Original Research

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

A. Astel[1]*, K. Astel[2], M. Biziuk[2], J. Namieśnik[2]

[1]Environmental Chemistry Research Unit, Biology and Environmental Protection Institute, Pomeranian Pedagogical Academy, 22a Arciszewskiego Str., 76-200 Słupsk, Poland
[2]Analytical 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 containig 1756 results of determination of disinfection by-products (THMs: $CHCl_3$, $CHBrCl_2$+C_2HCl_3, $CHBr_2Cl$, $CHBr_3$ and organohalogen compounds: CCl_4, CH_2Cl_2, C_2Cl_4, $C_2H_3Cl_3$) 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

Introduction

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 keeps 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

*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 interpretating 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

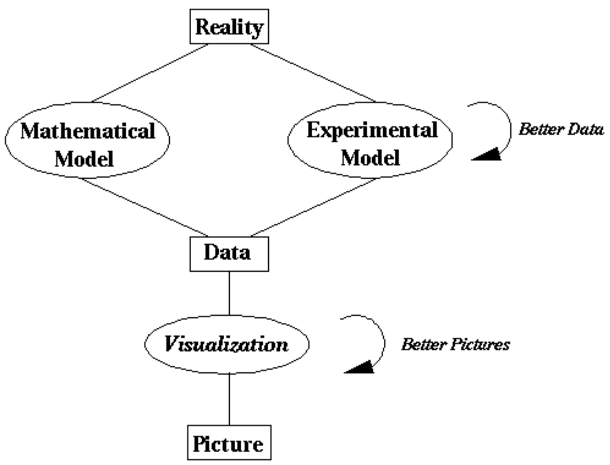

Fig. 1. The way of creating visual shape of reality.

(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]. VOCls 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 Voivodship 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, Niedźwiednik, 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 gravellers 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 $mg \cdot L^{-1}$ of chlorine to obtain a residual chlorine concentration at the level of 0.2-0.5 $mg \cdot L^{-1}$. 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_2HCl_3, $CHBr_2Cl$, $CHBr_3$) and organohalogen compounds (CCl_4, CH_2Cl_2, C_2Cl_4, $C_2H_3Cl_3$). Direct aqueous injection (DAI) into a capillary column of a Carlo Erba (Italy) Vega 6180 GC

