


SHORTER PAPERS

Context and cross-section data improve analyses of wine ratings

Jeffrey Bodington 

Bodington & Company, San Francisco, US

Email: jcb@bodingtonandcompany.com

Abstract

Much research shows that the ratings that critics, judges, and consumers assign to wines are heteroscedastic. A rating observed is one draw from a latent distribution that is wine- and judge-specific. Estimating the shape of a rating's distribution by minimizing a sum of cross entropies has been proposed and tested. This article proposes a method of improving the accuracy of that estimate by using information about the context of a wine competition or cross-section ratings data. Tests using the distributions implied by 90 blind triplicate ratings show that the sum of squared errors for the solution using context or cross-section information is 50% more accurate than not using such information and over 99% more accurate than ignoring the uncertainty about a rating.

Keywords: wine; judge; ratings; statistics; random

JEL classifications: A10; C10; C00; C12; D12

I Introduction

A difficulty in wine-ratings-related research, and in calculating consensus among judges, is that each rating is one observation drawn from a latent probability distribution that is judge- and wine-specific. One draw from a latent distribution is a tiny sample size so expert and consumer interpretations of wine ratings, analyses of ratings, predictions of ratings, and calculations of consensus among judges are infused with uncertainty that is difficult to quantify. Bodington (2022a) proposed to quantify that uncertainty using a maximum information entropy estimate of the latent distribution of a wine rating. Tests using ratings assigned to 1,599 blind triplicates assessed by judges at the California State Fair (CSF) Commercial Wine Competition and 30 blind triplicates assessed by judges at a tasting in Stellenbosch reported by Cicchetti (2014) showed that the result is much more accurate than ignoring the uncertainty about a rating.

This article shows that using information about context or cross-section ratings data improves the accuracy of an estimate of the maximum entropy distribution of the rating that a judge assigns to a wine. Tests using the Stellenbosch blind triplicates show that

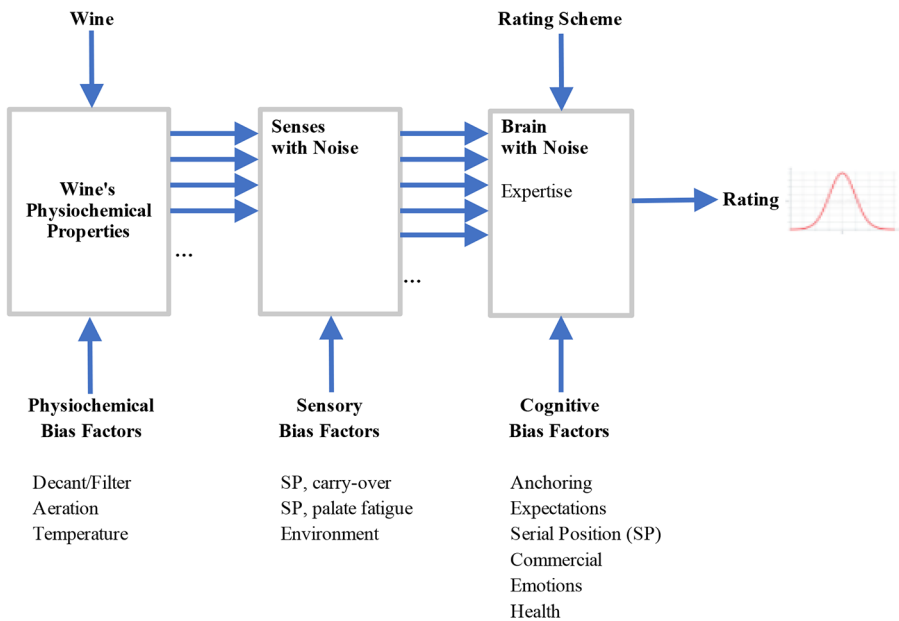


Figure 1. Summary of noise and biases that affect blind wine ratings.

the resulting distribution has the shape, mean, and standard deviation (SD) that are similar to the empirical distribution implied by blind triplicate ratings.

