UNIVERSITAT DE BARCELONA

Application of chemometric methods to water quality studies

Stefan Yordanov Platikanov

Application of chemometric methods to water quality studies

Stefan Yordanov Platikanov

Barcelona, 2016

University of Barcelona
Faculty of Chemistry
Analytical Chemistry Department

The Spanish National Research Council
Institute of Environmental Assessment and
Water Research (IDÆA)
Department of Environmental Chemistry

Application of chemometric methods to water quality studies

A thesis written by Stefan Yordanov Platikanov under the supervision of Professor Dr. Romà Tauler Ferré from IDAEA-CSIC and under tutoring of Professor Dr. María Teresa Galcerán from the Department of Analytical Chemistry, Faculty of Chemistry, University of Barcelona in fulfillment of the thesis requirement for the degree of PhD in Analytical Chemistry and Environment doctoral program.

Doctoral Program: Analytical Chemistry and Environment

Barcelona, June 2016

PhD student:

Stefan Yordanov Platikanov

Director of the thesis:

Dr. Romà Tauler Ferré
Environmental Chemistry Department,
Institute of Environmental Assessment and Water Research
CSIC

Tutor:

Dr. María Teresa Galcerán
Department of Analytical Chemistry,
Faculty of Chemistry, University of Barcelona

“*So much universe, and so little time*”

Terry Pratchett

Acknowledgements

I would like to express my gratitude and appreciation to Dr. Romà Tauler for his permanent support, sharing knowledge, helpful discussions and guidance throughout the course of this work, and in allowing me the opportunity to work on these projects. Outside work, I have been privileged to spend these years with someone of the most charismatic, talented and generous people that I have ever met and his contribution to my personality is immeasurable.

I would like to acknowledge the funding from the *Agency for Management of University and Research Grants*, Catalonia (AGAUR) and to the *Ministry of Economy and Competitiveness*, Spain during the period of my study.

My thanks also go for my university tutor Dr. Maria Teresa Galcerán, Dr. Anna de Juan and Dr. Silvia Lacorte for their efforts and permanent helps in resolving all administrative and academic tasks.

I am very thankful for having the opportunity to work together with Dr. Lleonart Matia, Dr. Ricard Devesa and Dr. Jordi Martín-Alosno from the *Aigües de Barcelona; with* Dr. Montserrat Batlle and Jordi Cros from the *ADASA S.A.*; with Dr. Jose Luis Cortina from the *UPC*. They have provided me with never-ending support, knowledge and encouragement to achieve my objectives.

My very special thanks to my wonderful friends Dr. Joaquim Jaumot, Dr. Antonio Checa, Dr Sílvia Mas, Dr. Marta Terrado and Dr. Mireia Farrés for their endless support, inspiring me with incredible amount of energy and happiness during these years.

I would like to take this opportunity to express my sincere thanks and appreciation to Dr. Ramon López-Roldán, Dr. Sara Rodríguez, Dr. Susana González and Dr. Nicolas de Arespacochaga and many colleagues and friends from *AGBAR*, *CETaqua*, *ICRA* and *IDAEA-CSIC* for their practical support in these projects and for their friendship outside work.

Finally, I acknowledge my family for their continual support – moral, financial and practical, during the course of this Thesis and in all other challenges in life.

Content

Abbreviations

AGBAR	Aiguas de Barcelona Company
AMS-LED	Automatic Monitoring Station with optical LED probe
ANN	Artificial Neural Networks
ANOVA	Analysis Of Variance
BB	Box-Behnken response surface design
BMA	Barcelona Metropolitan Area
BOD_5	Biological Oxygen Demand in 5 days
COD	Chemical Oxygen Demand
DBPs	Disinfection Bi-Products
DO	Dissolved Oxygen
DOC	Dissolved Organic Carbon
DoE	Design of Experiments
DOM	Dissolved Organic Matter
EDR	Electrodialysis Reversal
FPA	Flavour Profile Analysis
FTIR	Fourier Transform Infrared Spectroscopy
GAC	Granular Activated Carbon
GC-ECD	Gas Chromatograph with Electron Capture Detector
GC-MS	Gas Chromatography Mass Spectrometry
HAAs	HaloAcetic Acids
HPOF	Hydrophobic fraction of NOM
ICP-OES	Inductively Coupled Plasma Optical Emission Spectrometer
K-PLS	Kernel radial basis function Partial Least Squares regression
LC-MS/MS	Liquid Chromatography tandem Mass Spectrometry
LC-TOF/MS	Liquid Chromatography Time-of-Flight Mass Spectrometry
LIMS	Laboratory Information Management System
LOD	Limit of Detection
LVs	Latent Variables
MCAR	Missing values occurred Completely At Random
MLR	Multiple Linear Regression
NMAR	Missing values do not occur at random
NN	Neural Networks

NOM	Natural Organic Matter
OVAT	One single Variable At a Time
PB	Plackett–Burman screening design
PCA	Principal Component Analysis
PCR	Principal Component Regression
PCs	Principal Components
PLS	Partial Least Squares regression
RBF	kernel Radial Basis Functions
RMSEC	Root Mean Square Error in Calibration
RMSECV	Root Mean Square Error in Cross Validation
RMSEP	Root Mean Square Error in external validation Prediction
RO	Reverse Osmosis
RSM	Response Surface Methodology
SJD-DWTP	Sant Joan Despí Drinking Water Treatment Plant
SUVA	Specific UltraViolet Absorbance coefficient
SVD	Singular Value Decomposition
SVM	Support Vector Machine
SVR	Support Vector machine Regression
SWR	Stepwise Multilinear Regression
SWRO	Sea Water Reverse Osmosis DWTP at Llobregat village
TDS	Total Dissolved Solids
TOC	Total Organic Carbon
TSS	Total Suspended Solids
THMs	Trihalomethanes
TTHMs	Total sum of Trihalomethanes
USEPA	United States Environmental Protection Agency
UV-VIS	Ultraviolet-visible spectroscopy
VIP	Variable Importance of Projection
WDS	Water Distribution System
WFD	Water Framework Directive
WTP1	Cardedeu DWTP
WTP2	SJD-DWTP
WTP3-4	Abrera DWTP
WWTP	Waste Water Treatment Plant

List of Figures

List of Tables

Abstract

Several chemometric methods have been used to explore, analyze and interpret information regarding potable and waste water quality in this Thesis. The analyzed data were obtained from (a) drinking water disinfection processes, (b) wastewater treatment processes, (c) sensorial analysis comprising panelists' water taste evaluations, and (d) laboratory experiments.

This Thesis attempts to improve our knowledge regarding several common water quality problems, such as the formation of trihalomethanes (THMs) disinfection by-products (in the Sant Joan Despi Drinking Water Treatment Plant of Barcelona, SJD-DWTP) and the main factors affecting their formation. Furthermore, the Thesis illustrates how to facilitate the monitoring of water quality in a Wastewater Treatment Plant of Girona town (WWTP) by applying chemometric methods. Further objectives of the Thesis include the development of a chemometric method for source apportionment, where drinking waters with different origins were blended (as it usually occurs inside the Barcelona drinking Water Distribution System, WDS) using measured ultraviolet absorbance and physicochemical parameters. This Thesis additionally considered the problem of water taste by developing models, where water taste is explained and predicted based on the mineral content of tap and bottled waters using trained panelists.

The chemometrics methods, applied in this Thesis, have been applied to multi-parametric data matrices generated using different instrumental analyses techniques, such as laboratory UVVIS spectrophotometer, Gas Chromatograph with Electron Capture Detector (GC-ECD) and Inductively Coupled Plasma Optical Emission Spectrometer (ICP-OES). Additionally, data was obtained by implementation of standard methods for estimation of different physicochemical parameters or by multi-parametric data extractions from the Laboratory Information Management System (LIMS). Data was also acquired from an automatic multi-parametric station for online monitoring and from carefully designed sensorial experiments.

In this Thesis, different linear projection based methods, such as Principal Component Analysis (PCA), Principal Component Regression (PCR) and Partial Least Squares regression method (PLS), have been used and shown as appropriate for handling data. Different linear regression methods have been compared to powerful

nonlinear regression methods such as Kernel radial basis function Partial Least Squares (K-PLS) and Support vector machine regression (SVR) methods.

Among the most significant findings of this Thesis was the identification of a set of parameters, which are highly relevant for the trihalomethanes formation, such as water temperature, organic matter fractions and concentration, chlorine concentrations, turbidity, bromide/chloride ions concentrations, wells supply flow levels and carbon filters age. Chemometric models, with very low prediction errors for all four THMs species and their total sum, have been developed at SJD-DWTP. The most important physicochemical parameters for panellist water taste liking were found to be: HCO_3^-, SO_4^{2-}, Ca^{2+}, and Mg^{2+} at moderate concentration of the overall mineralization and pH. Temporal variation with a different data frequency (daily, monthly and annual cycles) were observed in WWTP water quality and suggested different plant management and operational procedures. A chemometric model was developed to predict source apportionment inside the Barcelona WDS. Five different water sources were detected in water blends.

Finally, different chemometric techniques for visualization and data interpretation have been tested and evaluated for their usefulness for water quality analyses. The prediction ability of linear or nonlinear regression methods have been compared when they were used to develop empirical models and predict water quality parameters such as THMs concentrations in drinking waters, nitrates, phenols, organic matter in wastewaters, water source apportionments in water distribution system and panelists taste ratings of water samples.

Last but not least, this Thesis had as an objective to demonstrate the advantages of using chemometric methods for water quality assessment. This work showed that complex problems can be resolved using a multivariate modeling methodology with a few experiments. Therefore, this methodology is a significant improvement over unvaried approaches which are based on expensive and time-consuming measurements.

Chapter 1

Objectives and Structure of the Thesis

1.1 General objectives

Different data sets obtained from different real water quality case studies at drinking and waste water treatments plants in Catalonia were analyzed using multivariate chemometrics methods.

The main objective of this Thesis was the development, application and promotion of practical chemometric methods for analysis and interpretation of data concerning sanitary and esthetic water quality problems in producing potable water and monitoring of wastewater quality.

This Thesis includes six different studies, more specifically five studies of potable water management and one study of wastewater management.

This main objective was divided as follows:

Objectives regarding water quality analysis and monitoring

- Development of reliable empirical regression models for predicting trihalomethane formation inside the Sant Joan Despí drinking water treatment plant, applying different linear and nonlinear chemometric methods and using environmental and operational water quality plant parameters that describe the disinfection process;
- Review and interpretation of the most important DWTP parameters for the THMs formation using chemometric visualization techniques, which facilitate water quality monitoring and control;
- Comprehensive assessment of natural organic matter (NOM) role in THMs formation and THMs speciation in disinfection conditions from specially designed experiments;
- Development of chemometric regression models which are able to predict water source apportionments of water blends from up to five different water sources, using their UV spectral profiles and other physicochemical parameters;
- Using chemometrics methods, the identification of the most influential physicochemical water parameters which can be associated with the panellists' water taste liking of mineral bottled and tap waters;

- Evaluation of chemometrics methods which can improve online monitoring and control of wastewater treatment plant management using different techniques and routines for regular water quality monitoring;
- Selection of a reduced number of UV spectral channels (wavelengths) for monitoring online WWTP operational processes.

Objectives regarding the application of chemometrics methods

- Development, application and validation of linear and nonlinear chemometrics methods in analysing water quality data, using parameters measured in-situ or in laboratory from different water treatment processes;
- Comparing predictive abilities of linear and non-linear regression methods for the formation of THMs;
- Identification of the most efficient techniques and tools in visualizing and evaluating the more important variables (parameters) in linear models;
- Application of visualization techniques for nonlinear regression techniques (K-PLS and SVR) and their subsequent comparison to linear methods;
- Adaptation of experimental design techniques (DoE) with the objective to obtain representative and economic calibration data sets;
- Evaluation of "Variable Importance of Projection" pre-selection technique in PLS modelling of water quality data and selection of a reduced number of variables with a sufficient strong predictive power.

1.2 Structure of the Thesis

This Thesis is divided into two main parts. The first introductory part contains a general description of common water quality problems such as the formation of THMs, and water taste esthetic issues in drinking water, as well as the presentation of existing methods for improving of wastewater quality monitoring. This part also presents a brief introduction of the applied chemometrics methods used in this Thesis. The second part contains the empirical results of performed studies. Published scientific articles, along with a background introduction and a short discussion of the obtained results, are

included. This second part concludes with the literature references cited throughout the Thesis.

The Thesis consists of the following six chapters:

- In Chapter 1, the objectives of the Thesis are presented, followed by the explanation of the Thesis structure. In this part, the list of publications is also included.
- In Chapter 2, the state of the art of the main water quality problems and chemometric methods applied in this Thesis are briefly outlined. This chapter contains the following blocks. First, previous research regarding THMs formation and the European legislation concerning regulated levels of THMs, and the importance of natural organic matter in THMs formation are presented, along with selected epidemiological studies regarding THMs sanitary risk. Second, the drinking water sources and WDS of Barcelona are discussed. Examples of a classical DWTP and of a WWTP treatments plants are provided. The main water quality parameters, regularly monitored in such facilities, are presented. A brief discussion of existing automatic systems based on UV sensors for online water quality monitoring is included. Third, relevant esthetical aspects regarding water taste are discussed. This chapter concludes with a review of the applied chemometrics methods used in this Thesis.
- In Chapter 3, the obtained results from the case studies, included in this Thesis, are discussed. This chapter is divided into three blocks as follows:
    1. In the first block, three articles that deal with chemometric modeling of the THMs formation from laboratory experiments and from a real monitoring of water quality in SJD-DWTP are included;
    2. In the second block, two articles that deal with the chemometrics study of drinking and waste water quality monitoring are presented. In the first article, a drinking water source apportionment study of the main water supply sources from Barcelona was performed. The second article presents a study that focuses on the chemometric modeling of wastewater quality

parameters using standard laboratory techniques and from on-line automatic monitoring system;

3. In the third block, the results from an article using a sensorial analysis of different bottled mineral and tap water samples are presented. The main objective of this study was to associate physicochemical parameters and ratings of trained panelists.

- In Chapter 4, the conclusions of this Thesis are presented.
- In Chapter 5, a brief summery in Spanish language is given.
- The Thesis finishes with the reference list.

1.3. List of scientific papers presented in this Thesis

1. **Article 1** – Platikanov, S., Puig, X., Martin, J. and R. Tauler. *Chemometric modeling and prediction of trihalomethane formation in Barcelona's water works plant.* **Water Research** 41 (2007) 3394-3406.

2. **Article 2** – Platikanov, S., Martin, J. and R. Tauler. *Linear and non-linear chemometric modeling of THM formation in Barcelona's water treatment plant.* **Science of Total Environment** 432 (2012) 365-374.

3. **Article 3** – Platikanov, S., Tauler, R., Rodriguez, P., Antunes, M., Pereira, D. and J. Esteves da Silva. *Factorial Analysis of the trihalomethane formation in the reaction of colloidal, hydrophobic and transphilic fractions of DOM with free chlorine.* **Environmental Science and Pollution Research** 17 (2010) 1389-1400.

4. **Article 4** – Platikanov, S., Garcia, V., Landeros, E., Devesa, R., Matía, L., Tauler, R., *Determination of water supply sources in the Barcelona distribution system by UV spectrophotometry and PLS.* **Water Science and Technology-Water Supply** 11 (2011) 45-54.

5. **Article 5** – Platikanov, S., Rodriguez-Mozaz, S., Huerta, B., Barcelo, D., Cros, J., Battle, M., Poch, G., Tauler, R. *Chemometrics quality assessment of wastewater treatment plant effluents using physicochemical parameters and UV absorption measurements.* **Journal of Environmental Management** 140 (2014) 33-44.

6. **Article 6** – Platikanov, S., Garcia, V., Fonseca, I., Rullan, E., Devesa, R., Tauler, R., *Influence of minerals on the taste of bottled and tap water: A chemometric approach.* **Water Research** 47 (2013) 693-704.

Chapter 2

Introduction

2.1 Literature review of the main water quality problems studied in this Thesis

2.1.1 Urban water cycle

The Barcelona Metropolitan Area (BMA) is characterized (see Figure 1) by its multi-source urban water cycle (Marín et al., 2012). The resources of raw water used for potabilization include surface water (i.e., the Ter River), brackish water (i.e., the Llobregat River), underground water (i.e., the Besòs and the Llobregat aquifers), and seawater (i.e., the Mediterranean Sea). There is a large variability in water quality among such sources, which have different water stressors' characteristics.

To remove contaminants and pathogens, raw water must be first treated prior to its distribution and use as potable water. The design of an appropriate treatment process is driven by the water quality of different sources. Today, drinking water treatment facilities in the BMA employ various treatment technologies such as: (a) a classical disinfection with chlorine, ozonation, granular activated carbon (GAC) filtration; (b) ultra-filtration; (c) brackish and seawater reverse osmosis (RO); (d) electrodialysis reversal (EDR) desalinization, and (e) water remineralization.

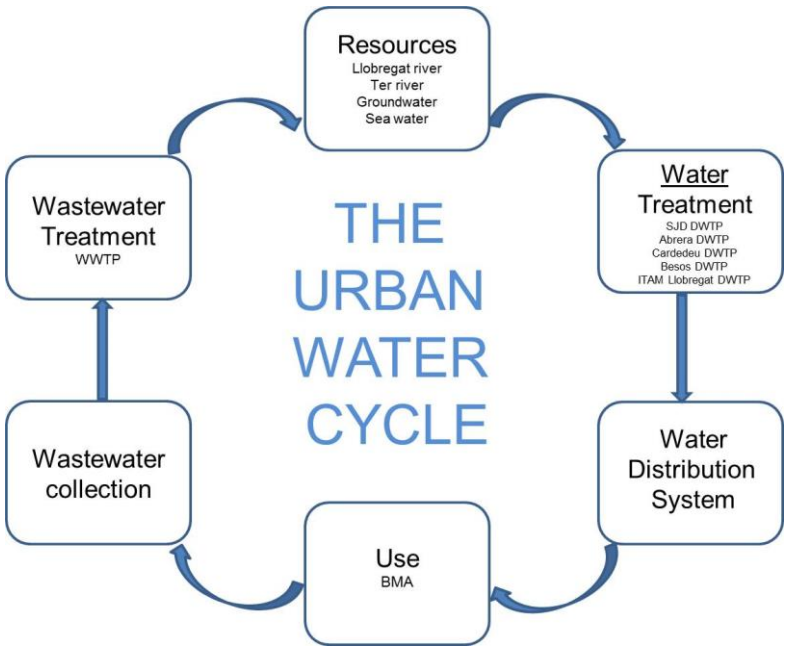


Figure 1. The urban water cycle for the Barcelona Metropolitan Area

After disinfection treatment, water is distributed to customers through a pressurized system of pipes, pumps, valves, and storage tanks - parts of the Water Distribution System (WDS) of BMA. The water distribution represents the second stage in the urban water cycle. The supplied potable water is used for various purposes, including industrial manufacturing, cleaning, cooking, bathing, laundry, drinking. After consumption, used water is transferred to the sewers system for wastewater collecting. The wastewater is conveyed by gravity to a wastewater treatment facility, using a network of increasingly large pipes.

A typical urban wastewater stream contains more than 99% water and approximately 1% waste (Venugopala Rao, 2005). To the extent that water quality has been seriously deteriorated, a wastewater treatment is required before water could be released in natural habitats, such as rivers or the Mediterranean Sea. As a result, wastewater treatment plants play an important role in the urban water cycle by implementing various physical, chemical and biological processes designed to remove wastes from the influent and restore water quality. Conditional on being effectively treated, the water effluent is returned into the rivers for example.

Although each component from the urban water cycle is designed such that to generate beneficial social and economic outcomes, it is a challenge to efficiently manage the process. Among the main difficulties to make the cycle functioning are the appropriate maintenance of all water treatment processes, and the enforcement of efficient water quality monitoring and control systems. In order to be able to develop sustainable water solutions, we first need to understand these challenges.

2.1.2 Drinking water quality

The access to drinking water is among the main drivers of the human development around the world. The general understanding is that the drinking water should be clear, pleasant to taste and without odors. The history of drinking water however is controversial, especially over the past centuries, when water actually has facilitated the spread of major diseases. The establishment of treatment procedures has influenced the attitude towards drinking water, leading to the predominant understandings nowadays

that drinking water treatments allow the consumption of water that is both microbiologically and chemically safe.

Aiguas de Barcelona (AGBAR) is responsible for the delivery of drinking water to more than 3 million residents (Paraira and West, 2015) of the Barcelona Metropolitan Area (BMA). Supplying drinking water to such a large population and simultaneously to every household in compliance with the strict regulatory norms is a major challenge from an operational management.

The list of main operational problems regarding water quality for the BMA includes:

- Pollution of raw river water with an industrial and agricultural character;
- Raw water scarcity primarily due to the over-exploitation of the main water sources, i.e., the two rivers nearby Barcelona;
- Formation of undesired disinfection bi-products (DBPs) in drinking water after disinfection process;
- Formation, distribution and behavior of these DBPs inside the water distribution network;
- Water taste improvement;

Among the main sources of river water pollution are (1) the concentration of industrialization and the growth rate of agricultural activities along the river basins, (2) financial and technological constraints, and (3) lack of–enforcement or bypassing of public laws. Furthermore, the weather conditions in Catalonia are often unfavorable from the perspective of DWTP management (Martin-Alonso, 2006). More specifically, scarce snowfalls are likely to cause long and severe droughts, whereas flushing spring/summer rainfalls may generate floods with an adverse effect on the water quality. Additionally, river water quality is negatively affected by accidental sewage discharges, industrial spills (for instance, mining activities in the middle part of the Llobregat River basin), and sources of pollution such as agricultural and urban run-off (Ginebreda et al., 2012). Although classical raw water disinfectants reduce the microbial risk, they have as a disadvantage the generation of disinfection by-products. Recent findings suggest that there are more than 250 different types of DBPs in drinking water (Sadiq and Rodriguez, 2004).

As a result, exhaustively and expensive physicochemical treatment procedures are required in order to overcome the strong fluctuations in the raw water quality and

the formation of undesired DBPs in drinking water. The existing practice in water companies is to implement advanced membranes technologies (Rahardianto et al., 2007; Valero et al., 2013; Wang et al., 2006) thus improving raw water quality and reducing DBPs formation. Such technologies however request an additional process of water remineralization (García et al., 2015), which improves the taste of drinking water.

The proper functioning of drinking water treatment and supply processes requires constant monitoring which is required to guarantee important aspects of the process, including hygiene, environment sanitation, storage and disposal, and the supply of high-quality drinking water to Barcelona.

2.1.3 Wastewater quality

An integrative part of the urban water cycle is the wastewater treatment and quality controls. To the extent that the demand for water sources is high, local authorities and business are interested in possible new water sources, conditional on being financially and socially acceptable in promoting efficiency and economic development (Hespanhol, 1997).

The ongoing trend towards increasing urban populations caused the accumulation of large quantities of wastewater at the municipal level. Because environmental and water pollution became a significant policy issue over the last decade, different parties in the water process increased their awareness of whether wastewaters are safely and economically beneficially disposed. Among the possible uses of wastewater is agriculture, especially when such disposals may compensate for the relatively scarce water resources in some areas (Al-Mutaz, 1989). The wastewater use in agriculture therefore may lead to the conservation of higher quality water and its efficient usage beyond irrigation. Although it has beneficial applications, wastewater treatment processes present a serious challenge in terms of implementation and control (Talaiekhozani et al., 2016). It is largely recognized that wastewater treatment processes are dynamic and complex, due to the variations of the quality in the municipal wastewater influent. The wastewater influent is characterized with a dynamic flow rate and a diverse chemical composition related to household and industrial activities, and urban runoffs during rain episodes. As a consequence, the municipal wastewater treated discharge can adversely affect the surface water quality of the rivers or the Mediterranean Sea (Amine et al., 2012). Therefore, the constant monitoring of

wastewater quality is highly relevant in managing more efficiently the urban water cycle, implementing environmental protection, and achieving sustainable growth.

2.1.4 European legislation

The EU environmental policy is focused on the water pollution and water resources deterioration. The first attempts of legislation were accepted by the European Council at 1973. In the last decades, the most important EU water directives had a focus on distinct issues and target specific problems:

- Directive concerning the quality required for surface water intended for the abstraction of drinking water in the Member States (75/440/EEC);
- Directive concerning the quality of bathing water (76/160/EEC);
- Directive on pollution caused by certain dangerous substances discharged into the aquatic environment of the Community (76/464/EEC);
- Directive relating to the quality of drinking water intended for human consumption (80/778/EEC prior to 2008 and 98/83/EC since then);
- Directive concerning urban waste water treatment (91/271/EEC).

There are a couple of European Union directives, which are directly related to the topic of this Thesis: a) Directive 98/83/EC on the quality of drinking water, and b) Directive 91/271/EEC concerning urban waste water treatment:

a) Directive 98/83/EC relating to the quality of drinking water intended for human consumption

Directive 98/83/EC has an objective to set new quality standards for many microbiological, chemical and organoleptic parameters, which must be implemented by the EU member countries in monitoring the quality of drinking water quality. The Directive 98/83/EC regulates a maximum concentration of total sum of THMs to 100 μg/l, which prior to the enactment of this Directive was 150 μg/l (Directive 80/778/EC). Spain adopted the European Directive in December 2003 (Real Decreto 140/2003), thus allowing for the total sum of THMs to be with a maximum concentration of 100 μg/l. The stated goal of this Directive is to protect consumers' health in the EU member countries by safety drinking water.

This Thesis includes studies on the regulated group of trihalomethanes (THMs) in Directive 98/83/EC. Three articles investigate THMs and their undesired formation, using chemometric modeling.

b) Directive 91/271/EEC concerning urban waste water treatment

Directive 91/271/EEC calls for a pollution reduction in surface waters, for example, by improving the quality of the discharged urban wastewaters. This directive calls for improvement of the technological processes inside the wastewater treatment plants (WWTP). An improvement of water quality monitoring process can contribute to the overall improvement of the technological process in WWTP.

In this Thesis, chemometric modeling of wastewater quality is proposed as a useful approach for monitoring. More specifically, water quality of the influent/effluent in an urban WWTP is monitored using chemometric methods with simultaneous modelling of various physicochemical parameters and pollutants. The application of multivariate analysis methods in the concept of continuous monitoring is proposed to facilitate and improve the control of WWTP operational processes, as well as to decrease the pollution effect specified in the Directive 91/271/EEC.

Besides the above mentioned directives, there are others which establish specific limits or emission standards for certain target compounds. Today, approximately 25 EU directives and decisions are directly or indirectly related to the management of the water resources. The large number of specific EU directives is a consequence of the lack of an integrated legal system at the level of the European Union. To address this problem of legislative multi-systems and the complexity of multilevel ecological systems of the water resources, the European Union adopted the Water Framework Directive (WFD 2000/60/EEC).

Today, the WFD is the most recent and comprehensive initiative of the European Union in the area of water protection. Its main goal was to overcome recurring problems in the integrative nature of water management for all EU members. The WFD represents a new fundamental approach of water legislation and includes many of the previously established regulation and specific directives (such as the above mentioned Directives directly related to the topic of the Thesis). WFD promotes for a new technical management control of river basins; endorses the protection and conservation of aquatic ecosystems by reducing the presence of long list of water pollutants. WDF also calls for

adequate technical systems for treatment of drinking and waste waters in order to meet new established standards.

The results and conclusions of this Thesis can contribute for the better understanding, efficient monitoring and control of water quality using chemometric modeling and analysis, thus facilitating the effective enforcement of the Water Framework Directive (2000/60/EEC).

2.1.5 Water resources and water treatment facilities for Barcelona Metropolitan Area (BMA)

AGBAR is one of the larger water suppliers, delivering drinking water to more than 3 million people of Barcelona through a large and complex water distribution system (WDS). The second largest water supply company operating in the BMA is Aigües Ter-Llobregat (ATLL).

There are several natural resources which are employed to supply water to the large Barcelona Metropolitan Area (BMA): (a) superficial water from two rivers (the Llobregat and the Ter), underground water from two aquifers (the Baix Llobregat and the Bessos), and the Mediterranean Sea.

Over the past years, the main source of potable water in Catalonia, Spain were two rivers (the Llobregat and the Ter). Largely driven by the growth in demand for urban consumption, the water supply became limited, especially in periods of serious natural droughts, such as in 2006 (Martin-Alonso, 2006). Both rivers have Mediterranean hydrological regimes characterized by irregular flows and seasonal fluctuations (i.e., from predominantly dry periods to sudden torrential flows).

The quality of potable water from the two rivers differs in terms of salinity. More specifically, the Llobregat River water is characterized by its higher salinity than the Ter River water. Also, the two rivers differ in water quality due to their different composition of the natural organic matter, largely explained by the different geomorphological, vegetation and industrial characteristics of their river basins.

The Llobregat River

The Llobregat River is a major water resource of the BMA. The river is 156.5 km long with a watershed of 4948 km^2. It covers around 40% of the total water demand

for Barcelona (Martin-Alonso et al., 2007). There are three dams along the river basin, located in the upper part, and many smaller contributors along its catchment area. In the upper and middle part of Llobregat basin, the mineral composition of water is strongly affected by the river passing through sedimentary rocks such as limestone, marls, gypsum, and halite. The most distinguished mineral composition includes: Ca^{2+}, Mg^{2+}, HCO_3^-, Sr^{2+}, Ba^{2+}, SO_4^{2-}, Na^+, K^+ and Cl^- (Marcé et al., 2012). The water mineral composition is also strongly enhanced by the potash mining activities in this area. Additionally, a very strong industrial area of Catalonia is situated in the Llobregat valley, where there are also extensive agricultural activities. The industrial area includes textile production, pharmaceutical industries and hydropower generation. There are two DWTPs (see Figure 2) which operate using the Llobregat river downstream water (the Abrera DWTP and the Sant Joan Despı DWTP).

The downstream raw Llobregat water is characterized by its low water quality, largely due to the high level of mineralization (dry residue 900 mg/l). The origin of such mineralization is the evaporate-bearing geological formations and the sodium and potassium chloride mining activities in the upper and middle parts of the river basin (Fernandez-Turuel et al., 2003). Several contamination episodes have been recorded and they were caused by industrial discharge of dicyclopentadiene and derivatives (Ventura et al., 1997), dioxanes/dioxolanes (Romero et al., 1998), and creosote (Ventura et al., 1998). Boleda et al. (2007) conducted a comprehensive review of detected taste and odor events in the BMA over the period 1990-2004 and suggested that industrial contamination generates more frequently contamination episodes in comparison to natural phenomena. Devesa et al. (2007) documented the contamination with diacetyl or butanedione and its further successive elimination. Both studies reviewed a large number of analytical chromatographic and spectral techniques, which were used in detecting possible sources of contamination produced by a large number of existing pollution compounds.

The Ter River

The Ter River is another major water resource of the BMA. The river is 208 km long with a watershed of 3010 km^2 and is situated in the North-East of Catalonia. Similarly to the Llovregat river, the Ter river has a Mediterranean-type regime. It is responsible for the supply of approximately 55% of the total water demand for Barcelona city (Céspedes et al., 2007). The river runs mostly over calcareous substratum (Sabater and

Armengol, 1986). In the upper part of the Ter river's basin, the water mineral composition charge is poor, largely because the river goes through thought siliceous bedrock. Although salinity and the organic matter are significantly higher in the medium and downstream part of the river basin, these levels are lower in comparison to the Llobregat river (dry residue around 400 mg/l). The most characteristic mineral constituents of the Ter water are Ca^{2+}, Mg^{2+}, HCO_3^-, Na^+ and K^+ (Sabater and Armengol, 1986). Urban and industrial pollution accidents usually occur in the middle and downstream part of the basin, where industry activities are more intensive (Sabater et al., 1990). The main pollution sources of Ter River's water quality are the intensity of agricultural activities in the area, as well as the presence of metallurgic, pulp mill, textile and tannery industries (Espadaler et al., 1997; Céspedes et al., 2006).

The quality of Ter water is positively affected by three reservoirs in the middle part of the basin. The Cardedeu DWTP (see Figure 2) receives the Ter raw water through a 60km-long pipeline directly from the reservoirs.

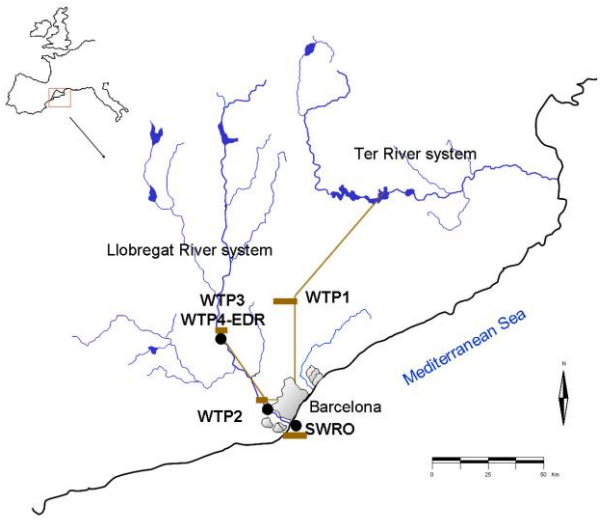


Figure 2. Drinking Water Treatment Plants supplying water for BMA. WTP1 refers to the Cardedeu DWTP; WTP2 – Sant Joan Despí DWTP; WTP3 and WTP4 – Abrera DWTP; SWRO- Llobregat Seawater Reverse Osmosis Desalination Plant.

The BMA is spread over 635 km2 with a population of approximately 4.5 million habitants (López-Roldán et al., 2015). Surface waters from the Ter and Llobregat rivers are used to supply around 85% of the total drinking water, whereas the remaining supply of approximately 15% is coming from groundwater. The serious drought in 2008 and the strong pressure stemming from the contamination, stimulated the adoption of new alternative water resources and the development of new infrastructures over the past years. Recently, a new modern DWTP supplying desalinated water from the Mediterranean Sea, using a reverse osmosis technology (SWRO) has been introduced. Furthermore, a new technology based on electrodialysis reversal (WTP4-EDR) has been installed in the Abrera DWTP. Some years ago, a reverse osmosis membrane technology has been also introduced in the SJD DWTP. At present, the two main DWTP installations include the following components:

- Abrera DWTP, has a new treatment line incorporating electrodialysis reversal; and
- Sant Joan Despí DWTP, has incorporated ultrafiltration and low pressure reverse osmosis.

Both DWTPs have implemented classical disinfection treatments with coagulants, flocculants, chlorine, chlorine dioxide, ozone and granulated carbon filtering, which are located in the lower-middle basin of the Llobregat River, and they supply approximately 40% of the drinking water to the BMA.

The other important DWTP installations of the BMA for water supply are:

- Cardedeu DWTP treating Ter River water with a classical process including coagulation, flocculation and activated carbon filtration;
- El Prat Seawater Reverse Osmosis Desalination Plant (SWRO);
- Numerous groundwater wells at the Llobregat and the Besòs Rivers basins (García et al., 2015).

The three DWTP in Cardedeu, Sant Joan Despí and Abrera (WTP1, WTP2, and WTP3 on Figure 2) are responsible for operating treatment in serving more than 95% of the potable water to the BMA. There are significant differences among the three DWTP in terms of treatment procedures, explaining the variation in drinking water quality supplied to the area.

Table 1. The main DWTP supplying water to BMA

Installation	Resource	Production capacity (hm^3)
DWTP Abrera	Llobregat River	126
DWTP Sant Joan Despí	Llobregat River Aquifer Llobregat	167
DWTP Cardedeu	Ter River	252
SWRO Llobregat	Mediterranean sea	60
DWTP Besòs	Aquifer Besòs	12

In continuation, the typical treatment processes at each facility are outlined (CETaqua and Chris Fife-Schaw, 2010).

a) Sant Joan Despí DWTP

The Sant Joan Despi DWTP (SJD-DWTP) is located nearby Sant Joan Despi, which is located in the outskirts of Barcelona city and is using the Llobregat River as a water source. The DWTP has a water production capacity of 5.3m3. Since its foundation in 1955, the DWTP has undertaken various reconstructions, re-modernizations and upgrades. The latest significant upgrade in the water treatment process of this plant was the implementation of ultrafiltration and reverse osmosis filtration in 2009. The purpose of this technological innovation is to eliminate the main THMs precursors, thus meeting the required sanitary limits in THMs formation.

In this Thesis, two studies are conducted using archival data from the end of 2003. The traditional operational stages of the water disinfection process in SJD-DWTP at that time were as presented in Figure 3.

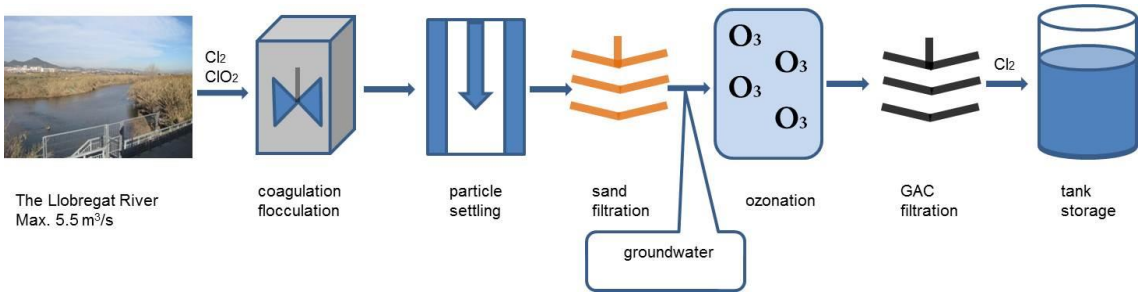


Figure 3. Traditional water treatment process in the Sant Joan Despí DWTP in 2003.

- **Captation:** The main water source is the raw river water from the Llobregat River. In the captation, raw water passes through a gallery of bar screen systems, removing sufficiently large solids. At this stage, groundwater from the Llobregat aquifer in Cornella might be directly used without the requirement to receive some treatment. This water source is employed only when surface river water flow is too poor or scarce.

- **Pre-treatment:** Following capitation, raw river water is subjected to preliminary pre-oxidation with gaseous chlorine and chlorine dioxide. Chlorine removes ammonia in water by producing chloramines. Additionally, chlorine dioxide is used as a disinfectant, primarily due to fact that it produces less THMs in comparison to the molecular chlorine. Despite its advantages, it cannot be employed in certain conditions, namely when the ammonia concentration in the incoming water is relatively high (more than 1 mg/L). In this situation, the water flow is slowed down to facilitate sedimentation and sand particles removal. Also, coagulants (aluminum salts) and flocculants are added to the process and water is pumped up into the clarifiers, where the precipitation of small particles takes place.

- **Sand filtration:** Following pre-treatment step, the water has to be filtered. The procedure requires that the water flow goes through sand filters, retaining all particles that were not precipitated and retained at earlier stages.

- **Groundwater addition:** Sand filtration may be followed by addition of groundwater from the Llobgregat aquifer. In this case, groundwater will be pumped into the main flow. River water and groundwater are usually mixed when the flow of the first one is not sufficiently high to meet water demand in the BMA.

  Disinfection process: Following the addition of groundwater (if applicable), the treated water is divided in two treatment lines: (a) a traditional treatment with ozone and filtration with granulated active carbon (GAC), and (b)ultrafiltration and reverse osmosis membrane filtration processes, which are followed by a specific reminalization procedure, which improve the water taste. The conventional treatment based on ozonation uses ozone as a biocide and oxidant. There are numerous microorganisms, which can be eliminated with ozone, and with the benefits of oxidizing the residual organic matter at the same time. As a result, the ozonation leads to significant improvements in the water organoleptic quality. The use of granulated active carbon filters additionally help in reducing organic matter and metal oxides (including iron, manganese, and nickel,). Finally, water from the GAC filtration and from the ozonation process is mixed with water from the ultrafiltration and RO membrane treatment. To improve the taste of drinking water, it is often and increasingly popular to have remineralization procedures with carbonic anhydride and using a calcic bed.

- **Post-chlorination:** Treated water is chlorinated with chlorine gas and stored in deposit tanks, where residual chlorine is maintained at certain levels so that water is disinfected prior to its distribution into the WDS.

The treatment and disinfection processes in SJD-DWTP are subjected to extensive controls and inspections. For example, physicochemical parameters - such as temperature, turbidity, color, Cr^{6+}, NH_4-N, UV absorption at 254nm, TOC, Cd, Pb, Ni, detergents, conductivity, pH - are constantly monitored in the raw river water. The purpose of such extensive checks and monitoring is to have the DWTP operation processes well-functioning by detecting unusual pollution events in advance. DWTP processes which are constantly monitored for these parameters are: (a) raw river water intake, (b) water from decanters, (c) water after sand filtration, (d) groundwater, (e) water after ozonation and GAC filtration, (f) fully disinfected water after

postchlorination and inside the WDS at different locations. Other physicochemical parameters which are monitored in the water treatment and disinfection are: dissolved oxygen, chlorides, bromides, heavy metals, odour, flocculation and coagulation doses, chlorine doses and the residual chlorine, THMs and other DBPs. To detect undesired pathogens growth and distribution, a regular permanent microbiological control is implemented

b) Abrera DWTP

The Abrera DWTP uses water from the Llobregat River and it is located nearby the municipality of Abrera. There are two water treatment and production lines. The first one is the conventional treatment process and it is composed by the following operations: a) water captation and bars screening which remove large suspended solids in water, b) sand decanters and preliminary treatment with $KMnO_4$ in order to oxidize organic and inorganic matter, c) pre-chlorination with chlorine and chlorine dioxide, d) flocculation and particle settling, e) chlorination with chlorine dioxide, f) sand and GAC filtration, g) post-chlorination with chlorine gas before to be transferred to storage tanks. The second line of production is the electrodialysis reversal treatment (EDR) line, which operates together with the traditional process. EDR process targets to reduce the high salinity (especially bromide concentrations) of the raw Llobregat river water and to reduce THMs formation. The incorporation of EDR processes required the incorporation of water remineralization processes also in the operational scheme of DWTP (Valero and Arbós, 2010).

c) Cardedeu DWTP

The DWTP Cardedeu is located in the municipalities of Cardedeu and Llinars La Roca del Vallès, Catalonia (Spain). It started its operation in 1966. Currently, it has a capacity of to manage 8m3/s. The treatment process contains the following phases and operations: (1) Water captation using an underground pipe and water from the three Ter river's dams (i.e., the Sau, the Susqueda and the El Pasteral); (2) An initial pre-oxidation of river water intakes using $KMnO_4$, NaClO and O_2.; and (3) Coagulation-

flocculation and first pre-chlorination with chlorine and chlorine dioxide, followed by decantation in tanks, GAC filtration and final post-chlorination.

d) Llobregat SWRO

The Llobregat SWRO is located nearby the Llobregat municipality, close to Barcelona. It was constructed after the severe drought in 2008 and has a water flow capacity of 2 m^3/s. The seawater desalination process includes: water extraction, flotation, microfiltration, reverse osmosis membrane treatment and posttreatment including remineralization processes, pH adjustments and disinfection procedures.

e) Besos DWTP

The Besos DWTP is in Barcelona city. In 2002, the DWTP was reopened as a source of supply water. It uses groundwater from the aquifer below the delta of the Besos River, Catalonia (Spain). The plant has a production line using nano-filtration with capacity of 5.200m3/d and three reverse osmosis lines with capacity of 8.700m3/d each.

2.1.6 Formation of trihalomethanes during water disinfection

Brief history of water disinfection

The ancient civilizations were largely concentrated around large sources of sweet water. The availability of large enough water quantity was the driving reason in choosing the location of ancestors' villages. The water quality was only of a secondary importance. The main criteria of quality for our predecessors was that the water is transparent, tasteless and without smell. For this reason, the river and lake waters were preferred over water from swamps. Several decades back in time, health problems stemming from poor water quality would not be a concern.

There are different sources (Kroehler, 2014.) which suggest that ancient Egyptians had some treatment practices, namely they were using the chemical alum to settle down particulates in the water, thus achieving some water visual effects

(turbidity). The ancient Greeks were using primitive aeration installations alongside their primitive water distribution systems, which transported water to the megapolises. The ancient Romans are known for using spring and wells water, which was transported via aqueducts, the first water distribution network system. Inside towns, the romans had plumbing systems with the purpose to avoid water contamination. In the Middle Ages, the lack of an appropriate disinfection and water distribution system was the reason for several epidemic diseases, including cholera and dysentery. Back then, primitive filters were first adopted. Historical documents suggest that in 1685, the first filter system consisted on settling and sand filtration was invented in Italy (AWWA, 2005). Various materials, including wool, sponges and charcoal, were used as a filter. Over time, around the beginning of 1800s, sand filtration was largely employed in the European countries.

The first scientific results in the field of water cleaning go back to Louis Pasteur (1822-1895), who provided evidence that microorganisms increase the probability of transmitting waterborne diseases. Over the next decades, the health effects of water pollution became a concern of an increasing importance. As a result of such trends, sand and charcoal filter plants were built in Scotland and France. However, only after chlorine was first used in 1854 as a response to the outbreak of cholera in London (White, 1986), the modern disinfection procedures have actually started. Soon after, the chlorination of potable water was adopted as a common practice throughout modern European countries, resulting in the successful reduction of typhoid, cholera and other water-borne disease outbreaks (Baxter, 1995).

Today, chlorine is the most used disinfectant in water treatment plants across the world (USEPA, 2006). In addition to chlorine, numerous other chemical reagents and techniques, including chloramines, chlorine dioxide, ozone and ultraviolet (UV) radiation, are largely employed nowadays for water disinfection (AWWA, 2000).

Chlorine disinfection

Numerous books and articles examine the basic chlorine chemistry for potable water disinfection. Two popular publications on the topic are: (a) *White's handbook of chlorination and alternative disinfectants"* with its 5th Edition by White (2010), and (b) the chapter titled "*Disinfection"* by Haas in "*Water quality and Treatment"* (1999).

There are three forms in which chlorine can be applied in water for disinfection purposes: as a compressed gas, as sodium hypochlorite solution, and as solid calcium hypochlorite. These forms can be presented as:

- Gaseous chlorine - $Cl_2 + H_2O \leftrightarrow H^+ + Cl^- + HOCl$ (1)
- Sodium hypochlorite – $NaOCl + H_2O \leftrightarrow H_2O + Na^+ + OCl^-$ (2)
- Calcium hypochlorite . $Ca(OCl)_2 + 2H_2O \leftrightarrow Ca^{2+} + 2OH^- + 2HOCl$ (3)

The chemical reaction of gaseous chlorine and calcium hypochlorite with water produces hypochlorous acid. In contrast, sodium hypochlorite in water gives hypochlorite ion. Furthermore, hypochlorous acid can also dissociate in water, thus producing a hypochlorite ion as well.

The above equations show that chlorination reactions are reversible and the dissociation of hypochlorous acid strongly depends on pH and on the temperature. More specifically, high pH will lead to the predominance of hypochlorite ion, whereas lower pH - to the predominance of hypochlorous acid (see Figure 4).

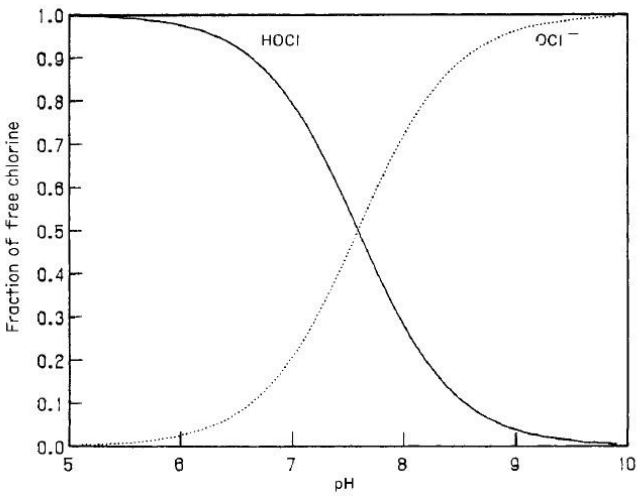


Figure 4. The pH effect of the distribution of hypochlorous acid (HOCl) and hypochlorite ion (OCl-) in water at different values and at 20 ºC (see Haas, 1999).

