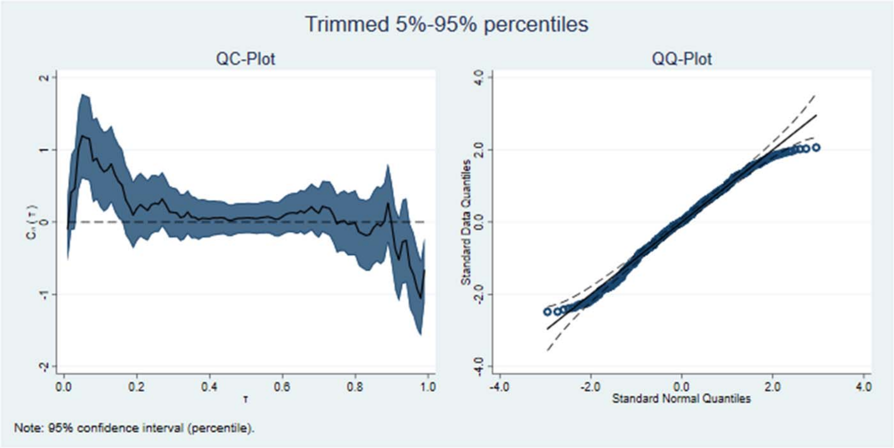
Table 5. Left bank wine scores, factors, and factor loadings.

Expert	Rated wines	Mean	Std. dev	Min	Max
WA	141	90.56	2.80	84.50	97.00
JL	149	90.86	2.66	84.00	97.00
JA	157	90.41	2.98	82.00	97.00
CK	147	88.88	2.77	81.00	95.00
JR	109	81.40	3.34	72.50	90.00
Total scores	703				
Spearman rank Correlation	WA	JL	JA	CK	JR
WA	1				
JL	0.8419	1			
JA	0.8601	0.8194	1		
CK	0.7481	0.7922	0.7756	1	
JR	0.6751	0.6294	0.6441	0.6907	1
Factor	Eigenvalue	Difference	Proportion	Cumulative	
Factor1	4.0730	3.6595	0.8146	0.8146	
Factor2	0.4134	0.1893	0.0827	0.8973	
Factor3	0.2242	0.0544	0.0448	0.9421	
Factor loadings					
WA	0.9245	−0.1284	0.2557		
JL	0.9288	−0.2364	0.0181		
JA	0.9220	−0.1658	0.0203		
CK	0.9121	0.0356	−0.3855		
JR	0.8208	0.5588	0.0971		

wine in the sample. The mean and standard deviation of the scores of experts are very similar except JR, where differences may be in part due to the smaller number of the wines JR evaluated. The rank correlation of scores among experts is fairly high. The estimation of common factors used by the experts indicates that about 94% of the variation of scores is spanned by three common factors. Looking at the experts' factor loadings, the magnitude of the loadings of the first factor is very similar across experts, while those of the second and third factors differ across experts, likely capturing different weights assigned to the latent quality factors used in experts' wine evaluations.

The scores of each expert were standardized relative to the panel's location and scale measures according to Equation 2. Bera et al. (2016) tests of normality rejects the null a 5% confidence level. By trimming the distribution of scores excluding scores below the 5% percentile and above the 95% percentiles (a total of 35 scores), and winsorizing variances, the modified F-tests based on the Yean statistics do not reject the null of normality at a 5% confidence level. This result is visually depicted by the QC and QQ plots of Figure 2, where both the QC and QQ plots exhibit confidence bands consistent with approximate normality.

https://doi.org/10.1017/jwe.2024.27



Note: 95% confidence interval (percentile).

Figure 2. Left Bank sample: Bera et al. (2016) **QC and QQ plots.**

The "robust" ANOVA table for this sample (not reported) delivers a value of the "robust" FLSD equal to 2.63. Recall that we can calibrate the MSD as a multiple of the FLSD, that is, $MSD\,(k) = kFLSD$, since the FLSD is the most liberal statistical criterion. By increasing k, the number of rating classes declines. The choice of $k \geqslant 1$ thus determines the desired number of classes according to the desired commercial objectives if the number of rating classes of the benchmark ($k = 1$) is deemed not optimal. As shown next, evaluating how rating classes change with different values of k can determine how sensitive is the rating of wines at the boundaries of the rating classes and the corresponding thresholds of the rating classes in terms of the numerical scale used. As observed earlier, the determination of rating class boundaries as reported in terms of the rating scale of reference have important commercial implications.

