

Statistical data analysis incorporating web text mining to establish correlations between grape wine taster comments and wine ratings

SUBANA SHANMUGANATHAN¹, PHILIP SALLIS¹ AND AJIT NARAYANAN²
¹*Geoinformatics Research Centre, ²School of Computing and mathematical Sciences,*
Auckland University of Technology

Private Bag 92006, Auckland 1142

NEW ZEALAND

Subana.shanmuganathan@aut.ac.nz : www.geo-informatics.org/

Abstract - *Wine taster comments and ratings provided to aid potential wine buyers are a useful source for modelling the effects of short term climate variations on wine quality. The paper describes how web text mining and statistical data analysis methodologies can be applied to studying the correlations between taster comments and ratings (or liking scores) of 1540 wines from New Zealand. With this new information on the correlations, it is possible to model the wine vintage-to-vintage variations for a wider project that is aimed at modelling the inter dependencies between grape wine sensory properties and “terrior” attributes (climate, soil and terrain) as well as “cultivars” (varieties) with data (historic) to predict grape vine varieties for short (micro) and long (macro) term future climate change scenarios.*

Keywords: WEBSOM

1 Introduction

Wine comments in free text format and ratings (numeric data) consist of valuable information for potential wine buyers as well as researchers. These comments and other structured data relating to the vintage of 1540 New Zealand wines being studied herein are from (www.wineenthusiast.com), a web magazine called ‘Wine Enthusiasts’ The wine comments are posted on the Internet by Master Sommeliers (famous wine tasters/ experts) based on wine sensory properties (such as ortho-, retro- nasal and taste), to aid consumers buying wine online, and these comments can be used to analyse short term climate/vintage-to-vintage variations in wine quality.

The work discussed in this paper is part of a wider research programme that is aimed at modelling the correlations between wine sensory properties and “terrior” attributes (such as climate i.e., weather conditions, environment, i.e., soil, terrain) and “cultivars” (varieties) for successful grape growing and winemaking. The overall aim of the overarching research project is to build computational models to further our knowledge on

the effects of various relevant and dependent factors that play a major role in grapevine phenology and thereby on wine quality. The models are built using data from historical records, remote weather monitoring devices and plant response sensors. The model dependencies established between relevant factors, such as climatic, and environmental conditions, plant growth and wine quality will then be applied to predicting grapevine varieties for future long and short term climate variation scenarios. See [1], [2] for more details on the main project concept.

1.1 Climate effects on grapes and winemaking

Grapevine growing for winemaking is limited to narrow geographical zones that provide the most favourable environmental conditions needed for ripening so-called “ideal” quality grapes i.e., with optimum sugar without compromising colour, aroma and flavour component proteins. Of the factors that are categorised under “terrior”, arguably climate exerts the major influence on a region’s or site’s ability to produce quality grapes and in turn finer wine in that particular appellation. While the average climate of a wine region determines the grapevine variety (for producing the wine appellation) that has been known to be suitable for the region, vintage-to-vintage wine production and quality are to a greater extent influenced by the factors that are site specific, viticulture practices and short term climate variability [3]. Based on this general notion the paper looks at the approaches investigated to analysing the vintage-to-vintage variations in wine sensorial properties (or descriptors) in text and ratings in numeric formats that could be ultimately extended for modelling the correlations between weather conditions that portray short term climate variations of a wine region or site.

1.2 Text mining the wine expert comments

Limited literature on the use of qualitative descriptors on wine sensory properties (extracted from expert comments) reveals that conventional data analysis and

mappings do not well reflect the correlations between the descriptors and their ratings. Of these work major approaches are outlined in section 2. The next section elaborates on the methodology and results achieved with web text mining (WEBSOM approach is based on an unsupervised neural network called Kohonen nets) [4], [5], and statistical methodologies for establishing the relationships between wine descriptors and ratings in this research. Section 4 consists of a discussion on the methods so far investigated and the paper concludes with major directions for future work.

