Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy 251 (2021) 119440

Contents lists available at ScienceDirect

Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy

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

Network analysis on Fourier-transform infrared (FTIR) spectroscopic data sets in an Eigen space layout: Introducing a novel approach for analysing wine samples

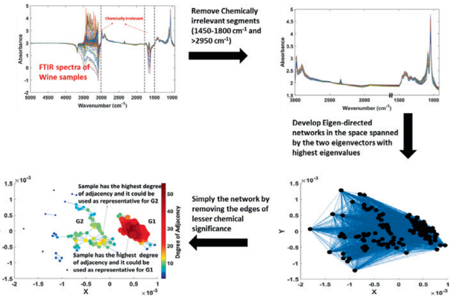
Keshav Kumar, Anja Giehl, Ralf Schweiggert, Claus-Dieter Patz *

Geisenheim University, Department of Beverage Research, Analysis and Technology of Plant-based Foods, Von Lade Str. 1, D-65366 Geisenheim, Germany

HIGHLIGHTS

- Eigen directed network analysis for FTIR spectroscopic data sets of wine samples was proposed.
- It was tested by analyzing a collection of 148 wine samples.
- It provided aesthetic values and chemical significance to the nodes positioning.
- It allowed an easy assessment regarding inter-and intra-group homogeneity of analyzed samples.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 21 October 2020
Received in revised form 30 December 2020
Accepted 3 January 2021
Available online 06 January 2021

Keywords:
Eigen-directed layout
Network
Wine
Homogeneity
Degree of adjacency

ABSTRACT

In the present work, Eigen-directed network analysis for Fourier-transform infrared (FTIR) spectroscopic data sets of wine samples was introduced. A network can generally be viewed as a collection of nodes connected to each other through links, often also called edges. Herein, each node in the network represents a sample and the dissimilarity weight associated with the difference between the two connected nodes is described by the edge. The utility of the approach was tested by analysing a collection of 148 wine samples. The networking on FTIR data sets of these samples in the Eigen space layout was found to impart required aesthetic values as well as the chemical significance to the nodes positioning. The proposed approach successfully captured the compositional differences among the analysed wine samples and classified them in two groups. The Eigen-directed network analysis also allowed a swift assessment regarding inter- and intra-group homogeneity. Homogeneous groups were found to contain nodes with high degree of adjacency and edges with smaller lengths. In comparative study, the proposed approach was found to outperform the network analysis in force-directed layout and principal component analysis. In summary, the proposed Eigen-directed network analysis provided a simplified illustration of highly correlated spectral data sets enabling a swift and intuitive interpretation.

1. Introduction

Wine is an hydroethanolic mixture of numerous chemical components such as sugars, glycerol, organic acids, and polyphenols in varying concentrations [1–6]. Generally, wine composition is heavily influenced by variety, geographical origin, vintage, and production procedures [5–12]. Thus, wine analysis becomes an inevitable step to ensure quality and authenticity to consumers. Fourier-transform infrared (FTIR) spectroscopy [13,14], a fast, simple, sensitive and non-destructive technique, has been successfully used for the wine analyses [1,5,6,9,10–12,15–20]. However, continuous improvements are required in the data analysis protocols implemented for FTIR spectroscopy in order to achieve better and

* Corresponding author.
E-mail address: claus.patz@hs-gm.de (C.-D. Patz).

https://doi.org/10.1016/j.saa.2021.119440

intuitive interpretations of the obtained results. Principal component analysis (PCA) [21–23] can discriminate the samples in an unsupervised manner, but it does not provide any information about group homogeneity (or heterogeneity). Such information, however, could be useful to understand the complexity of the sample set and to identify potentially most representative samples in a given group, i.e. to identify the "core" of a collection. Multidimensional scaling [6,24,25] can also successfully discriminate samples in an unsupervised manner, but the significance associated with the placement of the samples along horizontal and vertical axes of the target space remains unclear. Agglomerative cluster analysis [26,27] is another technique providing an unsupervised classification of samples. However, no information related to the inter-separation of the groups is obtained and, moreover, a combination of linkage criteria and distance metrics are required to be optimized, making the entire process intricate and subjective. Similarly, self-organizing maps (SOM) algorithms [27–30] can separate samples in different groups, but they do not provide any information regarding intra-variability. Thus, data classification approaches that not only provide an intuitive interpretation of the data by visualizing related groups, but simultaneously convey information regarding the homogeneity of the groups have been scarcely explored.

Therefore, Eigen-direct network analyses based on FTIR spectral data might be useful to achieve a better understanding of the compositional relationship of analyzed wine samples. Such networks are essentially a collection of the nodes connected through edges (syn. links) [31–40]. Each node of the network represents a sample, while the connecting edges weighted with dissimilarity values are indicative for the interaction between the samples [31–40]. For example, an edge associated with a weight value of a smaller magnitude will essentially indicate a stronger interaction between the connected nodes. The degree of adjacency for a given node is the total number of nodes that are directly connected to it [31–40]. A node having the highest degree of adjacency essentially suggests that it shares similarity with a maximum number of other nodes, thus presumably acting as the network's central point. Visual interpretations of the networks are quite subjective to the layout used to construct the network [41–45]. Some of the commonly used network layouts are (i) layered [46,47] (ii) subspace embedding [48] and (iii) force-directed ones [49,50]. Of these, force-directed layouts are most commonly used. They consider the nodes as "charges" and edges as "springs", creating the network by balancing electrostatic repulsive forces of the charges (nodes) with attractive forces exerted by the springs (edges). Furthermore, the force-directed layout minimizes the crossing of the edges, evenly distributes the nodes, ensures uniform edge length and conformation to frame area. It most frequently provides an aesthetically satisfying and simple network for the visualization of the data sets. However, force-directed layouts (and with other layouts) often tempt users to misinterpret the networks due to the following issues: (i) the spatial positioning of the nodes in these layouts does not reflect relationships among the samples, i.e. two similar nodes could be far from each other and, vice versa, two distinct nodes may appear in the neighbourhood of each other. This is mainly because the goal of the force-directed layout is providing visually aesthetic networks. (ii) As a result of (i), it is not always possible to assign a particular area in the coordinate frame to a particular group of samples. (iii) The placement of the nodes could be different even if the network analysis is carried out on analytical replica of identical sample data sets, i.e. it is difficult to re-generate the same network. These practical issues impose limitations on its utility as a user-friendly unsupervised pattern recognition technique.

In order to address these issues, the present work suggests an alternative approach that involves the construction of the networks in a layout spanned by eigenvectors. It is expected that Eigen-directed networking would ensure the desired aesthetic values to the developed networks and simultaneously provide meaning to the relative positions of the nodes. In other words, it essentially means that if two nodes in an eigenvector-spanned layout are spatially co-located closely, then it could safely be considered similar and vice versa. It would also enable grouping of the nodes appearing in different regions of the network and retaining the significance of the placement of the samples in horizontal and vertical axes. The specific objective of the present work is to propose a simple, cost-effective, and user-friendly approach by combining the FTIR spectroscopy with Eigen-directed network analysis for subsequent classification of the wine samples.

