Journal of Wine Economics, Volume 14, Number 3, 2019, Pages 234–251
doi:10.1017/jwe.2019.41

Predicting Italian Wine Quality from Weather Data and Expert Ratings

Alessandro Corsi [a] **and Orley Ashenfelter** [b]

Abstract

In this paper we estimate how a variety of subjective measures of quality taken from the published opinions of several experts on Italian wines (Barolo and Barbaresco) are determined by the weather conditions during the relevant season, in order to assess their reliability. Since these measures of quality are only ordinal, we estimate their determinants using an ordered probit model. The method provides measures of the determinants of vintage quality ratings and suggestions on the reliability of each expert. (JEL Classifications: D12, Q11, Q13)

Keywords: ordered probit, vintage ratings.

I. Introduction

Consumers show an increasing interest for high-quality wines. Higher-income levels favour a shift in food and beverage consumption towards higher-price/higher-quality products, and wine is no exception. On the supply side, producers perceive the shift of demand and strive to improve the quality of their products. Although quality is a multi-facet and slippery concept, it has a very sound economic counterpart, since it is reflected in wine prices.

The search for quality is a particularly difficult task for ageing wines. There is an intrinsic uncertainty in what the final outcome will be of so many years of biochemical processes; the real quality of mature wines is often revealed only in the long run.

This is a slightly revised version of the paper presented at the 7th VDQS Conference held in Reims (France), on May 11–13, 2000. We are grateful to the conference participants for their comments and to Karl Storchmann (the editor) for publishing an otherwise difficult-to-retrieve paper. We also wish to thank Federico Spanna of Regione Piemonte for providing the weather data.
[a] Department of Economics and Statistics, University of Torino, Lungo Dora Siena, 100/A, 10153 Torino, Italy; e-mail: alessandro.corsi@unito.it.
[b] Industrial Relations Section, Firestone Library, Princeton University, Princeton, NJ 08544-2098, USA; e-mail: c6789@princeton.edu.

This is the main reason both consumers and producers try to formulate predictions on future quality of these wines: producers anticipating their wine will be of a high quality will require higher prices even when they are younger, and consumers that make the same prediction will be ready to pay higher prices; the opposite is true when a lower quality is anticipated.

Saying that there is an intrinsic uncertainty on future quality, on the other hand, does not obviously mean that no prediction is possible. Consumers and producers do make predictions, and, moreover, experts provide opinions and ratings to this purpose. Much is known on the influence of several variables on wine quality and, among them, weather conditions during the growing and harvesting period are crucial. Information on weather conditions is largely available, but it is questionable whether experts efficiently exploit it. It has been shown, both for French Bordeaux and for Australian wines (Ashenfelter, Ashmore, and Lalonde, 1995; Byron and Ashenfelter, 1995), that weather data provide an accurate prediction of long-run wine prices (and, hence, of quality). Ashenfelter, Ashmore, and Lalonde's (1995) model explains 83% of total variation in log prices for Bordeaux wines, and Byron and Ashenfelter's 86% of total variation in log prices for Grange Hermitage.[1] More interestingly, it has been shown that these models give more accurate predictions of long-run prices than do market prices in the first years after the vintage. In other words, the market for immature wines over- or under-estimate the future quality and price of mature wines. Since the information on weather data on which the models are based is available, it is apparent that market operators do not efficiently use it.

Nevertheless, experts' opinions and ratings are often the only or the most at-hand information that consumers can have to formulate predictions and to make purchasing decisions. It is, therefore, important to assess to what extent experts' opinions are reliable. More so, when pricing data are not easily available. This is the case for Barolo and Barbaresco wines. Barolo and Barbaresco are produced in the hilly area of Langhe, in Piedmont, northwest of Italy, under strict appellation rules, and are well-known and prestigious wines, well fit for ageing. Although these wines for their better vintages can last as many years as the Bordeaux wines, there is no active market for immature wines, nor an auction system as the one that is active for mature Bordeaux. Of course, mature Barolo and Barbaresco are sold, but the market is very thin, and prices are, therefore, highly variable.

Indications on vintage quality mainly rely on experts' evaluations. Experts usually rate vintages along an ordinal scale, according to their judgement on the future evolution they are anticipating for the wines. Available information on which their

[1] Later papers on the relationship between weather variables and wine quality and price include, among others, Jones and Storchmann (2001), Schamel and Anderson (2003), Jones et al. (2005), Haeger and Storchmann (2006), Ashenfelter and Storchmann (2010, 2016), Oczkowski (2016), Ashenfelter (2017), and Niklas (2017).

evaluations are based include the evolution of weather during the growing season, which is largely known, and, possibly, tasting of immature wines. The issue therefore arises, whether experts' ratings efficiently exploit available information, much in the same sense as the issue whether the market efficiently exploits available information in the case of Bordeaux. If the information is common and is efficiently exploited, it should lead to a convergence of the experts' evaluations.[2] We will assess the way experts' ratings incorporate information on weather data when rating vintages. Since vintage quality ratings are only ordinal, we estimate their determinants using an ordered probit function that incorporates multiple indicators of quality and a common underlying structure of how the weather determines quality. Next, we evaluate the convergence across experts' vintage ratings.

The remainder of the paper is organised as follows: in Section II, the data are presented; Section III presents the results of the estimated models, separately for the three experts we considered; in Section IV, we compare the experts' vintage ratings to ascertain whether they are converging along a common pattern; some considerations conclude.

