Talanta 83 (2011) 1269–1278

Contents lists available at ScienceDirect

Talanta

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

# Self-Organizing Maps and Support Vector Regression as aids to coupled chromatography: Illustrated by predicting spoilage in apples using volatile organic compounds

Sim S. Fong[a], Virág Sági-Kiss[b], Richard G. Brereton[a,*]

[a] Centre for Chemometrics, School of Chemistry, University of Bristol, Cantocks Close, Bristol BS8 1TS, United Kingdom
[b] Corvinus University of Budapest, Department of Applied Chemistry, H-1118 Budapest, Villányi út 29-33, Hungary

ARTICLE INFO

*Article history:*
Available online 4 July 2010

*Keywords:*
Gas chromatography–mass spectrometry
Food spoilage
Volatile organic compounds
Support Vector Regression
Self-Organizing Maps
Pattern recognition

ABSTRACT

The paper describes the application of SOMs (Self-Organizing Maps) and SVR (Support Vector Regression) to pattern recognition in GC–MS (gas chromatography–mass spectrometry). The data are applied to two groups of apples, one which is a control and one which has been inoculated with *Penicillium expansum* and which becomes spoiled over the 10-day period of the experiment. GC–MS of SPME (solid phase microextraction) samples of volatiles from these apples were recorded, on replicate samples, over time, to give 58 samples used for pattern recognition and a peak table obtained. A new approach for finding the optimum SVR parameters called differential evolution is described. SOMs are presented in the form of two-dimensional maps. This paper shows the potential of using machine learning methods for pattern recognition in analytical chemistry, particularly as applied to food chemistry and biology where trends are likely to be non-linear.



## 1. Introduction

Over the past few years, analytical chemistry has increasingly been applied to problems in areas such as biology, medicine, food science and cultural studies [1–4]. Many applications involve using coupled chromatography to monitor processes, typically GC–MS (gas chromatography–mass spectrometry). There has been a substantial literature on the preprocessing of GC–MS systems, for example peak detection and resolution, but much less emphasis on the resultant pattern recognition. Typical approaches involve using linear models for visualizing and interpreting data such as Principal Component Analysis (PCA) [5–8], Partial Least Squares (PLS) [9–13] and Soft Independent Modelling of Class Analogy (SIMCA) [14,15]. These classical approaches are widely available in packaged software but also are computationally fast and effective when trends in data are linear.

Over the past two decades, several approaches from the machine learning community are becoming widespread in for example, biology, and economics, but are less well established in analytical chemistry. Two particularly widespread methods include those based on support vectors (SVs) [16–19] and Self-Organizing Maps (SOMs) [20–22]. These methods are more computationally intensive than traditional approaches. But with the rapid increase in computing power, approaches that were impractical and expensive on resources a decade or more ago can now be implemented easily on desktops and provide solutions within a short timescale. An especial advantage of these approaches is that they can tackle non-linear data. This is an important feature as it is likely that most natural processes are non-linear. For example we do not expect there to be a linear relationship between the amount of a volatile generated during fungal degradation of food and degradation time [23–28]. Another problem with most traditional approaches is that they are based on least squares methods that reduce the mean squared error between the data and a model: such methods can have problems if there are outliers which is quite common in many experimental studies. Whereas traditional approaches can, often with some difficulties, be adapted to these situations, they often become awkward and still have the same underlying assumptions. SVR and SOM based approaches make no such assumptions. In addition SOMs include a wide variety of approaches for visualization [12,22] and as such are much more flexible than principal components.

In this paper, we demonstrate the applicability of such approaches to a common problem – that of monitoring food spoilage using GC–MS, studying the change in volatile organic compounds (VOCs) monitored on the surface of the apples over time,

* Corresponding author.
*E-mail address:* r.g.brereton@bris.ac.uk (R.G. Brereton).

0039-9140/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.talanta.2010.06.051

**Table 1**
Distribution of samples according to sample type and level of spoilage.

| Sample type | Day | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | 2 | 3 | 4 | 5 | 6 | 8 | 9 | 10 |
| Inoculated | | | | | | | | |
| No. of samples | 4 | 4 | 4 | 4 | 2 | 4 | 4 | 4 |
| Spoilage | Healthy | Healthy | Healthy, 1st phase | 1st phase | 1st phase | 2nd phase | 2nd phase | 2nd phase |
| Control | | | | | | | | |
| No. of samples | 4 | 4 | 4 | 4 | 2 | 4 | 2 | 4 |
| Spoilage | Healthy | Healthy | Healthy | Healthy | Healthy | Healthy | Healthy | Healthy |

in a control group (that was not inoculated) and a treated group (inoculated with *Penicillium expansum*).

## 2. Experimental

### 2.1. Experimental procedure

Four granny smith apples (healthy, of diameters between 70 mm and 80 mm, and of the same colour) were selected. Each apple was stored separately in a commercial bottle jar (1700 ml) at 22 ± 2 °C during the experiment with a sampling outlet sealed with a polytetrafluoroethylene (PTFE) silicone septum (20 mm). A cotton wool plug was placed on the top of the bottle jars to ensure oxygen supply for microbes. Before the experiment, apples were decontaminated by washing with propanol and distilled water. After air drying, the apples were treated with low temperature steaming (60 °C for 20 min) for sterilization.

*Penicillium expansum* P1 strain, isolated from a spoiled plum, was maintained and cultivated on MG (malt extract–glucose) agar (0.5% glucose), 1.7% malt extract (MERCK) and 2% agar (MERCK) at 25 °C. Conidia of 5–7 days old culture were used for inoculation ($2.7 \times 10^6$ conidium/ml). The apples were inoculated with a four-tailed needle to make sure that all apples were inoculated with an equal amount of conidium. The needle was applied 12 times to each apple resulting in 48 penetration points.

All sampling was via headspace solid phase microextraction (HS-SPME): SPME fibres (polydimethylsiloxane/divinylbenzene (PDMS/DVB, 65 μm)) were purchased from Supelco (Bellefonte, PA). All fibres were preconditioned according to the manufacturer's recommendations (250 °C for 30 min) prior to their first use and reconditioned for 20 min in between each run to minimize carry-over effects. Extractions were performed at room temperature. The fibre is housed in a stainless-steel needle that allows penetration of the membrane covering the sample vial and the septum in the gas chromatography (GC) injection port. The SPME needle was passed through the septum of the flask and the fibre was gently pushed out of the needle to be exposed directly above the apples. After 25 min of extraction, the fibre was retracted into the cannula and removed from the flask. This was followed by desorption in the GC injector at 250 °C. A new fibre was used at the beginning of the analyses and was not changed throughout the experiments.

All extractions were analyzed on a Finnigan GCQ GC–MS system (Finnigan Mat., USA). For separation, a Rxi-5 ms capillary column (Restek Corporation, Bellefonte, PA) (30 m, 0.25 mm i.d., 0.40 μm, 95% dimethyl-/5% diphenyl poly-siloxane) with helium (purity 99.999%, 0.4 ml/min constant flow) as carrier gas. The GC oven program was chosen according to the following scheme: 40 °C for 3 min, 12 °C/min up to 235 °C. The desorption time and temperature were 20 min at 250 °C. The final temperature was maintained for 7 min. In total, the GC was run for 26.25 min. The split was kept closed in the beginning of the desorption and was opened after 1 min. For MS detection, electron ionization (EI) with 70 eV was applied and mass fragments were detected between 40 and 349 *m*/*z*. The ion source and transfer line temperature were 180 °C and 275 °C, respectively. The column was not changed throughout the experiment.

The extraction time was optimized in a separate experiment. The extraction time, desorption time and desorption temperature were evaluated based on a three-factorial experimental design. The fibre conditions however were according to that reported elsewhere [29]. The extraction temperature was fixed at room temperature.

### 2.2. Samples description

In this paper, there are a total of 58 VOC sampled from control and inoculated apples on 2nd, 3rd, 4th, 5th, 6th, 8th, 9th and 10th day. The number of samples obtained on each day is shown in Table 1. The apple samples were randomized each day for analysis using GC–MS and were analyzed on the day of sampling to avoid problems of SPME storage. Blank runs of the SPME fibre were performed before sample injection on each day.