2. Material and methods

2.1. Wine samples and FTIR measurements

In the present work, a set of 148 wine samples collected form German wine distributors were used. The brand names of the wine samples were anonymized. These 148 wine samples were all derived from grapevine (*Vitis vinifera* L.) could be segregated into six groups, i.e. red wine (48 samples), rosé wine (13 samples), white semi-sparkling wine (3 samples), red semi-sparkling wine (3 samples), white sparkling wine (10 samples) and white wine (71 samples). The name (incl. cultivar if given) and origin of each wine is reported in Table S1 of the supporting information file. Prior to spectral measurements, eventual undissolved impurities and CO_2 was removed by gravity filtration using the Whatman filter paper (Grade 595, 185 mm). The prepleated Whatman filter paper was fitted to a glass funnel and the wine samples were passed through it. FTIR spectral data sets for each sample were acquired using Winescan ™ SO_2 Foss instrument equipped with Foss Integrator software (Foss, Hillerød, Denmark). Each of the spectra consisted of 1060 data points acquired over the pin-numbers 240–1299. The pin numbers were converted to wavenumbers (925.92–5011.5 cm^{-1}) by multiplying each pin number with a constant factor of 3.858. The reproducibility of FTIR spectra for each of the 148 samples was confirmed by acquiring the data in duplicity. The detailed discussion on the analytical procedures that must be followed while carrying out the FTIR measurements of wine samples could be seen in our previously reported works [15,16].

2.2. Computational platform

The computational work and network analyses were carried out using MATLAB (2016b) platform. The developed MATLAB code is provided in the Appendix.

3. Results and discussion

3.1. FTIR data trimming and pre-processing

The FTIR spectra for each of the analyzed 148 wine samples are shown in Fig. 1a. Signals of several minor wine components such as phosphates, phenolics, phenyl derivatives, unsaturated lipids could mainly be seen in the wavenumber range < 1000 cm^{-1} [51]. Signals of glucose, oligosaccharides, polysaccharides, alcohols including ethanol could mainly be seen in the wavenumber range 970–1100 cm^{-1} [51]. Fatty acids, polyols *etc.* could be seen in the wavenumber range 2800–2935 cm^{-1} [51]. Apart from these chemical components, FTIR spectra also contain certain irrelevant segments that could mainly be seen in the wavenumber ranges 1450–1800 cm^{-1} and >2950 cm^{-1} that must be eliminated before subjecting the FTIR spectral data sets to Eigen-directed network

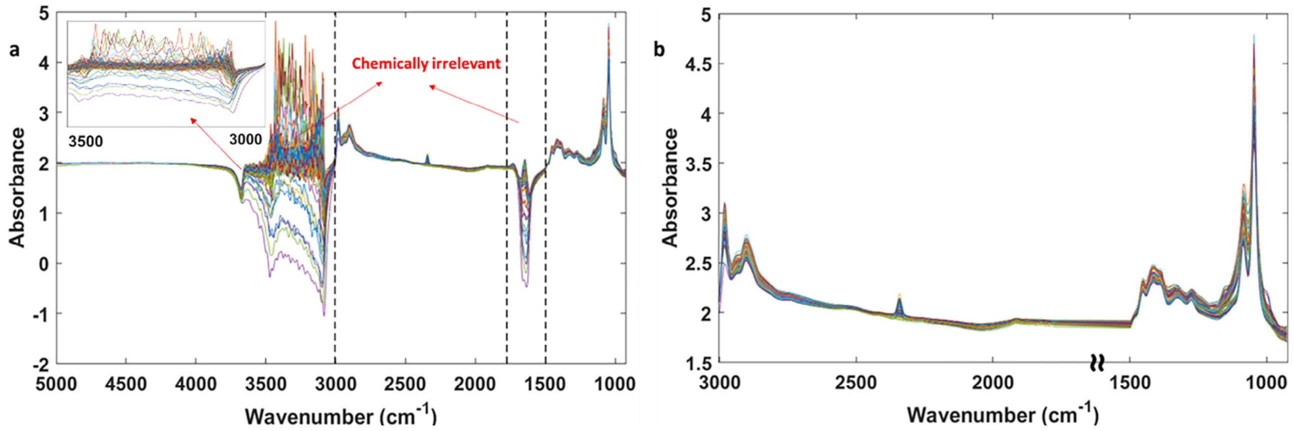


Fig. 1. (a) Raw (zoomed-in spectral range 3000–3500 cm^{-1} is shown in the inset) and (b) pre-processed FTIR spectra of 148 wine samples. The preprocessing involved the elimination of the chemically irrelevant segments (1450–1800 cm^{-1} and >2950 cm^{-1}) that mainly arise due to the absorbance of infrared radiation by water molecules.

analysis. The zoomed-in spectral range of 3000–3500 cm^{-1} is shown in the inset of Fig. 1a. These segments mainly contain signals arising due to absorbance of infrared radiation by water molecules. The trimmed spectra, shown in Fig. 1b, were then processed further as described in the following.

The trimmed FTIR spectral data sets of the wine samples were arranged in a two dimensional data matrix X of dimension 148 × 468 (sample × variable [absorbance at resp. wavenumber]). Prior to that, spectra were normalized to unit area by dividing each variable with the sum of all the variables. This normalization of each spectra was carried out to account for errors (or technical variabilities) introduced during sample preparation and instrument fluctuation.

3.2. Working scheme of Eigen-directed network analysis

Due to the above-mentioned issues with force-directed layouts of network analyses, we sought to create a network in an eigenvector-spanned spatial layout. The working scheme of this Eigen-directed network analysis can be described with the below listed steps:

(i) In the first step, the eigenvectors for the data matrix X ($I \times J$, i.e. sample × variable) must be obtained using the Eigen analysis. First, the data matrix X should be mean centered and a diagonalizable square matrix Y of dimension ($J \times J$) can be obtained (see MATLAB code in Appendix) using Eq. (1).

$$Y = \frac{X^T X}{I - 1} \tag{1}$$

Then, the square matrix Y is subjected to Eigen analysis (see MATLAB code in Appendix). The decomposition of matrix Y could be summarised using the Eq. (2), where E is the matrix with eigenvectors and λ is the matrix containing the eigenvalues.

$$YE = E\lambda \tag{2}$$

(ii) In the second step, the original data matrix X is projected into a two dimensional space spanned by the two most significant eigenvectors (i.e. the vectors of matrix E associated with the first two highest eigenvalues). These sets of eigenvectors serve as the basis for the layout to construct the network. It is to be noted that the data matrix X could basically also be projected into a space spanned by any other pair of eigenvectors. However, the first two eigenvectors often provide the most valuable information. Thus, in the following steps, we describe the scheme of Eigen-directed network analysis with them. The projection can be summarised using Eq. (3). The matrix N contains the node position for each of the analyzed sample in the Eigen-directed space.

$$N = Xe \tag{3}$$

(iii) In the next step, using the nodes position (summarised in matrix NN (see MATLAB code given in Appendix)) along the user specified eigenvectors, a full network is developed where all the nodes are connected to each other through edges. The Euclidean distance between a pair of nodes could be a measure of the weight associated with the edge connecting them. It can easily be followed that small Euclidean distance represent great similarity between the nodes and vice versa. Thus, one could further simplify the network by discarding connections of lower importance. This can be achieved by calculating the pairwise Euclidean distance matric for the given set of nodes and all the connections having the distances that are greater than a user-specified threshold value (see "Quantile" in MATLAB code given in Appendix) are discarded. The Euclidean distance between N_i th and N_k th nodes in the space spanned by a given pair of eigenvectors could be obtained using the Eq. (4).

$$d_{ik} = \sqrt{(N_{i1} - N_{k1})^2 + (N_{i2} - N_{k2})^2} \tag{4}$$

In the above equation, N_{i1} and N_{k1} are the positions along the first eigenvector and N_{i2} and N_{k2} are the position along the second eigenvector for the N_i th and N_k th nodes in the Eigen space.