2 Analysing wine taster comments

The following is a brief outline on literature reviewed on studying the correlations between wine descriptors and ratings made by wine tasters for potential online buyers

2.1 Rating of wines through scores and free-text assertions

In [6], by applying principal component analysis (PCA) authors created a wines x scores table, a low dimensional representation with preference mapping and synthetic scores formulated for wines analysed. In this study authors used wine descriptor words extracted from taster comments with conventional text mining approach to establish the correlation between free text comments and synthetic scores. This study demonstrated an interesting approach to overcome the issues relating to collectively analysing qualitative (in free text) and quantitative data.

2.2 Relationships between liking ratings and wine sensory descriptors

In another study [7] researchers used pairs of liking ratings and the intensity of 14 wine descriptors originally collected for statistical analysis (partial and least square regression) to create a second map with liking ratings on y axis and the sensory descriptive data of the wines on x axis. The model only accounted for 25% of the variation, described to be very low, hence the authors cautioned readers on the use of their results that failed to explain the driving factors for the remaining 75% of the variation, a majority of the wine taste data sample obtained from a trained panel to rate the intensity of each 14 wines used.

3 Methodology and data

This work is an extension of the use of web text mining approach investigated in [8], [9] with 95 New Zealand wines for analysing wine descriptors extracted from sommelier comments posted in an online web magazine called "Wine Enthusiast" [10]. In this paper, 1540 New Zealand wine comments from the same source are used for comparing the WEBSOM results with that of conventional statistical analyses. The following are the major steps adopted to produce a matrix of meaningful wine descriptor (weighted) based on their presence in the sommelier comments:

- 1) pre-process web text to a text file,

- 2) remove stop words
- 3) reduce verbs to lemmas (verb base) using standard a stemming algorithm i.e., porter stemmer.
- 4) discard very rare and common words based on wine chart of words listed in [11]
- 5) prepare wine descriptors x weight matrix based on the occurrence of each descriptor chosen, based on its presence in each wine comment and the collection of comments. The weighted wine descriptor frequency matrix is calculated using the well-known Salton vector space model based on formula (1) [12].

$$w_i = tf_i * \log\left(\frac{D}{df_i}\right) \quad (1)$$

Where,

tf_i = term frequency (counts) or number of times a term i occurs in a document.

df_i = document frequency or number of documents containing term i

D = number of documents in the collection/database.

- 6) merge text and numeric data into one table

At the end of step 3 the wine descriptor matrix had 3206 with misspelt words and step 4 brought down the matrix size to 234 wine descriptors.

3.1 WEBSOM approach

Using a SOM, the 234 wine descriptors are grouped into 20 clusters based on their weighted co-occurrence frequencies (Figure 1). The profiles of these wine descriptor groups (clusters) are analysed to model the vintage-to-vintage variations and are discussed herein.

3.1.1 Pinot Noir of Canterbury (1998-2004)

Ten Pinot Noir wines of Canterbury produced from 1998 - 2004 are analysed using the SOM components (Figure 2) as well as a graph (Figure 3) to see how any correlation between vintage-to-vintage wine descriptor frequencies and liking ratings could be modelled using the WEBSOM approach. The findings of this analysis are:

- 1) From the graph (Figure 3), it is clear that the presence of both clusters C3 and C4 descriptors raise the ratings to 86-87 in 3 of the 4 high rated wines in this category.
- 2) The presence of C4 descriptors alone increases ratings but not with that of C3 alone. C9, C13 and C15 descriptors are present in both high (86) and low (80) ratings. Hence, their presence seems to be "non contributory" in the rating scores.
- 3) Wine descriptors of C9 (*soft, solid, veget, winemak, bai, bottl and herbac*), C12 (*gooseberri, acid bright fruiti intens*) and C16 (*pear, chardonnai, oak, peach, pineappl, spice, vanilla, butter and toast*) certainly have negative influence in the ratings in this class.
- 4) C 10 (*firm and structur*) and C17 (*full*) have positive influence on the ratings.