The inoculated samples can be grouped according to the level of spoilage based on visual observation and numbers of days since the samples were inoculated. A sample is described as in the 1st phase of spoilage when white spots are observed and it is labelled as in the 2nd phase when white-green spots are formed. The control samples remain healthy throughout the experiment.

### 2.3. Data analysis

The chromatograms in netcdf format were converted into mat files using conversion tools available in the public domain [30]. The scanning rate of the chromatograms is 0.20 s/scan with *m*/*z* ranging from 40 to 349 resulting in GC–MS chromatograms of dimensions 7876 × 310.

The chromatograms are baseline corrected using asymmetric least squares [31] and pre-aligned [32] prior to peak detection and matching using our previously published method [33]; this results in a peak table which is a matrix whose columns represent the peak areas over all mass channels of all unique compounds detected and whose rows represent chromatograms or samples. The next phase is to remove all peaks that were present in 5 or less samples as these may be artifacts of the peak detection or analytical techniques and are not likely to be good markers, to give 200 unique compounds which results in a peak table of dimensions 58 × 200. Three of the 200 peaks are not detected in the inoculated samples. For the purposes of SOM visualization of the inoculated samples and SVR, a reduced peak table only those compounds detected in the inoculated samples, of dimensions 30 × 197 was employed, since variables that are not detected in a set of samples correspond to a row of 0 s, which cannot be standardized as the standard deviation is 0.

Further details of our methods for converting raw GC–MS data to peak tables have been described elsewhere [33] and are not repeated in this paper for brevity.

## 3. Multivariate analysis

### 3.1. Preprocessing

The peak table was square rooted (to reduce the influence of large variables and reduce heteroscedastic noise), and row scaled

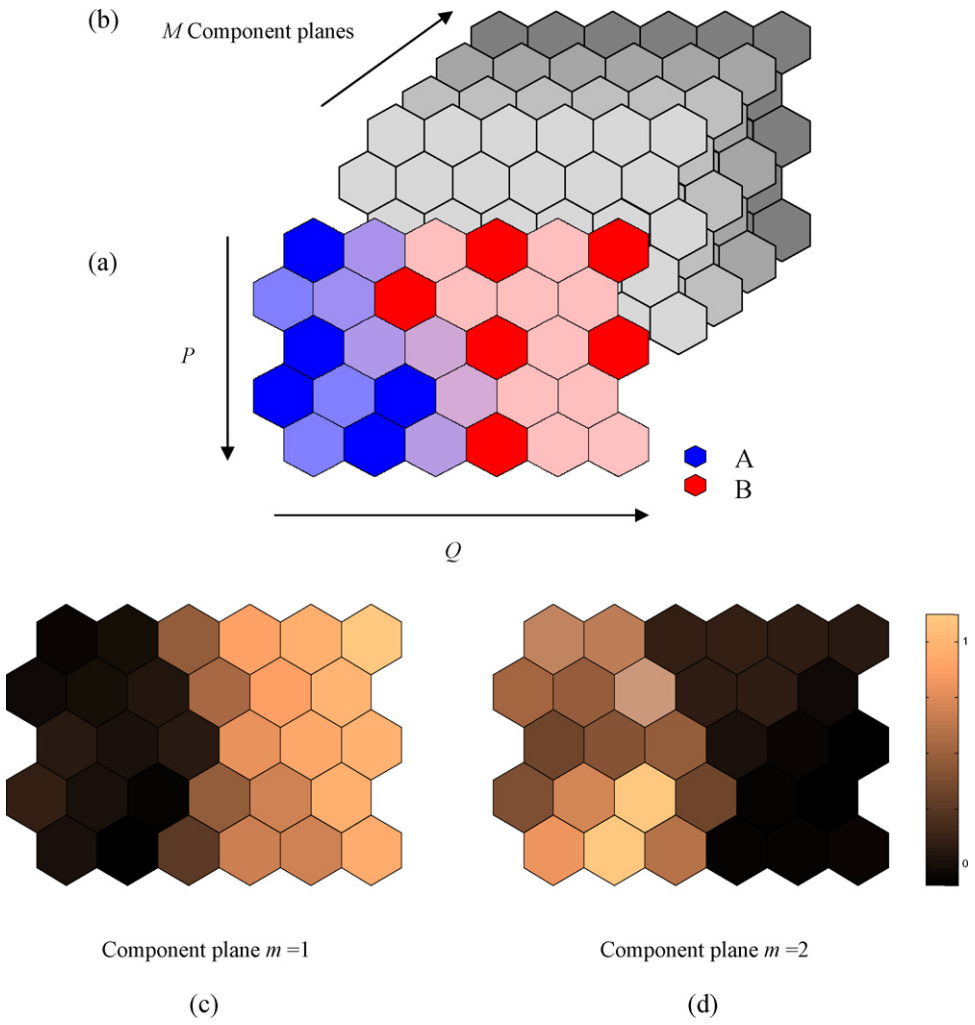



**Fig. 1.** Representation of a SOM for samples that fall into two classes with (a) the top layer the BMUs of the samples, (b) a series of *M* component planes corresponding to each variable or GC–MS peak, (c) the component plane of a variable number 1 and (d) the component plane of a variable number 2.

to constant sum (because the amount of gas sampled could not be controlled). Note that there are several alternative approaches for data preprocessing but we find that the method described in this paper is effective in a wide variety of studies, so restrict the discussion in this paper to one common approach for the sake of brevity. For SOMs, the variables were additionally standardized in order to allow each variable to have equal influence. The motivation behind choosing these preprocessing methods has been described elsewhere [12,34–36]. It is possible to compare many approaches but in this paper we are focusing on the SOMs and SVR reporting only one method for preparing the data.

### 3.2. Self Organizing Maps

#### 3.2.1. Algorithm

Self Organizing Maps (SOMs) are an approach for visualizing data [12,20–22,37] and have advantages over the more traditional approach of Principal Component Analysis (PCA) [5–8] in various ways. They are non-linear methods and it is likely that the relationship between spoilage and the concentration of volatiles is highly non-linear [23–28]. They use the full space for representation of data and there are many different ways for presentation of the maps. SOMs have been less commonly employed in analytical chemistry compared to traditional approaches partly due to the lower availability of user friendly software including graphical presentation and also because they are comparatively computationally intensive; however with the increasing powers of microprocessors they are far more feasible for real time calculations.

Maps can be of various geometries. The most traditional is rectangular maps, which are composed of cells that are either square or hexagonal in nature. In this paper, we use hexagonal cells (or units). We train maps of dimensions $15 \times 20$ ($P \times Q$) using 5000 iterations. The details of the algorithm is not included for brevity; they are described elsewhere [12,20–22]. Briefly, it involves initialization of a map of dimensions $P \times Q$ consisting of $K$ cells (300 in this application) each cell being initially characterized by randomly chosen weight vectors whose length equals that of the sample vectors (in this case, 200 variables, $M$ corresponding to the compounds detected in the samples), adding an extra "hidden" dimension to the map that is sometimes represented by component planes. A sample is randomly chosen to compare to the $K$ weight vectors and is used to find the cell that is characterized by the weight vector that is most similar to the sample called the best matching unit (BMU). This cell is then trained to resemble the randomly selected vector more closely. The neighbors of the BMU are also updated. For a map unit to be updated, its distance from the current BMU must be less than a parameter called the neighborhood width, otherwise it is not updated. The size of the neighborhood width reduces during training meaning that at later iterations the map is changed less. The entire process is repeated each iteration. The number of iterations far exceeds the number of samples, so all samples are chosen many times to have an influence over the training. Once the map is fully

trained, samples that resemble each other more closely are closer to each other on the map. When samples are grouped into classes, the trained map can be used to visualize which parts of the map correspond to each grouping in the data where an additional set of variables corresponding to the class membership of each sample is used for the visualization process. For unsupervised SOMs (as reported in this paper), class membership information is not used for the training, just to visualize the map. The class membership of a sample is characterized by an RGB (Red Green Blue) colour vector that represents a particular class. Each map unit therefore will have three additional weights corresponding to a trained RGB value that can be used to shade the map unit for visualization purposes. This allows all class regions to be displayed on the same map instead of using component planes which required a separate visualization of each class [12,22].

