Send Orders of Reprints at reprints@benthamscience.net

8 *The Open Food Science Journal,* **2013,** *7,* **8-22**

Open Access

Combined Cluster Analysis and Principal Component Analysis to Reduce Data Complexity for Exhaust Air Purification

Bastian Ebeling [a,*], Cristiam Vargas [b] and Simone Hubo [c,*]

[a]*Blohm + Voss Naval GmbH, Department AME, Hermann-Blohm-Strasse 3, D-20457 Hamburg, Germany*

[b]*Novartis Pharma Productions GmbH, Oeflinger Strasse 44, D-79664 Wehr, Germany*

[c]*Helmut-Schmidt-University/University of the Federal Armed Forces Hamburg, Institute of Thermodynamics, Holstenhofweg 85, D-22043 Hamburg, Germany*

Abstract: Anthropogenic and demographic processes cause worldwide air problems, giving rise to focus on exhaust air purification to counteract these effects. Due to the large number of substances found in exhaust air and the various operational parameters needed, a huge amount of often high dimensional data has to be analyzed. The ultimate goal is to finally reduce data complexity in terms of information reflecting the substances´ characteristics.

The Cluster Analysis (CA) of data from 30 exhaust air compounds with 11 indices representing both structural characteristics and physicochemical data resulted in 7 clusters. The Principal Component Analysis (PCA) led to the identification of 6 Principal Components (PCs) and therefore to a dimensional reduction compared to the originally used 11 indices. After re-gathering the total information of the original data-set upon the 6 PCs only, a re-clustering showed that we were able to restore the same cluster structure as in the original CA based on the 11 indices. This process is a first proof of principle in successful re-clustering after dimensional data reduction by our proposed combined CA-PCA method and hence a step towards a possible development of an adsorption method to selectively remove malodorous/toxic components from the exhaust air.

Keywords: Cluster Analysis, Dimensional Reduction, Exhaust Air Purification, Graph Theory, Odour Control, Principal Component Analysis.

1. INTRODUCTION

Air-quality has become a serious problem in many cities world-wide caused by the emission of air pollutants which are generated by different anthropogenic processes [1]. The tolerance against malodors in exhaust air streams especially in an urban environment is low, giving rise to the urgent need for the further development of industrial exhaust air purification processes [2]. In practice, control-ling air pollution is a very complex problem. In different industrial applications complex multicomponent mixtures have to be purified in industrial air purification systems. Several industrial applications have already been proposed and discussed, as e. g. the electrostatic precipitator (ESP) for cleaning gas streams which is mainly used in industrial dust collection and home ventilation systems [3]. The regeneration of activated carbon after adsorption of waste gas streams and waste water components by dielectric barrier discharge (DBD) oxidation opened up another way of industrial exhaust air purification [4]. Furthermore, there are technological approaches dealing with the specificity of the adsorbent to answer the question which components only need to be removed from the complex multicomponent mixture to set both malodor and toxicity of the exhaust air to a level of zero. Once a selective adsorbent is found, the possibility is to keep concentrations below the odor threshold by combining the selective adsorption process with on-site regeneration of the adsorbent by microwave, ultrasound, ultrasound with water, water desorption [5] or by the concept of a rotary adsorber [6]. Applying thermally stable adsorbents instead of highly selective ones can be used for the removal of volatile organic compounds in a hybrid adsorption/inci-neration multifunctional adsorber/reactor concept along with high energy efficiency [7].

All these recent developments in analytical methodologies contributed to improvements of odorous emissions in terms of understanding multicomponent composition and concentration. However, methods and studies applied beforehand the purification process, leading to classification and identification of the complex mixtures´ compounds, produce a massive quantity of high dimensional data [8]. To facilitate the process of selection, design and management within the industrial exhaust air purification, a precise analysis of these often high data amounts is an essential task to be performed in the future [9].

To analyze such data sets, the goal is to finally reduce data complexity in terms of information reflecting the

*Address correspondence to these authors at the Blohm + Voss Naval GmbH, Department AME, Hermann-Blohm-Strasse 3, D-20457 Hamburg, Germany; E-mail: bastian.ebeling@ingenieur.de; and Helmut-Schmidt-University/University of the Federal Armed Forces Hamburg, Institute of Thermodynamics, Holstenhofweg 85, D-22043 Hamburg, Germany;
E-mail: simone.hubo@hsu-hh.de

1874-2564/13

characteristics of the substances [10], as shown here by applying Cluster Analysis combined with a two-step Principal Component Analysis. Therefore, first a significant grouping of the compounds according to the characteristics of each substance needs to be performed in such a way that at the end they represent the original set as well as its diversity. A statistic tool helping to organize the observed data into meaningful subsets is Cluster analysis whereby a file of objects may be divided into several classes based on structural parameters as well as on physicochemical values [11]. The members of each class should be similar among each other and different from the members of the other classes [12]. The CA is used in many different fields of application and its approaches are numerous and diverse such as for example chemical research, drug discovery and many others, making the analysis of complex, high dimensional data sets easier [13,14]. Here, in our application, this statistic tool is followed by PCA which identifies Principal Components (PCs) representing as much substance information in as few Principal Components as possible. Hence, the PCA aims at reducing data complexity by dimensional reduction ideally without loss of any information [15].

The idea of this study is to establish our combined CA-PCA method to reduce data complexity already beforehand the development of an industrial exhaust air purification process.

2. THEORETICAL BACKGROUND

2.1. Descriptors in Graph Theory

Different structural indices are calculated translating the molecular properties into numerical values. This is achieved by representing the structural formula of the molecule by a chemical graph, replacing the atoms by numbered vertices and the molecular bonds by edges which associate two vertices with each edge [16-18]. A distance matrix is created upon the number of lines describing the shortest path between two points i and j in a simple graph. By definition, the distance of adjacent points equals one whereas the distance of non-adjacent points equals zero [19]. The generation of such a distance matrix completely defines the adjacency matrix that represents the chemical structure of substances used in the present study (Fig. **1**).

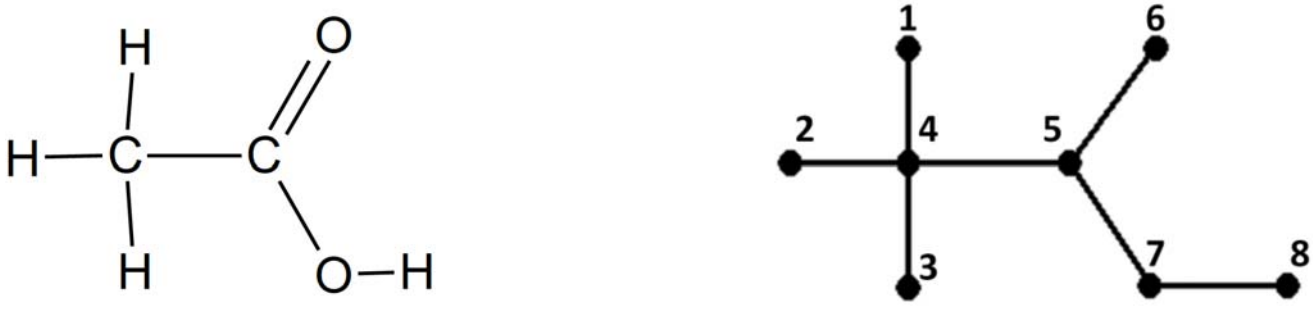
This method has already been used in different studies where chemical properties are predicted according to quantitative structure-activity relationships (QSAR) as well as quantitative structure-property relationship (QSPR) models [18,20].

