Food Quality and Preference 40 (2015) 240–251

Contents lists available at ScienceDirect

Food Quality and Preference

journal homepage: www.elsevier.com/locate/foodqual

Exemplarity measurement and estimation of the level of interjudge agreement for two categories of French red wines

Aurélie Loison [a,*], Ronan Symoneaux [b], Pascale Deneulin [c], Thierry Thomas-Danguin [a], Christine Fant [d], Laurence Guérin [e], Yves Le Fur [a]

[a] Centre des Sciences du Goût et de l'Alimentation, UMR 6265 CNRS, UMR 1324 INRA, Université de Bourgogne, Centre INRA, 17 rue Sully, BP 86510, F-21065 Dijon Cedex, France
[b] Groupe ESA, UPSP GRAPPE, SFR QUASAV 4207, 55 rue Rabelais, 49007 Angers, France
[c] University of Applied Sciences – Changins, Viticulture and Oenology, route de Duillier 50, case postale 1148, CH-1260 Nyon 1, Switzerland
[d] Département Sciences de l'Ingénieur et des Procédés, AgroSup Dijon, 26 boulevard Petit Jean, BP 87999, F-21079 Dijon, France
[e] Institut Français de la Vigne et du Vin, Pôle Val de Loire, 46 avenue Gustave Eiffel, F-37100 Tours, France

ARTICLE INFO

Article history:
Received 10 February 2014
Received in revised form 5 October 2014
Accepted 11 October 2014
Available online 31 October 2014

Keywords:
Wine
Exemplarity
Typicality
Interjudge agreement
Categorization
Protected Designation of Origin

ABSTRACT

Exemplarity measurements of wines by at least 20 judges are used to estimate the degree of interjudge (dis)agreement and to tell wines apart into two contrasting extremes. Two sets of French red wines – Pinot noir from Burgundy and Cabernet franc from the Loire Valley – are studied separately but by the same approach. Short-listing criteria are used to collate 40 starting-price and middle-range wines for each set differing a priori in olfactory terms. Wine professionals assess their local wines first orthonasally and then, independently, by global evaluation. A pool of descriptive and inferential statistics indicates there is generally neither complete divergence nor real agreement among judges. For Burgundy Pinot noir, the weak agreement observed and measured for orthonasal evaluation strengthens slightly for global evaluation. Contrariwise, for Loire Valley Cabernet franc wines, agreement is poorer for global evaluation than for olfactory evaluation. With orthonasal evaluation, responses are more consistent for Cabernet franc than for Pinot noir, whereas for global evaluation, the levels of interjudge agreement are of the same order for both sets of wines. The subjectivity of sensory responses is interpreted for each situation. The personal judgment of exemplarity (or typicality) may therefore be defined as a demanding cognitive decision varying with the circumstances of the experiment and responding to a process of perceptual categorization based on previous knowledge and on an intuitive comparison between a sample and an abstract but conscious image of the category. Agreement among judges is sufficient to bipolarize the wines. As the power to discriminate among the wines is related to the level of interjudge agreement, bipolarization is most marked for the Cabernet franc wines evaluated by olfactory evaluation.

Introduction

The aim of this paper is to reason on two different categories of red wines: Pinot noir (PN) from Burgundy and Cabernet franc (CF) from the Loire Valley. For each experiment, the purpose is to estimate the level of interjudge (dis)agreement in assessing the exemplarity of 40 wines. For that, different and complementary approaches are considered: Principal Components Analyses (PCA), Kendall's coefficients of concordance (W), Intraclass Correlation Coefficients (ICC) and, to a lesser extent, Eggshell plots. ICCs are used here for the first time for data from a sensory evaluation of exemplarity measurements. That is why several versions of ICC have been tested. Finally this paper also deals with categorization/bipolarization among the wines to produce two extremes of contrasting wines: those which are most and least representative of their respective categories.