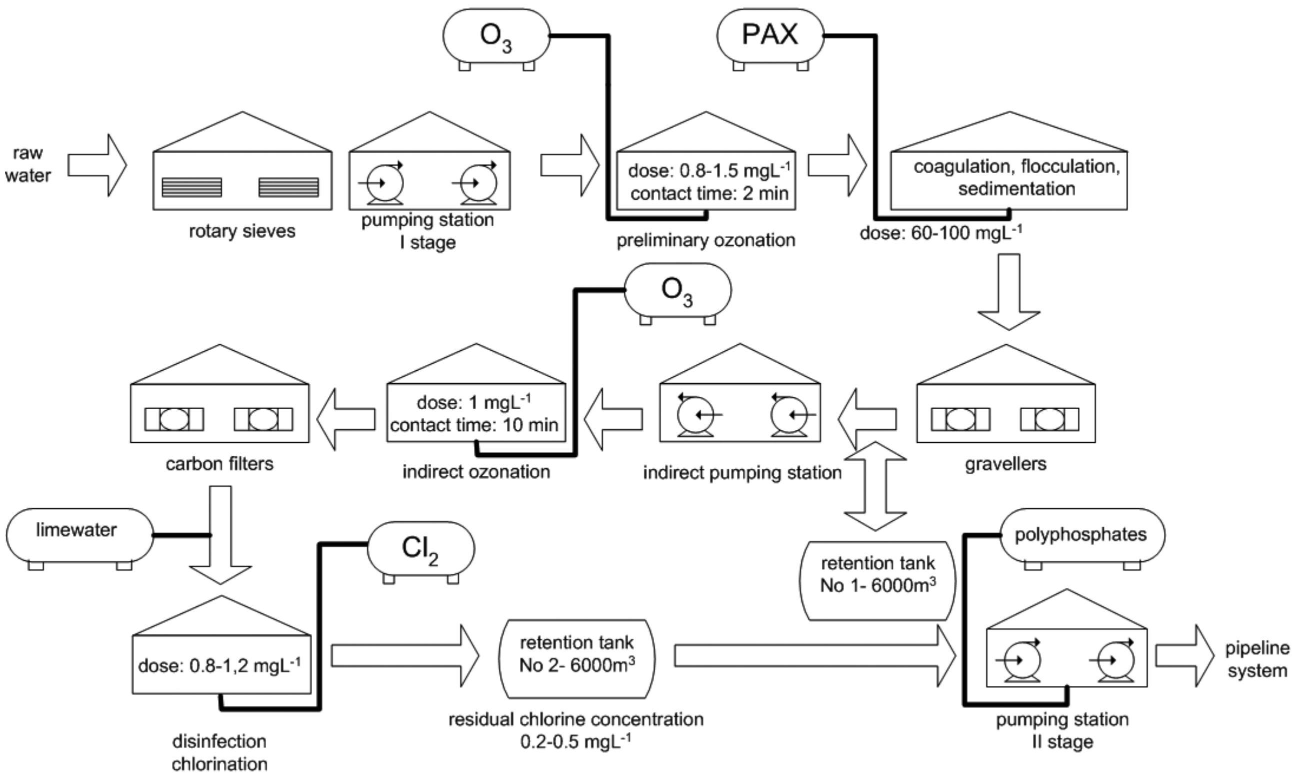

Fig. 2. The block diagram of the process line in Straszyn treatment plant.

system equipped with electron capture detection (ECD 40/400) was used for the determination of VOCls [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 $ml \cdot min^{-1}$); carrier gas, hydrogen at 0.4 $m \cdot s^{-1}$; injection volume, 2 µl. The detection limits of the DAI-ECD method, which were dependent on the species being determined, were approximately 0.01 $\mu g \cdot L^{-1}$ on average. A detailed description of the experimental procedure, including calibration and validation, can be found elsewhere [15-17].

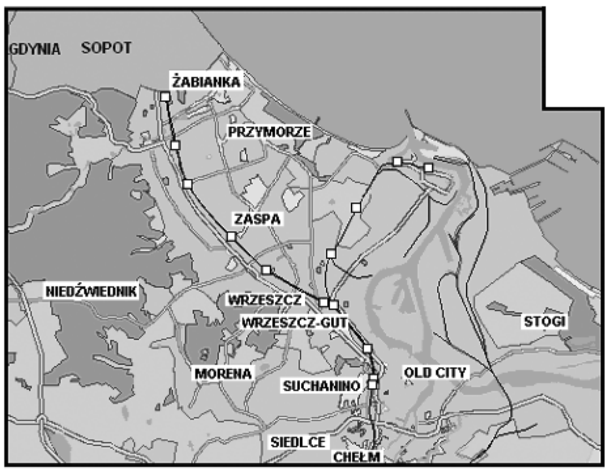

Fig. 3. Location of sampling sites.

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 organohalogen 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_2HCl_3, $CHBr_2Cl$, CH_2Cl_2, C_2Cl_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 $\mu g \cdot L^{-1}$ 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_2HCl_3 and $CHBr_2Cl$. 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 1. Results of determination of volatile organohalogen compounds ($\mu g \cdot L^{-1}$) in drinking water samples collected in various districts of Gdańsk.