2.2. Cluster Analysis

One way to quantify the degree of similarity or dissimilarity is given by the question of how close or apart two substances (objects) are from each other. This distance or proximity between two objects can be calculated by mathematical methods which can be visualized in a dendrogram according to the characteristics of each compound. Different types of clustering can be performed. Mainly, the hierarchical clustering (agglomerative) [21,22] and the nonhierarchical clustering or partitional clustering (k-means) is applied [21,23].

The hierarchical clustering method sets up the clusters using an iterative algorithm in a bottom up process. A matrix of similarity (or dissimilarity) measures between each pair of compounds is used first by merging individual compounds into clusters followed by merging clusters into super clusters. The final merge brings all the compounds into a single cluster. Considering the idea of ``distance" described above, a simplified way to see how clusters are generated with the hierarchical clustering method is first to consider each object as its own cluster. Two clusters, at this level still individual objects that have a small ``distance" between them, are going to be grouped into one cluster hosting two individual objects. Then the similarity matrix is recalculated considering the distances between new clusters (individual objects) and the cluster hosting already two individual objects in order to obtain the smallest distance to a new cluster and so on. This process is repeated until all objects are joined in one ``Megacluster".

The k-means clustering works on actual observations rather than on larger sets of dissimilarity measurements and therefore creates a single level of clusters in contrast to the agglomerative, hierarchical clustering method. In the k-means clustering process, k must be defined a priori to distribute all data points into the k clusters by choosing cluster centers. In the process presented here, k was chosen beforehand upon the number of clusters yielded by the hierarchical clustering method in an unbiased approach. Setting up the k-means clustering on this basis, an iterative algorithm over a given starting partition is used to minimize the sum of distances over all clusters from each object to its cluster centroid. Objects are sorted between clusters until the



(A) (B)

Fig. (1). (**A**) Chemical structure of acetic acid. (**B**) Chemical Graph structure of acetic acid.

sum cannot be decreased any further by the algorithm. A set of clusters that are as compact and well-separated as possible is given as the result. Applying several optional input parameters to k-means, the process can be controlled in detail, for example *via* the initial values of the cluster centroids coming from the hierarchical clustering as well as *via* the maximum number of iterations [22,24,25].

For a successful process of Cluster Analysis, three decisions must be taken [26]:

1. The concerns of a possible transformation of the variables
2. The selection of the algorithm (Clustering method)
3. The selection of the similarity measurement

2.3. Principal Component Analysis (PCA)

Principal component analysis is a multivariate statistical technique used in exploratory data analysis. Moreover, predictive models can be generated in which a data table representing observations described by several dependent, generally inter-correlated variables is elucidated. The purpose of PCA is to extract the most important information from the data set, the compression of its size as well as the generation of a simplified view of the original data set by dimensional reduction, ending up in an analysis of the structure of the observation and variables. In order to achieve this goal a set of new *orthogonal, non-correlated* variables (*q*) called Principal Components is calculated as linear combinations of the original variables (*p)* with $q < p$ [15].

The first Principal Component is required to have the largest possible variance and therefore this component will represent best the diversity of the given data. The second Principal Component is generated under the assumption of being orthogonal to the first component combined with also having the largest possible variance. The other components are calculated likewise. These components are exactly the eigenvectors of the correlation matrix which are employed as a basis for a new coordinate system [12,15,27].

Applying a basis transformation of the given data will result in new coordinates for each substance. Ideally, the first k calculated coordinates can be used to approximately describe the substances´ characteristics. Thus, for dimensional reduction it is possible to use only a subset of these coordinates. This does not necessarily result in a reduced number of indices needed to describe the substances´ characteristics, as generally all indices might be involved in those selected coordinates.

A PCA may be carried out on data from either a covariance or a correlation matrix differing basically in normalization. To use a covariance matrix the variables should have the same order of magnitude and no significant difference in the variances. If the data do not fulfill this condition the use of a covariance matrix can for the unfulfilled variables produce a greater variance and assignment of larger weights resulting in certain errors in the results. To avoid this it is preferred to perform the PCA on standardized data -- thus using the correlation matrix [28].

3. MATERIALS AND METHODOLOGY

3.1. Choice of substances

30 substances were chosen from a list of 155 substances based on the experimental and theoretical data collected for each substance. The 155 substances represent an average exhaust air stream of livestock facilities, fat refineries, and cocoa and coffee production plants [29]. The chosen 30 compounds were examined as regards to their structure, solubility and vapor pressure.

3.2. Cluster Analysis

Structural indices (calculated from the graph theory) as well as physicochemical values for the present work are described in Table **1**.

The first eight indices of Table 1 describe topological characteristics of the substances and belong to the tools applied as descriptors in Graph Theory (Fig. **1**) [29]. In contrast, the last three indices represent the substances´ chemical properties. A more detailed view of how the indices are employed in the present study in Graph Theory is given as follows:

The Zagreb Index measures the connectivity of a molecular graph (Fig. **1B**) by first counting the number of edges adjacent to each vertex calculating the vertex´s degree and afterwards summing up the square values of each vertex´s degree. Consequently, the Zagreb Index does not weight the vertices evenly as it leads to a higher weight for the more inner vertices. The Wiener Index gives a measure of the compactness for molecular structures as it leads to higher numbers for long chains and smaller numbers for branched structures. Although the Randic Index once more emphasizes the connectivity of a molecule like the Zagreb Index, the calculation always involves a pair of adjacent vertices. Hence, it is referred to as a second-generation index. Its numbers are calculated by the sum over all edges combined with the inverse square root of the vertices´ degrees being adjacent to the edge. In contrast to the Zagreb Index, the Randic Index therefore does not overvalue the inner vertices but the outer vertices due to the inverse of the square root. The Balaban Index resembles both the Wiener Index and the Randic Index but introduces a normalization factor ``A" with B being the number of edges and X representing the cyclomatic number of the graph, thus discriminating easily between cyclic and non-cyclic molecules. The Information Index on Atomic Composition discriminates among different elements within one molecule. According to the information theory the information content is dependent on the intramolecular variety, leading to the total number of atoms being expressed as ``N" in contrast to all atoms of the same type ``i" being expressed as ``N_i". In an analogous manner the Information Index on Molecular Bonds is calculated simply replacing the atoms by molecular bonds and their bond characteristics. The electron configuration of a molecule is considered by splitting up the electron distribution around the atomic nucleus into different partial bond spaces, while the carbon skeleton of the whole molecule forms a finite bond space. Basically, a distinction

Table 1. Structural Indices and Physicochemical Values Used in the Cluster Analysis. The Original Data Matrix Including All Single Index Values Listed Under their Abbreviations can be Found in the Supporting Information in Table S1

