The limits of human recognition abilities do not only depend on experiment performance and potential of human organs, especially of sight and audition, or the applied instruments, but also on the efficiency of the chemometric methods. Chemometrics is the area of science and technology devised to elicit all kinds of useful information from multidimensional data (measurement results), based on statistical and mathematical methods [1]. The importance of chemometrics is growing in science because it offers solutions to highly complex calculation problems required to obtain complete information about objects, processes and phenomena. At the present the chemometric methods have been indispensable in solving many scientific and practical problems in the fields of chemistry, environmental protection, medicine, biology, forensic science, industry and others [2-6].

The use of chemometric methods is indispensable not only for calculations. A number of simple but powerful visualization techniques give the possibility of creating and manipulating with graphical representations of data. Image content transcends the spoken word. For the viewer, this unspoken message can be more potent than spoken or written communication. We use these representations in order to gain better insight and understanding of the problem we are studying. Visual language is an effective vehicle of both context and content. Images can be visual renditions or representations of ideas, objects, dimensions and events. Tapping the power of artistic imagery to transmit quantitative information is a natural extension of visual display. Visual data set images show patterns of response not readily apparent as a page of numbers. It is a useful first to look at visualization in a general context. The process presented in

**Clasification of Drinking Water Samples Using the Chernoff’s Faces Visualization Approach**

A. Astel¹*, K. Astel², M. Biziuk², J. Namieśnik²

¹Environmental Chemistry Research Unit, Biology and Environmental Protection Institute, Pomeranian Pedagogical Academy, 22a Arciszewskiego Str., 76-200 Śluisk, Poland
²Analytical Chemistry Department, Chemical Faculty, Gdańsk University of Technology, 11/12 G.Narutowicza Str., 80-952 Gdańsk, Poland

Received: October 25, 2005
Accepted: March 14, 2006

**Abstract**

Our study shows the importance of drinking water monitoring using simple but powerful visualization tools to better understand spatial variations in water quality. The paper reports Chernoff’s Faces visualization approach applied for the classification of drinking water samples collected at twelve various districts of Gdańsk (Poland), over the period 1993-2000. A good visualization should give the viewer a rapid understanding of the data and the phenomenon behind the data. The complex data matrix containing 1756 results of determination of disinfection by-products (THMs: CHCl₃, CHBrCl₂+C₂HCl₄, CHBr₂Cl, CHBr₃ and organohalogen compounds: CCl₄, CH₂Cl₂, C₂Cl₄, C₂H₃Cl₃) was successfully treated with Chernoff’s approach, yielding two different groups of similarity among the sampling sites, and reflecting different types of drinking water supplies (surface and groundwater).

**Keywords:** drinking water, VOCl, visualization, Chernoff’s Faces, spatial changes in water quality

---

*Corresponding author; e-mail: astel@pap.edu.pl
Fig. 1 is iterative: different visualizations may be needed in order to get the best picture; or indeed improvements may be needed for the mathematical modeling or measurement process in order to gain better data.

Let’s take a given phenomenon, or reality that we are trying to understand; this might be the temperature in the upper atmosphere, or it might be the anatomy of a patient. In both cases, we have only partial knowledge of the phenomenon. In the atmospheric case, we would typically have temperature values predicted by some mathematical model at the nodes of a grid, or mesh. In the medical case, we would have readings from a scanner at discrete slices through the patient in a Magnetic Resonance Imaging scan this will be readings of water content. There is a certain difficulty in using and interpreting large sets of data resulting from long-term monitoring programs, especially if the number of variables is important. Many methods of multivariate graphical display attempt to reduce the information in the data set to a relatively few variables. This reduction to two or three variables causes both a loss of original information and clouds the interpretation of results or research findings [7].