Figure 4 shows that at pH value of 5, HOCl will be approximately 100% and OCl^- would not exist in water. In contrast, at pH value of 10, OCl^- will be close to 100%. Comparing both species, high concentration of HOCl is more important than

OCl^- for water disinfection purposes, because it has more disinfection power onto pathogens (i.e., neutral molecule HOCl penetrates easily the negatively charged pathogens). As a result, it is common nowadays to add chlorine at lower pH values than pH 9 in DWTPs. An important operational consideration is that when chlorine is applied as gas, the pH value of treated water decreases (AWWARF, 1996). Apart from pH, several other parameters should be also considered for the chlorine disinfection, such as temperature, initial chlorine concentration, contact time between chlorine and the pathogens, and the type of pathogens.

The sum of the three concentrations of Cl_2, HOCl and OCl^- (so-called free available chlorine) is a very important operational parameter in DWTP which is constantly monitored during the treatment process. When associated with other compounds (ammonia and organic matter), chlorine is called combined chlorine. The sum of free available and combined chlorine defines the total chlorine, which is one of the most important operational parameters in DWTPs' monitoring.

Chlorine reacts with many organic and inorganic compounds in water and acts as a non-selective oxidant. Vasconcelos *et al.* (1997) and Deborde and von Gunten (2008) suggested that some metals in reduced state of valence, such as iron and manganese or halides (e.g., bromide and sulphide), are among the most reactive inorganic compounds. The rates of such reactions can be very fast and may last for seconds or hours. In contrast, the reaction between chlorine and organic compounds is slower, requiring more time. Since organic and inorganic compounds have different concentrations and react differently with chlorine, it is expected that the concentration of chlorine in water changes over time. Clark and Sivaganesan (1998) showed that approximately the half-life time of chlorine in disinfected water can vary from hours to days. Therefore, it is of high importance to consider the effect of concentrations and reactions over time, when monitoring residual chlorine levels. This problem is especially relevant for the BMA, where the entire water distribution system (WDS) is long, thus requesting sophisticated maintenance and special considerations in research chlorine decay modeling.

Chlorine decay

Because of the permanent reactions between chlorine and organic, and inorganic substances in water, the loss of chlorine in the WDS is a constant process.

There are two types of chlorine decay, which have been identified in the existing literature: (a) bulk chlorine decay, referring to the consumption of chlorine in the bulk aqueous phase, and (b) wall chlorine decay referring to the chlorine consumption from the pipe walls due to biofilms *(*Vasconcelos et al., 1997). It is commonly accepted that the sum of the bulk and wall chlorine decay is called the chlorine demand, which is specific and characteristic for every water source (Warton et al., 2006)

The chlorine demand in bulk water varies with the water source, largely due to the different organic and inorganic compounds and their concentrations present in different water sources. Because it is practically impossible to isolate different reactions (i.e., between the chlorine and various, both organic and inorganic, compounds), the scientific attention is focused primarily on reaction rates at different levels of chlorine (Vieira et al., 2004).

Numerous studies model the reaction kinetics of chlorine decay. However, there are also specific models containing defined disinfection scenarios with water from various sources. Vieira et al. (2004) summarized different kinetics models regarding the chlorine decay, including zero-order decay, first order decay, parallel first order decay, second order reaction kinetics and others specific models. The general assumption is that chlorine decay is characterized by two types of kinetic reactions, namely a rapid reaction taking place very fast in the initial seconds, and a slow reaction which can continue for several hours (Clark and Sivaganesan, 2002; Vieira et al., 2004). According to USEPA (1992), the chlorine decay has three stages: (a) an initial rapid reaction (less than 5 minutes), (b) a medium fast reaction (between 5 minutes and 5 hours), and (c) a slow reaction (more than 5 hours).

Disinfection by-products (DBPs)

There are two studies, which independently reported the first DBPs in chlorinated drinking water, Rook (1974) and Bellar *et al.* (1974). The trihalomethanes (THMs) were the first to be identified as a DBP. Since then, there have been numerous studies which were primarily interested in finding other DBPs.

To the extent that DBPs have significant implications for public health, the importance of this topic in relation to the drinking water quality has increased over the past decade (Richardson, 2003). Richardson et al. (2007) showed that, in 2007, there were more than 1000 DBPs scientifically documented in the literature. Nevertheless, as

of today, more than 50% of the total possible DPBs in drinking water are still unknown. As a result, our knowledge about the potential toxicity of most DBPs and their impact on human health is limited.

The more common types of halogenated organic DBPs according to USEPA (2001) are: (1) Trihalomethanes (chloroform, bromodichloromethane, dibromochloromethane and bromoform);(2) Haloacetic acids (monochloroacetic, dichloroacetic, trichloroacetic, monobromoacetic and dibromoacetic acids); (3) Haloacetonitriles (dichloroacetonitrile, bromochloroacetonitrile, dibromoacetonitrile, trichloroacetonitrile); (4) Haloketones (1,1-dichloropropanone and 1,1,1-trichloropropanone); (5) Chlorophenols (2-chlorophenol, 2,4-dichlorophenol, 2,4,6-trichlorophenol), and (6) others such as chloropicrin, chloral hydrate and others. These groups are also among the more examined and regulated types of DBPs. Other compounds, which are also regarded as major DBPs, include disinfectant residuals, including free chlorine, chloramines, chlorine dioxide. Other inorganic species such as chlorite, chlorate and bromate ions, aldehydes, and ketones, are also classified as DBPs.

Trihalomethanes (THMs)

In the early 80s, extensive work on DBPs focused on the trihalomethanes (THMs): chloroform $CHCl_3$, bromodichloromethane $CHCl_2Br$, chlorodibromomethane $CHBr_2Cl$, and bromoform $CHBr_3$. Such compounds represent the largest fraction, approximately 50%, of all halogenated DBPs in treated water (Krasner et al., 1989). Currently, this group is in the scope of many international and local legislation entities (WHO, 2005). THMs are a set of compounds ubiquitously formed in water disinfection (Rook, 1974; Richardson et al., 2002). They have three halogen atoms and one carbon - CHX_3. The brominated and chlorinated forms are most frequently found in potable water in comparison to iodinated and fluorinated forms. The THMs are formed when individual carbon atoms are attacked with halogen disinfectants. Small hydrocarbon chains are cracked from natural organic matter (NOM) molecules and the reaction of the halogen species lasts until THMs are finally formed (USEPA 2001).

The chemical reaction leading to the formation of THMs during chlorination can be represented as follows:

- $HOCl + Br^- + I^- + NOM \rightarrow$ THMs + other Halogenated DBPs (4)

More specifically, the compounds considered as THMs are chloroform ($CHCl_3$), bromodichloromethane ($CHCl_2Br$), chlorodibromomethane ($CHBr_2Cl$), and bromoform ($CHBr_3$). The sum of their concentrations is suggested to be an important parameter in quality monitoring (Amy et al., 1987). Typically, the measurement of these four THMs quantities is largely dependent on water qualities (Clark et al., 2001). Although THMs are volatile compounds and soluble in most organic solvents, their solubility in water is quite limited being less than 1 mg/ml at 25°C (WHO, 2005).

Precursors for THMs formation

The formation of THMs and their speciation is largely dependent on the water quality and the operating conditions in the DWTP. The most important precursor for THMs formation is the content of natural organic matter content (NOM) and its concentration (Kitis et al., 2001). Recently, a large number of studies showed that NOM includes different organic fractions which affect the THMs formation and speciation in various ways (Reckhow et al., 2004). Other important THMs precursors and factors affecting the THMs formations are residual chlorine, reaction time, pH, and bromide concentration.

Natural organic matter (NOM)

The formation of THMs is mainly determined by the NOM's concentration and characteristics (Leenheer et al., 2001). In terms of concentration, high concentrations of NOM in water implied that the demand for chlorine is high, and therefore, it is likely to produce an increase of the THMs formation during disinfection. Regarding NOM characteristics, empirical work showed that they play an important role in the THMs formation, especially for THMs speciation (von Gunten et al., 2001). Croué et al., (1999), for instance, found that different fractions of NOM resulted in different chlorinated by-products when the chlorination conditions were quite similar.

Natural organic matter (NOM) is present in every kind of water. Despite that they are often the driving factor for organoleptic problems of drinking water, NOMs are not harmful. When NOM reacts with chlorine, the formation of chlorinated DBPs, such as trihalomethane (THMs) and haloacetic acids (HAAs) naturally occurs. The natural organic matter in water exists as a heterogeneous mixture of humic, fulvic acids, proteins, lipids, carbohydrates, carboxylic acids, amino acids and hydrocarbons (Leenheer, 2004). NOM involves particular and dissolved organic matters fractions

(POM and DOM). DOM is suggested to be the most important factor in the DBPs formation. In terms of physicochemical content, a major part of NOM comprises two fractions, namely humic substances, which are composed of fulvic and humic acids, and non-humic substances, which include carbohydrates, lipids, and amino acids (Thurman and Malcolm 1981). The humic fraction is typically heavy on molecular weight aromatic molecules and is less soluble in water. The non-humic fraction is soluble in water and contains smaller on size molecules. Each of the NOM fractions can be additionally divided into acidic, alkaline and neutral subgroups.

The humic fraction mainly consists of humic and fulvic acids. The humic acid fraction is regarding as being more reactive in comparison to the fulvic acid fraction. It additionally has a higher molecular weight, a larger size, and a lower solubility in water. In contrast, fulvic acid is less reactive, has a lower molecular weight, is smaller in size, and has higher water solubility (Krasner et al., 1996).

Typical water quality parameters representing NOM in water, which are constantly monitored in DWTPs, include: total organic carbon (TOC), dissolved organic carbon (DOC), the ultraviolet absorbance at 254nm (V_{254}), and water color (Owen et al., 1993). Another important parameter in monitoring is the specific ultraviolet absorbance (SUVA). It is calculated when the UV absorbance is divided by DOC and serves as an indicator for aromatic and hydrophobic nature of the organic matter (Eikebrokk et al., 2006). For monitoring purposes, TOC and DOC are measured using the amount of CO_2 produced by UV-oxidation or by the combustion of the organic matter, in both cases detected though specific analysers. UV absorption characterizes the different type of double bonds in aromatic rings of the organic matter. Colour is an indicator of molecular complexity of NOM, characterizing multiple bonds and highly substituted aromatic groups (Newcombe et al., 1997).

pH

Academic research demonstrated that the rate of formation and the quantities of THMs increases when pH is high (Reckhow and Singer, 1985; Krasner et al., 1989). According to Trussell and Umphres (1978), THM formation involves a hydrolysis step which is facilitated by high values of pH. In contrast, prior work found that lower levels of pH reduce THMs formation (Chowdhury and Champagne, 2008), but also lead to an increase of other DBPs, such as HAAs (Nokes et al., 1999).

Temperature

The temperature as a factor in THMs formation plays a dual role. Laboratory studies found that an increase in temperature of up to 30°C causes an increase in THMs formation (Krasner *et al*., 1989; Rodriguez and Serodes, 2001). Higher temperatures also increase the THM formation reaction rate. The increase in temperature is the cause of higher rates of hydrolysis, thus water molecules dissociate to hydrogen and hydroxyl ions (Garcia et al. 2005). This process facilitates the break of the aromatic bonds of organic matter molecules, thus leading to halogenation and THMs formation. Krasner (1999) used field data of 35 DWTP in the United States and demonstrated that the mean THM formation was higher in the summer and lower over the in the winter.

Although, temperature has a positive correlation on the THMs formation, it is plausible that this effect is not linear. For temperature levels above 40 °C, for instance, it is likely that there is a decrease of THMs concentrations in water because of the evaporation processes.

Bromide/chloride concentration ratio

Bromide ion concentration is another factor with a significant role in the formation of brominated THMs. Its importance is related to the fact that the presence of free bromide ions affects the chemistry of chlorine disinfection. Bromide ions substitute chlorine ions of the hypochlorous acid to form hypobromous acid (HOBr) and hypobromite ion (OBr^-), which is considered 20 times more reactive in comparison to its acid (Singer 1999; Chang et al., 2001). The reaction can be represented as follows:

- $HOCl + Br^- = HOBr + Cl^-$ (5)

Both, HOBr and OBr^- react with the organic matter to form mixed bromo chloromethanes and bromoform.

Over the past years, there has been an increasing interest in bromide ions in relation to the THMs formation and speciation (Nokes et al., 1999; Chang et al., 2001; Elshorbagy et al., 2000) since brominated THMs were considered also dangerous for the human health.

Free chlorine

The presence of chlorine has a strong role in the THMs formation. The general understanding is that the THMs formation is favored by the increase of free chlorine concentration (Adin et al., 1991). Such a positive correlation however can be observed only when the concentration of chlorine is enough to fulfill the chlorine demand. If there is an excess of free chlorine, no additional increase of THMs is observed. In this case, the formation reaction will be additionally limited by the NOM concentration (Carlson and Hardy, 1998). Moreover, Reckhow and Singer (1985) demonstrated that when there is an increase of chlorine dosage as an excess of free residual chlorine, haloacetic acids, instead of THMs, are likely to be formed.

Epidemiological studies of THMs

The academic interest in THMs has significantly increased over the past decade, especially after they were associated with cancer risks, various chronic human health problems, low birth weights and pre-term births, and neural tube defects, among others (Richardson et al. 2002; Villanueva et al. 2004). Human health can be exposed lifelong to the effects of THMs by drinking potable water, by inhalation during showering, or by swimming in pools, among others. This permanent and frequently contact with THMs increases the risk to human health. Several epidemiological and toxicological studies examine such exposures to THMs using experiments conducted with laboratory animals. Mills et al., (1998), Villanueva et al. (2004) and Wang et al. (2007) showed a significant association between different exposures, such as oral ingestion, inhalation and dermal absorption, and risk of cancer, particularly bladder and rectal cancers. Some authors, including King and Marrett (1996), directly attributed a number of human bladder cancers in Ontario to the higher conentrations of THMs in the drinking water. Nieuwenhuijsen et al. (2000), Graves et al. (2001) and Villanueva et al., (2007), among others, reported THMs reproductive effects, such as intrauterine growth retardation, low birth weight, preterm birth, congenital malformations, and stillbirth. Furthermore, Dodds and King (2001) found that THMs at exposure concentrations above 20 ug/L or higher increase the probability of developing neural tube defects.

Existing studies with animals suggest that there are at least three THMs - chloroform, bromodichloromethane and bromoform, which are carcinogens. Although dibromochloromethane is also considered to be a carcinogen, at present it is suggested in animal studies that its impact on health is not so severe; moreover, there is no

evidence from other studies with human participants that this compound behaves as a carcinogen (USEPA 2001). Table 2 shows the main four THMs and their toxicology characteristics according to USEPA (1999).

Although many epidemiological surveys that relate directly THMs with a particular disease or health effect, empirical results are not so conclusive yet. The majority of related studies have been conducted with animals and THMs at very high concentrations, which are not possible during the actual disinfection processes of potable water. In fact, levels of THMs normally found in water are of significantly lower concentrations (Freese and Nozaic, 2004).

Table 2. Toxicology for THMs (USEPA, 1999)

		Effects[1]		*Toxicity to human*[2]	
THMs	*Disinfectant*	*Animal*	*Human*	*RfD*	*SF*
Chloroform	Chlorine	Liver tumours	B2	0.01	0.01
Bromodichloromethane	Chlorine	Kidney tumours	B2	0.02	0.062
Dibromochloromethane	Chlorine	Liver tumours	C	0.02	0.0084
Bromoform	Chlorine, ozone	Colon tumours	B2	0.02	0.0079

[1] B2: Probable human carcinogen (sufficient laboratory evidence); C: Possible human carcinogen

[2] RfD: Reference dose (mg/kg day); SF: Slope factor $(mg/kg\ day)^{-1}$

European legislation for trihalomethanes

Under Directive 98/83/EC, the European Union recently modified the maximum allowed concentration of total sum of THMs from 150 μg/l (prior to 2009, under Directive 80/778/EEC) to 100 μg/l. Spain had adopted the European Directive in 2003, permitting the total sum of THMs to reach a maximum concentration of 100 μg/l (Real Decreto 140/2003). Table 3 presents the actual norms about THMs levels according to the currently existing legislation.

Table 3. Existing standards related to THMs (mg/l) according to the main international regulators

THMs	WHO(2005)[1]	USEPA(2006)	EU
Chloroform	0.300[2]		
Dichlorobormomethane	0.060		
Dibromochloromethane	0.100		
Bromoform	0.100		
TTHM		0.080	0.100

[1] Maximum contaminant level goals

[2] WHO: World Health Organization; USEPA: US Environmental Protection Agency; EU European Union

Modelling THMs formation

In the literature, there are three groups of predictive THM models (Chowdhury et al., 2009). The first group of predictive models has an objective to identify all significant natural and operational water quality parameters, which are likely to affect the THMs formation. The second group of models is centered on the kinetics of the THMs formation and the third group of models is developed to predict THMs concentrations in field monitoring studies. The second and third groups of models were developed as an alternative to the classical standard methods for THMs control, which are time-consuming and require quite expensive GC-MS methods.

The most of the predictive models are focused on parameters such as TOC, DOC, UV_{254}, SUVA, which are typical for NOM; natural water quality parameters such

as pH, temperature, raw water bromide/chloride concentration ratio; and operational parameters like chlorine doses and reaction time. The existing knowledge suggests that such parameters have a strong effect on THMs formation (Sadiq and Rodriguez, 2004).

Following the establishment of significant correlation, scientific studies aimed at establishing prediction models on the basis of important parameters, thus generating reliable estimations for THMs formation. In general, such predictive models are using data from the field or from laboratory-scaled studies.

Field-study data are obtained during the monitoring of disinfection processes within DWTPs or inside WDS. Among the modelled parameters, we can usually find raw water quality parameters, operational parameters from prechlorination and post chlorination processes, and some other water quality parameters from the finished drinking water. In contrast, laboratory-scaled models are based on an experimentally-designed, batch analysis, using raw or treated water with pre-selected quality. Laboratory-scaled studies usually include chlorine dose, pH, contact time, temperature, and bromide concentration as modelling parameters.

It is generally considered that laboratory-scaled empirical models are more reliable in comparison to field studies. The perceived advantage of laboratory-scaled models is that there is a complete control over the investigated parameters and that they have easily adjustable values and initial water quality information about the water. Nevertheless, such laboratory-scaled models have an important disadvantage which is the lack of information about further THMs formation along WDS, and the possible presence of unknown dynamic processes including multiple parameters of water treatment with a significant influence over THMs formation in DWTPs. Moreover, models from laboratory experiments can be only applied to small, in terms of scale, DWTPs, and only provide basic indications about THMs formation. The general expectation is that these models are unlikely to report highly accurate predictions, largely because they do not account for the large number of DWTP processes and the real correlations among all parameters, producing significant effects on the THMs formation along DWTPs (i.e., flocculation-granulation process, granulated active carbon filtering, among others).

A significantly improvement over laboratory-scaled models is the application of experimental design (DoE) strategies. More specifically, DoE can minimize the uncertainties in predicting THMs formation by using pre-determined concentration ranges of monitored parameters. Among the disadvantages of field-scaled models is the

lack of information about the contact time between disinfectants and organic matters. Such information is not available or difficult to obtain at the DWTP-level. Moreover, the complex nature of organic matters is oftentimes difficult to be characterized without comprehensive laboratory analyses, which are likely to be quite expensive. In addition, the uniqueness of a particular geographic area (affecting the NOM content) would make the raw water quality and the DWTP disinfection procedures site-specific, thus impeding the generalization of field-scaled developed models and their large-scale application (or transfer) to other distinctive settings (i.e., other plants or WDS) (Rodriguez et al., 2000). Furthermore, the seasonal variation of NOM content is also site-specific, and therefore it should be thoroughly considered when applied to field-based model predictions.

A comprehensive review of the existing literature suggested that a universal mechanistic-kinetic model has not been developed yet, primarily because of the uncertainty of the reaction between chlorine and complex organic matter. Over the past decades, several reviews summarize the research on predictive models (presented in a chronological order): Amy et al. (1987), Clark et al. (2001), Sadiq and Rodriquez (2004), Chowdhury et al (2009). Approximately fifty scientific publications reported more than 118 DBP formation models over the period 1983-2009 (Chowdhury et al., 2009). Forty-two out of these 50 articles contain THMs models. Additionally, approximately 49 models on TTHMs formation, 12 models - chloroform formation, 8 models - bromodichloromethane formation, 6 models - dibromochloromethane formation, and 6 models - bromoform formation, were reported.

Published models (see Table 4) usually have been based on empirical equations with THMs concentrations, and different raw water and operational parameters. Only few articles reported kinetic models and used factorial design approaches (e.g., Rodriguez et al., 2007). In these factorial-design studies, the authors investigated the effects of selected parameters as main factors, and studied the interaction between these parameters by varying their concentrations or values.

A significant part of developed predictive models has been based on empirical and semi-empirical kinetic equations. These equations have been developed using linear and non-linear regression analysis. The majority of such predictive models, based on multiple parameters, were following the proposed model by Amy et al. (1987), described as:

$$TTHM = k \times (t)^a \times (C_0)^b \times (TOC)^c \times (T)^d \times (pH)^e \times (UV_{254})^f \times (Br)^g \quad (6)$$

where TTHM stands for total trihalomethanes

k - Reaction constant

t - Time

C_0 -Initial chlorine concentration at t=0

TOC - Total Organic Carbon (mg/L)

T - Temperature (°C)

pH - pH

UV_{254} – UV absorbance at 254 nm

Br - Bromide concentration (mg/L)

a-g - Reaction constants

Table 4. Predictive models for trihalomethanes formation reported in the literature for the period of 1983-2009 (according to Chowdhury et al., 2009)

Authors, year	Model characteristics	R^2	Data source
Milnear and Morrow, 1983	TTHM = -3.91 + $(Br^-)^{0.15}$ + 0.23(log(Cl_2)) + 0.24(pH) + $10^{0.009T}$ + 0.26(NVTOC) in (µmol/L)	>0.9	Laboratory
Urano et al.,1983	TTHMs = 0.00082 (pH-2.8) TOC $(Cl_2)^{0.25}$ $(t)^{0.36}$	NR	Laboratory
Engerholm and Amy, 1983	$CHCl_3=k_1k_2(TOC)^{0.95}$ $(Cl_2/TOC)^{0.28}$ $(t)^z$	NR	Laboratory
Milnear and Morrow, 1987	TTHM = -3.91+$(Br^-)^{0.15}$ + 0.23(log(Cl_2)) + 0.24(pH) + $10^{0.009T}$ + 0.26(NVTOC) in (µmol/L) TTHM = -3.94 + $(Br^-)^{0.19}$ + 0.35(log(Cl_2)) + 0.24(pH) + $10^{0.009T}$ + 0.27(NVTOC) in (µmol/L) TTHM = -2.42 + $(Br^-)^{0.15}$ + 0.24(log(Cl_2)) + 0.24(pH) + $10^{-204.5T}$ + 0.25(NVTOC) in (µmol/L)	>0.9 NR NR	Laboratory
Amy et al., 1987	TTHM = $0.0031(UV_{254}.TOC)^{0.440}(Cl_2)^{0.409}(t)^{0.265}(T)^{1.06}$(pH-2.6)$^{0.715}(Br^-+1)^{0.0358}$ in (µmol/L)	0.9	
Adin et al.,1991	TTHM = $K_1.K_2$.TOC[(1/((K_1 + K_3)(K_2 + 0.19))) + (1/(K_1 + K_3 – K_2 – 0.19)) x (((1/(K_1 + K_3))$exp^{-(K_1 + K_3)(tc)}$) – ((1/(K_2 +0.19))$exp^{-(K_2 + 0.19)(tc)}$)] as K_1=4.38 x 10^{-8}(Cl_2), K_2=11.36x10^{-7}(Cl_2), K_3=7.14x$10^{-13}$$(Cl_2)^2$ in (µg/L)	0.9	Laboratory
Harrington et al., 1992	TTHM = $0.0039(TOC.UV_{254})^{0.44}(Cl_2)^{0.409}t^{0.265}T^{1.06}$ (pH-2.6)$^{0.715}(Br^- + 1)^{0.03}$ in (µg/L)	NR	Field
Malcolm Pirnie Inc., 1992	$CHCl_3$ = $0.078(TOC.UV_{254})^{0.616}(Cl_2)^{0.391}t^{0.265}T^{1.15}$(pH-2.6)$^{0.8}(Br^- + 1)^{-2.23}$ in (µg/L) BDCM = $0.863(TOC.UV_{254})^{0.177}(Cl_2)^{0.309}t^{0.271}T^{0.72}$(pH-2.6)$^{0.925}(Br^- + 1)^{0.722}$ in (µg/L) DBCM = $2.57(UV_{254}/TOC)^{-0.184}(Cl_2)^{-0.0746}t^{0.252}T^{0.57}$(pH-2.6)$^{0.8}(Br^- + 1)^{-2.23}$ in (µg/L)	NR	Field
Malcolm Pirnie Inc., 1993	THMs = $7.21(TOC)^{0.004}(UV_{254})^{0.534}(Cl_2-7.6\times NH_3-N)^{0.224}$ $(t)^{0.255}$ $(Br^-+1)^{2.01}$ $(T)^{0.480}$(pH−2.6)$^{0.719}$ in (µg/L) BDCM = $4.05(TOC)^{0.567}(UV_{254})^{0.567}(Cl_2-7.6\times NH_3-N)^{-0.351}$ $(t)^{0.366}$ $(Br^-)^{0.291}$ $(T)^{0.568}$(pH−2.6)$^{0.568}$ in (µg/L) $CHCl_3$ = $0.997(TOC)^{0.580}(UV_{254})^{0.580}(Cl^2)^{0.814}$ $(t)^{0.278}$ $(Br^-+1)^{-4.27}$ $(T)^{0.569}$(pH−2.6)$^{0.759}$ in (µg/L) DBCM=$22.9(TOC)^{0.253}(UV_{254})^{0.253}(Cl_2-7.6\times NH3-N)^{-0.352}$ $(t)^{-0.292}$ $(Br^-)^{1.04}$ $(T)^{0.491}$ (pH−2.6)$^{0.325}$ in (µg/L) $CHBr_3$=$1.28(TOC)^{-0.167}(UV_{254})^{-0.167}(Cl_2-7.6\times NH3-N)^{-2.22}$ $(t)^{0.294}$ $(Br^-)^{1.48}$ $(T)^{0.553}$ (pH−2.6)$^{0.198}$ in (µg/L)	NR NR NR NR NR	Field
Montgomery Watson, 1993	$CHCl_3$=$0.064(TOC)^{0.329}(UV_{254})^{0.874}(Br^-+0.01)^{0.404}$ $(pH)^{1.161}$ $(Cl_2)^{0.561}$ $(t)^{0.269}$ $(T)^{1.018}$ in (µg/L) BDCM=$0.0098(Br^-)^{0.181}(pH)^{2.55}(Cl_2)^{0.497}(t)^{0.256}(T)^{0.519}$ (for Cl2/Br^-<75) in (µg/L)	0.88 0.80	Laboratory Labor

	BDCM=1.325(TOC)$^{-0.725}$(Br$^-$)$^{0.794}$(Cl$_2$)$^{0.632}$(t)$^{0.204}$(T)$^{1.441}$ (for Cl$_2$/Br$^-$>75) in (µg/L) DBCM=14.998(TOC)$^{-1.665}$(Br$^-$)$^{1.241}$(Cl$_2$)$^{0.729}$(t)$^{0.261}$(T)$^{0.989}$ (for Cl$_2$/Br$^-$<50) in (µg/L) DBCM=0.028(UV$_{254}$)$^{-1.175}$(TOC)$^{-1.078}$(Br$^-$)$^{1.573}$(pH)$^{1.956}$(Cl$_2$)$^{1.072}$(t)$^{0.2}$(T)$^{0.596}$ (for Cl$_2$/Br$^-$>50) in (µg/L) CHBr$_3$=6.533(TOC)$^{-2.031}$(Br$^-$)$^{1.388}$(pH)$^{1.603}$(Cl$_2$)$^{1.057}$(t)$^{0.136}$ in (µg/L)	0.92 0.82 0.83 0.86	atory Labor atory Labor atory Labor atory
Lou and Chiang, 1994	TTHM=TTHMo+7.01(pH−2.3)$^{0.11}$ (NVTOC)$^{1.06}$ (t)$^{0.748}$(Cl$_2$)$^{0.764}$(β) in (µg/L)	NR	Field
Ibarluzea et al., 1994	CHCl3=10.8+0.04(Flu)+1.16(pH)+0.12(T)+1.91 (Co) in (µg/L)	0.82	Field
Siddiqui et al., 1994	CHBr$_3$=2.68(DOC)$^{1.28}$(pH)$^{-1.31}$(O$_3$)$^{0.742}$(Br$^-$)$^{1.55}$(T)$^{0.956}$(t$_m$)$^{0.353}$ in (µg/L)	0.78	Field
Rathbun, 1996a	CHCl$_3$=0.442(pH)^2(Cl$_2$)$^{0.229}$(DOC)$^{0.912}$(Br$^-$)$^{-0.116}$ in (µg/L) BDCM=17.5(pH)$^{1.01}$(Cl$_2$)$^{0.0367}$(DOC)$^{0.228}$(Br$^-$)$^{0.513}$ in (µg/L) DBCM=26.6(pH)$^{1.80}$(Cl$_2$)$^{-0.0928}$(DOC)$^{-0.758}$(Br$^-$)$^{1.2}$ in (µg/L) CHBr$_3$=0.29(pH)$^{3.51}$(C$_{l2}$)$^{-0.347}$(DOC)$^{-0.330}$(Br$^-$)$^{1.84}$ in (µg/L)	0.97 0.86 0.94 0.78	Labor atory Labor atory Labor atory Labor atory
Rathbun, 1996b	TTHM=14.6 (pH−3.8)$^{1.01}$ (Cl$_2$)$^{0.206}$ (UV$_{254}$)$^{0.849}$ (t)$^{0.306}$ in (µg/L)	0.98	Labor atory
Chang et al., 1996	TTHM=12.72 (TOC)$^{0.291}$ (t)$^{0.271}$ (Cl$_2$) $^{-0.072}$ in (µg/L) TTHM=108.8(TOC)$^{0.2466}$ (t)$^{0.2956}$(UV$_{254}$)$^{0.9919}$ (Cl$_2$)$^{0.126}$ in (µg/L) TTHM=131.75(t)$^{0.2931}$(UV$_{254}$)$^{1.075}$ (Cl$_2$)$^{0.1064}$ in (µg/L)	0.94 0.97 0.95	Labor atory Labor atory Labor atory
Garcia-Villanova et al., 1997a	ln(CHCl$_3$)=0.348+0.00059(T)^3−0.000023(T)^4+0.0237(pH)^2+d+e in (µg/L)	0.65	Field
Garcia-Villanova et al., 1997b	ln(CHCl$_3$)=0.81Y+0.162N+0.00047(T)^3−0.0000204(T)^4+0.00339(pH)^2+e in (µg/L)	0.86	Field
Golfinopoulos et al., 1998	TTHMs=13.5ln(Chla)−14.5(pH)+230(Br$^-$)−140(Br$^-$)^2−25.3(S)+110.6(Sp)−6.6(T.Sp)+1.48 (T.D) in (µg/L)	0.98	Field
Amy et al. (1998)	TTHMs=0.00412(DOC)$^{1.10}$(Cl$_2$)$^{0.152}$(Br$^-$)$^{0.068}$(T)$^{0.61}$ (pH)$^{1.60}$ (t)$^{0.260}$ in (µg/L)	NR	Labor atory
Rodriguez et al., 2000	TTHM=0.044(DOC)$^{1.030}$ (t)$^{0.262}$ (pH)$^{1.149}$ (Cl$_2$)$^{0.277}$ (T)$^{0.968}$ in (µg/L) TTHM=1.392(DOC)$^{1.092}$(pH)$^{0.531}$(T)$^{0.255}$ in (µg/L)	0.90 0.34	Labor atory Field
Sung et al., 2000	TTHM=a(OH$^-$)^i(Cl$_2$(1−e^{-kt}))(UV$_{254}$)^n(algae)^p CHCl$_3$=2.3×10^6(OH$^-$)$^{0.52}$(Cl$_2$(1−e^{-kt}))$^{0.56}$(UV$_{254}$)$^{0.57}$(algae)$^{-0.10}$ in (µg/L)	 0.93	Field
Westerhoff et al.,2000	TTHM=b$_o$+b$_1$(DOC)+b$_2$(Cl$_2$)+b$_3$(Br$_{-10\%}$)+b$_4$(T)+b$_5$(pH)+b$_6$(t) in (µg/L)	NR	Field
Elshorbagy et al. ,2000	TTHM$_{t+\Delta t}$=TTHM$_t$+0.582(Cl$_{t+\Delta t}$−Cl$_t$) in (µmol/L)	NR	Field
Golfinopoulos and Arhonditsis,2002	TTHM=−0.26chla+1.57pH+28.74Br−66.72Br^2−43.63S+1.13Sp+2.62T.S−0.72T.D in (µg/L) CHCl$_3$=−0.32chla+0.68pH+2.51Cl$_2$+1.93Sp−22.07S+1.38T.S−0.12T.D BDCM=−0.37chla+0.32pH+1.16Br−29.82Br^2+1.88Cl$_2$+5.17S−0.37T.Sp−0.12T.D	0.52 0.51 0.62	Field Field Field
Gang et al.,2002	TTHMs=αCl$_2$(1−fe^{-kt}r−(1−f)e^{-kt}s)	NR	Labor atory
Serodes et al., 2003	TTHM=16.9+16.0(TOC)+3.319(Cl$_2$)−1.135(T)+1.139(t) in (µg/L) log(TTHM)=−0.101+0.335(TTHM$_o$)+3.914(TOC)+0.117(t) in (µg/L) TTHM=21.2+2.447(Cl$_2$)+0.499(t) in (µg/L)	0.78 0.89 0.56	Labor atory Labor atory Labor atory
Nikolaou et al.,	logTTHM=0.33pH−0.02pH^2+0.12t−0.004t^2 in (µg/L)	0.53	Labor

2004	logTTHM=−0.44pH+7.53logpH−1.10Cl_2+0.2Cl_2^2 in (µg/L) logTTHM=0.98log(pH)+1.1log(t)−0.01(t).(Cl_2)+1.59log(Cl2) in (µg/L)	0.58 0.38	atory Labor atory Labor atory
Al-Omari et al., 2004	[TTHM]=4.527$t^{0.127}$ $Cl_2^{0.595}$ $TOC^{0.596}$ $Br^{0.103}$ $pH^{0.66}$ in (µg/L)	NR	Field
Kolla, 2004	TTHM=0.0001$Cl_2^{3.14}$ $pH^{1.56}$ TOC $^{0.69}$ $t^{0.175}$ in (µg/L)	0.77	Labor atory
Lekkas and Nikolaou, 2004	logTTHM=1.546+0.631pH^2+0.569log(t)+0.385log(Cl_2) in (µg/L)	0.87	Labor atory
Sohn et al., 2004	TTHM=$10^{-1.385}$$(DOC)^{1.098}$$(Cl_2)^{0.152}$ $(Br^-)^{0.068}$$(T)^{0.609}$$(pH)^{1.601}$$(t)^{0.263}$ in (µg/L) TTHM=0.42$(UV_{254})^{0.482}$$(Cl_2)^{0.339}$$(Br^-)^{0.023}$ $(T)^{0.617}$$(pH)^{1.601}$$(t)^{0.261}$ in (µg/L) TTHM=0.283$(DOC*UV_{254})^{0.421}$$(Cl_2)^{0.145}$ $(Br^-)^{0.041}$ $(T)_{0.614}$$(pH)^{1.606}$$(t)^{0.261}$ in (µg/L) TTHM=3.296$(DOC)^{0.801}$$(Cl_2)^{0.261}$ $(Br^-)^{0.223}$$(t)^{0.264}$ in (µg/L) TTHM=75.7$(UV_{254})^{0.593}$$(Cl_2)^{0.332}$ $(Br^-)^{0.0603}$$(t)^{0.264}$ in (µg/L) TTHM=23.9$(DOC*UV_{254})^{0.403}$$(Cl_2)^{0.225}$ $(Br^-)^{0.141}$$(t)^{0.264}$ in (µg/L) TTHM=$(TTHM_{pH=7.5,\ T=20\ °C})$*1.156$^{(pH-7.5)}1.0263^{(T-20)}$ in (µg/L)	0.90 0.70 0.81 0.87 0.90 0.92 0.92	database
Uyak et al.,2005	TTHM=0.0707$(TOC+3.2)^{1.314}$$(pH-4.0)^{1.496}$ $(Cl_2-2.5)^{-0.197}$$(T+10)^{0.724}$ in (µg/L)	0.98	Field
Uyak and Toroz, 2005	Log(TTHM)=1.078+0.398log(TOC)+0.158log(T)+0.702log(Cl_2) in (µg/L)	0.83	Field
Rodriguez et al., 2007	THMs=16.0+1.6FA+0.1Cl_2+0.3T−0.8FA×T−1.2FA^2−2.8Cl_2^2 in (µg/L) $CHCl_3$=3.5+0.8FA+0.02 Cl_2+0.07T−0.3T^2 in (µg/L) BDCM=4.5+0.7FA+0.04 Cl_2−0.8Cl_2^2+0.4T^2 in (µg/L) DBCM=4.0+0.4FA+0.05 Cl_2+0.1T−1.0Cl_2^2−0.7FA^2 in (µg/L) $CHBr_3$=4.0−0.2FA+0.03 Cl_2+0.09T−0.6 FA×T−0.5 FA^2−0.8Cl_2^2 in (µg/L)	NR NR NR NR NR	Labor atory Labor atory Labor atory Labor atory Labor atory
Hong et al., 2007	TTHM = $10^{-1.375}$ $t^{0.258}$$(Cl_2/DOC)^{0.194}$ $pH^{1.695}$ $T^{0.507}$$(Br^-)$ $^{0.218}$ in (µg/L) BDCM=$10^{-3.201}$$t^{0.297}$.$pH^{2.878}$$T^{0.414}$$(Br^-)^{0.371}$ in (µg/L) CHCl3 = $10^{-0.748}$$t^{0.210}$$(Cl_2/DOC)^{0.221}$ $pH^{1.374}$ $T^{0.532}$ $(Br-)^{-0.184}$ in (µg/L)	0.87	Labor atory
Semerjian et al., (2009)	$TTHM^2$=17.31+10.52Cl_2^2+259728.60$(SUVA)^2$ in (µg/L) $TTHM^2$=42.10+29.23Cl_2^2+353375$(UV_{254})^2$ in (µg/L) $TTHM^2$=−471.11+0.48t^2+1856.07$(Br^-)^2$+404.38Cl_2^2 in (µg/L)	0.39 0.33 0.31	Field, Lab
Chen and Westerhoff, 2010	CHCl3=1805$(DOC)^{0.11}$$(UV_{254})^{1.22}$$(Br^-+1)^{-2.19}$ BDCM=137$(DOC)^{0.16}$ $(UV_{254})^{0.94}$$(Br^-+1)^{3.66}$ TTHM=1147$(DOC)(UV_{254})^{0.83}$$(Br^-+1)^{0.27}$	0.88 0.60 0.87	lab
Chowdhury et al, 2010	TTHM=-51.408 + 8.449(DOC) + 13.529(Cl2) + 2.997(pH) +0.803(T) + 0.504(t) + 0.141((Cl2 -4.47)(T-15.03))		
Singh et al., 2012	TTHM=71.439 + 8.159(Cl_2/DOC) – 2.411(Cl_2/DOC)^2 + 31.014(pH) + 4.921$(pH)^2$ + 3.747(Br^-) + 3.061$(Br^-)^2$ +16.086(T) -2.424$(T)^2$ +26.467(t) -12.474$(t)^2$ + 2.775((Cl_2/DOC)(pH)) -2.036((Cl_2/DOC)(Br^-)) + 1.785 ((Cl_2/DOC)(T)) + 2.537((Cl_2/DOC)(t)) + 5.7004((pH)(Br^-) + 8.309((pH)(T)) + 14.961(pH)(t)) – 0.706((Br^-)(T)) + 1.717((Br^-)(t)) + 6.434((T)(t))	0.99	lab
Abdullah and El-dien Hussona,2013	TTHM= 1.58(UV254 x TOC)$^{0.38}$$(Cl_2)^{1.14}$$(t)^{0.6}$$(T)^{0.5}$$(pH-2.6)^{0.96}$$(Br^-)^{0.6}$	0.88	lab
Bach et al., 2015	CHCl3= 10.7 + 6 (HA) + 6.5 (ClO^-) + 4.3 pH CHBr3= 12.7 + 4.4(HA) + 7.8(ClO^-) + 8.7T + 5.4pH	0.9	Lab

TTHM= total sum of trihalomethanes; $CHCl_3$= chloroform; BDCM=bromodichloromethane; DBCM=Dibromochloromethane; CHB_3=bromoform; NVTOC=nonvolatile total organic carbon (mg/L); TOC=total organic carbón (mg/L);DOC=dissolved organic carbon (mg/L);UV_{254}= ultraviolet absorption at 254nm (cm^{-1}); SUVA =specific UV absorbance(L/mg-m) at 254nm in m^{-1} divided on TOC; Cl2=chlorine dose (mg/L); T = temperature in °C; t= reaction time (hours); tm=reaction time in minutes; t= reaction time (hours); tm=reaction time in minutes; *u*= rate constant(min^{-1}); C_{A0}= initial concentration of chlorine(mg/L); K=dimensionless constant; Br^- =bromide ion concentration (mg/L); NH3-N= ammonia nitrogen (mg/L); q=dimensionless time; C_0= residual chlorine at plant (mg/L); 0_3=ozone dose; Chla= chlororphyll-a (mg/m^3);OH^- hydroxide concentration; k=rate constant; ; Q outflow= outflows in the finished water reservoirs; V0 volume of the tank; FA=fulvic acid (mg/L); ClO^-hypochlorite ion (mg/L); γ= reactivity of HOBr times stronger than that of HOCl.

The work by Amy et al. (1987) has received high recognition in the literature and it is among the most frequently cited references on the topic of THMs formation modelling. Additionally, their modeling equation is included in the WRc OTTER water treatment modelling commercial package for DWTPs (Bridge, 2005). Their equation includes all parameters, which are known to affect the kinetic formation of TTHMs. Their model was based on the assumption that chlorine residual concertation was constant during the reaction period of a week, and that TTHMs are formed continuously during this period. Parameters were transformed into natural log forms, and multiple linear regression (MLR) was used. More specifically, the model explains the variance in the the natural log of TTHM formation as a function of the natural logs of time, temperature, initial chlorine concentration, the product of UV absorbance and TOC concentration, bromide concentration and pH. To obtain the actual TTHM concentration level, its anti-log transformation is calculated at the end of modeling.

A large number of models, both empirical (i.e., field studies) and kinetic (i.e., laboratory-scaled studies) were developed using multiple linear regression method. The main advantages of this regression method is that it allows obtaining and performing post-estimation diagnostics of the coefficients of determination (R-square), correlation coefficients, mean errors of prediction, thus evaluating model performance. MLR is also appropriate in many of these studies, given that they had a small number of parameters, which are likely not to be correlated. However, the high correlation existing between the independent parameters in the equation will present a serious concern in empirical analysis, giving unstable regression coefficients and unreliable standard errors. In addition to the possible multicollinearity concern, another limitation of a large number of published models was the absence of an external model validation procedure. Namely, the estimated models are not tested with external samples. This is required to obtain and report generalizable predictions. Furthermore, many models have been built using a field data sets obtained over a relatively short time interval. As a result, such models did not actually take into account the temporal changes in the investigated parameters, which may vary significantly over time, such as organic matter that can vary in its composition seasonally (Teixeira and Nunes, 2011).

2.1.7 Water distribution system (WDS)

Treated and disinfected water is transferred to large storage deposits before its final transfer to the WDS. Barcelona WDS includes more than 5500km of pipe network and approximately 150 smaller deposits. The two rivers (the Llobregat river and the Ter river), the underground water and the sea water have different water quality properties, which, along with the different treatment procedures in the five DWTPs explains why the supply of drinking water has different quality and organoleptic characteristics in the BMA. There are various blending processes, implemented in the distribution network to homogenize water quality prior its consumption. To facilitate this process, new distribution network interconnections were recently constructed with the objective to facilitate the blending of different water sources, to standardize the aesthetic water characteristics and to assure the constant supply of drinking water in the BMA (Valero and Arbós, 2010). An important element of the efficient management of this water distribution is the possibility to identify the water sources for the blending process inside the WDS. This analysis is critical for the proper functioning of the WDS, as well as the elimination of pipes corrosion (Lahav et al., 2009). Therefore, a methodology to distinguish different water sources would significantly facilitate the identification of the origins of contamination during accidents, as well as the source of organoleptic complains and illegal consumption alongside the network. Hence, the global management of WDS would significantly be improved with the application of such technology.

Currently, there are only few predictive models that are able to distinguish the water origins using hydraulic models and physiochemical tracers, such as conductivity parameter. Rubulis et al. (2011), for instance, studied the supply of drinking water to the city of Riga (Latvia), where six water sources with different conductivity values were mixed prior to consumption. Although the objective was to distinguish different water sources, this was not achieved for at least two reasons. First, the experiments were based only on one parameter, and second, sophisticated data analysis and modeling were required. In a different geographic region (Shanghai, China), Shu et al. (2010) developed prediction models using the decay of chlorine along the distribution network, among others. Similar to Rubulis et al. (2011), the models were not sufficiently precise in distinguishing water origins. The reason for such inaccurate predictions was that chlorine was incorporated in two instances (i.e., in the intake and during the water

transport in the WDS), resulting in a complex formulation which was notable to disguise the water origin.

2.1.8 Wastewater Treatment Plant of Girona, Trargisa S.A. (WWTP)

Wastewater is accumulated at the municipality level, by the industry in the area and from the urban runoff. After being aggregated, it is transported using sewers systems to the urban wastewater treatment plant (WWTP). WWTP is a complex of installations, where the wastewater is subjected to a series of physical, chemical and biological treatment processes for a removal of pollution, thus ensuring the quality of water, as required by the existing legislation, prior to its release in nature or reuse (Ostace et al., 2013).

The WWTP of Trargisa is situated nearby Girona city (Catalonia, Spain) and it has a capacity of up to 55,000 m^3/day. Figure 5 shows the main installations of the WWTP Trargisa . It can manage the collected wastewater from a population area with approximately 200.000 population equivalents. WWTP design includes a coarse pre-inlet pumping station, pre-treatment (screens removal, sand removal and degreasing), physico-chemical primary treatment, biological secondary treatment reactors, secondary clarifiers, and collecting purified water outlet to the River Ter. Perhaps, one of the most characteristic features of the WWTP (Trargisa) is the biological treatment with elimination of nitrogen and phosphorus, followed by a tertiary treatment.

Figure 5. Installations of WWTP TRARGISA at Girona, Catalonia, Spain (source Google maps).

Urban wastewater is primarily produced as a result of human activity. Water pollution in urban areas is associated to the use of water in domestic and public services, such as cleaning, bathing, and transport, the activities of households, and the urban runoff. The main contamination of domestic wastewater is from the degradable organic matter in suspension and in solution. Therefore, the primary contributors to the wastewater are:

- Fecal waters;
- Waters from domestic washing and small industrial processes;
- Water from the drainage system of streets;
- Rainwater and leachate.

Although urban waste water is largely homogeneous, in terms of composition and pollutant charge, such characteristics have a large variance. The reasons for this variance are the socio-economic characteristics of different municipalities, the nature and intensity of business activities in the area, and the climatic conditions, among others.

Different parameters can characterize a particular wastewater system (Metcalf abd Eddy, 2003). To quantify the level of wastewater contamination, usually, laboratory analyses of a set of parameters, which are commonly used as indicators of water quality

changes during the treatment processes, are performed. The parameters, commonly used to characterize and measure the water quality along the plant, are:

- *Biochemical oxygen demand* (BOD):

  BOD is defined as the amount of dissolved oxygen consumed by microorganisms (e.g., aerobic bacteria) in the oxidation of organic matter during the biological processes under certain conditions and at a given time. One of most important parameter to monitoring is called BOD5, which indicates the oxygen consumed by microorganisms in 5 days, during the oxidation process of organic matters.