Name (Abbreviation S1)	Description
Zagreb Index (ZI)	$M(G) = \sum_{i=1}^{N} g_i^2$
Wiener Index (WI)	$W(G) = \frac{1}{2}\sum_{ij}^{N} d_{ij}$
Randic Index (RI)	$X_R(G) = \sum_{Kanten} \frac{1}{\sqrt{g_i * g_j}}$
Balaban Index (BI)	$J(G) = A \cdot \sum_{adjacent(i,j)} \frac{1}{\sqrt{d_i \cdot d_j}}$ With the normalization factor A $A = \frac{B}{X+1}$ B= total of edges X= total of rings
Information Index on Atomic Composition (IIAC)	$I_{aZges} = N \cdot I_{aZ} = N \cdot \log_2 N - \sum_{i=1}^{k} (N_i \cdot \log_2 N_i)$
Information Index on Molecular Bonds (IIMB)	$I_{Bges} = B \cdot \log_2 B - \sum_{i=1}^{m} (B_i \cdot \log_2 B_i)$
Information Index on Electron Configuration (IIEC)	$I_{\varepsilon ges} = \log_2 P = \log_2 (N!/\prod_{i=1}^{k} N_i!)$
Topological Information Index (TII)	$I_0 = -\sum_{i=1}^{k} (\frac{C_i}{C} \cdot \log_2 \frac{C_i}{C})$
Normalized Molar Mass (MM)	$M = \frac{m}{n}$ m=Molar mass of molecule n= Number of atoms in molecule
Water Solubility (WS)	Maximum amount of a substance that can dissolve in water at equilibrium conditions.
Vapor Pressure (VP)	$\ln(p) = A - \frac{B}{(C+T)}$ A,B,C = Antoine constants T= Temperature (K)

is made between the valence bond space versus the non-valence bond space in closer proximity to the atomic nulei. The aim of the Topological Information Index is to generate equivalent classes of graph elements, i. e. vertices which can be substituted among each other without destroying or losing any of the graph properties. That means that a permutation of vertices must exist that maps the graph exactly to itself. If this requirement is fulfilled, vertices are grouped into one equivalent class subsequently subjected to the information -- theoretical equation expressing the Topological Information Index [17,29,30].

The data of the present study were analyzed by ClustanGraphics8§ released 2005 by Clustan -- A Class Act © 1998 Clustan Ltd., UK. If variables are measured in different scales, variables with large values would contribute more to the distance measure than variables with small values. Thus, each substance´s values for all eleven original variables (Table **S1**) were first put under the z-core standardization (decision no. 1) to avoid differences between magnitudes of data being a failure source when calculating the distance. As decision no. 2 the clustering method k-means (k = 7, selected due to the number of clusters yielded by the hierarchical clustering in an unbiased approach beforehand) was chosen with the Euclidean sum of squares as a clustering criterion and started out with a random partition of data (initial centers). After the clusters are grouped the proximities between the objects are calculated.

For continuous data, the Square Euclidian distance equation (1) is the one mostly used due to the fact that in chemical studies most of the data sets are continuous (decision no. 3) as in our case.

$$d^2(x,y) = \frac{(x_1 - y_1)^2 + \ldots + (x_n - y_n)^2}{v} = \frac{\sum_{i=1}^{n}(x_i - y_i)^2}{v} \quad (1)$$

As a result of this proximity calculation, a graphical output of the clustering process as a dendrogram is generated showing the arrangement of the clusters as a tree-like structure [31]. A dendrogram containing the clustering results of the study is presented in Fig. (**2**).

3.3. Principal Component Analysis (PCA)

Data of the structural as well as physicochemical characteristics of the given volatile substances analyzed in the present study were collected and organized in a matrix. The rows of this matrix correspond to the substances while the indices are organized in columns. So each row represents one substance with its own characteristic index values. The original data matrix can be found in Table **S1** of the supporting information. Subsequently, this matrix was imported into MATLAB®.

For each cluster a theoretically ideal substance (called main focus) for representation was selected *via* the arithmetic mean value over the substances. A real substance for follow-up experiments to represent each cluster should ideally combine the characteristics of the theoretically ideal substance or show at least the lowest deviance. The Principal Component Analysis was carried out twice in total. The first step was performed over the main foci as we chose the variance between the main foci being of greater importance than the variance among the substances to distinguish between the clusters first. Within the second PCA step it turned out to be sensible to repeat the whole process on the remaining data basis taking into account the complete variance of the data and not only the main foci. We established this process on the basis of a descriptive minimal mathematical example that can be found in the supporting information (chapter **S 3.3.1**) accompanying this study.

After a successful establishment of the minimal mathematical example, we applied the process to our experimental data set. Focusing on our 30 volatile substances, the corresponding data is set up in Matlab as a matrix

$$T \in \mathbb{R}^{c \times i} \quad (2)$$

containing the theoretically ideal substance for each cluster in each row. Thus, the number of rows equals the number of clusters c. The indices are organized in a number of i columns.

Next, data were shifted and scaled [32] in order to achieve a variance of 1 about zero mean among the indices. All the information by the indices have now an equal weight and can contribute equally to the following data processing. We defined the matrix C as the covariance matrix of z-scores, being symmetric positive semidefinite and resulting in

$$C \in \mathbb{R}^{i \times i} \quad (3)$$

The Principal Components of C are calculated according to equation 4. For all j with $j \in [1, i]$ applies:

$$C \cdot v_j = v_j \cdot \alpha_j \quad (4)$$

where v_j is the eigenvector to the eigenvalue α_j. The

Fig. (2). Dendrogram showing the 7 different clusters obtained from the Cluster Analysis. (* = 4-Isopropenyl-1-methyl).

eigenvectors are the Principal Components and each corresponding eigenvalue respectively represents their weight in terms of original information they contain. Without loss of generality we assume them sorted descending. To have them in the range of percentage, the eigenvalues were scaled as defined in equation 5:

$$\left[\overset{\approx}{\alpha}_j\right] \supset: \tilde{a} := \frac{a}{\|a\|_1} \quad \text{with } a := \lfloor \alpha_j \rfloor \tag{5}$$

As we regard covariance matrices having positive diagonal entries by nature the trace, which is the sum of diagonal entries being invariant under similarity transformations, equals the so called 1-norm (Manhattan Norm) of the vector a of eigenvalues. Hereby we took the decision to use this norm for normalization.

With these values α_j the significance of the new directions given by the eigenvectors can be distinguished. This method will yield at most

$$\min(c-1, i) \tag{6}$$

not vanishing eigenvalues and corresponding vectors [27].

As this method might not deliver enough Principal Components for successful reclustering, a cutoff γ on the eigenvalue (by nature $= 0$) is introduced to control which u eigenpairs of the first step are accepted ($= \gamma$) or rejected [33].

For the remaining subspace given by

$$V_r = \lfloor v_j \rfloor, \quad j > u \tag{7}$$

again the eigenfactorization is calculated but this time from the covariance matrix of all single substances scaled and shifted by means of the z-score from the main foci.

Using the four values 0.25, 0.50, 0.75 and 1 for the cutoff γ results in different numbers of PCs in each case. A value of $\gamma = 0.25$ results in seven PCs for successful reclustering while in all other cases only six PCs were necessary. Hence, selecting the cutoff $\gamma = 0.5$ (dotted line in Fig. **3**) yielded already the right result by the lowest possible γ-value at the same time: Using five PCs from the first step (calculated over the main foci) and one additional PC from the second step (calculated over all single substances) instead of the original eleven indices (Fig. **3**) lead to the same clustering result (Fig. **5**) as compared to the original result from Cluster Analysis (Fig. **2**). Thus, the benefit of our newly established two-step PCA process could clearly be underlined empirically by introducing this value for the cutoff γ. With this ordered set of new directions we introduce a new (orthogonal) basis. The variances (scaled to sum up to one) of all (scaled and shifted) substances in terms of this basis are plotted in Fig (**3**) with decreasing weight.