As applied to wines belonging to a category that is pre-established on the basis of non-sensory criteria, exemplarity measurement is a specific form of sensory evaluation of typicality. It consists in having wine sector professionals state an opinion holistically about the exemplary character of each of the wines presented. The aim is then to position those wines individually on what is known as an exemplarity scale thereby materializing the sensory distance there might be between one end for wines thought to be the poorest examples and the other end for wines thought to be the best examples of the category under study, that

* Corresponding author. Tel.: +33 (0)3 80 69 35 47; fax: +33 (0)3 80 69 32 27.
E-mail addresses: aurelie.loison@neuf.fr, aurelie.loison-bou-maroun@agrosupdijon.fr (A. Loison).

http://dx.doi.org/10.1016/j.foodqual.2014.10.001
0950-3293/

is, the wines that best exhibit the features and the sensory characteristics of the category in question. This original approach, developed by Ballester, Dacremont, Le Fur, and Etiévant (2005), and then taken up in particular by Parr, Green, White, and Sherlock (2007), Cadot, Caillé, Samson, Barbeau, and Cheynier (2010) and Pineau et al. (2011), leads indirectly to a judgement of the typicality of wines without having to address the concept of typicality head on.

Beyond the sensory characterization of wines based generally on their description, the measurement of exemplarity, and consequently typicality, provides more integrated information but raises real questions within the winemaking sector. Typicality, usually understood as the originality and the reputation of a product from a terroir, includes the product's sensory characteristics, the geographical origin both of the raw material and the transformation process, and its relationship with the social and cultural traditions of the area of production. Typicality is part of the label of origin concept, especially the Protected Designation of Origin (PDO), developed in the EU (Scintu & Piredda, 2007). Typicality relies on the properties of both affiliation and distinctiveness. Moreover, typicality may be one way of restoring consumer confidence. Recent agro-food literature reports growing consumer demand for such products, often as evidence of the emergence of a new rural development paradigm (Barham, 2003; Marsden, Banks, & Bristow, 2000; Murdoch, Marsden, & Banks, 2000; Van Der Ploeg & Renting, 2000; Van Der Ploeg et al., 2000).

In terms of method, typicality has been associated with sensory perception (Cadot, Caillé, Thiollet-Scholtus et al., 2012; Maitre, Symoneaux, Jourjon, & Mehinagic, 2010) with perceptual typicality being characterized by means of conventional sensory profiles and Just About Right (JAR) profiles (Cadot et al., 2010; Fischer, Roth, & Christmann, 1999). To assess the typicality of wines by a holistic approach, another tool, the napping test, has been proposed (Perrin et al., 2008).

The method proposed by Ballester et al. (2005) proved a real breakthrough. Exemplarity measurement can also be used to evaluate the level of agreement among local professionals about what they understand by exemplarity. Following on from Ballester et al. (2005) and Candelon, Ballester, Uscidda, Blanquet, and Le Fur (2004), Ballester, Patris, Symoneaux, and Valentin (2008) and more recently Jaffré et al. (2011) have looked at the fuzzy boundaries of the various sensory spaces associated with the use of certain pure grape varieties (Chardonnay, Sciaccarello and Melon de Bourgogne). These approaches might also prove valuable no longer at the scale of categorical segments based on the varietal but at the finer and more localized scale of the PDO. In this instance, the question of interjudge agreement or disagreement is crucial to determining the extent to which the professionals involved, as joint owners of the PDO, form a coherent and consistent community of thought, without necessarily being unanimous.

