The Test of Organic Solvents Vapours Based on Commercial Tin Dioxide Gas Sensors

Libor GAJDOSIK

1Department of Telecommunications, Faculty of Electrical Engineering and Computer Science, VSB–Technical University of Ostrava, 17. listopadu 15/2172, 708 33 Ostrava-Poruba, Czech Republic
libor.gajdosik@vsb.cz

Abstract. This article describes the experiments carried out with the system of three tin dioxide gas sensors MQR 1003, TGS 822 and TGS 813, as well as the achieved results. Vapours of five industrial organic solvents were tested. The measurement values were processed by a cluster analysis. The programmes STATGRAPHIC 5.0 Plus and MATLAB 7.11 were used for data processing. The achieved results show, that the used commercial sensors are able to successfully distinguish chosen industrial solvents and the results can be used for discrimination of organic solvents vapours in air.

Keywords
Cluster analysis, electronic nose, gas sensor, tin dioxide.

1. Introduction

Metal oxide gas sensors have been produced for many years. Tin dioxide gas sensors are the most popular. These devices are small, high sensitive and relatively cheap, which enables their usage in portable equipment. Their practical applications are either leakage detectors, combustion gases alarms or electronic noises.

The metal oxide gas sensors have relative wide selectivity. Their detection properties depend on the operating conditions and there also exists the possibility of poisoning the sensor. Long term stability of the detection properties is also important. The research efforts are focused on the improvement of the imperfections. One direction leads to the researching new materials [1], [2], [3], [4], the second direction is focused on proper operating regime [5], [6], [7], [8] and proper methods of data processing [9], [10], [11], [12]. The system of several sensors based on various principles is used in electronic noses, as well as statistical methods for data processing, e.g. the principal component analysis [13], [14], the cluster analysis [15], [16], pattern recognition by using an artificial neural network [17], [18], [19].

The research was motivated by the following questions. Is it possible to construct a simple electronic nose based on the commercial gas sensors? What is the sensitivity and discrimination ability of the used sensors to the chosen substances? The set of the substances was chosen purposely because of their easy availability and a wide range of application as paint thinners or stain cleaners.

2. Experiments

The measuring apparatus, the block diagram of which is in Fig. 1, was used. The apparatus is composed of a testing glass chamber of a known volume, and the sensors are placed in it. Dry laboratory air, dried in the unit 7 (20 % RH), is led into the chamber. The sensors are heated by the heating voltage from 0 to 5 V. For the response measurement (electrical conductance of the detection layer) an 8 bit A/D converter is used. The principle of the response measurement is shown in Fig. 2. The heating system of the sensor is galvanically insulated from the detection system of the sensor. The response is measured as the direct voltage $U_m$ across the resistor $R$. The detection system of the sensor and resistor $R$ act as a voltage divider and the changes of electrical conductance of the sensor are converted to the changes of $U_m$. Unit 5 liquid manometer keeps the atmospheric pressure in the gas chamber.

Sensors TGS813 and MQR 1003 are meant for the methane detection, sensor TGS 822 is meant for the ethanol detection. The testing was carried out with each of the 5 organic solvents separately: C 6000, S 6006, S 6300, P 6413, CIKULI. These are the commercial marks of the used industrial solvents. Each of these solvents is a mixture of several organic substances.
The saturated vapours of the solvents were obtained by the solvents evaporating above their liquid phase in a closed bottle at the constant ambient temperature 22°C. A proper volume (20 ml) of the saturated vapour was injected into the testing chamber (2700 ml) filled with dry clean air. The temperature of the apparatus and the closed bottle was kept at the ambient temperature.

The heating voltage was first increased to 5 V before each measurement of the given solvent to prepare the sensor for measurement [4]. The electrical conductance of the detection layer was tracked until it had reached the value in dry clean air. After that the heating voltage was decreased by step to 2 V and the saturated vapours were injected into the testing chamber. The temperature of the sensor is proper for chemisorption of the substance at this heating voltage [4], [10], [20]. Afterwards, the heating voltage was increased by step to 3 V. The range of 3 V up to 3.5 V was chosen purposely because it was experimentally discovered, that the response of the mentioned solvents is of the maximum value. Starting with the value of 3 V the heating voltage was increased every 10 seconds by 14 equal steps up to the value 3.5 V. The electrical value of the detected layer was sensed every 10 seconds after the change of the heating voltage, because of thermal inertia of the sensor described in [20]. At the end of the measurement the testing chamber was purged by dry clean air and the heating voltage was increased by step to 5 V to prepare the sensor for the following measurement. The measurement with each solvent, including its new preparation of the concentration, was repeated 10 times. The response of sensors for solvent S 6006 is shown in Fig. 3.