Section II of this short article presents a summary of the literature concerning the random nature of wine ratings and a maximum entropy estimate of the distribution of a rating. Section III presents a modification of that estimate to include information about context or cross-section data. That modification is tested in Section IV, and conclusions follow in Section V. The ratings data and MATLAB code concerning results reported in this article are available on request.

II Background

Although wine ratings are not merely random, evidence that ratings are heteroscedastic is abundant in wine-related academic literature and in the wine-trade press. Bodington (2022b) cites four experiments with blind replicates, three texts, and more than 30 articles showing that wine- and judge-specific heteroscedasticity is caused by the factors in Figure 1. And the uncertain nature of wine ratings is not unique. Kahneman et al. (2021, p. 80–86, 215–258) describe variance in wine ratings and other areas of human judgment including physicians' diagnoses, radiologists' assessments of x-rays, forensic experts' fingerprint identifications, and judges' sentencing of criminals. Recently, Barberá et al. (2023, p. 123) commented that wine rating procedures impose demands on experts that “may result in poor-quality expressions of actual opinions.”

A difficulty with the heteroscedasticity of wine ratings is that analysis of ratings is an inverse and ill posed problem. Inverse because parameters that describe the shape of the latent distribution of a judge's potential ratings on a wine must be inferred from

ratings observed. Ill posed because, except in the rare case of replicates, only one draw from the latent distribution is observed. With a sample size of one, there are not enough observations to statistically identify a unique distribution. Building on information theory developed by Shannon (1948), Jaynes (1957, p. 621–623) proposed that inverse and ill posed problems could be solved by maximizing information entropy subject to constraints that enforce what little may be known about the data. See numerous applications of that notion in Golan et al. (1996).

Bodington (2022a) employed Jaynes' notion to estimate the latent distribution of a rating by minimizing the sum of cross entropies shown in Equations (1) and (2). Any rating (x) can be assigned to wine " i " by judge " j " that is within a bounded set of ordered ratings ($\min \leq x_{ij} \leq \max$). The rating observed, or the set of ratings observed if there are replicates in a flight, is denoted x_{ij}^0 . Cross entropy (I) is the informational difference between two bounded and discrete distributions (estimate $p(x_{ij})$ and target $q(x_{ij})$) defined in Equation (1). The sample-size-weighted (n_{ij}) minimum of the difference between an unknown distribution (\hat{p}) and a uniform random distribution (u), plus the difference between that same unknown distribution \hat{p} and the distribution that is observed ($q|x_{ij}^0$), is defined in Equation (2). Using Stellenbosch blind triplicate data published by Cicchetti (2014) and CSF blind triplicate data provided by Robert Hodgson, Bodington (2024) showed that the solution to Equation (2) has less than 10% of the error made by the common practice of ignoring the uncertainty about a rating.

$$I(p, q) = - \sum_{\min}^{\max} q(x_{ij}) \ln(p(x_{ij})) \quad (1)$$

$$\arg[\hat{p}] = \operatorname{argmin} \left[\left(\frac{1}{1 + n_{ij}} \right) \cdot I(u, \hat{p}) + \left(\frac{n_{ij}}{1 + n_{ij}} \right) \cdot I(q|x_{ij}^0, \hat{p}) \right] \quad (2)$$

The information inputs to Equations (1) and (2) are the rating that a judge is observed to assign to a wine (x_{ij}^0) and the categories in an ordered rating system (\min to \max). But that is too narrow a view of what may be known. This article proposes in the next section to reconsider the random distribution u and then obtain a more precise estimate of the latent distribution of an x_{ij}^0 using context or cross-section data.