Costa-Santos, Antunes, Souto, and Bernardes (2010) claim there is no failsafe strategy for estimating interjudge (dis)agreement. Various complementary approaches have been used to evaluate the degree of agreement among judges. The use of Kendall's coefficient of concordance (W) has already been reported on sensory evaluations with judges rating alcoholic beverages especially (Ballester et al., 2008; Candelon et al., 2004; Le Fur et al., 2013; McDonnell, Hulin-Bertaud, Sheehan, & Delahunty, 2001; Schucany & Frawley, 1973). W was between 0.03 (Ballester et al., 2008) and 0.62 (Rodríguez, Albertengo, Vitale, & Agulló, 2003). Schmidt (1997) and McDonnell et al. (2001) argue that W values of less than 0.3 represent weak agreement. This cut-off point seems arbitrary and no claim is made that it is a universal threshold. Gisev, Bell, and Chen (2012) emphasize that it is increasingly hard to achieve high Ws as the number of judges increases. Ballester et al. (2008) describe a situation of real divergence among judges (white wines, orthonasal evaluation, 17 novice judges) since Ws ranged from 0.03 to 0.09. Finally, because Ws assume ordinal data, the interpretation of such results leads to a reduction in statistical power because information about the magnitude of the exemplarity scores is lost. Kottner et al. (2011) recommend calculating Intraclass Correlation Coefficients (ICCs) in addition to Ws for continuous measurements. The choice between ICC single and ICC average remains controversial (Shrout & Fleiss, 1979) and should be dictated primarily by the design of data. In the event that judges evaluate each sample once only, the ICC single seems more appropriate. Cicchetti and Sparrow (1981) seem to base their reasoning on the ICC single (A,1) to classify levels of ICC with: below 0.40: poor agreement; 0.40–0.59: fair; 0.60–0.74: good; and 0.75 and above: excellent agreement. This scale is consistent with the value of 0.75 that Burdock, Fleiss, and Hardesty (1963) proposed as reflecting a high level of agreement among raters. But what would reflect disagreement? For Giraudeau (1996) as for Costa-Santos et al. (2010), ICC ranges from 0 (no agreement) to 1 (perfect agreement). However, some authors evoke the possibility of obtaining ICCs of between −1 and 1 (Costa-Santos et al., 2010). In their work Cliff and King (1997) mentioned the usefulness of eggshell plots for wine competitions (large numbers of wines). Eggshell plots can be used to illustrate each judge's performance in ranking the wines. This representation has several advantages although it cannot substitute for the approaches described above. First of all, although the eggshell plot is a graphic technique based on ranks instead of using exemplarity scores directly, it provides a way of graphically evaluating interjudge agreement in discriminating among the wines (Hirst & Næs, 1994; Næs, 1998). Lastly, this representation provides a quick idea of individual performances and could, where required, identify those judges who deviate more than others from the consensus ranking. This method might also have the benefit of identifying those judges thought of as outliers. As such, another method for identifying outliers was proposed by Schlich (1994), then used by Parr, Green, and Geoffrey White (2006) and more recently by Urdapilleta, Parr, Dacremont, and Green (2011). The program is based on individual and global analyses of variance and summarizes results in a series of graphical judge scatterplots. However, that program was specifically developed for sensory profiling with repeat sessions during which all the judges give scores to all the products for a number of attributes.

Material and methods

Wines

Two sets of 40 French red wines were selected, one of Burgundy wines and one of Loire Valley wines. The aim was to bring together wines that reflect as accurately as possible the olfactory diversity on offer commercially. The main factors allowed for were (i) vintage (2009 and 2010), (ii) PDO, (iii) price, and (iv) human factor. Other sources of variability were considered: winegrowing area, winemaker's status, growing and winemaking practices. Excessively woody wines were excluded from the selection. The two sets of 40 wines were selected separately and by two slightly different approaches.

Burgundy wines

The wines were produced from pure PN. Only wines with regional and certain village PDOs within a price range of €4–€12 (average price €7) per bottle excluding taxes were selected. The number of wines from each PDO was proportional to its production volume, so PDOs were not equally represented. Some 28 wines out of 40 were from the 2009 vintage. The other 12 were from the 2010 vintage. The producers of the short-listed wines were interviewed to secure full information as to whether the wines complied with