O2 CHEM1

LINKING LINGUISTICS AND CHEMISTRY BY CHEMOMETRICS

M. Cocchi[1], N. Cavallini[2], R. Bro[3]

[1]*Dipartimento di Scienze Chimiche e Geologiche, Università di Modena e Reggio, Modena, Italy*
[2]*Department of Applied Science and Technology, Politecnico di Torino, Torino Italy*
[3]*Department of Food Science, University of Copenhagen, Frederiksberg C, Denmark*

In the last decades, there has been a great development of data analysis tools suitable to cope with heterogeneous data, i.e. data which come from distinct sources and differ not only in their scales and values but as well in their structure, despite they relate to the same phenomena. A salient example is the integration of text description with numerical data for the same entities. This task requires, as a first step, extracting information from text documents and converting it to a suitable data format, which can then be handled by data analysis tools.
In the present work, we propose a Chemometrics data analysis pipeline to link text data with analytical data and we evaluate it in the applicative context of food consumption and consumer preference/expectation.
The consumers' interest in how food is produced and prepared has recently strongly increased. Consumers tend nowadays to be more aware about the different aspects regarding food consumption and, in line with this trend, new-concept restaurants, new food production techniques and experiments on recipes and food pairings are constantly developed. This phenomenon is driven by high-quality standards and often speaks a language based on what can be called the "craft rhetoric", for which craft/handmade is opposed to industrial, and mass production is opposed to artisanal [1].
Analytical chemistry in synergy with advanced data analysis can be profitably used to build new tools to aid consumers when choosing and pairing foodstuff, and producers to meet the consumers' expectations. In this perspective, the aim of the present study is to investigate the links between the "objective" world of analytical chemical profiling – e.g. using spectroscopy – and the "subjective" world of consumers tasting and describing food.
Consumer's preferences are traditionally assessed by directly interviewing small groups of people, but with the growth of the Internet and its web communities, mining online-posted reviews has become an interesting approach for assessing product appreciation and reception. Huge amounts of user-generated data are available today in very different formats, such as numeric scores, logical scores (in the form of like/dislike), geotags and written descriptions.
These shifts in how food is chosen and consumed, has led the beer industry towards massive changes, propelled by the explosion of craft and micro-breweries and the spread of home brewing. In relation to this, a data set about beer was used as a benchmark: spectroscopic data were previously acquired and analyzed by us [2], while user-generated reviews were

mined from the RateBeer website (https://www.ratebeer.com/), a sort of social network for beer enthusiasts.
The proposed Chemometric strategy comprises:
i) Text analysis methods [2] to process the user-generated reviews and convert them into numeric format, by the bag-of-words approach [3].
ii) Principal component analysis–generalized canonical analysis (PCA–GCA,[4]) to investigate the links between spectral and text data.
Moreover, to select subsets of terms from the text data two approaches were used: topics extraction using penalized matrix decomposition [5] and manually-defined sets of terms related to specific aspects of beer making and tasting.
Overall, twenty topics were identified and correlated with spectral features, a representative example, topic "Hops" is shown in Figure 1.

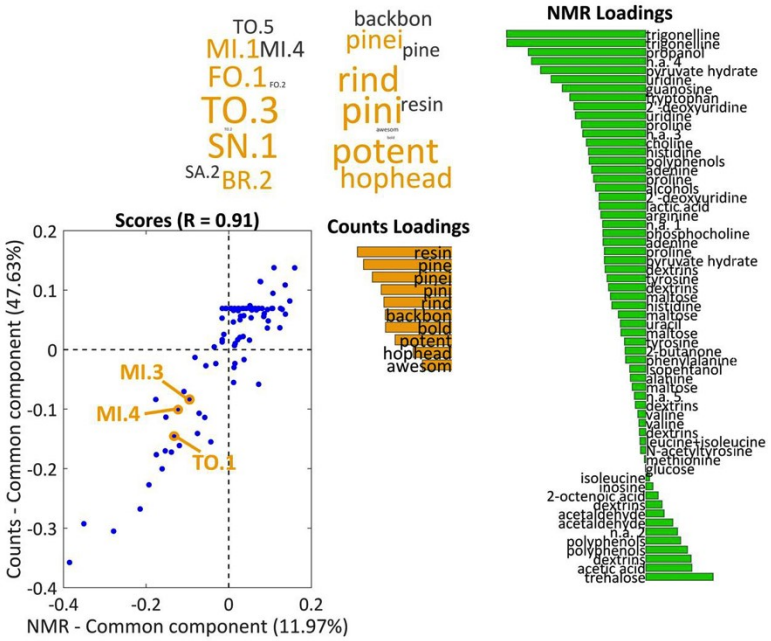

Figure 1. Results from PCA-GCA, on the bottom the scores of the common component , in the middle the loadings for text counts (orange) and for NMR features (green). On top the topic's characteristic terms and the main representative beer samples for this topic, represented as wordclouds.

References

[1] Rice J., Craft Rhetoric, Commun. Crit. Stud. 12, 2015, 218–222.
[2] Cavallini N., Bro R., Cocchi M., Anal. Chim. Acta , 1061, 2019, 70-83.
[2] Radovanović M., Ivanović M., Novi Sad J. Math. 38, 2008, 227–234.
[3] Banchs R.E., Text Mining with MATLAB®, Springer New York, New York, NY, 2013.
[4] A.K. Smilde, I. Måge, T. Naes, T. Hankemeier, M.A. Lips, H.A.L. Kiers, E. Acar, R. Bro, J. Chemom. 31, 2017, e2900.
[5] Witten D.M., Tibshirani R., Hastie T., Biostatistics, 10, 2009, 515–534.