3. Methods

Cluster analysis separates data into groups (clusters). This method is suitable for the data the natural property of which is forming of clusters. Generally, the goal of the cluster analysis is to separate the measured data into several fairly homogeneous clusters. There are two types of tasks. The goal of the first type is to split the data into the prescribed number of clusters, the goal of the second type of the task is to find out the optimal number of clusters for given data. The tasks with the prescribed number of clusters are solved by agglomerative or divisive methods [21].

Agglomerative clustering starts by considering each data point to be an independent cluster. Two clusters are merged at each step and the process is repeated until the desired number of clusters is obtained. The agglomerative methods, called the nearest neighbour, the furthest neighbour, centroid, median, group average and Ward’s method are used here.

Divisive clustering starts by putting all data in a single cluster. A single cluster is split into two clusters and the centroid of each one is calculated. New clusters are created by the division of the existing ones and the process is repeated until the desired number of clusters is obtained. Divisive method called k-means was used here.

For clustering either the data measured in their natural units or the normalized data can be used. In our case it is more convenient to use the normalized data, since the obtained results are independent on the used units. The normalized data are calculated by the fol-
The larger ones.

large clusters and the small clusters are merged into

disadvantage is that this method tends to break the
tends to form smaller more balanced clusters. The
method eliminates the disadvantage of the nearest
neighbour method. The furthest neighbour method
eliminates the disadvantage of the nearest

The following formulas were used:

\[ x_{ij} = \frac{z_{ij} - \bar{z}_j}{S(z_j)} \quad \text{for } i = 1, 2, \ldots, n \quad j = 1, 2, \ldots, p, \] (1)

\[ \bar{z}_j = \frac{1}{n} \sum_{i=1}^{n} z_{ij}, \]

\[ S^2(z_j) = \frac{1}{n-1} \sum_{i=1}^{n} (z_{ij} - \bar{z}_j)^2. \] (2)

where \( x_{ij} \) is the normalized value, \( z_{ij} \) is the original value, \( \bar{z}_j \) is the mean value, \( S(z_j) \) is the standard deviation, \( n \) is the total number of input data of all solvents, \( p \) is the number of used sensors.

Clustering is also influenced by the type of distance
calculation used between two clusters. Squared Eu-
clidean distance was chosen as a proper distance mea-
sure since it puts greater weight on the objects that
are farther apart and also it is also recommended for
some methods of clustering described in this article.
Squared Euclidean distance is calculated by formula:

\[ D(x_i, x_m) = \sum_{j=1}^{p} (x_{ij} - x_{mj})^2, \] (3)

where distance \( D(x_i, x_m) \) is calculated between i-th ob-
ject and m-th object.

The original clusters are merged into a new cluster
by a linkage method. For the nearest neighbour linkage
method this formula was used:

\[ D_{gm} = \min(D_{rm}, D_{sm}), \] (4)

where \( D_{gm} \) is the distance between new cluster g and
others clusters m, \( D_{rm} \) is the distance between origi-

The following formula was used for the furthest
neighbour linkage method:

\[ D_{gm} = \max(D_{rm}, D_{sm}). \] (5)

Two maximally distant clusters are merged. This
method eliminates the disadvantage of the nearest
neighbour method. The furthest neighbour method
tends to form smaller more balanced clusters. The
disadvantage is that this method tends to break the
large clusters and the small clusters are merged into the
larger ones.

This formula was used for the centroid linkage
method:

\[ D_{gm} = \frac{n_r D_{rm} + n_s D_{sm} - \frac{n_r n_s}{n_r + n_s} D_{rs}}{n_r + n_s}, \] (6)

where \( n_r \) is the number of the objects in cluster r,
and \( n_s \) is the number of the objects in cluster s. The
other symbols have the same meaning as in the pre-
vious paragraph. The distance between two clusters
is defined as squared Euclidean distance between their
centroids. One disadvantage of this method is that the
distance at which clusters are combined can decrease
from one step to the next one, thus the farther clusters
can be merged. The centroid method is more robust
to outliers than most of the other methods.

For the median method this formula was used:

\[ D_{gm} = \frac{D_{rm} + D_{sm}}{2} - \frac{D_{rs}}{4}. \] (7)

In case of the median method the same importance is
attached to two merged clusters regardless of how many
objects there are in each cluster. This is an advantage
compared to the centroid method. It is proper to use
squared Euclidean measure.