II. Data

Weather data are drawn from recordings of weather stations belonging to the Regional Service. The weather station in Castiglione Falletto was chosen as the most representative of the Barolo area, since it is included in the DOCG area for Barolo. Castiglione is located a few kilometres from Barolo, and the weather station is close to the top of the hill, at 360 m. above sea level, on the southwest side, and, therefore, represents rather well the typical situation of vineyards in the area. Records for Castiglione are nevertheless only available since 1980. For the previous periods, temperatures for Castiglione were estimated from data of the weather station in Cuneo, and rainfalls from data of the weather station in La Morra. Cuneo is about 50 kilometres from Castiglione, and is on the plain; nevertheless, temperatures are closely related, and the regression fits extremely well.[3] La Morra is very

[2] Some literature examines the consensus among experts on their ratings and evaluations, which is taken as an indicator of reliability. Ashton (2012, p. 70) reviews the research in this field, with a pessimistic conclusion ("Overall, little support is found for the idea that experienced wine judges should be regarded as experts"), a conclusion more nuanced when considering only prominent critics (Ashton, 2013), as confirmed by Luxen (2018). Cyr, Kwong, and Sun (2019) analyse the agreement among experts in their evaluation of *en primeur* wines. Masset, Weisskopf, and Cossutta (2015) evaluate the convergence of expert's evaluations and estimate the effect of climate on expert scores. All these papers refer to individual wines scores or ratings, not to overall vintages.

[3] The regression of Castiglione temperatures (C°) on Cuneo temperatures was estimated based on 1980–1996 monthly average temperatures, and is the following (t-values in parentheses):

$$\text{Castiglione} = -0.570\ (-3.118) + 1.130\ \text{Cuneo}\ (83.665),\ n = 222,\ R^2 = 0.97,\ \text{Adjusted-}R^2 = 0.97.$$

close to Castiglione, just the opposite side of the same valley; since rainfalls have a larger variation than temperatures, and since they may be quite different even at very small distances, the regression fits slightly worse, but still at very satisfactory levels.[4] Based on these regressions, fitted values were calculated for the 1970–1980 period, as well as for an individual missing datum (February 1988 temperature), and a complete 1970–1997 series of temperatures and rainfalls was made available.

We draw vintage ratings from three sources: the "Vini d'Italia" of the Gambero Rosso-Slow Food Arcigola wine guidebook (GR); ratings from Robert Parker on Barolo and Barbaresco (PK); and Sheldon and Pauline Wasserman's Italy's Noble Red Wines (WW) (Wasserman and Wasserman, 1991).

Gambero Rosso is probably the most influential Italian wine guidebook. It is issued every year, also in German and in English. Ratings for wines from each winery considered are given (in terms of up to three "bicchieri" or glasses), but also a general rating of each vintage for the best DOCG wines (Barolo, Barbaresco, Brunello di Montalcino, Chianti classico, Vino nobile di Montepulciano, Amarone) is provided. Other important Italian wine guidebooks do rate each wine, but not vintages as such. For this reason, we chose GR for our exercise. Vintage ratings are given by GR in points ranging from 1 to 5, where 1 is the lowest level, and 5 the highest. The first GR vintage ratings were published in 1989; vintages since 1970 were rated since the beginning. For Barolo and Barbaresco (vintage ratings for the two wines are basically the same, the only exception being 1971, when Barolo vintage was rated 5, and Barbaresco 4), ratings are usually given after three years: for instance, the 1999 issue, written in fall 1998, rated vintages until 1995. This respected the compulsory minimum ageing period of three years for Barolo to be marketed as a DOCG wine (two years are needed for Barbaresco). Nevertheless, the 2000 edition, closed in fall 1999, rates vintages until 1997. It should be noted that in this way ratings are based, apart from general available information on weather conditions, on wine tasting of relatively young wines. Secondly, vintage ratings are rarely revised over time, which means that the increasing available information on actual wine quality is not usually incorporated in ratings.

Parker's ratings are on a scale 1 to 4, and are available since 1957 and until 1993, but many observations are missing in the first years. Wasserman's ratings range from 1 to 12, and go back to 1945, but are only updated until 1990. Since weather data are available from 1970 to 1997, the coverage is not the same for the three series. For this reason, we will examine first separately the three rating systems, and then compare them.

[4] The regression of Castiglione rainfalls (mm.) on La Morra rainfalls was estimated based on 1980–1996 total monthly rainfalls, and is the following (t-values in parentheses):

Castiglione = 2.148 (1.149) + 0.984 La Morra (39.919), $n = 223$, $R^2 = 0.88$, Adjusted-$R^2 = 0.88$.

III. Results

A. Gambero Rosso Ratings

Since ratings are not an ordinal variable, but only indicate a ranking, and they are bounded from 1 to 5, the appropriate model is an ordered probit; Table 1 (first column, Model 1) shows the results of the model estimated from the weather data 1970–1997, where the dependent variable is represented by the GR ratings. The explanatory variables are basically the same as the ones used in Ashenfelter, Ashmore, and Lalonde (1995): WINTRAIN represents total rainfalls (in millimetres) between October and March of the season preceding the vintage; TEMPSPR is the average monthly temperature (in degrees centigrade) March to July; SUMMRAIN is total rainfalls in August and September; and SUMMTEMP the average temperature in the same months. The model also estimates the underlying implicit threshold values for the different ratings; since there are five different ratings, there are three threshold values.