- *Chemical oxygen demand* (COD)

  COD is a method for an indirect measurement of the amount of water pollution, which cannot be biologically oxidized. The COD measurement is based on the chemical decomposition of organic and inorganic contaminants, dissolved or suspended in water. COD value indicates the amount of water-dissolved oxygen (expressed as parts per million or milligrams per liter of water) consumed by the contaminants, during two hours of decomposition by a solution of boiling potassium dichromate.

- *Total suspended solids* (TSS)

  TSS is composed by organic and inorganic solid materials, including suspended in the water. Organic solids include proteins, hydrocarbons, fats and others that come from human activity. Inorganic solids include inert compounds such as salts, sands and soils. Aquatic microorganisms also play important role in TSS. This group includes viruses, bacteria, protozoa, algae, and others. A high concentration of TSS indicates a low water quality. TSS absorbs sunlight, which produces heating and a decrease of dissolved oxygen necessary for aquatic life. TSS also can clog rivers and even fish gills. Usually, TSS result from erosion from urban runoff, small industrial wastes, bank erosion, algae growth or wastewater discharges.

- *Temperature*

  The water temperature is one of the more important parameters for water quality, primarily because it has a strong influence on both, on the development of aquatic life, and on the chemical reactions and reaction rates. Additionally, biological treatment and nitrification processes are strongly affected by temperature changes. Severe temperature dynamics can cause increased mortality of the aquatic life. In a number of instances, high temperatures can lead to the proliferation of aquatic plants.

- *Dissolved oxygen (DO)*

  DO is critical for the development of living beings and therefore it is considered a basic water quality parameter for the control of wastewater as it is used for biological secondary treatment control. The increase in oxygen in water is due to:

  - Transportation of the oxygen through the interface surface water - air.
  - Photosynthesis, mainly due to green algae.
  - Lowering of temperature.
  - Dilution processes (usually when raining).

  In contrast, the amount of oxygen in the water decrease because of:

  - Micro and macro organisms respiration.
  - Temperature increase.
  - Chemical reactions.
  - Microorganisms enzyme reactions.

  Dissolved oxygen at high amounts is desirable and useful for prevention of the formation of unpleasant odors in wastewaters.

- *pH*

  Urban waters provide favorable conditions for the aquatic life at pH close to 7. If there is a significant increase or decrease in wastewater pH, this is an indication of possible industrial discharge pollution. It is necessary to control pH levels between 6.2 and 8.5, because fault inhibition of the biological processes would take place in this range of pH.

- *Nitrogen*

  Serving as a nutrient, nitrogen is an essential element for plants growth. To the extent that it is important for the protein synthesis, it is a critical parameter to follow its availability and concentration in water, thus evaluating biological processes and as a possibility of wastewater treatment. Lower concentrations of nitrogen would require an additional supply of nitrogen to the wastewater. Nitrogen is measured as total organic nitrogen (NT), ammonium-nitrogen (NH_4^+-N), nitrites and nitrates (NO_2^--N and NO_3^--N). In contrast, higher concentrations of nitrogen, released in nature, can cause excessive growth of algae and other plants, leading to accelerated eutrophication, and occasional loss of dissolved oxygen.

- *Phosphorus*

  Phosphorus is an important factor for microorganisms' growth in water. Because it acts as a nutrient for the microorganisms, it also affects biological treatment processes in the plant. Phosphorus is presented in water, either in dissolved form (phosphates or polyphosphates) or in suspension. In urban wastewaters, possible sources of phosphorus are detergents and septic tanks. Food and agricultural industry activities may also discharge phosphorus in wastewater.

There are also other parameters, which are constantly monitored in WWTP with the objective to characterize wastewater quality. Parameters such as the total organic carbon (TOC), conductivity, amounts of sand and fats, chlorides, are many times included in WWTP's water quality monitoring programs.

In conclusion, there are numerous interactions and possible pollutants in WWTP, which increase the complexity of control and monitoring activities. Therefore, the demand for chemometric methods to summarize and interpret simultaneously all variables and their interactions, is high and can significantly facilitate monitoring by describing and predicting various processes. The development of chemometric methods would be valuable for wastewater quality control, providing practical and effective applications in WWTP.

2.1.9 Monitoring techniques for drinking and wastewater water quality

In order to comply with the legislative norms, a constant quality control of the raw water, of the DWTP processes, of the water quality in WDS, and of the WWTP processes is required. A large number of environmental pollutants, DBP compounds, and water quality physicochemical parameters need to be permanently monitored. The most common techniques for monitoring include a comprehensive analysis using (a) standard laboratory analytical methods for measurements, (b) sensor measurements, and (c) sophisticated analytical instruments.

Physicochemical water quality parameters, such as water temperature, turbidity, salinity, TOC, conductivity or pH, can be measured using a sensor detection methods or applying a standard analytical methodology, which are usually not expensive and do not demand trained staff. On the contrary, the analysis of compounds, such as pesticides, pharmaceutical formulations, endocrine disruptors, DBPs and others, require the employment of sophisticated instrumental techniques, which can detect low concentrations (in pg/L) of them. However, such analysis turns to be expensive and only highly trained personal are capable of conducting it. The most popular instrumental analytical techniques for these compounds include: (a) a solid phase extraction followed by liquid chromatography tandem mass spectrometry (LC-MS/MS) analysis (Kuster et al., 2008), or (b) using the time-of-flight detection (LC-TOF/MS) technique (Martínez Bueno et al., 2007)

In-situ analyzers and portable water quality devices are generally preferred, when there are large and/or remote areas to be monitored in a short period of time. For example, the field test kit for determination of residual chlorine (based on the N, N-diethyl-p-phenylenediamine colorimetric method) is popular for the control in WDS (APHA 1995). Such tests are cheap and easy to use. A significant advantage of the in-situ sensors is the continuous measurements of water quality parameters providing data at real time. They are usually preferred when a high frequency of measurements is required like in storms, leaching, and chemical contamination events. The use of sensors also is a cost-efficient solution, because they do not require frequent field visits by technical staff but only maintenance. Another advantage of the sensors is that they can be simultaneously operated as automatic multiparametric water quality stations. The trends in the field of sensor monitoring technology are towards the development and

testing of new probes, new data recorders, and new telemetry equipment, which are likely to facilitate the monitoring of new parameters of water quality (WMO, 2013).

Recently, technologies, which are based on ultraviolet absorbance (UV), have become more popular, especially for applications where continuous monitoring of organic compounds in drinking and wastewater is required (Langergraber et al., 2004; Rieger et al., 2006). Furthermore, UV can be adapted for use in in-line monitoring instruments. Modern signal processing and high-technology optics have enabled the monitoring of a large number of chemical compounds in a single measurement. For example, nitrites, nitrates, and organic matter were monitored by UV online sensing. Because most organic compounds in raw water, drinking water and wastewater absorb UV radiation, the UV spectroscopy become an inexpensive alternative to other sophisticated instrumental techniques. Furthermore, sensors have another advantage: they can generate a large amount of data in a short period of time. The water quality data, which is recorded with sensors, can be further used for multivariate analysis of all parameters altogether. This data-collection strategy can be very useful in obtaining information about the evolution of the water quality over time, determining the spatial variability (i.e., sampling locations) of water quality in DWTP, WWTP or WDS. Moreover, multivariate analysis of water quality can be applied for real-time event detection, thus providing early-warning signals. Additionally, real-time multivariate models permit the generation of water quality predictions at different locations where a sensor or a multi-parametric automatic station is located.

2.1.10 Taste and odor

The classical physicochemical procedures of water treatment in DWTP serve to disinfect raw water and to eliminate pathogens. New technologies are continuously implemented with the objective to further improve raw water quality (Raich-Montiu et al., 2014). More specifically, membrane filtering (EDR and RO) methods can remove efficiently almost all potential contaminants and organic matter (Valero and Arbós, 2010). An inconvenience of all these procedures, however, is the significant effect which they may have on the organoleptic water properties. Supplied drinking water should not only comply with the legislative sanitary norms, but also it has to take into account consumers for its aesthetical characteristics, including taste, odor, and color.

Re-mineralization procedures are further required (Vingerhoeds et al., 2016), because the concentration of the main minerals decrease significantly during membrane filtering procedures, leaving water tasteless, odorless, and with worse health properties (e.g., due to the lack of minerals, which have vital nutritional effects). Therefore, re-mineralization provides consumers with the required concentration of a number of main salts and also reducing the probability of pipe corrosion. Remineralization takes place in calcite filters and in contact chambers (García et al., 2015), where the osmotic water passed through a bed of calcite (calcium carbonate, CaCO3). During this process, there is a possibility to add carbon dioxide (CO_2), and thus favor the dissolution of calcite required for the optimal remineralization level.

For several decades, the focus in the potable water production sector was on the water taste improvement. In 1980, a specialized "Off Flavors in the Aquatic Environment" group was created to deal with possible issues regarding water taste. Recently, a new specialized group called "Tastes, Odors, and Algal Toxins in Drinking Water Resources and Aquaculture" was formed by the International Water Association (IWA) in compliance with the guidelines developed by the first group in1980.

Water taste strongly depends on the chemical composition of dissolved minerals. Both cations and anions contribute in different ways to the formation of water taste. They can interact among themselves through synergism and antagonism (Burlingame et al., 2007). Apart from dissolved inorganic salts (i.e., total dissolved solids, TDS), volatile organic compounds also affect water taste. Such compounds can also affect retro-nasal mechanisms when drinking water (Dietrich, 2009). Therefore, the general perception of water, especially for drinking water, seems to be flavor rather than taste (Dietrich, 2006). The mineral content of bottled natural water is determined by the composition of rocks and geochemical processes (van der Aa, 2003). Similarly, the potable tap water is also characterized by its specific mineral and organic contents (Meng and Suffet, 1997), where the major source of raw water is the surface water influenced by the local vegetation. Finally, disinfection procedures with chlorine can also contribute to the ultimate flavor of drinking water.

TDS has become the most common and monitored parameter in water taste studies (Burlingame et al. 2007; Whelton et al. 2007; Devesa et al. 2010; Gallagher and Dietrich, 2014; García et al. 2015). For this reason, TDS has been regulated by several international legislative associations: (a) In the U.S. and Canada, USEPA (2015) and

Health Canada (2012) set the maximum level of TDS to 500 mg/L; (b) WHO Guidelines regulated the maximum levels of 1000 mg/L (WHO, 2011); and (c) In Europe, a TDS level of 1600 mg/L the established maximum (98/83/EC). It was shown that high levels of minerals (i.e., high concentration TDS) are disliked by the consumers (Teillet et al. 2010). Some specific salts were also shown to affect the water taste (Burlingame et al. 2007). These studies demonstrated that cations and anions are likely to interact among themselves, thus increasing the complexity of water taste even further. Rey-Salgueiro et al. (2013) focused their research on bottled water taste suggesting a new wheel of water descriptors especially modified for this type of water.

Barcelona Water Panel

Over the past three decades, the AGBAR Company has significantly advanced in both, research and control of the taste and odor of supplied drinking water. In 1989, the Barcelona Water Panel was established (Devesa, 2004). The participants in the panel are women and men (panelists), who are highly trained in the organoleptic analysis of natural and drinking water. Usually, small groups of 5 to 6 panelists participate in a taste- and-odor session. Following water tasting, the panelists would be required to share their opinion regarding the water taste. The panel would analyze the results using the modified Flavour Profile Analysis method (FPA), which is based on an individual explanation of the taste and odor of the sampled water, using series of descriptors and their intensities (APHA, 2005). The panel uses the wheel of water descriptors (see Figure 6), which has been modified by AGBAR so that to meet the goals of thier own water quality studies. In implementing this method, the results can be standard, because all panelists share a common language to describe the water taste. After data collection, a statistical analysis of the collected results follows.

The water taste and odor wheel contains four tastes (sweet, salty,bitter and acid) and all major odors and senstations. The area of odor is presented with the highest number of descriptors, divided in two subgroups: natural descriptors (soil, coolness, decay) and chemicals (chlorine, farmaceuticals, detergents, dissolvent, rubber). However, a further break-down to subdivisions is also possible and acceptable.

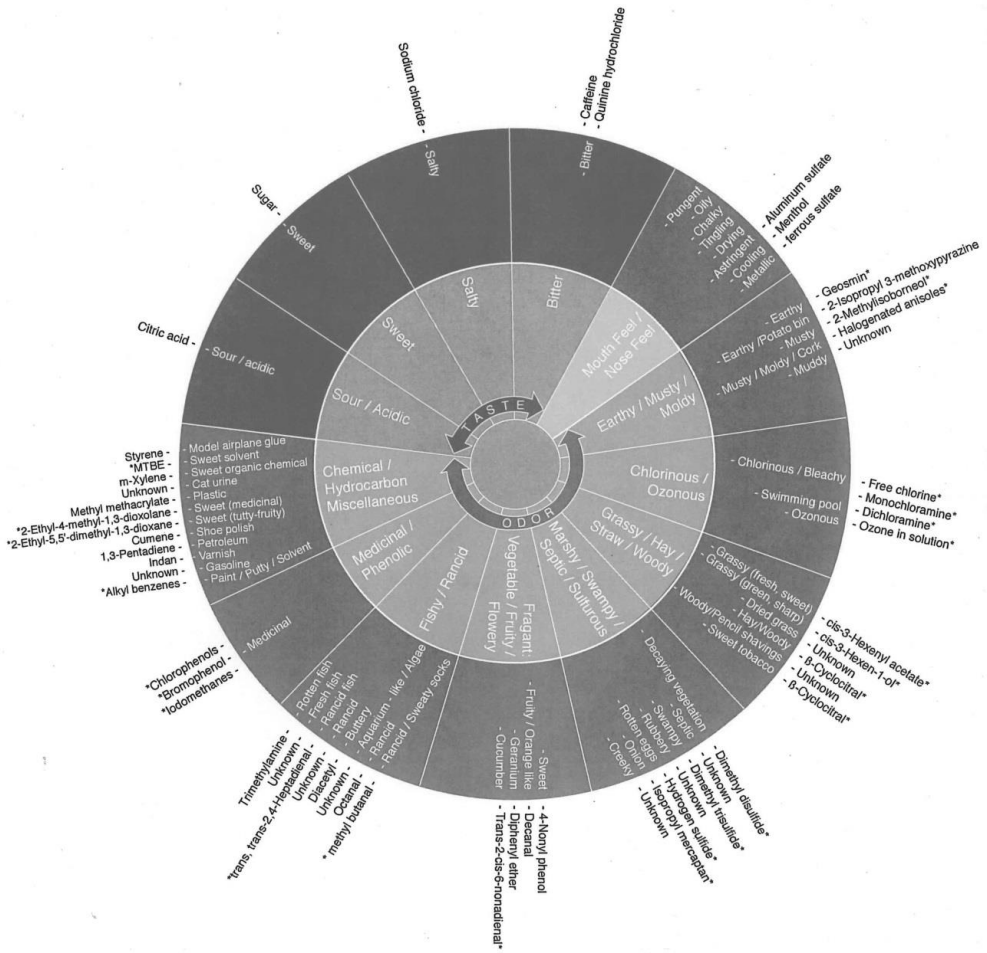


Figure 6. Wheel of water descriptors in order to perform organoleptic description (Suffet and Rosenfeld, 2007)

The water odor description is usually done in closed, glass Erlenmeyer flasks, served at 45 ºC. For taste sessions, water samples should be served at 25 ºC. The intensity usually is spanned from 1 (not detectable) up to 12(maximum intensity), but modifications are allowed.

The taste and odor sessions are organized in rooms, which are specially designed to provide comfort to the panelists. For instance, the water-tasting room has to be free of external odors and with sufficient individual space to each participant, because (s)he

would first work individually, before sharing opinions and discussing results with the rest of the group at the end of the water-tasting session.

There are different designs of the taste and odor experiments (Naes and Risvik, 1996). These discrimination tests include:

- Paired test, where several samples are presented in duplicates and the panelists have to pair the same water samples;
- Triangular test, where three water samples are provided to the panelists, with two of three being of the same origin and one sample being of a different origin. The task of the panelists is to identify the distinct sample;
- Double-triple test, where water samples is given as reference samples. Then, two water samples are served and one of them is the first reference sample. The task of the panelists is to detect the distinct water sample;
- Two from five tests, where two out of five water samples are the same and the other three samples have the same origin. The task of the panelists is to identify both groups;
- Yes/No test, where a reference water sample is presented to the panelists, who have to identify it in a group of different samples.

Another type of taste-and-odor tests of water samples are the tests of acceptance such as:

- Ranking test, where the panelists have to rank the proposed water samples according to their preferences
- Rating test, where the panelists have to score the presented water samples using a scale from 0 (worst) to 10(excellent)

In this Thesis, the rating test is used in developing chemometrics applications, because it provides a variable (i.e., panelists' ratings) which can be directly related to water minerals of the selected waters.

2.2 Chemometric methods

Svante Wold and Bruce R. Kowalski introduced the term of '*chemometrics*' for the first time in 1972 (Otto, 1999). In general, chemometrics is a chemical discipline that uses common methods from mathematics, statistics and information technology (Massart et al., 1988). For the last forty years chemometrics has become an important field of investigation in analytical chemistry, dealing with techniques such as UVVIS, NIR, Raman, nuclear magnetic resonance, fluorescence spectroscopy, chromatography and etc. (Brereton, 2003). The growth of chemometrics applications has become substantial over the last two decades handling large datasets acquired from environmental, pharmaceutical, food chemistry and other fields. Chemometrics has also many applications in analysis and control of industrial processes (Bakeev, 2005).

One major benefit of chemometrics is the possibility of obtaining useful information from raw data. Water quality monitoring involves the application of numerous laboratory instruments and sensors, which have become more and more complex and sophisticated in data acquisition. These instruments and sensors have led to the accumulation of raw data, which require a reliable interpretation, and hence, the application of chemometrics to extract the most relevant information.

In this Thesis, different chemometric techniques and methods from different areas, such as experimental design, exploratory data analysis and multivariate calibration, have been applied in several real case water quality studies.

2.2.1. Data structure

A successful application of chemometric methods requires having the experimental data properly structured.

Experimental data sets, analyzed in this Thesis, have been organized in two dimensional data tables, where in the rows are samples or observations and in the columns are usually a set of measured variables for each sample. These tables are also called data matrices, such as this one shown in Figure 7. Each data matrix has two directions or modes (two-way or two-mode data set).

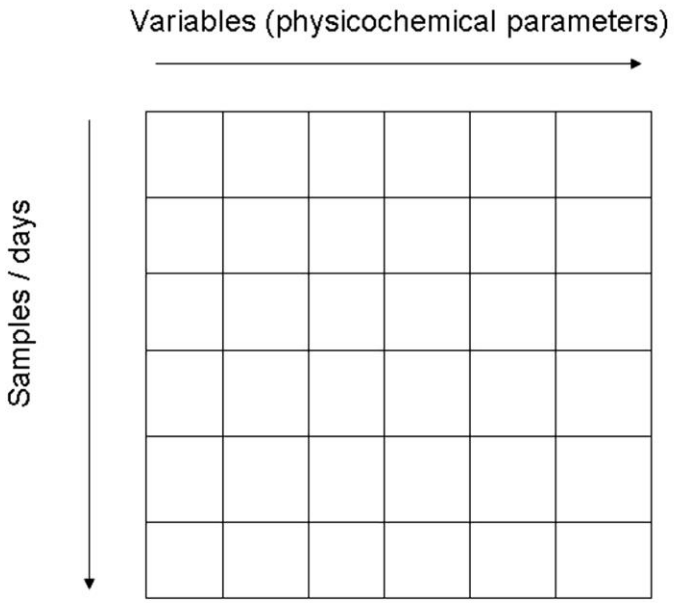


Figure 7. Structure of two-way experimental data, included in papers 1 and 2 from this Thesis. Samples at different days of measurements of multiple operational plant parameters in Sant Joan Despi DWTP.

The rows of these tables (samples) usually represented time series data (daily, hourly, weekly measurements over large period of time), or spatial distribution (samples taken at the inlet-outlet of WWTP) or batch measurements (experimentally designed mixture samples). Columns (variables) have measurements about certain set of physicochemical parameters, spectral absorbance or taste evaluations for a number of panellists.

When two or more data tables have the same number of rows or have the same number of variables, they can be appended in a row- or column-wise direction. Such a data manipulation could increase the possibility of richer interpretation (Måge et al., 2012). The combined use of two or more data sets from different analytical methods gives also a more robust strategy for data analysis than the use of a single data matrix.

In Paper 6 of this thesis, for instance, two tables (one containing measurements of physicochemical parameters – **A,** and another containing mean scores of panellists preferences – **B,** for the same set of water samples) were being concatenated in the row-wise direction as shown in Figure 8. This initial augmentation made possible the

application of PCA to relate the individual panellists' taste preferences to particular physicochemical parameters.

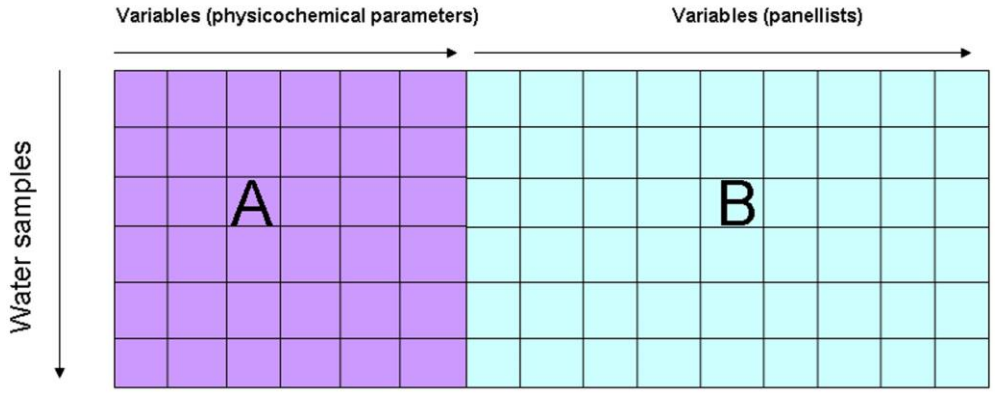


Figure 8. Row-wise augmented data set after concatenation of data matrix A (20 watere samples x 14 physicochemical parameters) and data matrix B (the same 20 water samples x 17 mean score panellists vectors).

2.2.2. Initial data treatment

In order to improve subsequent chemometric analysis and its proper interpretation, preliminary data treatment usually is required. Such data pre-treatment usually involves a process following various steps. The initial step of every data pretreatment is related to problems associated with data collection loss or absence of certain data values, which are called, missings (Stanimirova, 2013). Values below the instrumental limit of detection are considered censored data. The second step of data pretreatment includes application of different mathematical techniques to improve the quality of the measured signal (data). Properly applied pre-treatment procedures would allow for a better performance of the selected chemometrics methods and further improvement of results interpretation.

Missing data (missing values)

Missing values and censored data occur often in the analysis of multivariate environmental data sets and also have been frequent problems in this Thesis. In general, missing values and censored data encountered in data sets included in this Thesis could be characterized according the following situations (Rubin, 1976)

- Missing values occur Completely At Random, MCAR

- Missing values do not occur at random, NMAR
- and combinations from both above

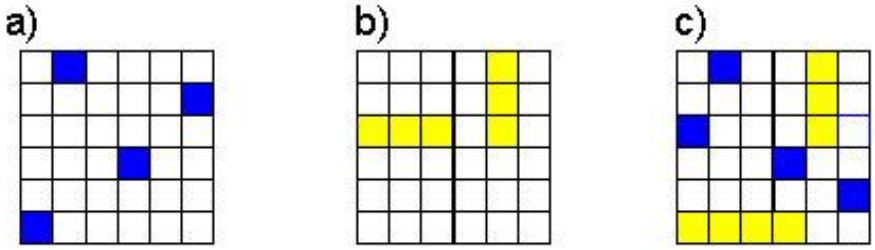


Figure 9. Graphical representation of missing values common for this Thesis. a) Missing values are distributed completely at random trough the table; b) missing values are distributed not at random; c) combination of both.

MCAR type of missing (see Figure 9a) gives independent observations of all variables. In this case, missings are distributed randomly without forming any particular pattern (Fig. 9a). MCAR observations observed in data sets from this Thesis were missing measurements for a particular physicochemical parameter monitored in the DWTP plant. Some parameters were not measured at specific time due to a malfunction of analytical instruments, or due to the high cost of a particular measurement.

In contrast, when data sets include incompletely observed variables and censored data (values below LOD), it is observed NMAR type of missing (see Figure 9b). In such case, NMAR values typically form distribution patterns. This is the case for example of values of particular physicochemical parameters below the instrumental limit of detection (LOD) for a long time. NMAR was also observed in the sensory analysis study, presented in paper 6. Some panellists could not attend a particular taste session. As a result, their rates for 5 water samples, presented in this taste session, were missing. Also, missing values of UVVIS spectral measurements in paper 5 of this Thesis were observed. The optical probe failed to take records when spectral readings were out of the detector range or due to an expected electrical failure for a long period of time.

Experimental data may contain either elements missing completely at random (MCAR), or not (NMAR), but they often are present simultaneously in concatenated data sets (Figure 9c).

Many methods for missing data imputation exist and many of them are based on iterative algorithms (Walczak and Massart, 2001a, b).

In this thesis, an imputation function, based on Principal component analysis (PCA) method (Jollife, 2002) for the estimation of the missing values of MCAR type, was used. In this imputation, initially the estimation was performed replacing empty values by zeroes and then performing PCA. Subsequently, zero values are replaced by the new predicted values by PCA with a determined number of components, and with these new predicted values, again, a new PCA model is recalculated. This process is repeated until the empty values estimated by PCA converge – giving the same values.

Measured values, observed below the LOD in this Thesis are considered of NMAR type. They were replaced with the half of their detection limit. This has been shown to be a proper method to deal with up to 30% below LOD values (Jain and Wang, 2008).

When missing values are observed in large regions of UVVIS spectral data, or when panelists were not attending at a particular taste session, the substitution method was rather subjective and it was decided in function of the size of missing data. Entire rows (entire spectrum) or columns (panelists evaluations for different sessions) were eliminated from data analysis. In cases of recorded spectral data with small number of missing values, PCA imputation method was preferred.

Data preprocessing

The application of data pre-processing is required in order to facilitate the extraction of useful information from experimental data. There is no single, universal method of pre-treatment to be generalized for all environmental data sets. The selection of the most appropriate pre-treatment techniques depends on several factors related to the data set structure, to the used techniques of measurement, to the type of information that is investigated, or to the prior knowledge about the data nature. With other words, the choice of the most appropriate method is subjective and depends mainly on the type of environmental problem or on the phenomenon under study. However, some recommendations can be given

Data pre-treatment methods, most frequently applied throughout this thesis have been autoscaling and meancentering procedures. Scaling methods were required since in all studies physicochemical parameters, UV spectral data or panellists rating data had differences in their units and scales.

- *Meancentering*

This method subtracts the total mean from a variable (in column direction) for each observation. Thus, all mean-centered variables have zero means:

$$z_{ij} = x_{ij} - \bar{x}_j \qquad i = 1,...,I \qquad j = 1,...,J \qquad \bar{x}_j = \frac{\sum_{i=1}^{I} x_{ij}}{I} \tag{7}$$

where z_{ij} is the new meancentered value, x_{ij} is the value of the sample *i* of the original variable *j* ; $\bar{x}_j$ is the average of values for the original variable j.

Meancentering performs a translation of the origin of coordinates from zero to the mean value of the data, thus variations from the mean are easier to visualize. This part of information, which does not change and is constant for the data matrix, is discarded. On the contrary, the preprocessed data emphasize the information related to the variance. Meancentering adjusts all values to vary around zero instead of around their mean. One drawback of this method is that it is not adequate for data with sub-populations with different variability (for example parameters measured in different units). Meancentering, then, is usually used in combination with other scaling methods.

- *Scaling*

Scaling is frequently used in environmental studies as a pre-treatment method. Its application is required when data sets consist of variables measured in different units, like data sets with physicochemical parameters such as pH, conductivity measured in S/m, concentrations of ions measured in mg/L and other parameters. Scaling is performed as follows:

$$z_{ij} = \frac{x_{ij}}{s_j} \qquad i = 1,...,I \qquad j = 1,...,J \tag{8}$$

where z_{ij} is the new scaled value and s_j is the standard deviation of all values of variable *j* .

This method divides all values of a variable by its standard deviation. The goal is to obtain variables with homogeneous distribution in order to be easily compared. Every variable has standard deviation of 1 after scaling. Thus all scaled variables in the final data set have the same relative importance.

- *Autoscaling*

This pre-processing method is a combination of meancentering and scaling. It is well–known in chemometrics. It consists of mean-centering followed by scaling of all variables, both in column-wise direction.

$$z_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j} \qquad i = 1,...,I \qquad j = 1,...,J \tag{9}$$

where z_{ij} is the new autoscaled value, x_{ij} is the value of the sample *i* of the original variable *j* ; $\bar{x}_j$ is the average of all values of variable *j* and s_j is its standard deviation.

As a result, the mean of the variables is zero and their variance is adjusted to one. Thus, this method adjusts the independent variables to similar variance and makes them comparable. The distribution of the values of variables obtained after applying autoscaling is similar to the case of scaled but at the same time, variables experience a translation of their origin due to their meancenter.

One possible pitfall of autoscaling is that sometimes meaningless variables can appear to be important just because scaling noise make them larger. In the framework of this thesis, autoscaling was used thoroughly as preliminary pre-processing technique for initial data screening and for more advanced chemometric analyses. Specifically, autoscaling was the preferred data pre-treatment for PCA, because in this case the main focus was on the description and investigation of data variance.

2.2.3 Experimental Design

Experimental design (Fisher, 1971) or *design of experiments* (DoE) is an important area of chemometrics. In a particular laboratory experiment, one or more variables (or factors) are usually changed to observe the effect that these changes have on one or more response variables (Martens and Naes, 2001). Using DoE appropriately the obtained data will lead to valid conclusions.

In this Thesis, Paper 3 includes a DoE study to assess the effect of several important factors on trihalomethanes (THMs) formation in water treatment process. The investigated factors were the concentration of different DOM fractions, chlorine disinfectant dose, pH, bromide anion concentration and temperature. DoE offered the possibility to avoid expensive numerous experiments in order to obtain an optimal combination of investigated parameters. DoE designed a procedure in which the overall number of measurements was significantly reduced, thus, it resulted to be less

expensive, time efficient method. DoE resulted to be a very useful and effective approach compared to traditional techniques of analyzing one single variable at a time (OVAT). In summary, the benefits from DoE over OVAT in this study were:

1. DoE permitted to investigate several variables simultaneously and take into account their interactions under investigation (OVAT could not);
2. DoE provided a global knowledge (in the whole experimental domain), while the OVAT would give only a local knowledge (only where the experiments have been performed);
3. The quality of the information obtained by DoE in each point of the experimental domain was higher.
4. The number of experiments required by the experimental design was optimal at reduced time.

As a first step of DoE, the objective function should be defined. In our case of study, the objective was to investigate the conditions, which favor the formation of target disinfection by-products (particular THMs). The second step included the determination of the set of factors (such as temperature, pH, organic matter concentration and etc) that affect the target and to select the most important of them for the study. The third step included the selection and planning of the experimental design (DoE permits laying out a detailed experimental plan in advance of doing the experiment). The fourth step included performing the experiment and the fifth step was analysis of the results using statistical techniques such as the analysis of variance (ANOVA), multiple linear regression (MLR) and response surface strategies.

Two designs - the Plackett–Burman (AMC, 2013) and the Box-Behnken design (Box and Behnken, 1960), have been used in Paper 3 in order to analyze the effects of factors such as concentration, temperature, pH on THMs formation.

Plackett–Burman screening design (PB) was chosen as a very economical DoE approach for preliminary screening among several factors with the goal of selection of the most influential. PB gives information only on the effects of single factors with two levels: high level encoded as +1 and low lever encoded as –1. PB allows studying the main effects of a large number (n) of variables with no more than n + 4 experiments. This method is well suited to establish whether the formation of a particular THM was affected by changes in the investigated factors without having preliminary knowledge. In Paper 3, five parameters (concentrations of organic matter, bromide ion, chlorine,

temperature and pH) were initially being considered, resulting with just eleven experiments to be carried out. The major drawback of PB design is that it does not account for the interactions between factors and therefore a new design for a more detailed factor investigation is needed.

A Box-Behnken design (BB) was chosen as a response surface design. The Box-Behnken design enabled precisely to study the effect of the selected 3 factors -organic matter concentrations, chlorine and temperature, as well as to obtain response surfaces with a relatively few number of experiments. Three levels for each of the factors were used. It was especially useful to study the quadratic behavior of the factors. One advantage of BB was that extreme combinations were avoided. The corner points of the design are extreme points in terms of design region (concentration, or temperature range) where the experiment is performed (see Figure 10a). Also, BB has the possibility of a detailed assessment of the data information using the response surface methodology (RSM). The main goal of RSM was to visualize the surface of response of the selected parameters (see Figure 10b) in order to quantify their relationship and their response surfaces (Kwak, 2005).

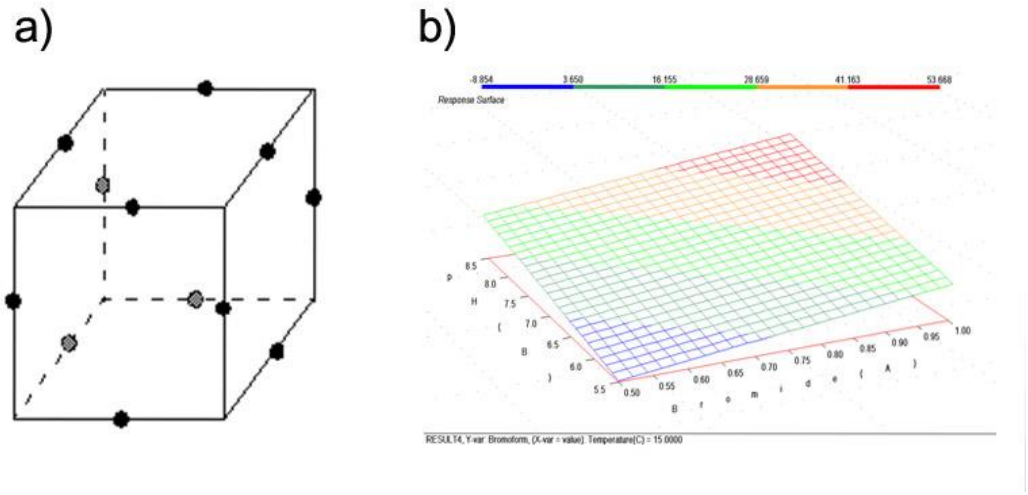


Figure 10. a) Graphical representation of Box-Behnken design; b) Response surface methodology plot of CHBr3 formation versus pH and bromide.

2.2.4 Principal Component Analysis

Principal Component Analysis (PCA) (Jollife, 2002) is probably the most used method in multivariate statistical analysis of data in the laboratory and in environmental monitoring studies.

PCA is based on the hypothesis that there are a small number of dominant factors (components) in the original data set with significant influence and which present the main sources of data variation. These factors cannot be measured directly and usually are called hidden factors, since they cannot be directly experimentally observed.

Generally, a large part of the information contained in the experimentally measured variables is redundant (instrumental noise, natural variance, variables correlated with other variables and etc.) and therefore irrelevant for the problem under study. The purpose of PCA is to find and extract a new set of orthogonal coordinate axes called principal components (PCs), based on the linear combination of the original variables. The projection of the original data onto these new axes provides a better and easier interpretation of the underlying phenomena or sources that are causing the observed data variance (see Figure 11). One advantage of PCA is that these new principal components are orthogonal. This means that the variance (information) explained by one principal component is different to the variance explained by another principal component, avoiding the overlapping of the information. In addition, the first component is calculated in the direction that explains the largest amount of variance, the second component the same for the residual variance, and successively the same for the following components.

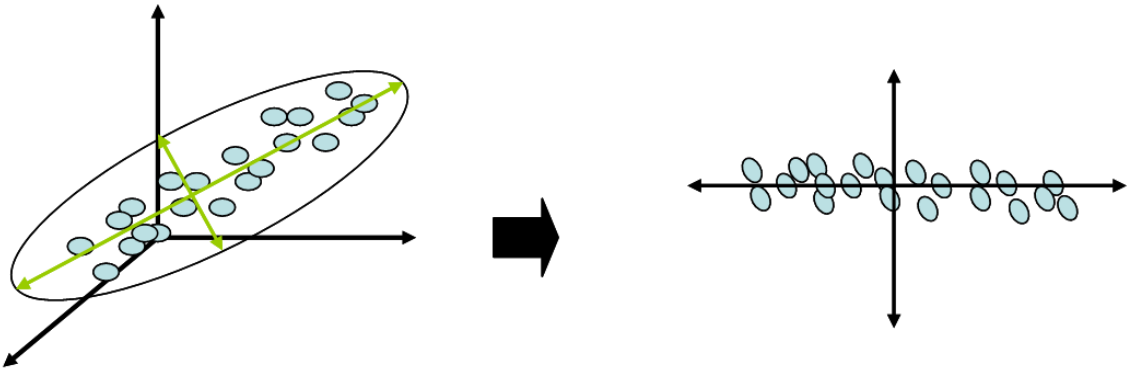


Figure 11 . Graphical representation of dimension reduction in PCA. New orthogonal axes calculated as linear combination of original variables.

In matrix notation, the original data matrix is decomposed using a bilinear model, giving the product of two orthogonal matrices, **T** and $\mathbf{P}^T$

$$\mathbf{D} = \mathbf{TP}^T + \mathbf{E} \quad (10)$$

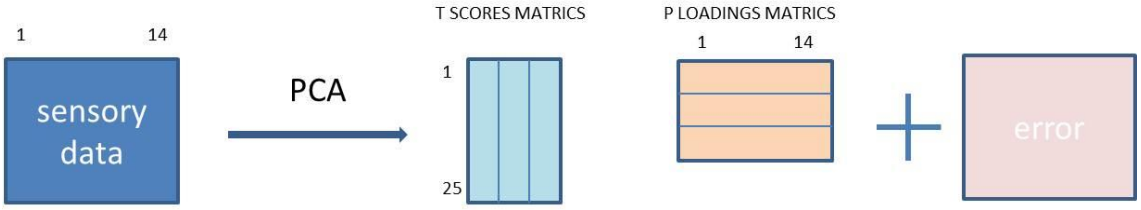


Figure 12. PCA decomposition of sensory data matrix using three principal components

where, **D** (for example – a sensory data containing fourteen panellist evaluations for a set of 25 water samples, Figure 12) contains the original experimental data; **T** is the *scores* matrix (map of samples), $\mathbf{P}^T$ is the matrix of *loadings* (map of the variables) and **E** is the matrix of residuals. The product of the *scores* and *loadings* matrices for a defined number of PCs gives back to the original data matrix after noise reduction.

Also, it is possible to write the PCA decomposition of **X** as the sum of a number of $\mathbf{t}_i$ and $\mathbf{p}_i$ vectors, where *r* is this number or the rank of the data matrix **X,**

$$\mathbf{X} = \mathbf{t}_1\mathbf{p}^T_1 + \mathbf{t}_2\mathbf{p}^T_2 + ... + \mathbf{t}_k\mathbf{p}^T_k + ... + \mathbf{t}_r\mathbf{p}^T_r + \mathbf{E} \quad (11)$$

r is usually much lower than the smaller dimension of **X** (rows or columns). $\mathbf{t}_i$, $\mathbf{p}_i$ score and loading vector pairs are ordered by the amount of captured variance. $\mathbf{t}_i$ score vectors give information on how the samples relate to each other. **p**i loading vectors give information on how the variables relate to each other.

The two most commonly used methods for data matrix decomposition are 1) the eigenvector decomposition of the covariance or correlation matrix and 2) the singular value decomposition or SVD of the data matrix.

- *Eigenvector decomposition of the covariance or correlation matrix*

The classical algorithm of data decomposition by PCA is based on the eigenvector decomposition of the covariance (when data have been mean-centered) or correlation matrix (when data have been autoscaled) of the variables. It is a relevant algorithm

when the number of samples is huge and the number of variables is small. For a given data matrix **X** with *m* rows and *n* columns, the covariance matrix of **X** is defined as:

$$\mathrm{cov}(X) = \frac{X^T X}{m-1}, \quad (12)$$

In the PCA decomposition, the $\mathbf{p}_i$ vectors are called *eigenvectors* of the covariance matrix; that is, for each $\mathbf{p}_i$:

$$\mathrm{cov}(\mathbf{X})\mathbf{p}_i = \lambda_i \mathbf{p}_i , \quad (13)$$

where λ_i is the associated eigenvalue to the eigenvector $\mathbf{p}_i$. The scores then are calculated as linear combination of the original X variables defined by the corresponding eigenvector:

$$\mathbf{X}\mathbf{p}_i = \mathbf{t}_i , \quad (14)$$

The *scores* $\mathbf{t}_i$ are just projections of **X** onto the $\mathbf{p}_i$ (Wise et al., 1990).

- *Singular value decomposition*

The most frequently used algorithm to decompose the original data matrix is the singular value decomposition or SVD (Golub and Van Loan, 1996). This data matrix decomposition is described by the following equation:

$$\mathbf{D} = \mathbf{U}\,\mathbf{\Lambda}^{1/2}\,\mathbf{V}^{T}, \quad (15)$$

where $\mathbf{\Lambda}^{1/2}$ is the diagonal matrix with singular values as diagonal elements (square root of eigenvalues).

The link between the eigenvector decomposition of covariance matrix in PCA and SVD of the original data matrix can be figured out as follow:

$\mathbf{D}=\mathbf{T}\mathbf{P}^{\mathbf{T}}$ (for eigenvector decomposition) (16)

$\mathbf{D}= \mathbf{U}\,\mathbf{\Lambda}^{1/2}\,\mathbf{V}^{T}$ (for SVD decomposition), and consequently:

(14)

$$\mathbf{T}= \mathbf{U}\,\mathbf{\Lambda}^{1/2} \quad (17)$$

$$\mathbf{P}=\mathbf{V} \quad (18)$$

In PCA modelling, the most important step is to select a reduced number, *r*, of the most important principal components explaining meaningful information (variance). The remaining components (usually explaining only a small amount of the total data variance) give the residual PCA unexplained data matrix **E**. Thus, the main advantage of PCA is to summarize only the relevant information that is contained in the original data matrix (eliminating the corresponding noise, error and natural variance).

The selection of a number of components, throughout the studies of this Thesis, was done mainly by observing the size of the eigenvalues associated with these components. The eigenvalues are the square of the singular values found by SVD, and they reflect the amount of variance explained by each new component. So, the first eigenvalues are larger, and the rest of components vary very little for every new component. The number of components was chosen in a way that the addition of a new component provided additional relevant information in the context of the problem, or on the contrary, it was discarded and considered that it explained only experimental noise. The selected number of principal components has to be related to the number of sources and patterns of lineally independent variation present in the analyzed data.

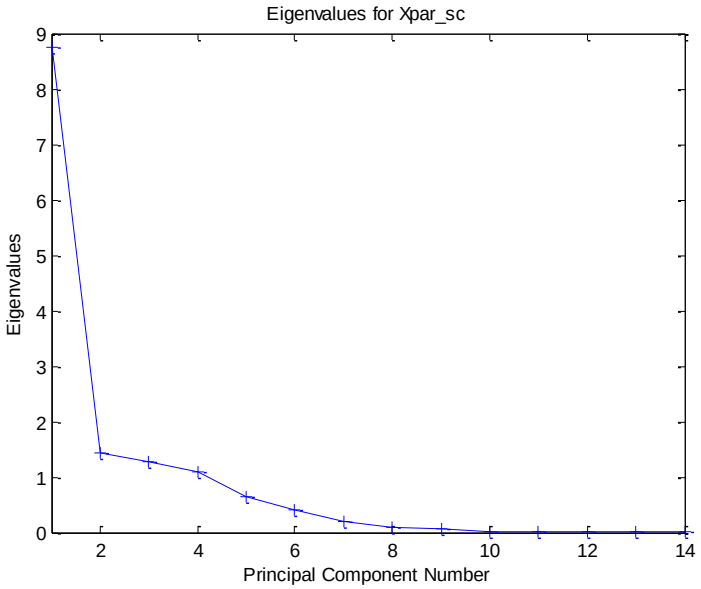


Figure 13. Plot of the eigenvalues versus principal components calculated for the water physicochemical parameters data example.

For example, Figure 13 reports eigenvalues assigned for every PCs in the decomposition of the pre-treated original data matrix containing physicochemical parameters, measured for a set of water samples (analysed data in Paper 6 , Material and Methods section). The selected number of components was four in this case. The 2nd, 3rd and fourth PC explain rather similar amounts of variance and since all al them presented an associated eigenvalue larger than 1 (with explained variance higher than the average per component), a model with four PC was selected.

From the fifth component, the eigenvalues decreased smoothly and changed very little. Depending on the magnitude of the explained variance and on the individual contribution of original variables on each PC, it is possible to deduce the importance of the various environmental components, and also it is possible to determine the noise level of the experimental data. It is also useful to investigate the shape (profile) of the *score* and *loading* vectors.

Loadings values ($\mathbf{P}^T$ matrix of *loadings*) indicate the magnitude of the contribution of every original variable to every principal component. In case of water quality data, *loadings* will indicate the chemical composition or chemical profile of the identified sources. Variables with large values of *loadings* on the same component are assumed to correlate. If they present the same sign, they present a positive correlation. On the contrary if they present opposite signs, they correlate inversely (negatively). On Figure 14, two groups of inversely correlated physicochemical parameters are shown. Chloride, sodium and silica ions are positively correlated between them, but they are inversely correlated to pH, magnesium, calcium, bicarbonate and nitrate ions, which on the other side correlate positively between them.

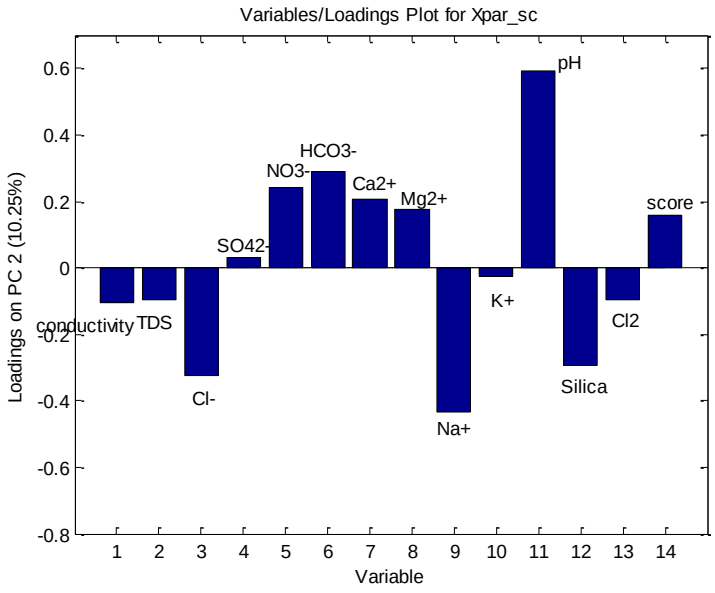


Figure 14. PC2 loadings plot for the water physicochemical parameters data set.

The projections of the samples in the new space defined by the new principal components give the *scores* (**T** *scores* matrix). According to their distribution it is possible to group some samples according to the similarity of their *score* values. Also it is possible to elucidate the presence of samples with extreme *score* values. In the water

quality studies, the scores contained information about the water samples' spatial distribution, time distribution or about the different water origins, as hidden patterns in the data. For example in Figure 15, two sets of samples – bottled waters in red triangles, and tap waters in green asterisk, are easily distinguished by their visualization of the PC1 versus PC3 space.

