3rd International Conference on Advanced Research Methods and Analytics (CARMA2020)
Universitat Politècnica de València, València, 2020
DOI: http://dx.doi.org/10.4995/CARMA2020.2020.11647

Proposal of a composite indicator for measuring social media presence in the wine market

Andrea Conchado Peiró[1], José Miguel Carot Sierra[1], Elena Vázquez Barrachina[1], Enrique Orduña-Malea[2]

[1]Department of Applied Statistics and Operational Research, and Quality, Universitat Politècnica de València, Spain, [2]Department of Audiovisual Communication, Documentation and History of Art, Universitat Politècnica de València, Spain.

Abstract

Cybermetrics field is attracting considerable interest due to its utility as a data-oriented technique for research, though it may provide misleading information when used in complex systems. This paper outlines a new approach to market research analysis through the definition of composite indicators for cybermetrics, applied to the Spanish wine market. Our findings show that the majority of cellars were present in only one or two social media networks: Facebook, Twitter or both. Besides, the presence on the Web can be summarized into three principal components: website quality, presence on Facebook, and presence on Twitter. Three groups of cellars were identified according to their position in these components: cellars with a high number of errors in their website with complete absence of information in social media, cellars with strong presence in social media, and cellars in an intermediate position. Our results constitute an excellent initial step towards the definition of a methodology for building composite indicators in cybermetrics. From a practical approach, these indicators may encourage cellar managers to make better decisions towards their transition to the digital market.

Keywords: *Composite indicator; principal component analysis; wine; cybermetrics; social media.*

1. Introduction

There is evidence of a growing interest in research related to the impact of new products, brands and firms through their websites and social media profiles (Orduna-Malea and Alonso-Arroyo, 2017). Most of these studies have been traditionally based on link analysis or Search Engine Optimization (SEO) techniques, usually with a marketing-oriented approach. However, the majority of these analyses do not follow empirically tested or validated methodologies by the scientific community. On the contrary, they are mainly focused on narrow preestablished objectives or specific case studies, such as the analysis of a unique brand, biased studies skipping other potential emerging competitors or qualitative analysis with non-representative customer profiles.

Cybermetrics is the field that studies how to build and use web resources, structures and technologies (Björneborn e Ingwersen, 2004). Quantitative advances in this area are mostly addressed to solve research problems in social sciences through the application of quantitative research methods (Thelwall, 2009). The huge dependence on the availabily and variability of both metrics and sources has jeopardized the use of Cybermetrics (Thelwall, 2010), mostly limited to link analysis. However, this particular technique is insufficient when it comes to analyse complex environments, which cannot be describe with simple metrics.

Recent developments have led to the design and assessment of measurement systems composed by different indicators, with the aim of comparing results within different units of analysis (Orduna-Malea and Alonso-Arroyo, 2017). These measurement tools are based on composite indicators (or indexes), which are defined as the combination or mathematical aggregation of a set of simple indicators aiming to summarize a multidimensional concept into a simple or one – dimensional index, based on an underlying theoretical model (Nardo et al., 2008). This system of composite indicators represents a sound procedure for measuring websites' impact on the basis of its ability to identify profiles and trends as regards to the supply (contents generated by companies) and demand (contents searched by users) in particular economic sectors. This method might help to provide useful information for the commercialization of new products. This work precisely aims to apply this method to the Spanish wine sector.

Recent evidences show that the international wine market can be divided into two clusters: countries with a traditional approach as regards wine commercialization, like France, Italy and Spain (Old World), and countries with stronger presence in online international markets, such as United States, Argentina, Chile, South Africa and New Zeland (New World). Following that logic, the presence of Spanish cellars both on the Web and social media channels is expected to be quite low, as previous findings have evidenced (Compés and Castillo, 2014).

In the light of these events, the main objective of this paper is to outline a new approach to market research analysis through the definition of Cybermetric composite indicators, to be applied to the Spanish wine market.