The overall fit of the model can be tested with a likelihood ratio test, which gives a χ^2 of 12.757, significant at a 2% level. Nevertheless, the only significant parameters at the conventional levels are the variable referring to summer rainfalls, and the threshold values[5]; also, the Pseudo-R^2 value (McFadden, 1974) is quite low. The negative parameter of the summer rainfall variable indicates that vintages are rated high when August and September are dry. The other variables are not significant, but the signs indicate that a wet winter preceding the vintage increases the rating, as well as a warm spring. However, contrary to expectations, the summer temperature variable displays a negative sign.

Since the purpose of the model is essentially practical, one might wonder how good is the model in predicting the ratings. This is shown in Table 2, where the observed and the predicted ratings are reported, as well as the probability of the predicted ratings. Predicted ratings are the ones that have the highest probability. Although there is a general correspondence between predicted and observed ratings, it is far from perfect. Observed and predicted ratings exactly match in 13 over 28 years, and on the average the model rates are 1.1 places above or below the observed ratings; in some years the difference is rather strong (e.g., 1977, 1987, 1992, 1995).

Of course, the limited fit of the model can stem from two opposite reasons: (a) weather data are unable to adequately explain vintage quality; or (b) GR ratings do not appropriately represent actual vintage quality. We can take advantage of other pieces of available information to try to ascertain which is the case.

[5] To control for possible measurement error, we also tested the inclusion of a dummy variable, taking the value 1 for the years when weather data for Castiglione were estimated from La Morra and Cuneo data, and 0 otherwise, but the variable was never significant, and created problems with shorter series (see under), and it was, therefore, dropped.

Table 1
Results of the Ordered Probit Models for Gambero Rosso Ratings, 1970–1997

	Model 1 1970–1997		*Model 2 Multiplicative Heteroscedasticity 1970–1997*		*Model 3 1970–1985*	
Variable	*Parameter*	*t-ratio*	*Parameter*	*t-ratio*	*Parameter*	*t-ratio*
Constant	−2.549	−0.472	−6.391	−0.307	−1.999	−0.191
WINTRAIN	0.002	1.094	0.005	0.816	0.002	0.612
SUMMRAIN	−0.011	−2.675	−0.035	−1.552	−0.018	−1.970
TEMPSPR	0.451	1.162	0.565	0.480	0.125	0.231
SUMMTEMP	−0.101	−0.376	0.343	0.361	0.208	0.353
Mu(1)	1.887	2.632	5.649	1.517	2.162	2.288
Mu(2)	2.375	3.207	7.371	1.604	2.493	3.083
Mu(3)	3.062	3.692	10.025	1.712	4.478	3.429
			Variance Function			
TIME			0.0820	2.097		
Log-L	−34.217		−30.810		−14.663	
Log-L(0)	−40.596		−40.595		−22.268	
McFadden's Pseudo R^2	0.157		0.241		0.342	
LR test of the model: χ^2 (d.f.)	12.757 (4)		19.571 (5)		15.210 (4)	
P-value	0.0125		0.0015		0.0043	
LR test of model 1 against model 2: χ^2 (d.f.)			6.813 (1)			
P-value			0.009			

Notes: WINTRAIN = total rainfall October–March, SUMMRAIN = total rainfall August–September, TEMPSPR = average monthly temperature April–July, SUMMTEMP = average monthly temperature August–September, TIME = YEAR–1969, Mu(i) = threshold values.

Specifically, we know that GR vintage ratings were first issued in 1989, with ratings starting from 1970. Therefore, unlike in more recent years, at that time information based on actual vintage quality, as revealed by mature wine, was available for more distant vintages. We therefore estimated two further models: the first one is estimated on the same series 1970–1997, and with the same explanatory variables, but with a further assumption of multiplicative heteroscedasticity, that is, that the variance of the random term is an increasing function of the time past from the first vintage.[6] This is equivalent to assuming that ratings are less and less accurate from the beginning of the series on. The results are reported in Table 1, Model 2. Again, the overall model is significant, now at the 1% level; the Pseudo-R^2 value is also higher. Nevertheless, the parameter of the SUMMRAIN variable is now only significant at a 12% level and the threshold variable roughly at the same level or slightly better. The model can be contrasted to the previous one, both with a t-test on the

[6] More specifically, the assumption is: $\varepsilon \sim N[0, (e^{(g^{*}time)})^2]$, where ε is the random term of the rating equation, time is (vintage year – 1969), and g is a parameter to be estimated (Greene, 1995).