Table 6 reports statistics of standardized scores in each rating class for all $k \in \{1, 1.5, 2\}$ obtained by computing the corresponding QVs. The minimum and maximum of each rating class are the rating class boundaries. Is a 90-point wine strictly better than an 89-point wine? Under the FLSD "currency" ($MSD\,(1)$), these two wines are rated D, since the D class includes 43 wines with scores in the interval (89.5–92.0). Therefore, their quality is rated equally. Under the $MSD\,(1.5)$ "currency," a 90-point wine belongs to rating class B, which includes 26 wines with scores in the interval (90.7–94.4), whereas an 89-point wine belongs to rating class C, which includes 67 wines with scores in the interval (86.8–90.2). Thus, under the $MSD\,(1.5)$ "currency," these two wines belong to different rating classes. This latter comparison also holds under the $MSD\,(2)$ "currency."

Summarizing, rating boundaries will crucially depend on the "evaluation style" of experts that compose the tasting panel since their evaluations will result in different "price systems," leading to different distributions of QVs that determine the rating classes. From a commercial perspective, any issuer of ratings from expert panels adopting this rating system (a wine publication, or a rating website, for example), will

Table 6. Left Bank sample: Ratings with MSD(k) = kFLSD, for k = 1, 1.5, 2.

	No. of wines	Mean score	Minimum	Maximum	Rank	Rating
MSD(1) = 2.63	2	97.7	97.7	97.7	1	A
	12	96.1	94.8	97.1	2	B
	10	93.9	92.4	94.4	3	C
	43	90.5	89.5	92.0	4	D
	40	88.2	86.8	89.3	5	E
	42	85.4	84.2	86.8	6	F
	16	83.0	81.8	84.1	7	G
	5	80.9	80.4	81.4	8	H
MSD(1.5) = 3.95	14	96.3	94.8	97.7	1	A
	26	92.3	90.7	94.4	2	B
	67	88.9	86.8	90.2	3	C
	52	85.0	82.9	86.8	4	D
	11	81.6	80.4	82.8	5	E
MSD(2) = 5.26	14	96.3	94.8	97.7	1	A
	53	91.1	89.5	94.4	2	B
	82	86.8	84.2	89.3	3	C
	21	82.5	80.4	84.1	4	D

have to determine the relevant "price system" that fulfills some commercial objectives. Disclosures of these choices may strengthen the reliability and the reputation of the ratings, enhancing their desired marketing impact.

A. Price–quality rating

Bordoverview.com reports the "average initial consumer price "en-primeur" in euros with tax included" for a subsample of wines, which are suggested to be used as a general guideline. To illustrate how price information can be embedded in the rating system, we consider the rating classes corresponding to the $MSD\,(1.5)$ value. We first estimate the missing prices with a standard hedonic model, and then derive the "+" rating subdivisions via quantile regressions.

To obtain the missing prices, we estimated the following hedonic regression:

$$\log P_j = \alpha + X_j\beta + \sum_R \gamma^R I_R + I_j \tag{6}$$

where the vector X_j includes indicator variables that index AOC, classification, and size of production, and I_R is an indicator variable equal to 1 if wine j belongs to rating class $R \in \{A, B, C, D, E\}$. The estimated coefficients are used to estimate the missing prices in the sample. As expected, this distribution of wine prices exhibits high right-skewness due to very expensive wine belonging to the historical 1855 Bordeaux classification.

The quantile regressions are specified as in Equation 5, and they are estimated for $q = 0.10$ and $q = 0.25$. Table 7 reports the results for both the standard hedonic regression and the quantile regression. The explanatory power of these regressions is fairly high, ranging from an R2 of 0.87 for the standard hedonic regression, to values of the Pseudo-R2 of 0.67 and 0.71 for the $q = 0.10$ and the $q = 0.25$ quantile regressions respectively. The p-values of the coefficients associated with AOC and Classification variables are all significant. Moreover, the coefficients of quality ranking are significant and indicate a positive qualit—price relationship.