Compound	n.o.	Mean value	n.o.	Mean value	n.o.	Mean value	n.o.	Mean value	n.o.	Mean value	n.o.	Mean value
	1993		1994		1995		1996		1999		2000	
$CHCl_3$	78	3.260	83	5.493	50	4.506	56	4.424	37	3.325	34	3.114
$CHBrCl_2$+C_2HCl_3	78	2.371	84	2.997	50	2.404	56	2.614	37	1.537	34	0.773
$CHBr_2Cl$	78	0.465	84	0.716	50	0.673	56	1.078	37	0.506	34	0.202
THMs	78	6.145	84	9.414	50	7.378	56	8.105	37	5.527	34	4.135
CH_2Cl_2	-	-	70	0.239	50	0.232	56	0.859	37	0.065	34	0.220
C_2Cl_4	-	-	70	0.168	50	0.090	-	-	-	-	34	0.074

n.o. – number of observations

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

District	$CHCl_3$	$CHBrCl_2$+C_2HCl_3	$CHBr_2Cl$	THMs	CH_2Cl_2	C_2Cl_4
Chełm	6.675	3.031	1.283	11.114	0.551	0.016
Morena	6.966	3.373	1.625	12.090	0.752	0.001
Suchanino	7.946	3.937	1.301	13.586	0.464	0.044
Niedźwiednik	9.291	4.253	1.055	14.715	0.747	0.020
Przymorze	0.990	0.859	0.193	2.042	0.021	0.140
Old Town	0.808	0.500	0.278	1.600	0.069	0.040
Zaspa	0.899	0.479	0.179	1.557	0.086	0.063
Żabianka	0.954	4.170	0.048	4.678	0.116	0.141
Wrzeszcz-GUT	3.163	1.713	0.373	5.287	0.264	0.088
Wrzeszcz	4.205	2.052	0.827	7.076	0.339	0.052
Stogi	1.299	0.003	0.003	1.295	0.140	0.090
Siedlce	10.819	5.911	2.069	18.800	1.474	0.265

in most drinking water samples, except for $CHBr_3$. In 1993, tetrachloroethane and dichloroethane, and in 1996 and 1999 tetrachloroethane were not detected at any of the examined locations. CCl_4, $C_2H_3Cl_3$ and $CHBr_3$ 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 in districts supplied by drinking water from surface source is 12.87 $\mu g \cdot L^{-1}$, while from underground sources it is 3.03 $\mu g \cdot L^{-1}$. In the case of the rest of organohalogen compounds (CH_2Cl_2 + C_2Cl_4) the mean value in districts supplied by drinking water from the surface source is 0.64 $\mu g \cdot L^{-1}$ and from underground sources 0.21 $\mu g \cdot L^{-1}$, 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 $CHCl_3$ and $CHBrCl_2$+C_2HCl_3. The wide nose for this category is connected with higher concentrations of $CHBr_2Cl$. 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 $mg \cdot L^{-1}$ of chlorine. With the exception of Siedlce, the concentration of C_2Cl_4 and CH_2Cl_2 represented by angle and length of eyebrows is not quite a good feature to classify drinking water samples due to the concentrations of CH_2Cl_2 and C_2Cl_4 not exceeding a value of 1 $\mu g \cdot L^{-1}$. 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 CH_2Cl_2 and C_2Cl_4 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 $CHCl_3$ being equal to 9.29 $\mu g \cdot L^{-1}$ 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