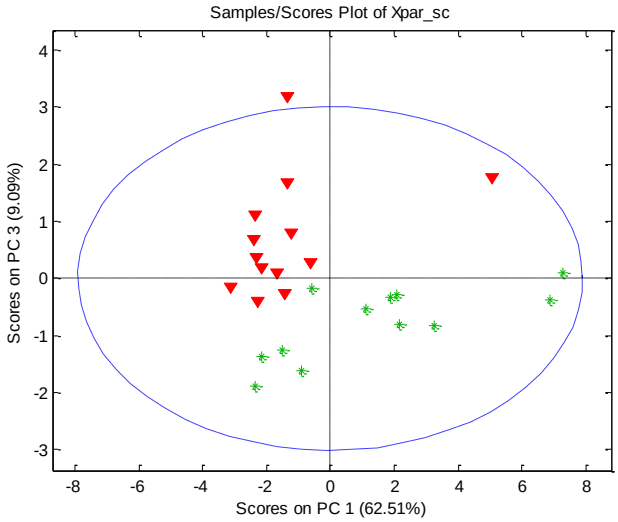


Figure 15. PC1 versus PC3 loadings plot for water samples presenting two types of water – bottled mineral (red triangles) and tap (green asterisks).

Generally, for a correct interpretation of the hidden data patterns, *scores* and *loadings* plots are analyzed together.

To resume the main benefits that can be obtained from the analysis of score and loading plots in PCA, the following aspects are considered

On *scores* plot graphics:

1. The distance among samples indicates similarity. Closely distributed samples show higher similarity than others plotted at a certain distance.
2. It is possible to detect clusters of samples that show higher similarity among themselves.
3. Using external information (meta-analysis), it is possible to distinguish the origin of these clusters (natural origins, chemical composition, physicochemical characteristics and etc.)
4. Samples that are distributed far away from the centre of the plot can be considered as extreme observations or outliers.

On the *loadings* plot:

1. The most important variables for the model should show larger loading absolute values.
2. Variables close to the centre of the plot (close to zero on the two PC axes) do not contain relevant information to describe variance patterns.
3. It is possible to detect direct or inverse correlations among variables

2.2.5 Multivariate calibration. Linear regression methods

Multivariate Calibration (Martens and Naes, 1991) is a very important area of chemometrics. This Thesis includes various studies which required the application of multivariate calibration methods in order to achieve prediction models. Different concentrations of target analytes such as THMs, TOC and others (as predicted variables) were used in multivariate calibration methods to investigate analytical datasets, containing various quality control parameters (as predictors) such as physicochemical parameters or UVVIS spectral data. Also, the sensorial responses of a group of panellists were related to physicochemical parameters of different sets of water samples. The obtained, regression models could be used next to predict unknown concentrations or unknown sensorial responses from new samples. Generally, these regression models predict values of dependent (predicted) variable, indicated by the vector **y,** using a regression function, which is applied to the variables of the block of independent (predictor) variables, which are collected into the data matrix **X** (Martens and Naes, 1991).

Multivariate calibration finds a mathematical relationship (regression) between these two blocks of variables:

$$\mathbf{y} = f(\mathbf{X}), \quad (19)$$

Among the most used multivariate calibration methods are the multiple linear regression (MLR), the principal component regression (PCR), the partial least squares regression (PLS) and the support vector machine regression (SVR) methods (the last for non-linear type of data).

Multiple linear regression (MLR)

MLR relates the concentrations of a target analyte **y** (for example one of the four target trihalomethanes concentrations, THMs, modelled in Paper 1 of this Thesis) to a

series of recorded DWTP parameters collected in **X** matrix, using a regression vector **b** (Brereton, 2003):

$$\mathbf{y = Xb + E,} \quad (20)$$

where **E** is a residual matrix with the same dimensions as **X**.

Each row of **X** corresponds to a set of DWTP operational parameters measured at different plant locations at define time. Each row of **y** corresponds to the concentrations of a specific THMs target analyte measured at the same time at the exit of plant.

MLR models the relationship between these DWTP predictor parameters and the predicted target analyte values by fitting a linear equation. MLR maximizes the covariation between **X** and **y** to obtain the best estimation of **y**.

The **b** regression coefficient vector is estimated by the equation:

$$\mathbf{b} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}, \quad (21)$$

where $\mathbf{X^T}$ is the transposed matrix of **X** (rows in **X** become columns and vice versa). The "$^{-1}$" indicates that $(\mathbf{X}^T\mathbf{X})$ have been inversed. $(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ is called the pseudo inverse matrix of **X**, because **X** is not square in general and cannot be inverted.

Each b_i regression coefficient represents the change in y_i relative to a one unit change in the respective X_i independent variables. For instance, b_1 is the change in y_1 relative to a one unit change in X_{11}, holding all other independent variables constant with fixed values. Very important step in MLR modelling is to undergo statistical test to assess whether each regression coefficient is significantly different from zero. This process selects the more important DWTP parameters (**X** variables) for the target THM compound formation.

In the second step of multivariate calibration, the prediction of the concentration of the target analyte is possible from a new dataset by multiplying the new dataset by the previously obtained regression vector during calibration,

$$\mathbf{y_{unknown} = X_{new}\, b,} \quad (22)$$

The main drawback of MLR is the co-linearity problem in **X** variables. This happens in situations where some of the variables (columns in **X**) are linear combination

of the other variables apart from the noise. In other words, when a linear dependence exists among the variables (Martens and Naes, 1991). Many multicollinear variables would result in an unstable regression equation (because of the difficulty in precise estimation of **X** pseudoinverse) with a consequent difficult interpretation of regression coefficients and unreliable prediction (Todeschini et al., 2004).

Stepwise Multilinear Regression (SWR)

Stepwise Multilinear Regression (SWR, Draper and Smith, 1981) is a particular case of MLR, based on the forward (or backward) selection, which consists of first classifying the predictor variables according to their statistical significance, and next including one variable at a time at different steps. At the end of the process only statistically significant variables are included to build the model. Again, problems may happen when predictor variables are highly correlated and when there is the possibility of one input variable masking the effect of another input variable. Very often, the achieved models include variables depending on starting choices and insertion strategies (Esbensen et al, 2000). Regression coefficients are obtained using finally selected variables and ordinary least squares estimation.

Principal component regression (PCR)

Principal component regression combines first PCA data compression of the predictor data matrix **X** and a further MLR regression step after data compression (Geladi and Esbensen, 1991). Calculation of the pseudo inverse matrix of the data with orthogonal PCs is improved, rather than with the original X variables. PCR estimates first PCA *scores* (**T**) and *loadings* (**P**), and then then as a second step, MLR is carried out using the following equation:

$$\mathbf{y} = \mathbf{Tq} + \mathbf{e}, \quad (23)$$

where **y** denotes actual values of the predicted variable, **q** are the y-*loadings* for PCR model and **e** is the error vector (unexplained part of **y** by the PCR model). The y-*loadings* **q** are calculated using the least squares approach:

$$\mathbf{q} = \mathbf{T}^{+}\mathbf{y}\ (\mathbf{T}^{T}\mathbf{T})^{-1}\mathbf{T}^{T}\mathbf{y} \quad (24)$$

In case of multicolinearity, PCR provides a very stable matrix inversion $(\mathbf{T}^{T}\mathbf{T})^{-1}\mathbf{T}^{T}$ in contrast to the MLR matrix inversion shown in Equation (21), because the PCA scores are orthogonal to one another.

Regression coefficients can be calculated as follow:

$\mathbf{b}_{pcr}=\mathbf{Pq}$, (25)

and they are used for further prediction of unknown samples.

The main drawback of PCR is the possible data overfitting in case of too many PCs included in the model, leading to a very sensitive model for unforeseen disturbances. The proper selection of PCs includes several model validation procedures (discussed in section 2.2.8). Another drawback of PCR modelling is that it relies on using the principal components as predictors for the responses, but principal components do not necessarily correlate well with **Y**. As possible solution to overcome such a problem, some authors (Mason and Gunst, 1985) suggested selecting only the latent variables, correlating maximally with the responses, or more common in chemometrics using the PLS method (see below)

Partial least squares regression (PLS)

MLR and PCR show some limitations because both assume that there is no error within **y** measurements. Thus a new chemometric method dealing with the experimental noise in **y** such as the partial least squares regression (PLS) method was needed.

The PLS regression (Geladi and Kowalski, 1991) belongs to the family of inverse regression methods in which the calculated model relates the latent variables **X** (matrix of predictor variables) with the **y** (predicted variable), intending to maximize the covariance between **X** and **y**.

The general equation of the inverse regression methods model is defined as:

$\mathbf{y = Xb_{pls}}$, (26)

Where **X** (matrix of predictor variables with dimensions m x n) and **y** (vector of predicted variable with dimensions m x 1) are represented by their latent variables:

$\mathbf{X= T_{pls}P_{pls}' + E}$, (27)

$\mathbf{y= T_{pls}q + f}$, (28)

and **b** (m x 1) is the vector of calculated regression coefficients in the calibration step.

The matrix **W** (weight matrix) in the next equations reflects the covariance structure between the **X** predictors and **y** predicted variable and it is calculated and used in the estimation of the regression vector.

$\mathbf{b_{pls}= X^{+}y}$, (29)

$\mathbf{b_{pls}=W_{pls}(P^{T}_{pls}W_{pls})^{-1}q}$, (30)

Therefore, the **b** regression vector can be expressed as function of **T**, **P**, **W** and **q**.

There are two main versions of PLS models based on the number of predicted variable. In PLS1, only one variable (**y** vector) has to be predicted and in PLS2, several variables (**Y** matrix) are simultaneously predicted. In this Thesis only PLS1 method was used to predict one by one different target properties such as concentrations of individual THMs; the mean liking vector of all panelists ratings of water samples and nitrate and total organic matter concentrations in wastewater samples.

The interpretation of latent variable scores, loading and weights plots are similar to the interpretation of score and loading plots from PCA model.

Several criteria have been followed to build a PLS model:

1. Data pretreatment. This is an important step in order to achieve a good PLS model. Both **X** matrix and **y** vector should be pretreated simultaneously in order to compensate offset differences (meancentering) or differences in the measurement scale of all variables (autoscaling). However, other methods are useful also and depend on the case of study.

2. Selection of the number of latent variables in PLS modeling. One common method to select the number of LVs is the cross-validation in the internal calibration of the PLS model, in which different samples are used to construct the model and to validate it. The optimal number of selected LVs of PLS model has to report the smallest residual variance of **y** in the cross-validation. However for a particular practical application, the best selection of latent variables will be achieved when external validation procedures are used (see section 3.8). In both cases, the percentage residual variance is calculated as:

$$\%\ \text{Residual variance of y} = \frac{\sum_{i,j}\left(\hat{y}_{ij} - y_{ij}\right)^2}{\sum_{i,j} y_{ij}^2} * 100, \quad (31)$$

where $\hat{\mathbf{y}}_{\mathbf{ij}}$ is the predicted value and $\mathbf{y}_{\mathbf{ij}}$ is the actual (measured) value.

Due to the large number of variables (physicochemical parameters or spectra) included in the studies of this Thesis, a variable selection method was needed to identify the most relevant variables to give a better interpretation of the results or to conduct more specific work as it was described in paper number 5 (development of an automatic system with an optical probe based on the measurement of a reduced number of UV wavelengths).

The Variables Importance in Projection method (VIP) is a variable-selection technique, used to summarize the influence of individual X-variables on the PLS model (Wold et al., 2001). VIP is calculated from the weighted sum of squares of the PLS weights, **w***, which take into account the amount of explained **y** variance by each latent variable. VIP scores works as a summary of the variables which contribute to the most of the **y** variance. For a given model and problem there is one VIP-vector, summarizing the importance of **X** variables for the prediction of **y.**

The VIP value for the j[th] variable is given as

$$VIP_j = \sqrt{\frac{\sum_{f=1}^{F} w_{jf}{}^2 \cdot SSY_f \cdot J}{SSY_{total} \cdot F}} \quad (32)$$

Where $\boldsymbol{w}_{jf}$ is the weight value for variable *j* component *f*, SSY_f is the sum of squares of explained variance for the *f*th component and *J* number of variables. SSY_{total} is the total sum of squares explained of the dependent variable, and *F* is the total number of considered components. VIP_j is a measure of the contribution of each **X** variable according to the **y** variance explained by each PLS model. Since the average of squared VIP scores equals to 1, the 'greater than one rule' is generally used as a criterion for variable selection (Chong and Jun, 2005). Figure 16 shows a typical plot of VIP for physicochemical parameters used as predictors (**X** data matrix) in the PLS modelling of overall water taste liking (**y** predicted vector) in Paper 6. Parameters like Cl-, NO3, Na

and others present VIP scores above the threshold value of one and, thus they are considered to be significant for the PLS model.

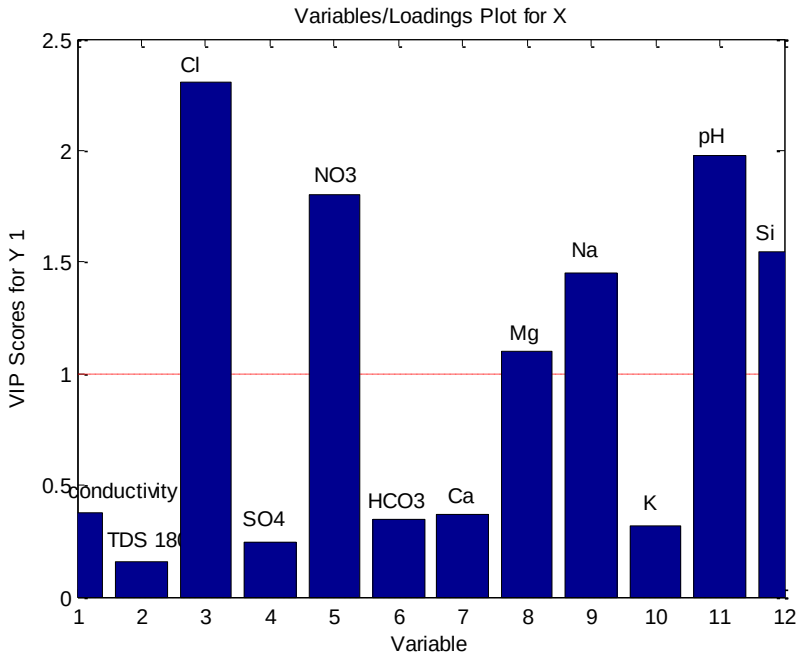


Figure 16. Plot of VIP (variable importance in projection) scores for physicochemical parameters related to the water taste liking. Parameters with VIP scores above the threshold value of one (red dotted line) were considered significant in the PLS model.

2.2.6 Outliers inspection

The detection and elimination of outliers is a crucial step in many circumstances for a proper PCA, PCR and PLS modelling.

A very useful chart for the detection of outliers is presented on Figure 17, where $\mathbf{T^2}$ hotelling values (leverage) presents the sum of the normalized squared *scores* and Q residuals, which is the measure of the difference, or residual, between a sample and its projection on the k principal components retained in the model.

For a suspected sample to be considered as an outlier, it is recommended to observe its leverage in the model and to check if the model explains it correctly.

Leverages of a PCA model indicate how much influence each sample has on the model. The leverage of a sample is related to the magnitude of the score for this sample. Samples with high leverage cause the principal components to rotate towards them. Leverages can be used to find very important observations as well as detecting potential outliers.

Another useful statistics for outlier detection is Q statistics of a sample. The Q statistics indicates how well a particular sample conforms to the model. It gives a measure of the difference, or *residual,* between a sample and its projection on the *k* principal components retained in the model. Samples with very high residual are not well explained by the model.

On the graphics of leverages versus residuals (Figure 17), samples that present a large leverage and a large residual (upper right part of the plot) can be considered to be outliers and should be omitted from the data, and the model should be recalculated. Samples with high leverage but with low Q residuals (bottom right part of the plot), can be considered to be extreme objects and to be the most influential in the model, however, after subsequent investigation of their origins and nature. Samples with high Q residuals and with low leverages (upper left part of the plot) present observations that are not well explained by the model. After subsequent inspection of their nature and origins, they can be retained as part of the data for analysis, expecting to feature a new pattern, still not explained by the actual model. Samples with low Q residuals and low T^2 (bottom left part of the plot) can be considered to be well explained by the model and should be retained.

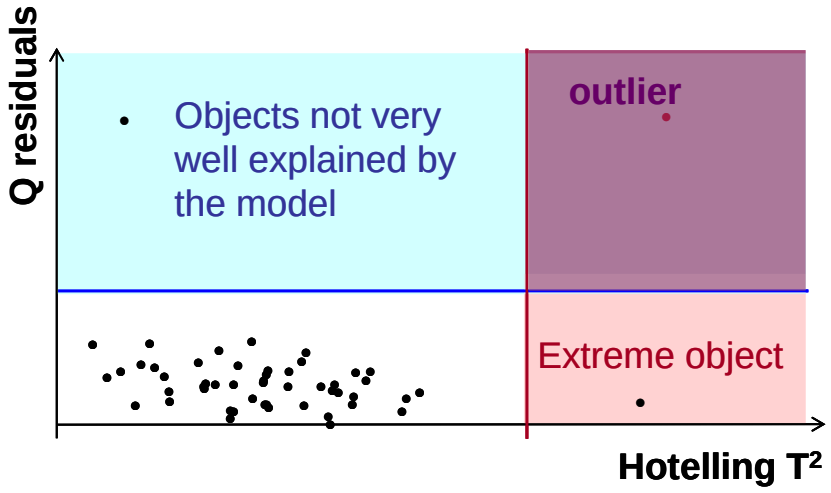


Figure 17. Plot of Hotelling's T^2, and of Q residuals contributions of different samples. Four different cases can be encountered for different combinations of T^2 and Q.

2.2.7 Multivariate calibration. Non-linear regression methods

Multivariate linear techniques, such as MLR, PCR and PLS, are usually used in water quality investigations, but they strictly rely upon the hypothesis that relationships between the predictor variables and the target analyte have a linear dependence with little departure of this condition. In practice, however, some water chemical systems and processes can display nonlinear relationships (Milot et al., 2002; Baxter et al., 2001). Thus, linear methods may be inappropriate for modelling them.

Modelling nonlinear chemical systems is a challenging task, since the analytical form of the relationship between measured variables and the target property of interest is generally unknown.

Several nonlinear techniques have been proposed to cope with nonlinearity in chemical data such as: smoothing with a multiple additive regression technique, SMART (Friedman , 1984); multivariate adaptive regression splines, MARS (Friedman , 1991); artificial neural networks, ANN (Himmelblau, 2008); and radial basis function networks ,RBFN (Moody and Darken, 1989) and others.

These nonlinear methods can handle a large number of nonlinear chemical problems, but also they show significant drawbacks. For example, many of these methods require many terms and parameters to be adjusted. This task is usually time-

consuming and tedious, thus they are difficult to be used for routine applications. On the other side, ANN suffers from local minima and usually reports too optimistic results (due to over-fitting problems) resulting in not stable, inaccurate models, and showing poor prediction results for samples not included in the calibration set (Shawe-Taylor and Cristianini, 2004).

Among the possible nonlinear modelling techniques to use for the goals of this study, a particular class of methods, i.e. kernel-based methods like Support Vector Machines, SVM (Vapnik, 1998) and kernel-PLS methods (Dayal and MacGregor, 1997) were used in this Thesis. These two methods were distinguished from the rest because they are relatively easy to use and they usually report very accurate prediction results.

The advantage of these two methods comes from: 1) the original data are transformed using the so called kernel trick; 2) a linear PLS regression or in case of SVM – the e-insensitive linear lost function (Vapnik, 1995) are applied to the kernel transformed data. Thus, if non-linear aspects exist in the data, they can be captured by the kernel, and the simplicity and good statistical properties of linear regression techniques can still be reserved.

Many kernel functions can be used and the choice of kernel transformation is user dependent. The simplest kernel is just the dot product of the data matrix by its transpose (the covariance data matrix), called also linear kernel. Probably the most popular kernel function is the radial basic function RBF (Rosipal and Trejo, 2001). RBF has a tuning parameter - the Gaussian width, which is necessary to be optimized, because it has an influence on the predictive ability of the model.

Radial Basis Functions - Partial Least Squares Regression (RBF-PLS)

Walczak and Massart (1996a) proposed a radial basis function kernel to perform this kernel transformation, followed by the application of PLS regression. RBF kernel performs a nonlinear transformation of the input variables **X** into a new feature higher dimensional space, which is called the activation matrix **A**, which uses the following radial basis mapping function,

$$A = \begin{bmatrix} \Theta_1(x_1)\Theta_2(x_1)...\Theta_m(x_1) \\ \Theta_1(x_2)\Theta_2(x_2)...\Theta_m(x_2) \\ \\ \Theta_1(x_m)\Theta_2(x_m)...\Theta_m(x_m) \end{bmatrix} \quad (33)$$

where Θ is the RBF, characterized by the center and the width parameters. The most frequently used type of RBF is the Gaussian function, which calculates every element from **A** as it is follows:

$$a_{ij} = \Theta(\|x_j - x_i\|) = \exp \frac{-(\|x_j - x_i\|)^2}{\sigma_j^2}, \quad (34)$$

where i = 1,2, …,m; and j = 1,2,…m. The center and width of the jth radial basis function are x_j and σ_j respectively. The Euclidian distance measure is denoted as $\| \|$.

The obtained activation matrix **A** is a squared symmetrical matrix (m x m) with ones on its diagonal. **A** is also independent of the number of variables in the original data, and it is only determined by the number of samples in the data (Walczak and Massart 1996b).

Such kernel-transformed data is multicollinear and the regression problem cannot be solved with the standard multiple linear regression method (MLR). PLS regression offers a good solution for handling multicollinearity and it is then applied to model the relation between the matrix **A** and the target property, **y**.

Centered data from **A** and **y** are then projected onto the low dimensional score matrices T and U respectively (see equations 35 and 36), during PLS modeling.

A=**TP'**+**E**, (35)

y= **Uq'** + **f**, (36)

The linear inner relation between the score matrices T and U became as follows:

U =**T** +**H**, (37)

and finally

y= **Tq'** + **f,** (38)

where **E**, **f** , **f*** and **H** are residual vectors or matrices.

This avoids calculating the coordinates in the feature space which is a rather difficult task for a highly dimensional feature space. The major benefit of RBF-PLS is that using the kernel function (i.e. the dot products in the feature space) avoids non-

linear optimization procedures and allows the use of much simpler and more reliable linear PLS regression algorithm later.

A cross-validation procedure (please refer to section 2.2.8) should be used to select the optimal number of factors. The performance of RBF-PLS depend significantly on the width of the Gaussian functions (Walczak and Massart, 1996a) and this parameter has to be optimized by searching over a predefined range without any a priori knowledge about the data distribution. Once the RBF-PLS model is constructed, it can be used for prediction of new samples.

A user-friendly graphical interface, based on a collection of MATLAB m-files, called TOMCAT, Toolbox for Multivariate Calibration Techniques, was used in the work for RBF-PLS calculations (Daszykowski et al., 2007). The graphical user interface and their routines are freely available.

Support vector machines regression (SVR)

Support vector machine (SVM) is a method developed in the frame of the machine learning theory (Vapnik, 1995), implemented for structural risk minimization.

SVM gained recently interest because of its advantages to find global optima, and to provide a good generalization with a reduced number of samples in sparse and high-dimensional spaces (Cristianini and Shawe-Taylor, 2000). Initially, SVM was applied to solve classification problems, but later it was extended to solve non-linear regression problems with the introduction of the ε-insensitive loss function for regression.

Support vector regression methods (SVR) start with the mapping of the original data matrix **X** (in a similar way as RBF-PLS) into a higher-dimensional feature space using a kernel function and then, a linear regression is performed. Using mathematical notification, the linear model $f(\mathrm{x}, \omega)$ in the feature space can be written as follows:

$$f(\mathrm{x}) = \sum_{i,j=1}^{m} \omega \left\langle \varphi(x_i)\varphi(x_j) \right\rangle + b, \qquad (39)$$

where $\varphi_{(x)}$, is the nonlinear transformation (nonlinear mapping function) of the original variables **X**, and ω (weight vector) and b (bias) are the coefficients of the linear model

that can be obtained by solving a quadratic programming optimization problem. The usual process of building SVM models includes an approximation performed by using different types of kernel functions. The SVR regression equation described above, can be rewritten including the mapping function (kernel function) notation as follows:

$$f(x)= \sum_{j=1}^{m} \omega K(x_i, x_j) + b \ , \tag{40}$$

where the kernel function $K(x_i,x_j)$ can be a linear, a polynomial, a sigmoid or the one previously mentioned and used in this Thesis – the radial basis function (RBF). In this way, the nonlinear separable problem becomes linearly separable once the original data is mapped onto a high-dimensional feature space.

The coefficients of the regression model are estimated by minimizing a square error function, which can be defined as an empirical risk loss function, which indicates the quality of the estimation (Kao et al., 2013). SVR calculates a loss function (ε-insensitivity loss function) defined as the one which at most has an ε- deviation from the expected values, for all the training data, and at the same time it is as flat as possible (Martinčič et al, 2015). The loss function (L) used to measure the quality of the estimation is:

$$L(f(x), y) = \begin{cases} |f(x) - y| - \varepsilon & for |f(x) - y| \geq \varepsilon \\ 0 & otherwise \end{cases}, \tag{41}$$

Where, y is the target property (the concentrations of a particular THM), and ε is a user –defined parameter for the region of ε-insensitivity. Zero is observed when the predicted values do not exceed the defined band region. On the contrary, if the predicted values are out of the band region, the loss equalizes the difference between the predicted values and the margin (Kao et al., 2013).

The weigh vector ω is calculated by minimizing the following regularized empirical risk function as follows:

$$R = C\frac{1}{n}\sum_{i=1}^{n} L(f(x_i), y_i) + \frac{1}{2}\|\omega\|^n \ , \tag{42}$$

Where, the objective function $\|\omega\|^n$ is a regularization term, which specifies the trade-off between the model complexity and the approximation to the accuracy of the model. Thus, this model would show better generalization properties. C is a user defined regularization parameter, which affects the trade-off between the regularization term and the empirical risk.

Taking into account the empirical risk and the structure risk, the optimal solution is transformed into the following constrained expression with slack variables (Vapnik, 1998), which can be solved as a quadratic optimization problem:

$$\text{Minimize } \frac{1}{2}\|\omega\|^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^*), \tag{43}$$

subjected to the constraints:

$$\begin{cases} y_i - (\omega \cdot \Phi(x_i) + b \le \varepsilon + \xi_i \\ (\omega \cdot \Phi(x_i) + b) - y_i \le \varepsilon + \xi \\ \quad \xi_i \xi_i^* \ge 0, i = 1....., n \end{cases} \tag{44}$$

Where $\xi_i \xi_i^*$ are called the slack variables and they should be positive on sign. They are used to measure the deviation of training samples outside the ε-insensitive region. In other words, the model complexity consists of the error of the training data with an added penalty term as $\xi_i \xi_i^*$ slack variables (Liu et al., 2014).

As mentioned before, Equation 43 can be solved as quadratic programing problem using the Lagrange multipliers. The general expression of SVR function can be presented as follows:

$$f(x) = \sum_{j=1}^{n}(\alpha_i - \alpha_i^*)K(x_i, x_j) + b, \tag{45}$$

where $\alpha_i, \alpha_i^* \ge 0$ are called the Lagrange multipliers , which satisfy the equalities $0 \le \alpha_i \le C; 0 \le \alpha_i^* \le C$ and C is the above mentioned regularization parameter, that specifies the tradeoff between the model simplicity (and hence its generalization), and the training error, allowing for some data fit losses (Üstün et al., 2007). Some of the

Lagrangian multipliers are zero corresponding to the data concerning the inside of the e-insensitive tube. Data values with nonzero Lagranjgian multipliers α_i, α_i^* are called support vectors. In other words, support vectors are those data points that "support" the construction of the regression function. There are two types of support vectors (Noori et al., 2011). A first type of support vectors have values of the weights less than C and thus, are called margin support vectors. A second type of support vectors have values equal to C and thus, are called error support vectors. The margin support vectors are found on the margin of the ε-insensitive tube, while the error support vectors are outside the tube (see Figure 18).

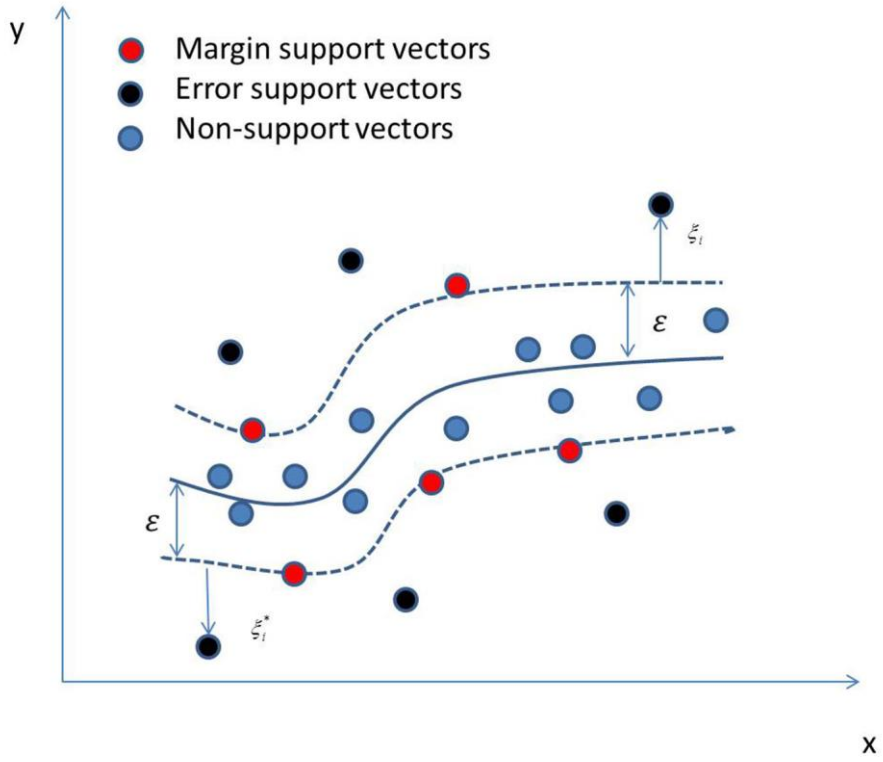


Figure 18. SVM non-linear regression function with ε-insensitive band and ξ slack variables

The complete SVR equations can be found in Schölkopf and Smola, (2002), Vapnik (1995) and Vapnik (1998).

Once SVR calibration model is constructed, it can be used to predict unknown **y** values from new **X** values.

SVR performance (estimation accuracy) depends on the proper adjustment of the parameters *C*, ε and of the kernel parameters, which are usually user-defined. The most

popular method for parameter optimization (also applied in this Thesis) is the cross validation procedure, using a comprehensive grid search procedure over all possible values of the parameters (Luts et al., 2010). Many software packages like STATISTICA (StatSoft, Inc., Tulsa, OK, USA), PLS_toolbox (Eigenvector Research, Manson, WA, USA) and also toolboxes such as libSVM (Chang and Lin, 2001) or the SVM package of Steven Gunn (Gunn, 1998), provide MATLAB routines for SVM regression calculations.

2.2.8 Model validation and error measurements

The purpose of model validation is to evaluate the performance of multivariate models for discrimination and prediction purposes. The goal is to obtain a good model able to make accurate predictions when applied to new data and to give reliable and interpretable results.

Usually, first PCs or LVs capture (when dealing with PCA or PLS respectively) the most significant part of the variance, while the noise is being modelled by the other higher PCs or LVs. To reach the optimal number of PCs and LV to build a good model, its proper validation is required. There are many validation methods to optimise a particular model and thus, to assess its future performance. In the six case-studies of this Thesis, different techniques have been used for validation such as the prediction of training set of data or self-prediction, the internal cross validation and the external model validation, which are described next in more details.

- *Prediction of training set of data* (self-prediction)

This is obtained using the *root mean square error of calibration* (RMSEC). RMSEC is a measure of how well the model fits the data:

$$RMSEC = \sqrt{\frac{\sum_{i=1}^{n} (\hat{y}_i - y_i)^2}{n}} \quad , \tag{46}$$

where the $\hat{y}_i$ are the values of the predicted property (for example trihalomethanes concentrations) and y_i are the actual measured experimental values for calibration (training) samples included in the development of the model and n was the number of them. Usually as the number of components increases, the RMSEC decreases and therefore this method was not preferred due to the always improving fitting just by increasing the number of components, leading to overfitting.

- *Internal model validation (cross-validation, CV)*

The use of internal model validation or cross validation is usually an important step in the correct selection of the number of components. Using this method (Brereton, 2003), a few samples are left out from the calibration data set and the model is built with the remaining samples. Then, the values of the left out samples are predicted and the prediction errors calculated. The process is repeated then with another subset of the calibration samples until all samples had been left out.

Various schemes of samples subsetting for CV are available. A very popular method is the *Leave-One-Out* (LOO) CV method. However it works properly only for small data sets (citation). Figure 19 illustrates graphically the process of this method:

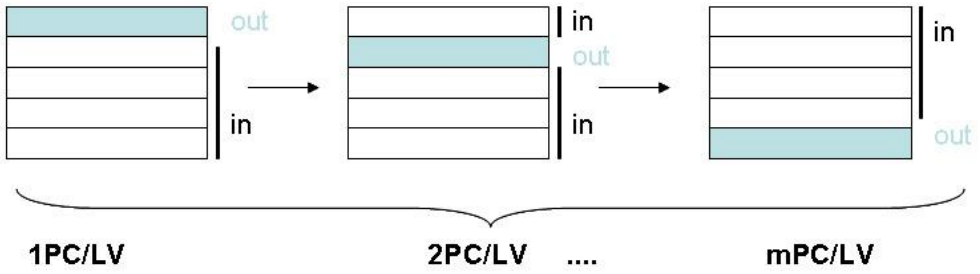


Figure 19. Graphical representation of leaving-one-out procedure.

LOO method extracts one sample (row) from both **X** and **y**, which is left out, then a new model is built (with *a* number of components) with the remaining samples, and used to predict the **y** data from all **X**. Finally, the predicted residual error sum of squares (PRESS) is calculated as follows:

$$cumPRESS(m) = \sum_i (y_i - \hat{y}_i)^2 \quad , \tag{47}$$

PRESS is calculated from the squared differences of actual (true) y_i and predicted *values* y_i for every particular sample extracted and for every model with m principal components or latent variables. This process is repeated with all of the samples so that each row will have

been left out and predicted once. The sum of *PRESSm values* gives the cumulative PRESS (*CUMPRESS*) value for the model with *m* components.

The *root mean square error of cross validation* (RMSECV) is obtained using the expression as follows:

$$RMSECV = \sqrt{\frac{cumPRESS_m}{I}} \quad , \tag{48}$$

where *I* is the total number of samples in the calibration set. This is repeated for the different tested models with different number of components.

Other methods for cross-validation are:

1. Segmented cross-validation, where data are divided in certain random segments (typical number for instance is 10 segments or 10% of samples per segment)

2. Systematic segmented cross-validation, where a constant number of samples is left out at a time (typical example includes samples with replicates).

RMSECV is plotted as a function of the number of components. The minimum of this curve usually agree with the optimum number of components (or latent variables) to consider in the final model.

- *External model validation*

This is certainly the most accurate and reliable method of model validation. The models should be successful in prediction of new, unknown samples. When CV is used, the influence of intrinsic sources of errors and instrument noise can be diminished, however, CV is not able to account for the correlation among dependent variables **Y**. In order to validate externally a model, the data have to be split (usually randomly) first on two data subsets. The most frequently used method for random splitting of data on two sets has been the Kennard –Stone method (Kennard and Stone, 1969). The first set is called the calibration (or training) subset and the second is called the validation subset (test data).

The root mean square error of prediction (RMSEP) is calculated as follows:

$$RMSEP = \sqrt{\frac{\sum_{i=1}^{n} (\hat{y}_i - y_i)^2}{n}} , \tag{49}$$

where the $\hat{y}_i$ are the values of the predicted values by the model, y_i are the actual measured values of the external validation (test) samples and *n* was the number of them.

In practice, RMSEC, RMSECV and RMSEP (if possible) are plotted against the number of components (as in Figure 20). Whereas RMSEC (very often) and RMSECV (sometimes) can give misleading results and overfitting problems, the independent external dataset validation is more accurate.

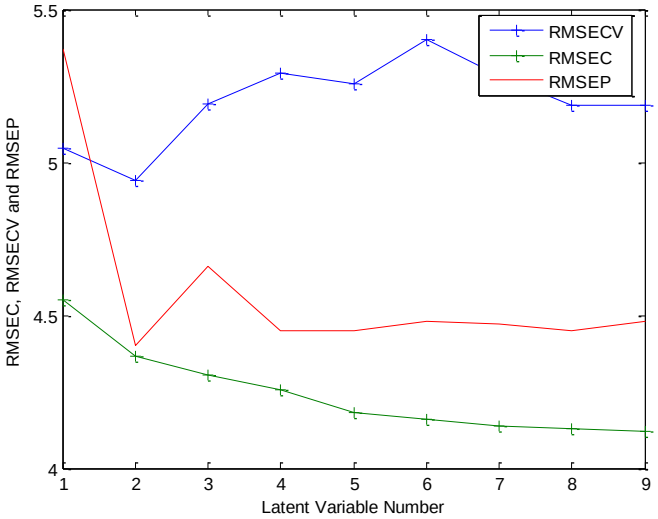


Figure 20. Examples of PLS prediction errors in calibration (RMSEC), in internal cross validation (RMSECV) and extrernal validation (RMSEP).

From the plot above (Figure 20), RMSEC does not suggest a define number of LVs to be selected. On the contrary, RMSECV and RMSEP clearly point out that the correct number for LVs would be 2.

Some others errors estimations used for model efficiency diagnostics are:

- Relative errors of concentration prediction in percentage, for both, calibration and prediction step:

$$Rel.\ error\ in\ \% = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{\sum_{i=1}^{n}(y_i)^2}}\ \text{x}100, \qquad (50)$$

- Bias or average values of residuals between the actual and predicted values:

$$bias = \frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)}{n}\ , \tag{51}$$

where $\hat{y}_i$ and y_i were defined above:

- SEP, Standard error of performance, which is calculated as follows:

$$SEP = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i - y_i - bias)^2}{n-1}}\ , \tag{52}$$

Chapter 3

Results and Discussion

This section summarizes six articles suggesting the application of various chemometric methods for water quality research. The articles are organized in three main blocks according to the main objectives. More specifically, the blocks contain the presentation and analysis of experimental data obtained: 1) in the Sant Joan Despí drinking water treatment plant (SJD-DWTP); 2) in laboratory experiments with simulated water disinfection conditions; 3) in laboratory experiments with water sources from the water distribution system of Barcelona (WDS); 4) in the TRARGISA wastewater treatment plant (WWTP); and 5) in sensory tasting experiments.

3.1 Chemometrics modeling of the trihalomethanes formation in a DWTP and in laboratory conditions

The formation of trihalomethanes (THMs) in Sant Joan Despí drinking water treatment plant (SJD-DWTP) Barcelona is associated with potabilization procedures employed to guarantee sufficiently high quality of drinking water. To the extent that such procedures may result in the formation of high THMs concentrations and may impede compliance with the established sanitary limits according to the EU guidance, it is important to identify which are main factors affecting THMs formation.

In this subsection, the empirical results of three papers dealing with the THMs formation are discussed. The papers comply with the main objectives of the Thesis in the following aspects:

2. Development of reliable empirical models for trihalomethanes formation based on multivariate analysis of real DWTP parameters data by applying and comparing different linear and nonlinear chemometric methods;
3. Review of the most important parameters influencing THMs formation from the achieved models;
4. Comprehensive assessment of natural organic matter (NOM) role in THMs formation when DWTP operational disinfection conditions were simulated in specially designed experiments.

The first two papers present results from chemometric modeling of THMs disinfection by-products ($CHCl_3$, $CHBr_3$, $CHCl_2Br$, and $CHBr_2Cl$) generated in water

treatment processes implemented at the SJD-DWTP. Both papers investigate the linear and nonlinear relationships among THMs in finished drinking water and various operational parameters, monitored in different treatment processes and plant locations. Various chemometrics techniques, such as principal component analysis (PCA), multilinear regression (MLR), stepwise MLR (SWR), principal component regression (PCR) and partial least squares regression (PLS), Kernel type (radial basis functions) Partial Least Squares (K-PLS) regression and Support Vector Machines (SVR) regression, have been used and compared to model, predict and visualize the complex behaviour observed for the measured trihalomethane concentrations. The major operational parameters have been determined in order to match targeted objectives.

The third article of this block proposes a chemometrics assisted methodology for the interpretation of THMs speciation, when natural organic matter (NOM) reacts with chlorine in simulated disinfection reactions, conducted in laboratory. Two experimental design strategies were applied in order to evaluate and describe some of the major factors determining the reaction of THMs formation. Results suggest that NOM fractions are of high relevance for THMs speciation during their formation.

3.1.1 Article 1 – Platikanov, S., Puig, X., Martin, J. and R. Tauler. *Chemometric modeling and prediction of trihalomethane formation in Barcelona's water works plant.* Water Research 41 (2007) 3394-3406.

Introduction

It is of great importance to set different treatment or operational strategies in DWTP that reduce THMs formation in finished drinking water. The literature provided different types of THMs formation models such as kinetic models in laboratory conditions, deterministic models from laboratory and field studies, and multivariate empirical regression equations from field studies. Although traditional empirical regression models can be adequate in some cases, such models require the use of many operational parameters that are monitored over a large period of time.

Chemometrics provides different tools and techniques which may be particularly useful to work with multivariate data obtained from DWTP processes. The most popular chemometrics methods to empirically analyse THMs formation according to the literature as shown in Table 4 of Section 2.1.6, were multiple linear regression (MLR)

and nonlinear neural networks regression approach (NN). Because MLR has limitations (such as the inclusion of a large number of multicorrelated parameters in the estimated model) and Neural Networks is a rather complex approach for application on a daily basis, other chemometrics regression methods are proposed. In this paper, chemometrics regression techniques such as Principal Component Regression (PCR) and Partial Least Squares Regression (PLS) were suggested for data modelling and compared with MLR.

The data consisted of twenty-three physicochemical parameters generated in the SJD-DWTP operational processes (i.e., explanatory variables), along with four THMs concentrations and their total sum (i.e., dependent variables). A feature of the data that is worth mentioning is that these data contained operational parameters monitored at the sand filtration, at the carbon filtration, after post-chlorination and at the exit of the plant (please refer to Figure 1 of the Article 1). Data were generated for 162 days including all four seasons during one year. Data were split into a calibration subset, containing samples for 144 days, and a subset for external model validation, containing samples for 18 days. Detailed information about all analyzed parameters is provided in Table 1 of Article 1. After data were conveniently organized for multivariate analysis in matrices, they were autoscaled with the objective to eliminate offsets and changes in measurement units prior to the multivariate analysis.

The first stage was to perform multivariate data exploration by examining descriptive statistics and by applying Principal Component Analysis (PCA) method. At the second stage, a linear regression analysis was conducted. More specifically, the estimated model had DWTP operational parameters as independent variables and the THMs concentrations as dependent variables. In this part, four chemometrics linear regression methods were used and compared: Multilinear Regression (MLR), Stepwise Regression (SWR), Principal Component Regression (PCR) and Partial Least Squares Regression (PLS). The regression analysis included first, model calibration, and then, a subsequent model validation.

WATER RESEARCH 41 (2007) 3394–3406

Available at www.sciencedirect.com

ScienceDirect

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

Chemometric modeling and prediction of trihalomethane formation in Barcelona's water works plant

Stefan Platikanov[a], *Xavier Puig*[a], *Jordi Martín*[b], *Romà Tauler*[a,*]

[a]*Department of Environmental Chemistry, IIQAB-CSIC, Jordi Girona, 18-26, Barcelona 08026, Spain*
[b]*AGBAR, Av. Diagonal 2090-211, 08018 Barcelona, Spain*

ARTICLE INFO

Article history:
Received 7 February 2007
Received in revised form
21 April 2007
Accepted 24 April 2007
Available online 18 May 2007

Keywords:
Trihalomethanes
Water disinfection
PCA
Regression methods

ABSTRACT

Formation and occurrence of trihalomethanes ($CHCl_3$, $CHBr_3$, $CHCl_2Br$, and $CHBr_2Cl$) are investigated in water chlorination disinfection processes in the Barcelona's water works plant (WWP). Twenty-three WWP variables were measured and investigated for correlation with trihalomethane formation. Multivariate statistical methods including principal component analysis (PCA), multilinear regression (MLR), stepwise MLR (SWR), principal component regression (PCR) and partial least squares regression (PLSR) have been used and compared to model and predict the complex behavior observed for the measured trihalomethane concentrations. The results, obtained by PCA as well as the evaluation of the statistical significance of the coefficients in the linear regression vectors, revealed that the most important WWP variables for trihalomethane formation were: water temperature, total organic carbon, added chlorine concentrations, UV absorbance and turbidity at different sites of the WWP, as well as other variables like wells supply flow levels and carbon filters age. Overall, MLR and PLSR methods performed the best and gave similar good predictive properties. Best results were obtained for the total sum of trihalomethane concentrations, TTHM, with average modeling and prediction relative errors of 12% and 16%, respectively. Among the individual trihalomethanes, the concentrations of $CHBr_3$ were the worst predicted ones with average modeling and prediction relative errors between 21–25% and 29–31%, respectively, followed by $CHCl_2Br$ with 23–26% and 25–27%. Better predictions were obtained for the concentrations of $CHBr_2Cl$ with relative modeling and prediction errors varying between 14–17% and 21%, and for the concentrations of $CHCl_3$ with 21–24% and 23–25% errors, respectively.

1. Introduction

Trihalomethane compounds (THMs) are a set of compounds ubiquitously formed by the interaction between organic chemical compounds present in surface river waters with disinfection oxidant products like chlorine and chlorine dioxide (Rook, 1974; Richardson, 2002). In particular, the compounds considered in this study are chloroform ($CHCl_3$), bromodichloromethane ($CHCl_2Br$), chlorodibromomethane ($CHBr_2Cl$) and bromoform ($CHBr_3$). The presence of these compounds in domestic waters is undesirable because of their negative effects on human health (McGeehin et al., 1993; Simpson and Hayes, 1998). Their formation is a consequence on one hand of the presence of organic matter in surface waters and, on the other hand, of potabilization procedures usually used in water supply plants. Since these procedures are usually necessary to guarantee the quality of consume waters, the main question is therefore, to have quality control

*Corresponding author. Tel.: +34934006140; fax: +34932045904.
E-mail address: rtaqam@iiqab.csic.es (R. Tauler).
0043-1354/$ - see front matter © 2007 Elsevier Ltd. All rights reserved.
doi:10.1016/j.watres.2007.04.015

procedures to predict and avoid the raising of trihalomethane concentrations to levels below the recommended limits for sanitary purposes. In this context, it is important to point out the existence of the EU guidance (EEC, 1998) setting the limits of these compounds for consume water (150 μg/l until December 31st 2008 and 100 μg/l from 2009).

In this work, the case of study is the formation of THMs in the Sant Joan Despí Water works plant (WWP) in Barcelona, Spain, which supplies around 30% of the consumed water in the Barcelona metropolitan area. The input water source to the plant is mostly from Llobregat river, located in Catalonia, NE Spain. The hydraulic regime of the Llobregat River is of the Mediterranean type: low average flows during ordinary conditions with peak events of heavy floodings. The water flow range is extremely wide: from absolute dryness up to 2000 m^3/s. When this source is scarce, water from wells in the Baix Llobregat large aquifer (Martín-Alonso, 2003) is also pumped and provides a very important complementary source of raw water to the plant. Despite its limited extension, the presence of halides (including chloride and bromide) in a limited amount in the upper part of the basin has had a dramatic impact on water quality since the establishment of three salt mines. The halides and the organic matter in the river lead to a surplus of chlorinated and brominated compounds when treated with disinfection products (Ventura and Rivera, 1985).