(iv) The collection of the nodes that are appearing in different regions of the Eigen space could be classified as separate groups.

(v) The degree of adjacency of a node, that is equal to the sum of the number of nodes directly attached to it, is generally shown in the network by making the nodes of different sizes in a color-coded manner. A node in the network, with biggest size and darkest color would indicate a high degree of adjacency, i.e. that a sample corresponding to such a node have the highest similarity with other samples and can safely be considered as the group's most representative sample.

The above described steps are summarised in Fig. 2. The MATLAB code developed for carrying out the network analysis in the Eigen defined space is given in the Appendix of the current work.

3.3. Application of the Eigen-directed network analysis on FTIR spectral data sets of wine samples

After pre-processing FTIR spectral data sets of 148 wine samples and their arrangement in a two-dimensional matrix X of size 148 × 468 (sample × variable), the mean-centered data was subjected to the Eigen analysis and used for Eigen-directed network analysis according to the above-described procedure. The positions

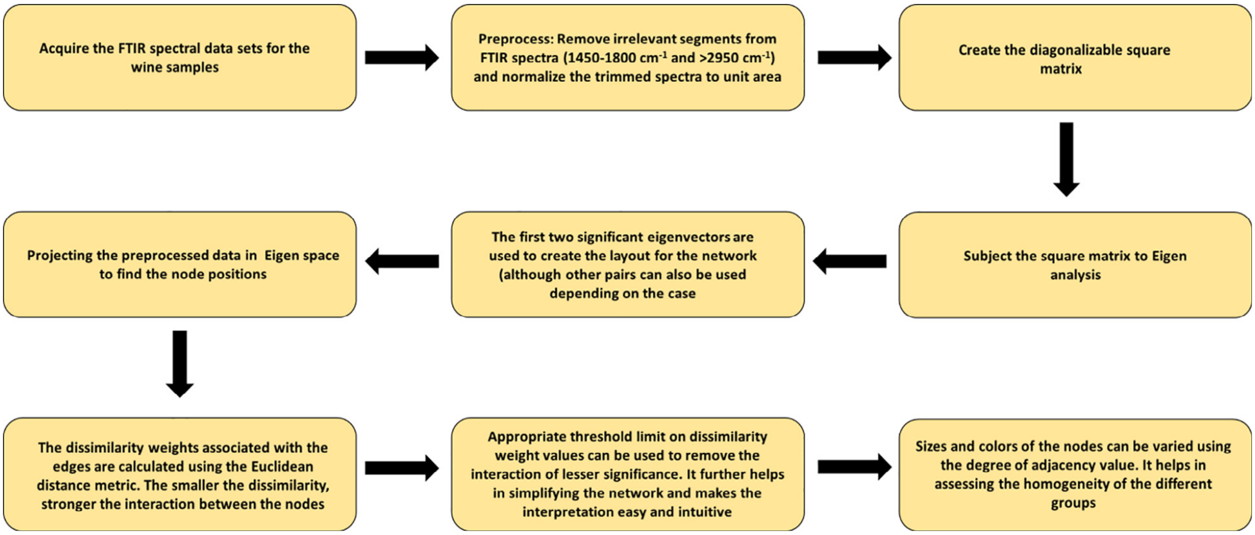

Acquire the FTIR spectral data sets for the wine samples
Preprocess: Remove irrelevant segments from FTIR spectra (1450-1800 cm-1 and >2950 cm-1) and normalize the trimmed spectra to unit area
Create the diagonalizable square matrix
Subject the square matrix to Eigen analysis
The first two significant eigenvectors are used to create the layout for the network (although other pairs can also be used depending on the case
Projecting the preprocessed data in Eigen space to find the node positions
The dissimilarity weights associated with the edges are calculated using the Euclidean distance metric. The smaller the dissimilarity, stronger the interaction between the nodes
Appropriate threshold limit on dissimilarity weight values can be used to remove the interaction of lesser significance. It further helps in simplifying the network and makes the interpretation easy and intuitive
Sizes and colors of the nodes can be varied using the degree of adjacency value. It helps in assessing the homogeneity of the different groups

Fig. 2. Schematic of Eigen-directed network analysis for assessing FTIR spectral data sets of wine samples.

for the 148 nodes in the Eigen-directed layout was obtained using Eq. (3). The resulting network having all possible aesthetic values is shown in Fig. 3a. In this first network provided, each node is connected to the remaining 147 nodes. To achieve a better interpretation of the network, it was further simplified by removing certain edges (links) having large associated distance (i.e. weight) values. The weight associated with each edge ($^{148}C_2$= (148! / (146! × 2!)) = 10,878) was plotted in Fig. 3b. To reduce the number of edges to be plotted, the quantile values corresponding to 0.95, 0.90, 0.75, 0.50, 0.25, 0.20, 0.15, 0.10 and 0.05 were selected as the thresholds to create nine different networks, shown in Fig. 4-a-i, with varying complexity. It could be seen that complexity of the network reduces as the threshold values were varied from the higher to lower quantile values. The networks corresponding to the threshold values of 0.95, 0.90, 0.75 and 0.5 quantiles (Fig. 4) are complex and difficult to interpret due to an overwhelming number of displayed edges (links). The network corresponding to a quantile of 0.20 (Fig. 4) was simplified to a large extent, but it still contained some edges that had connected the two distinct groups. The network corresponding to threshold value of 0.15 quantile was found to take care of this issue and allowed the removal of these undesired edges connecting two apparently distinct groups, thus consequently enhancing the interpretability of the network. The network corresponding to the threshold values of 0.10 and 0.05 was found to be oversimplified due to insufficiently connected nodes within the groups. The oversimplification yielded an apparently wrong impression, suggesting high heterogeneity in the collection of the analyzed samples.

Ideally, a network must (i) reflect the major differences among the analyzed sample set, (ii) have minimum number of small groups, and (iii) have a chemical significance to node positioning. Based on the above specfied criteria, one could easily select the network corresponding to the threshold value of 0.15 (Fig. 4) quantile value as the optimum model.