This work is based on the experience of the authors as the coordinator team of the research funded project: 'eMarketwine – Design of a method and an online tool for information intelligence addressed to geolocalized recommendations in the wine industry' (Ref. CSO2016-78775-R). The Spanish wine sector (including only bottled wine) was chosen because it is one of the most relevant sectors of the Spanish industry. This project follows on a previous project, 'Trademetrics – Methodological proposal of a cybermetric analysis of products, branches, people and firms in the Spanish online market' (Ref. CSO2013-46138-P), also funded by the Ministry of Economy and Competitiveness.

2. Methodology

2.1. Data

The sample consisted of 3,164 Spanish cellars, collected between 2018 and 2019. As a result, 23 variables were collected, of which seven were quantitative variables (metrics), and the rest were qualitative variables (Yes / No).

Metrics gathered are divided into two main categories. First, metrics related to the technical quality of cellars' websites. These data were obtained from W3C Markup Validation Service and Link Checker (https://validator.w3.org). Second, metrics related to the activity of cellars on Twitter and Facebook. These data was extracted via API. The type, description and nature of all the variables considered for each cellar are included in Table 1.

Table 1. Type, label, description and nature of variables analysed.

Type	Variable	Description	Nature
Identification	Id	Numeric code	Numeric
	Name	Name of cellar	Character
Characteristics	Zip code	Location	Character
	D.O.	Denomination of origin	Qualitative
Metrics	broken_links	Number of broken links	Quantitative (integer)
	html_errors	Number of HTML error codes	
	html_warnings	Number of HTML warnings	
	Likes	Number of Likes in Facebook	
	fb_followers	Number of Facebook followers	
	tw_posts	Number of posts in Twitter	
	tw_followers	Number of Twitter followers	
Social media presence	Facebook	The cellar has a profile on …	Qualitative (Yes/No)
	Twitter		
	Linkedin		
	Pinterest		
	Flickr		
	Youtube		
	Instagram		

2.2. Methods for building composite indicators

Principal Component Analysis (PCA) was chosen to analyse the seven quantitative variables, since it is one of the most practical ways to measure constructs that cannot be directly measured (so–called latent variables). This technique is useful for understanding the structure of a set of variables, such as the set quantitative metrics considered. Besides, it provides a tool for reducing this data set to a more manageable size, while retaining as much of the original information as possible.

This data analysis technique is concerned with establishing the underlying linear components that exist within the data, and how much each variable contributes to each component. With this aim, the eigenvalues of the correlation matrix representating the relationship between

variables are calculated. These eigenvalues are subsequently used to calculate eigenvectors in such a way that eigenvalues represente a measure of the substantive importance of the associated eigenvector. According to Kaiser (1960), only components with eigenvalues greater than 1 should be retained. This simple rule of thumb has generated a high number of recommendations about the appropriate number of components to retain. Once factors have been extracted, factor rotation procedures rotate factor axes such that variables are loaded maximally on only one factor. Among available rotation procedures, we can chose between orthogonal methods (including varimax, equamax or quartimax procedures) or oblique rotation, depending on whether components are allowed to correlate.

Log–transformation was needed for dealing with the highly skewed data contained in metrics. Next, normalization procedures were applied in order to guarantee the robustness of the analysis against the presence of outliers in the data (Ebert & Welsch, 2004). Thanks to this transformation, the composite indicator was comparable across different groups. Among all available methods, min–max normalization was chosen due to the simplicity of interpretation of the resulting composite indicator.

3. Results and discussion

3.1. Quality of data in metrics (technical website quality and social media activity)

As expected, Facebook and Twitter were the social media networks with a stronger presence of Spanish cellars. Figure 1 shows that other networks had neglible presence of cellars compared to this results, like Flickr, Pinterest or Linkedin. The majority of cellars were present in only one or two social media networks.