For the group average method this formula was used:

\[ D_{gm} = \frac{D_{rm} + D_{sm}}{n_r + n_s}. \] (8)

The distance between two clusters is the average dis-
tance between pairs of objects, one of which is placed
in each cluster. This method is characterized by a
compromise between the nearest and the farthest neigh-
bours. The method is less susceptible to outliers. Its
disadvantage is the tendency to form globular clusters.
The group average tends to join clusters with small
variances, and it is slightly biased toward producing
clusters with the same variance.

For Ward’s method this formula was used:

\[ D_{gm} = \frac{[(n_r + n_m)D_{rm}] + (n_s + n_m)D_{sm} - n_m D_{rs}}{n_r + n_s + n_m}. \] (9)

The distance is calculated as a total sum of the
squared deviations from the mean of the cluster. Two
clusters having the smallest possible increase in the
error sum of squares are merged. Ward’s method tends
to form clusters of an equal number of objects and to re-
duce small clusters. The method is similar to the group
average method and the centroid method. Therefore,
it is proper to use squared Euclidean measure. The
method is sensitive to outliers. This method is very
efficient and fits well if each cluster contains the equal number of objects.

K-means method does not require computation of all possible distances. The objects are assigned to the prescribed number of k clusters. Each object is assigned to the cluster with the shortest distance to the cluster mean. The distance between object \( x_i \) and cluster mean \( c_m \) is calculated by the squared Euclidean formula:

\[
D(x_i, c_m) = \sum_{j=1}^{p} (x_{ij} - c_{mj})^2.
\]  

\( \text{Equation 10} \)

The cluster mean is calculated by the following formula:

\[
c_j = \frac{1}{n} \sum_{i=1}^{n} x_{ij},
\]  

\( \text{Equation 11} \)

where \( x_{ij} \) are the objects assigned to the cluster, \( n \) is the number of the objects in the cluster. The algorithm starts with the initial set of means and it classifies the objects based on their distances to the cluster means. Cluster means are computed again using the objects that are assigned to the cluster. Afterwards, all objects are reclassified on the basis of a new set of the means. These steps are repeated until the cluster means change. Finally, all cases are assigned to their permanent clusters. It follows from the algorithm that during the analysis the same object can move from one cluster to another one.

The cophenetic correlation coefficient is used for the measure of validity of a clustering structure. The coefficient is calculated by this formula For Ward’s method this formula was used:

\[
R = \frac{a \sum_{j=1}^{n-1} \sum_{j=1}^{n} D_{ij} C_{ij} - \mu_D \mu_C}{\sqrt{a \sum_{j=1}^{n-1} \sum_{j=1}^{n} D_{ij}^2 - \mu_D^2} \sqrt{a \sum_{j=1}^{n-1} \sum_{j=1}^{n} C_{ij}^2 - \mu_C^2}}
\]  

\( \text{Equation 12} \)

where \( D_{ij} \) and \( C_{ij} \) are the elements of the distance matrix \( D \) and cophenetic matrix \( C \). Matrix \( D \) contains the distances between i-th and j-th object, matrix \( C \) contains the distances at which these two objects are first joined together. The \( \mu_D \) and the \( \mu_C \) are the mean values of the matrixes. For good validity \( R \) should be \( R > 0.75 \) and close to 1.

The correlation coefficient of the used sensors can be calculated by the formula:

\[
r_{ik} = \frac{\sum_{j=1}^{n} (x_{ij} - \bar{x}_i) (x_{kj} - \bar{x}_k)}{\sqrt{\sum_{j=1}^{n} (x_{ij} - \bar{x}_i)^2} \sqrt{\sum_{j=1}^{n} (x_{kj} - \bar{x}_k)^2}}
\]  

\( \text{Equation 15} \)

where \( x \) are the normalized values of the responses and \( i, k \) go from 1 up to \( p \). It is obvious from the presented properties, that used linkage methods differ from each other. Therefore, a successive application of these methods on the same measured data generally leads to different assignment into the prescribed number of clusters. As the correct cluster assignments are given by the numbers and kinds of the tested solvents, it is easy to check if the used clustering method splits data into the clusters correctly or not. If a larger number of the used methods splits data into the clusters correctly, it means that such data are better distinguishable in comparison with the case in which only a smaller number of methods splits data correctly. It is possible to decide according to the value of \( R \) which linkage method fits better. The correlation coefficient \( r_{ik} \) is the measure of mutual similarity of the used sensors.