Then, the network was further enriched with the information by varying the size and color intensity in proportion to their degree of adjacency. The modified network consisting of varying size and the colour-coded node is shown in Fig. 5. One set of nodes labeled as G1 could mainly be seen in the region spanned over X (0.00094, 0.00016) and Y (-0.00057, 0.00052). The second set of nodes labelled as G2 could be seen in the region spanned over X (-0.0012, 0.00015) and Y (-0.00066, 0.00092). The nodes in the G1

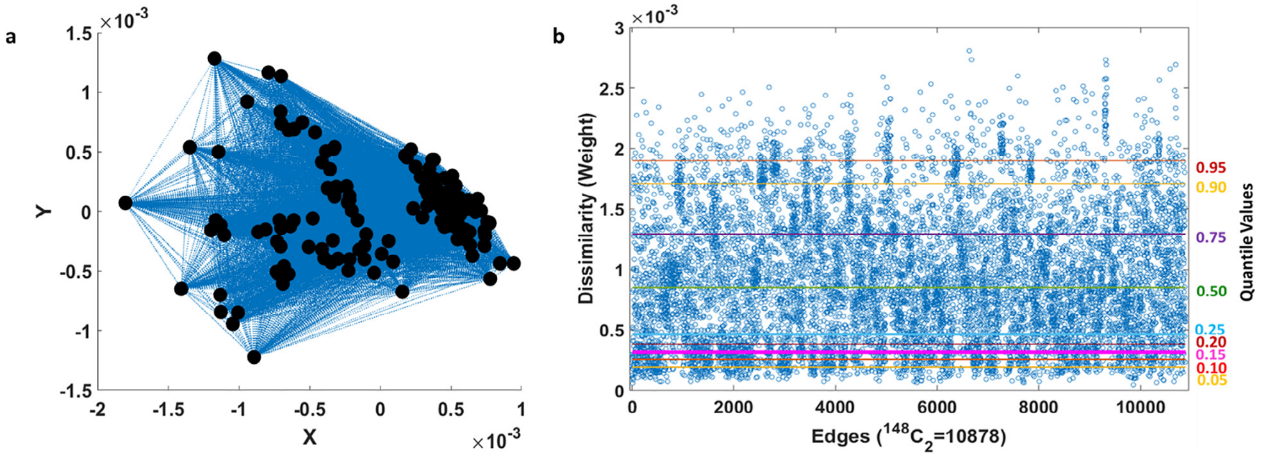


Fig. 3. (a) The Eigen-directed network analysis of FTIR spectral data sets of 148 wine samples, showing the network plot within the space spanned by the two eigenvectors of the sample matrix with the highest eigenvalues. Each one of the 148 nodes is attached to all of the remaining 147 nodes. The network could further be simplified by removing the edges that are chemically irrelevant (cf. Fig. 4). It can be achieved by using a suitable threshold value on the dissimilarity weights associates with different edges. (b) The dissimilarity weight associated with different edges. The threshold limits are represented by different quantile values, being used for the simplification of the networks.

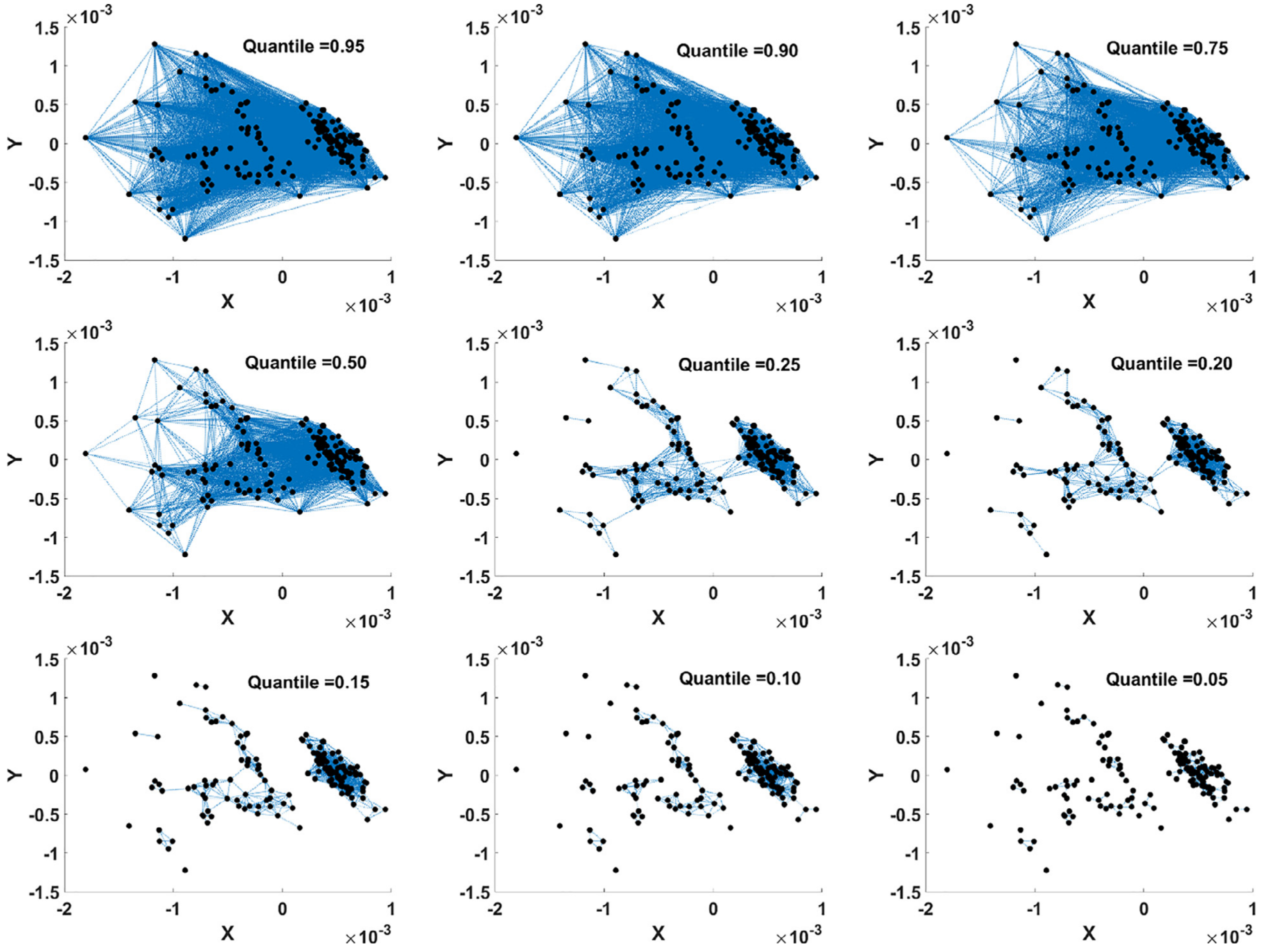


Fig. 4. The networks corresponding to 0.95, 0.90, 0.75, 0.50, 0.25, 0.20, 0.15, 0.10 and 0.05 quantile values used as the thresholds for removing the edges of lesser significance. It can be seen that for the quantile range 0.95–0.5, the network is complex, making the interpretation difficult. The network corresponding to the quantile values 0.25 and 0.20 are largely simplified, but they still contain edges connecting two apparently distinct groups of node collections. In contrast, the network corresponding to quantile values 0.10 and 0.05 values are oversimplified and overestimate the heterogeneity in the analyzed wine samples. The network corresponding to 0.15 quantile value provide a clear separation of the node collection in two groups. It could be preferred over other networks for analyzing the FTIR data sets of the wine samples.