The ratings incorporating price information are reported in Table 8, which shows the mean, minimum, and maximum prices in each rating class. Interestingly, price ranges overlap in the B and C rating classes, suggesting the existence of some overpricing in each of the two classes, likely due to the reputation effects of the classifications. As expected, the number of underpriced wines in each rating class that deserve a "+" increases with the chosen q.

To summarize, using a sample which includes price information, we have illustrated the usefulness of using a rating system based on statistically significant and commercially relevant criteria to provide information about wine quality. This information is conditional on both the sample of rated wines and the number of experts involved in the evaluation. The commercial importance of setting the MSD parameters and the choice of quantile levels has been stressed. As in rating systems used in many other commercial fields, transparent information to the public about these choices may enhance the trustworthiness and reputation of the rating system in providing information about the quality of wines for producers and consumers.

V. The ONAV wine sample

The ONAV database contains a large number of Italian wines rated on a (75–100) scale and classified by standard typologies (still white wines, still red wines, sparkling wines, etc.), vintage year, and denomination of origins.[9] We selected from the database a sample of 2,485 wines, composed of 986 still white wines and 1,499 still red wines.

Table 9 shows the distribution of wine scores, vintage year, classification, and price ranges of the selected sample. First note that wine scores are in the subset of the reference rating scale, ranging from a minimum of 83 to a maximum of 96. Second, wine scores are obtained by teams of experts who follow a wine evaluation template similar to the one used by OIV (2022), but we do not have information on the scores assigned

[9] We consider four denominations: IGT, IGP, DOC/DOP, and DOCG. IGT (Indicazione Geografica Tipica) and IGP (Indicazione Geografica Protetta) are both geographical classifications. The IGT differs from IGP due to fewer restrictions for the bottling, labeling and production of grapes encompassing a very large area of production. The IGP classification is more restrictive, since the wine must be created or transformed in the production area indicated by the specification. DOC and DOP classifications are equivalent (DOC was earlier established in Italy, and subsequently incorporated in DOP according to the European-wide wine classification,The DOCG (Denominazione di Origine Controllata and Garantita observe more production restrictions summarized by the "guaranteed" term.

https://doi.org/10.1017/jwe.2024.27

Table 7. Left Bank wines: Hedonic linear and quantile regressions.

Linear Regression			Quantile Regressions			
N	123		*N*	170	170	
R^2	0.8724		Pseudo-R^2	0.6764	0.7103	
Adj. R^2	0.8458		Quantile	0.10	0.25	
	Coef.	*p*-value	Coef.	*p*-value	Coef.	*p*-value
Production Size	0.002	0.29	−0.001	0.33	0.000	0.80
AOC						
2	−0.047	0.83	0.039	0.82	0.004	0.98
3	0.283	0.04	0.332	0.00	0.328	0.00
4	0.434	0.00	0.335	0.00	0.404	0.00
5	0.038	0.80	0.215	0.07	0.170	0.14
6	0.347	0.06	−0.047	0.75	0.322	0.03
7	0.430	0.04	0.536	0.01	0.520	0.01
8	−0.134	0.52	−0.117	0.34	−0.104	0.39
9	0.313	0.06	0.347	0.01	0.393	0.00
Classification						
2	−0.916	0.00	−1.386	0.00	−1.231	0.00
3	−1.091	0.00	−1.533	0.00	−1.481	0.00
4	−1.048	0.00	−1.537	0.00	−1.246	0.00
5	−1.157	0.00	−1.562	0.00	−1.350	0.00
6	−1.388	0.00	−1.928	0.00	−1.675	0.00
7	−1.013	0.00	−1.381	0.00	−1.538	0.00
8	−0.779	0.00	−1.656	0.00	−1.227	0.00
9	−1.435	0.00	−1.901	0.00	−1.700	0.00
Rankings						
B	−0.906	0.00	−0.945	0.00	−0.836	0.00
C	−1.436	0.00	−1.264	0.00	−1.374	0.00
D	−1.733	0.00	−1.510	0.00	−1.608	0.00
E	−1.867	0.00	−1.541	0.00	−1.673	0.00
Constant	5.924	0.00	6.164	0.00	6.041	0.00

to each of the quality factors by the teams, and how they were aggregated. Hence, our rating system is applied directly to the distribution of reported wine scores. Moreover, prices of each wine are not reported individually but are just placed in six price ranges. About 95% of the wines have prices not greater than 40 euros, indicating that the bulk of the sample includes wines from low to medium–high price points. Despite the unavailability of individual expert scores, we can construct a "star" rating system identifying ranked disjoint equivalent classes by using a simple version of a Finite Mixture Model (FMM). Such a model allows to separate observations in subpopulations using

https://doi.org/10.1017/jwe.2024.27

Table 8. Left Bank wines: Prices, ratings, and +ratings, MSD(1.5).