It is not necessary to restrict maps to rectangular geometry, and a common alternative is to represent samples on the surface of a sphere. Rectangular maps have the advantage that outlying samples can be represented at the edges of the map and they can be used to represent other structures in the data for example spoilage of samples as a function of time. However, they have the disadvantage that some samples are always chosen to be in the centre or the edges, and spherical representation does not have such a restriction. In this paper, we focus on rectangular maps.

#### 3.2.2. Self-Organizing Map Discriminant Index

In addition to using SOMs for representing similarities between samples, they can also be used to determine which variables (in this case, GC–MS peaks) are most responsible for separation. A SOM can be represented either as the top layer which usually is illustrated using the BMUs of the samples, or as a series of hidden layers (component planes). Each layer consists of a cross section through the corresponding weight vectors, so each variable (corresponding to a GC–MS peak) has its own component plane. Fig. 1 shows the representation of a SOM for two classes (A and B) characterized by $M$ variables. Each variable is represented by a single component plane where the relationship between samples and variables can be visualized – the darker the shading, the less important the variable is for describing the corresponding region of the SOM. Two component planes are illustrated in Fig. 1(c) and (d) corresponding to variable 1 which is characteristic of class B and variable 2 for class A. The concept of a SOMDI (Self-Organizing Map Discrimination Index) [37,38] has been defined to numerically describe how well the component plane of a variable corresponds to the SOM when BMUs are distinguished according to class. The SOM class map is transformed into two-layer maps, one layer representing an "in group" [12,39] and the other the "out group" for comparison to the component planes of each variable. When there are more than two groups, this comparison is performed separately for each group.

Previously, we have reported both the use of unsupervised SOMs for ranking variables when there are just two groups in the data [38] and supervised SOMs for ranking and determining the number of significant variables when there are more than two classes and several factors that may influence the data [37]. In this paper, we use unsupervised SOMs as the groups are well separated but extend the approach by dividing the data into three groups corresponding to the degree of spoilage with an indicator of significance attached. A total of 100 maps are trained, each differing according to the random start: note that each map will be different according to the random seeded weight vectors but we expect real trends to be stable over all maps. Each map is in itself trained using 5000 iterations involving in a total 500,000 iterations to result in a consensus view. The BMUs in each map are characterized according to their class (or origins); when there are more than one class, one versus all [37] comparisons are performed for each of the classes separately. If there are three classes, comparisons between class 1 and the rest, class 2 and the rest and class 3 and the rest are performed on each of the 100 maps. Variables are ranked according to the magnitude of average SOMDI ($\Delta\bar{d}$) over 100 iterations. Which variables are significant are determined based on the Hodges Lehmann method [40]. If there are $M$ variables, there will be $M$ average SOMDI. The Hodges Lehmann method defines $Y$ ($Y = M(M+1)/2$) Walsh averages, $W_y$.

$$W_y = \frac{m_i + m_j}{2}$$

where $i \geq j$, $i = 1 \ldots, M$, $j = i \ldots M$, and $m$ is the average SOMDI.

The $100(1-\alpha)\%$ confidence interval of the null distribution is determined from the ordered statistic of Walsh averages.

The compounds deemed to be significant were tentatively identified by comparing their mass spectra to the mass spectra of reference compounds in a reference standard library (NIST MS Search 2.0, National Institute of Standards and Technology).

### 3.3. Support Vector Regression

#### 3.3.1. Algorithm

In this paper, we want to determine whether there is a relationship between GC–MS peak areas and spoilage time. The relationship is likely to be non-linear and in addition there may be outliers, for example the peak detection algorithm may occasionally miss or misassign peaks or there could be some biological or analytical factors that influence the occurrence of individual peaks. Sometimes it is possible to detect outliers by visual inspection but if there are 200 peaks in 30 chromatograms, this would involve 6000 visual checks which at 5 min per check would take 500 h and still may not be perfect. Hence, when trying to relate 200 peaks to spoilage time using a relationship that may be multilinear, traditional linear least squares methods may not be the most suited. In this paper, we report an alternative.

Support Vector Regression (SVR) is a method originating from machine learning that can be employed for regression problems [19,41–44]. Below we formulate SVR models by assuming that we are trying to relate a GC–MS peak area, which we call $x$ to spoilage time, which we call $c$. We are trying to determine whether the GC–MS area can be determined by the spoilage time i.e., to determine whether $x$ is a function of $c$. Note that this application is called classical calibration rather than inverse calibration. We use the $x/c$ notation in this paper, as it is conventional to refer to GC–MS peak areas as the "$x$" block. In SVR, the input data vector is mapped onto multi-dimensional feature space using kernel functions and a linear model is constructed in this feature space.

We try to find a relationship as follows:

$$\hat{x} = (w \cdot \Phi(c)) + b$$

where $w$ denotes is a weight vector, $b$ is a constant (the bias), $\hat{x}$ is the predicted peak area, and $\Phi$ is called a kernel that can be used for non-linear models. For a linear model, $\Phi(\boldsymbol{c}) = \boldsymbol{c}$. In this paper, we employ a radial basis kernel [12,19] which can be used to introduce non-linearity into the relationship, the value of $\sigma$ being in units of the standard deviation of the overall data. We use soft margin SVR (which is normal) characterized by a penalty error $C$ and a value for the width of the margins $\varepsilon$. The main parameters are illustrated in Fig. 2. Samples on or outside the margins are considered Support Vectors. The penalty error $C$ relates to the relative weight attached to samples outside the margins; it controls the trade-off between the margin and the size of the slack variables. In practice, a small value of $C$ will increase the training error and a large value will lead to behaviour similar to that of hard margin SVs where all samples are on the correct side of the margins. The penalty error $C$ can take values between 0 and $\infty$. For $\varepsilon$ which corresponds to the width of the margins, the