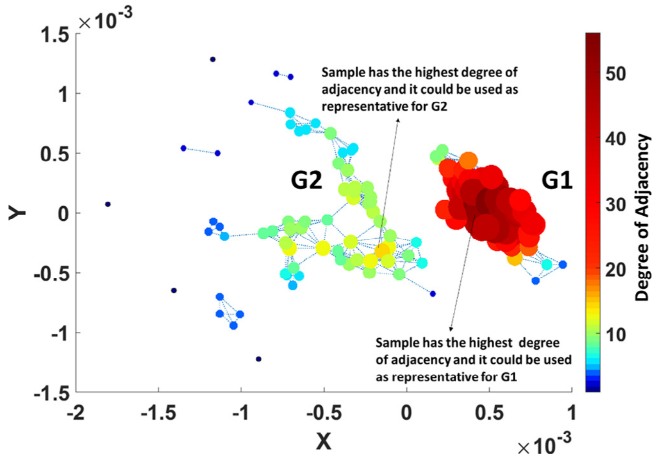


Fig. 5. Eigen-directed network corresponding to the threshold derived from a 0.15 quantile value. The network has the desired aesthetic value (minimum edge crossing, chemical relevance to the node positioning etc.) and provides simple and intuitive interpretation of the analyzed samples. The sizes of the nodes are varied and color-coded based on their degree of adjacency. The network clearly shows two separate collections (groups) of nodes labelled as G1 and G2. The G1 collection was found to be highly homogenous, where the nodes have high degree of adjacency.

collections were closely co-located with a high degree of the adjacency (>40), suggesting the wine samples corresponding to these sets of nodes have high compositional similarities. The sample corresponding to the node with 55 degrees of adjacency, marked in the network, could be taken as the representative sample (i.e. core) for the G1 collection. On the other hand, the nodes in the G2 collection were widely dispersed and the associated edges have higher dissimilarity weights. Furthermore, the degrees of adjacency for the nodes in G2 collection were mainly found to vary in a low range of 5 to 10. Compared to G1, such low degree of adjacency clearly reflected the heterogeneity among the samples of the G2 collection. The heterogeneity also made it difficult to find a representative sample for this collection in an unambiguous manner. Nevertheless, the G2 sample corresponding to the node with highest degree of adjaceny of 15 (marked in Fig. 5) could be considered as the representative of G2 collection. As discussed earlier, the Eigen-directed networking imparts the chemical significance to the node-positions. Thus, all the samples appearing in the region spanned over X(−0.002, −0.001) and Y(−0.0015, 0.0015) were found to belong to the G2 collection.

To understand the spectral and chemical basis for the node positioning of the G1 and G2 collection, their normalized FTIR spectral profiles (Fig. 6) were analysed. It can be seen that two groups mainly differed from each other in the wavenumber range

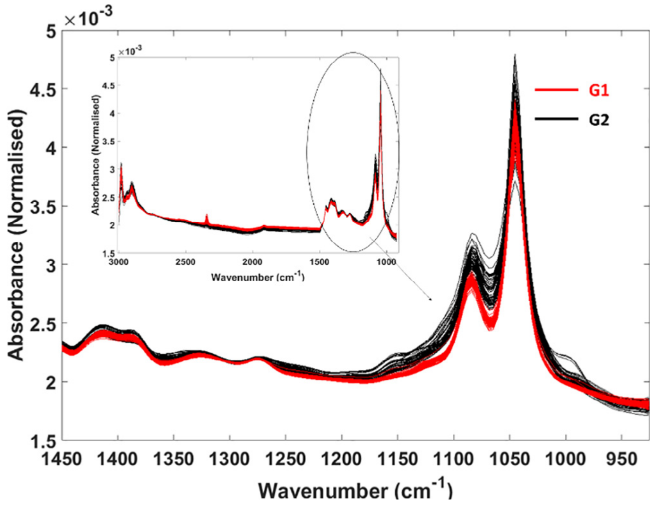


Fig. 6. Normalized FTIR spectra of the G1 and G2 groups. The two groups G1 and G2 differ in the spectral region 970–1100 cm^{-1}. It can be seen that compared to those of the G2 collection, FTIR spectra of the samples of the G1 collection show a lower absorbance in this region, suggesting relatively low abundance of ethanol, phosphates, phenolics, phenyl derivatives, unsaturated lipids and saccharides.

970–1100 cm^{-1} that mainly correlate to ethanol, phosphates, phenolics, phenyl derivatives, unsaturated lipids, saccharides, oligosaccharides, and polysaccharides. Compared to the G2 collection, FTIR spectra of the samples belonging to the G1 collection have in common a rather low absorbance in the aforementioned region, suggesting a relatively low abundance of above-mentioned chemical components. The samples of G2 were also found to have relatively higher extract than those in G1 samples. The obtained results suggest that Eigen-directed network analysis on FTIR spectral profiles can capture and illustrate relevant compositional differences, and hence could serve as a simple, fast and user-friendly analytical approach for wine analysis.

3.4. Comparing Eigen-directed network analysis with force-directed network analysis and PCA

In order to demonstrate potential advantages and limitations, we compared our proposed approach of Eigen-directed networks analyses with force-directed network analysis and PCA. The network resulting from force-directed network analysis, wherein the nodes were placed in a manner that they are as far as possible with minimum edge crossing, is given in Fig. 7a. The shown network was found to have a high aesthetic value. However, to facilitate its interpretation, it was necessary to further simplify by removing edges being chemically irrelevant. By analogy to the above described Eigen-directed network analysis, nine different networks were created (Fig. 8), in which edges with dissimilarity weights greater than 0.95, 0.90, 0.75, 0.50, 0.25, 0.20, 0.15, 0.10 and 0.05 quantiles, respectively, were considered insignificant and removed. Please note that, in Fig. 8, the color and size of nodes were varied based on their degree of adjacency. Unlike Eigen-directed network analysis, wherein the node positioning was invariant to the threshold used for simplifying the network, here the node positions were found to vary in different networks, rendering the interpretation related to the placement of nodes along the horizontal (X) and vertical (Y) axes highly intricate. Furthermore being unlike Eigen-directed network analysis, here the displacement between the nodes was also found to be uninterpretable. It was also difficult to assign a particular space in the networking plot to a particular group of wine samples. In the network employing a 0.20 quantile, the samples were found to be separated in two groups, although it still had edges joining them. However, it failed to provide visual depiction of intra-group variability present among the analyzed wine samples. In addition, it was also difficult to assign a total of fourteen samples (two subgroups of four wine samples, one subgroup of two wine samples and 4 independent samples) to any of the two major groups, because the node positioning in the force-directed layout lacked the chemical relevance and interpretation. Whereas, in the Eigen-directed network, it was possible to assign each sample to a particular group based on their node positioning. It was also possible to visualize and make interpretation regarding the within group variability of the analyzed wine samples. Lastly, a PCA score plot comprised by the first (PC1) and second (PC2) principal components, shown for comparison in Fig. 7b, discriminated the samples in same manner, as both PCA and Eigen-directed network analysis involved the Eigen analysis of the same covariance matrix (obtained from the same data pre-processing steps normalization followed by data mean-centering). However, the PCA score plot failed to provide any measure of group homogeneity and level of interactions among the analyzed wine samples. These issues were easily addressed in the Eigen-directed network analysis where the group homogeneity was assessed using the degree of adjacency. The group homogeneity information