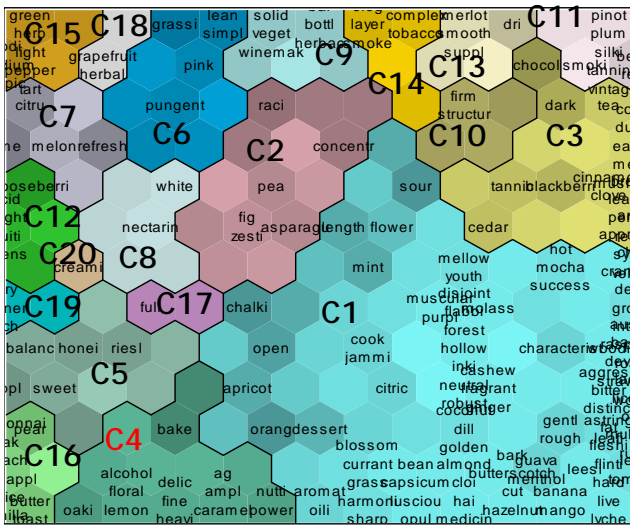


Figure 1: SOM of 234 weighted wine descriptor frequencies in 1540 New Zealand wines being analysed.

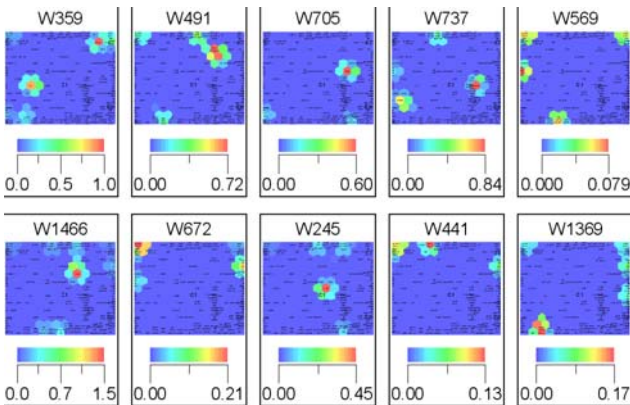


Figure 2. SOM components of wine descriptors in Pinot Noir wines from Canterbury, New Zealand (1998-2004).

C 1: sour, length, flower, mint, mellow youth, hot mocha success, chalki, muscular purpl, molass, anis approach char cranberri deep group integr raspberri roast strawberri wood, open, cook jammi, disjoint flabbi forest hollow inki neutral robust, characterist, woodi, apricot, citric, cashew fragrant ginger, auster bacon develop lactic licoric oliv rhubarb rubi, orang, dessert, coconut dill golden, gentl rough, astring leafi, guava menthol, leesi, leaf tomato, aromat oili round, blossom currant grass harmoni sharp thick ting warm, bean capsicum luscious opul sweati viscou, almond cloi hai medicin quinc syrupi, bark butterscotch cut hazelnut slight thin tree, banana mango pure, aggress bitter distinct fat fleshi flinti hard live lyche perfum petal rose steeli tight variet young **C 2:** raci, concentr, pea, fig zesti, asparagu **C 3:** chocol, dark, tea, coffe dusti earthi meati mushroom, tannic, blackberri, cinnamon clove, cedar, brown leather persist readi syrah velveti **C 4:** bake, oaki, alcohol floral lemon linger spici, delic fine heavi modest, ag ampl caramel gri subtl, nutti power strong **C 5:** balanc, honei, riesl, appl, sweet **C 6:** grassi, lean simpl, pink, pungent **C 7:** citru, lime, melon, refresh **C 8:** white, nectarin **C 9:** soft solid veget winemak, bai bottl herbac **C 10:** firm structur **C 11:** black cola noir pinot plum silki tannin vintag, smoki, berri red **C 12:** gooseberri, acid bright fruiti intens **C 13:** cabernet merlot smooth suppl, dri **C 14:** eleg layer smoke, complex tobacco **C 15:** clean crisp fresh green herb light pepper tart, bodi medium tropic **C 16:** pear, chardonnai oak peach pineappl spice vanilla, butter toast **C 17:** full **C 18:** grapefruit herbal **C 19:** dry miner rich **C 20:** creami