RATINGS Rating class	No. of wines	Mean	Min	Max	No. of wines		No. of wines	
		Price			$Pq = 0.1$	$q = 0.1$ price	$Pq = 0.25$	$q = 0.25$ price
A	13	330	114	684	2	153	4	186
B	20	93	45	242	2	51	10	73
C	54	46	13	233	13	29	15	32
D	32	25	13	51	8	19	12	22
E	4	17	14	20	0	13	1	15

estimates of the latent distributions that compose the mixture distribution of the wine scores.[10]

Denote with y_j the score of wine j and with $\mathbf{y}$ the N-dimensional vector of wine scores. The density $f\,(\mathbf{y})$ of $\mathbf{y}$ is assumed to come from R distinct classes of densities $f_1, f_2, .., f_R$ in proportions $\pi_1, \pi_2, \dots, \pi_R$. A general specification of an R-component FMM conditional on a linear model of a vector of X covariates is given by:

$$f\,(\mathbf{y}) = \sum_{i=1}^{R} \pi_i f_i\left(\mathbf{y}|X^T\beta_i\right) \tag{7}$$

where $\pi_i \in [0, 1]$ is the probability for the ith class, $\sum_{i=1}^{R} \pi_i = 1$, and $f_i\left(\mathbf{y}|X^T\beta\right)$ is the probability density function of $\mathbf{y}$ in the ith class conditional on the vector X, where the T superscript denotes a transpose.

The estimation of the probabilities of each component and the relevant conditional density function is interpreted as arising from the different unobservable weighting of wine quality factors assigned by the experts and summarized by a wine score. Since all wines are evaluated blind, we estimate the model with no covariates (the vector X includes only a constant), under the assumption that a wine score issued by a panel of experts summarizes all relevant information leading to a rating of a wine, including denomination of origin, vintage, and other unreported features of the wines that experts used.

The probabilities of the latent classes π_i are estimated using a multinomial logit, where $\pi_i = \frac{exp(y_i)}{\sum_i exp(y_i)}$. The choice of the family of densities to use for $f_i\,(.)$ depends on the structure of the data. As shown in Figure 3, the log scores appear to be well approximated by a (truncated) normal distribution. Hence, we use log-normal densities for $f_i\,(.)$ in the estimation of the parameters of Equation 5.1 In other words, the density of the vector of wine scores is approximated by a linear mixture of lognormal densities.

[10] For a review of FMM models, see McLachlan et al. (2019). An example of application of finite mixture models to wine data is in Cao (2014), who focused on a model designed to identify common versus random components of tasters' evaluations.

https://doi.org/10.1017/jwe.2024.27

Table 9. ONAV wine sample: Scores, vintages, classification and price ranges.

Score	No. of wines.	Percent	Vintage	No. of wines	Percent
83	272	10.95	2007	6	0.24
84	271	10.91	2008	8	0.32
85	336	13.52	2009	29	1.17
86	293	11.79	2010	42	1.69
87	448	18.03	2011	111	4.47
88	346	13.92	2012	166	6.68
89	174	7	2013	335	13.49
90	182	7.32	2014	277	11.15
91	62	2.49	2015	554	22.3
92	67	2.7	2016	739	29.75
93	20	0.8	2017	217	8.74
94	9	0.36			
95	2	0.08			
96	3	0.12			
Cassification	No. of wines	Percent	Price range (euro)	No. of wines	Percent
IGT	352	14.16	10	753	30.39
IGP	64	2.58	10–20	1,186	47.86
DOC/DOP	1,360	54.73	20–40	432	17.43
DOCG	709	28.53	40–60	79	3.19
			60–80	15	0.61
			80	13	0.52

The likelihood function is computed as the sum of the probability-weighted conditional likelihood from each latent class, and estimation is iterative. The maximum of the predicted posterior probabilities across classes determine the partition of scores in rating classes. A key choice of the estimation procedure is the determination of the number of rating classes. As a baseline, we chose the number of classes as determined by standard AIC and BIC criteria. We computed both the AIC and BIC statistics for a number of classes ranging from 2 to 5, and found that both AIC and BIC statistics are minimized for $R = 4$.