Table 2

Observed and Predicted Vintage Ratings, Gambero Rosso 1970–1997 and 1970–1985

Year	*Gambero Rosso Ratings*	*Predicted Ratings Model 1*	*Estimated Probability Model 1*	*Absolute Deviations Model 1*	*Predicted Ratings Model 2*	*Estimated Probability Model 2*	*Absolute Deviations Model 2*	*Predicted Ratings Model 3*	*Estimated Probability Model 3*	*Absolute Deviations Model 3*
1970	4	2	0.350	2	4	0.727	0	4	0.678	0
1971	4	5	0.494	1	4	0.587	0	4	0.586	0
1972	2	2	0.455	0	2	0.935	0	2	0.718	0
1973	2	2	0.599	0	2	0.955	0	2	0.660	0
1974	4	2	0.301	2	4	0.524	0	4	0.662	0
1975	2	2	0.595	0	2	0.596	0	2	0.575	0
1976	2	2	0.631	0	2	0.887	0	2	0.642	0
1977	2	5	0.296	3	4	0.395	2	4	0.614	2
1978	5	5	0.321	0	4	0.469	1	4	0.616	1
1979	4	2	0.619	2	2	0.757	2	2	0.719	2
1980	4	2	0.350	2	4	0.409	0	4	0.638	0
1981	3	5	0.358	2	4	0.374	1	4	0.637	1
1982	5	5	0.835	0	5	0.745	0	5	0.656	0
1983	4	5	0.684	1	5	0.528	1	4	0.591	0
1984	1	2	0.645	1	2	0.577	1	2	0.720	1
1985	5	5	0.632	0	5	0.598	0	5	0.533	0
1986	3	5	0.475	2	5	0.347	2			
1987	2	5	0.306	3	5	0.423	3			
1988	5	5	0.615	0	5	0.537	0			
1989	5	5	0.565	0	5	0.483	0			
1990	5	5	0.449	0	5	0.453	0			
1991	3	5	0.288	2	5	0.472	2			
1992	2	5	0.354	3	5	0.434	3			
1993	3	5	0.398	2	5	0.393	2			
1994	2	2	0.563	0	1	0.304	1			
1995	5	2	0.622	3	1	0.368	4			
1996	5	5	0.461	0	5	0.397	0			
1997	5	5	0.617	0	5	0.514	0			
Total				**31**			**25**			**7**
Average				1.107			0.893			0.438

TIME variable of the variance function, and with a likelihood ratio test on the overall model. Both tests reject the restriction implied in the previous model. Moreover, the rating predictions are now more accurate: they exactly match the observed ratings in 15 years, and on the average vintages are rated 0.9 places above or below the observed ones.

Another way to see whether ratings in the first period are more accurate than in the later one is to estimate a model like the first one, but on the 1970–1985 period. The idea here is to estimate a model for which GR vintage ratings were mainly based on the observation of the actual wine quality, since enough time passed in 1989, when the first ratings were published, to allow for tasting mature wines for vintages 1970 to 1985.

The results of this model are reported in Table 1, Model 3. The overall model is significant at the 1% level, and the significant parameters include the summer rainfall variable, and the thresholds; the Pseudo-R^2 value is now 0.342. As compared to the model estimated on the 1970–1997 period, in this one all other parameters, although not significant at the usual levels, exhibit the expected signs: as already found by Ashenfelter, Ashmore, and Lalonde (1995), better vintages are the ones when the preceding winter was wet, spring and summer temperatures were high, and summer was dry. The model estimated on the 1970–1985 period also has much better predictive power (see Table 2): 11 over 16 vintage ratings are exactly predicted, and on the average predicted ratings are 0.4 places higher or lower than the observed ones. Even considering the shorter time span, this represents a dramatic increase in the accuracy of the model.

Both considering the model allowing for heteroscedasticity depending on time, and considering the model estimated on the first period, the conclusion is unavoidable that there is a decreasing correspondence between what weather data would predict and what the GR ratings are.

This conclusion supports the view that GR experts' opinions, like market prices, do not fully exploit available information: when the relevant information on mature wines is available, ratings do correspond more strictly to what weather data would suggest.

B. Parker Ratings

The ordered probit model estimated from Parker ratings for the period 1970–1993[7] gave the results reported in Table 3, which also reports the same model estimated for the same years with GR ratings, so to allow a comparison. Since PK ratings only range from 1 to 4, there are only two threshold values, as compared to the three for GR. The results for both rating systems are very similar; only the summer rainfall variable, and the thresholds, are significant at the usual values, and the parameter

[7] Excluding year 1973, for which no rating was available.

Table 3

Results of Ordered Probit Models for Parker and Gambero Rosso Ratings 1970–1993 (1973 Excluded)

	PARKER				*GAMBERO ROSSO*			
	Model 1		*Model 2 Multiplicative Heteroscedasticity*		*Model 1*		*Model 2 Multiplicative Heteroscedasticity*	
Variable	*Parameter*	*t-ratio*	*Parameter*	*t-ratio*	*Parameter*	*t-ratio*	*Parameter*	*t-ratio*
Constant	4.372	0.574	2.304	0.497	1.437	0.210	–0.691	–0.043
WINTRAIN	0.002	0.715	0.002	1.039	0.003	1.458	0.004	0.911
SUMMRAIN	–0.023	–4.017	–0.015	–1.401	–0.020	–3.419	–0.030	–1.321
TEMPSPR	0.154	0.324	0.157	0.526	0.274	0.581	0.327	0.353
SUMMTEMP	–0.212	–0.591	–0.168	–0.652	–0.107	–0.319	0.117	0.156
Mu(1)	0.660	1.909	0.412	0.944	2.032	2.301	3.571	1.263
Mu(2)	1.576	2.593	0.966	1.181	2.655	3.046	4.860	1.153
Mu(3)					3.709	3.490	7.000	1.240
			Variance Function				Variance Function	
TIME			–0.037	–0.544			0.054	0.733
Log-L	–22.602		–22.459		–26.367		–25.869	
Log-L(0)	–31.281		–31.281		–34.320		–34.320	
McFadden's Pseudo R^2	0.277		0.282		0.232		0.246	
LR test of the model: χ^2 (d.f.)	17.359 (4)		17.645 (5)		15.906 (4)		16.901 (5)	
P-value	0.002		0.003		0.003		0.005	
LR test of model 1 against model 2: χ^2 (1 d.f.)			0.285				0.996	
P-value			0.593				0.318	

Notes: WINTRAIN = total rainfall October–March, SUMMRAIN = total rainfall August–September, TEMPSPR = average monthly temperature April–July, SUMMTEMP = average monthly temperature August–September, TIME = YEAR–1969, Mu(i) = threshold values.

values for summer rainfall are negative for both, as expected, and very similar. The other variables are not significant, but the signs indicate that rainfalls during the winter preceding the vintage increase the rating, as well as a warm spring. Again, contrary to expectations, the parameter of the summer temperature variable is negative.