In order to monitor water quality and to control water treatment procedures, different variables and parameters are continuously measured along the water treatment plant (WWP variables) as well as formation and concentrations of trihalomethane compounds (THM variables). The first set of variables was considered to be the predictor block of variables and the second set was the predicted block of variables. The goal of this study is to build a predictive model, which relates the two data blocks, i.e. a model, which may help to predict the formation of THMs from the WWP measured variables. First, an exploratory multivariate data analysis has been performed using principal component analysis (PCA) (Jolliffe, 1986) of the augmented data set formed by the two data blocks $[\mathbf{X}, \mathbf{Y}]$. And second, different multivariate regression methods, including multilinear regression (MLR) (Massart et al., 1998), stepwise regression (Draper and Smith, 1981), principal component regression (PCR) (Naes and Martens, 1988) and partial least squares regression (PLSR) (Geladi and Kowalski, 1986; Höskuldsson, 1988) have been used to build predictive models to correlate the two data blocks, WWP variables and THM concentrations.

Several investigations have been carried out to improve the understanding of the relation between water-quality parameters, WWP managing, and concentrations of THMs in drinking water. Natural organic matter (NOM) in surface waters is a heterogeneous mixture of substances, such as humic acids, fulvic acids and others. As reported by Gallard and von Gunten (2002) and Rodrigues et al. (2007), different fractions in humic and fulvic substances, have been considered as the main precursors of THMs. Other parameters that influence the formation of THMs are chlorine residual concentrations, reaction time, pH and temperature. Brominated THMs are also formed especially in waters containing bromide ion. According to Clark et al. (2001) and Nikolaou (2004) the presence and concentration of bromide ion affects the overall formation of halogenated THMs. There are different types of THMs formation models described in recent literature. Data are obtained from field studies (Garcia-Villanova et al., 1997; Rodríguez et al., 2003a) or at laboratory-scale (Nikolaou, 2004). Some models, based on studying the kinetic reaction of THMs formation, were developed (Clark, 1998; Li and Zhao, 2006). Other models are based on multivariate empirical regression equations including a number of operational and water-quality parameters as predictors and the generated THM concentrations as predicted variables (Amy et al., 1987; Rodríguez and Sérodes, 2001). Until now, chemometric methods more widely applied for modeling formation of THMs have been MLR analysis (Golfinopoulos et al., 1998; Golfinopoulos and Arhonditsis, 2002; Nikolaou et al., 2004) and non-linear methods such as neural networks and logistic regression analysis (Milot et al., 2002; Rodríguez and Sérodes, 2004). Most of these previous investigations have been performed under rather limited experimental conditions in the laboratory or under well-controlled conditions and short monitoring time periods and situations. There is still an urgent need to develop reliable and robust models to predict THM formation in WWP. These models should be applied under real plant conditions and should allow for the on-line control of the formation of THMs and to obtain the conditions under which their formation is limited to levels below the established legal and sanitary limits. This investigation has been oriented to improve quality assessment and optimal management of Barcelona's WWP disinfection procedures and to provide more efficient tools to avoid possible human health risks produced by consume waters contaminated by THM compounds.

2. Materials and methods

2.1. *Experimental data set*

Water samples were collected from the Sant Joan Despí WWP, Barcelona, Spain. It has a design capacity of 500 000 m^3/d and it is of conventional design. Major features include pre-chlorination sites, coagulation, flocculation, sedimentation, sand filtration, ozonation, carbon filtration and post-chlorination. Fig. 1 shows the process and the location of sampling and chlorination points. Samples were taken at the following locations: after sand filtration, after carbon filtration, after post-chlorination and at the exit of the plant.

In Table 1, WWP variables measured at the plant are given.

UV absorbance, total organic carbon (TOC) concentrations as well as residual chlorine concentrations, chloride concentration and water turbidity, were measured after sand and carbon filtration at one or two sites. Wells water amount, added to the system to supply the insufficient river water level, was measured too. Water temperature was measured after the location where mixing of the well water with the water coming after the sand filtration was taking place. Finally the influence of emergently added chlorine was assessed immediately after post-chlorination location point at two sites of the plant. Concentrations of different THMs were measured at the water exit of the plant.

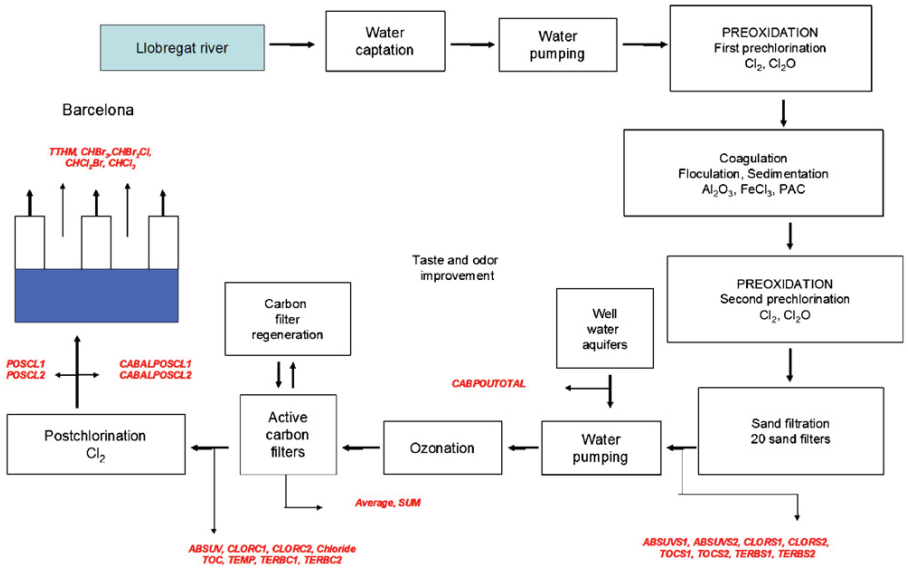


Fig. 1 – Barcelona's water works plant scheme and locations of sampling. For variable abbreviations, see Table 1.

All these variables were measured at the plant over the whole year 2003 and were used to build up the block of **X** variables (data matrix **X**). For a total number of 162 days during 2003, concentrations of the different THMs were measured (THM variables, block of **Y** variables). Since measurements of THM concentrations were performed only for some days, only those measurements of the WWP variables (**X**) being coincident in time (the same day) with measurements of THM concentrations ($\mathbf{y}_i$) were finally considered, where *i* refers to concentration of the considered trihalomethane compound, $CHCl_3$, $CHCl_2Br$, $CHBr_2Cl$, $CHBr_3$ or TTHM. These 162 measurements were spread over the whole year, although for some months there were more measures (21 in January) than for others (5 in August). For one month, measurements were not available (June). Concentrations of the four THMs were determined using standard chromatographic methods (Head Space GC-ECD, Gallard and von Gunten, 2002). Samples were analyzed using a Fisons 8130 gas chromatograph with a DB-624 30 m × 0.32 mm × 1.8 mm μm film thickness fused silica column. Automatic injections of 0.5 ml were made in split mode ($\frac{1}{10}$), with helium as a carrier gas and nitrogen as a make up gas. Detection and quantification limits were estimated to be around 0.10 and 0.25 μg/l. Trihalomethane standard solutions were prepared in the range of 0.5−8 μg/l . The total amount of THMs (variable TTHM) was calculated as the sum of the concentrations of the four individual compounds. Therefore, a total number of five $\mathbf{y}_i$ variables were defined. UV absorbance measurements in Abs/100 cm units were carried out at 254 nm using 1-cm quartz cells and then reporting values for a 100 cm path length (i.e. multiplying them by 100). Residual free chlorine was measured according to the DPD colorimetric method and chloride was measured volumetrically using Mohr method (APHA et al., 1998). TOC concentration in mg/l of the samples was determined using a TOC analyzer. Turbidity was measured in FNU using nephelometry. Estimated measurement errors for quality control were below 35% for THM concentrations, below 25% for turbidity, below 15% for halides, below 15% for chlorine and below 15% for TOC. Only accredited and validated (ISO17025, ISO9000) methods were used and applied for routine treatment control. Influence of chlorine dioxide was studied and if the assay is run properly (i.e. reaction times) there was no interference at the levels of dosage.

Two data subsets were prepared, one to build the model and another to validate the model. Validation data are a reduced data subset of the whole set of values selected to cover the different situations observed during different months and seasons of the year (one or two samples per month), where temperature and river flow may change considerably. The data subset for model calibration covered WWP and THM concentration variables measured for a total number of 144 days. The validation data set covered the same variables measured for the rest of days, with a total number of 18 days.

2.2. *Data treatment*

Preliminary exploratory data analysis was performed using univariate descriptive statistics. Each variable was plotted individually for the different measurements performed during the whole period of investigation. Histograms and box plots were analyzed for tendencies and outliers. Pair-wise correlations among all the variables were also evaluated.

Once experimental data were properly arranged in data matrices, they were autoscaled (column mean centered and scaled). This preliminary data treatment eliminated offsets, changes in measurement units and focused the analysis on proper modeling of observed variances in measured variables. This data pretreatment is frequently used in multivariate data analysis (Massart et al., 1998).

Table 1 – System variables (variables 1–23) and predicted variables (variables 24–28) measured in the water works plant

Variable	Abbreviation	Description	Average	Minimum	Maximum	Standard deviations
1	ABSUVS1	UV absorbance after sand filters at site 1	8.05	5.26	12.36	1.42
2	ABSUVS2	UV absorbance after sand filters at site 2	8.04	4.43	12.05	1.36
3	CLORS1	Residual Cl_2 concentration in mg/l after sand filters at site 1	0.194	0.017	7.925	0.635
4	CLORS2	Residual Cl_2 concentration in mg/l after sand filters at site 2	0.160	0.004	1.581	1.358
5	TOCS1	Organic carbon total concentration in mg/l after sand filters at site 1	5.1	2.4	7	0.6
6	TOCS2	Organic carbon total concentration in mg/l after sand filters at site 2	5.2	2.6	7.1	0.7
7	TERBS1	Turbidity measured in FNU after sand filters at site 1	0.71	0.23	3.05	0.44
8	TERBS2	Turbidity measured in FNU after sand filters at site 2	0.63	0.17	5.75	0.60
9	ABSUV	UV absorbance after carbon filters	4.00	1.2	7.8	1.0
10	CLORC1	Residual Cl_2 concentration in mg/l after carbon filters site 1	0.97	0.80	1.71	0.11
11	CLORC2	Residual Cl_2 concentration in mg/l after carbon filters site 2	0.96	0.82	1.23	0.08
12	Chloride	Chloride concentration in mg/l after carbon filters	289.7	160	532	67.3
13	TOC	Total organic carbon concentration in mg/l after carbon filters	3.47	2.07	4.83	0.58
14	TEMP	Water temperature in Celsius after carbon filters	14.5	7.3	25.2	4.9
15	TERBC1	Turbidity measured in FNU after carbon filters site 1	0.18	0.12	0.34	0.04
16	TERBC2	Turbidity measured in FNU after carbon filters site 2	0.18	0.12	0.33	0.04
17	CABPOUTOTAL	Input wells water total amount per day in liters	784984.5	0	3240700	883577.9
18	POSCL2	Emergency added Cl_2 concentrations in mg/l after carbon filters site 2	0.30	0	31.86	2.64
19	POSCL1	Emergency added Cl_2 concentrations in mg/l after carbon filters site 1	0.02	0	0.67	0.10
20	CABALPOSCL2	Emergency added Cl_2 volume in liters/day after carbon filters site 2	40.4	0	2793.3	295.6
21	CABALPOSCL1	Emergency added Cl_2volume in liters/day after carbon filters site 1	2.3	0	192.3	17.7
22	AVERAGE	Average time of life of carbon filters from last regeneration in days	214.14	155.25	295.95	32.99
23	SUM	Total time of life of carbon filters from last regeneration in days	4282.8	3105	5919	659.9
24	TTHM	Sum of the concentrations in μg/l of the four trihalomethane compounds measured at the exit of the water treatment plant	75.58	40.87	121.55	14.91
25	$CHBr_3$	Concentration in μg/l of bromoform measured at the exit of the water treatment plant	18.75	6.67	61.00	5.51
26	$CHBr_2Cl$	Concentration in μg/l of chlorodibromomethane measured at the exit of the water treatment plant	24.41	12.55	45.00	5.87
27	$CHCl_3$	Concentration in μg/l of chloroform measured at the exit of the water treatment plant	15.19	6.00	30.85	4.79
28	$CHCl_2Br$	Concentration in μg/l of bromodichloromethane measured at the exit of the water treatment plant	17.24	3.67	35.70	6.28

2.2.1. *Multivariate data exploration*

Preliminary multivariate data exploration was performed using PCA (Jolliffe, 1986). This is a projection method, which gives information about the latent (hidden) structure of the data set. It transforms a large number of (possibly) correlated original variables into a smaller number of uncorrelated,

orthogonal variables explaining maximum variance called principal components. Samples or objects are projected on them giving the samples scores. Plots of variable loadings in principal components allow interpretation of main sources of data variance and describe the relationships among the variables. The main advantage of PCA is that it reduces the dimensionality of the problem (number of variables) but retains most of the original variability in the experimental data and filters noise and minor sources of variance. Therefore, PCA allows for a simpler interpretation of variance sources in a particular data set.

2.2.2. Multivariate regression methods

The main goal of this study was to build a multivariate model able to explain and predict the changes observed in the concentration of the different investigated trihalomethane compounds ($\mathbf{y}_i$ variables) as a function of the measured WWP variables (**X** block of variables). This is to find a mathematical relationship between these two data sets of variables. In particular, in this paper, four different multivariate linear regression methods are evaluated for the modeling of trihalomethane formation by WWP disinfection procedures.

(a) Multilinear regression (MLR): This linear method maximizes the covariation between **X** and **y** data sets to obtain best estimations of **y**. Difficulties will come if there are many highly correlated variables, which will lead to an unstable regression equation with a difficult interpretation of regression coefficients. These regression coefficients are estimated by ordinary least squares.

(b) Stepwise multilinear regression (SWR): Stepwise Multilinear Regression (Draper and Smith, 1981), based on forward selection, consists of first classifying the explanatory variables according to their statistical significance and next including one variable at a time at different steps. At the end only statistically significant terms are used to build the model. Again, problems may happen when predictor variables are highly correlated and when there is the possibility of one input variable masking the effect of another input variable. Very often, the achieved models include variables depending on starting choices and insertion strategies. Regression coefficients are then obtained using selected variables and ordinary least squares estimation.

(c) Principal component regression: PCR (Naes and Martens, 1988) is a widely used technique to build regression models when independent or predictor variables are strongly correlated. PCR uses PCA decomposition of the WWP variables **X** data before the regression with **y** variables. The regression vector is calculated using the loadings and score matrices of the PCA decomposition of **X** data.

(d) Partial least squares regression: PLSR (Geladi and Kowalski, 1986) is related to both PCR and MLR. PLSR attempts to maximise the covariance between **X** and **y**. PLSR searches for the factor space most congruent to both matrices, and its predictions are sometimes better than using PCR (Donachie et al., 1999). A new matrix of weights (reflecting the covariance structure between the **X** predictors and **y** response variables) is calculated and included for the estimation of the regression vector.

The selection of optimal number of components (latent variables) in PCR and PLSR was done using cross validation (leaving-out-one sample at a time) and optimal prediction of $\mathbf{y}_i$ values.

2.3. Figures of merit

The following figures of merit were calculated to evaluate and validate the different applied methods:

(a) Root mean squared error of calibration and prediction (RMSEC and RMSEP), calculated as follows:

$$\mathrm{RMSEC} = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{n}}, \tag{1}$$

where $\hat{y}_i$ are the values of the predicted trihalomethane concentrations and y_i are the actual measured values, when calibration samples are included in the development of the model and n is the number of samples. RMSEC is a measure of how well the model fits experimental THM concentrations. RMSEP is calculated exactly as RMSEC except that the estimates are now the values for external validation samples. RMSEP is a measure of how well the model will make predictions of THM concentrations.

(b) Relative errors of THM concentrations in percentage, for both calibration and prediction steps are calculated as follows:

$$\text{Rel. error in\%} = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{\sum_{i=1}^{n}(y_i)^2}} \times 100. \tag{2}$$

(c) Bias or average values of residuals (non-explained differences) between the actual and predicted THM concentration values are calculated as follows:

$$\mathrm{bias} = \frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)}{n}, \tag{3}$$

where $\hat{y}_i$ and y_i were defined above.

(d) SEP, standard errors of prediction, values are calculated as follows:

$$\mathrm{SEP} = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i - y_i - \mathrm{bias})^2}{n-1}}. \tag{4}$$

Quality assessment of the obtained results was discussed by comparison of predicted values versus measured values, both for calibration and validation data sets. For this purpose, different tools were used, like plots of predicted versus measured values and linear regression between them, to estimate the slope, offset, Pearson correlation and determination coefficients of the best fitting line. A slope close to one, an offset close to zero and a correlation coefficient close to 1, mean a good agreement between calculated and experimental values. This evaluation is performed for calibration and validation data.

2.4. Software

Initial data preparation and data arrangement of different **X** and $\mathbf{y}_i$ data sets were performed using EXCEL (Microsoft, Redmon, WA, USA). All calculations were performed using PLS Toolbox 3.5 (Eigenvector Research, Manson, WA, USA) and MATALAB 6.5 software (The Mathworks, Natick MA, USA).

3. Results and discussion

3.1. *Descriptive statistical data analysis (WWP and THM variables)*

Fig. 2 shows that $CHClBr_2$ seasonal average concentration levels were higher than for other THMs. Its concentration varied considerably from 12.55 to 45 μg/l, with its highest average value in spring. $CHCl_2Br$ and $CHBr_3$ concentrations ranged from 3.7 to 35.7 μg/l and between 6.7 and 61 μg/l, respectively. The maximum average concentration for $CHCl_2Br$ was in the spring again and followed the same trend levels than for $CHClBr_2$, also for the other of seasons. The average seasonal $CHBr_3$ concentration levels varied lower in comparison to those of the other three THMs. $CHCl_3$ concentrations showed a different seasonal behavior. Its maximum average concentration level was found in the summer, while for the other three seasons, the average concentration levels of this compound were almost equal or lower. $CHCl_3$ concentrations varied between 6 and 30.9 μg/l with significant fluctuations.

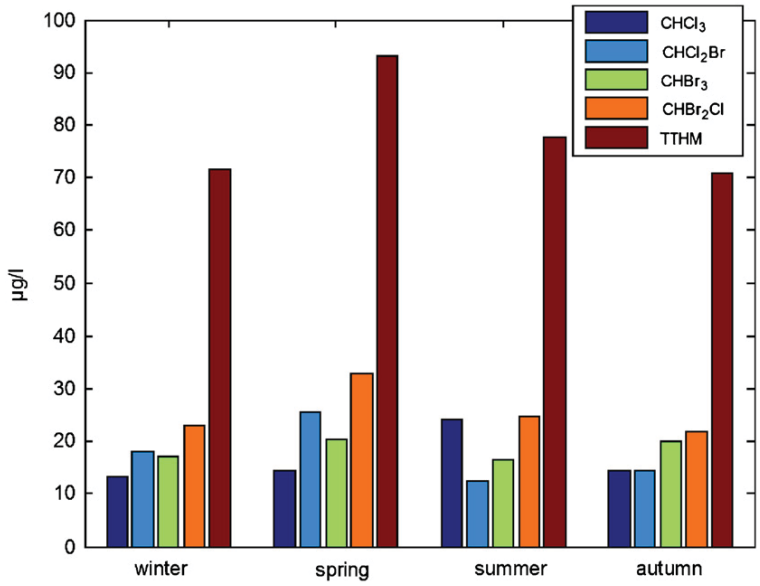


Fig. 2 – Seasonal variation of average THM concentrations.

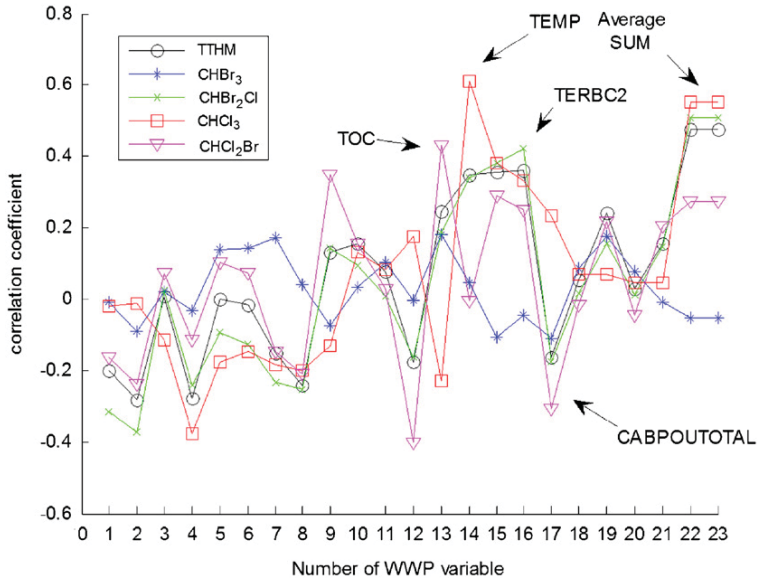


Fig. 3 – Pair-wise correlations among WWP variables and THMs. For variable abbreviations, see Table 1. All correlation coefficients among arrow marked variables are statistically significant ($p < 0.05$).

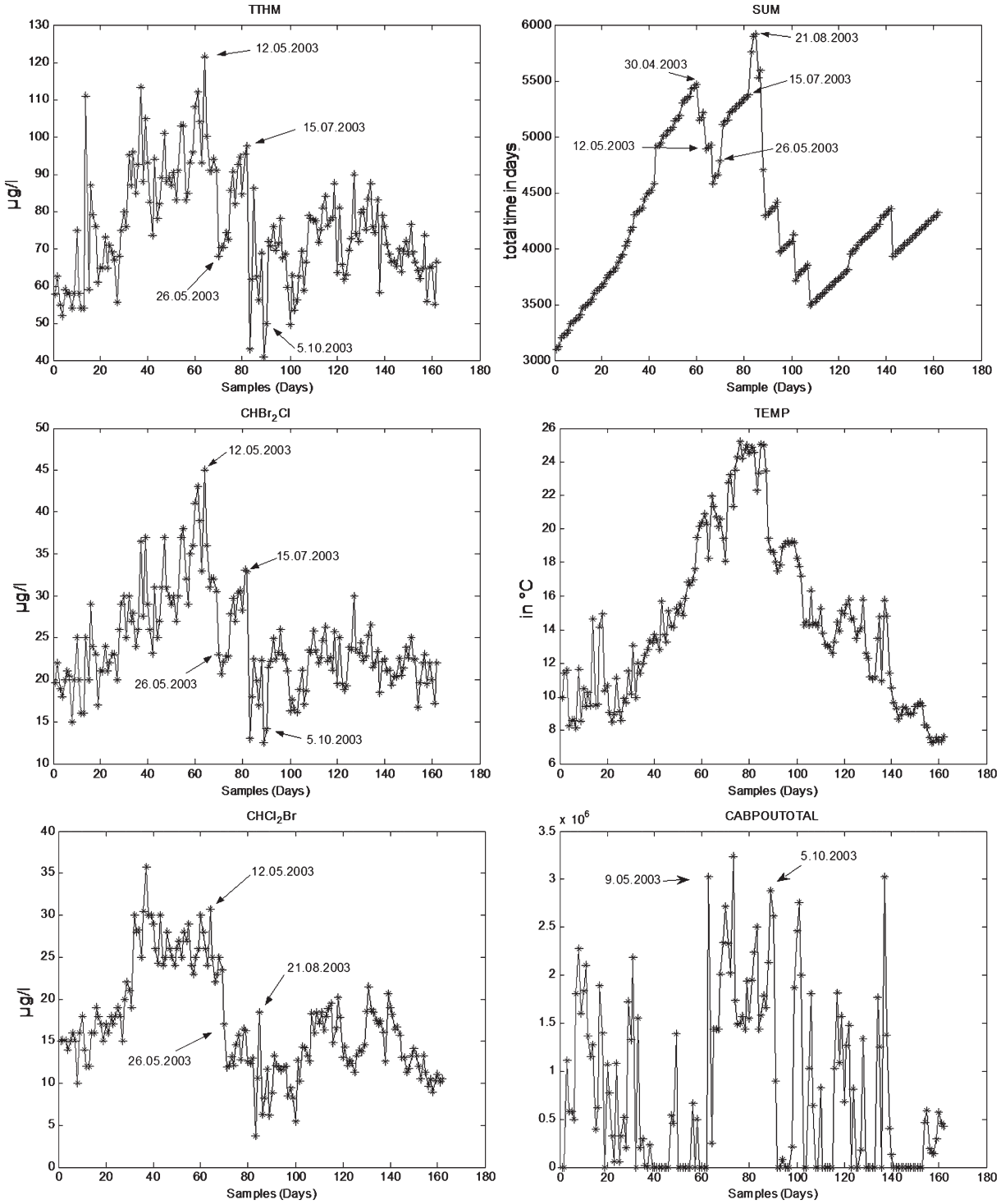


Fig. 4 – Annual trends of some important WWP and selected THM variables. Some dates with significant trends in the WWP or THM variables are given in the plots. For variable abbreviations, see Table 1.

Since three of the THM compounds were found to have maximum average concentration levels in the spring, the total concentration of them or TTHM average concentration level had its maximum in this season too. Summer time was the next season with higher TTHM average concentration values. TTHM concentrations varied between 40.9 and 121.6 µg/l. Colder seasons, autumn and winter, were characterized by relatively lower THM concentration levels

compared to spring and summer. Overall, the average concentration levels of brominated THM compounds exceeded those of $CHCl_3$ for 2003.

Fig. 3 shows pair-wise correlations among THMs and WWP variables. The examination of these relationships revealed that the highest positive pair-wised correlations were found between $CHCl_3$, $CHBr_2Cl$ and TTHM on one side and carbon filter age variables (Average and SUM) on the other side. Water temperature (TEMP) was also found to have a high positive correlation with $CHCl_3$ concentration levels. High positive correlations were also found between TOC concentrations and $CHCl_2Br$, and between water turbidity after carbon filters at site 2 (TERBC2) and $CHBr_2Cl$. Highest negative correlation coefficients were found for the input wells water amounts (CABPOUTOTAL) and all five THM variables. This would indicate that input of wells water could prevent the formation of THMs.

Changes in concentrations of $CHCl_2Br$, $CHBr_2Cl$ and TTHM follow the same trends as carbon filter age and temperatures (see Fig. 4). Higher input wells water values agree with those days with THM decreasing concentration values. Rising of THM concentrations with increasing carbon filters age and temperature was also clearly detected. After carbon filter regenerations (30.04.2003), and even with water temperature increasing, THMs concentrations felt down suddenly (12.05.2003). Then, THMs concentrations kept growing until the time of new carbon filters regeneration took place. It was obvious, that during summer time, addition of wells water and regeneration of carbon filters were crucial points for the relatively low THMs concentrations encountered during this season, in spite of the higher water temperatures and of the relatively constant trends of the other WWP variables.

3.2. *PCA exploration of WWP and THM variables*

In Table 2, results of PCA analysis of the complete (autoscaled) data set including 162 daily observations of the 28 measured variables (23 WWP variables and five THM concentrations) are given. For these autoscaled variables, eight components were needed (with an eigenvalue >1) to explain 80% of the data variance (see Table 2). The four more important principal components explained around 57% of the total data variance, with the first one explaining approximately 21% of it. This indicates that the measured variables are rather independent.

Table 2 – PCA analysis for the complete data set

PCA results [**X**, **Y**] (162, 28)			
PC Nr	λ	% var	% cum var
1	5.84	20.86	20.86
2	4.06	14.49	35.35
3	3.43	12.24	47.59
4	2.59	9.26	56.85
5	2.02	7.20	64.04
6	1.91	6.81	70.86
7	1.56	5.56	76.42
8	1.08	3.85	80.27

PC Nr, number of principal component; λ, eigenvalue; % var, percentage of variance of each eigenvalue; % cum var, cumulative percentage of variance.

Fig. 5 shows PCA loadings plots. Variables are well distributed on PC1–PC2 subspace, with no one being predominant (Fig. 5a). The first PC1 is differentiating WWP variables measured after sand filters (giving negative loadings) from WWP variables measured after carbon filters (giving positive loadings). Moreover, most of THM concentrations (except for $CHBr_3$) exhibit also relatively large positive PC1 loadings. This indicates that the formation of most of THMs is favored by high water temperatures (TEMP), carbon filters age (Average and SUM) and water turbidity after carbon filters (TERBC1 and TERBC2). The opposite happens for all measured variables after sand filters. According to PC2 (as they give loadings with different signs), the formation of TTHM, $CHCl_2Br$ and $CHBr_2Cl$ is negatively correlated with input wells water flows, chloride concentrations and variables measured after sand filters. It is on the other hand outstanding the high correlation (according to PC1 and PC2) between the formation of $CHCl_3$ with turbidity after carbon filters (TERBC1 and TERBC2) and with water temperature (TEMP). It has to be pointed out, however, that the amount of explained variance considering PC1 and PC2 is still rather low (35.5%) and therefore, that the other PC plots should be explored to understand the whole problem. PC1–PC3 loadings plot (Fig. 5b) reveals similar trends for TTHM, $CHCl_2Br$ and $CHBr_2Cl$ according to PC1. According to PC3, it is obvious that water turbidity after carbon filters are not so influential for the formation of $CHCl_3$, since they take a different sign. Interesting information is obtained from PC2 to PC3 loadings plot (Fig. 5c). In this case, TTHM, $CHCl_2Br$, $CHBr_2Cl$ and even $CHBr_3$ (in smaller extend) show a high correlation with TOC and UV absorbance (ABSUV), measured after carbon filters. On the other hand, the formation of $CHBr_3$ did not exhibit any significant correlation for the first four principal components, showing a rather independent and a more complex behavior of this compound, compared to the other THMs, and also in relation to the other measured WWP variables. Everything explained until now confirms that formation of THMs exhibit a very complex pattern.

3.3. *Modeling and prediction of trihalomethane formation concentrations using multivariate regression methods*

Table 3 gives the detailed results obtained by the application of different multivariate linear regression methods for the modeling and prediction of THM formation from WWP system variables. Two data sets were used; the first one was for calibration of the model and the second one for validation of the model (see Materials and Methods section). First of all, it has to be emphasized that all data values came from real measured data values in the WWP, and that they were not obtained in the laboratory under well-stipulated and controlled conditions. It is considered that all possible sources of variance during the investigated year were included. It is expected therefore, that these data will be more complex and

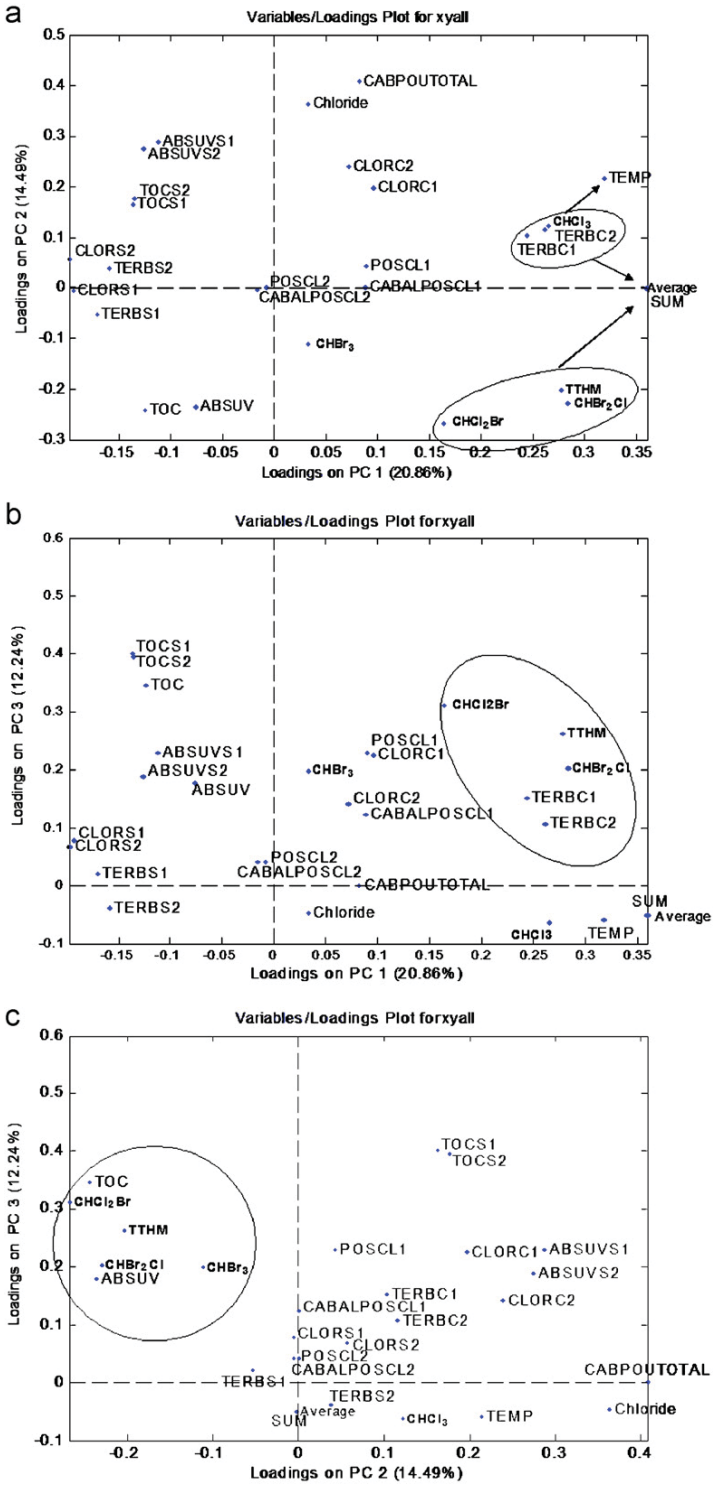


Fig. 5 – Loadings plots of the first principal components for WWP variables and THMs. Encircled or pointed by arrows are some WWP variables and some trihalomethanes, which resulted to be strongly correlated. For variable abbreviations, see Table 1.

difficult to interpret than those reported in previous related works obtained under much more controlled conditions (Garcia-Villanova et al., 1997).

Overall, MLR and PLSR performed best and gave similar good predictive properties. Both methods, MLR and PLSR, explained relatively the same percentage of variance of the

Table 3 – Comparison between predicted and actual values with four chemometric analyses for the calibration and validation data

	Calibration data					Validation data				
	Statistics	MLR	SW	PCR	PLS	Statistics	MLR	SW	PCR	PLS
Bromoform	RMSEP	4.0844	4.6825	4.0867	4.1093	RMSEC	5.1099	5.9032	5.1372	5.0628
	SEP	4.0986	4.6989	4.1010	4.1236	SEC	5.2472	5.9729	5.2709	5.1826
	Bias	0.0000	−0.0000	0.0000	−0.0000	Bias	−0.3289	1.0742	−0.3900	−0.5148
$CHBr_3$	Rel. error in %	21.2340	24.5181	21.2465	21.3694	Rel. error in %	28.5879	30.9379	28.8088	28.5520
	Slope	1.0000	1.0000	1.0000	1.0000	Slope	0.5077	−0.4741	0.4968	0.5333
	Offset	−0.0000	−0.0000	−0.0000	0.0000	Offset	8.9814	26.9037	9.2043	8.6310
	Corcoef	0.6712	0.5270	0.6707	0.6661	Corcoef	0.3227	−0.1822	0.3251	0.3612
	R^2	0.45047	0.2777	0.44985	0.44374	R^2	0.10417	0.0331	0.10571	0.13046
Chloroform	RMSEP	3.2501	3.4756	3.6738	3.3000	RMSEC	4.2615	4.2188	3.7916	4.0561
	SEP	3.2615	3.4877	3.6866	3.3115	SEC	4.3094	4.3350	3.8959	4.1378
	Bias	−0.0000	−0.0000	0.0000	0.0000	Bias	0.7884	0.2243	−0.2026	0.5308
$CHCl_3$	Rel. error in %	20.8751	22.3932	23.7404	21.2095	Rel. error in %	24.7859	25.3746	23.4355	23.9391
	Slope	1.0000	1.0000	1.0000	1.0000	Slope	1.0155	1.0295	1.2167	1.0640
	Offset	−0.0000	0.0000	−0.0000	0.0000	Offset	−1.0492	−0.7032	−3.2222	−1.5889
	Corcoef	0.7242	0.6755	0.6265	0.7140	Corcoef	0.6650	0.6601	0.7495	0.6981
	R^2	0.5245	0.4563	0.39245	0.50981	R^2	0.44219	0.4358	0.56182	0.48739
Bromo-dichloromethane	RMSEP	4.1160	4.6055	4.5685	4.1824	RMSEC	4.4372	4.9474	4.8295	4.4480
	SEP	4.1303	4.6215	4.5844	4.1970	SEC	4.3680	5.0763	4.9000	4.4145
	Bias	−0.0000	−0.0000	−0.0000	0.0000	Bias	−1.2922	−0.3728	−0.8047	−1.1743
$CHCl_2Br$	Rel. error in %	23.2742	26.2218	25.9968	23.6706	Rel. error in %	25.0698	26.7312	26.8763	24.9708
	Slope	1.0000	1.0000	1.0000	1.0000	Slope	1.1910	1.1249	1.4628	1.1821
	Offset	−0.0000	0.0000	−0.0000	0.0000	Offset	−1.9926	−1.8906	−7.3791	−1.9794
	Corcoef	0.7499	0.6723	0.6788	0.7403	Corcoef	0.7677	0.6560	0.7183	0.7608
	R^2	0.56228	0.4520	0.46074	0.54804	R^2	0.58943	0.4303	0.51589	0.57889
Dibromo-chloromethane	RMSEP	3.4480	3.6559	4.0807	3.5512	RMSEC	5.3420	5.4852	5.4745	5.3062
	SEP	3.4600	3.6686	4.0949	3.5636	SEC	5.4655	5.6442	5.6324	5.4523
	Bias	0.0000	0.0000	0.0000	−0.0000	Bias	−0.5700	0.0008	0.0920	−0.2828
$CHBr_2Cl$	Rel. error in %	13.9238	14.7811	16.5432	14.3491	Rel. error in %	21.1285	21.2906	21.2386	20.7773
	Slope	1.0000	1.0000	1.0000	1.0000	Slope	1.1016	1.1885	1.3856	1.1443
	Offset	0.0000	−0.0000	−0.0000	0.0000	Offset	−1.9526	−4.7896	−9.9233	−3.3421
	Corcoef	0.7970	0.7680	0.6993	0.7830	Corcoef	0.7018	0.6827	0.7035	0.7062
	R^2	0.63521	0.5899	0.48905	0.61304	R^2	0.49251	0.4660	0.49491	0.49872
Total Trihalomethanes	RMSEP	8.9115	9.4654	10.6972	9.0330	RMSEC	12.5534	13.1849	13.0648	12.7487
	SEP	8.9427	9.4985	10.7346	9.0645	SEC	12.8337	13.5669	13.4427	13.0574
	Bias	0.0000	0.0000	0.0000	0.0000	Bias	−1.4257	−0.0757	−0.1515	−1.2274
TTHM	Rel. error in %	11.6692	12.4053	14.0497	11.8304	Rel. error in %	16.2255	16.7954	16.6992	16.4405
	Slope	1.0000	1.0000	1.0000	1.0000	Slope	0.9017	0.9297	1.0901	0.8913
	Offset	0.0000	−0.0000	0.0000	−0.0000	Offset	8.9348	5.5400	−6.8426	9.5531
	Corcoef	0.7980	0.7683	0.6905	0.7918	Corcoef	0.6666	0.6126	0.6222	0.6524
	R^2	0.63688	0.5903	0.47678	0.62692	R^2	0.44431	0.3753	0.38717	0.42564

RMSEC and RMSEP, root mean squared error of calibration and prediction; SEP, standard error of performance; Bias, average values of residuals; Rel.error in %, relative error of concentration prediction in percentage; Slope, the slope; Offset, the intercept; Corcoef, correlation coefficient; R^2, the statistical coefficient of determination.

THM concentrations (see R^2 in Table 3). Calibration errors (RMSEC), prediction errors (RMSEP) and percentage relative errors were also similar for these two methods. For instance, for the prediction of total trihalomethane concentrations, TTHM, modeling errors in calibration were, respectively (RMSEC and Rel. error %) 8.9% and 11.7% for MLR and 9.0% and 11.8% for PLSR, and in validation were, respectively (RMSEP and Rel. error %) 12.5% and 16.2% for MLR and 12.7% and 16.4% for PLSR. Similar tendencies were observed for the prediction of individual THM concentrations. Errors in calibration and validation obtained by PCR and SWR were always higher in comparison to MLR and PLSR errors. The only exception was for $CHBr_3$ using PCR, which gave more accurate estimations than using MLR or PLSR. Bias in calibration was always practically zero, whereas in validation bias values were always different than zero, although it is difficult to infer any systematic trend.

In Table 3, slopes, offsets and correlation coefficients of the best lines, obtained when measured and predicted values of the trihalomethane concentrations were compared, are given. For calibration data, slope and offset values were always one and zero, respectively, for all the compared methods and for all THMs. Correlation coefficients were always around 0.7 and 0.8, which can be considered rather good taking into account the nature of the data. In all cases, the two best methods were MLR and PLSR, especially for total trihalomethane concentrations, TTHM (0.798 and 0.792), and for $CHBr_2Cl$ (0.797 and 0.783). The same tendency was observed for validation data, although now with worse values for the slope, offset and correlation coefficient values. Especially bad was the case of $CHBr_3$, where the results were rather poor, with slopes around 0.5, offsets around 9 and correlation coefficients around 0.3. The models for $CHBr_3$ always failed because of the difficulty to model the presence of a little number of days with very extreme concentration values, while the rest of days showed practically constant $CHBr_3$ concentration values without showing any clear dependency with WWP variables. On average, $CHBr_3$ prediction errors for validation data were around 29–30% for all methods. Slightly better predictions were obtained for $CHCl_3$ with average relative errors around 23–25% for validation data, and also with better agreement between experimental and predicted concentrations, with slopes close to one, lower offsets and correlation coefficients around 0.7. Even better results were obtained for $CHBrCl_2$ and $CHBr_2Cl$, with average relative prediction errors closer to 20% for validation data. For calibration data the improvement of figures of merit followed the ranking $CHBr_3$ (worse), $CHBrCl_2$, $CHCl_3$, $CHBr_2Cl$, TTHM (best).

Comparison between predicted and actual experimental values of total trihalomethane (TTHM) concentrations using the external validation data set is shown in Fig. 6a for all the tested methods. Predicted values for TTHM concentrations followed the same trend than experimental values. However, predictions were worse in some cases for some extreme experimental observations. Five days can be distinguished in this validation data set where all methods gave systematically lower concentrations than measured ones and three other days where the opposite occurred, giving systematically higher concentrations than they should be. These differences were, however, always around or below 20% of the total concentration. When WWP variables were carefully examined for these five extreme observations, nothing unusual could explain why these THM concentrations were so extreme for these days. There was no method, which could explain consistently why these extreme high and low concentrations did happen because WWP variables did not show any trend for these days. For calibration data, PLSR predictions (Fig. 6b) were better, but the presence of extreme concentrations for some days was also present. Therefore, the final conclusion is that these extreme values could not be well explained by the measured WWP variables whatever was the regression method used.

When loading weights for this first PCR and PLSR latent variables together with statistically significant coefficients at

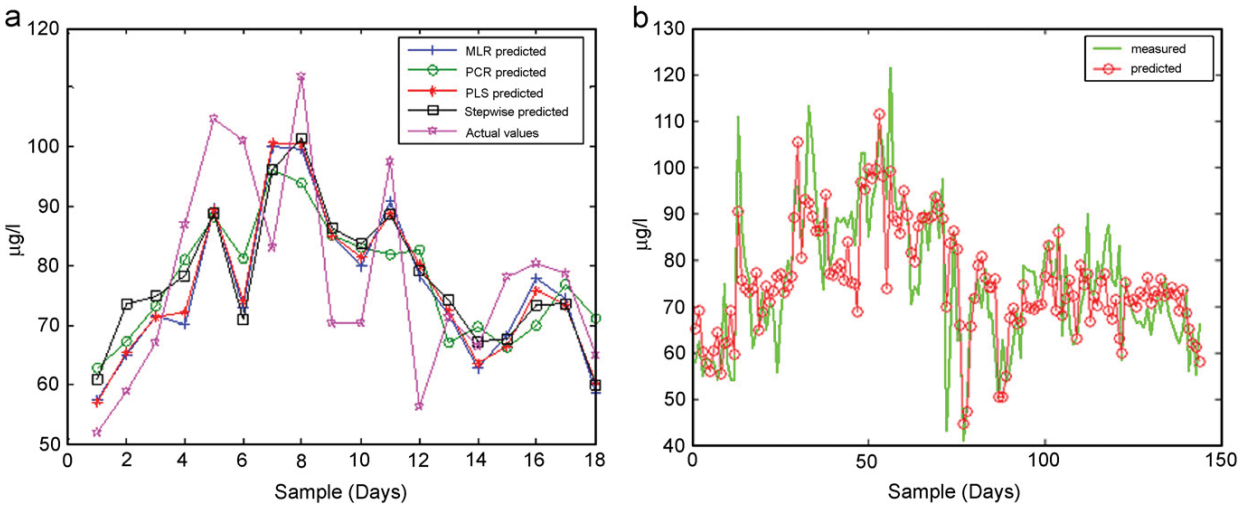


Fig. 6 – Predicted and actual experimental values of TTHM: (a) using external validation data set and all regression methods; (b) using calibration data set and PLSR.

the 5% level, obtained with MLR and SWR, were examined, the more influential WWP variables for $CHCl_2Br$, $CHCl_2Br$, $CHCl_3$ and TTHM formation were UV absorbance after sand filters (at site 2) and after carbon filters, total organic carbon concentration after carbon filters, total amount of input wells water, water temperature, water turbidity after carbon filters, carbon filters age and post-chlorination at site 1. It was difficult to reveal the most important WWP variables in the case of $CHBr_3$ formation since according to their loading weights, they had very little influence on it.

Some possible model improvements were also investigated. First, the possibility of a time lag between measurements was considered. In fact, measurements were not performed at exactly the same time instant. In some cases, they were averaged over the whole day and in other cases they were simply punctual measured at a particular time instant. However, attempts to improve these results using different time shifts and measure delays did not improve significantly the results. A second possibility was the presence of experimental measurement errors in THM concentrations, but lab analysis did confirm that extreme values were correct. Even after log or other non-linear transformation of $\mathbf{X}$ and $\mathbf{y}_i$ data were applied, no significant improvement in THM concentration predictions were obtained. Selection of variables using methods like genetic algorithms (Leardi, 2001) did not either improves the results in terms of predictive properties for the different validation data sets.

4. Conclusions

The main conclusions derived from the present work are the following:

(1) Due to water disinfection techniques and procedures implemented in Sant Joan Despí WWP (Barcelona, Spain), THM compounds are ubiquitously formed at relative large amounts, especially in spring season, and with predominance of brominated type of compounds over chloroform.
(2) Twenty-three different WWP variables were measured and investigated for their possible correlation with measured concentrations of the different THM compounds at the exit of the plant. Among these WWP variables, water temperature, carbon filters age, water turbidity after carbon filters, as well as input wells water amount, UV absorbance, total concentration of organic carbon (the two later after carbon filters) and post-chlorination were found to have significant correlations and influence formation of THM compounds. Some of these variables, like carbon filter age and input wells water can be modulated by appropriate plant management to decrease the levels of THM concentrations at the exit of the plant.
(3) Using these WWP variables, multivariate regression techniques, such as MLR and PLS, allowed for a good average THM concentration modeling and prediction. Average relative errors ranged between 12% and 23% for the modeling of THM concentrations in calibration samples and between 16% and 29% in external validation samples for all investigated trihalomethanes. Total trihalomethane concentrations were predicted the best, showing an average effect for this variable compared to the other individual THM concentration variables, more difficult to predict in general, and especially for bromoform.
(4) Some extreme THM concentration values (very low or very high) were not well explained yet, especially for bromoform and chloroform concentrations. One possible explanation for this is that additional WWP variables reflecting better the changes in trihalomethane concentrations were needed. Current research work is also performed about the possible application of non-linear modeling methods such as neural networks (Rodríguez et al., 2003b), or of other more complex non-linear methods (Lin and Yeh, 2005). Future work will also consider the possible use of bromide anion concentrations, pH, temperature of wells water, reaction times after chlorination, potential of THM formation and others, which have been also suggested to be influential (Nikolaou, 2004).