4. Result

Measured values of the responses were processed by a cluster analysis in programme STATGRAPHIC 5.0 Plus. Coefficient \( R \) was calculated by programme MATLAB 7.11 as the above-mentioned STATGRAPHIC is not able to calculate it. The measured values of the responses of three sensors for the tested substance at the given heating voltage present a cluster of 10 points in a three-dimensional area. Theoretically, there is a total of 5 clusters for 5 solvents if the sensors show mutual different sensitivity to the above-mentioned solvents. If the sensitivity to some solvents was the same for all sensors, the clusters of the solvents would fuse and could not be discriminated.

The ability of the sensor system to successfully discriminate the given number of the tested solvents under the given conditions was investigated by the cluster analysis. Successful discrimination means the faultless classification of the given number of substances into the same number of clusters.

In program STATGRAPHIC 5.0 Plus seven methods of clustering were used: the nearest neighbour method, the furthest neighbour method, centroid, median, group average, Ward’s method and k-means. Squared Euclidean as a distance metric and standard-
ized data preprocessing were considered to be appropriate and that is why used here. When the number of successful discriminations in the given case increases, the clusters are better discriminable. The data taken at a certain level of the heating voltage represent one case of the analysis. Thus there is a total of 14 cases for each response. It follows from the obtained results that there is an optimal value of the heating voltage, at which the possibility of discriminating the tested solvents is the best one. In the heating voltage levels ranging between 3 V and to 3,5 V the responses of the sensors show the maximal values for the tested solvents. At this operating point the sensors show the maximal sensitivity according to the equation [15] in [22]. It means that this is the best operating point:

$$ S = \frac{G - G_0}{G_0}, $$

where $S$ is the sensor sensitivity, $G_0$ is the electrical conductance of the sensor in clean air, $G$ is the electrical conductance of the sensor in tested gas. Thus, in the file of the measured data the maximum value of the reached electrical conductivity was searched out for each response and the reached value was used as the input data for processing.

At first, the data obtained from 5 solvents measured by 3 sensors were evaluated. All 7 methods were used simultaneously to faultlessly distinguish the 5 clusters. Six methods, except k-means, were successful in case of 5 solvents and 3 sensors. Coefficient $R$ ranged between 0,8213 and 0,9914 for all successful methods. The correlation matrix was calculated by using STATGRAPHIC. The results are shown in Tab. 1.

**Tab. 1**: Correlation coefficient $r$ of the used sensors. All 50 measured values were used.

<table>
<thead>
<tr>
<th></th>
<th>MQR 1003</th>
<th>TGS 813</th>
<th>TGS 822</th>
</tr>
</thead>
<tbody>
<tr>
<td>MQR 1003</td>
<td>1</td>
<td>0.7552</td>
<td>0.9514</td>
</tr>
<tr>
<td>TGS 813</td>
<td>0.7552</td>
<td>1</td>
<td>0.8093</td>
</tr>
<tr>
<td>TGS 822</td>
<td>0.9514</td>
<td>0.8093</td>
<td>1</td>
</tr>
</tbody>
</table>

It is possible to imagine this result as 5 separated clusters in three-dimensional space, where each sensor represents one coordinate. A scatter plot in three-dimensional space can be drawn as a system of three two-dimensional graphs. It is shown in Fig. 4, 5 and 6.

In the next step the data of 5 solvents taken from each one pair of sensors were processed. For three pairs of used sensors we have three results. Overview of the number of successful methods for cluster separation is shown in Tab. 2.

It can be seen that pair 2 shows low discrimination ability, while pairs 1 or 3 show higher ability. Compared to the case where all three sensors were used it follows that the discrimination ability of the three sensors system is practically given by pair 1 or pair 3. This opinion is confirmed by the values of $r_{ik}$ in Tab. 1 because pair 1 and pair 3 have lower values of $r_{ik}$.

It is possible to improve the number of the successful methods of each pair of the sensors if the number of the tested solvents is reduced. Thus, if the number of the tested solvents in the file of the measured data is reduced to 4, it is possible to constitute 5 combinations. Each combination contains 4 solvents. The combinations are itemized in Tab. 2. The described
methods of the cluster analysis were used for the data of each combination. A number of successful methods used for separation into 4 clusters was evaluated for each combination. The obtained results are in Tab. 3.

It follows from the Table 3, that the most successful pair are pairs 1 and 3, because of the maximum number of successful methods. Pair number 1 successfully detected the group of solvents in three cases (combinations 1, 2 and 5), as well as pair number 3 (combinations 1, 2, 4 and 5). The most proper group of the solvents are the combinations 2 and 5 because they are detectable by two pairs of sensors by 7 methods. The range of cophenetic coefficients R of the used linkage methods for combinations 2 and 5 and pair 1 and the pair 3 are shown in the Tab. 4.