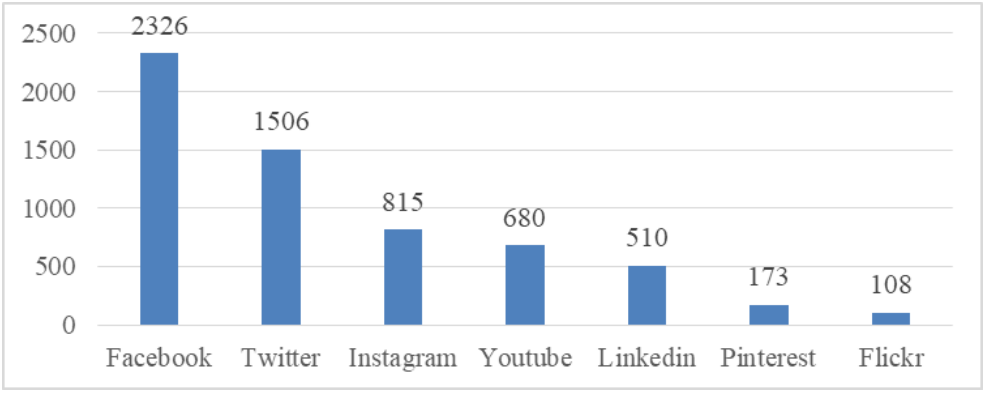

Figure 1. Number of cellars present in each social networking site.

There was a high level of heterogeneity among data metrics concerning technical quality of the website, intensity or volume of use in Facebook and Twitter (Table 2). There was a high

level of correlation between the number of Facebook's likes and Facebook's followers (r ~ 1), and a moderate level of correlation between the number of Twitter's posts and Facebook's followers (r = 0.67). Due to the high number of outliers with a high number of followers both in Facebook and Twitter, data metrics presented high level of asymmetry and kurtosis.

Table 2. Quality of data metrics.

Parameter	**Website**			**Social media**			
		Technical quality		**Facebook**		**Twitter**	
	Broken links	**HTML error**	**HTML warnings**	**Likes**	**Followers**	**Likes**	**Followers**
Valid (%)	90%	93%	93%	60%	60%	46%	46%
Min	0	0	0	0	0	0	0
Max	197	790	377	381,000	381,000	38,717	24,017
Mean	8	24	24	2,632	2,677	1,260	1,337
SD	13	54	28	14,249	14,266	2,564	2,480
Asim Std.	5	6	3	20	20	7	4
Kurt Std.	49	57	22	458	456	66	24

Among the data set composed of 3,164 cellars, we selected those cellars with valid information in data metrics (N = 3,052) and presence in social media networks (N = 2,478). Next, we applied PCA to the data set of seven metrics, with varimax rotation procedures and pairwise selection methods for missing data (KMO = 0.57, Barlett's test $\boldsymbol{X^2 = 12.487, p = 0.000}$). As a result, three principal components were extracted which explained 72.5% of variability contained within the initial data set.

Table 2. Quality of data metrics.

Metrics	PC1	PC2	PC3
broken_links	0.006	0.168	0.427
html_errors	-0.005	-0.035	0.730
html_warnings	0.003	-0.044	0.789
fb_likes	0.987	0.157	0.001
fb_followers	0.987	0.159	0.001
tw_ followers	0.130	0.907	-0.009
tw_post	0.175	0.881	0.117

According to the loading scores obtained through PCA, the first principal component (PC1) was labelled as 'Presence in Facebook', PC2 was labelled as 'Presence in Twitter' and PC3 was labelled as 'Website's technical quality'. Consistently with the correlation coefficient, both Likes in Facebook and Followers in Facebook showed similar values for their loading scores, whereas Followers in Twitter presented a higher contribution in its corresponding principal component. As regards to the third component, websites' technical quality, the number of broken links presented the lowest contribution to these latent variables, as evidenced in the low value of communality (0.189). All other values of communalities were higher than 0.5.