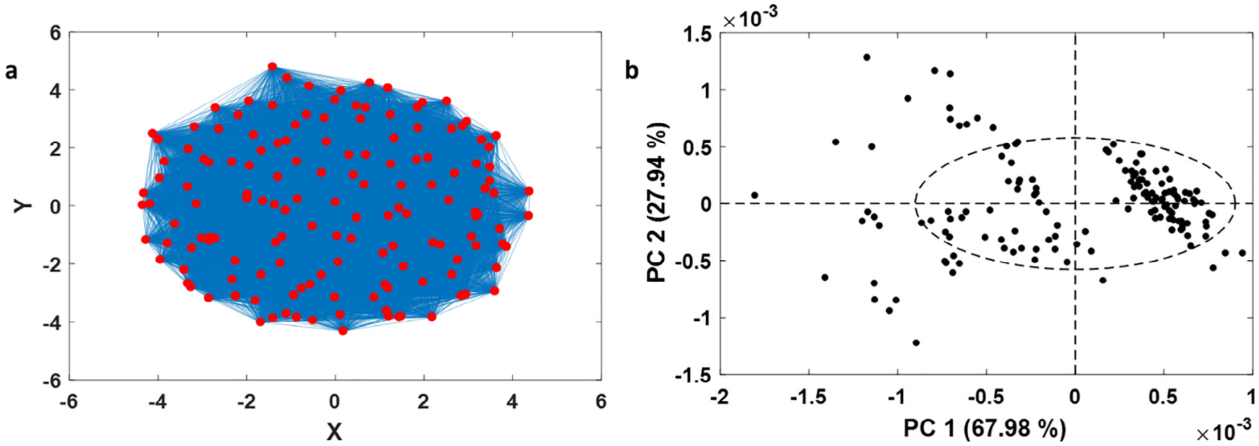


Fig. 7. (a) Network in force-directed layout with each node connected to remaining 147 nodes. The nodes are placed with minimum edge crossing so that network has all the aesthetic values, however a simplification is required to make any chemical interpretation (b) PCA score plot comprising PC1 and PC2 provided the same classification that was obtained from Eigen-directed network analysis (as both approaches involved Eigen analysis of the same covariance matrix). However, PCA failed to provide the visual assessment (i) of group homogeneity and (ii) level of interactions among the analyzed wine samples.

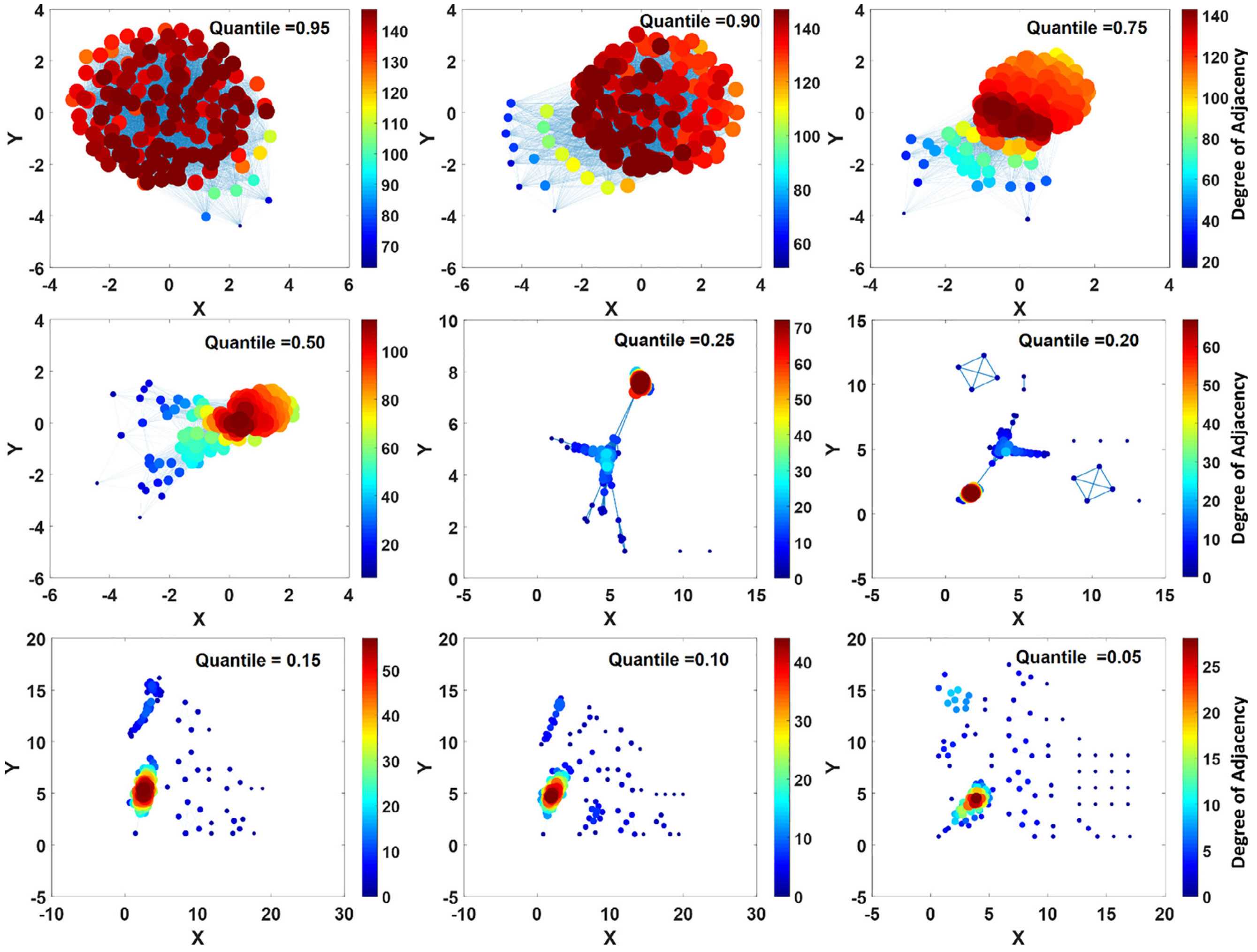


Fig. 8. The networks in force-directed layout corresponding to 0.95, 0.90, 0.75, 0.50, 0.25, 0.20, 0.15, 0.10 and 0.05 quantile values used as the thresholds for removing the edges of lesser significance. The node positioning was found to be quite sensitive to the threshold limits used for simplifying the network. The node positions along the X and Y coordinate had no chemical interpretation.

was also successfully depicted by color-coding and varying the node sizes for visual perception. The Eigen-directed network wherein the nodes having chemically meaningful interactions were shown to be connected through edges. The length of edge in Eigen-directed layout provided a means of assessing the level of interactions i.e. shorter the length stronger the interaction. The comparative study among the Eigen-directed network analysis, force-directed layout network analysis and PCA have shown that Eigen-directed network analysis provides simple and meaningful means of analyzing the FTIR spectral data sets of wine samples.