The values in Tab. 4 are a bit lower, those obtained when using a full number of the above mentioned sensors (R=0.8213 and 0.9914) for 5 solvents. It seems that the pair 2 also assists to the detection ability despite of its higher coefficient of correlation \( r_{ik} \).

An example of the cluster scatter plot for successful case is shown in the Fig. 8.

The achieved results show, that system of two properly chosen commercial sensors is able to successfully distinguish 4 solvents if a proper choice of solvents is made.

The obtained results show that the detection system using commercial sensors has its value. It is possible to construct an electronic nose of two or three commercial sensors of a described type to distinguish four or five mentioned solvents separately.

---

**Fig. 7:** The cluster scatter plot of 5 solvents in the system of 3 sensors. Each cluster contains 10 points. The dependency obtained from TGS 822 and TGS 813.

**Tab. 2:** Results of the discrimination of the normalized data of 5 solvents by a pair of sensors. Numbers of successful methods for separation into 5 clusters.

<table>
<thead>
<tr>
<th>Pair 1</th>
<th>Pair 2</th>
<th>Pair 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>MQR 1003</td>
<td>MQR 1003</td>
<td>TGS 813</td>
</tr>
<tr>
<td>TGS 813</td>
<td>TGS 822</td>
<td>TGS 813</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td>Except k-mean</td>
<td>Only Ward’s</td>
<td>Except k-mean</td>
</tr>
</tbody>
</table>

**Tab. 3:** Results of the discrimination of normalized data of 4 solvents by a pair of sensors. Numbers of successful methods for separation into 4 clusters.

<table>
<thead>
<tr>
<th>Pair 1</th>
<th>Pair 2</th>
<th>Pair 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>MQR 1003</td>
<td>MQR 1003</td>
<td>TGS 813</td>
</tr>
<tr>
<td>TGS 813</td>
<td>TGS 822</td>
<td>TGS 822</td>
</tr>
</tbody>
</table>

Combination of solvents number 1
S 6006, S 6300, CIKULI, P 6413

6

Only Nearest, Ward’s, k-mean

Combination of solvents number 2
C 6000, S 6300, CIKULI, P 6413

7

Except k-mean

Combination of solvents number 3
C 6000, S 6006, S 6300, CIKULI

6

Except k-mean

Combination of solvents number 4
C 6000, S 6006, S 6300, CIKULI

6

Except k-mean

Combination of solvents number 5
C 6000, S 6006, CIKULI, P 6413

7

Except k-mean

**Tab. 4:** Range of values of cophenetic coefficient R for the combinations 2 and 5.

<table>
<thead>
<tr>
<th>Pair 1</th>
<th>Pair 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Combination 2</td>
<td>0.8201 – 0.8773</td>
</tr>
<tr>
<td>Combination 5</td>
<td>0.8853 – 0.9591</td>
</tr>
</tbody>
</table>

---

**Fig. 8:** The cluster scatter plot of 4 solvents in the system of 2 sensors. Each cluster consists of 10 points. The combination number 5 and the pair number 3 are used.
5. Conclusion

This article deals with the detection properties of commercial tin dioxide gas sensors. The properties were successfully tested on 5 chosen industrial solvents. It follows from this that it is possible to construct a relative simple discrimination system of industrial organic solvents with commercial tin dioxide gas sensors originally determined for methane or ethanol detection. The nose could use two or three commercial tin dioxide gas sensors and could discriminate four or five frequently used organic solvents. The methods described in this article can be used in practical applications.

Acknowledgment

This article was enabled due to project GAP 108/11/057 entitled Synthesis, structure and properties of nano composites conducting polymer/phylllosilicate provided by GACR (The Czech Science Foundation). This article was created with the active support by the Ministry of Education of the Czech Republic within the project no. SP2013/69 of the VSB–Technical University of Ostrava. The research has been partially supported by the project No. CZ.1.07/2.3.00/20.0217 "The Development of Excellence of the Telecommunication Research Team in Relation to International Cooperation” within the frame of the operation programme "Education for competitiveness" financed by the European Structural Funds and from the state budget of the Czech Republic.

References


About Authors

Libor GAJDOSIK graduated in telecommunication engineering from the Czech Technical University (Prague, Czech Republic) in 1983. He received his Ph.D. in 1998 from the VSB–Technical University of Ostrava (Ostrava, Czech Republic). His thesis was focused on the detection properties of tin dioxide gas sensors at the dynamic operating modes. He has been an assistant professor at the VSB–Technical University of Ostrava since 1989. His main interests are electronics, circuit theory and chemical sensors.