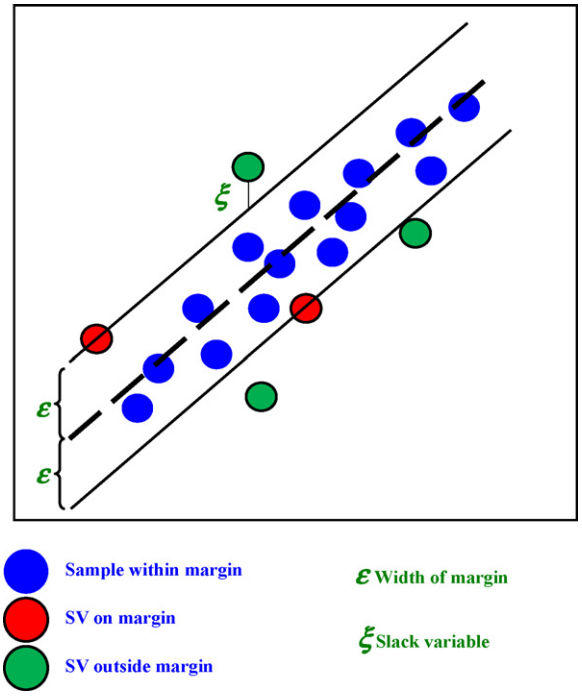



**Fig. 2.** Main parameters for SVR.

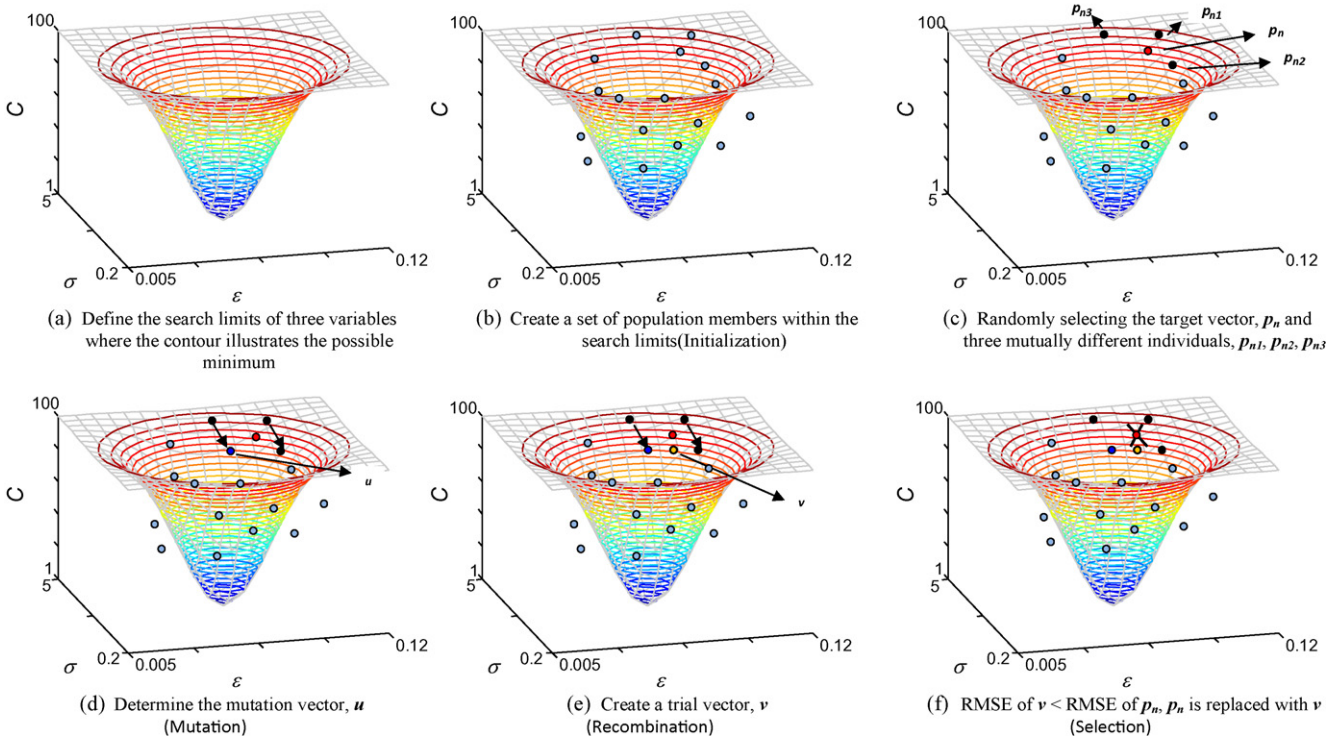
value influences the number of SVs used to construct the margins. The larger the value of $\varepsilon$, the fewer the SVs resulting in a smoother regression function but larger prediction error. For the kernel width $\sigma$, a large value results in a linear function whereas a small value indicates a more complex one. Note that some combinations of values of these three tuneable parameters are not valid, for example it may not be possible to produce a linear model (high $\sigma$) combined with a high value of $C$ and low $\varepsilon$ (corresponding to tight boundaries and all samples within these boundaries) if the data are curved.

### 3.3.2. Differential Evolution Algorithm for optimization

The quality of SVR models depends on an optimal setting of SVR parameters: $C$, $\sigma$ and $\varepsilon$. There is relatively limited discussion in the analytical chemistry literature about how to refine such calibration models. In this paper, Differential Evolution (DE) algorithm is employed to determine the optimum SVR parameters [45,46]. The DE algorithm is a simple evolutionary optimization algorithm first reported by Storn and Price [47]. The algorithm begins to explore the search space by randomly choosing a set of population with the number of members, $N$, within defined search limits of the three variables where the population members then experience mutation, recombination and selection until the stopping criterion i.e., a maximum number of iterations, $G$ is met.

Fig. 3 summarises the DE algorithm. The algorithm begins by defining the upper, $U$ and lower limits, $L$ for each parameter, $d - [L_d, U_d]$ where $d = 1, \ldots, D$. $D$ is the total number of parameters ($D = 3$) and refers to $C$, $s$ and epsilon, $\varepsilon$. Fig. 3(a) illustrates the search space with the possible minimum. The search limit of $C$ is set between 0.1 and 100 (and the value is optimized using a log scale). For $\sigma$, the value is varied between 0.2 and 5 (similarly on a log scale) and the $\varepsilon$ value involves the range of square root and row scaled GC–MS peak areas (0.005–0.12 on a linear scale). Note that these limits are quite liberal, designed to cover all possible ranges of the three parameters.

A population, $P$ ($N \times D$ where $N = 20$ and $D = 3$ corresponding to the number of parameters in the model), is initialized by a set of randomly generated sets of values of each parameter using a uniform random number generator within the search limit of the three variables ($C$, $\sigma$ and $\varepsilon$) where each variable is denoted by $d$ ($d = 1, \ldots, D$) and the lower and upper limits $L$ and $U$ as described above (Fig. 3(b)). The GC–MS peak areas, $x$ are divided into training, $\boldsymbol{x}_{\text{train}}$ (20 samples) and test set, $\boldsymbol{x}_{\text{test}}$ (10 samples) where the corresponding spoilage times are $\boldsymbol{c}_{\text{train}}$ and $\boldsymbol{c}_{\text{test}}$, respectively.




(a) Define the search limits of three variables where the contour illustrates the possible minimum

(b) Create a set of population members within the search limits(Initialization)

(c) Randomly selecting the target vector, $\boldsymbol{p}_n$ and three mutually different individuals, $\boldsymbol{p}_{n1}$, $\boldsymbol{p}_{n2}$, $\boldsymbol{p}_{n3}$

(d) Determine the mutation vector, $\boldsymbol{u}$ (Mutation)

(e) Create a trial vector, $\boldsymbol{v}$ (Recombination)

(f) RMSE of $\boldsymbol{v}$ < RMSE of $\boldsymbol{p}_n$, $\boldsymbol{p}_n$ is replaced with $\boldsymbol{v}$ (Selection)

**Fig. 3.** The schematic of the differential evolution algorithm.

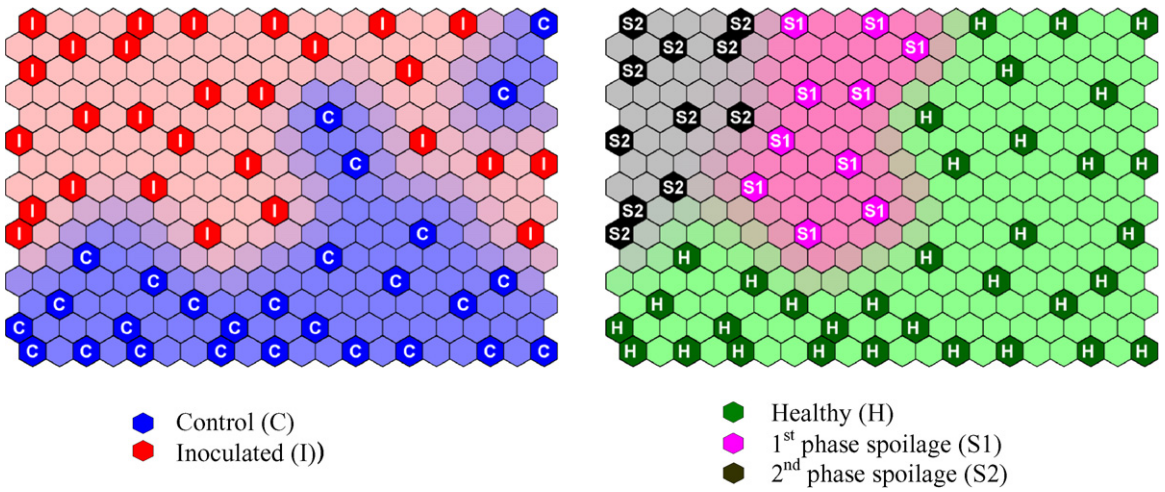



**Fig. 4.** The trained SOMs of the control and inoculated samples using a rectangular grid represented according to the sample type and level of spoilage (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of the article.).