Acknowledgments

AGBAR (Barcelona's Water Company) is acknowledged for financial help to carry out this research and especially to the people working at the 'Àrea de Control del Tractament' (Antoni Bernal) and at the 'Àrea de Química Orgànica' (Francesc Ventura), who provided the experimental data sets used for this study.

REFERENCES

Amy, G.L., Chadyk, P.A., Chowdhury, Z.K., 1987. Developing models for predicting trihalomethane formation potential and kinetics. J. Am. Water Works Assoc. 79 (7), 89–97.
APHA, AWWA, WEF, 1998. Standard Methods for the Examination of Water and Wastewater, 20th ed. American Public Health Association, Washington, DC.
Clark, R.M., 1998. Chlorine demand and TTHM formation kinetics: a second order model. J. Environ. Eng. ASCE 124 (1), 16–24.
Clark, R.M., Thurnau, R.C., Sivaganesan, M., Ringhand, P., 2001. Predicting the formation of chlorinated and brominated by-products. J. Environ. Eng. ASCE 127 (6), 493–501.
Donachie, A., Walmsley, A.D., Haswell, S.J., 1999. Application and comparisons of chemometric techniques for calibration modelling using electrochemical/ICP-MS data for trace elements in UHQ water and humic acid matrices. Anal. Chim. Acta 378 (1–3), 235–243.
Draper, N., Smith, H., 1981. Applied Regression Analysis. Wiley, New York.
EEC, 1998. Council Directive 98/83/EC of 3 November 1998 on the quality of water intended for human consumption. Official Journal of the European Communities L 330/32, 5.12.98.
Gallard, H., von Gunten, U., 2002. Chlorination of natural organic matter: kinetics of chlorination and of THM formation. Water Res. 36 (1), 65–74.
Garcia-Villanova, R.J., Garcia, C., Gomez, J.A., Garcia, M.P., Ardanuy, R., 1997. Formation, evolution and modelling of trihalomethanes in the drinking water of a town: I. At the municipal treatment utilities. Water Res. 31 (6), 1299–1308.
Geladi, P., Kowalski, B.R., 1986. Partial least-squares regression: a tutorial. Anal. Chim. Acta 185, 1–17.
Golfinopoulos, S., Arhonditsis, G., 2002. Multiple regression models: a methodology for evaluating trihalomethane

concentrations in drinking water from raw water characteristics. Chemosphere 47 (9), 1007–1018.
Golfinopoulos, S., Xilourgidis, N., Kostopoulou, M., Lekkas, T., 1998. Use of a multiple regression model for predicting trihalomethane formation. Water Res. 32 (9), 2821–2829.
Höskuldsson, A., 1988. PLS regression methods. J. Chemometr. 2 (3), 211–228.
Jolliffe, I.T., 1986. Principal Component Analysis. Springer, New York.
Leardi, R., 2001. Genetic algorithms in chemometrics and chemistry: a review. J. Chemometr. 15 (7), 559–569.
Lin, Y.C., Yeh, H.D., 2005. Trihalomethane species forecast using optimization methods: genetic algorithms and simulated annealing. J. Comput. Civil Eng. 19 (3), 248–257.
Li, X., Zhao, H., 2006. Development of a model for predicting trihalomethanes propagation in water distribution systems. Chemosphere 62 (6), 1028–1032.
Martín-Alonso, J., 2003. Combine use of surface and groundwater for drinking water production in the Barcelona metropolitan area. In: Proceedings of the II International Riverbank Filtration Conference Cincinnati, USA.
Massart, D.L., Vandeginste, B.G.M., Buydens, L.M.C., de Jong, S., Lewi, P.J., Smeyers-Verbeke, J., 1998. Handbook of Chemometrics and Qualimetrics. Elsevier Science, Amsterdam.
McGeehin, M.A., Reif, J.S., Becher, J.C., Mangione, E.J., 1993. Case-control study of bladder cancer and water disinfection methods in Colorado. Am. J. Epidemiol. 138 (7), 492–501.
Milot, J., Rodríguez, M.J., Sérodes, J.B., 2002. Contribution of neural networks for modeling trihalomethanes occurrence in drinking water. J. Water Res. Pl. ASCE 128 (5), 370–376.
Naes, T., Martens, H., 1988. Principal component regression in NIR analysis: viewpoints, background details and selection of components. J. Chemometr. 2 (2), 155–167.
Nikolaou, A.D., 2004. Investigation of the formation of chlorination by-products in water rich in bromide and organic matter content. J. Environ. Sci. Health A39 (11–12), 2835–2853.
Nikolaou, A.D., Golfinopoulos, S.K., Arhonditsis, G.B., Kolovoyiannis, V., Lekkas, T.D., 2004. Modeling the formation of chlorination by-products in river waters with different quality. Chemosphere 55 (3), 409–420.
Richardson, S., 2002. The role of GC-MS and LC-MS in the discovery of drinking water disinfection by-products. J. Environ. Monit. 4 (1), 1–9.
Rodríguez, M.J., Sérodes, J.B., 2001. Spatial and temporal evolution of trihalomethanes in three water distribution systems. Water Res. 35 (6), 1572–1586.
Rodríguez, M.J., Sérodes, J.B., 2004. Application of back-propagation neural network modeling for free residual chlorine total trihalomethanes and trihalomethanes speciation. J. Environ. Eng. Sci. 3 (S1), S25–S34.
Rodríguez, M.J., Vunette, Y., Sérodes, J.B., Bouchard, C., 2003a. Trihalomethanes in drinking water of greater Quebec region (Canada): occurrence, variations and modelling. Environ. Monit. Assess. 89 (1), 69–93.
Rodríguez, M.J., Milot, J., Sérodes, J.B., 2003b. Predicting trihalomethane formation in chlorinated waters using multivariate regression and neural networks. J. Water Supply Res. Technol. 52 (3), 199–215.
Rodrigues, P.M.S.M., Esteves da Silva, J., Antunes, M.C., 2007. Factorial analysis of the trihalomethanes formation in water disinfection using chlorine. Anal. Chim. Acta, in press, doi:10.1016/j.aca.2006.12.031.
Rook, J.J., 1974. Formation of haloforms during chlorination of natural waters. Water Treat. Exam. 23 (2), 234–243.
Simpson, K.L., Hayes, K.P., 1998. Drinking water disinfection by-products: an Australian perspective. Water Res. 32 (5), 1522–1528.
Ventura, F., Rivera, J., 1985. Factors influencing the high content of brominated trihalomethanes in Barcelona's water supply (Spain). B. Environ. Contam. Toxicol. 35, 73–81.

Results and Discussion

- Results of descriptive statistical analysis of THMs

Preliminary data inspection was performed to detect and remove outliers, as well as to preliminarily identify data trends.

The preliminary inspection and visualisation of the observed concentrations of all four target THMs and of their total sum (TTHM) revealed that the concentrations of mixed chloro-bromo THMs were higher than the concentrations of chloroform and bromoform during the sampling year. However, the main problem was detected to be the formation of high concentrations of multi-brominated THMs in the SJD DWTP. Their average concentration levels were found to exceed those of the multi-chlorinated THMs, most likely due to the high concentrations of bromide ions, which are naturally present in the Llobregat River.

The seasonal average concentrations of $CHClBr_2$ were found to be the highest concentration during the sampled period. The $CHClBr_2$ considerably varied from 12.55 to 45 µg/l, with the highest average value found in spring. The second and third levels of concentration formation were $CHCl_2Br$ and $CHBr_3$, ranging from 3.7 to 35.7µg/l and between 6.7 and 61 µg/l respectively. $CHCl_3$ formation was lower in SJD DWTP but an increase in its concentrations was observed for the summer season. $CHCl_3$ concentrations varied between 6 and 30.9 µg/l. Additionally, a seasonal trend in THMs formation was found, where the highest concentrations of total THMs were detected during spring. The next season with highest total THMs concentrations was summer. Winter and autumn seasons were characterized with lower formation of THMs concentrations. TTHM concentrations varied between 40.9 and 121.6 µg/l.

- Results of pair-wise correlation statistical analysis of DWTP operational parameters and THM concentrations

The pair-wise correlations between $CHCl_3$, $CHBr_2Cl$ and TTHM concentrations and the age of carbon filters variables were found high and positive (please refer to Figure 3 of Article 1). Water temperature was also found to have a high positive correlation with $CHCl_3$. High positive correlations were also observed between $CHCl_2Br$ and total organic carbon, and between $CHBr_2Cl$ and water turbidity measured after carbon filters. The highest negative correlation was observed between the added amounts of underground water and all four THMs variables.

An alternative presentation of the levels and significance of pairwise correlations is Figure 4 of Article 1, where annual trends of some of the most important operational

parameters and selected THMs variables were plotted over time. The figure displays, for example, that the annual trend of carbon filter age and temperature were largely similar to those observed for $CHCl_2Br$, $CHBr_2Cl$ and TTHM concentrations. Higher input of underground water amounts (i.e., a large increase of the trend line) coincide with THMs decreased concentrations (i.e., a large decrease in the trend line). An interesting trend was observed on 30.04.2003, when a significant decrease of THMs was captured and this pattern conceded in time with carbon filters regeneration. On the basis of these trends, a conclusion can be drawn that large volumes of underground water pumped to fulfil water demand in summer and regeneration of carbon filters, result in a reduction of THMs concentrations.

These findings suggested that the aging of carbon filters is significantly associated with a decrease of their efficiency, resulting in increased turbidity. Thus, the formation of THMs was favoured by the turbidity increase. It was additionally observed that higher temperature in summer season had also favoured THMs formation. This factor was particularly relevant for CHCl3 formation. The observation that pumping underground water reduced THMs could be attributed to the produced dilution effect on THMs precursors such as organic matter (measured as TOC) and Br^- ions concentrations thus preventing the formation of THMs.

The univariate statistics analysis was extended to multivariate data analysis to better understand the SJD DWTP processes.

- Results of Principal Component exploratory data Analysis (PCA) of 23 SJD DWTP operational parameters and THMs concentrations

PCA analysis was performed on the autoscaled dataset containing 162 daily observations of the 28 measures (23 DWTP variables and 5 THM concentrations).The obtained model was found to be very complex, because eight principal components (PCs) with an eigenvalues larger than 1 were required to explain 80% of data variance (please refer to Table 2 in Article 1). The first four PCs explained approximately 57% of the total data variance, with PC1 explaining approximately a third (21%).

The detailed examination of PC1 loadings suggested that the formation of THMs was favoured by high water temperatures, carbon filters age and water turbidity after carbon filters, and that it was negatively correlated with DWTP parameters monitored after sand filters (the previous process in the SJD-DWTP operational scheme, see Figure 3). According to PC2 and PC3 loadings, the formation of TTHM, $CHCl_2Br$ and

$CHBr_2Cl$ was negatively correlated with input underground water amounts, chloride and chlorine residual concentrations after carbon filter. In contrast, TTHM, $CHCl_2Br$, $CHBr_2Cl$ and even $CHBr_3$ (to a smaller extend) showed a high positive correlation with total organic carbon (TOC) and UV absorbance, measured after carbon filters. Taken together, the analysis suggested that the formation of brominated THMs is likely to be highly dependent on the residual amounts of organic matter, left after filtration processes.

- *Results of application of linear empirical models for trihalomethanes formation prediction*

The comparison of linear regression results obtained following the application of SWR, MLR, PCR and PLS regression methods are presented in Table 3 of Article 1. The results obtained from MLR and PLS predictions suggested that both methods had good predictive properties for THMs modeling and performed better than SWR and PCR. For instance, prediction errors for TTHM calibration (RMSEC/Rel. error %) were of 8.9 - 11.7% for MLR and of 9.0-11.8% for PLS. At the external validation stage, prediction errors (RMSEP/Rel. error %) were respectively of 12.5-16.2% for MLR and of 12.7-16.4% for PLS. Similar results were observed for the prediction of the individual THM concentrations. Prediction errors in calibration and external validation obtained with PCR and SWR methods were always higher (around 3-4 %) in comparison to MLR and PLS prediction errors, except for $CHBr_3$ using PCR, which gave more accurate estimations than using MLR or PLS. Interestingly, MLR performed well in these cases. Most probably, this fact was observed because of two reasons: 1) the most important operational parameters for the THMs formation were rather independent among them, and 2) a large number of samples have been included in the calibration data.

On average, at the validation stage, the $CHBr_3$ prediction errors were around 29-30% for all methods. Better predictions were obtained for $CHCl_3$ with average relative errors of approximately 23-25% for validation data, and also with better agreement between experimental and predicted concentrations, with slopes close to one, smaller offsets and correlation coefficients of about 0.7. The best results were obtained for $CHBrCl_2$ and $CHBr_2Cl$, with average relative prediction errors closer to 20% for validation data. For calibration data, the improvement of figures of merit had the

following order (worse to best): $CHBr_3$ (worse) > $CHBrCl_2$ > $CHCl_3$ > $CHBr_2Cl$ > TTHM (best).

Figure 6a of Article 1 compares predicted and actual experimental values of total trihalomethane (TTHM) concentrations using the external validation data set for all tested regression methods. The obtained prediction values for TTHM concentrations followed a trend similar to such of the experimentally observed values. However, predictions were worse in some cases, especially for extreme experimental observations. Five days, when all methods gave systematically lower concentrations than the measured ones, were distinguished in this validation data set. In three different days, the reverse trend was observed, where predicted concentrations were systematically higher than the real data. However, such differences were always around 20% of the total concentration. No method could consistently explain such extreme high and low concentrations, likely because DWTP variables did not exhibit a particular trend during these days. For calibration data, PLS predictions were better, but extreme concentrations for some days were also present (please refer to Fig. 6b of Article 1). Thus, the final conclusion is that these extreme values could not be explained by the measured DWTP variables, independently of the linear regression method used.

- *Evaluation of more important variables in linear regression models for 23 DWTP operational parameters and THMs*

The most influential DWTP parameters for $CHClBr_2$, $CHCl_2Br$, $CHCl_3$ and TTHM formation were found after examining loadings and loadings weights for the first PCR and PLS latent variables, along with the statistically significant coefficients (obtained by MLR and SWR). UV absorbance measurements after sand and carbon filters, total organic carbon concentration after carbon filters, total amount of input underground water, water temperature, water turbidity after carbon filters, carbon filters age and post-chlorination were among the most important and statistically significant DWTP operational parameters for the THMs formation. In the case of $CHBr_3$ formation, the most important WWP variables were difficult to identify.

3.1.2 Article 2 – Platikanov, S., Martin, J. and R. Tauler. *Linear and non-linear chemometric modeling of THM formation in Barcelona's water treatment plant.* Science of Total Environment 432 (2012) 365-374.

Introduction

This paper builds on our previous research work by augmenting the model with new parameters using the data from DWTP in Sant Joan Despí, Barcelona. In the first study, 23 WTP variables representing the late stages of the disinfection process in the plant (such as after the sand filtration process) were monitored inside the plant (see Figure 3). Predictions for externally validated data had average errors between 15 and 30% for the concentrations of all THMs. However, several important DWTP parameters were previously ignored in the analysis and such parameters were expected to be critical for improving the predictive power of estimated models. The literature review suggested that some DWTP parameters, including raw water quality and the prechlorination pre-treatment, were important for the THMs formation process. As a result, eighteen new plant parameters from the two earlier stages of the DWTP operational processes were included in the new dataset analysis in order to improve the THMs prediction results of the first article. The new investigation, included in second article, considered the interactions among 41 DWTP parameters over a period of 162 days, similar to the first study. A significant improvement over the first study was the inclusion in the regression models of all observable DWTP operational parameters: from raw water quality at the entrance to the drinking water at the exit of the plant. The expectation was that the inclusion of a comprehensive set of DWTP parameters would describe better the THMs formation reaction in the plant.

Another novelty of this study was the application of non-linear regression techniques, such as Kernel type Partial Least Squares Regression (K-PLS) using radial basis functions kernel and Support Vector Machine Regression (SVR), to model possible non-linear relationships among all monitored DWTP variables and THMs concentrations in finished drinking water. Additionally, a new visualization technique was applied to capture complex relationships among operational variables and THMs concentrations.

Science of the Total Environment 432 (2012) 365–374

Contents lists available at SciVerse ScienceDirect

Science of the Total Environment

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

Linear and non-linear chemometric modeling of THM formation in Barcelona's water treatment plant

Stefan Platikanov [a], Jordi Martín [b], Romà Tauler [a,*]

[a] Department of Environmental Chemistry, IDAEA-CSIC, Jordi Girona, 18-26, Barcelona 08026, Spain
[b] Fundació AGBAR, Sant Joan Despí, 1, 08940 Cornellà de Llobregat, Barcelona, Spain

ARTICLE INFO

Article history:
Received 12 March 2012
Received in revised form 22 May 2012
Accepted 31 May 2012
Available online 1 July 2012

Keywords:
Drinking water
Disinfection by-products
Trihalomethanes
Linear models
Kernel-PLS
SVM regression

ABSTRACT

The complex behavior observed for the dependence of trihalomethane formation on forty one water treatment plant (WTP) operational variables is investigated by means of linear and non-linear regression methods, including kernel-partial least squares (K-PLS), and support vector machine regression (SVR). Lower prediction errors of total trihalomethane concentrations (lower than 14% for external validation samples) were obtained when these two methods were applied in comparison to when linear regression methods were applied. A new visualization technique revealed the complex nonlinear relationships among the operational variables and displayed the existing correlations between input variables and the kernel matrix on one side and the support vectors on the other side. Whereas some water treatment plant variables like river water TOC and chloride concentrations, and breakpoint chlorination were not considered to be significant due to the multi-collinear effect in straight linear regression modeling methods, they were now confirmed to be significant using K-PLS and SVR non-linear modeling regression methods, proving the better performance of these methods for the prediction of complex formation of trihalomethanes in water disinfection plants.

1. Introduction

Water disinfection procedures using chlorine are a global practice for reducing the health risk of pathogenic growth in drinking water treatment processes. Despite the crucial importance of this strategy, several classes of undesirable disinfection by-products (DBP) are usually identified in potable waters (Rook, 1974). Chlorine reacts with natural organic matter (NOM) in raw water, resulting in the formation of trihalomethanes (THMs), haloacetic acid, haloacetonitriles and other chemical compounds (Richardson, 2003). This study focuses on the formation of trihalomethanes, in particular, of chloroform ($CHCl_3$), dichlorobromomethane ($CHCl_2Br$), dibromochloromethane ($CHBr_2Cl$), bromoform ($CHBr_3$) and of the total sum of them (TTHM) during a classical water treatment procedure. The presence of THMs in drinking water is undesirable since they are considered possible carcinogen products and they have been related with reproductive and development problems (Calderon, 2000; Reif et al., 1996). Health-risk concerns have resulted in setting out new fixed limits for maximum TTHM concentration levels equal to 100 μg/l since the year 2009 in EU countries (EC, Council Directive 98/83/EC).

THM formation depends significantly on raw water quality as well as on water treatment procedures. Natural organic matter and especially its dissolved fractions in water (DOM, humic and fulvic acids) have long been recognized as the main precursors for THM formation (von Gunten et al., 2001). Recently, it was found that the concentration, functionality and aromaticity characteristics of the organic matter are important due to different reactivity of DOM fractions with chlorine, originating from different DBP (Croué, 2004). Among other important factors in THM formation, the presence of relatively high bromide anion concentrations, water temperature, ammonia and other variables have been described to be important (Nikolaou, 2004). Also, THM formation depends on operational procedures implemented in water treatment plants (WTP), such as chlorine dose, contact time between chlorine and organic matter, pH and others (Navalon et al., 2008; Sadiq and Rodriguez, 2004).

Previous investigations have been carried out to obtain knowledge about the existing relationship among raw water variables, operational plant managing variables and THM formation in drinking water. Different types of models for THM formation have been described in the literature both, in WTP real studies (Rodríguez et al., 2003), and in laboratory-scale studies (Nikolaou, 2004). Some of these models have been based on the study of the kinetics of the formation of THMs (Gallard and von Gunten, 2002). Other models have been based on chemometric multivariate empirical regression methods including a number of operational and water quality variables as predictors and the generated THM concentrations as predicted variables (Rodríguez and Sérodes, 2001; Toroz, and Uyak, 2005). Different chemometric methods have been already proposed

* Corresponding author. Tel.: +34 934006140; fax: +34 932045904.
E-mail address: rtaqam@idaea.csic.es (R. Tauler).

0048-9697/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
doi:10.1016/j.scitotenv.2012.05.097

for THM formation modeling, such as the multilinear regression analysis (Golfinopoulos and Arhonditsis, 2002), MLR, and partial least squares regression, PLSR (Serrano and Gallego, 2007). Additionally, non-linear methods such as the neural networks and logistic regression analysis (Milot et al., 2002; Rodríguez and Sérodes, 2004) have also been shown to give reasonably good predictions. A lot of information about the existing models of disinfection by-product formations is collected in the review article of Chowdhury et al. (2009)

Most of the existing investigations at present have been performed over short monitoring times using a reduced number of experimental and operational WTP variables. In THM research field, there is still a demand for reliable and robust models for in-situ prediction of THM formation. These models should be applied under real time plant operations and should allow for the on-line control of THM formation. Also, information about spatial and temporal THM formation conditions at the plant, when their concentrations are too high or when they should be limited to levels below established legal and sanitary limits, is an important aspect to consider (Rodríguez and Sérodes, 2001). In this work, an investigation is performed about which factors determine the THM formation in WTP by classical chlorination disinfection procedures. This study continues our previous research work (Platikanov et al., 2007) taking as model example the WTP located in Sant Joan Despí, Barcelona, Spain. Nowadays, this plant provides drinking water for around 50% of Barcelona's area population of 3,000,000. In the first investigated time period, only 23 WTP variables were monitored inside the plant and they were selected during the late stages of the disinfection process in the plant, e.g. after the prechlorination step. Predictions for externally validated data gave average errors between 15 and 30% for the total concentrations of all THMs. Several important missing WTP variables were suspected to be influential for a possible improvement of prediction models. Especially, the information related with incoming raw water quality (Ventura and Rivera, 1985) and variables that characterize the prechlorination pre-treatment procedures were missing, although they were considered to be influential in THM formation (Nissinen et al., 2002).

It is for this reason that in this new study, eighteen new plant variables were included in the data analysis to improve previous results and allow for an optimal management and design of water treatment plant and prediction of THM formation under disinfection procedures. The present investigation monitors the interactions among 41 WTP variables (in previous study they were only 23) including those from incoming raw river water, at the prechlorination step, filtering and at the postchlorination step. This new study investigates the effects of including these new variables and the possible non-linear relationships among all monitored WTP variables. Therefore, in this work, apart from the prediction abilities of linear models like MLR and PLSR, new nonlinear modeling methods like the support vector machine, SVR, (Smola and Schölkopf, 2004) and kernel partial least squares, K-PLS (Walczak and Massart, 1996a), regression methods have been used for optimal prediction of THM formation at plant conditions. Support vector machine methods are now already well established chemometric procedures to solve complex classification and regression problems. It is preferred to other methods like the artificial neural networks (ANN) for its higher generalization performance and for its ability to model non-linear relationships in a unique and global manner (Scholkopf and Smola, 2002). K-PLS, on the other side, has been shown (Walczak and Massart, 1996b) to be also a very powerful and easy to apply regression technique, with equal or even better prediction abilities of non-linear data sets than other methods. Both nonlinear techniques, SVR and K-PLS, use a kernel function making possible the non-linear relationship modeling in an optimal way. This kernel function projects input data variables, usually non-linearly related to the associated output, into a multidimensional feature space where the non-linear relationship is represented in a linear form. In this study, SVR and K-PLS were applied as prediction regression tools. In recent literature (Krooshof et al., 2010; Postma et al., 2011; Üstün et al., 2007), different procedures have been proposed to visualize and understand K-PLS and SVR models based on the visualization and interpretation of kernel matrix correlations with input variables. One of these techniques has been also tested and discussed in this work and it is proposed as a useful tool for process monitoring and control of water-disinfection plant performance.

2. Theory

The main goal of this study is to build a multivariate regression model able to explain and predict the formation and changes observed in the concentration of the different investigated trihalomethane compounds ($\mathbf{y}_i$ variables) as a function of 41 measured WTP variables (**X** block of variables). This implies finding a mathematical relationship between these two sets of variables, **X** and **y**. In particular, in this work, four different multivariate regression methods have been tested for optimal modeling of trihalomethane formation in WTP disinfection procedures: multilinear regression (MLR), partial least squares regression (PLSR), kernel partial least squares regression (K-PLS) and support vector machine regression (SVR). The first two are examples of linear regression methods and the latter two are examples of non-linear regression methods (Rosipal and Trejo, 2001). Only a brief description of the main aspects of them will be given and previous references are given for a more detailed description of them.

2.1. Multilinear regression (MLR)

MLR method maximizes the covariance between **X** and **y** data sets to obtain optimal estimations of **y**. MLR predictive models will suffer from the presence of many highly correlated variables in the **X** data matrix block, which leads to unstable regression equations with a difficult interpretation of regression coefficients. These regression coefficients are estimated by ordinary least squares (Massart et al., 1998).

2.2. Partial least squares regression (PLSR)

PLSR (Geladi and Kowalski, 1986) maximizes the covariance between **X** and **y**, as it searches for the factor subspace of latent variables most congruent to both matrices, and its predictions are usually expected to be better than using MLR, when variables in the **X** matrix are highly correlated. A matrix of weights (reflecting the covariance structure between the **X** predictors and the **y** response variables) is calculated providing rich factor interpretation information. The selection of the optimal number of components (latent variables) in PLSR is done by using cross validation techniques (leave-one out sample at time) and optimal prediction of $\mathbf{y}_i$ values for external validation samples. Recently, the variable importance in projection or VIP scores have been proposed as a useful tool for the interpretation of PLSR models (built from several latent variables), (Chong and Jun, 2005). The interpretation of VIP scores, obtained for a particular regression model, is a useful tool for the evaluation of the importance achieved by each variable in the final PLS projection, and it may be also used for variable selection procedures. As a general rule, it is considered that a variable with a VIP score greater than 1 (one) is being considered to be highly significant for a given model.

2.3. Kernel partial least squares regression (K-PLS)

Walczak and Massart (1996a) proposed a nonlinear extension of the PLS regression method using RBF kernels. In this approach, a nonlinear transformation of the input variables **X** into a new feature

higher dimension space, **A**, by means of a kernel function like the radial basis mapping function, RBF, is assumed

$$\phi : \mathbf{X} \in \Re^N \rightarrow \phi(\mathbf{X}) \in \mathbf{A}.$$

In other words, the goal of the K-PLS models is to build a linear PLS regression model into a higher dimension feature space, **A**. The algorithm, named RBF-PLS, performs PLS regression on φ (**X**). In the calculation of the PLS components, the matrix product **XX**′ is replaced by the new matrix product:

$$\varphi(\mathbf{X})\varphi(\mathbf{X})',$$

using the so-called kernel trick, which allows for the calculation of dot products in the higher dimensional feature space using simple dot product functions defined on pairs, *i*,*j*, of input individual sample patterns (matrix **X** rows):

$$\varphi(\mathrm{x}_i)\varphi\left(\mathrm{x}_j\right)' = K\left(\mathrm{x}_i, \mathrm{x}_j\right).$$

This avoids calculating the coordinates in the feature space which could be a difficult task for a highly dimensional feature space. The benefit of this procedure is that using the kernel functions corresponding to the dot products in the feature space avoids non-linear optimization procedures and allows the use of much simpler and more reliable linear PLS regression algorithm.

2.4. Support vector machine regression (SVR)

Originally developed to solve classification problems, support vector machines can also be adapted to non-linear regression problems with the introduction of the so called ε-insensitive loss function. The theory of SVR has been discussed in details in Vapnik (1995, 1998). Briefly, in SVM regression analysis, the data matrix **X** is first mapped (in a similar way like in K-PLS) into a higher-dimensional feature space by the use of a kernel function and then a linear regression is performed in the feature space. In this case, the linear model, $f(\mathrm{x}, \omega)$, proposed in the feature space is as follows:

$$f(\mathrm{x}, \omega) = \sum_{i,j=1}^{m} \omega\phi(\mathrm{x}_i)\phi\left(\mathrm{x}_j\right) + b,$$

where $\phi(\mathrm{x})$, is a nonlinear transformation of the original variables **X**, and ω and b are the coefficients of the linear model that can be obtained by solving a quadratic programming optimization problem. The appropriate non-linear mapping $\varphi(\mathrm{x})$ is in general unknown in advance and it is therefore difficult to determine. An approximation is performed using different types of kernel functions. Therefore, the SVM regression equation described above, can be rewritten as

$$f(\mathrm{x}, \omega) = \sum_{j=1}^{m} \omega K\left(\mathrm{x}_i, \mathrm{x}_j\right) + b$$

where the kernel function $K(\mathrm{x}_i,\mathrm{x}_j)$ can be a linear, a polynomial, a sigmoid or a more sophisticated function as the one previously mentioned and often used in K-PLS, the radial basis function, RBF. The Lagrangian multipliers λ obtained in the solution of the quadratic programming problem, are then used for the calculation of the weights in the regression equation, ω and the function $f(\mathrm{x})$ can be rewritten as follows:

$$f(\mathrm{x}) = \sum_{i,j=1}^{m} (\lambda_i - \lambda_i^*) K\left(\mathrm{x}_i, \mathrm{x}_j\right) + b,$$

where $\lambda_i\lambda_i^* = 0$, $\lambda_i, \lambda_i^* \geq 0$ and the data points corresponding to non zero values for $(\lambda_i - \lambda_i^*)$ are called support vectors. After building a model on the calibration data set, SVR can be used to predict unknown **y** values from the new **X** values. The complete SVR equations are fully described in Smola and Schölkopf (2004) and Vapnik (1995, 1998) and they are not repeated here in this work for brevity. Another parameter, typically denoted as C in the SVM literature, sets the tradeoff between the model simplicity (and hence its generalization), and the training error, allowing for some data fit losses (Üstün et al., 2007). Although several kernel functions can be used, in this work, the radial basis function, RBF, has been selected as a preferred kernel function to be compared with the results obtained using the K-PLS method. All parameter optimization and selection were done by cross validation, using a comprehensive grid search procedure over all possible values of the two parameters, ε and C.

2.5. Details of data modeling using K-PLS and SVR procedures

Both, K-PLS and SVR, include first the mapping of the input data variables to the feature space by using a kernel function and a kernel matrix. This mapping leads to linearization of the problem: the non-linear regression problem is transformed into a high dimensional (feature) space in which the solution of the problem is directly considered to be a linear problem. The choice of the kernel function and especially its parameter settings is an important aspect to consider in order to solve the problem adequately. If the kernel parameters of the selected kernel function are not chosen properly, valuable information is lost, and the application of K-PLS or of SVR methods will not solve the regression problem adequately. Thus, the selection of the optimal kernel function and of its accompanying parameters to describe the nature of the original input data is a crucial step. Many techniques have been described in the literature about the kernel selection and about the kernel parameter adjustment, using for instance exhaustive grid search or genetic algorithms (GA) followed by an optimization procedure like simplex or others. In this work, we have used a systematic grid search procedure for the RBF function σ (function spread) parameter optimization and for the selection of C and ε-loss function values.

Calibration and predictions in K-PLS have been performed using the TOMCAT toolbox (Daszykowski et al., 2007) and MATLAB. The RBF Gaussian width was optimized using a grid search procedure starting from 0.1 to 1.0, with increments of 0.01. A leave-one out cross validation was considered in the modeling.

Calibration and predictions in SVM modeling were performed using STATISTICA (STATSOFT, Inc.) commercial software. The default value for the of γ optimization parameter was set to be equal to $1/k$, where k is the number of input variables (in this work k was equal to 41). The parameter C in the grid search procedure was from 0.1 to 10, with increments of 0.1 and the parameter ε was from 0.01 to 0.5 with increments of 0.01. A tenfold cross-validation was employed in this study.

The RBF kernel defines a square symmetrical matrix, called activity matrix **A**, which represents the similarity between pairs of objects. Each row (and column) in the activity matrix represents the similarity of a specific object with all other objects in the training set. Since the RBF function is used here, the element values of this activity matrix will vary between zero and one. A value close to zero indicates that two objects are very different, whereas a value close to one corresponds to two almost similar objects. As mentioned above the mapping of the original input data variables into the activity matrix, will cause the loss of information about them. This information can be however very relevant to interpret adequately what original variables are finally significant in the built regression model. It was shown (Üstün et al., 2007) that the correlation between each column (variable) of the original input data matrix ($\mathbf{X}(n \times m)$) with each row or column of the kernel matrix, $\mathbf{A}(n \times n)$, can be used for the recovery of the information about the relation between the kernel matrix and the input variable space. Since each column of **X** contains the information of the input variables and each row/column of **A** represents the similarity between the objects, the elements of the calculated

correlation matrix ($\mathbf{R}(m\times n)$) will reveal the individual contribution of each input WTP variable to the kernel matrix. A correlation value close to zero indicates that the respective input variable has no relevant contribution to the specific row of the kernel matrix, **A**, while a value close to +1 or −1 indicates that this is an important variable.

To calculate the contribution of each individual variable in the final K-PLS or SVM regression models is necessary to make explicit what is the relation between the obtained support vectors (λ-values in SVR) and *b*-regression vector values (in K-PLS) with the original variables in **X**. Since the kernel matrix, **A**, does not contain information about the original variables; it has no sense to relate the λ-values directly to the kernel matrix rows. The relationship between the λ-vector values and the original input data variables has to be investigated in another way. The quadratic programming optimization part of the SVR algorithm returns a vector of λ-values (which is comparable to the *b*-regression vectors in MLR, PLS and K-PLS) having a length equal to the number of objects whose elements satisfy the constraint $\lambda_i\lambda_i^*=0$, $\lambda_i,\lambda_i^*\geq 0$. As a consequence, each sample (object) in the original input space is weighted by its assigned λ-value. The samples having λ-values equal or very close to zero are not important because these samples do not contribute to the SVR model. By calculation of the inner product between the contributing input samples (original rows of data matrix **X**) and their corresponding λ-vector, a new vector **p** with the length of the number of input variables is obtained. Plotting this new vector will reveal the profile of the variables which contribute more significantly to the overall model.

For brevity, in this study the relative prediction errors of THM concentrations in percentage are included and compared, for both calibration and prediction steps and they are calculated as follows:

$$\text{Rel. error in } \% = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{\sum_{i=1}^{n}(y_i)^2}} \times 100.$$

3. Material and methods

Forty-one WTP operational variables were measured and investigated as possible factors affecting trihalomethane formation within the Sant Joan Despí WTP in Barcelona, Spain. Table 1 and Fig. 1 show measured plant variables and locations of sampling and chlorination in the plant operational scheme. Variables selected for this study were obtained from the laboratory information management system (LIMS). These variables give analytical information from several locations at the plant and they cover the entire water treatment process (input river water quality, first pre-chlorination, after sand filtration, after carbon filtration, after post-chlorination, and output supplied water quality). The investigated variables were analyzed in the plant over the whole year 2003 and were used to build-up the block of **X** variables (data matrix **X**).

For a total number of 162 days during 2003, concentrations of the different THMs were measured (THM variables, block of **y** variables) at the exit of the plant.

Concentrations of the four THMs were determined using standard chromatographic methods (Head Space GC-ECD, Ventura and Rivera, 1985). Samples were analyzed using a Fisons 8130 gas chromatograph with a DB-624 30 m × 0:32 mm × 1:8 μm film thickness fused silica column. Automatic injections of 0.5 ml were made in a split mode (1:10), with helium as a carrier gas and nitrogen as a make up gas. Detection and quantification limits were estimated to be around 0.10 and 0.25 μg/l. Trihalomethane standard solutions were prepared in the range of 0.5–8 μg/l. The total amount of THMs (variable TTHM) was calculated as the sum of the concentrations of the four individual compounds. Therefore, a total number of five $\mathbf{y}_i$ variables were defined. UV absorbance measurements in Abs/100 cm units were carried out at 254 nm using 1-cm quartz cells and then reporting values for a 100 cm path length (i.e. multiplying them by 100). Residual free chlorine was measured according to the DPD colorimetric method and chloride was measured volumetrically using the Mohr method (APHA et al., 1998). TOC concentration in mg/l of the samples was determined using a TOC analyzer. Turbidity was measured in FNU using nephelometry. Estimated measurement errors for quality control were below 35% for THM concentrations, below 25% for turbidity, below 15% for halides, below 15% for chlorine and below 15% for TOC. Only accredited and validated (ISO17025, ISO9000) methods were used and applied for routine treatment control. Influence of chlorine dioxide was studied and if the assay is run properly (i.e. reaction times) there was no interference at the levels of dosage.

Two data subsets were prepared, one to build the calibration model (144 days) and another to validate the model (18 days).

Initial data preparation and data prearrangement of different data blocks (**X** and $\mathbf{y}_i$ data sets) were performed using EXCEL (Microsoft, Redmon, WA, USA). All calculations were performed using PLS Toolbox 5.8 (Eigenvector Research, Manson, WA, USA), TOMCAT toolbox (Daszykowski et al., 2007) and MATLAB 7.0 software (The Mathworks, Natick MA, USA) and STATISTICA 8 for Windows (StatSoft, Inc., Tulsa, OK, USA).

4. Results and discussion

4.1. Linear and non-linear regression predictions

Table 2 summarizes the prediction errors obtained for the modeling of THM concentration changes as well as for the prediction of their total sum. MLR and PLSR predictions were now better than in our previous study (Platikanov et al., 2007). Prediction errors for chloroform and bromoform formation were now around 4% lower than in our previous work (Platikanov et al., 2007). When considering the performance of nonlinear regression techniques, the prediction results achieved by K-PLS and SVR were better than those obtained previously by MLR and PLSR, with 4–5% lower prediction errors for external validation samples. It was possible to conclude that nonlinear interactions among the variables could be better captured and modeled now by these two nonlinear techniques. Comparing the results obtained by them, K-PLS predictions were slightly better than those from SVR, although this could be due to the difficulty in the optimal tuning of SVM parameters during the modeling step. Moreover, K-PLS modeling provides additional features which facilitate better modeling performance and interpretability (see below).

The average calibration and external validation relative prediction errors for the total trihalomethane (TTHM) concentrations over the full investigated period (1 year) were equal to 14%, using either K-PLS or SVR methods. For the individual trihalomethane compounds, a significant improvement in the prediction of bromoform concentrations was obtained in external validation samples, with an error decrease from 29% (before) to 21% (now), using the proposed nonlinear techniques. Also, the prediction of chloroform did improve when K-PLS was used with an almost 10% overall prediction error improvement for external validation samples compared to previous results. Finally, the ranking of prediction performances for the different THMs, in both, calibration and external validation steps, was: $CHBr_3$ (the worst), $CHBrCl_2$, $CHBr_2Cl$, $CHCl_3$and TTHM (the best). Again, as an additional remark, prediction results using the whole 41 plant variables resulted to be better than those previously reported (Platikanov et al., 2007) using only 23 of the plant variables, especially for the modeling of bromoform and chloroform formation. Recently, the application of KPLS–SVM has been proposed and its possible application in the context of similar studies is a matter of discussion (Rosipal et al., 2003; Jin, 2009).

Table 1
Variables monitored in the water disinfection process at Sant Joan Despí WTP.

Variable	Process stage	Abbreviation	Description	Average	Minimum	Maximum	Standard deviations
1	1. River water quality	TEMPRiv	River water temperature in Celsius	14.38	5.45	28.22	5.84
2		ChlorideRiv	River chloride concentration in mg/l	253.73	140.00	449.67	54.59
3		absUVRiv	River water UV absorbance in Abs/100 cm	10.74	7.27	22.60	2.25
4		BRKPTRiv	Chlorination breakpoint	9.35	3.50	32.60	4.61
5		FlowRiv	Incoming river flow in m^3/s	11.72	2.00	70.00	9.65
6		CONDRiv	River water conductivity (20 °C) in μS/cm	1319.34	948.00	2064.50	157.41
7		TOCRiv	River total organic carbon concentration in mg/l	6.3	3.39	11.40	1.17
8		NH3Riv	River ammonia concentration in mg/l	0.97	0.08	4.54	0.74
9		OXIDAB	River water oxidability ($KMnO_4$)	4.16	1.62	10.70	0.85
10		TURBRiv	River water turbidity in UNF	87.64	4.58	1597.10	158.48
11	2. Pre-chlorination[a]	CLconsumL1-1	Chlorine consumption in kg/h Linia 1 at site 1	68.10	0.00	112.22	23.96
12		CLconsumL1-2	Chlorine consumption in kg/h Linia 1 at site 2	15.39	0.00	72.37	18.17
13		CLconsumTotL1	Total chlorine consumption in kg/h Linia 1	83.49	4.09	178.95	31.84
14		DoseTotPreClL1	Total prechlorination dose in mg/l Linia 1	11.50	0.71	55.89	5.83
15		CLconsumL2-1	Chlorine consumption in kg/h Linia 2 at site 1	51.96	0.00	100.16	23.07
16		CLconsumL2-2	Chlorine consumption in kg/h Linia 2 at site 2	7.59	0.00	47.81	10.29
17		CLconsumTotL2	Total chlorine consumption in kg/h Linia 2	59.55	4.55	108.10	22.56
18		DoseTotPreClL2	Total prechlorination dose in mg/l Linia 2	9.24	1.91	20.54	3.19
19	3. Sand filtration	absUVs1	UV absorbance after sand filters at site 1 in Abs/100 cm	8.05	5.26	12.36	1.42
20		absUVs2	UV absorbance after sand filters at site 2 in Abs/100 cm	8.04	4.43	12.05	1.36
21		Clors1	Residual Cl_2 concentration in mg/l after sand filters at site 1	0.194	0.017	7.925	0.635
22		Clors2	Residual Cl_2 concentration in mg/l after sand filters at site 2	0.160	0.004	1.581	1.358
23		TOCs1	Organic carbon total concentration in mg/l after sand filters at site 1	5.1	2.4	7	0.6
24		TOCs2	Organic carbon total concentration in mg/l after sand filters at site 2	5.2	2.6	7.1	0.7
25		TURBs1	Turbidity measured in FNU after sand filters at site 1	0.71	0.23	3.05	0.44
26		TURBs2	Turbidity measured in FNU after sand filters at site 2	0.63	0.17	5.75	0.60
35	4. Water pumping from well water aquifer	Wellswat	Input well water total amount per day in liters	784,984.5	0	3,240,700	883,577.9
27	5. Active carbon filtration	absUV	UV absorbance after carbon filters in Abs/100 cm				
28		Chlorc1	Residual Cl2 concentration in mg/l after carbon filters at site 1	0.97	0.80	1.71	0.11
29		Chlorc2	Residual Cl2 concentration in mg/l after carbon filters at site 2	0.96	0.82	1.23	0.08
30		Chloride	Chloride concentration in mg/l after carbon filters	289.7	160	532	67.3
31		TOC	Total organic carbon concentration in mg/l after carbon filters	3.47	2.07	4.83	0.58
32		TEMP	Water temperature in Celsius after carbon filters	14.5	7.3	25.2	4.9
33		TURBc1	Turbidity measured in FNU after carbon filters at site 1	0.18	0.12	0.34	0.04
34		TURBc2	Turbidity measured in FNU after carbon filters at site 2	0.18	0.12	0.33	0.04
40		Average	Average time of life of carbon filters from the last regeneration in days	214.14	155.25	295.95	32.99
41		SUM	Total time of life of carbon filters from the last regeneration in days	4282.8	3105	5919	659.9
36	6. Postchlorination	POSCl1	Emergency added Cl_2 concentrations in mg/l after carbon filters at site 1	0.30	0	31.86	2.64
37		POSCl2	Emergency added Cl_2 concentrations in mg/l after carbon filters at site 2	0.02	0	0.67	0.10
38		FlowPOSCl2	Emergency added Cl_2 volume in liters/day after carbon filters at site 1	40.4	0	2793.3	295.6
39		FlowPOSCl1	Emergency added Cl_2 volume in liters/day after carbon filters at site 2	2.3	0	192.3	17.7
42	Water at the plant exit	$CHCl_3$	Concentration in μg/l of chloroform measured at the exit of the water treatment plant	15.19	6.00	30.85	4.79
43		$CHCl_2Br$	Concentration in μg/l of bromodichloromethane measured at the exit of the water treatment plant	17.24	3.67	35.70	6.28
44		$CHBr_2Cl$	Concentration in μg/l of chlorodibromomethane measured at the exit of the water treatment plant	24.41	12.55	45.00	5.87
45		$CHBr_3$	Concentration in μg/l of bromoform measured at the exit of the water treatment plant	18.75	6.67	61.00	5.51
46		TTHM	Sum of the concentrations in μg/l of the 4 trihalomethane compounds measured at the exit of the water treatment plant	75.58	40.87	121.55	14.91

[a] Variable position in the **X** matrix.

4.2. WTP variables which were more significant on THM formation

4.2.1. MLR statistical significant variables and PLSR–VIP scores

The global contribution of all 41 individual variables to the linear regression models was estimated by the evaluation of the statistical significance of the MLR regression coefficients or by using VIP scores in PLSR analysis (Chong and Jun, 2005). Table 3 summarizes what WTP variables resulted to be more influential for the modeling and prediction of trihalomethane formation during the different plant treatments using both linear models.

Table 3 shows that some variables like TEMP, TURBc1, TURBc2, Average, Wellswat, Chlorc1 and Chlorc2 characterized the late stages of the disinfection processes at the plant and they are important for the prediction of all THMs. According to the MLR and PLSR results, some of the new included variables from the incoming river water quality and from the prechlorination step, resulted to be also significant, especially for the formation of multichlorinated THMs (like TEMPRiv, OXIDAB, TURBRiv, CONDRiv, CLconsumL1-1 and ClconsumL1-2, see Table 3).