4. RESULTS AND DISCUSSION

The results of the Cluster Analysis are described and discussed followed by the results´ analysis and discussion of

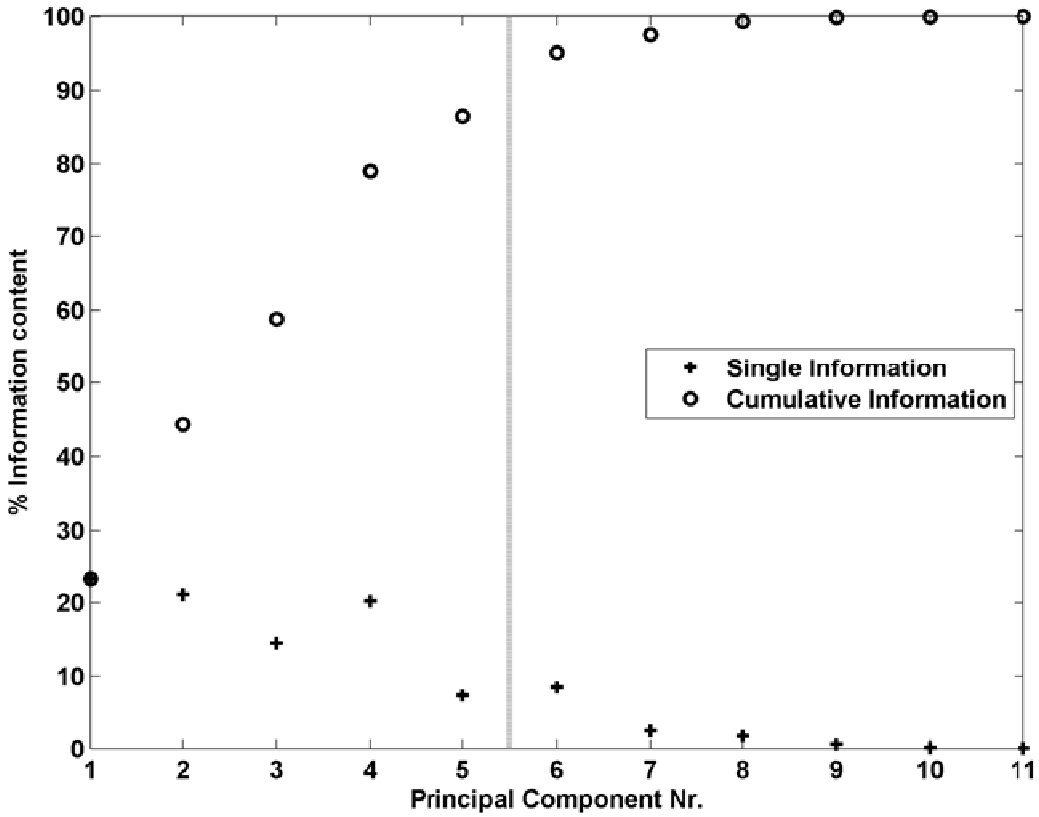


Fig. (3). Information content of the eleven Principal Components. Cross markers reflect the single information each Principal Component contains, circles represent the cumulative percentage of information from the first PC up to the current one. The dotted line marks the selected cutoff γ (cf. 3.3) within the two-step PCA process.

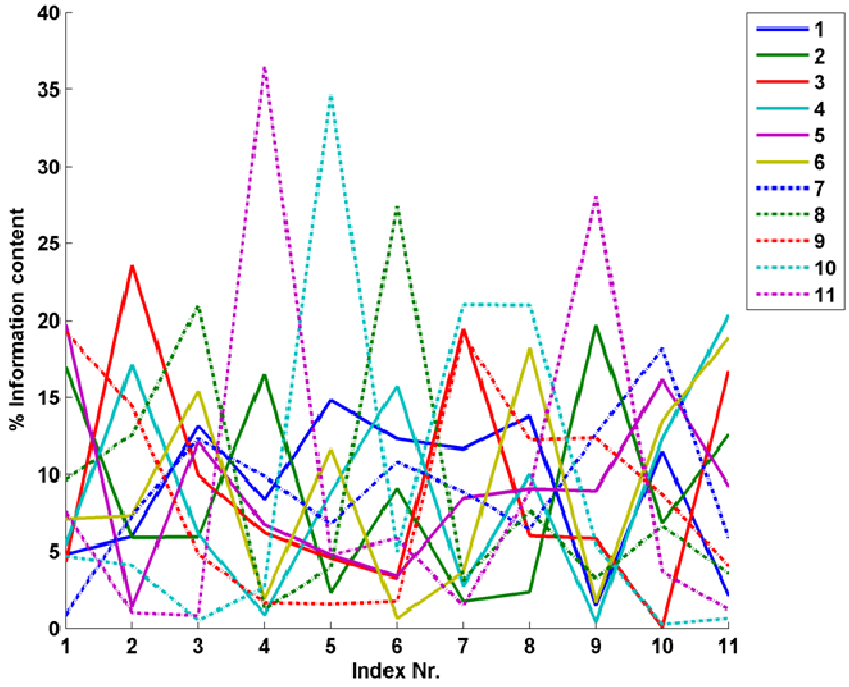


Fig. (4). Visualization of the Principle Components (PCs). Each colored line represents one PC and shows its decomposition into the originally used index quantities. Continuous lines (PC #1-6) carry in total already about 95 % of the original information, dotted lines (PC #7-11) do nearly carry no information as can be obtained from Figure 4.

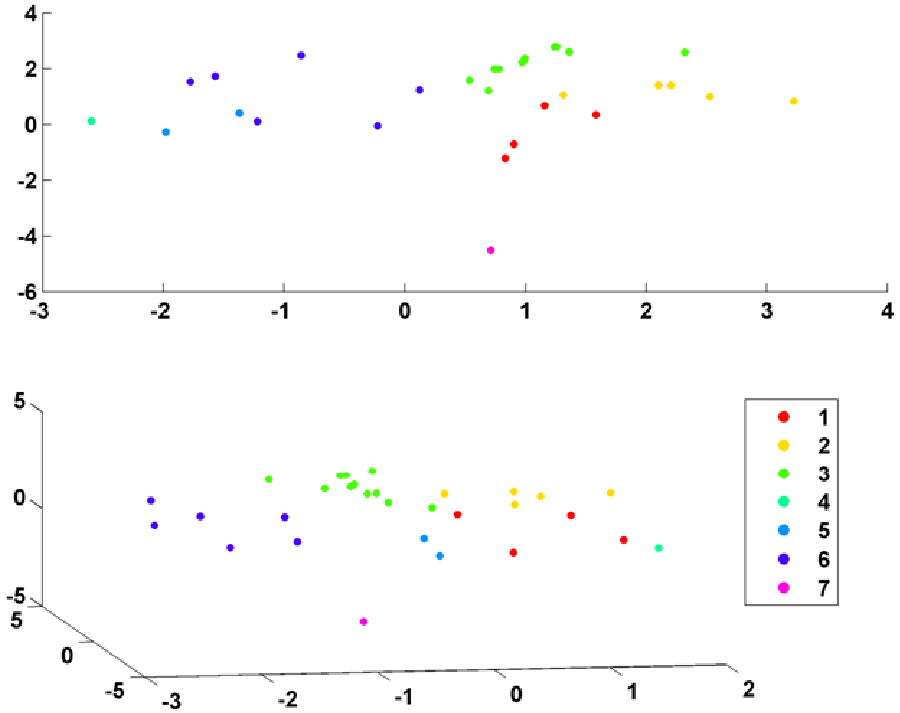


Fig. (5). Scatterplot of the Re-Clustering after the two-step PCA process. **a**): Two dimensional scatterplot made up of the first two PCs reflecting about 71 % of the original data information. **b**): Three dimensional scatterplot made up of the first three PCs covering about 85 % of the original information. The cluster numbers 1-7 represent the original clusters in Figure 2.

the Principal Component Analysis. We describe how the clusters were calculated after transformation of the data, the performance of the Principal Component Analysis, the recollection of information for each substance after dimensional reduction as well as the reproduction of the original cluster structure on the reduced data set.