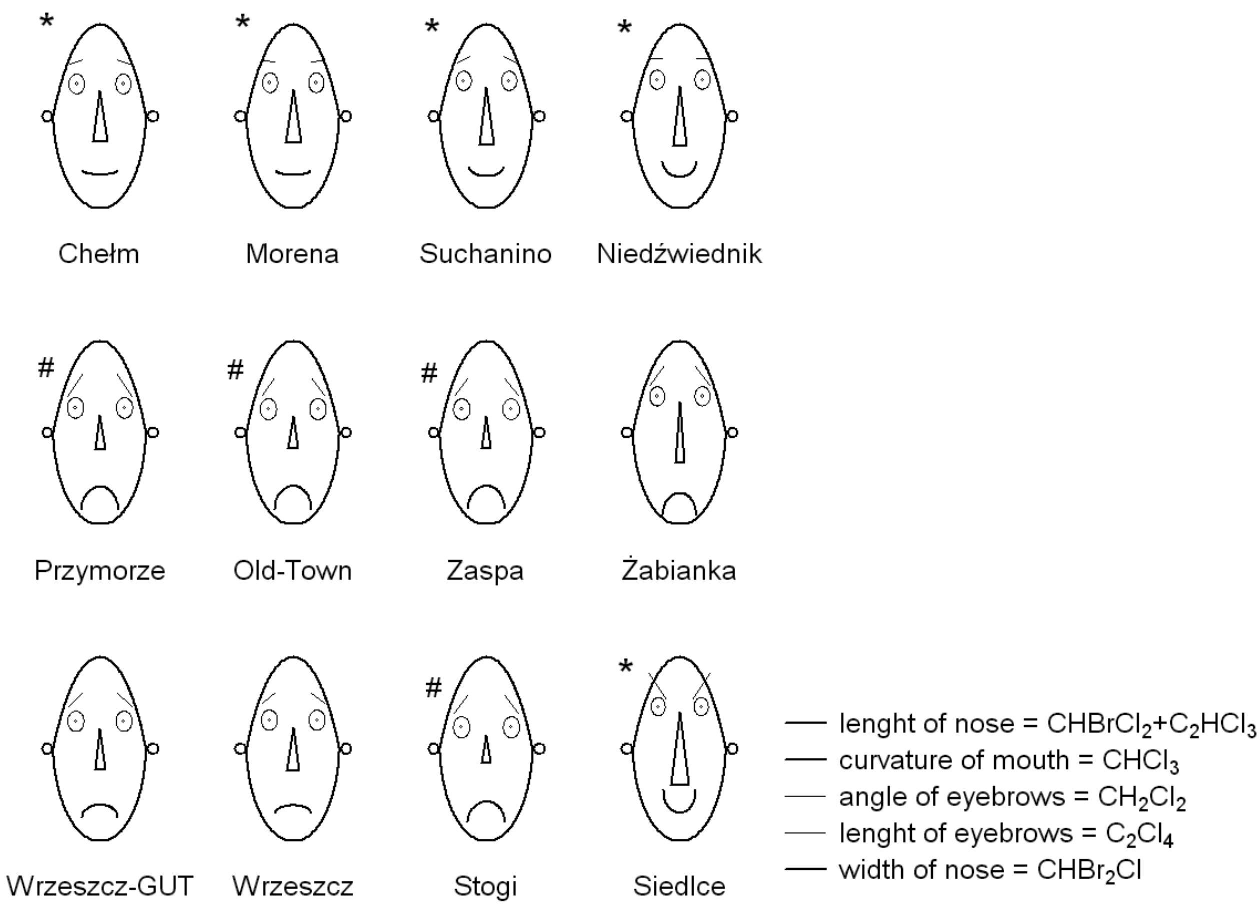

Fig. 4. Classification of drinking water samples in accordance to location of divisions by Chernoff's Faces visualization approach.

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 $CHCl_3$ and $CHBrCl_2$+C_2HCl_3 and $CHBr_2Cl$. 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 $CHCl_3$, but a long nose in the case of Żabianka and Wrzeszcz-GUT indicates a middle range concentration of $CHBrCl_2$+C_2HCl_3. 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 concentration of major VOCls (THMs, DBPs) in drinking water in a metropolitan area (Gdańsk, Poland). VOCl 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 $CHCl_3$ and $CHBrCl_2$+C_2HCl_3, while those characterized by "sad" looks (short and narrow noses), are characteristic of drinking water samples with low concentrations of $CHCl_3$ and $CHBrCl_2$+C_2HCl_3 and $CHBr_2Cl$.