III. More information

We do have *information* in addition to the observed value of x_{ij}^0 and the categories in the rating system. That information includes the context of a tasting, and it may include cross-section data when panels of judges assess flights of wines. Context and cross-section data contain information that may enable a more precise, lower entropy, estimate of the distribution of u in Equation (2). From here forward, that more-informed distribution will be denoted as u' .

a. Context information

Information about the context for a rating may include the name and reputation of a wine critic or a sponsoring organization and its judges. For example, the CSF judges

Table 1. Distributions of CSF triplicates that include a specified ordinal category

Ordinal	Category rank			
Category	<i>n</i>	Rank	Mean	SD
Gold+	10	1	3.3	2.6
Gold	486	2	5.0	3.1
Gold−	200	3	5.4	2.6
Silver+	142	4	5.9	2.3
Silver	557	5	6.0	2.3
Silver−	301	6	6.4	2.2
Bronze+	553	7	6.7	2.2
Bronze	708	8	6.9	2.3
Bronze−	160	9	7.9	1.9
No award	736	10	7.6	2.8
All	1,599	–	6.5	2.2

are trained and tested, and historical CSF data provide information about the past dispersions of judges' ratings. The means and SDs of the ratings that judges assigned to CSF blind triplicates appear in Table 1. For example, the pool of 557 triplicates that included a gold medal has a mean rank of 5.0 and an SD of 3.1 ranks.

The data in Table 1 yield an estimate of u' . For example, using $(x_{ij}^0 = \text{Gold}) \sim u'$ (5.0, 3.2) yields a solution to Equation (2) that is consistent with historical CSF data. Whatever information may be available and employed, including information about the expertise of a particular judge, u' is an explicit and flexible description of that information's analytical impact on the estimated distribution of a rating.

b. Cross-section information

When a panel of judges assesses a flight of wines, the result is cross-section information. For every rating x_{ij}^0 , x_i^0 is the vector of ratings assigned to wine “ i ” by all the judges, x_j^0 is the vector of ratings assigned to all the wines by judge “ j ,” and x^0 is the union of ratings assigned by all judges to all wines.

Weighting survey data to calculate estimates that are representative of a target population is a common practice. See, for example, Valliant and Dever (2018) and Mercer et al. (2018). The target in this application is a maximum entropy estimate of the distribution of the ratings that one judge assigns to one wine (u'_{ij}) that, following Jaynes, is consistent with known data (x^0). A start on an estimate of u'_{ij} is the distribution of the ratings assigned by all the judges to the subject wine ($d|x_i^0$). But the factors in Figure 1, and Berg et al. (2022), show that some pairs of judges' ratings are more highly correlated than others, some judge's ratings are uncorrelated with any other judges, and some judges' ratings are indistinguishable from random assignments. Thus, some judges within $d|x_i^0$ may not be representative of the target u'_{ij} . To evaluate a survey of demographic and political views, Mercer et al. (2018, p. 4) dealt with unrepresentative observations by assessing the similarity of actual survey demographics to

target demographics, used the most similar observations, and “discarded” the rest. An arithmetical variation of that approach is employed in this effort to estimate u'_{ij} . The information, if any, in the set of ratings assigned by judge k (x_k^0) about x_j^0 is captured by regressing x_k^0 on x_j^0 , using the regression parameters to estimate $\hat{x}_{ij|k}^0$, and then weighting the observed distribution of $\hat{x}_{ij|k}^0$ ($d|\hat{x}_{ij|k}^0$) by the coefficient of determination ($R_{j|k}^2$) for the respective regression. In Equation (3A), that approach has logical boundary properties. If the regression fails to capture any explanatory information ($R_{j|k}^2 = 0$), the weight applied to $d|\hat{x}_{ij|k}^0$ is zero. If the regression does capture explanatory information ($R_{j|k}^2 > 0$), the weighting tends toward unity. That regression and weighting have the effect of making u'_{ij} more representative of the target judge than merely $d|x_i^0$.