Table 10 reports the results. As shown in the upper panel, wine scores fall into four rating classes in proportions 0.20, 0.21, 0.22, and 0.37 respectively. The mean scores are 83.47 in class 1, 85.29 in class 2, 87.27 in class 3, and 88.83 in class 4. The predicted posterior probabilities, denoted by pp1, pp2, pp3, and pp4, measure the probability of a wine with score x belonging to each of the four classes. As shown in the lower panel, the partition of ratings in terms of score intervals is simply determined by the maximum predicted probability of a wine of a given score to belong to each class. One star is given to wines with scores in the 83–84 range, two stars are given to wines with scores in the 85–86 range, three stars are given to wines with scores in the 87–88 range, and four stars are given to wines with scores greater than or equal to 89.

https://doi.org/10.1017/jwe.2024.27

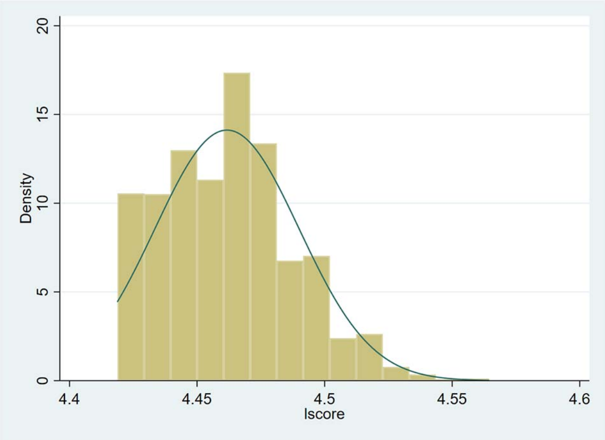


Figure 3. ONAV wine sample: log(score) distribution.

Table 10. ONAV Wine Sample: FMM predicted posterior probabilities and ratings.

Class	π_i	Mean				STARS
		Score				
1	0.20	83.47				*
2	0.21	85.29				**
3	0.22	87.27				***
4	0.37	88.83				****
No. of wines	Score	pp1	pp2	pp3	pp4	
272	83	0.99	0.00	0.00	0.01	
271	84	0.82	0.14	0.00	0.04	*
336	85	0.02	0.89	0.00	0.09	
293	86	0.00	0.60	0.16	0.24	**
448	87	0.00	0.02	0.71	0.27	
346	88	0.00	0.00	0.52	0.48	***
174	89	0.00	0.00	0.05	0.95	
182	90	0.00	0.00	0.00	1.00	
62	91	0.00	0.00	0.00	1.00	****
67	92	0.00	0.00	0.00	1.00	
20	93	0.00	0.00	0.00	1.00	
9	94	0.00	0.00	0.00	1.00	
2	95	0.00	0.00	0.00	1.00	
3	96	0.00	0.00	0.00	1.00	

https://doi.org/10.1017/jwe.2024.27

Table 11. Italian wines sample: Wine distribution by rating and price range.

	10	10–20	20–40	40–60	60–80	80	
Price ranges	1	2	3	4	5	6	Total
STARS							
*	189	258	75	13	3	4	542
*	187	314	113	10	4	1	629
**	284	372	108	17	3	5	789
***	93	242	136	39	5	3	518
Total	753	1,186	432	79	15	13	2,478
Percent of wines in price range							
*	0.25	0.22	0.17	0.16	0.20	0.31	0.22
*	0.25	0.26	0.26	0.13	0.27	0.08	0.25
**	0.38	0.31	0.25	0.22	0.20	0.38	0.32
***	0.12	0.20	0.31	0.49	0.33	0.23	0.21
Percent of wines in rating classes							
*	0.35	0.48	0.14	0.02	0.01	0.01	1
*	0.30	0.50	0.18	0.02	0.01	0.00	1
**	0.36	0.47	0.14	0.02	0.00	0.01	1
***	0.18	0.47	0.26	0.08	0.01	0.01	1

A. Price–quality rating

Turning to the relationship between ratings and prices, Table 11 reports the distribution of wines by rating and price range under the four-star rating systems constructed with the FMM model.