4.1. Cluster Analysis

Clustering of the 30 substances resulted in 7 clusters graphically displayed in form of a dendrogram (Fig. **2**). The corresponding structure of each molecule is shown in Table **S2** of the supplement. The dendrogram offers a view of the

``distances" between molecules, reflecting how similar or dissimilar they are. If two molecules are bound by the same line (level) they are more similar or chemically closer to each other than other molecules in the same cluster, not connected directly. This applies also for distances between two clusters.

Since some of the variables applied to the clustering process explain structural parameters, by examining the substances of each cluster some characteristics can be observed:

Cluster 1 presents four members, two of them (pyrrolidine and piperidine) are cyclic molecules with a density of 0.86 g/mL each and provide a characteristic amine odor, and two group members are of linear structure with a density of 0.79 g/mL each and a pleasant odor. The only difference among the cyclic ones is one additional carbon substitution at the ring; hence they are together at the same level of the dendrogram but different to acetone and methanol (Fig. **2**). All four molecules are water-miscible, colorless and highly flammable.

Substances with aliphatic intramolecular bonding architecture are grouped in cluster 2. The first three substa-nces are water soluble and have only saturated hydrocarbons in their molecular structure. The last two substances, triethylamine and 1,3-pentadien show amine groups, and double bonds. They are insoluble in water. Most of cluster 2 substances have an unpleasant odor.

Cluster 3 hosts molecules with a carbonyl group, one cyclic molecule without (4-Isopropenyl-1-methylcyclohexane), one molecule with an alcohol group (1-octen-3-ol) and one with two terminal amine groups (cadaverine). It is interesting to note that almost all substances in this cluster are colorless to pale yellow liquids having very pleasant, in most cases fruit-like, odors. Besides, they all show quite similar densities (0.82-0.84 g/mL) except for n-decane. Only n-decane has a mild gasoline-like odor and a density of 0.73 g/mL, cadaverine has a decomposition characteristic odor. 1-Octen-3-ol is the only molecule without a fruit-type odor; it has rather an earthy, mushroom-like characteristic odor.

It is worth noting that dimethyldisulfide forms a single cluster (cluster 4) being the only molecule containing sulfur atoms in the structure and an alliaceous onion odor.

In cluster 5, only two substances differ as regards to their functional group from the other substances as they are the only carboxylic acids in the list, both with pungent vinegar odor, soluble in water, ethanol, diethylether and other organic solvents. Their aliphatic chains vary in length in only one carbon atom.

All molecules with aromatic ring systems are in cluster 6. Toluene and benzene share the same level in the dendrogram due to their similar structure. Both are insoluble in water, have a water-clear appearance, are highly flammable and give off a characteristic sweet odor.

Water as well stays alone, building up cluster 7, showing its particular structure and unique behavior compared with the other substances.

A chemical analysis of the substances within each cluster was conducted and physicochemical as well as odorous characteristics were found that validate the cluster analysis in the first step. According to Linusson *et al.*, this kind of analysis can be taken as a validation procedure for clustering of chemical data regarding the chemical interpretation of the substances in each cluster [34]. Although not all molecules fit perfectly into the clusters, analyzing them in this case as described above, all clusters obtained do make sense: all molecules with fruit odor were grouped into one cluster (cluster 3) except for cadaverin but only concerning its odor. Taken together, similarities in structural and topological parameters led to correct clustering of substances according to their similarities in physicochemical parameters, as regards density or appearance.

4.2. Principal Component Analysis

The Principal Component Analysis was carried out to check the possibility of dimensional reduction that could be successfully verified in a repeated Cluster Analysis of the data set re-generated after PCA and dimensional reduction. Due to this fact, an experimental verification is currently being performed on all substances presented in this paper.

In total, eleven Principal Components were calculated based on originally eleven indices used in the CA in the first part of this study. At the beginning, each of the indices carried one eleventh ($\approx$ 9 %) of the total information. Some of the newly calculated Principal Components carry significantly more than one eleventh of the original total information as shown in Fig. (**3**).

Fig. (**3**) depicts the information content of all eleven Principal Components calculated. They are sorted descending according to their percentage of single information shown by the cross markers. The cumulative percentage of information (circle markers) is calculated as the sum of percentages of information using all PCs from the first to the current one. Each of the first four PCs contains more than one eleventh of the total original information whereby one eleventh equals a cutoff value $\gamma = 1$ in our two-step PCA process (cf. 3.3). PC number 5, 6 and 7 still carry information contributing to a significant increase in the cumulative percentage of information. PCs No. 8-11 contain neglectable information. It can be deduced from Fig. (**3**) that the first six Principal Components already reflect a total of about 95 % of the original information. The PCs No. 8-11 do not add any major information as the percentage for each of these components is near 0 and their cumulative percentage of information stays close to 100 %. The dotted line between PCs No. 5 and 6 represents the selected cutoff $\gamma = 0{,}5$ yielding best results by the lowest possible γ-value at the same time in our two-step PCA process (cf. 3.3): Whereas the first five PCs were calculated over the variance of the main cluster foci (theoretically ideal substances), PCs No. 6-11 were calculated over the variance among all 30 substances. Graphically spoken, the eleven Principal Components define new orthogonal axes of the eleven dimensional vector space over the original indices representing one dimension each [35].

The proportion of each original index represented by each PC can be depicted in Fig. (**4**). This visualization shows one PC by each colored line. The original indices are marked on the abscissa whereas the ordinate shows the proportion of

each index, e. g. the first PC (displayed in a continuous blue line) combines about 5 % of index 1, 6 % of index 2 and so on, comprising a total of already 43 % of the original information. Those PCs contributing significantly to the total of about 95 % of the original information are drawn in continuous lines (PC #1-6), the dotted lines (PC #7-11) do nearly carry no more information and therefore do not contribute significantly to the cumulative percentage of information as shown in Fig (**4**).

Each of the eleven Principal Components displays quite an even mixture of the original index quantities. There is no PC that represents mainly one index but each PC comprises a significant part of each original one. This finding may lead to difficulties in dimensional reduction, as the Principal Components are calculated under the condition to maximize the original data information each component covers and thus, this information cannot be deduced to derive simply from mainly a single index.

However, it was possible to reduce dimensions and therefore complexity of the original data set. This was tested and verified by applying the k-means clustering process once again to the substance data set fully regenerated after the combined CA-PCA process, reproducing the same clustering result as obtained from the original Cluster Analysis (Fig. **2**). For visualization, the two dimensional scatterplot made up of the first two PCs already shows a clear differentiation between the 7 clusters (Fig. **5a**), reflecting about 71 % of the original data information. The three dimensional scatterplot covers about 85 % of the original information and is built upon the first three PCs (Fig. **5b**). In both projections a clear grouping of the substances belonging to the 7 different clusters marked in different colors can be noted. Therefore, this process could be taken as a proof of principle for the functionality of the proposed method.

Moreover, no loss of information occurred if a reduction of the data to six dimensions, represented by the first six PCs, is performed as deduced from the cumulative percentage of information in Fig. (**3**). The result showed exactly the same structure of clusters containing exactly the same substances within each cluster as obtained from the Cluster Analysis in the first part of this study (Fig. **2**).

5. CONCLUSIONS AND OUTLOOK

We conclude that a dimensional reduction to six PCs is sufficient to represent the chemical behavior of all single substances for successful reclustering in our study, although we cannot deduce the information of a single PC to derive mainly from one index as the PCs display quite an even mixture of the original indices. Hence, all eleven indices are independent from each other (orthogonal) according to the analysis by our combined CA-PCA method.