Further support for the transform in Equation (3A) is a corollary to the definition of an intelligent machine set forth by Alan Turing. Turing (1950) proposed that a computer can be described as an intelligent machine if, in a typewritten conversation, a computer can imitate a human so well that a computer's and a human's responses are indistinguishable. A corollary to that test suggested here is that if the ratings a wine judge assigns are indistinguishable from random ratings, then that judge can be described as one who assigns ratings randomly. This article takes the position that a judge k who assigns ratings randomly should be ignored and, for $R_{j|k}^2 = 0$, such ratings are disregarded in Equation (3A). In addition, Equation (3A) has the essential asymptotic properties that $u'_{ij} = u$ for judge sample size $n_j = 0$, that u'_{ij} tends to the weighted sum of $d|\hat{x}_{ij|k}^0$ as the sum of information $\sum_{k=1}^J R_{j|k}^2 \rightarrow \infty$, and through Equation (3B), that $\hat{p}_{ij} \rightarrow d|x_{ij}^0$ as the replicate sample size $n_{ij} \rightarrow \infty$.

$$\begin{aligned} u'_{ij} &= \left(\frac{1}{1 + \sum_{k=1}^J R_{j|k}^2} \right) \cdot u + \left(\frac{\sum_{k=1}^J R_{j|k}^2}{1 + \sum_{k=1}^J R_{j|k}^2} \right) \cdot \left(\frac{1}{\sum_{k=1}^J R_{j|k}^2} \right) \cdot \sum_{k=1}^J (R_{j|k}^2 \cdot d|\hat{x}_{ij|k}^0) \\ &= \left(\frac{1}{1 + \sum_{k=1}^J R_{j|k}^2} \right) \cdot \left(u + \sum_{k=1}^J (R_{j|k}^2 \cdot d|\hat{x}_{ij|k}^0) \right) \end{aligned} \quad (3A)$$

$$u'_{ij} = \left(\frac{1}{1 + \sum_{k=1}^J R_{j|k}^2} \right) \cdot u + \left(\frac{\sum_{k=1}^J R_{j|k}^2}{1 + \sum_{k=1}^J R_{j|k}^2} \right) \cdot \left(\frac{1}{\sum_{k=1}^J R_{j|k}^2} \right) \cdot \sum_{k=1}^J (R_{j|k}^2 \cdot d|\hat{x}_{ij|k}^0) \quad (3B)$$

Standing back, the goal of the solution to Equation (2) was to obtain an estimate of the distribution of x_{ij}^0 that is more accurate than ignoring uncertainty about x_{ij}^0 . The goal of Equation (3) is to improve the accuracy of Equation (2) by considering what information may be within cross-section ratings data about the latent distribution of x_{ij}^0 . The results for tests of those objectives appear in Section VI.

VI. Tests

This section presents tests of the relative accuracy of ignoring uncertainty and the solutions to Equations (2) and (3).

Cicchetti (2014) published the scores assigned to two flights of eight wines each at a tasting in Stellenbosch, South Africa. Each flight was assessed by 15 judges, each judge