The aim of the present paper is to demonstrate the opportunities offered by Chernoff’s Faces visualization approach to classify the long-term monitoring data obtained by determining the main class of volatile chlorination byproducts in drinking water, which is trihalomethanes (THMs). The group of THMs was also the first category of disinfection by-products (DBPs) identified in water. The presence of bromide in water reservoirs used as sources of drinking water can significantly contribute to the formation of brominated and mixed bromo/chloro DBPs during chlorination [8]. This phenomenon is not limited to surface water only as a result of bromide occurrence in the groundwater of coastal areas due to seawater intrusion [9]. Brominated by-products are suspected to be more harmful to health, and also much stronger carcinogens and mutagens than their chloride-containing analogs [10]. According to the U.S. Environmental Protection Agency, chloroform, dichlorobromomethane and bromoform belong to group B2 (probable human carcinogens), while dibromochloromethane, dichloroacetonitrile, dibromoacetonitrile and chloral hydrate belong to group C (possible human carcinogens) [11]. Epidemiological studies have suggested a possible link between chlorination and chlorination by-products and an increased risk of bladder and rectal cancer [12, 13]. VOCs can be found in chlorinated water supplies and in the indoor air where running water and showers release the chemicals into the room; however, this airborne exposure is minimal compared to that from drinking water [14].

Materials and Methods

Sampling Sites

The long-term monitoring program was implemented on the basis of bilateral agreement between Gdańsk Voivodeship city office and the Department of Analytical Chemistry (Chemical Faculty, Gdańsk University of Technology). The study was carried out over a period of six years (1993-2000) with the exceptions of 1997 and 1998. Water samples were collected seasonally from twelve various districts of the Gdańsk area (Wrzeszcz, Zaspa, Niedźwiędnik, Morena, Chełm, Żabianka, Przymorze, Suchanino, Wrzeszcz-Gdańsk University of Technology, the Old City, Stogi, and Siedlce). Drinking water in the Tricity area is supplied from a raw surface source (Straszyn lake reservoir, located south of the Siedlce and Chełm divisions), from underground sources and by mixing both kinds of water. The raw surface water is pretreated in Straszyn treatment plant prior to use. After rotary sieving, the first step of disinfection is preliminary ozonation. Next, the raw water is subjected to coagulation, flocculation and sedimentation processes. After filtering through gravelers the indirect ozonation is performed and directly before feeding drinking water to the pipeline system the water is slightly, but continuously, chlorinated with a dose of 0.8-1.2 mg L$^{-1}$ of chlorine to obtain a residual chlorine concentration at the level of 0.2-0.5 mg L$^{-1}$. The capacity of the processing line is estimated at 6,000 m$^3$/day and presented in Fig. 2 in the form of block diagram. Groundwater is collected from quaternary, tertiary and cretaceous springs. Water delivered from underground sources is subjected to degassing, deironizing and demanganizing processes. In consequence, the Fe and Mn are immobilized on filter packing and periodically backwashed. Disinfection by sodium hypochlorite (NaOCl) is limited only to the cases of biological contamination of water. The sampling points in various parts of Gdańsk are presented in Fig. 3.

Determination of Analytes

The following analytes were determined in water samples: THMs (CHCl$_3$, CHBrCl$_2$+C$_2$HCl$_3$, CHBr$_3$, CHCl$_2$) and organohalogen compounds (CCl$_4$, CH$_2$Cl$_2$, C$_2$Cl$_4$, C$_2$HCl$_4$). Direct aqueous injection (DAI) into a capillary column of a Carlo Erba (Italy) Vega 6180 GC
system equipped with electron capture detection (ECD 40/400) was used for the determination of VOCs [15]. The chromatographic conditions were as follows: a 30 m x 0.32 mm I.D. fused silica capillary column, coated with bonded 5-μm nonpolar DB-1 phase (J&W Scientific); 2 m x 0.32 mm I.D. fused silica precolumn; temperature program, 102°C isothermally, injection system, cold on-column with secondary cooling; detector, ECD operated at 350°C with pure nitrogen (99.999%) as a make-up gas (30 ml∙min⁻¹); carrier gas, hydrogen at 0.4 m∙s⁻¹; injection volume, 2 μl. The detection limits of the DAI-ECD method, which were dependent on the species being determined, were approximately 0.01 μg∙L⁻¹ on average. A detailed description of the experimental procedure, including calibration and validation, can be found elsewhere [15-17].