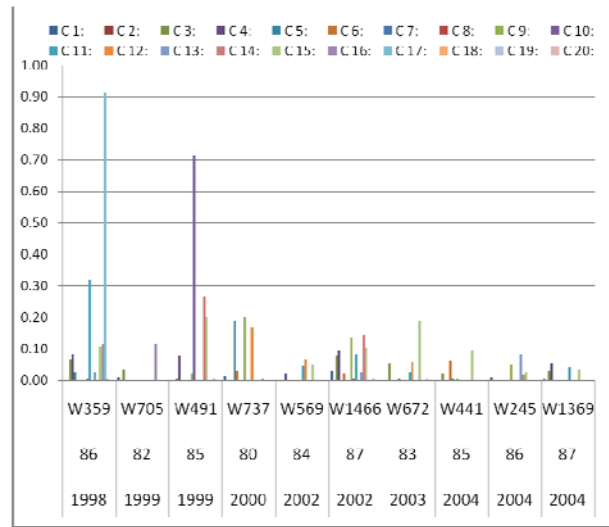


Figure 3. Graph showing average values of wine descriptor frequencies (clusters 1-20) in Canterbury, New Zealand Pinot Noir wines of 1998-2004

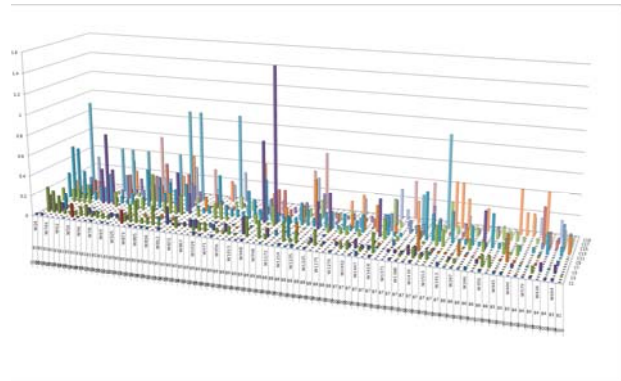


Figure 4. Graph showing average frequencies of wine descriptors extracted from sommelier comments made for Pinot Noir (central Otago NZ 1998-2008). On x axis from left to right, are wines from high to low ratings in descending order. Higher ratings contain more of C3 and C4 and low rated have more of C18 and less of C3 & C4.

3.1.2 Pinot Noir (PN) of Central Otago (1998-2006)

In graph (Figure 4) of 84 Pinot Noir wines of Central Otago (from 1998 to 2008) the presence of certain wine descriptors clearly depicts the low and high ratings.

- 1) In this graph, the wines with low ratings (81-87 on the right of the graph) do not consist of C3 or C4. The above wines as well have higher average values for C18 (*grapefruit and herbal*) descriptors implying as having negative influence on the wine taste.
- 2) Furthermore, C10 (*firm and structur*) and C17 (*full*) descriptors are hardly used for this rating range.
- 3) Mid range wine of the category do have a mix of both descriptors.

Conventional statistical data analysis methodologies are carried out to further refine and verify the results obtained with the WEBSOM approach and are elaborated upon in the next section.

3.2 Statistical data analysis

The results of discriminant and linear regression analysis thus far conducted in this research show promising results in predicting the ratings within the mid range.

3.2.1 Discriminant analysis results

Stepwise discriminant analysis ran with the 234 wine descriptors for the 1540 New Zealand wines produced 11 steps with 0.564 Wilk's Lamda for 11th word (*full*). Classification results of the linear regression showed 15 rating classes (80 - 94) and cross validation done with the 11 functions classified 19.8 % correctly for grouped cases of ratings with 88-91.

3.2.2 Regression analysis

Linear regression ran on the wine descriptor matrix produced 66 models and the final model consists of 67 predictors. Of these wine descriptors "*rich*" has the highest positive correlation and "*tart*" highest negative.