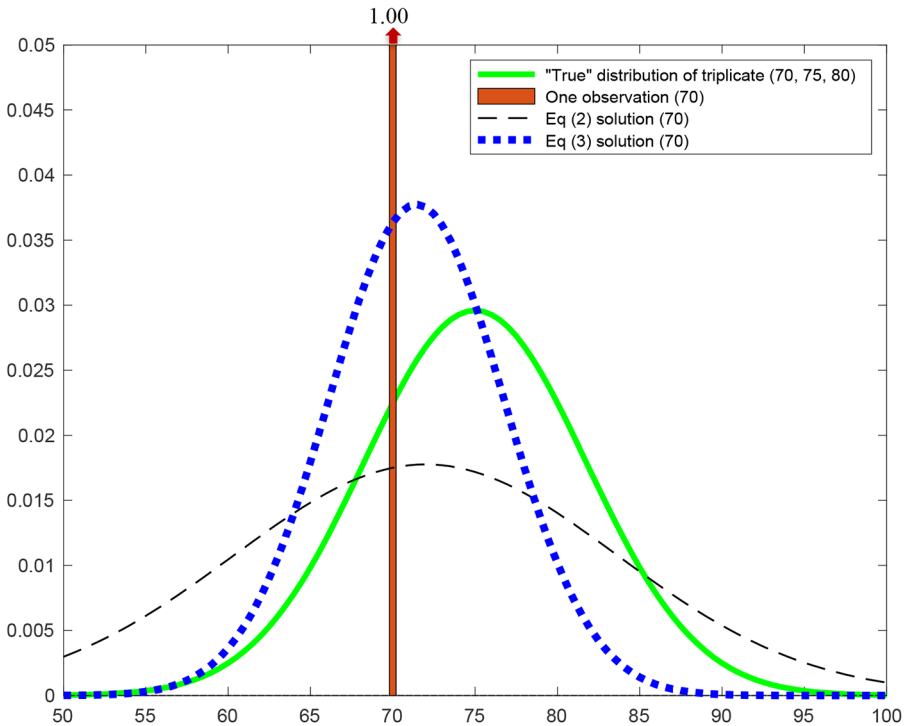


Figure 2. Example of the “true,” observed, Equation (2), and Equation (3) distributions.

assigned a score within the range of 50 to 100 to each wine, and each flight contained a set of blind triplicates. Although it probably understates true variance, assume here that the distribution of the triplicate results for each judge describes the “true” distribution of that judge’s ratings on a triplicate wine. But suppose that wine did not appear in triplicate and that only one of the scores was observed (x_{ij}^0 where $n_{ij} = 1$). In that case, ignoring the uncertainty about a rating is equivalent to assuming that x_{ij}^0 has a degenerate or one-hot distribution. A measure of the error made when assuming the rating has a one-hot distribution is the difference between the “true” distribution and the one-hot distribution. Likewise, a measure of the error made when using the solution to Equation (2) or (3) is the difference between the “true” distribution and the distribution implied by (2) or (3).

A visual explanation appears in Figure 2. Stellenbosch judge #2 assigned ratings of 70, 75, and 80 to a blind triplicate of Pinotage. The “true” distribution implied by those ratings is shown as the solid line in Figure 2. Suppose only a rating of 70 had been observed. Ignoring uncertainty about that rating yields a one-hot distribution for which the probability of observing 70 is 1.00 rather than approximately 0.022 according to the “true” distribution. The solution to Equation (2) is less inaccurate but still much broader than “true.” Adding the information within cross-section data, the solution to Equation (3) moves substantially toward the “true” distribution.