Visualization

In 1971, displaying multivariate information was about to improve with the new ideas of Herman Chernoff. He offered a new approach, namely, "Chernoff’s Faces", which are simplified, cartoon-like faces that can be used to graphically display complex multivariate data (n-variables on a two-dimensional surface) and thus help to find similarity in data variability, classify the data and improve conclusions in environmental impact studies. They draw upon the human mind’s innate ability to recognize small differences in facial characteristics and to assimilate many facial characteristics at once. Each of several variables is assigned to a facial characteristic and a face is then generated for each condition [18]. Certain facial features are adjusted within acceptable range of realistic to semi-realistic values. The number of features to adjust are only limited by the same factors that would make one face different from another, however certain features are much more distinguishable than others. In order of importance: area of face, shape of face, length of nose, location of mouth, curve of smile, eyes (location, separation, angle, shape, and width), pupil location and eyebrows (location, angle, and width). A major advantage of this visualization method is the capability to store and show the values of up to 18 different variables per face display. Chernoff’s visualization approach allowed us to detect multivariate
similarities between the content of various organohalo-
gen compounds in drinking water at different districts and thus to classify the drinking water samples. Five features characterized by the maximum level of variation (concentrations of CHCl$_3$, CHBrCl$_2$+C$_2$HCl$_3$, CHBrCl, CH$_2$Cl, C$_2$Cl$_4$) were assigned to facial characteristics: length and width of nose, angle and length of eyebrows and curvature of mouth. Because faces must be created properly the remaining facial characteristics, i.e. shape of face and eyes, width of face, etc., were left as default values.

**Results and Discussion**

The results of determination of VOCl compounds in drinking water in Gdańsk are given in Table 1 and additionally in Table 2. In cases when the analytes were not detected in a sample, the value of one-third of LOD was inserted in the data set due to chemometric requirements [19]. In this study, a commercial statistics software package, Statistica 6.0 for Windows, was used for data visualization [20].

In the Gdańsk area, the total concentration of THMs in drinking water fluctuates between undetectable levels and 51 μg·L$^{-1}$ and is comparable to the concentrations determined and presented elsewhere [9, 21-23]. In general, the THMs often occurring in water were: CHCl$_3$, CHBrCl$_2$+C$_2$HCl$_3$ and CHBr$_2$Cl. For each year investigated, chloroform was more abundant than the brominated species. Although the concentrations of the brominated compounds were usually an order of magnitude lower comparing to CHCl$_3$, they were detected

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>CHCl$_3$</td>
<td>78</td>
<td>3.260</td>
<td>83</td>
<td>5.493</td>
<td>50</td>
<td>4.506</td>
</tr>
<tr>
<td>CHBrCl$_2$+C$_2$HCl$_3$</td>
<td>78</td>
<td>2.371</td>
<td>84</td>
<td>2.997</td>
<td>50</td>
<td>2.404</td>
</tr>
<tr>
<td>CHBrCl</td>
<td>78</td>
<td>0.465</td>
<td>84</td>
<td>0.716</td>
<td>50</td>
<td>0.673</td>
</tr>
<tr>
<td>THMs</td>
<td>78</td>
<td>6.145</td>
<td>84</td>
<td>9.414</td>
<td>50</td>
<td>7.378</td>
</tr>
<tr>
<td>CH$_2$Cl$_2$</td>
<td>-</td>
<td>-</td>
<td>70</td>
<td>0.239</td>
<td>50</td>
<td>0.232</td>
</tr>
<tr>
<td>C$_2$Cl$_4$</td>
<td>-</td>
<td>-</td>
<td>70</td>
<td>0.168</td>
<td>50</td>
<td>0.090</td>
</tr>
</tbody>
</table>

n.o. – number of observations

Table 2. Results of determination of volatile organohalogen compounds (μg·L$^{-1}$) in drinking water samples collected in various districts of Gdańsk.