4. Conclusion

In the present work, an Eigen-directed network analysis approach for the analysis of FTIR spectral profiles of wine samples was introduced. The Eigen-directed layout was found to provide an aesthetic network with minimum edge crossing. It also imparted the chemical significance to the node positioning that was subsequently used to differentiate the wine samples. The Eigen-directed networking also allowed the assessment of the homogeneity of the different groups. A homogenous group was found to have multiple nodes with a high degree of adjacency and edges with smaller lengths. In summary, the present work provided a simple and easily interpretable approach for the analysis of FTIR spectral data sets of the wine samples. The approach and the code is given and can easily be applied for the analysis of the different type of data sets acquired for different kind of samples.

CRediT authorship contribution statement

Keshav Kumar: Conceptualization, Methodology, Data curation, Visualization, Investigation, Data Analysis, Validation, Writing- Reviewing and Editing. **Anja Giehl:** Methodology, Investigation, Writing - Reviewing and Editing. **Ralf Schweiggert:** Conceptualization, Methodology, Visualization, Investigation, Writing- Reviewing and Editing. **Claus-Dieter Patz:** Conceptualization, Methodology, Visualization, Investigation, Writing - Reviewing and Editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

Dr. Keshav Kumar is thankful to the Hessen State Ministry for Higher Education, Research and the Arts, Hessen, Germany for financial support within a post-doctoral research program.

Appendix A

MATLAB Code for Eigen-directed Network analyses

```
X =input(’data=’);
% The data matrix X be such that each row must represent a sample
[I,J]= size(X);
% The data matrix X is normalized to unit area
for i=1:I
  X(i,:)=X(i,:)./(sum(X(i,:)));
end
% The data matrix X is mean centered
mn = mean(X);
X = X - repmat(mn,I,1);
% A covariance matrix Y of dimension J×J is obtained
Y=(X’*X)./(I-1);
% Eigenvector decomposition of covariance matrix Y
[E,lambda] = eig(Y);
% E contains the eigenvectors and
% lambda is a diagonal matrix containing the Eigen values
% Sorting the eigenvalues from high to low and accordingly arrange
% the eigenvectors in the matrix E
lambda = diag(lambda);
[∼, ind] = sort (-1*lambda);
lambda = lambda(ind);
E = E(:,ind);
% Networking of the samples in the Eigen-directed layout
N=X*(E); % Projecting the data X in the space spanned by the eigenvectors
% choose the two indices of the eigenvectors that must be used from the
% Eigen matrix to find the locations of the nodes. Usually first and second % indices are the most informative
fi= input(index of the first eigenvector to be used=’);
si=input(index of the second eigenvector to be used=’);
NN = N(:,[fi si]); % The matrix NN contains the node positions along the
% user specified eigenvectors
dd=pdist2(NN,NN);
for i=1:I;
  for j=1:I;
    if j<=i;
      dd(i,j)=0;
    end
  end
end
dd=reshape(dd,I*I,1);
dd=sort(dd);
dd(find(dd==0))=[];
% to obtain the full network where each node is connected to remaining set % of nodes please set the quantile to 1
qua=input(’specify the quantile value=’)
% it will be used as the threshold to find the significant interactions
% to simplify the network or in other words to have only significant
% interactions a lower threshold i.e. quantile values 0.1-0.3
% could be used
dist_mat=[];% it contains the information regarding the linkage
figure(1);
clf;
hold on;
for i = 1:I % I is the number of nodes (=number of samples)
  for j = 1:I
    dissimilarity = sqrt((NN(i,1) - NN(j,1))^2 + (NN(i,2) - NN(j,2))^2);
      if dissimilarity <= quantile(dd,qua)
        dist_mat(i, j) = 1; % there is a link;
        line([NN(i,1) NN(j,1)], [NN(i,2) NN(j,2)],’LineWidth’,1,’LineStyle’, ’:’);
      else
        dist_mat(i, j) = 0;
      end;
  end;
end;
scatter(NN(:,1),      NN(:,2),50,’MarkerEdgeColor’, [0 0 0],’MarkerFaceColor’,[0 0 0])
% modifying the sizes of the nodes based on the degree of adjacency
% information
dg=sum(dist_mat); % contains the degree of adjacency
colormap (’jet’) % other commonly used options are ’hot’, ’cool’, ’grey’, % ’winter’, ’bone’
colorbar
scatter(NN(:,1), NN(:,2), dg.*35,dg,’filled’)% color coded plot
% a factor of 35 is multiplied so that size of the nodes are appropriately % amplified, however, user can adjust the factor as per their requirement
```

References

[1] D. Cozzolino, R.D. Damberges, L. Janik, W.U. Cynkar, M. Gishen, Analysis of grapes and wine by near infrared spectroscopy, J. Near Infrared Spectrosc. 14 (2006) 279–289.

[2] I.J. Košir, J. Kidric, Use of modern nuclear magnetic resonance spectroscopy in wine analysis: determination of minor compounds, Anal. Chim. Acta 458 (2002) 77–84.

[3] H.K. Yıldırım, B. İşçi, A. Altındişli, Chemometric and phenolic attributes of blended wines assessed by multivariate techniques, J. Inst. Brew. 121 (2015) 636–641.

[4] C.G. Jares, S.G. Martín, R.C. Torrijos, Analysis of some highly volatile compounds of wine by means of purge and cold trapping injector capillary gas chromatography. Application to the differentiation of rias baixas Spanish white wines, J. Agric. Food Chem. 43 (1995) 764–768.

[5] K. Kumar, A. Giehl, C.-D. Patz, Chemometric assisted Fourier Transform Infrared (FTIR) Spectroscopic analysis of fruit wine samples: Optimizing the initialization and convergence criteria in the non-negative factor analysis algorithm for developing a robust classification model, Spectrochim Acta Part A 209 (2019) 22–31.

[6] K. Kumar, A. Giehl, R. Schweiggert, C.-D. Patz, Multidimensional scaling assisted Fourier-transform infrared spectroscopic analysis of fruit wine samples: introducing a novel analytical approach, Anal. Methods 11 (2019) 4106–4115.

[7] A.P. Umali, S.E. LeBoeuf, R.W. Newberry, S. Kum, L. Tran, W.A. Rome, T. Tian, D. Tiang, J. Hong, M. Kwan, H. Heymann, E.V. Anslyn, Discrimination of flavonoids and red wine varietals by array of differential sensors, Chem. Sci. 2 (2011) 439–445.

[8] H.V.D. Voet, D.A. Doornbos, The use of pattern recognition techniques in chemical differentiation between Bordeaux and Bourgogne wines, Anal. Chim. Acta 159 (1984) 159–171.