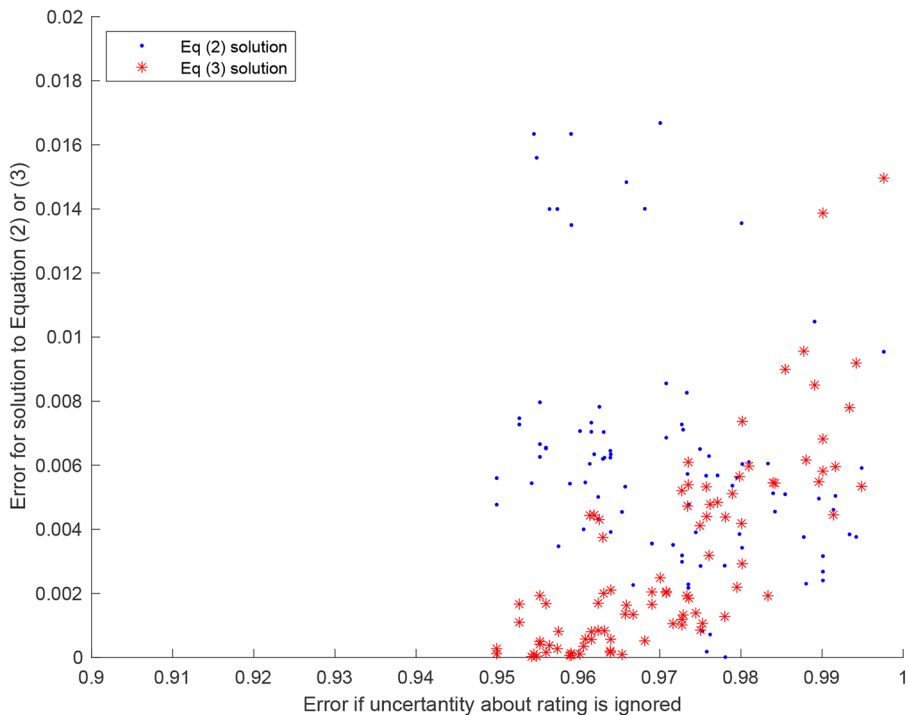


Figure 3. Sums of squared errors for Stellenbosch blind triplicate data, 90 observations.

For all 90 Stellenbosch blind triplicates, the differences between the distributions are quantified as sums of squared errors (SSEs) and shown in Figure 3. The error if uncertainty about a rating is ignored, the SSE is the difference between the “true” distribution and a one-hot distribution. For the solutions to Equations (2) and (3), the SSE is the difference between the “true” distribution and the distribution implied by the solution to either Equation (2) or (3).

The striking implication of Figure 3 is that the solutions to Equations (2) and (3) are substantially more accurate than the common practice of ignoring uncertainty about a rating by assuming a one-hot distribution. And the solution to Equation (3) is usually more accurate than the solution to Equation (2). In Figure 3, almost every dot moves down to a lower-error asterisk.

Additional test results appear in Table 2. The sample mean and SD of the Stellenbosch blind triplicates are, respectively, 78.4 and 4.7. The “true” distributions estimated using Equation (3) have a mean and SD of, respectively, 78.0 and 7.1. Ignoring the uncertainty about a rating yields an SSE of 0.971 and that is near the theoretical maximum of 0.980.¹ The solution to Equation (2) has a lower-than-true mean, a higher-than-true SD, and a much lower average SSE of 0.006. It’s

¹ Maximum SSE is the difference between a one-hot distribution and a uniform random distribution. For a score range of 50 through 100 thus 51 potential ratings, $50 * ((1/51) - 0.0)^2 + 1 * ((1/51) - 1.0)^2 = 0.980$.

Table 2. Summary of errors in estimates of “true” distributions for Stellenbosch blind triplicates

Probability mass function, $p(x_{ij}^0)$	Average error in estimate of “true,” SSE Mean and SD of estimates, $\mu = 78.4$ and $\sigma = 4.7$
$p(x_{ij}^0) = \text{“true” distribution of blind triplicate}$	SSE = 0 Sample: $\mu = 78.4$ and $\sigma = 4.7$ Equation (3): $\mu = 78.0$ and $\sigma = 7.1$
$p(x_{ij}^0) = 1$, ignore uncertainty about x_{ij}^0	SSE = 0.971, see dispersion in Figure 3. $\mu = 78.4$ and $\sigma = 0$
$p(x_{ij}^0) = f(x_{ij}^0, u)$, solution to Equation (2)	SSE = 0.006 see dispersion in Figure 3. $\mu = 76.3$ and $\sigma = 11.5$
$p(x_{ij}^0) = (x_{ij}^0, x^0, u)$, solution to Equation (3)	SSE = 0.003, see dispersion in Figure 3. $\mu = 77.8$ and $\sigma = 6.3$

$(1 - (0.006/0.971)) = 99\%$ more accurate than ignoring uncertainty, but the SD indicates that the distributions are too broad. The solution to Equation (3) has a more accurate mean and SD, and the SD is within the range of the SD estimates for the “true” distribution. And it’s $(1 - (0.003/0.006)) = 50\%$ more accurate than the solution to Equation (2).