Moreover, substances belonging to the same cluster should behave similar so that the results obtained for one representative substance can be assigned to other substances in the same cluster. The behavior of substances could be predicted if new substances are added to the clusters. Upon the analysis of new exhaust air it should ideally be possible to assign certain substances to already existing clusters solely upon their physicochemical values which are not any more to be analyzed by each single original index but only by the reduced set of PCs according to our combined CA-PCA process.

Based on the knowledge about their predicted physicochemical behavior it might be possible to give a first estimation about which substances need necessarily to be removed by e. g. a selective adsorption process from the exhaust air to guarantee the removal of any malodor or even toxic substance, respectively. In the future, it might therefore not be necessary to analyze all substances but just a few representative substances in experiments. Both the addition of new substances to the already existing cluster structure and their experimental validation of the chemical behavior are currently being investigated in our laboratories. In total, this would lead to a significant reduction of money (time, complexity and number of experiments) in purification process development for selective adsorption of malodorous or toxic substances from exhaust air.

ABBREVIATIONS

CA = Cluster Analysis

DBD = Dielectric barrier discharge

ESP = Electrostatic precipitator

PC = Principal Component

PCA = Principal Component Analysis

QSAR = Quantitative structure-activity relationship

QSPR = Quantitative structure-property relationship

CONFLICT OF INTERESTS

The author(s) confirm that this article content has no conflicts of interest.

ACKNOWLEDGEMENTS

We thank the Helmut-Schmidt-University/ University of the Federal Armed Forces Hamburg for the financial support of this study.

APPENDIX

Supplementary Material

The original data matrix including all single index values is shown in Table S1 of the supplementary material. The list of 30 substances along with their structures and cluster numbers can be found in Table S2. Chapter S 3.3.1 describes the minimal mathematical example we constructed and calculated to first establish our process.

Table S1. Original Data Matrix Including All Single Index Values Subsequently Subjected to the Combined Cluster Analysis and Principal Component Analysis. This Table was Imported into MATLAB § acco Rding to Chapter 3.3.

	ZI	WI	RI	BI	IIAC	IIMB	IIEC	TII	MM	WS (g/L)	VP (hPa)
Acetone	1,2955	0,4529	1,5710	8,6095	4,8000	11,1000	0,4155	0,4231	5,908	10000,0000	246,00
Cadaverine	1,2060	0,0000	2,8685	9,6166	5,3333	37,3333	0,4232	0,2900	4,866	1021,8000	0,90
Dimethyldisulfide	1,3710	0,0000	1,3710	16,3241	4,6000	12,7000	0,4207	0,3563	9,420	2,5000	38,00
Acetic Acid	1,5000	0,5177	2,4056	11,5733	4,2500	7,8750	0,4352	0,4529	7,506	602,9000	16,00
2 -- Heptanone	1,1433	0,2636	3,6635	9,9016	5,4545	40,7727	0,4161	0,2781	5,190	4,3000	4,50
n -- Heptane	0,8865	0,0000	2,8057	7,9391	5,5652	42,9565	0,4130	0,2770	4,357	0,0500	47,00
Piperidine	1,1661	0,0000	2,7933	9,1015	5,8824	23,0588	0,4209	0,2096	5,009	10000,0000	33,00
Propionic Acid	1,4354	0,4264	2,8454	11,3557	4,7273	13,0909	0,4301	0,4044	6,735	20,2500	4,00
Pyrrolidine	1,1981	0,0000	2,3788	8,7377	5,8571	16,7143	0,4218	0,2418	5,080	10000,0000	170,00
Skatol	1,2448	1,0526	3,9977	13,0239	5,5789	30,6316	0,4519	0,1275	6,904	0,4500	1,33
Toluene	0,9968	0,9710	3,0566	10,2249	5,2000	21,2000	0,4450	0,1979	6,143	0,4700	29,00
Triethylamine	1,0907	0,0000	2,0254	8,9828	5,4545	38,4545	0,4144	0,2970	4,600	133,0000	69,00
Water	0,9183	0,0000	0,9183	4,7354	2,0000	1,3333	0,4714	0,5443	6,007	10000,0000	23,40
Benzene	1,0000	1,0000	1,0000	10,5061	5,0000	14,5000	0,4553	0,2300	6,509	1,8000	100,00
4 -- Isopropenyl -- 1 -- Methylcyclohex-ene	0,9612	0,3912	4,0270	9,8942	5,6923	49,0000	0,4266	0,1500	5,240	0,0300	1,90
Ethanol	1,2244	0,0000	2,4194	7,3672	4,6667	9,1111	0,4234	0,4605	5,119	10000,0000	58,00
Pyridine	1,3486	0,9940	2,7322	12,5100	4,9091	12,7273	0,4579	0,2387	7,191	10000,0000	20,50
2, 5 - Dimeth-ylpyrazine	1,4056	0,9544	2,4056	13,6475	5,2500	25,5000	0,4395	0,1743	6,759	32,0000	4,00
Hexanal	1,1674	0,2933	3,5766	9,8437	5,3684	31,2632	0,4181	0,3115	5,272	5,0000	12,00
1, 3 -- Pentadiene	0,9612	0,6000	3,1808	8,3981	4,6154	18,7692	0,4366	0,3336	5,240	0,6900	530,00
n -- Hexane	0,8813	0,0000	2,4464	7,4547	5,5000	33,8000	0,4125	0,3041	4,309	0,1600	162,00
Methylcyclopentane	0,9183	0,0000	2,7947	7,7379	6,0000	24,9444	0,4167	0,2084	4,676	0,0420	147,00
Heptanal	1,1433	0,2636	3,7888	10,0792	5,4545	40,0909	0,4179	0,2834	5,190	1,2500	0,86
Benzaldehyde	1,2958	1,2958	3,2359	13,6649	5,0000	19,1429	0,4553	0,2044	7,580	3,3000	1,30
n -- Decane	0,8960	0,0000	3,2028	9,0423	5,6875	76,4375	0,4141	0,2186	4,446	0,00005	1,00
3 -- Octanone	1,1239	0,2399	3,8635	10,1408	5,4167	49,3333	0,4194	0,2495	5,128	4,5000	2,00
1 -- Octen -- 3 -- ol	1,1239	0,2399	4,0537	10,4005	5,3600	50,1600	0,4266	0,2588	5,128	0,0000	1,00
Octanal	1,1239	0,2399	3,9737	10,2971	5,5200	49,9200	0,4177	0,2600	5,128	0,5600	1,00
Nonanal	1,1078	0,2204	4,1375	10,4996	5,5714	60,7500	0,4176	0,2401	5,080	0,0960	0,35
2 -- Nonanone	1,1078	0,2204	4,0391	10,3601	5,5714	61,5000	0,4162	0,2370	5,080	0,3700	0,83

Table S2. Molecular Structure of the Substances Analyzed in the Combined Cluster Analysis and Subsequent Two-step Principal Component Analysis

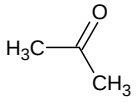
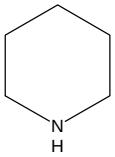
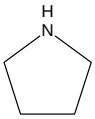
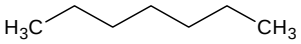
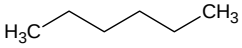
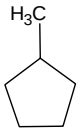
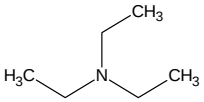
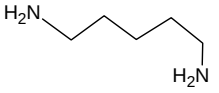
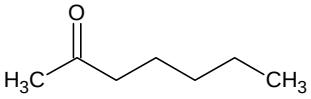
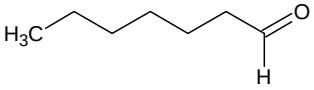
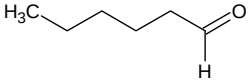
Name	Structure
Cluster 1	
Acetone	
Name	**Structure**
Ethanol	CH_3CH_2OH
Piperidine	
Pyrrolidine	
Cluster 2	
n-heptane	
n-hexane	
Methylcyclopentane	
Triethylamine	
1,3-pentadien	
Cluster 3	
Cadaverine	
2-heptanon	
Heptanal	
Hexanal	

Table S2. Contd…..