[9] D. Picque, T. Cattenoz, G. Corrieu, Classification of red wines analysed by middle infrared spectroscopy of dry extract according to their geographical origin, J. Int. Sci. Vigne Vin 35 (2001) 165–170.

[10] M. Palma, C.G. Barroso, Application of FT-IR spectroscopy to the characterization and classification of wines, brandies and other distilled drinks, Talanta 58 (2002) 265–271.

[11] D. Cozzolino, H.E. Smyth, M. Gishen, Feasibility study on the use of visible and near-infrared spectroscopy together with chemometrics to discriminate between commercial white wines of different varietal origins, J. Agric. Food Chem. 51 (2003) 7703–7708.

[12] D.M.A.M. Luykx, S.M. van Ruth, An overview of analytical methods for determining the geographical origin of food products, Food Chem. 107 (2008) 897–911.
[13] B. Smith, Infrared Spectral Interpretation: A Systematic Approach, CRC Press, Boca Raton, Florida, 1999.
[14] C.N. Banwell, Fundamentals of Molecular Spectroscopy, 3rd edn., McGraw-Hill, London, 1983.
[15] C.D. Patz, A. David, K. Thente, P. Kürbel, H. Dietrich, Wine analysis with FTIR-spectrometry, Vitic. Enol. Sci. 54 (1999) 80–87.
[16] C.D. Patz, A. Blieke, R. Ristow, H. Dietrich, Application of FT-MIR spectrometry in wine analysis, Anal. Chim. Acta 513 (2004) 81–89.
[17] H.H. Nieuwoudt, I.S. Pretorius, F.F. Bauer, D.G. Nel, B.A. Prior, Rapid screening of the fermentation profiles of wines yeasts by Fourier transform infrared spectroscopy, J. Microbiol. Methods 67 (2006) 248–256.
[18] M. Friedel, C.D. Patz, H. Dietrich, Comparison of different measurement techniques and variable selection methods for FT-MIR in wine analysis, Food Chem. 141 (2013) 4200–4207.
[19] H.H. Nieuwoudt, R. Bauer, J. Kossmann, FTIR spectroscopy for grape and wine analysis, Anal. Chem. 80 (2008) 1371–1379.
[20] D.W. Lachenmeier, Rapid quality control of spirit drinks and beer using multivariate data analysis of Fourier transform infrared spectra, Food Chem. 101 (2007) 825–832.
[21] R. Kramer, Chemometric techniques for quantitative analysis, Marcel Dekker, New York, 1998.
[22] S. Wold, K. Esbensen, P. Geladi, Principal component analysis, Chemom. Intell. Lab. Syst. 2 (1987) 37–52.
[23] R. Bro, A.K. Smilde, Principal component analysis, Anal. Methods 6 (2014) 2812–2831.
[24] G. Young, A.S. Householder, Discussion of a set of points in terms of their mutual distances, Psychometrika 3 (1938) 19–22.
[25] J.C. Gower, Some distance properties of latent root and vector methods used in multivariate analysis, Biometrika 53 (1996) 325–338.
[26] L. Kaufman, P.J. Rousseeuw, Finding groups in data: An introduction to cluster analysis, John Wiley & Sons, New Jersey, 2005.
[27] G.R. Brereton, Chemometrics for pattern recognition, John Wiley & Sons, Chichester, 2009.
[28] T. Kohonen, Self-Organizing Maps, Springer, Berlin, 2000.
[29] T. Kohonen, Self-organized formation of topologically correct feature maps, Biol. Cybern. 43 (1982) 59–69.
[30] G.R. Lloyd, R.G. Brereton, J.C. Duncan, Self Organising Maps for distinguishing polymer groups using thermal response curves obtained by dynamic mechanical analysis, Analyst 133 (2008) 1046–1059.
[31] J. Dalege, D. Borsboom, F. van Harreveld, H.L.J. van der Mass, Network Analysis on attitudes: A brief tutorial, Soc. Psychol. Pers. Sci. 8 (2015) 528–537.
[32] D. Borsboom, A.O.J. Cramer, Network Analysis: An Integrative approach to the structure of psychopathology, Annu. Rev. Clin. Psychol. 9 (2013) 91–121.
[33] S. Brohee, K. Faust, G. Lima-Mendez, G. Vanderstocken, J. van Helden, Network analysis tools: from biological networks to clusters and pathways, Nat Protoc. 3 (2008) 1616–1629.
[34] S. Vishveshwara, K.V. Brinda, N. Kannan, Protein structure: insights from graph theory, J. Theor. Comput. Chem. 1 (2002) 187–212.
[35] W.D. Wallis, A beginner's guide to graph theory, Birkhauser Boston, Berlin, 2007.
[36] R. Diestel, Graph theory, Springer-Verlag, New york, 2000.
[37] N. Deo, Graph theory with applications to engineering and computer science, Prentice Hall of India Private Limited, New Delhi, 1984.
[38] O. Ivanciuk, A.T. Balaban, Encyclopaedia of computational chemistry, Graph theory in chemistry, John Wiley & Sons, New York, 1998.
[39] R.J. Wilson, Introduction to graph theory, Addison Wesley Longman Limited, Essex, England, 1998.
[40] G.D. Battista, P. Eades, R. Tamassia, I.G. Tollis, Algorithms for drawing graphs: an annotated bibliography, Comput. Geom. 4 (1994) 235–282.
[41] R. Davidson, D. Harel, Drawing graphs nicely using simulated annealing, ACM Trans. Graph 15 (1996) 301–331.
[42] J. Diaz, J. Petit, M. Serna, A survey of graph layout problems, ACM Comput Surv. 34 (2002) 313–356.
[43] C. Gotsman, Y. Koren, Distributed graph layout for sensor networks, Lecture notes in Computer Science, Springer, Berlin, Heidelberg, 2005.
[44] K. Misue, P. Eades, W. Lai, K. Suiyama, Layout adjustment and the mental map, J. Vis. Lang. Comput. 6 (1995) 183–210.
[45] E. Gansner, E. Koutsoftos, S. North, K.P. Vo, A technique for drawing directed graphs, IEEE Trans. Software Eng. 19 (1993) 214–230.
[46] W. Barth, M. Juenger, P. Mutzel, Simple and efficient bilayer cross counting, J. Graph Algorithm Appl. 8 (2004) 179–194.
[47] Y. Koren, Drawing Graphs by Eigenvectors: Theory and Practice, Comput. Appl. Math. 49 (2005) 1867–1888.
[48] T.M.J. Fruchterman, E.M. Reingold, Graph drawing by force-directed placement, Software Pract. Ex. 21 (1991) 1129–1164.
[49] T. Kamada, S. Kawai, An algorithm for drawing general undirected graphs, Inf. Process. Lett. 31 (1989) 7–15.
[50] R. Banc, F. Loghin, D. Miere, F. Fetea, C. Socaciu, Roamanian wines quality and authenticity using FT-MIR spectroscopy coupled with multivariate data analysis, Not Bot Horti Agrobo 42 (2014) 556–564.