V. Conclusions and discussion

Solutions to Equations (2) and (3) show that context and cross-section data can improve analyses of wine ratings. From Table 1, the reduction in error due considering cross-section information compared to ignoring uncertainty is approximately 99%.

Regarding interpretation, the maximum entropy estimates are not assertions that judges’ ratings are merely random. The distributions are assertions that they are the most precise distributions that can be supported by evident knowledge about the ratings observed, context, judges, and wines. They are maximum humility estimates of what is known about a rating, they don’t assume the precision that is implicit when uncertainty is ignored.

With an estimate of the distribution of each rating in hand, a next step is to use those distributions to calculate a consensus among judges about a wine and/or a wine’s ranking among other wines. The maximum entropy distributions could be employed to apply Cochran (1937)’s inverse variance and to calculate expectations of sums, ranks, Borda counts, Shapley values, the normalized aggregates proposed by Gergaud et al. (2021), and the categories proposed by De Nicolò (2023). Analysis of those and other applications, and replication of these results, are suggested here as future research.

Acknowledgments. The author thanks an anonymous reviewer and the attendees at the 16th Annual Conference of the Association of American Wine Economists in Lausanne, Switzerland, for insightful and constructive comments.

References

Barberá S., Bossert W., and Moreo-Ternero J. D. (2023). Wine rankings and the Borda method. *Journal of Wine Economics*, 18(2), 122–138.

- Berg E. C., Mascha M., and Capehart K. W. (2022). Judging reliability at wine and water competitions. *Journal of Wine Economics*, 17(4), 311–328.
- Bodington J. (2022a). A maximum entropy estimate of uncertainty about a wine rating: What can be deduced about the shape of a latent distribution from one observation? *Journal of Wine Economics*, 17(4), 296–310.
- Bodington J. (2022b). Stochastic error and biases remain in blind ratings. *Journal of Wine Economics*, 17(4), 345–351. doi:10.1017/jwe.2022.53
- Bodington J. (2024). A maximum entropy estimate of a wine rating's distribution: Results of tests using large samples of blind triplicates. *Journal of Wine Research*, 35(4), 1–7.
- Cicchetti D. (2014). Blind tasting of South African wines: A tale of two methodologies. AAWE Working Paper No. 164, 15.
- Cochran W. G. (1937). Problems arising in the analysis of a series of similar experiments. *Supplement to the Journal of the Royal Statistical Society*, 4(1), 102–118.
- De Nicolò G. (2023). Wine ratings and commercial reality. AAWE Working Paper No. 276, 37.
- Gergaud O., Ginsburgh V., and Moreno-Ternero J. (2021). Wine ratings: Seeking a consensus among tasters via normalization, approval, and aggregation. *Journal of Wine Economics*, 16(3), 321–342.
- Golan A., Judge G., and Miller D. (1996). *Maximum Entropy Econometrics*, 307. John Wiley & Sons.
- Jaynes E. T. (1957). Information theory and statistical mechanics. *Physics Review*, 106, 620–630.
- Kahneman D., Sibony O., and Sunstein C. R. (2021). *Noise, a Flaw in Human Judgement*, 454. Little, Brown Spark, Hachette Book Group.
- Mercer A., Lau A., and Kennedy C. (2018). How different weighting methods work. Pew Research Center. Downloaded 8 February 2024. Available at <https://www.pewresearch.org/methods/2018/01/26/how-different-weighting-methods-work/>, 10.
- Shannon C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, 27, 379–423, 623–656. Reprinted with corrections, 33 pages, downloaded on 12 August 2020 from <http://web.mit.edu/6.976/www/handout/shannon.pdf>.
- Turing A. M. (1950). Computing machinery and intelligence. *Mind, New Series*, 59(236), 433–460.
- Valliant R., and Dever J. (2018). *Survey Weights: A Step by Step Guide*, 183. Stata Press.