Three main results emerge from this table. First, as shown in the upper panel, there are 93 four-star wines in the lowest price range, suggesting that these wine would be natural candidates of an extra "+" mark indicating underpricing relative to quality. The number of four-stars wine in the next higher price range is also substantial. Second, when we look at the fraction of wines in the lowest rating class by price range, this fraction declines only slowly with the increase in price ranges, suggesting wine overpricing relative to quality for a substantial number of wines. Relative wine overpricing similarly occurs for the intermediate rating categories (two and three stars). For the highest rating class, the fraction of wines increases and then decreases with the price range. Third, a significant fraction of wines of the highest rating classes 3 and 4 are in the lowest and next to the lowest price ranges, suggesting significant underpricing relative to quality. The evidence suggests that for this sample, a positive relationship between price and wine quality might not be as strong as predicted by standard hedonic price equations. This might be due to the pervasive wines' underpricing and overpricing. Yet, our rating methodology is fairly flexible, since the number of rating classes can be increased or reduced in the estimation, and other conditioning covariates might be introduced in the X matrix to obtain finer partitions of rating classes consistent with specific commercial objectives.

The application of our rating system to this sample of Italian wines has shown that useful information about wine quality and its relationship with prices can be extracted from the data even without information about the scores assigned by the panels of experts involved in wine evaluations. The availability of this information, as well as the use of individual prices of wines, would undoubtedly improve the assessment of wine quality and the identification of wine underpricing.

VI. Conclusion

This paper has constructed a wine rating system based on scores assigned by a panel of wine experts to set wines. Standardized expert scores and a novel measure of wine QVs deliver ranked disjoint equivalent quality rating classes, which can be expanded into subcategories using a standard model of the price–quality relationship.

We have applied the system to the 1976 "Judgment of Paris" wine competition using only wine score information, to a sample of Bordeaux *en-primeur* wines and prices, illustrating the incorporation of price information in the ratings, and to a sample of Italian Wines, where the rating was constructed solely based on final scores by teams of experts. The application of the proposed rating system to these datasets shows that the system provides an informative assessment of wine quality for producers and consumers and a flexible template for wine rating reports.

All the datasets we have employed report data of the experts' score of each wine but do not report the separate scores of wine quality factors customarily employed in professional wine tastings. An extension of the proposed rating system adapted to these more detailed data is part of our research agenda.

Supplementary material. The supplementary material for this article can be found at https://doi.org/10.1017/jwe.2024.27.

Acknowledgement(s). I would like to thank an anonymous reviewer, Magalie Dubois, and participants in the 15th Annual Conference of the American Association of Wine Economists (AAWE) in Stellenbosch, South Africa, for their comments and suggestions.

References

Amédée-Manesme, C.-O., Fayeb, B., and Le Furb, E. (2020). Heterogeneity and fine wine prices: Application of the quantile regression approach. *Applied Economics*, 52(26), 2821–2840.

Amerine, M. A., and Roessler, E. B. (1983). *Wines: Their Sensory Evaluation*. W.H Freeman and Company.

Ashenfelter, O., and Quandt, R. (1999). Analyzing a wine tasting statistically. *Chance*, 12(3), 16–20.

Balinski, M., and Laraki, R. (2007). A theory of measuring, electing, and ranking. *Proceedings of the National Academy of Sciences of the United States of America*, 104(21), 8720–8725.

Balinski, M., and Laraki, R. (2010). *Majority Judgment*. The MIT Press.

Balinski, M., and Laraki, R. (2013). How best to rank wines: Majority judgment. In: Giraud-Héraud E and Pichery MC (eds.), *Wine Economics. Applied Econometrics Association Series*. Palgrave Macmillan (pp. 149–172).

Bera, A. K., Galvao, A. F., Wang, L., and Xiao, Z. (2016). A new characterization of the normal distribution and test of normality. *Econometric Theory*, 332(5), 1216–1252.