<table>
<thead>
<tr>
<th>District</th>
<th>CHCl$_3$</th>
<th>CHBrCl$_2$+C$_2$HCl$_3$</th>
<th>CHBrCl</th>
<th>THMs</th>
<th>CH$_2$Cl$_2$</th>
<th>C$_2$Cl$_4$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chełm</td>
<td>6.675</td>
<td>3.031</td>
<td>1.283</td>
<td>11.114</td>
<td>0.551</td>
<td>0.016</td>
</tr>
<tr>
<td>Morena</td>
<td>6.966</td>
<td>3.373</td>
<td>1.625</td>
<td>12.090</td>
<td>0.752</td>
<td>0.001</td>
</tr>
<tr>
<td>Suchanino</td>
<td>7.946</td>
<td>3.937</td>
<td>1.301</td>
<td>13.586</td>
<td>0.464</td>
<td>0.044</td>
</tr>
<tr>
<td>Niedźwiednik</td>
<td>9.291</td>
<td>4.253</td>
<td>1.055</td>
<td>14.715</td>
<td>0.747</td>
<td>0.020</td>
</tr>
<tr>
<td>Przymorze</td>
<td>0.990</td>
<td>0.859</td>
<td>0.193</td>
<td>2.042</td>
<td>0.021</td>
<td>0.140</td>
</tr>
<tr>
<td>Old Town</td>
<td>0.808</td>
<td>0.500</td>
<td>0.278</td>
<td>1.600</td>
<td>0.069</td>
<td>0.040</td>
</tr>
<tr>
<td>Zaspa</td>
<td>0.899</td>
<td>0.479</td>
<td>0.179</td>
<td>1.557</td>
<td>0.086</td>
<td>0.063</td>
</tr>
<tr>
<td>Żabianka</td>
<td>0.954</td>
<td>4.170</td>
<td>0.048</td>
<td>4.678</td>
<td>0.116</td>
<td>0.141</td>
</tr>
<tr>
<td>Wrzeszcz-GUT</td>
<td>3.163</td>
<td>1.713</td>
<td>0.373</td>
<td>5.287</td>
<td>0.264</td>
<td>0.088</td>
</tr>
<tr>
<td>Wrzeszcz</td>
<td>4.205</td>
<td>2.052</td>
<td>0.827</td>
<td>7.076</td>
<td>0.339</td>
<td>0.052</td>
</tr>
<tr>
<td>Stogi</td>
<td>1.299</td>
<td>0.003</td>
<td>0.003</td>
<td>1.295</td>
<td>0.140</td>
<td>0.090</td>
</tr>
<tr>
<td>Siedlce</td>
<td>10.819</td>
<td>5.911</td>
<td>2.069</td>
<td>18.800</td>
<td>1.474</td>
<td>0.265</td>
</tr>
</tbody>
</table>
in most drinking water samples, except for CHBr. In 1993, tetrachloroethane and dichloroethane, and in 1996 and 1999 tetrachloroethane were not detected at any of the examined locations. CCl₄, C₂H₃Cl, and CHBr were not detected at all. The complex data matrix containing 1756 results of determination of disinfection by-products were further visualized in the Chernoff’s Faces plot to classify the drinking water samples and explore their spatial trends (Fig. 4).