Observed and predicted ratings are reported in Table 4. For the PK model, predicted ratings are equal to observed ratings in 15 years over 23, and on the average the model rates 0.63 points above or below the observed rating; again, 1977, 1979, 1987 are years for which the difference between observed and predicted ratings is strong. For the GR model, 11 over 23 ratings are perfectly predicted, and the average absolute difference is slightly larger (0.67; but the range of ratings is larger, 1 to 5 instead of 1 to 4). Apparently, the overall fit is better for PK than for GR, also considering that both the log-likelihood and the Pseudo-R^2 are larger for the former.

Finally, for both rating systems, a model with multiplicative heteroscedasticity was estimated, but for both likelihood ratio tests and t-tests could not reject the restriction implied in the model without heteroscedasticity.[8] For GR, since this series is shorter than the previous one, this result is not inconsistent with the explanation given earlier. As to Parker ratings, there is no apparent improvement along time in the correspondence between what weather data would predict and actual ratings.

C. Wasserman Ratings

The series on which a model can be estimated with WW ratings is even shorter, since it only goes from 1970 to 1990. The relevant results are presented in Table 5, along with an ordered probit model estimated on GR ratings for the same period.[9] Not surprisingly, given the short period and the number of parameters involved, the WW model is overall highly significant, but no parameter is significant at the usual levels (only two thresholds are marginally significant). Nevertheless, the parameter for summer rainfalls has a reasonably large t-value, and has the same sign and magnitude of the corresponding parameter estimated on GR ratings, which is the only significant one, apart from the threshold values.

Predicted ratings, given the larger range, show on the average larger differences, relative to observed ratings, than the other rating systems, since they are 1.9 places above or below the observed ones (Table 6). The largest differences are in years 1974, 1979, and 1987; with the exception of the last one, they are years different from the ones where the largest differences are observed for the other rating systems. In this case, the fit is worse for WW than for GR.

[8] The likelihood ratio tests provided χ^2 values of 0.285 and 0.996, with 1 d.f., for Parker and Gambero Rosso models, respectively.

[9] Since WW ratings include 12 values, but only 10 of them were given in the 1970–1990 period, ratings were rescaled so to give a complete ordering: for instance, since rating 2 was not given, previous ratings 3 became 2, and so on. Since ratings only have an ordinal meaning, this has no effect on the estimation.

Table 4
Observed and Predicted Vintage Ratings, Parker and Gambero Rosso 1970–1993 (1973 Excluded)

	PARKER							GAMBERO ROSSO						
		Model 1			Model 2				Model 1			Model 2		
Year	Observed Ratings	Predicted Ratings	Prob. of predicted Ratings	Abs. Deviat.	Predicted Ratings	Prob. of Predicted Ratings	Abs. Deviat.	Observed Ratings	Predicted Ratings	Prob. of Predicted Ratings	Abs. Deviat.	Predicted Ratings	Prob. of Predicted Ratings	Abs. Deviat.
1970	3	3	0.349	0	1	0.318	2	4	4	0.358	0	4	0.638	0
1971	4	4	0.676	0	4	0.628	0	4	5	0.529	1	4	0.531	0
1972	1	1	0.756	0	1	0.642	0	2	2	0.636	0	2	0.849	0
1974	3	1	0.318	2	1	0.342	2	4	4	0.370	0	4	0.501	0
1975	1	1	0.815	0	1	0.817	0	2	2	0.686	0	2	0.663	0
1976	1	1	0.957	0	1	0.925	0	2	2	0.627	0	2	0.731	0
1977	1	4	0.356	3	4	0.426	3	2	4	0.400	2	4	0.457	2
1978	4	4	0.666	0	4	0.642	0	5	5	0.474	0	4	0.463	1
1979	3	1	0.896	2	1	0.905	2	4	2	0.673	2	2	0.704	2
1980	2	1	0.334	1	1	0.375	1	4	4	0.371	0	4	0.446	0
1981	1	1	0.468	0	1	0.523	0	3	2	0.390	1	4	0.372	1
1982	4	4	0.741	0	4	0.772	0	5	5	0.724	0	5	0.685	0
1983	3	3	0.352	0	4	0.409	1	4	5	0.452	1	5	0.435	1
1984	1	1	0.913	0	1	0.966	0	1	2	0.635	1	2	0.576	1
1985	4	4	0.478	0	4	0.480	0	5	5	0.496	0	5	0.532	0
1986	2	4	0.436	2	4	0.429	2	3	4	0.402	1	4	0.331	1
1987	2	4	0.495	2	4	0.395	2	2	4	0.402	2	5	0.384	3
1988	4	4	0.778	0	4	0.850	0	5	5	0.656	0	5	0.550	0
1989	4	4	0.782	0	4	0.834	0	5	5	0.557	0	5	0.483	0
1990	4	4	0.576	0	4	0.517	0	5	4	0.399	1	5	0.412	0
1991	2	3	0.305	1	2	0.355	0	3	4	0.371	1	5	0.387	2
1992	1	1	0.301	0	2	0.367	1	2	4	0.343	2	5	0.335	3
1993	3	1	0.304	2	2	0.368	1	3	4	0.356	1	5	0.299	2
Total				**15**			**17**				**16**			**19**
Average				0.625			0.708				1			0.792