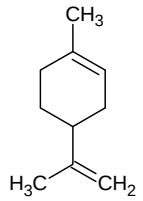
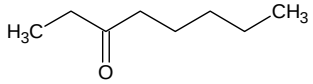
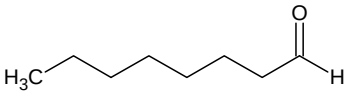
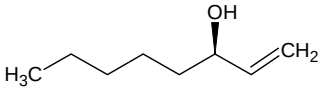
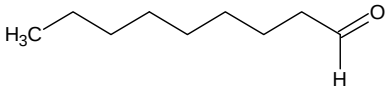
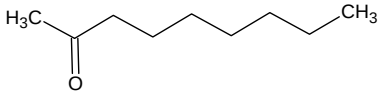
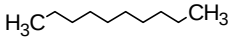
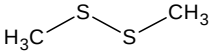
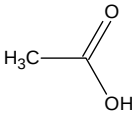
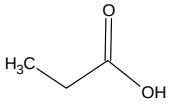
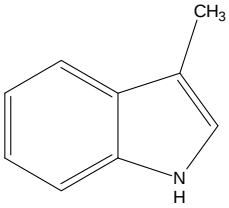
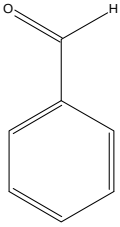
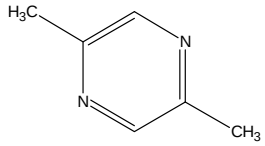
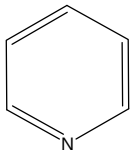
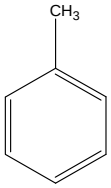
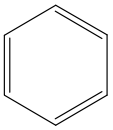
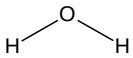
4-isopropenyl-1-methylcyclohexene	
3-octanone	
Octanal	
1-octen-3-ol	
Nonanal	
2-nonanon	
n-decane	
Cluster 4	
Dimethyldisulfide	
Cluster 5	
Acetic acid	
Propionic acid	
Cluster 6	
Skatole	
Benzaldehyde	

Table S2. Contd…..

2,5-dimethylpirazine	
Pyridine	
Toluene	
Benzol	
Cluster 7	
Water	

S 3.3.1. Descriptive Minimal Mathematical Example

Here we present our procedure first on a constructed data set. The set is minimal as we need at least three indices for being able to apply our two-step procedure without getting trivial intermediate results for our Principal Components. Furthermore, we need at least 3 different clusters which should ideally be represented by only two Principal Components. Of course, those clusters should not all contain only one substance. Thus, we set up with a minimum of four substances spread over three clusters. The three index values for each substance s (index number indicates cluster) can be mathematically expressed *via* the following vectors:

$$s_{1a} = \begin{pmatrix} 15 \\ -9 \\ 21 \end{pmatrix};; s_{1b} = \begin{pmatrix} 21 \\ -9 \\ -3 \end{pmatrix};; \; s_2 = \begin{pmatrix} 21 \\ -6 \\ 12 \end{pmatrix} and \; s_3 = \begin{pmatrix} 9 \\ -18 \\ 0 \end{pmatrix}$$

The three main foci of the clusters result in:

$$c_1 = \begin{pmatrix} 18 \\ -9 \\ 9 \end{pmatrix};; \; c_2 = \begin{pmatrix} 21 \\ -6 \\ 12 \end{pmatrix} and \; c_3 = \begin{pmatrix} 9 \\ -18 \\ 0 \end{pmatrix}$$

In order to achieve an equal weight and subsequently an equal contribution to the further data processing, a ``z-score" transformation is applied. First, this includes a shift among the arithmetic mean

$$\mu = \begin{pmatrix} 16 \\ -11 \\ 7 \end{pmatrix}$$

and as a second step, data are scaled regarding the standard deviation in each component. As the standard deviation is already equal in each component of our minimal example, we passed on without the scaling step as it would not influence our result at all but facilitate further calculations. The calculated index values for the transformed data resulted in the shifted vectors for the substances:

$$\tilde{s}_{1a} = \begin{pmatrix} -1 \\ 2 \\ 14 \end{pmatrix};; \tilde{s}_{1b} = \begin{pmatrix} 5 \\ 2 \\ -10 \end{pmatrix};; \tilde{s}_2 = \begin{pmatrix} 5 \\ 5 \\ 5 \end{pmatrix} \text{ and } \tilde{s}_3 = \begin{pmatrix} -7 \\ -7 \\ -7 \end{pmatrix}$$

and the main foci:

$$\tilde{c}_1 = \begin{pmatrix} 2 \\ 2 \\ 2 \end{pmatrix};; \tilde{c}_2 = \begin{pmatrix} 5 \\ 5 \\ 5 \end{pmatrix} \text{ and } \tilde{c}_3 = \begin{pmatrix} -7 \\ -7 \\ -7 \end{pmatrix}$$

The correlation matrix of the main foci is given by

$$\begin{pmatrix} 78 & 78 & 78 \\ 78 & 78 & 78 \\ 78 & 78 & 78 \end{pmatrix}$$

and the corresponding eigenpairs are calculated to the following eigenvalues and eigenvectors:

$$\lambda_1 = 234, v_1 = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix};; \lambda_{2,3} = 0, v_2 = \begin{pmatrix} 1 \\ 0 \\ -1 \end{pmatrix} \text{and } v_3 = \begin{pmatrix} 1 \\ -2 \\ 1 \end{pmatrix}$$

Each eigenvector is a Principal Component with its eigenvalue representing its weight in terms of original information it contains. So the first Principal Component in our minimal descriptive example is as follows:

$$pc_1 = v_1 = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$$

With the first Principal Component all transformed data of the substances can be expressed as:

$$\left(\tilde{s}_{1a}\ \tilde{s}_{1b}\ \tilde{s}_2\ \tilde{s}_3\right) = \left(\begin{array}{ccc} 1 & 1 & -1 \\ 1 & 0 & 2 \\ 1 & -1 & -1 \end{array}\right) \left(\begin{array}{cccc} 5 & -1 & 5 & -7 \\ -7.5 & 7.5 & 0 & 0 \\ -1.5 & 1.5 & 0 & 0 \end{array}\right)$$

It is clear, that $\tilde{s}_{1a}$ and $\tilde{s}_2$ are not discriminable as they both have the same value of ``5'' for their first index entry. The matrix decomposition did not lead to the identification of any further Principal Components. Hence, it is sensible to repeat the whole process on the remaining data basis taking into account the complete variance of the data and not only the main foci, as it was carried out in the first round of PCA. The remaining data basis is given as

$$\begin{pmatrix} -7.5 & 7.5 & 0 & 0 \\ -1.5 & 1.5 & 0 & 0 \end{pmatrix}$$

with its correlation matrix

$$\begin{pmatrix} 225/2 & 45/2 \\ 45/2 & 9/9 \end{pmatrix}$$

and the corresponding eigenpairs (containing again an eigenvector and its eigenvalue):

$$\nu_1 = 117, u_1 = \begin{pmatrix} 5 \\ 1 \end{pmatrix};; \nu_2 = 0, u_2 = \begin{pmatrix} -1 \\ 5 \end{pmatrix}$$