The training set consists of two-third of the samples whilst the test set consists of the remaining one-third of the samples and is independent of the training set. The model is optimized in autopredictive mode on the training set, and then the test set is used to determine how well the model performs. This distinction is important, optimizing and testing the model on the test set may result in overfitting of the models. The split into test and training sets is performed 100 times so that there are 100 separate models, as described in Section 4.3. Below we describe the algorithm as applied to a single test and training set split.

For each of the $N$ sets of parameters (or population members) the root mean square error, RMSE for the training set is obtained as follows:

$$RMSE_n = \sqrt{\frac{\left[\sum_{z=1}^{Z}(x_{z,\text{train}} - \hat{x}_{nz,\text{train}})\right]^2}{Z}}$$

where there are $Z$ samples in the training set and $\hat{x}_{nz,\text{train}}$ is the prediction of the $z^{\text{th}}$ training set sample using the model using the $n^{\text{th}}$ set of parameters. A population member, $\boldsymbol{p}_n$ (the target vector of dimensions $1 \times D$ (or $1 \times 3$ in our case)) and three mutually different population members, $\boldsymbol{p}_{n1}$, $\boldsymbol{p}_{n2}$, $\boldsymbol{p}_{n3}$ (where $n1$, $n2$, $n3 \neq n$) are randomly selected (Fig. 3(c)). A mutation vector (of dimensions $1 \times 3$) $\boldsymbol{u}$ is created where $\boldsymbol{u} = \boldsymbol{p}_{n3} + F(\boldsymbol{p}_{n1} - \boldsymbol{p}_{n2})$ and $F$, the differentiation constant, is set to 0.95. The scaled difference between two randomly chosen members of the population, $F(\boldsymbol{p}_{n1} - \boldsymbol{p}_{n2})$, is used to define the direction and length of the search step for the third population member $\boldsymbol{p}_{n3}$ (Fig. 3(d)).

In the next step, the algorithm evaluates each element of the mutation vector $\boldsymbol{u}$ in turn. If a randomly selected number is greater than the crossover constant $R$ ($R$ is set to 0.5 in this paper), $u_d$ replaces $p_{(n,d)}$. The algorithm is designed to replace at least one of the elements of $\boldsymbol{p}_n$ with the corresponding element of $\boldsymbol{u}$. The regenerated target vector is called the trial vector, $\boldsymbol{v}$. The elements of the trial vector are checked to ensure they are within the search limits of the variables. If not, the element(s) exceeding the search limit are regenerated, $L_d + [(U_d - L_d)r]$ where $r$ is a random number between 0 and 1 generated using a uniform distribution: note that it is rare that an element of the trial vector exceeds the experimental limit, but this procedure is introduced to avoid the occasional boundary problem. The RMSE obtained using $\boldsymbol{v}$ and $\boldsymbol{p}_n$ are compared. If the RMSE calculated with $\boldsymbol{v}$ is greater than that obtained with $\boldsymbol{p}_n$, $\boldsymbol{p}_n$ is retained otherwise it is replaced by $\boldsymbol{v}$. Note that in each of the iterations, a maximum of one of the $N$ population vectors will be changed. Fig. 3(e) shows a case where the RMSE of $\boldsymbol{v}$ less than the RMSE of $\boldsymbol{p}_n$ therefore $\boldsymbol{p}_n$ is replaced with $\boldsymbol{v}$.

The algorithm is repeated until the maximum number of iterations, $G$ (=1000 in this paper) is reached or the difference between the minimum RMSE and the second minimum RMSE value within the $N$ members of the population is less than $1 \times 10^{-5}$. Note that the value $1 \times 10^{-5}$ relates to the range of peak areas employed in this study (in practice it is a small number relative to the experimental data) and in other applications may be changed according to the scale of the raw data. The population member with the minimum RMSE within the population is considered the best combination.

When the data are divided into test and training sets, the RMSEP can then be calculated on the test set using the optimized parameters of the training set for each of the 100 test/training set splits as follows, $\text{RMSEP} = \sqrt{\sum_{t=1}^{T}(x_{t,\text{test}} - \hat{x}_{t,\text{test}})^2/T}$, where there are $T$ samples in the test set.

Note that this algorithm can also be used in autopredictive mode in which case the optimized RMSE is employed as a measure of predictive power.

In this paper, we illustrate the application of SVR for prediction using non-linear models. We investigate the use of regression models for peak areas over time during spoilage time. It is anticipated that there will be definable trends and changes in concentrations of volatiles as there is decomposition of the fruit. It is most unlikely that these trends will be a linear function of time; in addition as in many such experiments there may be occasional outlying sample for a variety of common experimental reasons including problems with GC–MS peaks overlapping and subsequent quantification by peak deconvolution algorithms.

#### 3.3.3. Partial Least Squares regression

In this paper, we also use Partial Least Squares regression (PLSR) to predict the row scaled root square area of styrene. The prediction is performed on 100 identical training/test splits of SVR using 1 PLS component. The algorithm is described in detail elsewhere [48,49].

## 4. Results and discussions

### 4.1. Exploratory analysis

The trained SOMs clearly indicate that samples are visually distinguishable according to sample types and level of spoilage as illustrated in Fig. 4 using a rectangular grid. We can also train the inoculated samples to examine the level of spoilage according to observation of fruit and days. The SOM visualization map demon-

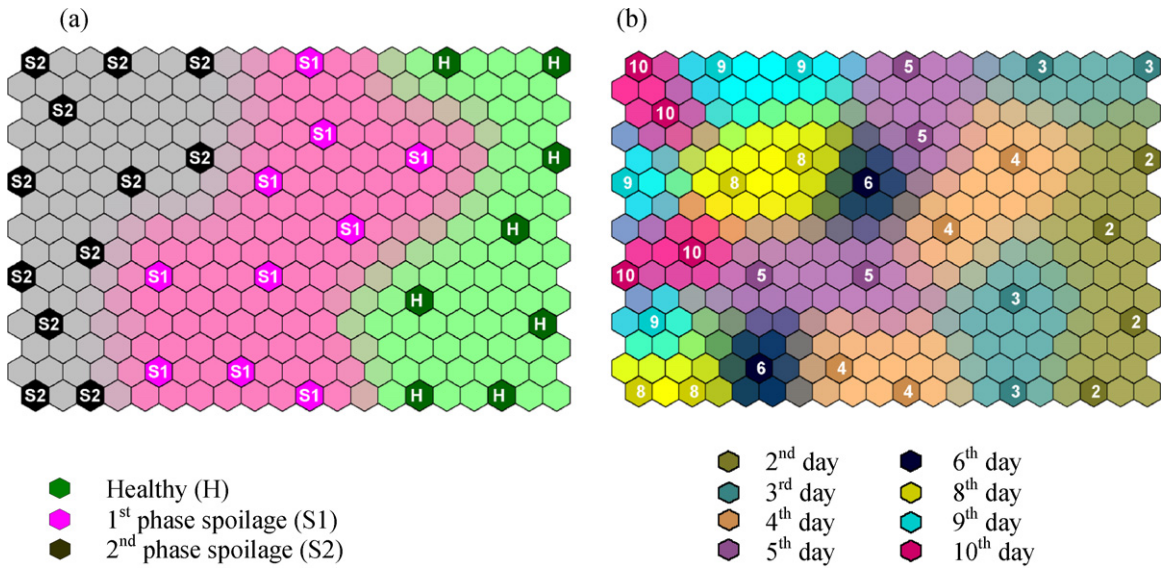



**Fig. 5.** The trained SOMs of the inoculated samples according to level of spoilage: (a) the nest matching units are labelled according to spoilage level and (b) the bst matching units are labelled according to the number of days the samples are inoculated. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of the article.)

strates that the inoculated samples are well clustered according to different phases of spoilage. In terms of days, the visualization map illustrates that samples from 2nd and 3rd day are likely to be healthy. When the inoculated samples have reached 4th, 5th and 6th day, the samples are characterized by white spots (1st phase spoilage) and on 8th, 9th and 10th day, white-green spots (2nd phase spoilage). The visualization maps of the inoculated samples with the BMU indicated according to phases of spoilage and days are illustrated in Fig. 5. The SOM visualization has been very useful as the apple samples can be differentiated according to control and inoculation consistent with visual inspection, and the spoilage time i.e., how many days have the samples been exposed to fungal degradation, can also be visualized in the maps.