Table 5

Results of Ordered Probit Models for Wasserman and Gambero Rosso Ratings 1970–1990

	WASSERMAN				*GAMBERO ROSSO*			
	Model 1		*Model 2 Multiplicative Heteroskedasticity*		*Model 1*		*Model 2 Multiplicative Heteroskedasticity*	
Variable	*Parameter*	*t-ratio*	*Parameter*	*t-ratio*	*Parameter*	*t-ratio*	*Parameter*	*t-ratio*
Constant	6.518	0.766	7.090	0.451	−1.862	−0.211	−9.344	−0.257
WINTRAIN	−0.002	−0.477	−0.002	−0.292	0.002	0.967	0.006	0.567
SUMMRAIN	−0.021	−1.428	−0.022	−0.727	−0.016	−2.658	−0.049	−0.600
TEMPSPR	0.080	0.143	0.081	0.129	0.245	0.463	0.539	0.268
SUMMTEMP	−0.167	−0.446	−0.178	−0.420	0.083	0.204	0.708	0.376
Mu(1)	0.561	1.091	0.594	0.686	2.190	2.617	7.620	0.551
Mu(2)	1.046	1.544	1.107	0.825	2.600	3.180	9.397	0.545
Mu(3)	1.490	1.316	1.586	0.711	3.805	3.611	14.298	0.577
Mu(4)	1.748	0.802	1.871	0.545				
Mu(5)	2.068	1.008	2.225	0.652				
Mu(6)	2.652	1.367	2.864	0.717				
Mu(7)	3.191	1.678	3.452	0.728				
Mu(8)	4.363	1.876	4.759	0.671				
			Variance Function				Variance Function	
TIME			0.007	0.073			0.123	0.821
Log-L	−33.646		−33.633		−22.061		−19.360	
Log-L(0)	−45.661		−45.661		−30.297		−30.297	
McFadden's Pseudo R^2	0.268		0.263		0.272		0.361	
LR test of the model: χ^2 (d.f.)	24.029 (4)		24.055 (5)		16.472 (4)		21.875 (5)	
P-value	0.000		0.000		0.003		0.001	
LR test of model 1 against model 2: χ^2 (1 d.f.)			0.026				5.402	
P-value			0.872				0.020	

Notes: WINTRAIN = total rainfall October–March, SUMMRAIN = total rainfall August–September, TEMPSPR = average monthly temperature April–July, SUMMTEMP = average monthly temperature August–September, TIME = YEAR–1969, Mu(i) = threshold values.

Table 6

Observed and Predicted Vintage Ratings, Wasserman and Gambero Rosso 1970–1990

	WASSERMAN							*GAMBERO ROSSO*						
		Model 1			*Model 2*				*Model 1*			*Model 2*		
Year	*Observed Ratings*	*Predicted Ratings*	*Prob. of Predicted Ratings*	*Abs. Deviat.*	*Predicted Ratings*	*Prob. of Predicted Ratings*	*Abs. Deviat.*	*Observed Ratings*	*Predicted Ratings*	*Prob. of Predicted Ratings*	*Abs. Deviat.*	*Predicted Ratings*	*Prob. of Predicted Ratings*	*Abs. Deviat.*
1970	8	7	0.226	1	7	0.244	1	4	4	0.447	0	4	0.962	0
1971	11	9	0.265	2	9	0.268	2	4	5	0.492	1	4	0.769	0
1972	1	1	0.837	0	1	0.852	0	2	2	0.721	0	2	0.991	0
1973	1	1	0.920	0	1	0.930	0	2	2	0.615	0	2	0.978	0
1974	5	1	0.414	4	1	0.406	4	4	4	0.431	0	4	0.707	0
1975	1	1	0.457	0	1	0.455	0	2	2	0.651	0	2	0.587	0
1976	1	1	0.911	0	1	0.918	0	2	2	0.682	0	2	0.863	0
1977	1	3	0.181	2	3	0.179	2	2	4	0.422	2	4	0.481	2
1978	9	7	0.229	2	7	0.234	2	5	4	0.431	1	4	0.567	1
1979	6	1	0.636	5	1	0.639	5	4	2	0.726	2	2	0.726	2
1980	4	1	0.313	3	1	0.306	3	4	4	0.453	0	4	0.463	0
1981	2	4	0.176	2	4	0.174	2	3	4	0.435	1	4	0.419	1
1982	10	9	0.385	1	9	0.388	1	5	5	0.795	0	5	0.658	0
1983	4	4	0.173	0	4	0.168	0	4	5	0.534	1	5	0.467	1
1984	2	1	0.491	1	1	0.493	1	1	2	0.724	1	2	0.449	1
1985	11	7	0.230	4	7	0.225	4	5	5	0.637	0	5	0.551	0
1986	9	7	0.228	2	7	0.222	2	3	4	0.453	1	5	0.351	2
1987	5	9	0.342	4	9	0.341	4	2	4	0.440	2	5	0.441	3
1988	10	9	0.348	1	9	0.346	1	5	5	0.622	0	5	0.504	0
1989	12	9	0.439	3	9	0.427	3	5	5	0.514	0	5	0.468	0
1990	11	9	0.422	2	9	0.408	2	5	4	0.429	1	5	0.458	0
Total				**39**			**39**				**13**			**13**
Average				1.857			1.857				0.619			0.619

IV. A Comparison between Expert Opinions

An interesting issue is detecting whether the experts have a common ground in evaluating vintages, or each of them has specific information and/or evaluation. Of course, a common ground does exist, based on weather data: so the issue is more precisely whether the experts have specific information and evaluation apart from weather data. If they have, their ratings, once the variation due to weather is removed, should be uncorrelated. By contrast, if the ratings, controlling for weather, are still correlated, a common ground does exist: this may be due to a common misperception of actual wine quality, for instance, because experts communicate among them, or because opinions are mutually reinforcing; or, the weather data, as modelled, do not appropriately represent all determinants of wine quality, and some specific, and common to all experts, quality determinant is left out.