From the first eigenpair of this second round of PCA the next Principal Component (second in the total process) can be calculated as follows:

$$pc_2 = \begin{pmatrix} 1 & -1 \\ 0 & 2 \\ -1 & -1 \end{pmatrix} \begin{pmatrix} 5 \\ 1 \end{pmatrix} = \begin{pmatrix} 4 \\ 2 \\ -6 \end{pmatrix} \propto \begin{pmatrix} 2 \\ 1 \\ -3 \end{pmatrix}$$

The transformed data are now applied to this new basis given by the Principal Components (supplemented with the orthogonal complement) and calculated as follows:

$$\left(\tilde{s}_{1a}\ \tilde{s}_{1b}\ \tilde{s}_2\ \tilde{s}_3\right) = \left(\begin{array}{ccc|c} 1 & 2 & 4 & \\ 1 & 1 & -5 & \\ 1 & -3 & 1 & \end{array}\right) \left(\begin{array}{cccc} 5 & -1 & 5 & -7 \\ -3 & 3 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{array}\right)$$

As in this minimal descriptive example the last row of the matrix only shows entries of ``0", the four substances can be correctly displayed by the two Principal Components, as they already cover 100 % of the original information.

REFERENCES

[1] Mayer H. Air pollution in cities. Atmos Environ 1999; 33(24-25): 4029-37.

[2] Wolterbeek B. Biomonitoring of trace element air pollution: principles, possibilities and perspectives. Environ Pollut 2002; 120(1): 11-21.

[3] Chaichanawong J, Tanthapanichakoon W, Charinpanitkul T, Eiadua A, Sano N, Tamon H. High-temperature simultaneous removal of acetaldehyde and ammonia gases using corona discharge. Sci Technol Adv Mater 2005; 6(3-4): 319.

[4] Qu GZ, Li J, Li GF, Wu Y, Lu N. DBD regeneration of GAC loaded with acid orange 7. Asia-Pacific J Chem Eng 2009; 4(5): 649-53.

[5] Robers A, Figura M, Thiesen PH, Niemeyer B. Desorption of odor-active compounds by microwaves, ultrasound, and water. AI ChE J 2005; 51(2): 502-10.

[6] Konrad G, Eigenberger G. Rotary adsorbers for waste air purification and solvent recovery. Chemie Ingenieur Technik 1994; 66(3): 321-31.

[7] Salden A, Eigenberger G. Multifunctional adsorber/reactor concept for waste-air purification. Chem Eng Sci 2001; 56(4): 1605-11.

[8] Munoz R, Sivret EC, Parcsi G, *et al*. Monitoring techniques for odour abatement assessment. Water Res 2010; 44(18): 5129-49.

[9] Zvinavashe E, Murk AJ, Rietjens IMCM. On the number of EINECS compounds that can be covered by (Q)SAR models for acute toxicity. Toxicol Lett 2009 10; 184(1): 67-72.

[10] Gardner JW. Detection of vapours and odours from a multisensor array using pattern recognition Part 1. Principal component and cluster analysis. Sensors Actuators B: Chem 1991; 4(1-2): 109-15.

[11] Yau SS, Chang SC. A direct method for cluster analysis. Pattern Recognition 1975; 7(4): 215-24.

[12] Wold S. Pattern recognition by means of disjoint principal components models. Pattern Recognit 1976; 8(3): 127-39.

[13] Rannar S, Andersson PL. A Novel Approach Using Hierarchical Clustering To Select Industrial Chemicals for Environmental Impact Assessment. J Chem Inf Model 2010; 50(1): 30-6.

[14] Willett P. Clustering tendency in chemical classifications. J Chem Inf Comput Sci 1985; 25(2): 78-80.

[15] Abdi H, Williams LJ. Principal component analysis. WIREs Comput Stat 2010; 2(4): 433-59.

[16] Balaban AT. Applications of graph theory in chemistry. J Chem Inf Comput Sci 1985; 25(3): 334-43.

[17] Basak SC, Magnuson VR, Niemi GJ, Regal RR, Veith GD. Topological indices: their nature, mutual relatedness, and applications. Math Model 1987; 8(0): 300-5.

[18] Basak SC, Magnuson VR, Niemi GJ, Regal RR. Determining structural similarity of chemicals using graph-theoretic indices. Dis Appl Math 1988; 19(1-3): 17-44.

[19] Balaban AT, Bonchev D, Scitz WA. Topological/chemical distances and graph centers in molecular graphs with multiple bonds. J Mol Struct 1993; 280(2-3): 253-60.

[20] Basak SC, Niemi GJ, Veith GD. A graph-theoretic approach to predicting molecular properties. Math Comput Model 1990; 14: 511-6.

[21] El-Sonbaty Y, Ismail MA. On-line hierarchical clustering. Pattern Recognit Lett 1998; 19(14): 1285-91.

[22] Questier F, Walczak B, Massart DL, Boucon C, de Jong S. Feature selection for hierarchical clustering. Anal Chim Acta 2002; 466(2): 311-24.

[23] Likas A, Vlassis N, Verbeek J. The global k-means clustering algorithm. Pattern Recognit 2003; 36(2): 451-61.

[24] Barnard JM, Downs GM. Clustering of chemical structures on the basis of two-dimensional similarity measures. J Chem Inf Comput Sci 1992; 32(6): 644-9.

[25] Pollard D. Strong Consistency of K-Means Clustering. Ann Stat 1981; 9(1): 135-40.

[26] Massart B, Guo Q, Questier F, *et al*. Data structures and data transformations for clustering chemical data. TrAC Trends Anal Chem 2001; 20(1): 35-41.

[27] Turk M, Pentland A. Eigenfaces for Recognition. J Cogn Neurosci 1991; 3(1): 71-86.

[28] Forgács E, Cserháti T. Use of cluster and principal component analysis in quantitative structure-retention relationship study. Anal Chim Acta 1997; 348(1-3): 481-7.

[29] Mahlke IT, Thiesen PH, Niemeyer B. Chemical Indices and Methods of Multivariate Statistics as a Tool for Odor Classification. Environ Sci Technol 2007; 41(7): 2414-21.

[30] Gutman I, Trinajstic N. Graph theory and molecular orbitals. Total f-electron energy of alternant hydrocarbons. Chem Phy Lett1972; 17(4): 535-8.

[31] Dubes R, Jain AK. Clustering techniques: The user's dilemma. Pattern Recognit 1976; 8(4): 247-60.

[32] Jain A, Nandakumar K, Ross A. Score Normalization in Multimodal Biometric Systems. Pattern Recognit 2005; 38: 2270-85.

[33] Tu IP, Chen H, Chen X. An Eigenvector Variability Plot. Stat Sinica 2009; 19(4): 1741-54.

[34] Linusson A, Wold S, Nordén B. Fuzzy Clustering of 627 Alcohols, Guided by a Strategy for Cluster Analysis of Chemical Compounds for Combinatorial Chemistry. Chemometr Intell Lab Syst 1998; 44: 213-27.

[35] Tipping ME, Bishop CM. Probabilistic Principal Component Analysis. J R Stat Soc Series B Stat Methodol 1999; 61(3): 611-22.

Received: March 22, 2013 Revised: May 28, 2013 Accepted: June 10, 2013