Thus, the findings using PCA confirm the strength of the association between similar variables concerning these three principal components. According to these results, it can be concluded that each principal component can be summarized through its scores. However, observed variables are preferred for building composite indicators, rather than latent variables, as those obtained with PCA. Thus, accordingly we decided to select the number of followers in Facebook and Twitter to build the composite indicator. Both variables presented high values of communalities resulting from the extraction of the previous three components. Besides, both were referred to people using different social media which resulted in higher values of internal consistency (4 items, Cronbach's $\alpha = 0.714$) than the solution including items concerning websites' technical quality (7 items, Cronbach's $\alpha = 0.649$), despite the low number of items.

As previously shown, variables about presence in Facebook and Twietter were highly skewed due to the presence of outliers. Thus, log–transformation of data was applied in order to deal with these non–normal distributions. Additionally, min-max normalization was also used for building a composite indicator with two components, each of them weighted using a rank between 0 and 50. This latter decision was taken as a practical way to generate an indicator

with a range of variation of 100 points, through the sum of two different components with independent ranges from 0 to 50. This method of weighting was chosen because it was the most practical ways to create a meaningful indicator to be used and applied by managers and steering committees of commercial wine cellars. Though many other methods of weigthing are available for composite indicators, researchers have not reached an agreement about the criteria to priorize among them. Thus, we considered different solutions and finally decided to follow this procedure on account of its feasibility for this environment.

Thus, the resulting formula for our proposal of a composite indicator measuring presence in social media networks is the following:

$$\boldsymbol{IFT_i = 50\frac{LogF_{fi}}{\max_i(LogF_{fi})} + 50\frac{LogT_{fi}}{\max_i(LogT_{fi})}}$$

Where:

F_f= Followers on Facebook

T_f= Followers on Twitter

The graphical representation of the IFT_i score of all cellars in this composite indicator in a scatterplot with websites' technical quality allows the reader to identify three groups of cellars (Figure 2). First, cellars with a high number of errors in their website with complete absence of information in social media. These non-digital companies have been marked in red. Second, and on the opposite side, there is another group of cellars with a strong presence in social media. The majority of them also have high quality websites (in terms of technical developement), though some outliers might consider to invest in the technical improvement of their websites. Third, in an intermediate position, we find cellars with moderate presence in social media and different levels of technical quality in their websites.