The formation of multibrominated compounds appeared to take place predominantly at the very last stages of the disinfection process at the plant. Dibromochloromethane formation did not show any clear dependence with the variables measured in the incoming river water or with variables monitored at the prechlorination stage, whereas, the formation of bromoform did appear as a very complex process. Many WTP variables were found to be significant for the bromoform formation by the MLR and PLSR models. At the begging of the disinfection process, the most important variables for the bromoform formation appeared to be the incoming water conductivity. Probably, this fact indicates the dependence of bromoform formation from the quality of organic matter in the incoming river water. Also, it was found that chlorine consumption in the prechlorination

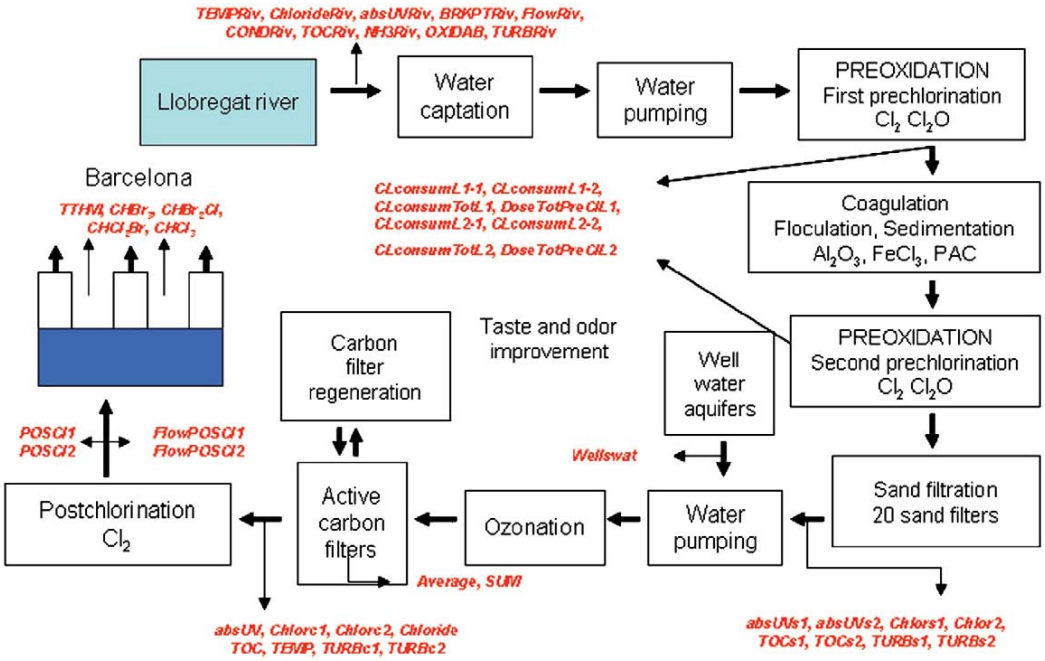


Fig. 1. Plant organization scheme and plant locations where the 41 WTP variables were measured.

step (CLconsumL1-1 and CLconsumL2-1 variables) were important for bromoform formation. Finally, many WTP variables like water temperature inside the plant, UV absorbance and turbidity at different sites of the WTP, TOC, post added chlorine concentrations, as well as well-supply flow levels, and carbon filter age, were found affecting significantly the formation of multibrominated THMs.

4.2.2. Visualization of kernel matrix in K-PLS modeling of TTHM

Since the nonlinear techniques appeared to predict more accurately than the MLR and PLSR linear methods, a visualization of the WTP variable contribution from nonlinear method application would be of interest for a better understanding of the different THM formation steps.

The correlation plot depicted in Fig. 2, distinguished those variables with higher correlation values (both positive and negative) from those variables with very low correlation values (close to zero) after kernel transformation in K-PLS modeling of TTHM. These plant variables with high positive or negative correlation values are relevant for the kernel function, thus are also important for THM formation, in contrast to variables with low correlation values close to zero.

In Fig. 2, variables number 1 (river temperature) and numbers 40 and 41 (carbon filter life) show very strong positive (dark red, close to +1) and negative (dark blue, close to −0.8) correlation values.

Table 2
Relative prediction errors in % using different chemometric methods for calibration and validation.

	MLR[a]	MLR	PLSR	K-PLS	SVR
Calibration					
Chloroform[b]	20.9	17.2	20.4	10.6	9.4
Dichlorobromomethane[b]	23.3	17.7	18.6	5.1	6.1
Dibromochloromethane[b]	13.9	12.6	13.2	5.0	6.7
Bromoform[b]	21.2	17.9	18.8	5.5	10.9
Total trihalomethanes[b]	11.7	10	10.9	7.3	2.5
External validation					
Chloroform	24.8	20.9	21.5	14.1	17.4
Dichlorobromomethane	25.1	22.4	21.6	16.1	16.5
Dibromochloromethane	21.1	21.2	20.7	15.6	15.9
Bromoform	28.6	24.6	25.8	20.8	20.8
Total trihalomethanes	16.2	17.5	17.5	13.6	13.9

[a] Predictions obtained by previous MLR analysis of 23 plant variables (Platikanov et al., 2007).

[b] Parameter optimizations: Chloroform: PLSR 4LVs; K-PLS 6LVs, σ0.5; SVR γ0.024 C10 ε0.1 SVs101. Dichlorobromomethane: PLSR 5LVs; K-PLS 6LVs, σ0.64; SVR γ0.024 C3.1 ε0.2 SVs79. Dibromochloromethane: PLSR 6LVs; K-PLS 7LVs, σ0.64; SVR γ0.024 C2.4 ε0.16 SVs70. Bromoform: PLSR 7LVs; K-PLS 8LVs, σ0.56; SVR γ0.024 C10 ε0.45 SVs50.

Table 3
MLR and PLSR results. WTP variables which were more important on the THM formation.

MLR results	
Compound	Variables[a] statistically significant at 5%
Chloroform	TEMPRiv; OXIDAB; TURBRiv; CLconsumL1-1; ClconsumL1-2;TEMP; TURBc1; TURBc2; Average;
Dichlorobromomethane	TEMPRiv; CLconsumL1-1; CLconsumL1-2;absUVs2; Chlors2; absUV; TEMP; FlowPOSCl1; Average;
Dibromochloromethane	TEMP; TURBc2; TEMPRiv;
Bromoform	TEMPRiv; CONDRiv; CLconsumL1-1; CLconsumL2-1; TURBs1; absUV; Chlorc1; Chlorc2; TOC; TEMP; TURBc1; TURBc2; Wellswat; POSCl1; FlowPOSCl1;
Total trihalomethanes	TOC; TEMP; POSCl1; Average; TEMPRiv
PLSR results	
Compound	Variables[a] with VIP scores more than threshold value of one
Chloroform	TEMPRiv, OXIDAB; CLconsumL1-1;CLconsumTotL1; ClconsumL2-1; CLconsumTotL2; DosetotPreClL2; Chlors2; TOC; TEMP; Average; SUM
Dichlorobromomethane	ChlorideRiv; CONDRiv; CLconsumL1-1; CLconsumTotL1; absUVs2; absUV; Chloride; TOC; TEMP; TURBc1; TURBc2; Wellswat; POSCl1; FlowPOSCl1; Average; SUM
Dibromochloromethane	TEMPRiv; absUVRiv; absUVs1; absUVs2;TOCs1; TOC; TEMP; TURBc1; TURBc2; Average; SUM
Bromoform	CONDRiv; CLconsumL2-1; absUVs2; TOCs1, TOCs2; TURBs1; absUV; Chlorc1; Chlorc2; TOC; TEMP; TURBc1; TURBc2; Wellswat; POSCl1; FlowPOSCl1;
Total trihalomethanes	TEMPRiv; absUVRiv; absUVs2; Chlors2; TOC; TEMP; TURBc1; TURBc2; POSCl1; Average; SUM

[a] For variable identification, see Table 1.

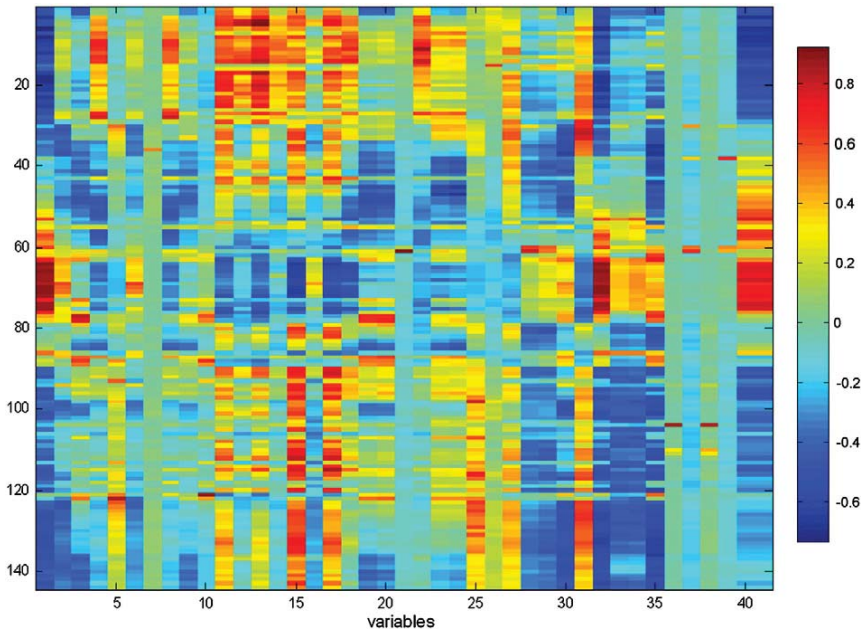


Fig. 2. Correlation plot between the original matrix **X** (WTP variables) and the kernel matrix achieved after kernel transformation in K-PLS modeling of TTHM. The y-axis is the samples. The x-axis depicts the original variables. For variable identification, see Table 1.

These three variables were considered to be very relevant. On the other hand variables with numbers 38 and 39 (emergency added Cl_2 volumes) had very close to 0 correlation values (light green color) and therefore they could not be considered relevant for the kernel function, therefore not important either for THM formation.

Some of the new variables with higher positive or negative correlation values like the river water temperature, river water chloride concentration and break point chlorination did appear to be very relevant for the better performance of K-PLS regression.

Many of the new variables monitored at the prechlorination stage like total chlorine amounts and total chlorine doses were found to be relevant for the kernel transformation. The correlation plot given in Fig. 2 shows a rather complex behavior for the monitored variables during the last disinfection stages since many variables on the plot did show higher correlation values. Variables like TOC, turbidity at specific locations, added amounts of well water and especially, the carbon filter age, resulted to be also very relevant in the kernel transformation.

4.2.3. Interpretation of SVR models

Fig. 3 shows the plots of the **p**-vectors calculated by SVR for the prediction of the trihalomethane concentrations. In these plots some WTP variables had higher contributions to the SVR model (high arbitrary values on the y axis) than the others (close to 0 values on the y-axis). The SVR model for chloroform formation (Fig. 3a) suggests the importance of new variables like river water TOC and breakpoint of chlorination, and also confirmed the importance of the other river water quality variables which were already reported to be significant in linear models. Also some other variables monitored during the prechlorination stage like total chlorine doses were found to play an important role in the modeling of chloroform concentration. SVR modeling also underlined those variables from the filtration and postchlorination stages that were also reported as significant in the linear models.

Very similar patterns can be noticed for the SVR results obtained for the concentrations of mixed chloro-bromomethanes and bromoform (Fig. 3b, c and d). Variables that were monitored throughout the entire disinfection process from the input river water quality to the last postchlorination stages could be correlated to the concentration of one of the THMs containing bromine. It must be highlighted that the presence of variable concentrations of chloride in river water had a strong effect on the SVR modeling of all brominated THMs. Although in this study the bromide concentration could not be directly incorporated as a known independent variable to the models (they were not experimentally available), its effects can be deduced indirectly due to the known fact that the bromide concentrations are very closely related to the chloride concentrations of input river water since the bromide/chloride concentration ratio (Davis et al., 1998) should be rather constant in the river waters. In the case of Barcelona WTP, this ratio is known to be approximately equal to 0.002.

Fig. 3e summarizes the average effect of the WTP variables in the SVR modeling of the TTHM concentrations. In addition to the results mentioned above in the section Visualization of kernel matrix in K-PLS, strong contributions to the SVR model for the prediction of TTHM concentrations were found for water oxidability and river flow volumes (a variable especially important for TTHM concentrations). Other WTP operational variables from the prechlorination, filtering and postchlorination steps, previously reported as important in the linear modeling and in the kernel visualization approach, were found to be also important in SVR modeling.

4.3. Interpretation of the achieved results

Analyzing all these results, we can conclude that the THM formation in water treatment plants is a really complex process and that the eighteen variables newly incorporated in this study, characterizing the river water quality and first prechlorination steps, did allow for a better description of the THM formation in Barcelona's plant. These results did suggest that the formation of multichlorinated and multibrominated compounds did occur at different plant locations during the whole disinfection process.

Multichlorinated THMs were already generated at the beginning of the disinfection process immediately after the implementation of the

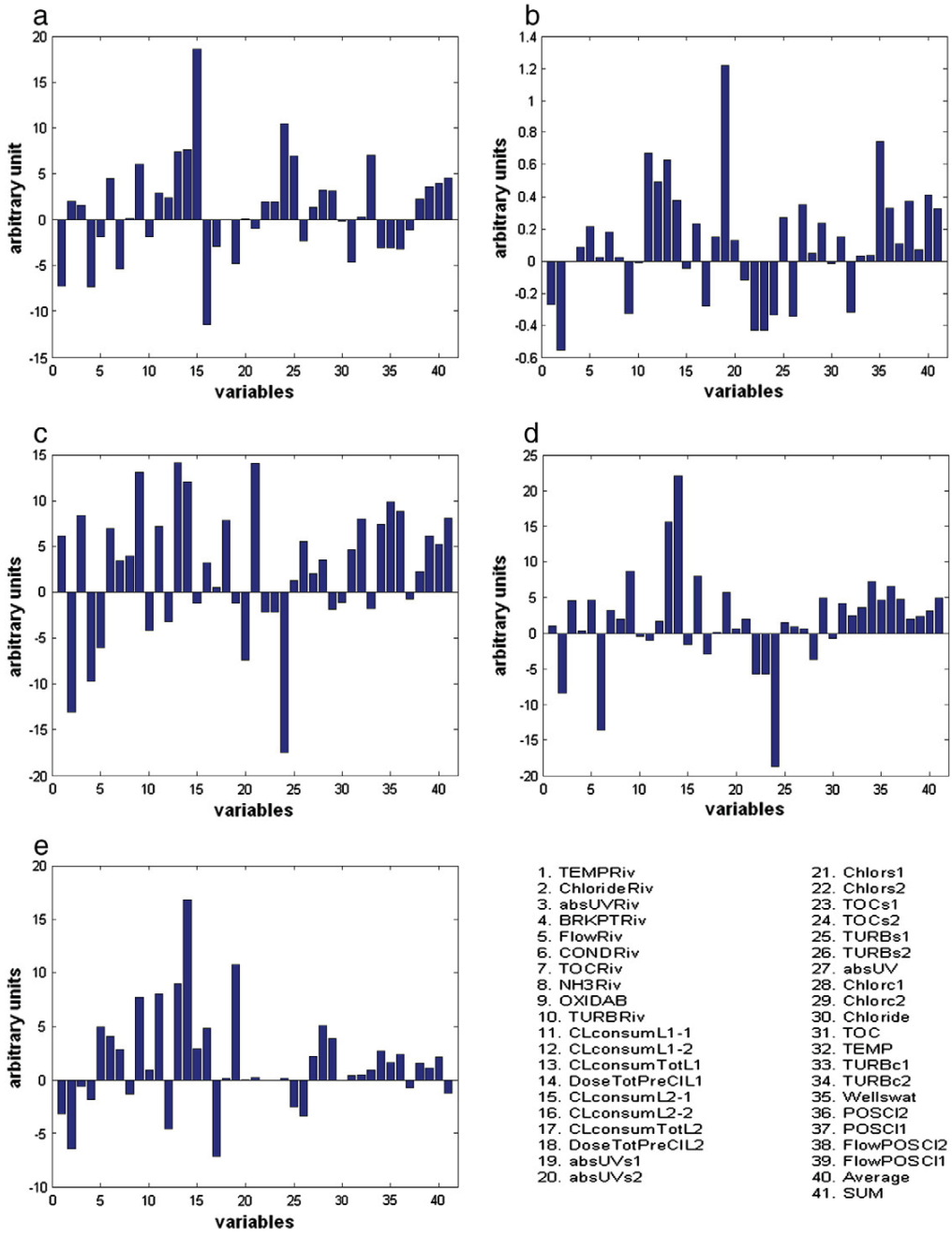


Fig. 3. Plot of inner-products (**p**-vectors) between the original calibration data set matrix **X** and the λ-vectors obtained by each SVR model: a) for chloroform; b) for bromoform; c) for dichlorobromomethane; d) for dibromochloromethane and e) for TTHM. The y-axis is given in arbitrary units. The x-axis depicts the original variables. For variable identification, see Table 1.

first two prechlorination steps (see Fig. 1). Variables characterizing the incoming river water quality were not possible to be changed during the management, and therefore, special attention should be paid to their natural variation, mostly due to seasonal changes. Since the Mediterranean climate area determines very different conditions for Barcelona's WTP operational procedures (from very intensive long drought periods to short huge flooding periods), significant fluctuations of river water flow and quality challenge the plant management, affecting the THM formation in different ways that have to be considered. A long term monitoring program, properly planned and established, on the river water quality variables in relation to THM formation would be very interesting for the optimal design of further consecutive disinfection procedures.

In order to avoid the formation of multichorinated THMs at the beginning of the plant, proper WTP management procedures should be considered at the prechlorination stage, especially in applying a proper chlorine dosage.

On the other hand, since multibrominated THM formation depended in a much higher extent from variables measured at the latest stages of the disinfection process like filtering, and postchlorination, many variables like total organic matter, adsorption effectiveness of sand or carbon filters, added post chlorination amounts, and added

amounts of well water became important now and they should be managed properly. For instance by mixing river and well water when high concentrations of TOC are occurring, to avoid the increasing formation of multibrominated THMs at the last stages.

The fact that the total amount of TOC in incoming river water did not appear as a significant parameter in linear regression models, can be explained considering the fact that this parameter was rather constant for all the analyzed data and also considering that more than the total amount of incoming organic matter, what has been shown to be important for THM formation is the chemical composition, of the different organic matter fractions (Croué, 2004). This fact was clearly corroborated in this work by the significance of variables like OXIDAB (permanganate oxidability) and CONDRiv (conductivity or river water) in the results of the linear models.

5. Conclusions

Formation of trihalomethanes is shown to depend strongly on several environmental and/or operational water treatment plant variables monitored during the disinfection processes. It has been concluded that trihalomethane formation takes place predominantly at three distinct locations of water treatment plants, i.e. after the first pre-chlorination stage, after the carbon filtration stage – probably due to desorption of already retained trihalomethanes – and after the post-chlorination stage. This formation also strongly depends on the quality of incoming raw water. One of the most important variables to be monitored in water treatment plants for trihalomethane formation predictions resulted to be input water temperatures at the beginning of the disinfection process and inside the plant. Some other variables associated to the organic matter quality, such as water oxidability and water conductivity resulted to be also important, as well as prechlorination, especially for the nonlinear models. Carbon filter aging, water turbidity after carbon filters, as well as the amount of input well water, UV absorbance, total organic carbon concentrations (the latter two after carbon filters) and post-chlorination variables were also confirmed to be important in both linear and nonlinear models. General improvement for external validation predictions for the concentration of chloroform and bromoform was achieved with the inclusion of these new variables and with the use of nonlinear modeling methods, with prediction errors decreasing up to 10%, in comparison to previous application of linear modeling methods. Total trihalomethane concentration was the best predicted variable, followed by concentrations of mixed bromo–chloro trihalomethanes and chloroform. Bromoform concentration was still the worst predicted parameter, probably (at least in part) due to a lower precision of its reference values. Results obtained by K-PLS and SVR methods confirmed the presence of nonlinear interactions among the operational variables, and the visualization of kernel transformations proved to be especially useful for this purpose.

Developed models for the Barcelona's WTP could be adapted to other WTPs. There will be specific characteristics of the incoming raw water and of the particular treatment procedures, like for instance addition of well water, which will change from plant to plant, and need model recalibration and updating. A proper WTP management plan should be the result of the implementation of on site-specific models covering all possible local conditions of work, but they will use similar chemometric methods (like KPLS and SVR) and similar calibration–validation strategies like those used in this work.

Acknowledgments

AGBAR (Barcelona's Water Company) is acknowledged for its financial help to carry out this research and especially to the people working at the 'Àrea de Control del Tractament' (Antoni Bernal) and at the 'Àrea de Química Orgànica' (Francesc Ventura), who provided the experimental data sets used for this study.

References

APHA, AWWA, WEF. Standard methods for the examination of water and wastewater. 20th ed. Washington, DC: American Public Health Association; 1998.
Calderon R. The epidemiology of chemical contaminants of drinking water. Food Chem Toxicol 2000;38:S13–20.
Chong I, Jun C. Performance of some variable selection methods when multicollinearity is present. Chemometr Intell Lab Syst 2005;78:103–12.
Chowdhury S, Champagne P, McLellan PJ. Models for predicting disinfection byproduct (DBP) formation in drinking waters: a chronological review. Sci Total Environ 2009;407:4189–206.
Croué J. Isolation of humic and non-humic NOM fractions: structural characterization. Environ Monit Assess 2004;92(1–3):193–207.
Daszykowski M, Serneels S, Kaczmarek K, van Espen P, Croux C, Walczak B. TOMCAT: a MATLAB toolbox for multivariate calibration techniques. Chemometr Intell Lab Syst 2007;85:269–77.
Davis S, Whittemore D, Fabryka-Martin J. Uses of chloride/bromide ratios in studies of potable water. Ground Water 1998;36:338–50.
EC, Council Directive 98/83/EC of 3 November 1998 on the quality of water intended for human consumption. Off J Eur Comm L 330/32, 5.12.1998.
Gallard H, von Gunten U. Chlorination of natural organic matter: kinetics of chlorination and of THM formation. Water Res 2002;36:65–74.
Geladi P, Kowalski BR. Partial least-squares regression: a tutorial. Anal Chim Acta 1986;85:1-17.
Golfinopoulos S, Arhonditsis G. Multiple regression models: a methodology for evaluating trihalomethane concentrations in drinking water from raw water characteristics. Chemosphere 2002;47:1007–18.
Jin P. New method for quantitative analysis of multi-component alkane gaseous mixture by FTIR spectroscopy. J Instrum Anal 2009;10:18. (http://en.cnki.com.cn/Article_en/CJFDTOTAL-TEST200910018.htm).
Krooshof PW, Üstün B, Postma G, Buydens L. Visualization and recovery of the (bio) chemical interesting variables in data analysis with support vector machine classification. Anal Chem 2010;82:7000–7.
Massart DL, Vandeginste BGM, Buydens LMC, de Jong S, Lewi PJ, Smeyers-Verbeke J. Handbook of chemometrics and qualimetrics. Amsterdam, Holland: Elsevier Science; 1998.
Milot J, Rodríguez MJ, Sérodes JB. Contribution of neural networks for modelling trihalomethanes occurrence in drinking water. J Water Resour Plann Manag-ASCE 2002;128:370–6.
Navalon S, Alvaro M, Garcia H. Carbohydrates as trihalomethanes precursors. Influence of pH and the presence of Cl− and Br− on trihalomethane formation potential. Water Res 2008;42:3990–4000.
Nikolaou AD. Investigation of the formation of chlorination by-products in water rich in bromide and organic matter content. J Environ Sci Health A 2004;39:2835–53.
Nissinen TK, Miettinen IT, Martikainen PJ, Vartiainen T. Disinfection by-products in Finnish drinking waters. Chemosphere 2002;48:9-20.
Platikanov S, Puig X, Martin J, Tauler R. Chemometric modeling and prediction of trihalomethane formation in Barcelona's water works plant. Water Res 2007;41:3394–406.
Postma GJ, Krooshof PW, Buydens L. Opening the kernel of kernel partial least squares and support vector machines. Anal Chim Acta 2011;705:123–34.
Reif JS, Hatch MC, Bracken M, Holmes L, Schwetz BA, Singer PC. Reproductive and developmental effects of disinfection by-products in drinking water. Environ Health Perspect 1996;104:1056–61.
Richardson S. Disinfection by-products and other emerging contaminants in drinking water. Trends Anal Chem 2003;22:666–84.
Rodríguez MJ, Sérodes JB. Spatial and temporal evolution of trihalomethanes in three water distribution systems. Water Res 2001;35:1572–86.
Rodríguez MJ, Sérodes JB. Application of back-propagation neural network modeling for free residual chlorine, total trihalomethanes and trihalomethanes speciation. J Environ Eng Sci 2004;3:S25–34.
Rodríguez MJ, Vunette Y, Sérodes JB, Bouchard C. Trihalomethanes in drinking water of greater Québec region (Canada): occurrence, variations and modeling. Environ Monit Assess 2003;89:69–93.
Rook JJ. Formation of haloforms during chlorination of natural waters. Water Treat Exam 1974;23:234–43.
Rosipal R, Trejo L. Kernel partial least squares regression in reproducing kernel Hilbert space. J Mach Learn Res 2001;2:97-123.
Rosipal R, Trejo L, Matthews B. Kernel PLS–SVC for linear and nonlinear classification. Proceedings of the Twentieth International Conference on Machine Learning (ICML-2003), Washington DC; 2003.
Sadiq R, Rodriguez MJ. Disinfection by-products (DBPs) in drinking water and the predictive models for their occurrence: a review. Sci Total Environ 2004;321:21–46.
Scholkopf B, Smola AJ. Learning with kernels. Cambridge, USA: MIT Press; 2002.
Serrano A, Gallego M. Rapid determination of total trihalomethanes index in drinking water. J Chromatogr A 2007;1154:26–33.
Smola A, Schölkopf B. A tutorial on support vector regression. Stat Comput 2004;14:199–222.
Toroz I, Uyak V. Seasonal variations of trihalomethanes (THMs) in water distribution networks of Istanbul City. Desalination 2005;176:127–41.

Üstün B, Melssen W, Buydens L. Visualisation and interpretation of support vector regression models. Anal Chim Acta 2007;595:299–309.
Vapnik V. The nature of statistical learning theory. New York, USA: Springer-Verlag; 1995.
Vapnik V. Statistical learning theory. New York, USA: John Willey & Sons; 1998.
Ventura F, Rivera J. Factors influencing the high content of brominated trihalomethanes in Barcelona's water supply (Spain). Bull Environ Contam Toxicol 1985;35: 73–81.
von Gunten U, Driedger A, Gallard H, Salhi E. By-products formation during drinking water disinfection: a tool to assess disinfection efficiency. Water Res 2001;35: 2095–9.
Walczak B, Massart D. The radial basis functions — partial least squares approach as a flexible non-linear regression technique. Anal Chim Acta 1996a;331:177–85.
Walczak B, Massart DL. Application of radial basis functions — partial least squares to non-linear pattern recognition problems: diagnosis of process faults. Anal Chim Acta 1996b;331:187–93.

Results and Discussion

- Prediction improvement after the insertion of the new 18 DWTP operational parameters in linear regression models

MLR and PLS predictions with the augmented model in Article 2 were better than such obtained with 23 plant variables in Article 1. The inclusion of new variables measuring the raw river quality and representing the pre-chlorination process had a positive effect on the THMs prediction. For example, the prediction results for chloroform improved around 4-5% in the external validation. This result was consistent with previous literature because the kinetic reaction of $CHCl_3$ formation is fast and occurs immediately when DOM and chlorine disinfectant enter in contact (Johnson and Jensen, 1885) during the prechlorination stage. Moreover, in the case of the SJD DWTP, the results suggested that $CHCl_3$ and $CHCl_2Br$ are predominantly formed at the prechlorination stage of the process. The model predictions regarding the formation of $CHBr_3$ and $CHClBr_2$ have not significantly improved with the extended set of 41 variables included in the model.

This result is likely due to the fact that multibrominated THMs form at lower rate and hence such compounds occur at later stages of the disinfection process (Al-Omari et al., 2014). Among the main conclusions for the SJD DWTP operational management was that there is a fast formation of $CHCl_3$ and $CHCl_2Br$ at the prechlorination step, and that formation of brominated THMs was common at other stages of potabilization procedures.

- Comparison of the linear and non-linear prediction models

Article 2 examined the predictive power of two well-known nonlinear techniques (SVR and K-PLS) for the THMs modelling in the SJD DWTP. These two methods proved to be a useful approach in modelling the THMs formation (for a particular compound or for their total sum), because such methods resulted in lower prediction errors in comparison to the the lineal methods in the external validation. The prediction errors obtained with K-PLS or SVR were lower by approximately 4–5% compared to the errors obtained with the linear MLR and PLS methods. The lower prediction errors suggested that a nonlinear interaction between DWTP variables and THMs existed and that such relationships were better modelled by applying the two nonlinear techniques.

The comparison of the prediction results between the two nonlinear techniques suggested that K-PLS model was slightly better than the SVR model. This result could be explained by the difficult optimization of the three SVR tuning parameters and the K-PLS reliance on adjustments of only two parameters.

The largest improvement of the relative predictions in the external validation was found for the total sum of trihalomethanes concentrations (TTHM), which was found to have a relative prediction error equal to 14%, using either K-PLS or SVR methods. The prediction of bromoform concentrations were also significantly improved, because its relative prediction error decreased from 29% (i.e., with 23 variables used as predictors) to 21% (i.e., with 41 variables used). The prediction of chloroform also improved when using K-PLS, where prediction errors improved by approximately 10%for external validation samples compared to the errors obtained with linear regression methods. The final ranking of prediction errors improvement was not significantly different in comparison to the previous work and it was as follows: $CHBr_3$ (worse) > $CHCl_3$ > $CHBrCl_2$ > $CHBr_2Cl$ > TTHM (best).

Taken together, the four regression techniques used in the two articles, namely MLR, PLS, K-PLS and SVR, proved to be useful in accurately modelling of THMs formation and prediction in the SJD DWTP. When various parameters, including incoming raw water quality and measures obtained at the last postchlorination step, were modelled, external validation predictions of total sum of THMs reached values around 10% of relative error. To the extent that this parameter (i.e., total sum) is the only parameter regulated by the EU legislation, it is critical to have a good prediction model. Concentrations of mixed bromo-chloro trihalomethanes and chloroform were relatively well predicted by most of the models employed in the studies. Bromoform concentration was the most difficult parameter to be predicted, perhaps partially due to lowest precision of its reference values. The application of nonlinear regression techniques such as K-PLS and SVR methods resulted useful because nonlinear interactions among operational variables and THMs could be expected.

- Evaluation of more important variables in linear regression models

Similar to Article 1, the contribution of 41 individual DWTP parameters to the linear regression models was evaluated. For this purpose, MLR regression coefficients and the

so-called VIP scores in PLS modelling were assessed for their importance and statistical significance.

The most important DWTP parameters for the modelling and prediction of trihalomethanes formation during the different plant treatments using both linear models are presented in Table 3 (Article 2) As displayed, variables such as water temperature, turbidity after carbon filters, amount of added wells water and chlorine concentrations after the carbon filters were found to be important for predicting THM formation. Interestingly, some of the new included DWTP parameters, specifically from the incoming river water quality and from the prechlorination step, were found to play a significant role in the formation of multichlorinated THMs, namely river water temperature, water oxidability, river water turbidity, river water conductivity, and the chlorine consumption at the first prechlorination step.

According to the results of the linear models, the formation of multibrominated compounds occurred at the last stages of the disinfection process in the plant. Many DWTP variables were found to be significant for the bromoform formation in MLR and PLS models. This observation is suggestive that bromoform formation presented a complex behaviour with nonlinear relationships with operational parameters. The most important variable for bromoform formation appeared to be the incoming water conductivity parameter. This fact meant that bromoform formation strongly depended on the quality of organic and inorganic matter in the incoming river water. Additionally, it was found that chlorine consumption in the prechlorination step was important for bromoform formation.

Other DWTP influential variables observed for multibrominated THMs formation were confirmed to be relevant also in this second work. They were water temperature inside the plant, UV absorbance and turbidity at different sites of the DWTP, TOC, post added chlorine concentrations, as well as the pumped underground water quantities, and carbon filters age.

- *Evaluation of more important variables in non-linear regression models*

Additional information about nonlinear relationships among operational DWTP parameters and THMs was obtained after the visualization of kernel matrix parameters in K-PLS modelling of TTHM (see more in the Methodology section). The correlation plot in Figure 2 (Article 2) displays the most important DWTP parameters. The parameters which had high positive or high negative correlation values were considered

to be relevant for the kernel function and therefore important for TTHM formation, in contrast to parameters with low correlation values close to zero.

In Figure 2 of Article 2, three parameters appeared to be very relevant for TTHM formation, river temperature and carbon filters age parameters. In contrast, emergency added Cl_2 volumes were not relevant and therefore they were not considered important for the kernel function and therefore not important either for THMs formation.

New variables with higher positive or negative correlation values were found to be river water temperature, river water chloride concentration and break point chlorination. Parameters from the prechlorination stage such as total chlorine amounts and total chlorine doses were also detected to be relevant for the kernel transformation, although they were underestimated in the linear modelling. The correlation plot displayed in Figure 2 (Article 2) suggested that there is a complex nonlinear behaviour regarding the trihalomethane formation in the last treatment stages, because various variables on the plot were found to be relevant in the kernel transformation. Parameters such as TOC, turbidity at specific locations, added amounts of wells water and especially, carbon filters age resulted to be highly relevant and important for THMs formation in both type of models, linear and nonlinear.

The interpretation of SVR models presented in Figure 3 (Article 2) was performed using the vizualisation of **p**-vectors (see more in the Methodology section of Article 2). The significance of DWTP variables was inferred from the higher contributions of such variables to the SVR model. Based on this analysis, in respect to the chloroform formation (see Figure 3a, Article 2), parameters such as river water TOC and the breakpoint of chlorination were found to be important, along with total chlorine doses and other parameters already detected to be significant in the linear models. Similar patterns were observed in the SVR results obtained for the concentrations of mixed chloro-bromomethanes and $CHBr_3$ (please refer to Figure 3b, c and d of Article 2). The most relevant information was gathered from the chloride concentration in river water, which was found to contribute significantly in the SVR modelling of all brominated THMs. Although the bromide concentration parameter was not measured and reported in the data, its indirect effect could be inferred from chloride concentrations of input river water to the extent that bromide/chloride concentration ratio is rather constant in the river water system under study (Davis et al., 1998). The results also suggested that water oxidability and river flow volumes were important for THMs formation.

Table 5. Summary of the most important operational parameters for the THMs formation

MLR results	
Compound	**Variables[1] statistically significant at 5 %**
Chloroform	TEMPRiv; OXIDAB; TURBRiv; CLconsumL1-1; ClconsumL1-2;TEMP; TURBc1; TURBc2; Average;
Dichlorobromomethane	TEMPRiv; CLconsumL1-1; CLconsumL1-2;absUVs2; Chlors2; absUV; TEMP; FlowPOSCl1; Average;
Dibromochloromethane	TEMP; TURBc2; TEMPRiv;
Bromoform	TEMPRiv; CONDRiv; CLconsumL1-1; CLconsumL2-1;TURBs1; absUV; Chlorc1; Chlorc2; TOC; TEMP; TURBc1; TURBc2; Wellswat; POSCl1; FlowPOSCl1;
Total Trihalomethanes	TOC;TEMP; POSCl1; Average; TEMPRiv
PLS results	
Compound	**Variables[1] with VIP scores more than threshold value of one**
Chloroform	TEMPRiv, OXIDAB; CLconsumL1-1;CLconsumTotL1; ClconsumL2-1; CLconsumTotL2; DosetotPreClL2; Chlors2; TOC; TEMP; Average; SUM
Dichlorobromomethane	ChlorideRiv; CONDRiv; CLconsumL1-1; CLconsumTotL1; absUVs2; absUV; Chloride; TOC; TEMP; TURBc1; TURBc2; Wellswat; POSCl1; FlowPOSCl1; Average; SUM
Dibromochloromethane	TEMPRiv; absUVRiv; absUVs1; absUVs2;TOCs1; TOC; TEMP; TURBc1; TURBc2; Average; SUM
Bromoform	CONDRiv; CLconsumL2-1; absUVs2; TOCs1, TOCs2; TURBs1; absUV; Chlorc1; Chlorc2; TOC; TEMP; TURBc1; TURBc2; Wellswat; POSCl1; FlowPOSCl1;
Total Trihalomethanes	TEMPRiv ; absUVRiv ; absUVs2 ; Chlors2; TOC; TEMP; TURBc1; TURBc2; POSCl1; Average; SUM
	K-PLS results
Compound	**Variables relevant in the kernel matrix visualization**
Chloroform	*Same parameters as for TTHMs*
Dichlorobromomethane	*Same parameters as for TTHMs*
Dibromochloromethane	*Same parameters as for TTHMs*
Bromoform	*Same parameters as for TTHMs*
Total Trihalomethanes	TEMPRiv, TOC, TURBs1, Average, SUM, ChlorideRiv, BRKPTRiv, CLconsumTotL2, CLconsumL2-1, Wellswat, TEMP

	SVR results
Compound	**Variables[1] with higher contributions to the SVR model**
Chloroform	TEMPRiv, BRKPTRiv,TOCRiv,CLconsumTotL1,DoseToTPreClL1CLconsumL2-1,TOCs2, TURBs2,TURBc2
Dichlorobromomethane	ChlorideRiv, BRKPTRiv,Oxidab,CLconsumTorL1,DoseTotPreClL1,Chlors1,TOCs2, TEMP,TURBc2, Wellswat, Average , SUM
Dibromochloromethane	Chloride, CONDRiv,CLconsumTotL1,DoseTotPreClL1,TOCs2
Bromoform	ChlorideRiv, CLconsumL1-1, CLconsumL1-2, absUVs1,Chlors2, TOCs1,Wellswat, Average, SUM
Total Trihalomethanes	ChlorideRiv, OXIDAB, CLconsumTotL1,DoseTotPreClL1,absUVs2,

For variable identification, see Table 1 , Article 2.

The results suggested that the THMs formation in the Sant Joan Despí DWTP (Barcelona, Spain) presented a seasonal trend with the highest amounts in spring season. Due to natural characteristics of the Llobregat River water, the formation of brominated THMs compounds is favored. The formation of trihalomethanes is strongly dependent on several natural and operational DWTP parameters, which are monitored at the plant and can be modelled. These parameters capture the entire water treatment process from the incoming raw water quality to the last post-chlorination process, before drinking water to be realised into the WDS. It was deduced that the most important parameters found in these two studies are these related to the kinetic reaction of THMs formation. Parameters, such as the natural organic matter quality as a precursor, the chlorine doses to disinfect, the temperature and the bromide concentration, were shown to be very influential for the THMs formation. In addition, DWTP operational procedures, such as pumping underground water and carbon filters frequent regeneration, were suggested to reduce significantly THMs formation and occurrence.

THMs formation reaction begins at locations where chlorine is added, i.e. after first pre-chlorination (especially, for $CHCl_3$ and $CHCl_2Br$) and after the post-chlorination and occurs permanently (especially, for $CHClBr_2$ and $CHBr_3$). The granular activated carbon filtration can be considered critical for the reduction of THMs. Two processes favouring THMs formation may take place along with the aging of carbon filters. The first process is the desorption of THMs in the carbon filters. The

second process is the saturation of carbon filters with organic matter which explains the filters' decreased efficiency in retaining THMs precursors.

3.1.3 Article 3 – Platikanov, S., Tauler, R., Rodriguez, P., Antunes, M., Pereira, D. and J. Esteves da Silva. *Factorial Analysis of the trihalomethane formation in the reaction of colloidal, hydrophobic and transphilic fractions of DOM with free chlorine.* Environmental Science and Pollution Research 17 (2010) 1389-1400.

Introduction

In the previous two articles, several important DWTP parameters for the THMs formation have been detected. In both studies, linear and nonlinear regression techniques have been applied to model the relationships between THMs and operational parameters. The formation of THMs was demonstrated to be highly dependent on several natural and operational DWTP parameters. Among the parameters, total organic carbon concentrations (TOC) and the water absorbance at UV 254nm, both describing natural organic matter in water, were distinguished to be important precursors for the THMs formation.

Scientific research has concluded that NOM (its quantity and quality) should be the most important precursor of THMs formation. An extensive scientific literature has been focused on the characteristics of NOM and its various forms (such as the organic matter dissolved in water, DOM), because different fractions of DOM were suggested to contribute to the formation and speciation of THMs. Some studies have attempted to relate specific characteristics of organic matter, such as functionality and aromaticity, to THM formation (Gallard and von Gunten 2002). Other studies have largely focused on the influence of operational parameters, including chlorine dose, water temperature, pH, and reaction time, on the THM formation (Radiq and Rodriguez 2004). The contribution of our article is that it combines the two lines of research to gain better understanding about the influence of NOM precursor and the operational conditions of disinfection, jointly determining the THMs formation. More specifically, Article 3 focused on the importance of different DOM fractions for the THMs formation and investigated the main determinants of THMs formation. In this article, results of laboratory experiments, where different DOM fractions (colloidal, hydrophobic, and transphilic fractions) were disinfected by chlorine under target conditions, are presented.

This study was conducted during my research stay at the University of Oporto (Portugal) by invitation of Professor Joaquim Esteves da Silva (Faculty of Science). DOM fractions were obtained and fractionated for further analysis from filtered lake water from Caldeirão dam in Guarda (Portugal). About 200 L of water were initially concentrated using a portable reverse osmosis system that consisted of an electric pump, ionic exchange resins, and a reverse osmosis membrane. The fractionation system is presented in Figure 1 of Article 3. The DOM concentrated water sample was collected after the osmosis process and was subsequently acidified to pH 2. Following the fractionation method of Leenheer (2004), DOM was separate into four fractions: colloidal, hydrophobic (HPOF), transphilic, and hydrophilic. For the purpose of this study, three fractions were chosen, namely colloidal, HPOF and transphilic. Solution samples with DOM fractions were then frozen, lyophilized and characterized with Fourier transform infrared spectroscopy (FT-IR) and elemental analysis. Information about the chemical structure of each DOM fraction was found to be relevant in explaining THMs formation.

The investigation included a screening Placket-Burman factorial analysis design, which considered five factors: DOM fraction concentration, chlorine dose, temperature, pH, and bromide concentration. A Box-Behnken design was applied to assess the effect of the most important factors (DOM fraction concentration, chlorine dose, and temperature) using a response surface strategy.

Results from the linear models and response surface plots for THMs formation revealed that the formation of THMs was complex and largely dependent on DOM fraction concentration, chlorine dose, and temperature. Formation reactions were found strongly determined by individual factors and their corresponding interactions.

Environ Sci Pollut Res (2010) 17:1389–1400
DOI 10.1007/s11356-010-0320-4

RESEARCH ARTICLE

Factorial analysis of the trihalomethane formation in the reaction of colloidal, hydrophobic, and transphilic fractions of DOM with free chlorine

Stefan Platikanov · Roma Tauler · Pedro M. S. M. Rodrigues · Maria Cristina G. Antunes · Dilson Pereira · Joaquim C. G. Esteves da Silva

Received: 30 April 2009 / Accepted: 24 February 2010 / Published online: 25 April 2010

Abstract
Background, aim, and scope This study focuses on the factors that affect trihalomethane (THMs) formation when dissolved organic matter (DOM) fractions (colloidal, hydrophobic, and transphilic fractions) in aqueous solutions were disinfected with chlorine.
Materials and methods DOM fractions were isolated and fractionated from filtered lake water and were characterized by elemental analysis. The investigation involved a screening Placket-Burman factorial analysis design of five factors (DOM concentration, chlorine dose, temperature, pH, and bromide concentration) and a Box-Behnken design for a detailed assessment of the three most important factor effects (DOM concentration, chlorine dose, and temperature).
Results The results showed that colloidal fraction has a relatively low contribution to THM formation; transphilic fraction was responsible for about 50% of the chloroform generation, and the hydrophobic fraction was the most important to the brominated THM formation.
Discussion When colloidal and hydrophobic fraction solutions were disinfected, the most significant factors were the following: higher DOM fraction concentration led to higher THM concentration, an increase of pH corresponded to higher concentration levels of chloroform and reduced bromoform, higher levels of chlorine dose and temperature produced a rise in the total THM formation, especially of the chlorinated THMs; higher bromide concentration generates higher concentrations of brominated THMs. Moreover, linear models were implemented and response surface plots were obtained for the four THM concentrations and their total sum in the disinfection solution as a function of the DOM concentration, chlorine dose, and temperature. Overall, results indicated that THM formation models were very complex due to individual factor effects and significant interactions among the factors.
Conclusions In order to reduce the concentration of THMs in drinking water, DOM concentrations must be reduced in the water prior to the disinfection. Fractionation of DOM, together with an elemental analysis of the fractions, is important issue in the revealing of the quality and quantity characteristics of DOM. Systematic study composed from DOM fraction investigation and factorial analysis of the responsible parameters in the THM formation reaction can, after an evaluation of the adjustment of the models with the reality, serves well for the evaluation of the spatial and temporal variability in the THM formation in dependence of DOM. However, taking into consideration the natural complexity of DOM, different operations and a strict control of them (like coagulation/flocculation and filtration)

Responsible editor: Philippe Garrigues

S. Platikanov · R. Tauler
Department of Environmental Chemistry, IIQAB-CSIC,
Jordi Girona 18-26,
08026 Barcelona, Spain

P. M. S. M. Rodrigues
Research Unit for Inland Development, UDI,
Instituto Politécnico da Guarda,
Av. Dr. Francisco Sá Carneiro 50,
6301-559 Guarda, Portugal

M. C. G. Antunes
Chemistry Department,
Universidade de Trás-os-Montes e Alto Douro,
5000-911 Vila Real, Portugal

D. Pereira · J. C. G. Esteves da Silva (✉)
Departamento de Química, CIQ(UP),
Faculdade de Ciências da Universidade do Porto,
R. Campo Alegre 687,
4169-007 Porto, Portugal
e-mail: jcsilva@fc.up.pt

has to be used to quantitatively remove DOM from the raw water.

Recommendations and perspectives Assuming that this study represents a local case study, similar experiments can be easily applied and will supply with relevant information every local water treatment plant meeting problems with THM formation. The coagulation/flocculation and the filtration stages are the main mechanisms to remove DOM, particularly the colloidal DOM fraction. With the objective to minimize THMs generation, different unit operation designed to quantitatively remove DOM from water must be optimized.

Keywords Factorial analysis · Response surface methodology · Chlorine water disinfection · Colloidal · Hydrophobic · Transphilic · Trihalomethanes formation · Disinfection by products

1 Background, aim, and scope

Since the first study conducted by Rook in 1974, it has been established that the use of chlorine for disinfecting drinking water leads to the formation of various disinfection byproducts (DBPs) potentially harmful for human health (Bellar et al. 1974; Nieuwenhuijsen et al. 2000; Chang and Young 2000; Richardson and Thruston 2003; Richardson et al. 2007). Regardless, chlorine remains the most commonly used disinfectant because it is effective, relatively inexpensive and has a disinfection residual property, which is important to prevent possible sources of contamination in the distribution system. Among the DBP groups identified in chlorinated water (Hrudey 2009), only the trihalomethane (THM) family is regulated by European Community legislation, which includes chloroform, bromodichloromethane, chlorodibromomethane, and bromoform (Council Directive 98/83/EC). THMs constitute an important matter of public health concern, since they are regarded as carcinogens and, more recently, epidemiological studies indicate that they are also associated with reproductive and developmental problems (McGeehin et al. 1993; Simpson and Hayes 1998; Lewis et al. 2006; Savitz et al. 2006). However, many epidemiologists and other scientists contest many of epidemiological studies involving DBPs in drinking water, particularly those involving acute exposure, which suffer from the misclassification of exposure (Reif et al. 1996). Actually, a review of epidemiological studies about cancer risks found only a somewhat consistent association among chlorinated surface waters and bladder cancer. Also, weak to moderate bladder cancer risks were found associated with long-term exposure to chlorinated surface water and THM (Villanueva et al. 2007; Hamidin et al. 2008). The health risk concern from exposure to THM forced the European Union to establish a new drinking water quality regulation that changed the maximum levels of total THM (TTHM, the sum of all individual trihalomethanes) allowed in drinking water from 150 μg L^{-1} to 100 μg L^{-1} (Council Directive 1998). However, to strictly follow this directive and apply the practices in municipal treatment plants that supply safe and potable water, understanding the process of THM formation is crucial.

Natural organic matter dissolved in water (DOM) is usually considered the precursor of DBP (von Gunten et al. 2001; Rostad et al. 2000; Leehneer et al. 2001; Panyapinyopol et al. 2005). DOM is a complex mixture of various compounds with very different chemical properties. Many efforts have been made to characterize DOM in order to improve its removal and reduce DBP formation during water disinfection (Croué et al. 1999; Croué 2004). The most common practice for the isolation and fractionation of DOM from water is using XAD resins and ion-exchange resins (Leenheer et al. 2000; Leenheer 2004). Recently, Leenheer (2004) proposed an operational scheme to separate DOM into four fractions: colloidal, hydrophobic, transphilic, and hydrophilic. A number of studies have attempted to correlate some specific characteristics of organic matter, functionality, and aromaticity with THM formation (Norwood et al. 1980; Gallard and von Gunten 2002; Dickenson et al. 2008). Likewise, many investigations have focused on operational parameters such as chlorine dose, water temperature, pH, and reaction time, which are regarded as influential for THM formation (Peters et al. 1980; Radiq and Rodriguez 2004). In addition, the bromide ion in raw water may also play an important role in the THM formation reaction, leading to a predominance of brominated THMs (Xue et al. 2008; Nikolaou 2004). As a result of the intense research in this area, during the last years, many mathematical models have been developed for predicting DBP and THM formations (Sohn et al. 2004; Platikanov et al. 2007). These models mainly focused on the prediction of total THM or chloroform formation. In spite of the large number of studies examining DBP formation of the isolated DOM fractions from different water sources by chlorination in different conditions, there are still contradictory results, mainly in the disinfectant dosage and pH effect (Nikolaou 2004; Lu et al. 2009).