After creating a series common to all three rating systems (1970 to 1990, except for 1973), the issue was addressed in two ways. The first one is estimating linear regressions of the different ratings on weather data, and calculating the correlation coefficients between the residuals of the three regressions; this gives an approximate evaluation of how the ratings are correlated, apart from the weather. The relevant correlation coefficients between the residuals are 0.806 between GR and WW, 0.876 between GR and PK, and 0.870 between PK and WW.

The second way is estimating the same ordered probit model for the three rating systems, and calculating a chi-square statistic between the residuals. Theoretically, the residuals are the difference between the ratings predicted by a model including weather variables (which are the common ground for the experts), and ratings predicted by a model without weather variables; since our models only include weather variables, the ratings predicted by a model without weather variables are simply the observed ratings, as they have the highest probability. A chi-square statistic is preferable to a correlation coefficient to measure the association between residuals, because residuals are difference in ordering, and not cardinal values.

Table 7 presents the results of the ordered probit models estimated for the same years for the three rating systems, and Table 8 the residuals. The results of the ordered probit models are not very different from the ones presented earlier; the only significant parameters for GR and PK concern summer rainfall, while no parameter is significant for WW.[10]

[10] It should be noted that the best fit (both in terms of log-likelihood and of Pseudo R^2) is for the Parker's ratings, followed by Gambero's, while Wasserman and Wasserman's comes last. Of course, this means that Parker's ratings are better explained by the weather variables; it could be concluded that they are the most accurate only if it is assumed (or proved) that weather variables, as modelled, correctly predict the quality of future mature wine.

Table 7

Results of the Ordered Probit Models for Gambero Rosso, Parker and Wasserman Ratings, 1970–1990 (1973 Excluded)

	GAMBERO		*WASSERMAN*		*PARKER*	
Variable	*Parameter*	*t-ratio*	*Parameter*	*t-ratio*	*Parameter*	*t-ratio*
Constant	−1.481	−0.167	6.308	0.732	2.320	0.285
WINTRAIN	0.003	1.129	−0.002	−0.496	0.001	0.446
SUMMRAIN	−0.018	−2.816	−0.020	−1.348	−0.020	−3.348
TEMPSPR	0.296	0.543	0.061	0.110	0.105	0.215
SUMMTEMP	0.014	0.032	−0.143	−0.372	−0.070	−0.183
Mu(1)	1.963	2.272	0.573	1.097	0.629	1.468
Mu(2)	2.372	2.803	1.058	1.565	1.477	2.172
Mu(3)	3.598	3.360	1.501	1.318		
Mu(4)			1.756	0.818		
Mu(5)			2.069	1.024		
Mu(6)			2.642	1.371		
Mu(7)			3.176	1.679		
Mu(8)			4.342	1.861		
Log-L	−21.436		−33.550		−18.557	
Log-L(0)	−28.980		−44.142		−26.702	
McFadden's Pseudo R^2	0.260		0.240		0.305	
LR test of the model: χ^2 (4 d.f.)	15.088		21.184		16.289	
P-value	0.005		0.000		0.003	

Notes: WINTRAIN = total rainfall October–March, SUMMRAIN = total rainfall August–September, TEMPSPR = average monthly temperature April–July, SUMMTEMP = average monthly temperature August–September, TIME = YEAR–1969, Mu(i) = threshold values.

The chi-square test of independence of the residuals gives values of 24.167 (12 d.f.) between GR and PK, 42.071 (28 d.f.) between GR and WW, and 36.190 (21 d.f.) between PK and WW. All values are significant at a 5% level or better.[11]

Both approaches lead to the same conclusion that the hypothesis of independence between ratings of the different systems has to be rejected. Therefore, either the experts have some information in common that is not available to researchers, or they have a common misperception of how weather affects mature wine quality.

V. Summary and Conclusions

Different ordered probit models have been estimated of the influence of weather data on vintage ratings given by different experts. Since all models are overall significant, this shows that weather data actually are among the determinants of vintage ratings,

[11] The correlation coefficients between the residuals are 0.70 between GR and PK, 0.67 between GR and WW, and 0.71 between PK and WW.

Table 8
Residuals of the Ordered Probit Models 1970–1990 (1973 Excluded)

Year	*GAMBERO*	*PARKER*	*WASSERMAN*
1970	0	1	1
1971	1	0	0
1972	0	0	0
1974	0	–2	–3
1975	0	0	0
1976	0	0	0
1977	2	2	2
1978	–1	0	0
1979	–2	–2	–4
1980	0	1	–2
1981	1	0	2
1982	0	0	1
1983	1	1	1
1984	1	0	–1
1985	0	0	–2
1986	1	2	0
1987	2	2	5
1988	0	0	1
1989	0	0	–1
1990	–1	0	0

but the only individual weather significant parameter is summer rainfalls, which may indicate that in evaluating vintages all experts attach a particular and overwhelming importance to August and September rainfalls.