At the back of the map are 200 component planes, each corresponding to a single detectable GC–MS peak as illustrated in Fig. 1. The distribution of samples in the map can be compared to the distribution of intensity in the component planes, as described in Section 3.2.2 and discussed below, to determine which variables are significant.

### 4.2. Significant variables for inoculation

The variables that appear to have a significant difference between the control and the inoculated samples are determined using SOMDI with an unsupervised SOM. Styrene and 1-methoxy-3-methylbenzene are found to be important for describing the spoiled samples as reported elsewhere [50–53]. From other studies, styrene is well recognized as a compound found in apple spoilage due to *P. expansum* [50,51]. This compound is characterized as weakly toxic; the US EPA has described styrene as a suspected carcinogen and a suspected toxin to the gastrointestinal, kidney, and respiratory systems. Generally, the significant variables either accumulate or decrease in concentration over time after samples are inoculated; the distributions of the significant compounds determined using the methods of Section 3.2 in control and inoculated samples from 2nd to 10th day are illustrated in Fig. 6, together with their identities as determined using manual interpretation of mass spectra and their aligned RT (this varies slightly according to chromatogram but we have discussed about our peak detection and alignment method elsewhere [33]). These compounds are in agreement with those reported elsewhere. Pentanoic acid ethyl ester [54] and hexyl propanoate [55] are found to reduce whilst methyl 3,3-dimethylbutanoate, ethyl hexanoate, benzene-1-methoxy-3-methyl and phenyl acetic acid [51] increase in the inoculated samples over time. Appearance and disappearance of the variables are expected as the fungi grow and metabolize. For the control samples, the concentrations of the significant compounds appear to be quite consistent throughout the experimental period suggesting the reproducibility of the sampling and instrumentation.

The component planes of the significant variables have clustering similar to the class map in Fig. 4 with the intensity of the map units indicates the presence/absence of the variable in a corresponding group. The darker the shading, the less important the variable for describing the group. This is illustrated by the component planes of 2 significant variables at RT 6.45 min and 7.27 min (Fig. 7) where the former compound is predominantly found in the inoculated samples and the later is more likely to be present in the control samples. Note that some inoculated samples remain healthy and are comparable to the control samples.

### 4.3. Prediction

In SVR, optimization of three parameters ($C$, $\sigma$ and $\varepsilon$) is important. The parameter $\varepsilon$ regulates the radius of the margin around the regression function; a large value will result in a smoother regression function but the model may not be applicable. The parameter $C$ however, determines the trade-off between the smoothness of the regression function and the amount up to which deviation larger than $\varepsilon$ are tolerated.

The optimization of SVR parameters is performed using 100 training/test splits. The results suggested that the parameters are selected over a range within the boundary constraints. Fig. 8 illustrates the optimized values $C$, $\sigma$ and $\varepsilon$ over 100 iterations with the dashed lines indicating the parameters at 25%, 50%, 75% and 95% quantile of the range.

Fig. 9 shows the SVR solutions for prediction of styrene using the optimized SVR parameters over 100 training/test splits. It is shown that the models are non-linear and the samples could be modelled appropriately with the selected parameters; the mean RMSE is 0.0109 in units of square root and row scaled peak area corresponding to 17.68% of the mean of the square root row scaled

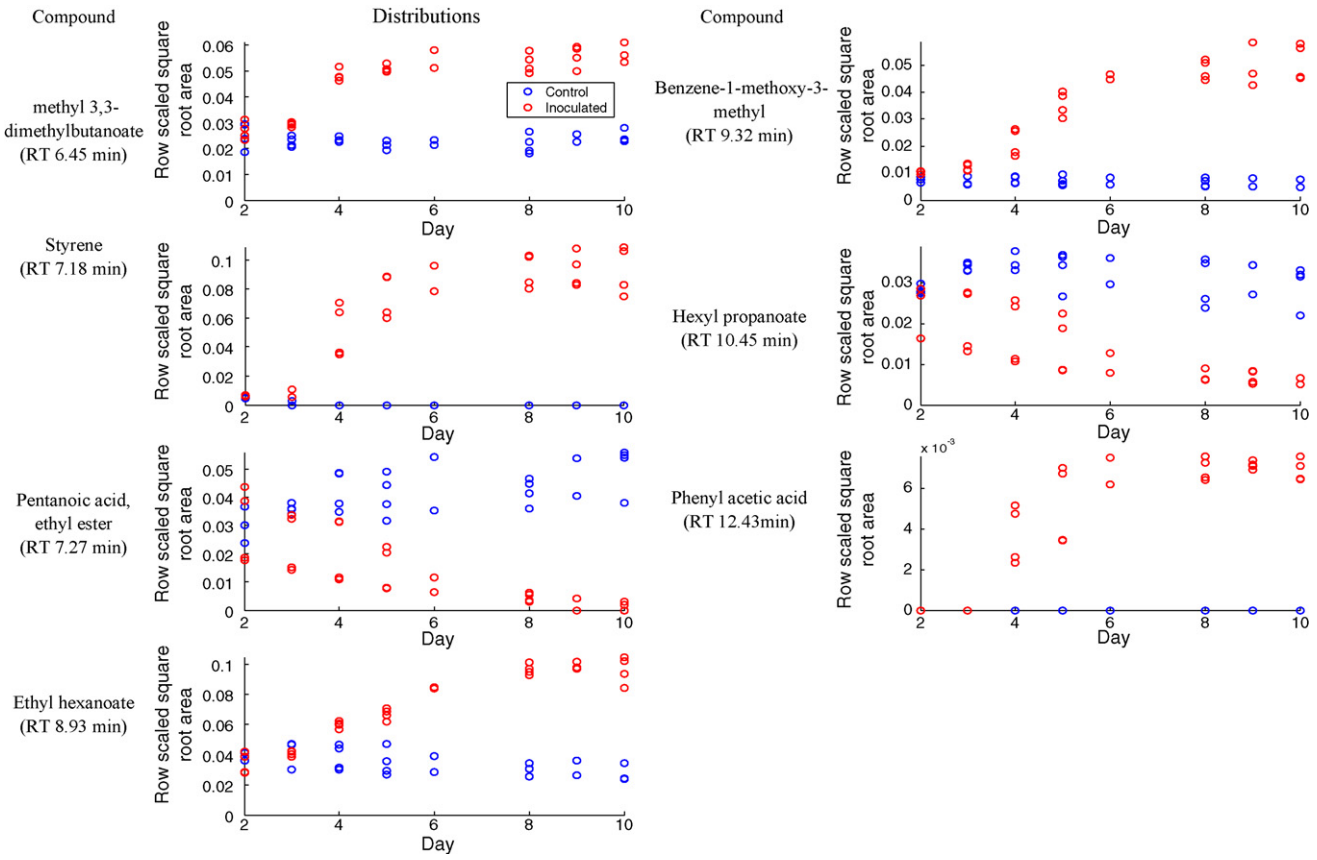



**Fig. 6.** The distribution of significant compounds in control and inoculated samples from 2nd to 10th day using the row scaled square root peak area.