4 Discussion

WEBSOM approach to modelling wine descriptors is seen to be useful in establishing the correlations between descriptors and ratings. Graphs of wine descriptor group profiles (of WEBSOM) in Pinot Noir of both Canterbury and Central Otago show clear distinction between the comments given for high and low rated wines. However, statistical methodologies, discriminant and linear regression, thus far investigated show the wine descriptors consist of positive and negative correlation to liking ratings and vintage-to-vintage variations but with very low percentages and only for the ratings within mid range.

5 Conclusions

Both WEBSOM and statistical analysis results show that wine comments do have a correlation with ratings as well as vintage-to-vintage variability but further research is warranted on validating the results. Hence, it is concluded that further analysis is carried out using the same techniques and wine comments but with wine comments be analysed separately with respect to their region and style or appellation.

References

[1] Sallis, P.J., Shanmuganathan, S., Pavesi, and L., and Jarur, M. A system architecture for collaborative environmental modelling research. The 2008 International Symposium on Collaborative Technologies and Systems (CTS 2008), Eds., Waleed W. Samari and William McQuay, A publication of the IEEE, New Jersey, USA. ISBN: 978-1-4244-2248-7, Irvine, California, pp 39-47. May 19-23 2008

[2] Shanmuganathan, S., Sallis, P.J., Pavesi, L. and Jarur, M. Computational intelligence and geo-informatics in viticulture, in proceedings of the Second Asia

International Conference on Modelling & Simulation. CD version published by IEEE computer society. Eds., Edited by David Al-Dabass, Steve Turner, Gary Tan and Ajith Abraham Kuala Lumpur, Malaysia, pp 480-485. 13-15 May 2008

[3] Gregory Jones, "How Hot Is To Hot?" Wine Business Monthly, pp.1-4, February 2005.

[4] Honkela, T., Kaski, S., Lagus, K., and Kohonen, T., Newsgroup exploration with WEBSOM method and browsing interface. Technical Report A32, Helsinki University of Technology, Laboratory of Computer and Information Science, Espoo. (1996)

[5] Lagus, K., Honkela, T., Kaski, S., Kohonen, T., (1996) WEBSOM - A Status Report, in proceedings of STeP'96. Jarmo Alander, Timo Honkela and Matti Jakobsson (eds.), Publications of the Finnish Artificial Intelligence Society, pp. 73-78.

[6] Mo'nica Be'cue-Bertaut, Ramo'n A' lvarez-Esteban and Je'rome Page's (2008) Rating of products through scores and free-text assertions: Comparing and combining both in Food Quality and Preference 19 122-134 pp. 2008.

[7] Frøst, M. B., Noble, (2002) A., Preliminary study of the effect of knowledge and sensory expertise on liking for red wines. American Journal of Enology and Viticulture. vol. 53(4) pp. 275-284.

[8] Sallis, P.J., Shanmuganathan, S., Pavesi, L., and Jarur, M. (2008) Kohonen Self-organising maps in the mining data mining of wine taster comments. Data Mining IX, Data Mining, Protection, Detection and other Security Technologies 2008. Cadiz, Spain, 26-28 May 2008. Eds., A Zanasi, D Almorza Gomar, N F F Ebecken and C A Brebbia. ISBN: 978-1-84564-110-8, ISSN (print): 1746-4463, ISSN (on-line): 1743-3517 Transactions on information and Communication Technologies, Vol. WIT press. 40 pp 125-139.

[9] Shanmuganathan, S., and Sallis P., Modelling climate change effects on wine quality based on expert opinions expressed in free-text format: the WEBSOM approach for publication in proceedings of 15th International Conference on Neural Information Processing of the Asia-Pacific Neural Network Assembly (ICONIP 2008) Nov 25-28, 2008, Auckland, New Zealand pp265-266

[10] www.winemag.com/buyingguide/search.asp?db=

[11] Campbell Bob 2000, "Essential Wine Tasting Guide" 2000.

[12] Salton, Gerard. Introduction to Modern Information Retrieval. McGraw-Hill, 1983.