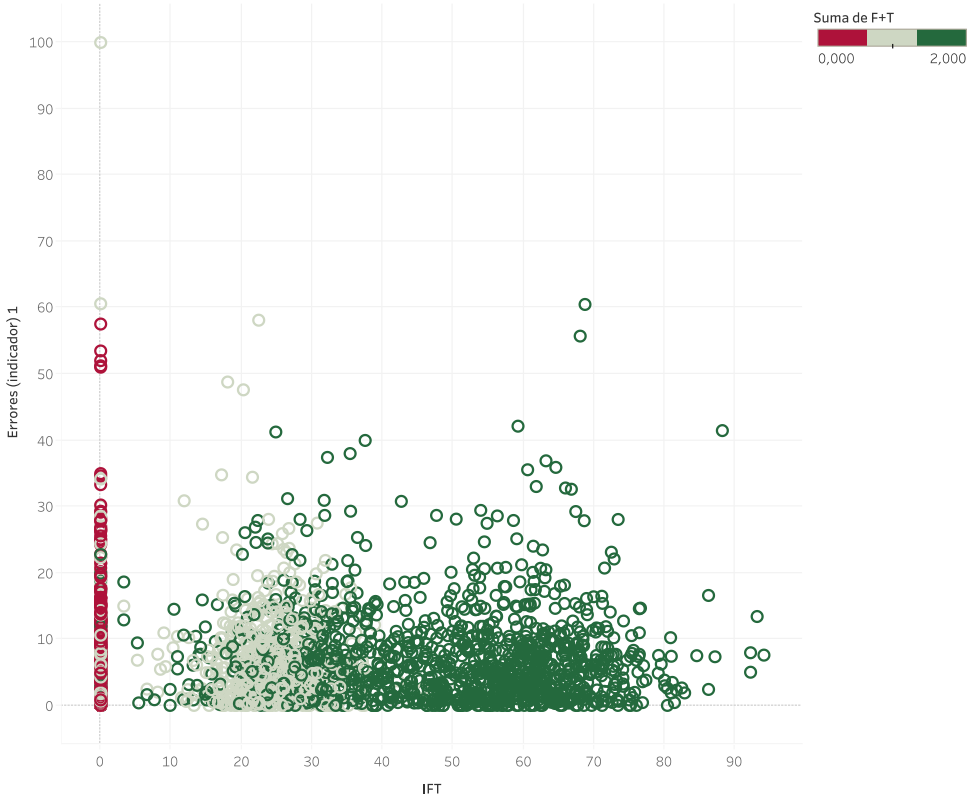

Figure 2. Scatterplot of IFT versus Websites' technical quality.

4. Conclusions

This study has highlighted the importance of composite indicators in cybermetrics when dealing with highly skewed data and the need to apply specific methods of data analysis for extracting useful information in decision making. We have devised a methodology which can be useful in other research areas concerning cybermetrics. This work has some limitations, such as the reduced number of metrics involved in the analysis. Research is underway to overcome this shortcoming. However, in our view, these results constitute an excellent initial step towards the definition of a methodology for building composite indicators in cybermetrics.

Specifically, results have evidenced that activity in social networking sites is not strongly related to the websites' technical quality for Spanish cellars. This result is unexpected as it could be assumed a priori that top companies would be good elsewhere, either creating websites without errors or showing high activity in the social media channels. Otherwise, the presence of Spanish cellars has been shown to be concentrated in Facebook and Twitter.

Moreover, the number of followers in these two social networking sites is enough to explain Spanish cellars' impact variability on these online spaces.

These results have important managenerial implications for cellars' managers, who can monitor the market activity on social media, and systematically carry out benchmarking analyses using robust statistical methods. Likewise, results can be used for advisory and consultancy activities oriented to strategic decision making and online marketing. The future inclusion of aggregated data, such as denomination of origin or region, will expand the capabilities of the system.

Acknowledgements

This work was carried out within the framework of a Spanish research project 'eMarketwine: diseño de un método y una herramienta de online information intelligence orientada a la recomendación geolocalizada para el mercado del vino' (Ref. CSO2016-78775-R), founded by the Ministerio de Economía y Competitivad (Spanish Ministry of Economy and Competitiveness).

References

Björneborn, L., & Ingwersen, P. (2004). Toward a basic framework for webometrics. *Jasist*, 55(14), 1216-1227.

Compés López, R., & Castillo Valero, J.S. (2014). *La economía del vino en España y en el mundo*. Alicante: Cajamar Caja Rural.

Ebert, U., & Welsch, H. (2004). Meaningful environmental indices: a social choice approach. *Journal of Environmental Economics and Management*, 47(2), 270-283.

Feng, C., Wang, H., Lu, N., Chen, T., He, H., Lu, Y., & Tu, X. M. (2014). Log-transformation and its implications for data analysis. *Shanghai archives of psychiatry*, 26(2), 105–109. https://doi.org/10.3969/j.issn.1002-0829.2014.02.009

Kaiser, H. F. (1960). The application of electronic computers to factor analysis. *Educational and psychological measurement*, 20(1), 141-151.

Nardo, M., Saisana, M., Saltelli, A., Tarantola, S., Hoffman, A., & Giovannini, E. (2008). *Handbook on constructing composite indicators: methodology and user guide*, OECD Statistics Working Paper, STD/DOC (2005)3, OECD Publishing, Paris.

Orduna-Malea, E., & Alonso-Arroyo, A. (2017). *Cybermetric techniques to evaluate organizations using web-based data*. Cambridge: Elsevier.

Thelwall, M. (2009). *Introduction to webometrics: Quantitative web research for the social sciences*. San Rafael, CA: Morgan Claypool.

Thelwall, M. (2010). Webometrics: Emergent or Doomed?. *Information Research*, 15(4), 1-10.