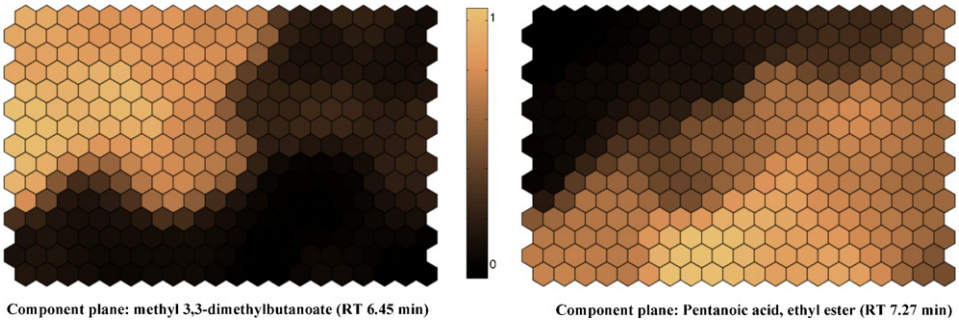



**Fig. 7.** The component planes of two significant variables at RT 6.45 min and 7.27 min.

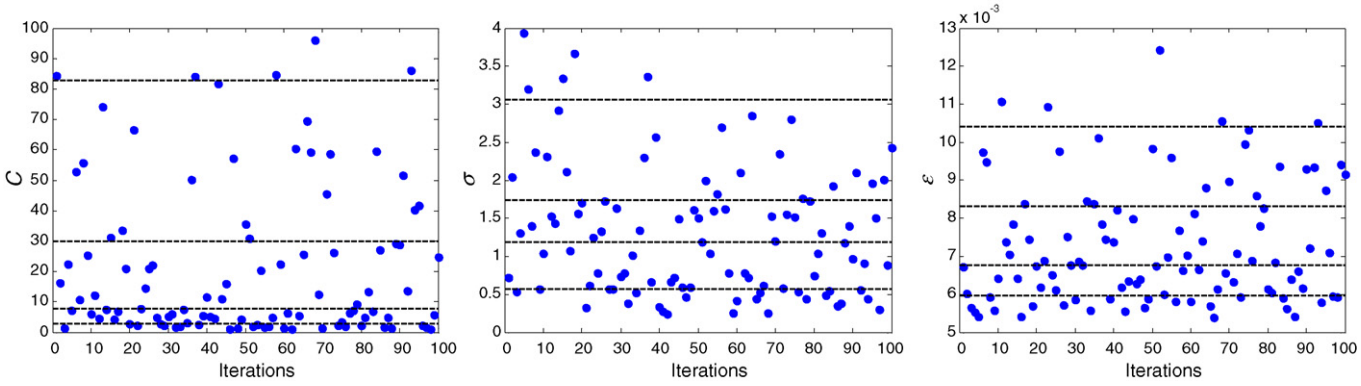



**Fig. 8.** The SVR parameters chosen for prediction of styrene over 100 training/test splits, horizontal lines illustrating the 95%, 75%, 50% and 25% quantiles of the range.

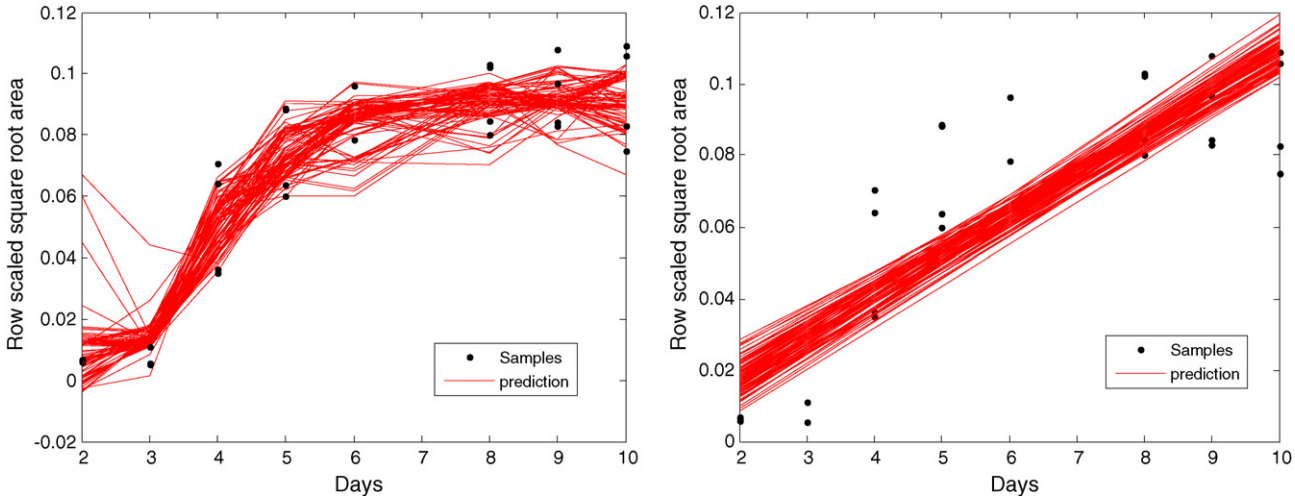



**Fig. 9.** The PLSR and SVR solutions for prediction of styrene over 100 iterations of training/test splits.

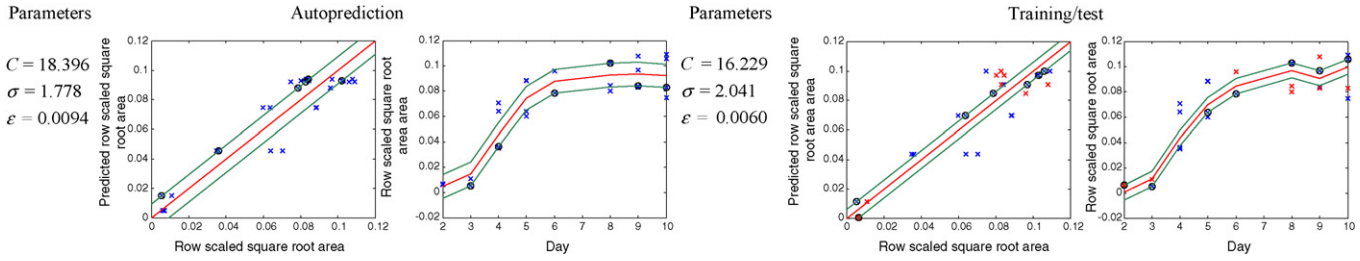



**Fig. 10.** SVR regression lines for styrene. Left: autoprediction, right: using one training/test set split. In all cases parameters are optimized as described in the text. SVs are indicated with circles, the best fit straight line in red and the margins in green. Test set samples are indicated in red as appropriate. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of the article.)

GC–MS peak area for styrene. The corresponding values of RMSEP are 0.0158 corresponding to 25.58%.

The prediction can be performed on the identical training/test splits using PLSR (using one PLS component) to illustrate the difference between SVR and a linear method. For PLSR models, in comparison the RMSEP obtained is 0.0205 corresponding to an error of 33.17% which is greater than that achieved with SVR.

Fig. 10 illustrates the autopredictive and a training set model for the optimal values of the three tuneable parameters obtained as discussed in Section 3.3.2. Apart from being able to accommodate the non-linearity of the data, the fit of the model can be adjusted accordingly using the three tuneable parameters.

## 5. Conclusion

We can see in this paper that modern approaches based on machine learning are readily applicable to mining complex chromatographic datasets, such as the one illustrated relating to fruit spoilage over time. Some adaptations have been described, including a method for optimizing the SVR parameters. SOMs are particularly powerful approaches for visualizing analytical chemical data. It is anticipated that in the next few years both SVR and SOM based methods will have a major role to play in analytical chemistry as the application areas and chromatographic or spectroscopic datasets become even more complex and so harder to interpret by eyeballing because of the volume of data, but computers become more powerful and so able to perform calculations rapidly on a desktop at speeds inconceivable one or two decades ago.

In addition to illustrating the applicability of such approaches to analytical data, we have also described a novel approach for optimizing SVR parameters, and used computationally intense enhancements such as splitting the data 100 times into test and training sets in order to produce consensus models. These additional adaptations of SVR are important in order to apply the available methods to analytical chemistry.

In this paper, SOMs are demonstrated on a relatively simple dataset containing two groups (control and inoculated samples) one of which is further characterized by level of spoilage i.e., healthy, 1st phase and 2nd phase for visualization and variable selection. Since the number of groups is small enough and the data structure is clearly distinguishable, unsupervised SOMs are adequate even for variable selection. When multiclass data with greater complexity is involved, SOMs can be extended to a supervised approach where component planes corresponding to class membership are included for training [37] so the ideas in this paper can be extended to more complex datasets.