Capehart, K. W. (2021). Willingness to pay for wine bullshit: Some new estimates. *Journal of Wine Economics*, 16(3), 260–282.

Cardebat, J.-M., and Paroissien, E. (2015). Standardizing expert wine scores: An application for Bordeaux en primeur. *Journal of Wine Economics*, 10(3), 329–348.

Carlson, K., P., Kopalle, A., Riddel, D., Rockwell, and P., Vana, 2023, Complementing human effort in on-line reviews: A deep learning approach to automatic content generation and review synthesis. *International Journal of Research in Marketing*, Vol. 40, 54-74.

Castriota, S., Corsi, S., Frumento, P., and Ruggeri, G. (2022). Does quality pay off? "Superstar" wines and the uncertain price premium across quality grades. *Journal of Wine Economics*, 141–158.

Cicchetti, D. V. (2006). The Paris 1976 tasting revisited once more: Comparing ratings of consistent and inconsistent tasters. *Journal of Wine Economics*, 1(2), 125–140.

FitchRatings. (2022). The rating process: How Fitch assigns credit ratings. February, www.firtchratings.com.

Gergaud, O., Ginsburgh, V., and Moreno-Ternero, J. D. (2021). Wine ratings: Seeking a consensus among tasters via normalization, approval, and aggregation. *Journal of Wine Economics*, 16(3), 321–342.

Hayter, A. (1986). The maximum familywise error rate of the fisher's least significance difference. *Journal of the American Statistical Association*, 81(396), 1000–1004.

Hulkower, N. (2009). The Judgment of Paris according to Borda. *Journal of Wine Research*, 20(3), 171–182.

Jackson, R. S. (2017). *Wine Tasting: A Professional Handbook* (3rd ed.). Academic Press, Elsevier Ltd.

Jackson, R. S. (2020). *Wine Science: Principles and Applications* (5th ed.). Academic Press, Elsevier Ltd.

Koenker, R. (2005). *Quantile Regression*. Cambridge University Press.

Lawless, H. T., and Heymann, H. (2010). *Sensory Evaluation of Food: Principles and Practice* (2nd ed.). Springer.

Lesschaeve, I., and Noble, A. C. (2022). Sensory Analysis of Wine, Chapter 7. In: Andrew GR (ed.), *Managing Wine Quality. Volume I: Viticulture and Wine Quality* (2nd ed., pp. 243–277). Woodhead Publishing, Elsevier.

Masset, P., Weiskopf, J.-P., and Cardebat, J.-M. (2023). Efficient pricing of Bordeaux en primeur wines. *Journal of Wine Economics*, 18, 39–65.

McLachlan, G. J., Lee, S. X., and Rathnayake, S. I. (2019). Finite Mixture Models. *Annual Review of Statistics and Its Applications*, 6, 355–378.

Miller, J. R., Stone, R. W., and Stuen, E. T. (2015). When is a wine a bargain? A comparison of popular and regression-based approaches. *Journal of Wine Research*, 26(2), 153–168.

Morgan Stanley Capital International. (2022). MSCI ESG Ratings. www.msci.com.

Núñez, J., Martín-Barroso, D., and Velázquez, F. J. (2024). The hedonic price model for the wine market: A systematic and comparative review of the literature. *Agricultural Economics*, 55, 247–264.

OIV, 2022, Standard for international wine and spirituous beverages of vitivinicultural origin competitions, www.oiv.int/sites/default.

Quandt, R. E. (2006). Measurement and inference in wine tasting. *Journal of Wine Economics*, 1(1), 7–30.

Sauder, D. C., and DeMars, C. E. (2019). An updated recommendation for multiple comparisons. *Advances in Methods and Practices in Psychological Science*, 2(1), 26–44.

Storchmann, K. (2012). Wine economics. *Journal of Wine Economics*, 7(1), 1–33.

Taber, G. M. (2005). *Judgment of Paris*. Sribner.

Wilcox, R. R. (2022). *Introduction to Robust Estimation and Hypothesis Testing* (5th ed.). Academic Press.

Cite this article: De Nicoló, G. (2024). Wine ratings and commercial reality. *Journal of Wine Economics*, 1–25. https://doi.org/10.1017/jwe.2024.27