The contents of organohalogen compounds in drinking water samples collected in various districts of Gdańsk is diversified. The mean value of THM in water collected by drinking water from surface source is 12.87 μg∙L⁻¹, while from underground sources it is 3.03 μg∙L⁻¹. In the case of the rest of organohalogen compounds (CH₂Cl₂ + CH₂Cl₄) the mean value in districts supplied by drinking water from the surface source is 0.64 μg∙L⁻¹ and from underground sources 0.21 μg∙L⁻¹, respectively. Using Chernoff’s visualization approach the locations with different VOCls content can be clearly distinguished. As a result the drinking water samples collected at the metropolitan area can be classified into two main types of faces. One of them (marked with a star) is characterized by a smile and long noses, which corresponds to higher concentrations of CHCl₃ and CHBrCl₂. The wide nose for this category is connected with higher concentrations of CHBrCl. Districts classified by these faces are supplied with drinking water from Straszyn lake (surface source), slightly but continuously chlorinated before feeding to the pipeline system with a dose of 0.8-1.2 mg∙L⁻¹ of chlorine. With the exception of Siedlce, the concentration of C₂Cl₄ and CH₂Cl₄ represented by angle and length of eyebrows is not quite a good feature to classify drinking water samples due to the concentrations of CH₂Cl₂ and C₂Cl₄ not exceeding a value of 1 μg∙L⁻¹. Siedlce district is supplied by Straszyn lake after water treatment in the treatment plant and thus the total concentration of chlorinated disinfection by-products is the highest. In this case the concentrations of CH₂Cl₂ and C₂Cl₄ were classified as playing a minimal role as the factor used to assess the similarity/dissimilarity between districts. The comparison of the look of faces with linear distance from Straszyn lake allows the assumption that concentrations of residual chlorine and chlorination by-products decrease as water flows through the pipe line system and, particularly in large utilities, may become very low and even undetectable at the extremities. This is why we explain the occurrence of the concentration of CHCl₃ being equal to 9.29 μg∙L⁻¹ visualized by the curvature of the mouth in Niedźwiednik (located in the most northerly direction) as a reason for secondary chlorination in indirect drinking water treatment station. Faces marked

![Chernoff's Faces visualization approach](image-url)
with a hash are characterized by “sad” looks, short and narrow noses. This appearance corresponds to the drinking water samples characterized by low concentrations of CHCl₃ and CHBrCl₂+CH₃Cl and CHBrCl. This is due to the origin of water coming from underground sources and, additionally, disinfection by sodium hypochlorite (NaOCl) limited only to the cases of biological contamination of water. Faces not marked represent the samples characterized by low concentrations of CHCl₃, but a long nose in the case of Żabianka and Wrzeszcz-GUT indicates a middle range concentration of CHBrCl₂+CH₃Cl. These districts are located along the main highway and due to a long distance from the underground sources are supplied partially by the surface water, or by mixed water but the mixing ratio is limited by the occurrence of peak demand.

Conclusions

This study covered long-term monitoring of concentrations of major VOCs (THMs, DBPs) in drinking water in a metropolitan area (Gdańsk, Poland). VOC concentration monitoring programs generate multidimensional data that in some cases requires sophisticated techniques for their visualization and interpretation. In this case Chernoff’s visualization approach helped to group the twelve examined sampling sites into two main groups of similar characteristics pertaining to concentrations of THMs in water resulting from the kind of water system supply (surface water and underground sources). The analysis of facial characteristics allowed finding the relationship between THM concentrations and the geographical location of the districts. Chernoff’s visualization approach is an example of a simple but powerful visualization technique that successfully create graphical representation of data. Districts characterized by smiles and long noses corresponded to higher concentrations of CHCl₃ and CHBrCl₂+CH₃Cl, while those characterized by “sad” looks (short and narrow noses), are characteristic of drinking water samples with low concentrations of CHCl₃ and CHBrCl₂+CH₃Cl and CHBrCl.

References

21. TURGEON S., RODRIGUES M.J., THERIAULT M., LEVALLOIS P., Perception of drinking water in Quebec City region (Canada): the influence of water quality and consumer