The aim of this study was to utilize a method (Rodrigues et al. 2007) that has been proposed to determine the factors that affect the formation of the four THMs by chlorine disinfection of different DOM fractions (hydrophobic, colloidal, and transphilic) in a prototype laboratory simulation. DOM fractions were extracted from water samples of the Caldeirão dam (Guarda, Portugal) by a reverse osmosis water pre-concentration procedure, followed by dialysis and adsorption resins (Leenheer and Croué 2003).

THM formation is a complex process that depends on several factors and usually involves interactions of those factors. This study uses a factorial analysis strategy in order to identify the most THM formation-relevant factors and the way they influence THM formation (Rodrigues et al. 2007; Esteves da Silva et al. 2001). Two experimental designs, based on a Placket-Burman design of five factors (DOM fraction, chlorine dose, temperature, pH, and bromide ion concentration) and a Box-Behnken design for the analysis of three factors (DOM fraction concentration, chlorine dose, and temperature), were used to identify the most important factors in the formation of the four THM species and in the calculation of the corresponding response surfaces. A Box-Behnken design was chosen because it enabled a more precise study of the effect of several factors, as well as to obtain response surfaces with a relatively few number of experiments and with only three levels for each of the factors under analysis.

2 Materials and methods

As before mentioned, DOM fractions were obtained from Caldeirão dam in Guarda, Portugal. To be brief, a known volume of water (about 200 L) was concentrated using a reverse osmosis system. This system consisted of an electric pump, ionic exchange resins, and a reverse osmosis membrane. The concentrated water collected after the osmosis process was filtered using 0.45 μm Whatman cellulose acetate membranes and acidified to pH 2 with 6 M hydrochloric acid.

DOM fraction isolation was carried out in several stages: (1) deposition of the concentrated water solution, acidified to pH 1, in a dialysis bag (Spectrum, Spectra/Per), with a 3.5 kDa cutoff, (2) immersion during 36 h (three times 4 L) in a 0.1 M HCl solution (Merck); (3) immersion of the dialysis bag in 0.2 M HF solution followed by immersion of the dialysis bag in deionized water, The dialysis bag retained the colloid fraction, which was frozen and lyophilized, and (4) the dialyzed solution was sequentially eluted by XAD-8 (Fluka) and XAD-4 (Sigma) column, which adsorbed the hydrophobic (HPOF) and transphilic fractions, respectively. The HPOF and transphilic fraction, adsorbed onto XAD-8 and XAD-4 columns, respectively, were then eluted with a mixture of acetonitrile (Merck) and water in a proportion of 75% and 25%, respectively (Fig. 1). The solutions with the fractions were frozen and lyophilized (B. Braun, Christ LDC-1). Fourier transform infrared (FT-IR) spectra of the DOM fractions were done with a Bruker, vector 22 model, FTIR spectrophotometer.

THM ($CHCl_3$, chloroform; $CHBrCl_2$, bromodichloromethane; $CHBr_2Cl$, dibromochloromethane; and $CHBr_3$, bromoform) 200 g L^{-1} standard solution in methanol (SUPELCO, Bellefonte, USA) was used for the preparation of the aqueous standard solutions in the μg L^{-1} range (0.5–30 μg L^{-1}). All reagents were of analytical grade quality. The sodium hypochlorite used was a commercial solution.

2.1 Laboratory simulation of a water disinfection process

The disinfection process of the water sample containing DOM followed the following steps: we (a) placed a reaction vessel of 250 ml volume, with an aqueous solution of DOM (concentrations of 0.5, 2.7, and 5 mg L^{-1}), in a water bath at a constant temperature; (b) added to the DOM fraction solution a volume of sodium chloride to achieve a final concentration of 10 mg L^{-1} of chloride anion and a predetermined volume of potassium bromide (final concentration of 0.1, 0.55, and 1.0 mg L^{-1}); (c) adjusted pH with hydrochloric acid and/or sodium hydroxide to predetermined values (pH 6.0, 7.0, and 8.0); (d) added a predetermined amount of sodium hypochlorite to begin the disinfection reactions; (e) kept the sample at a constant temperature in a water bath; (f) 20.00 mL were removed at times zero (after sodium hypochlorite addition), 5 and 30 min to perform the THM analysis (after sample

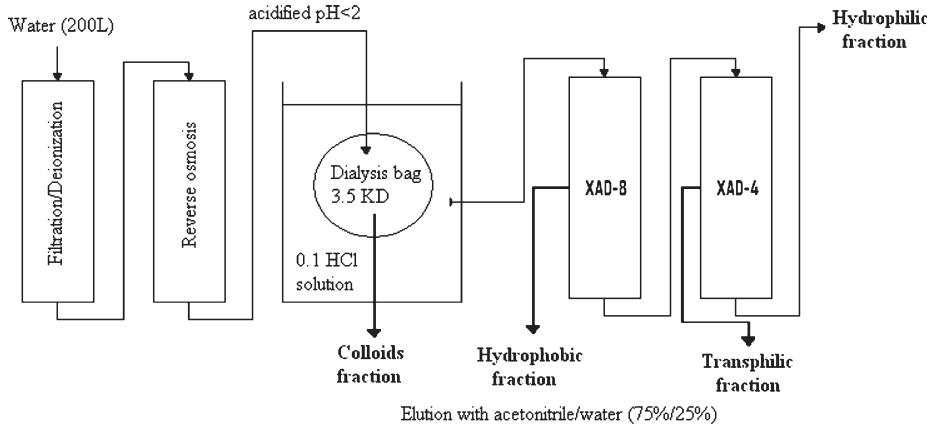


Fig. 1 Isolation protocol for colloids, hydrophobic, and transphilic fractions from water

collection, 30 μL of a solution 2 M sodium thiosulphate were added to eliminate free chlorine); (g) free chlorine was analyzed in all samples using a portable photometer kit (ELE International Limited, England).

THMs were analyzed by gas chromatograph-electron capture detector (GC-ECD). Gas chromatographic analyses were performed with a Chrompack CP9003 GC gas chromatograph equipped with a ^{63}Ni electron capture detector and a split/splitless injector. The column used was a Chrompack CP-Sil 13CB (25 m×32 mm, 1.2 μm) fused-silica column. Headspace analysis and GC-ECD parameters are shown in (Rodrigues et al. 2007). The limits of detection for the four THM of the HS-GC-ECD were in the range 0.3-1.4 μg L^{-1} and were calculated using the following criteria: $LOD = (a + 3Sy/x)$, where a is the intercept of the calibration curve and Sy/x is the random error in the y direction (Miller and Miller 2000).

2.2 Organization of the study

THM formation study was organized using two different experimental designs. First, a preliminary screening analysis was performed following a Placket-Burman design, focusing on the effect's evaluation of the five main factors: DOM fraction concentration, chlorine dose, pH, water temperature, and bromide anion concentration. Second, this preliminary study was followed using a Box-Behnken experimental design to estimate the effects of the two factors that have a natural variability (DOM concentration and water temperature) and one operationally controlled factor (chlorine dose) in the water treatment plant.

Table 1 shows concentrations and volumes of the factors under investigation, which were used for the preparation of simulated disinfection experiments. The DOM, chloride and bromide concentrations, and temperature were chosen to represent the natural variation of these parameters along the year in waters of the Caldeirão Dam. The concentrations of sodium hypochlorite were chosen such that there was always an excess of free chlorine. The minimum concentration of free chlorine (0.4 mg L^{-1}) led to free residual chlorine of 0.01 mg L^{-1} at the end of the experiment and the maximum concentration of free chlorine (2.4 mg L^{-1}) led to free residual chlorine of 0.1 mg L^{-1}. These values of the free residual chlorine did not decrease markedly after 60 min of subsequent reaction (the experimental time was about 90 min). This is in agreement with the fact of most THM growth rate was higher during the first 69–90 min (Korshin et al. 2002; Fabbricino and Korshin 2005; Fabbricino and Korshin 2009).

All calculations and data analysis were done using peak areas obtained from the recorded chromatogram using Chrompack CP-Maitre I/II software (version 2.5). The experimental design formulation and the corresponding analysis of the effects (ANOVA) and response surface calculations were done using The Unscramble v9.2 (CAMO PROCESS AS, Oslo, Norway).

3 Results

3.1 Characterization of the DOM fractions

To obtain information about the chemical structure of the investigated fractions and to relate it to the THM formation afterwards, an elemental analysis and FT-IR spectroscopy was performed. Elemental analysis and the H/C and the C/N atomic ratios of the three DOM fractions are presented in Table 2. The analysis of this table shows that the main differences among the three DOM fractions are the following: (1) higher elemental percentages of nitrogen and sulfur are detected in the colloidal fraction; (2) the H/C ratios in the HPOF fraction are lower than that in the others; (3) C/N ratio increases according to this order: colloidal fraction (lowest), transphilic fraction (middle) and HPOF (highest). Similar trends were observed for fractions analyzed by Leenheer and others (2000). These results show that the colloidal fraction is characterized with higher amounts of protein residuals in their molecules and a lesser

Table 1 Experimental designs, factors, and corresponding levels

Factors	Levels		
Placket-Burman design (8 + 3 center experiments)[a]			
DOM fraction concentration in mg L^{-1}	0.5	2.75	5
Bromide anion concentration (Br^-) in mg L^{-1}	0.1	0.55	1.00
pH	6.0	7.0	8.0
Water temperature (T) in °C	10	17.5	25
Chlorine (Cl_2) in mg L^{-1}	0.4	1.4	2.4
Box-Behnken design (12 + 3 center experiments)[a,b]			
DOM fraction concentration in mg L^{-1}	0.5	2.75	5
Water temperature (T) in °C	10	17.5	25
Chlorine (Cl) in mg L^{-1}	0.4	1.4	2.4

[a] A constant background concentration of 10 mg L^{-1} chloride anion was used in all experiments

[b] A constant background concentration of 10 mg L^{-1} chloride and 0.1 mg L^{-1} bromide anions were used in all experiments

Table 2 Elemental composition (mass %) and atomic DOM fractions ratios

DOM fraction	N	C	H	S	H/C	C/N
Colloidal	4.3	40.2	5.6	0.9	1.7	10.9
Transphilic	3.1	54.4	6.6	<0.3	1.5	20.5
Hydrophobic	1.3	56.2	5.9	0.7	1.3	50.4

amount of carbohydrate structures. However, the HPOF is dominated more with condensed aromatic structures (Leenheer et al. 2000; Leenheer 2004). The transphilic fraction takes intermediate position, possessing lower amounts of protein structures than the colloids and lower amounts of condensed aromatics than the HPOF.

IR spectra of the three fractions are shown in Fig. 2. The band at about 3,300 cm^{-1} is generally attributed to OH groups and bands at 2,900–2,930 cm^{-1} are assigned to CH, CH_2, and CH_3 stretching of the aliphatic groups. The bands at 1,640–1,680 cm^{-1} and 1,560–1,551 cm^{-1} are attributed to CO stretching vibration of carboxylic acids and ketones/quinones, respectively. The bands at about 1,450 cm^{-1} and 1,410 cm^{-1} are attributed to CH deformation of aliphatic and CH_3 groups, respectively. Also, bands in the 1,280–1,137 cm^{-1} regions are attributed to CO stretching of esters, ethers and phenols, and the band at about 830 cm^{-1} can be assigned to OH stretching vibration of carboxylic groups. In the IR spectrum of the colloidal fraction, the band located at 1,050 cm^{-1} due to CO groups is particularly important because these groups are indicative of the presence of N-acetylglucosamine (Croué 2004), formed from the oxidation of carbohydrates with amino groups from the bacterial cell wall structure (Hwang et al. 2001; Leenheer 2004). In the case of hydrophobic and transphilic fractions, there is a strong intensity band near 1,720 cm^{-1} which suggests a relatively greater abundance of carbonyl groups.

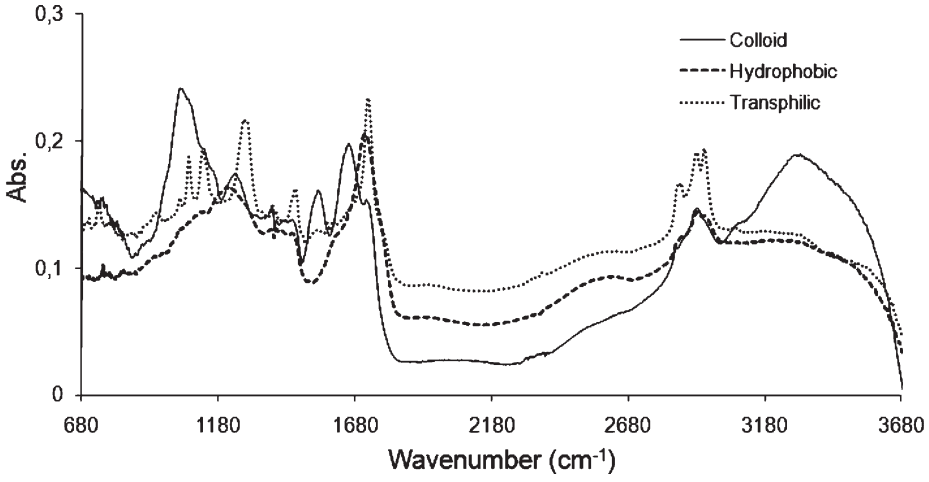


Fig. 2 FT-IR spectra of colloid, hydrophobic, and transphilic fractions

3.2 Qualitative analysis of factor effects using a Placket-Burman design

Table 1 shows the factors and levels used for the evaluation of the main effects on THM formation. The five parameters were studied using a Placket-Burman design (eight plus three center experiments). In this screening analysis, only the HPOF and colloid fractions were studied because the available quantity of transphilic fraction was very low. Table 3 shows the analysis of the effects of the five parameters on the four individual THMs and their total sum using Placket-Burman design experiments. The experimental error was estimated using replicated center samples.

Some results from the analysis in Table 3:

1. Higher the concentration of the colloidal and hydrophobic fraction greater the total THM production. This fact was expected because DOM concentration is the main precursor from which THM originates (Leenheer 2004; Lu et al. 2009). HPOF concentration was a very significant parameter for the formation of multichlorinated trihalomethanes, whereas colloidal fraction was more influential in the formation of mixed bromochloromethanes. Bromoform formation did not show any significant dependence on the two fraction concentrations.
2. pH positively affects the formation of $CHCl_3$ and $CHBrCl_2$ and slightly affects the formation of $CHBr_2Cl$ for both DOM fractions. In general, a pH increase (above pH 7) resulted in a reduction in the concentration of the brominated species and an increase in the concentration of $CHCl_3$. This occurs, possibly, because the formation of the hypochlorite ion ($Cl_2 + OH^- \rightleftharpoons OCl^- + Cl^- + H^+$ or $Cl_2 + H_2O \rightleftharpoons HOCl + Cl^- + H^+$ and $HOCl \rightleftharpoons OCl + H^+$) is shifted to the right with increasing pH (i.e., increasing OH^-). Consequently, hypochlorite ion concentration increases, leading to

Table 3 Qualitative analysis of the effects of the five parameters on the four THM and TTHM for HPOF and Colloidal fractions

Factor	$CHCl_3$	$CHBrCl_2$	$CHBr_2Cl$	$CHBr_3$	TTHM
HPOF fraction					
pH	+	+	NS	NS	+
Chlorine	++	++	+	NS	+++
Temperature	NS	NS	NS	NS	NS
Bromide concentration	−	−	NS	+	−
HPOF concentration	+++	+++	+	NS	+++
Colloidal fraction					
pH	++	+++	++	NS	+
Chlorine	++	+++	−	NS	+
Temperature	++	++	++	NS	+
Bromide concentration	−	−	+++	+	NS
Colloids concentration	++	+++	+++	NS	++

NS not significant factor; + means a positive effect, − means negative effect. *More than one* + *or* − signs mean stronger effects

predominance of chlorinated species (Nikolaou 2004). Also, the decreasing concentration of brominated THMs at pH values above 7–8 may be due to the following disproportionate reaction of the hypobromite ion at basic pH values (Bard et al. 1985): $OBr^- + 2HOBr \rightarrow BrO_3^- + 2Br^- + 2H^+$. From this equation, the OBr^- is disproportionate to bromate and bromide ions, neither of which reacts to organic matter. Chlorination is a typical electrophilic substitution which occurs in many steps, for example, in phenol groups the H^+ is release from phenolic ring to the solution. Thus, pH would affect the equilibrium of the reaction. The effects of pH on chlorination process must be explained simultaneously by the deprotonation of hypochlorite and/or the organic compound which may change the reaction kinetic (Westerhoff et al. 2004; Ge et al. 2006).

3. Similar behavior was observed for the chlorine dose used in the disinfection. The higher dose of Cl_2 generates higher concentrations of TTHM, especially $CHCl_3$ and $CHBrCl_2$. This is also expected since higher chlorine doses lead to an increase of hypochlorite ion concentration, as found in previous results in the literature (Rook 1974; Sohn et al. 2004).
4. Increasing the temperature produces an increase in the concentration of $CHCl_3$ and of mixed bromochloromethanes when the colloidal fraction is oxidized. The temperature parameter however does not have a significant effect during the chlorination of HPOF.
5. Bromide concentration produces a similar effect in THM formation of the two DOM fractions. High concentrations of bromide produce high concentrations of brominated THMs and relatively low concentrations of $CHCl_3$ and $CHBrCl_2$. It is well-known that $CHCl_3$ is formed in the reaction of DOM with OCl^- and $CHBr_3$ with OBr^-, and the amounts of $CHCl_3$ and $CHBr_3$ depend on the concentration of OCl^- and OBr^-, respectively. Higher concentrations of OBr^- are present in the case of higher concentration of bromide anion, resulting in the formation of a higher concentration of $CHBr_3$. As the concentration of OBr^- increases, the amount of $CHCl_3$ will decrease in response.

3.3 Preliminary analysis of the effect of the DOM fraction on the THM formation

Box-Behnken design analyses where performed to investigate the effect of DOM concentration, chlorine dose and temperature factors on THM formation. Table 1 shows the levels of these three factors under analysis. In this analysis, bromide concentration was kept constant at 0.1 mg L^{-1} since the natural water from the Caldeirão Dam has low concentrations of it due to an absence of geological or anthropological sources of bromide ions. Also, pH was kept constant at 7.0 because this is a common operational procedure implemented in water treatment plants. In spite of temperature parameter in Placket-Burman design was not a significant factor in the formation of THM in the HPOF fraction, we consider this in the Box-Behnken design because it may be important in the formation of THM in the other fractions and because this parameter have a great variability in water treatment plant along year seasons.

As shown in Table 4, the amount of generated THMs in the experiment was characterized by a rather large range in

Table 4 Concentration (μg L^{-1}) ranges of the four THM and total THM generated from the disinfection of aqueous solutions of the three DOM fractions

THM	Colloidal	HPOH	Transphilic
$CHCl_3$	3.3–6.0	1.6–4.6	4.0–21.3
$CHBrCl_2$	3.5–4.5	3.6–11.0	37–8.0
$CHBr_2Cl$	4.0–5.4	4.3–15.0	4.4–13.1
$CHBr_3$	3.2–7.1	5.1–12.4	3.2–15.7
TTHM	14.8–22.2	15.0–42.9	17.7–39.6

concentration. This results show the relevance of this three factors under investigation for THM generation. Figure 3a shows the pie plots with the percentage contributions of the three DOM fractions in the production of the four THM. A preliminary analysis of Fig. 3a shows the following trends: (1) the transphilic fraction is responsible for the production of about half the amount of chloroform, followed by the colloids and HPOF; (2) there is an increase in the percentage of the most brominated THM in the HPOF and transphilic DOM fractions and there is a decrease in their percentages in the colloidal DOM fraction.

Figure 3b shows the pie plot with the percentage contribution in the three DOM fractions for the total production of THM. This plot suggests the following trend in the order of total THM production: transphilic > hydrophobic > colloidal. These results are in agreement with the conclusions of Marhaba et al. (2006), where the differences of DBPs yields between the fractions are possibly due to their different characteristics of functional groups and structures. Indeed, the colloidal fraction shows a lower amount of aromatic and polyphenolic compounds than the transphilic and hydrophobic fraction, which can explain a greater THM generation by the last two fractions.

3.4 Response surface analysis in the formation of the four individual THMs

ANOVA of factor effects in the formation of individual THMs was done as well and linear models were obtained (data not shown). Included in the models where the coefficients of the factors that were statistically significant at the 5% level, as well as coefficients with absolute values higher than the corresponding standard deviations.

Colloidal fraction models

$$\mathrm{CHCl_3}\left(\mu\mathrm{gL^{-1}}\right) = 4.8 + 0.2\,\mathrm{colloids} + 0.2\mathrm{T} \times \mathrm{Cl} - 0.25\ \mathrm{colloids} \times \mathrm{T} + 0.2\,\mathrm{Cl}^2$$

$$\mathrm{CHBrCl_2}\left(\mu\mathrm{gL^{-1}}\right) = 3.7 + 0.03\,\mathrm{colloids} + 0.01\,\mathrm{T} + 0.01\mathrm{Cl} + 0.1\,\mathrm{T} \times \mathrm{Cl} + 0.1\,\mathrm{Cl}^2 - 0.06\ \mathrm{colloids}^2$$

$$\mathrm{CHBr_2Cl}\left(\mu\mathrm{gL^{-1}}\right) = 4.6 + 0.005\,\mathrm{T} + 0.02\,\mathrm{Cl} + 0.1\ \mathrm{colloids} - 0.04\,\mathrm{T} \times \mathrm{Cl} - 0.03\,\mathrm{T} \times\ \mathrm{colloids} + 0.3\,\mathrm{Cl} \times \mathrm{colloids} + 0.06\,\mathrm{Cl}^2 - 0.1\ \mathrm{colloids}^2$$

$$\mathrm{CHBr_3}\left(\mu\mathrm{gL^{-1}}\right) = 3.5 + 0.05\,\mathrm{Cl} + 0.3\,\mathrm{colloids} + 0.6\,\mathrm{Cl} \times \mathrm{colloids}$$

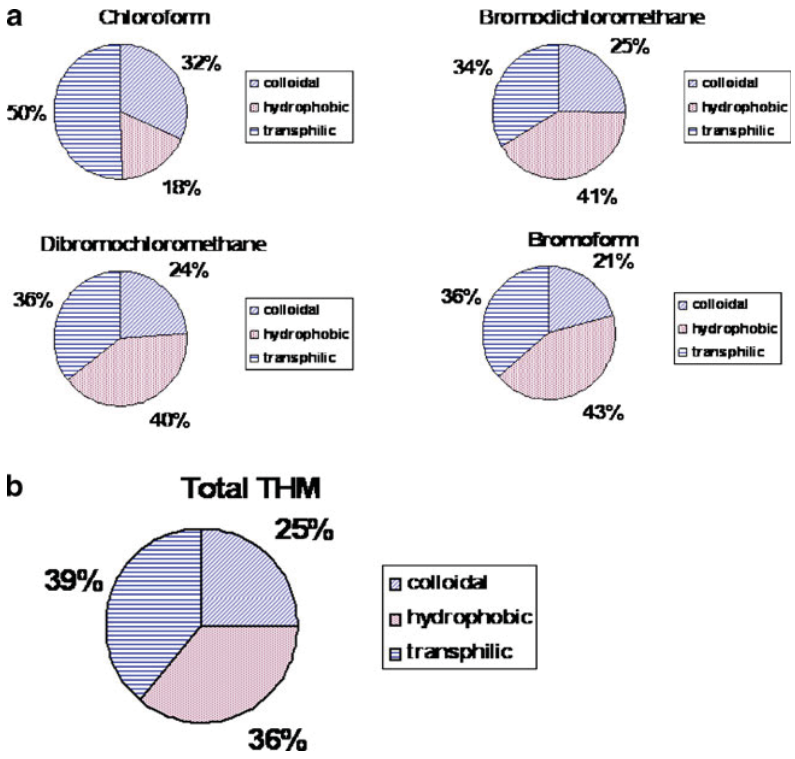


Fig. 3 Percentages of the four THM (**a**) and TTHM (**b**) generated from the three DOM fractions

The three factors under investigation play significant roles in a quite complex THM generation. The results show that the amount of chlorine positively affects the formation of the four THMs.

Increasing the colloid concentration leads to a decrease in the production of mixed chloro-bromomethanes, but produces a slight increase in the production of chloroform and a strong increase in the bromoform formation. This observation is contrary to the Placket-Burman screening analysis that the colloidal fraction is more influential in the formation of mixed bromochloromethanes than bromoform. This erroneous result is a consequence of a lack of degrees of freedom of the screening design which results in an unreliable detailed factorial analysis as consequence of the mixing effect of the factors. This erroneous result must be solved in the future by doing more experimental analysis in the same conditions and under less variable factors.

The temperature factor plays a controversial role. It has a strong independent effect and also interacts with chlorine in the production of $CHBrCl_2$. It has a positive influence as an independent factor, but the interaction with the chlorine produces the opposite effect on the formation of $CHBr_2Cl$. The role of the global temperature balance is therefore not clear in the formation of chloroform and bromoform during the colloid fraction chlorination.

HPOF fraction models

$$CHCl_3\left(\mu gL^{-1}\right) = 2.8 + 0.02\,T + 0.01\,Cl + 0.5\,HPOF + 0.1\,T \times HPOF - 0.1\,Cl^2 + 0.1\,HPOF^2$$
$$CHBrCl_2\left(\mu gL^{-1}\right) = 6.4 + 0.1\,T + 0.1\,Cl + 1\,HPOF + 0.3\,T \times HPOF + 0.3\,Cl \times HPOF - 0.5\,Cl^2$$
$$CHBr_2Cl\left(\mu gL^{-1}\right) = 9 + 0.1\,T + 0.1\,Cl + 1.4\,HPOF + 0.6\,T \times HPOF + 0.5\,Cl \times HPOF - 1.2\,Cl^2$$
$$CHBr_3\left(\mu gL^{-1}\right) = 10.7 + 0.1\,T + 0.1\,Cl + 1.1\,HPOF + 0.4\,Cl \times HPOF - 0.6\,T^2 - 1.2\,Cl^2 - 0.7\,HPOF^2$$

A consistency can be found in all four models: any excess of the amount of added chlorine generally leads to an increase in THM concentration. The increase of HPOF concentration independently, or when HPOF concentration interacts with the other two factors, increases the formation of $CHCl_3$, $CHBrCl_2$ and $CHBr_2Cl$. An interesting result is that high levels of HPOF fraction concentration, chlorine dose and temperature will reduce the formation of $CHBr_3$. These surprising results can be explained by the smaller bromide ions concentration, available in solution, and not by the HPOF concentration (Marhaba et al. 2006).

Also worth noting is that a positive interaction exists between HPOF fraction concentration and chlorine for the formation of brominated THMs. Moreover, in the disinfection of HPOF, it should be mentioned that temperature positively affects the production of chloroform and mixed chloro-bromomethanes.

Transphilic fraction models

$$CHCl_3\left(\mu gL^{-1}\right) = 6.1 - 0.4\,T + 0.1\,Cl - 1.8\,T \times Cl + 1.7\,Cl^2$$
$$CHBrCl_2\left(\mu gL^{-1}\right) = 4.8 + 0.04\,T + 0.1\,Cl + 0.5\,transphilic + 0.4\,T \times transphilic + 0.5\,Cl \times transphilic$$
$$CHBr_2Cl\left(\mu gL^{-1}\right) = 8 + 0.1\,T + 0.1\,Cl + 0.9\,transphilic + 0.6\,T \times transphilic + 1\,Cl \times transphilic - 0.7\,T^2$$
$$CHBr_3\left(\mu gL^{-1}\right) = 8.7 + 0.3\,T + 1\,transphilic + 1.6\,T \times Cl + 1.4\,T \times transphilic - 1.2\,Cl^2 - 1\,transphilic^2$$

Analysis of the three models for the formation of $CHCl_3$ shows different trends. The disinfection of colloidal and HPOF fractions derived very complex models. Controversially, the obtained model for the transphilic fraction chlorination depends only on chlorine dose and temperature and not on DOM concentration.

Another difference, compared to the other two experiments, is the interaction of the transphilic concentration with the other two factors (Cl and T) in the formation of brominated trihalomethanes.

In general, the global analysis of the DOM fraction disinfection reveals a high model complexity, i.e., many factor interactions were involved. Some models show significant lack of fit. Nevertheless, this result may be due to the relatively high precision of the THM measurements comparatively to a less precision in the control of operational factors, like kinetic time reaction, in the experimental procedures (Rodrigues et al. 2007). Even if we accept the existence of model misadjustment, factor effects on the THM formation are still realistic. A common feature to all DOM fractions is that the

highest values of all factors are responsible for the higher concentration of all THMs. Some exceptions to the above mentioned fact can be noticed for $CHCl_3$ formation, when transphilic fraction solution is disinfected. Relatively high values of the three factors cause the disinfection of HPOF and transphilic fractions solutions to form higher concentrations of $CHBr_3$. In contrast, this fact was not valid when colloids were oxidized. Also, relatively similar THM concentrations are formed when HPOF is disinfected and the factor levels are at a similar degree. Globally reducing the amount of HPOF and transphilic concentrations, together with keeping the temperature low, will yield a lower concentration of brominated THMs.

3.5 Response surface analysis of TTHM formation

Since EU regulation considers the sum of all individual THMs, the effects of DOM concentration, chlorine dose and temperature on the total THM formation was analyzed. ANOVA of factor effects for the formation of TTHM was calculated and linear models were obtained as well as response surfaces (Fig. 4).

$$\text{Colloidal fraction model}: \text{TTHM}\left(\mu\text{gL}^{-1}\right) = 16.6 + 0.06\,\text{T} + 0.1\,\text{Cl} + 0.65\ \text{colloids} + 0.3\,\text{T} \times \text{Cl} + \text{colloids} \times \text{Cl} + 0.6\,\text{Cl}^2$$

$$\text{HPOF fraction model}: \text{TTHM}\left(\mu\text{gL}^{-1}\right) = 28.9 + 0.3\,\text{T} + 0.3\,\text{Cl} + 4\,\text{HPOF} + 1.2\,\text{T} \times \text{HPOF} + 1.2\,\text{Cl} \times \text{HPOF} - 3\,\text{Cl}^2$$

$$\text{Transphilic fraction model}: \text{TTHM}\left(\mu\text{gL}^{-1}\right) = 27.5 + 0.3\,\text{Cl} + 2.2\,\text{transphilic} + 3.2\,\text{T} \times \text{transphilic} - 2.3\,\text{Cl}^2$$

The analysis of these models and, especially, the large contribution of interaction effects among the factors confirm that TTHM formation is a complex process. This formation depends globally not only on the type of DOM fraction and its concentration, but as well as on the individual chlorine dose and temperature and on the involved interactions among these three factors. An easier visualization of the combined effects of the three factors can be observed in Fig. 4. The most

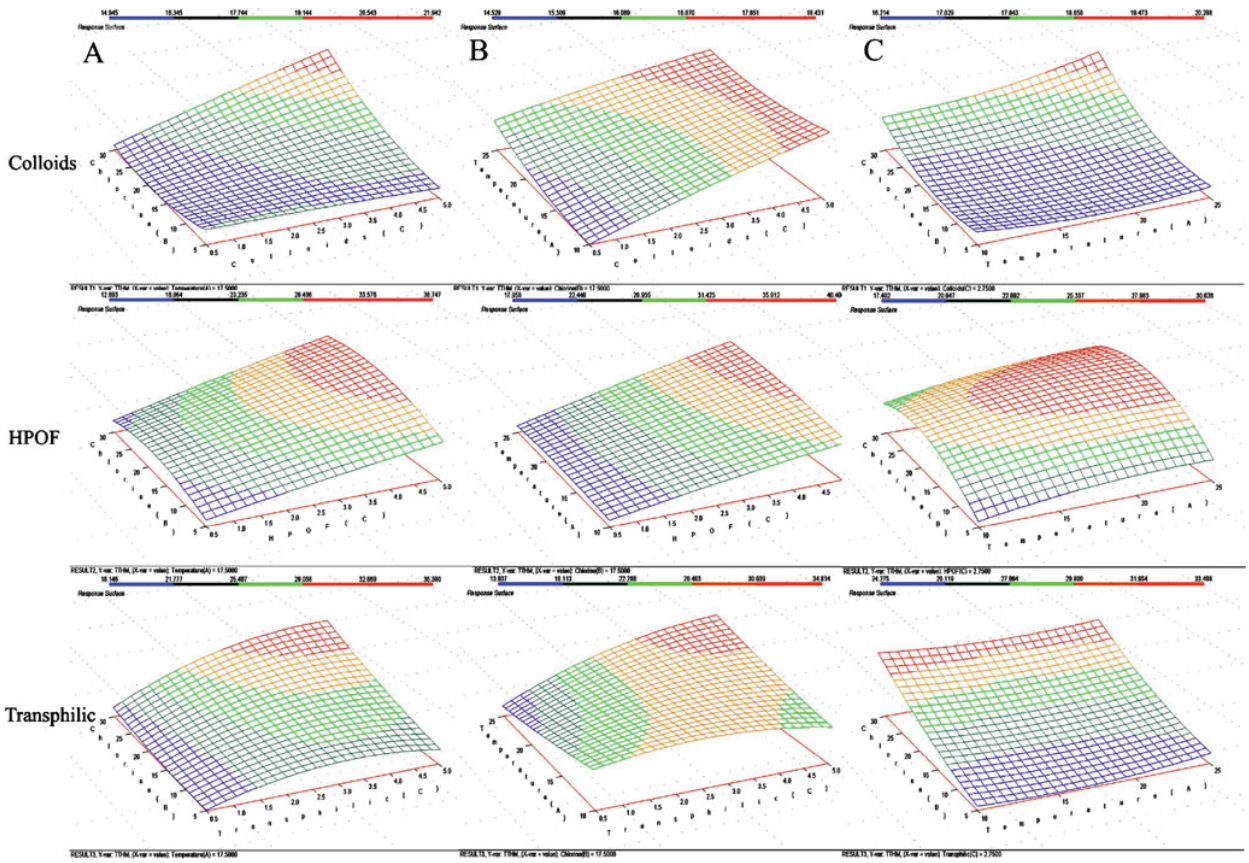


Fig. 4 Response surface of TTHM formation as a function of the three factors and DOM fraction type: **a** DOM fraction vs. chlorine; **b** DOM fraction vs. temperature; **c** chlorine vs. temperature

important is DOM concentration. It positively affects the formation of TTHM, solely or when interacted with chlorine dose and temperature. Although less important, temperature and chlorine also affect TTHM formation.

4 Discussion

Factorial analysis of the water disinfection process is a useful approach for a more comprehensive understanding. This work illustrates a particular study of a local disinfection process in Portugal that leads to the formation of THMs, but it can be extended to other real water works plant management systems. Moreover, it can incorporate the investigation of various factors and their interaction. Combined use of fast screening Placket-Burman and of detailed assessed Box-Behnken experimental designs allowed optimal inferences about more influential parameters in the formation of trihalomethans and of their effects and interactions. Whereas the fast screening did not allow for the detection of the effects of the parameters on bromoform formation probably because of the small number of samples used, the use of the more detailed Box-Behnken design showed clearly this dependence. Therefore, the combined use of both approaches provided a better assessment and reliability of the finally obtained results.

Valuable information was obtained regarding the effect of the DOM fraction type and concentration. The colloidal fraction, richest in nitrogen atoms and poorest in carbon atoms, is approximately responsible for 20–30% of the formation of each individual THM. On average, it contributes a quarter of the formation of the total sum of THMs. A possible explanation for this low contribution in THM production is that this colloidal fraction is more responsible for the production of the other DBPs such as haloacetonitriles and others including N-atoms in their molecules (Ueno et al. 1996).

The fraction where the formation of brominated THMs is more important is the hydrophobic fraction, which is the most carbon enriched fraction. $CHCl_3$ formation strongly depends on the disinfection of the transphilic fraction. Both fractions take similar percentages in the total sum of THM formation—about 40%.

Special attention must be paid when bromide anions are present in raw water. This study reveals that even when only a small quantity of bromide anions exists in the water, the formation of brominated trihalomethanes is highly favored no matter what DOM fraction has been oxidized. This is a consequence of the rapid oxidation of bromide to bromine (hypobromous acid and hypobromite ion). Once formed, bromine is capable of participating in reactions analogous to those of chlorine. The presence of both halogens leads to competition for substitution at suitable carbon atoms in the DOM. Hypobromous acid is a more powerful halogenating agent than hypochlorous acid and this result in a greater incorporation of bromine into DOM. This result is very relevant for risk assessment management since brominated trihalomethanes are considered stronger carcinogen agents than chloroform (Muellner et al. 2007).

5 Conclusions

In order to reduce the concentration of THMs in drinking water, DOM concentrations should be reduced in the water prior to the disinfection. However, taking into consideration the natural complexity of DOM, different operations have to be used to quantitatively remove DOM from the raw water. In fact, the information resulting from this work is in agreement with our previous knowledge about water disinfection. DOM fraction concentration is the most important factor among the investigated ones. No matter which DOM fraction was used, a higher concentration leads to the production of higher amounts of all THMs. Furthermore, water disinfection efforts should focus on the elimination of higher concentrations of each DOM fraction prior to the chlorination. Since the transphilic and HPOF fractions generated 75% of TTHM formed, should the effort to remove the DOM focus on these two organic fractions. The coagulation/flocculation and the filtration stages are the main mechanisms, in a classic water plant treatment, to remove DOM in particular the colloidal and the hydrophobic fraction with a removal of about 70%. The efficiency of the alum treatment for the fractions more hydrophilic is only about 16% (Kim and Yub 2005; Bose and Reckhow 2007). The minimization of the DOM in public water depends, mainly, of a good control of the alum coagulant quantity and raw water pH value.

Special attention must be also paid to the chlorine dose used in disinfection processes. Formation of all THMs is favored by high amounts of chlorine. However, its use is undoubtedly important for the oxidation of raw water and for the disinfection and future avoidance of pathogen regrowth in the distribution system. Chlorine levels should be reduced as low as possible without compromising the microbiological quality of the supplied drinking water, which is the primary concern in the delivery of safe drinking water. In real water works plant management, an investigation of chlorine-DOM fraction type interactions should be undertaken. Temperature appeared also to be significant in THM formation, especially when the DOM concentration and chlorine dose were controlled and constant. Actually, it increases the speed of THM forma-

tion; but the response surface plots reveal that temperature is less significant when chlorine dose and DOM fraction concentrations are low. However, it has been established that temperature decreases the water solubility of THMs and in warm conditions water aeration after chlorination reduces the total amount of THMs present in water.

6 Recommendations and perspectives

The methodology used in this paper is appropriate and can be used in the analysis of other groups of DBPs, mainly the emerging DBPs, like, for example the haloacetic acids, haloacetonitriles, haloketones, or haloacids, which some of them have a more toxic and harmful effect in human health.

The THM reduction in consumption water can be achieved reducing the DOM concentration (mainly the hydrophobic and transphilic fraction) and chlorine dose without compromise the water microbiology quality. Bromide ion concentration control is also very important to minimize the brominated THM formation. These can be previously minimized if the water source contains low concentration of organic and inorganic matter. The use of granular activated carbon and membrane filtration prior the pre-oxidation/disinfection can reduce DOM and consequently the DBPs formation. Moreover, none of the currently available treatment approaches can completely remove pathogens and the precursors to DBP formation. At this moment the solution to minimize the problem is to get a good control in all the process and operational parameters of water treatment.

Acknowledgment Financial support from Fundação para a Ciência e Tecnologia (Lisboa) (FSE-FEDER) (Project PTDC/QUI/71001/2006) is acknowledged. Research grant CTQ2006-15052-C02-01 from the Spanish government is acknowledged to provide research funds to perform this investigation.

References

Bard AJ, Parsons R, Jordan J (1985) Standard potentials in aqueous solution. Marcel Dekker, Inc, New York, pp 70–83

Bellar TA, Lichtenberg JJ, Kroner RC (1974) The occurrence of organohalides in chlorinated drinking water. J Am Water Works Assoc 66:703–706

Bose P, Reckhow DA (2007) The effect of ozonation on natural organic matter removal by alum coagulation. Water Res 41:1516–1524

Chang PB, Young TM (2000) Kinetics of methyl tert-butyl ether degradation and by-product formation during UV/H_2O_2. Water Treat Water Resour 34:2233–2240

Council Directive 98/83/EC of 3 November 1998 on the quality of water intended for human consumption. Official Journal of the European Communities L 330/32, 5.12.98.

Croué JP (2004) Isolation of humic and non-humic nom fractions: structural characterization. Environ Monit Assess 92:193–207

Croué JP, Violleau D, Bodaire C, Leqube B (1999) Removal of hydrophobic and hydrophilic constituents by anion exchange resin. Water Sci Technol 40:207–214

Dickenson E, Scott Summers R, Croué JP, Gallard H (2008) Haloacetic acid and trihalomethane formation from the chlorination and bromination of aliphatic β-dicarbonyl acid model compounds. Environ Sci Technol 42:3226–3233

Esteves da Silva J, Dias J, Magalhaes J (2001) Factorial analysis of a chemiluminescence system for bromate detection in water. Anal Chim Acta 450:175–184

Fabbricino M, Korshin GV (2005) Formation of disinfection by-products and applicability of differential absorbance spectroscopy to monitor halogenation in chlorinated coastal and deep ocean seawater. Desalination 176:57–69

Fabbricino M, Korshin GV (2009) Modelling disinfection by-products formation in bromide-containing waters. J Hazard Mater 168:782–786

Gallard H, von Gunten U (2002) Chlorination of natural organic matter: kinetics of chlorination and of THM formation. Water Res 36:65–74

Ge F, Zhu L, Chen H (2006) Effects of pH on the chlorination process of phenols in drinking water. J Hazard Mater B133:99–105

Hamidin N, Yu QJ, Connell DW (2008) Human health risk assessment of chlorinated disinfection by-products in drinking water using a probabilistic approach. Water Res 42:3263–3274

Hrudey SE (2009) Chlorination disinfection by-products, public health risk tradeoffs and me. Water Res 43:2057–2092

Hwang CJ, Krasner SW, Sclimenti MJ, Amy GL, Dickenson E, Bruchet A, Prompsy C, Filippi G, Croué JP, Violleau D, Leenheer JL (2001) Polar NOM: characterization, DBPs, treatment, AWWA Research Foundation and American Water Works Association (USA)

Kim HC, Yub MJ (2005) Characterization of natural organic matter in conventional water treatment processes for selection of treatment processes focused on DBPs control. Water Res 39:4779–4789

Korshin GV, Wu WW, Benjamin MM, Hemingway O (2002) Correlations between differential absorbance and the formation of individual DBPs. Water Res 36:3273–3282

Leenheer JA (2004) Comprehensive assessment of precursors, diagenesis, and reactivity to water treatment of dissolved and colloidal organic matter. Water Sci Technol Water Supply 4:1–9

Leenheer JA, Croué JF (2003) Characterizing aquatic dissolved organic matter. Environ Sci Technol 37:18A

Leenheer JA, Croue JP, Benjamin M, Korshin GV, Hwang CJ, Bruchet A, Aiken G (2000) Comprehensive isolation of natural organic matter for spectral characterization and reactivity testing. In: Barrett S, Krasner SW, Amy GL (eds) Natural organic matter and disinfection by-products. American chemical society symposium series 761, Washington DC

Leehneer JA, Rostad CE, Barber LB, Schroeder RS, Anders R, Davisson ML (2001) Nature and chlorine reactivity of organic constituents from reclaimed water in groundwater, Los Angeles County, California. Environ Sci Technol 35:3869–3876

Lewis C, Suffet IH, Ritz B (2006) Estimated effects of disinfection by-products on birth weight in a population served by a single water utility. Am J Epidemiol 163:38–47

Lu J, Zhang T, Ma J, Chen ZH (2009) Evaluation of disinfection by-products formation during chlorination and chloramination of disolved natural organic matter fractions isolated from a filtered river water. J Hazard Mater 162:140–145

Marhaba TF, Mangmeechaib A, Chaiwatpongsakorn C, Pavasant P (2006) Trihalomethanes formation potential of shrimp farm effluents. J Hazard Mater A136:151–163

McGeehin MA, Reif JS, Becher JC, Mangione EJ (1993) Case-control study of bladder cancer and water disinfection methods in Colorado. Am J Epidemiol 138:492–501

Miller JN, Miller JC (2000) Statistics and chemometrics for analytical chemistry, 4th edn. Pearson Prentice Hall, Dorchester, pp 107–150

Muellner M, Wagner ED, Mccalla K, Richardson SD, Woo YT, Plewa MJ (2007) Haloacetonitriles vs. regulated haloacetic acids: are nitrogen-containing DBPs more toxic? Environ Sci Technol 41:645–651

Nieuwenhuijsen MJ, Toledano MB, Eaton NE, Fawell J, Elliot P (2000) Disinfection by-product in water and their association with adverse reproductive outcomes: a review. Occup Environ Med 57:73–85

Nikolaou AD (2004) Investigation of the formation of chlorination by-products in water rich in bromide and organic matter content. J Environ Sci Health A39:2835–2853

Norwood D, Johnson J, Christman R, Hass J, Bobenrieth M (1980) Reactions of chlorine with selected aromatic models of aquatic humic material. Environ Sci Technol 14:187–190

Panyapinyopol B, Marhaba TF, Kanokkantapong V, Pavasant P (2005) Characterization of precursors to trihalomethanes formation in Bangkok source water. J Hazard Mater 120:229–236

Peters CJ, Young RJ, Perry R (1980) Factors influencing the formation of haloforms in the chlorination of humic substances. Environ Sci Technol 14:1391–1395

Platikanov S, Puig X, Martin J, Tauler R (2007) Chemometric modelling and prediction of trihalomethane formation in Barcelona's water works plant. Water Res 41:3394–3406

Radiq S, Rodriguez MJ (2004) Disinfection by-products (DBPs) in drinking water and predictive models for their occurrence: a review. Sci Total Environ 321:21–46

Reif JS, Hatch MC, Bracken M, Holmes L, Schwetz BA, Singer PC (1996) Reproductive and developmental effects of disinfection by-products in drinking water. Env Health Persp 104:1056–1061

Richardson SD, Thruston AD (2003) Tribromopyrrole, brominated acids and other disinfection byproducts produced by disinfection of drinking water rich in bromide. Environ Sci Technol 37:3782–3793

Richardson SD, Plewa MJ, Wagner ED, Schoeny R, DeMarini DM (2007) Occurrence, genotoxicity, and carcinogenicity of regulated and emerging disinfection by-products in drinking water: a review and roadmap for research. Mutat Res 636:178–242

Rodrigues P, Esteves da Silva J, Antunes M (2007) Factorial analysis of the trihalomethanes formation in water disinfection using chlorine. Anal Chim Acta 595:266–274

Rook JJ (1974) Formation of haloforms during chlorination of natural waters. Water Treat Exam 23:234–243

Rostad CE, Martin BS, Barber LB, Leehneer JA, Daniel SR (2000) Effect of a constructed wetland on disinfection byproducts: removal processes and production of precursors. Environ Sci Technol 34:2703–2710

Savitz DA, Singer PC, Herring AH, Hartmann KE, Weinberg HS, Makarushka C (2006) Exposure to drinking water disinfection by-products and pregnancy loss. Am J Epidemiol 164:1043–1051

Simpson KL, Hayes KP (1998) