CMA4CH 2012, Mediterraneum Meeting
Use of Multivariate Analysis and Chemometrics in Cultural Heritage and Environment
4th ed., Rome, Italy, Europe, 27-30 May 2012
lecture

# Quantitative Application of Self Organising Maps

**Richard G. Brereton**
School of Chemistry, Cantocks Close, University of Bristol, Bristol BS8 1TS, UK
and Brereton Consultancy, New Bond House, Bond Street, Bristol BS2 9AG, UK

## Abstract

Self Organising Maps (SOMs) are usually introduced as methods for visualisation but can be extended to quantitative situations, for example, to determine the class membership of a sample or which variables are most significant markers. This presentation will discuss the enhancement of SOMs.

The methods will be illustrated by two case studies. The first involves an NMR study of salival metabolites and the second involves studying fungal degradation of apples.

## Introduction

SOMs were first proposed by the Finnish scientist Teuvo Kohonen [1], and are sometimes also called Kohonnen maps. The first applications were very computer intense and primarily related to visualisation of data. Unlike most traditional statistical methods, such as PCA (principal components analysis), they do not assume that trends are linear or that the measurements are normally distributed. Most real world case studies (e.g. in metabolomic profiling, or cultural heritage studies), fail these assumptions, hence conventional approaches are often inappropriately applied to chemometric data. Of particular importance is that quantitative decisions are made when samples are far away from the mean e.g. in classification or quality control, yet it is very hard to model probabilities on the wings of a distribution unless a very large training set (often many thousands) is available.

A limitation of conventional uses of SOMs, is that they are not quantitative, yet much of chemometrics involves obtaining quantitative information about samples. This presentation discusses the extension of SOMs to situation where the aim is to obtain some level of numerical information about samples or variables, hence increasing their utility to chemometricians faced with complex data.

Supervised SOMs are an enhancement of the original method and involve creating additional layers involving information about class membership. It is necessary to optimise them by determining a relative weight between the classifiers and variables. The optimised maps can then be used to predict the class membership of unknown samples.

Supervised SOMs also have an important role in variable or feature selection, that is, determining which variables are most influential in discriminating between two or more groups of samples. Variables can be ranked in order of relative significance. By repeating the learning many times over, one can obtain a numerical value of confidence of significance.

Finally supervised SOMs can be used when there are several factors that influence the data, by supervising the learning for each factor independently, and therefore form an alternative to, for example, multiway ANOVA. Many real life applications fail tests of multivariate normality, and as such traditional approaches such as ANOVA followed by F statistics are not suitable. Furthermore variables (or features) often interact and as such methods that test for the significance of each variable independently, although they take into consideration interactions between factors, do not take into consideration interactions between variables. Supervised SOMs overcome these limitations of traditional approaches, but are computationally intense.

## Materials & Methods

Two case studies are used to illustrate the methods.

The first involves the NMR studies of saliva [2]. An aim is to determine whether mouthwash has an effect on the salival signal. Bacteria in the mouth should be killed and this reduces the bacterial metabolites found in mouthwash. However there will be other factors influencing the salival profile.

Three factors are studied, namely the use of mouthwash (against controls), the individual donor, and a dummy factor, the sampling day.

The second involves studying fungal degradation of apples [3]. Apples were divided into control and treated groups. The treated group was injected with penicillum (fungus), both groups were studied for 10 days. The volatiles were monitored using GCMS.

## Results

By comparing component planes to overall maps, it is possible to determine which variables are most diagnostic of a specific group or class. Figure 1 illustrates this for the apples. The method can be sophisticated by developing quantitative indices of similarity between the component plane and the map, and reforming the map many times (typically 100). If a variable remains discriminating for all 100 maps it is a strong discriminator.

Supervised SOMs can force a map to appear distinguish different classes, and hence are prone to overfitting. Unlike approaches such as PLS-DA or LDA, it is possible to weight the classifier at any desired level relative to the variables, a high weight separating even random classes, and low weight largely ignoring class information. However by leaving out samples in a test set it is possible to determine the quality of the classifier. By repeating the maps many times over, different samples can be left out and so classification ability assessed.

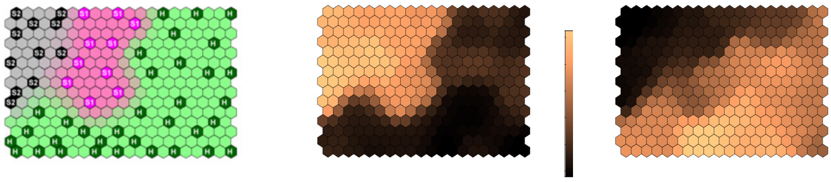

Figure 1: Apple spoilage. Left, a map, H = healthy, S1 = early spoilage, S2 = late spoilage, compared, right, to component planes of two variables,

Finally, in the case of the NMR of saliva, there are three factors (treatment, donor and sampling day) that can be used to train supervised maps. It is therefore possible to determine which variables are most influential for each of the three factors, and also whether these factors have a significant effect, by assessing classification ability using repeated generation of test sets. This is an alternative to ANOVA based approaches and does not assume normality or linearity. It can be shown that in the case of the NMR (as in most real world situations) the distribution of the variables (in this case NMR peak intensities) fail the assumptions of normality.

## Conclusions

SOMs provide a valuable alternative to traditional chemometrics methods. With the advance of rapid computer power, they are feasible tools for real time assessment of data, and carry none of the assumptions of traditional approaches.

## Acknowledgements

For the NMR studies, Martin Grootveld (University of Bolton, UK) and Chris Silwood (London South Bank University, UK) are thanked, for the apple studies Virag Sagi-Kiss (Corvinus University, Budapest, Hungary). Computations were performed by Gavin Lloyd, Kanet Wongravee and Sim Fong of the University of Bristol.

## References

1) T. Kohonen, Self-Organized Formation of Topologically Correct Feature Maps, *Biol. Cybern.*, 43(1), (1982) 59-69

2) K. Wongravee, G.R. Lloyd, C.J.L. Silwood, M. Grootveld, R.G. Brereton, Supervised Self Organizing Maps for classification and determining potentially discriminatory variables: illustrated by application to NMR metabolomic profiling, *Anal. Chem.*, 82(2), (2010) 628-638

3) S.S. Fong, V. Sagi-Kiss, R.G. Brereton, Self Organizing Maps and Support Vector Regression as aids to Coupled Chromatography: illustrated by Predicting Spoilage in Apples using Volatile Organic Compounds Reference, *Talanta*, 83(4), (2011) 1269-1278