## Acknowledgements

We thank Dr. Andrea Pomázi and Dr. Anna Maráz for microbiological preparation. The measurements of the samples was sponsored by Regional Knowledge Centre “Food chain” (project OMFB-01555/2006).

## References

[1] G. Musumarra, M. Fichera, Chemom. Intell. Lab. Syst. 44 (1998) 363–372.
[2] K. Varmuza, Pattern Recognition in Chemistry, Springer, Berlin, 1980.
[3] D. Coomans, I. Broeckaert, Potential Pattern Recognition in Chemical and Medical Decision Making, Research Studies Press, Letchworth, 1986.
[4] L. Munck, L. Nørgaard, S.B. Engelsen, R. Bro, C.A. Andersson, Chemom. Intell. Lab. Syst. 44 (1998) 31–60.
[5] S. Wold, K. Esbensen, P. Geladi, Chemom. Intell. Lab. Syst. 2 (1987) 37–52.

[6] J.E. Jackson, A User's Guide to Principal Components, Wiley, New York, 1991.
[7] I.T. Jolliffe, Principal Components Analysis, second ed., Springer, Berlin, 2002.
[8] K. Esbensen, Multivariate Analysis in Practice, third ed., CAMO, Oslo, 1998.
[9] P. Geladi, B.R. Kowalski, Anal. Chim. Acta 185 (1986) 19–32.
[10] R.G. Brereton, Analyst 125 (2000) 2125–2154.
[11] M. Baker, W. Rayens, J. Chemom. 17 (2003) 166–173.
[12] R.G. Brereton, Chemometrics for Pattern Recognition, Wiley, Chichester, 2009.
[13] S.J. Dixon, Y. Xu, R.G. Brereton, A. Soini, M.V. Novotny, E. Oberzaucher, K. Grammer, D.J. Penn, Chemom. Intell. Lab. Syst. 87 (2007) 161–172.
[14] S. Wold, M. Sjostrom, ASC Symposium, vol. 52, Washington, 1977.
[15] S. Wold, Pattern Recogn. 8 (1976) 127–139.
[16] V.N. Vapnik, The Nature of Statistical Learning Theory, second ed., Springer, New York, 2000.
[17] N. Cristianini, J. Shawe-Taylor, An Introduction to Support Vector Machines and Other Kernel based Learning Methods, Cambridge University Press, Cambridge, 2000.
[18] B. Schölkopf, A.J. Smola, Learning with Kernels, MIT Press, Cambridge, 2002.
[19] R.G. Brereton, G.R. Lloyd, Analyst 135 (2010) 230–267.
[20] T. Kohonen, J. Hynninen, J. Kangas, J. Laaksonen, Technical Report A31, Helsinki University of Technology Laboratory of Computer and Information Science, Espoo, Finland, 1996.
[21] T. Kohonen, Self Organizing Maps, third ed., Springer, Berlin, 2001.
[22] G.R. Lloyd, R.G. Brereton, J.C. Duncan, Analyst 133 (2008) 1046–1059.
[23] L.V. Jørgersen, H.H. Huss, P. Dalgaard, J. Agric. Food Chem. 49 (2001) 2376–2381.
[24] R.L. Wierda, G. Fletcher, L. Xu, J.P. Dufour, J. Agric. Food Chem. 54 (2006) 8480–8490.
[25] A. Boschetti, F. Biasioli, M. van Opbergeb, C. Warneke, A. Jordan, R. Holzinger, P. Prazeller, T. Karl, A. Hansel, W. Lindinger, S. Iannotta, Postharvest Biol. Technol. 17 (1999) 143–151.
[26] H.J.D. Lalel, J. Singh, S.C. Tan, Postharvest Biol. Technol. 27 (2003) 323–336.
[27] P. Ragaert, F. Devlieghere, E. Devuyst, J. Dewulf, H. Van Langenhove, J. Debevere, Int. J. Food Microbiol. 112 (2006) 162–170.
[28] Y.Y. Voon, N.S.A. Hamid, G. Rusul, A. Osman, S.Y. Quek, Postharvest Biol. Technol. 46 (2007) 76–85.
[29] R. Bouvier-Brown, K. Holzinger, A. Palitzsch, H. Goldstein, J. Chromatogr. A 1161 (2007) 113–120.
[30] Sourceforge.net: http://mexcdf.sourceforge.net/.
[31] H.F.M. Boelens, R.J. Dijkstra, P.H.C. Eilers, F. Fitzpatrick, J.A. Westerhuis, J. Chromatogr. A 1057 (2004) 21–30.
[32] J.W.H. Wong, C. Durante, H.M. Cartwright, Anal. Chem. 77 (2005) 5655–5661.
[33] S.J. Dixon, R.G. Brereton, H.A. Soini, M.V. Novotny, D.J. Penn, J. Chemom. 20 (2006) 325–340.
[34] R.G. Brereton, Data Analysis for the Laboratory and Chemical Plant, Wiley, Chichester, 2003.
[35] S.J. Dixon, Y. Xu, R.G. Brereton, H.A. Soini, M.V. Novotny, E. Oberzaucher, K. Grammer, D.J. Penn, Chemom. Intell. Lab. Syst. 87 (2007) 161–172.
[36] R.A. Van den Berg, H.C.J. Hoefsloot, J.A. Westerhuis, A.K. Smilde, M.J. van der Werf, Genomics 7 (2006) 142–157.
[37] K. Wongravee, G.R. Lloyd, C.J. Silwood, M. Grootveld, R.G. Brereton, Anal. Chem. 82 (2010) 629–639.
[38] G.R. Lloyd, K. Wongravee, C.J.L. Silwood, M. Grootveld, R.G. Brereton, Chemom. Intell. Lab. Syst. 98 (2009) 149–161.
[39] D. Li, G.R. Lloyd, J.C. Duncan, R.G. Brereton, J. Chemom. 24 (2010) 96–110.
[40] N. Das, Int. J. Adv. Manuf. Technol. 41 (2009) 799–807.
[41] S.R. Gunn, Support vector machine for classification and regression, Technical Report, University of Southampton, 1998.
[42] A.J. Smola, B. Schölkopf, Stat. Comp. 14 (2004) 199–222.
[43] S.M. Clarke, J.H. Griebsch, T.W. Simpson, J. Mech. Des. 127 (2005) 1077–1087.
[44] Y. Xu, S. Zomer, R.G. Brereton, Crit. Rev. Anal. Chem. 36 (2006) 177–188.
[45] J. Li, Z. Cai, Advances in Computation and Intelligence, Springer, Berlin, Heidelberg, 2008, pp. 510–519.
[46] S.K. Lahiri, K.C. Ghanta, Chem. Ind. Chem. Eng. Quart. 14 (2008) 191–203.
[47] R. Storn, K. Price, J. Global Optim. 11 (1997) 341–359.
[48] H. Wold, Encyclopedia of Statistical Sciences, Wiley, New York, 1984.
[49] P. Geladi, B. Kowalski, Anal. Chim. Acta 185 (1986) 19.
[50] T.O. Larsen, J.C. Frisvad, Mycol. Res. 99 (1995) 1153–1166.
[51] K. Karlshøj, P.V. Nielsen, T.O. Larsen, J. Agric. Food. Chem. 55 (2007) 4289–4298.
[52] R. Schwalbe, M. Moller, R. Ostrowski, W. Dott, Chemosphere 39 (1999) 795–810.
[53] I. Lara, J. Graell, M.L. López, G. Echeverría, Postharvest Biol. Technol. 39 (2006) 19–28.
[54] C. Garcia, A. Martin, M.L. Timón, J.J. Córdoba, Lett. Appl. Microbiol. 30 (2000) 61–66.
[55] M. Moalemiyan, A. Vikram, A.C. Kushalappa, Postharvest Biol. Technol. 45 (2007) 117–125.