References

1. MAZERSKI J., Fundamentals of chemometry, GUT Press, Gdańsk, Poland, pp. 200, **2000**.
2. MEINRATH G., LIS S., BUT S., ELBANOWSKI M., Chemometric and Statistical Analysis of Polyoxometalate Interaction with Lanthanide(III) Ions, Talanta, **55,** 371, **2001**.
3. HERNGREN L., GOONETILLEKE A., AYOKO G. A., Understanding heavy metal and suspended solids relationships in urban stormwater using simulated rainfall, J. Environ. Manage., **76**, 149, **2005**.
4. BROŻEK-MUCHA Z., ZADORA G., Grouping of ammunition types by means of frequencies of occurrence of GSR, Forensic Science International, **135 (2)**, 97, **2003**.
5. LOPES J.A., MENEZES J.C., Industrial fermentation end-product modeling with multilinear PLS, Chemom. Intell. Lab. Syst., **68**, 75, **2003**.
6. WORKMAN J. JR, The state of multivariate thinking for scientists in industry: 1980-2000. Chemom. Intell. Lab. Syst., **60**, 13, **2002**.
7. http://www2.sas.com/proceedings/sugi26/p195-26.pdf
8. SIMPSON K.L., HAYES K.P., Drinking water disinfection by-products: an Australian perspective, Wat. Res., **32 (5)**, 1522, **1998**.
9. KAMPIOTI A.A., STEPHANOU E.G., The impact of bromide on the formation of neutral and acidic disinfection by-products (DBPs) in Mediterranean chlorinated drinking water, Wat. Res., **36 (10)**, 2596, **2002**.
10. NOBUKAWA T., SATOSHI S., Effect of bromide ions on genotoxicity of halogenated by-products from chlorination of humic acid in water, Wat.Res., **35 (18)**, 4293, **2001**.
11. NIKOLAU A.D., LEKKAS T.D., GOLFINOPOULOS S.K., KOSTOPOLOU M.N., Application of different analytical methods for determination of volatile chlorination by-products in drinking water, Talanta, **56**, 717, **2002**.
12. CLARK R.M., GOODRICH J.A., DEININGER R.A., Drinking water and cancer mortality, Sci.Total. Environ., **53**, 153, **1986.**
13. MORRIS R.D., AUDET A.-M., ANGELILLO I.F., CHALMERS T., MUSTELLER F., Chlorination, chlorination by-products, and cancer: a meta-analysis, Amer. J., Public Health, **82**, 955, **1992**.
14. POLKOWSKA Ż., KOZŁOWSKA K., MAZERSKA Z., GÓRECKI T., NAMIEŚNIK J., Relationship between volatile organohalogen compounds in drinking water and human urine in Poland, Chemosphere, **53 (8)**, 899, **2003**.
15. BIZIUK M., NAMIEŚNIK J., CZERWIŃSKI J., GORLO D., MAKUCH B., JANICKI W., POLKOWSKA Ż., WOLSKA L., Occurence and determination of organic pollutants in tap and surface waters of the Gdańsk district, J.Chromatogr. A., **733**, 171, **1996.**
16. BIZIUK, M., POLKOWSKA, Ż., GORLO, D., JANICKI, W., NAMIEŚNIK, J., Determination of Volatile Organic Compounds in Water Intakes and Tap Water by Purge and Trap and Direct Aqueous Injection-Electron Capture Detection Techniques, Chem. Anal., **40**, 299, **1995**,
17. BIZUK, M., POLKOWSKA, Ż., Determination of trihalomethanes in Gdańsk water pipe system, Pollut. Environ., **3**, 12, **1993**.
18. CHERNOFF H., The use of faces to represent points in k-dimensional graphically, J. Amer. Stat., **68**, 361, **1973**.
19. ASTEL A., MAZERSKI J., POLKOWSKA Ż., NAMIEŚNIK J., Application of PCA and time series analysis in studies of precipitation in Tricity (Poland), Adv. Environ. Res., **8**, 337, **2004.**
20. StatSoft, Inc. (2001). STATISTICA (data analysis software system), version 6. www.statsoft.com;
21. TURGEON S., RODRIGUES M.J., THERIAULT M., LEVALLOIS P., Perception of drinking water in Quebec City region (Canad): the influence of water quality and consumer

location in the distribution system, J. Environ. Manag., **70**, 363, **2004**.
22. WILLIAMS D.T., LEBEL G.L., BENOIT F.M., Disinfection by-productes in canadian drinking water, Chemosphere, **34 (2)**, 299, **1997**.
23. PALACIOS M., PAPILLON J.F., RODRIGUEZ M.E., Organohalogenated compounds levels in chlorinated drinking waters and current compliance with quality standards throughout the European Union, Wat. Res., **34 (3)**, 1002, **2000**.