The analysis on Gambero Rosso ratings shows a decreasing correspondence between predicted and observed ratings along time. This is consistent with what was found both by Ashenfelter, Ashmore, and Lalonde (1995) for Bordeaux wines, and by Byron and Ashenfelter (1995) for Australian wines: prices in the first years after the vintage over- or under-estimate future prices of mature wines and, since market prices are based on predictions on future wine quality, this means that market operators' predictions in those periods are not accurate. Much in the same way, qualitative ratings given in the first years after the vintage may not accurately reflect future mature wine quality.

A comparison of the models estimated from different rating systems indicates that the evaluations of the different experts are highly correlated, so that either there is a common information not available to the larger public (researchers included), or there is a common misperception of actual vintage quality, or possibly a self-reinforcing imitation process in giving ratings.

Ashenfelter and Jones (2013) showed that expert opinion does not fully incorporate all available information. Expert opinion on Italian wines do not seem to behave differently. Of course, much greater caution is convenient in the case we are examining, because we do not have the counterpart of the "objective" evaluation of quality given by market prices for mature Bordeaux and Australian wines; and our conclusions, therefore, heavily rely on the assumption that weather, *as modelled*, does determine vintage quality. There is no doubt that weather is a determinant of wine quality, but many other weather characteristics may possibly have an important influence, in particular in connection with plant pathologies. Anyway, if the assumption is true, experts of Italian wines seem to incorporate less of the weather information than their French equivalents: for instance, Ashenfelter and Jones (2013) estimated ordered probit models of the ratings given by French experts to Bordeaux vintages, and they found that both average temperature and harvest rain were highly significant determinants of experts' quality index, while winter rain was marginally significant. By contrast in our estimates winter rain and average temperatures are never statistically significant, and even rain in August and September is not significant in all models. Experts' ratings of Italian wines seem to keep in mind the overall weather course (as shown by the models being overall highly significant), but not in such a way that the effect of specific characteristics are reflected in statistically significant parameters.

References

Ashenfelter, O. (2017). The hedonic approach to vineyard site selection: Adaptation to climate change and grape growing in emerging markets. *Journal of Wine Economics*, 12(1), 3–15.

Ashenfelter, O., Ashmore, D., and Lalonde, R. (1995). Bordeaux wine vintage quality and the weather. *Chance*, 8(4), 7–14.

Ashenfelter, O., and Jones, G. (2013). The demand for expert opinion: Bordeaux wine. *Journal of Wine Economics*, 8(3), 285–293.

Ashenfelter, O., and Storchmann, K. (2010). Using hedonic models of solar radiation and weather to assess the economic effect of climate change: The case of Mosel valley vineyards. *Review of Economics and Statistics*, 92(2), 333–349.

Ashenfelter, O., and Storchmann, K. (2016). Climate change and wine: A review of the economic implications. *Journal of Wine Economics*, 11(1), 105–138.

Ashton, R. H. (2012). Reliability and consensus of experienced wine judges: Expertise within and between? *Journal of Wine Economics*, 7(1), 70–87.

Ashton, R. H. (2013). Is there consensus among wine quality ratings of prominent critics? An empirical analysis of red Bordeaux, 2004–2010. *Journal of Wine Economics*, 8(2), 225–234.

Byron, R. P., and Ashenfelter, O. (1995). Predicting the quality of an unborn Grange. *Economic Record*, 71(212), 40–53.

Cyr, D., Kwong, L., and Sun, L. (2019). Who will replace Parker? A copula function analysis of Bordeaux *en primeur* wine raters. *Journal of Wine Economics*, 14(2), 133–144.

Greene, W. H. (1995). *LIMDEP Version 7.0 User's Manual*. Bellport, NY: Econometric Software.

Haeger, J. W., and Storchmann, K. (2006). Prices of American Pinot Noir wines: Climate, craftsmanship, critics. *Agricultural Economics*, 35(1), 67–78.

Jones, G. V., and Storchmann, K. H. (2001). Wine market prices and investment under uncertainty: An econometric model for Bordeaux Crus Classés. *Agricultural Economics*, 26(2), 115–133.

Jones, G. V., White, M. A., Cooper, O. R., and Storchmann, K. (2005). Climate change and global wine quality. *Climatic Change*, 73(3), 319–343.

Luxen, M. F. (2018). Consensus between ratings of red Bordeaux wines by prominent critics and correlations with prices 2004–2010 and 2011–2016: Ashton revisited and expanded. *Journal of Wine Economics*, 13(1), 83–91.

Masset, P., Weisskopf, J. P., and Cossutta, M. (2015). Wine tasters, ratings, and *en primeur* prices. *Journal of Wine Economics*, 10(1), 75–107.

McFadden, D. (1974): The measurement of urban travel demand. *Journal of Public Economics*, 3(4), 303–328.

Niklas, B. (2017). Impact of annual weather fluctuations on wine production in Germany. *Journal of Wine Economics: Selected Proceedings*, 12(4), 436–445.

Oczkowski, E. (2016). The effect of weather on wine quality and prices: An Australian spatial analysis. *Journal of Wine Economics*, 11(1), 48–65.

Schamel, G., and Anderson, K. (2003). Wine quality and varietal, regional and winery reputations: Hedonic prices for Australia and New Zealand. *Economic Record*, 79(246), 357–369.

Wasserman, S., and Wasserman, P. (1991). *Italy's Noble Red Wines*. New York: Macmillan Publishing Company.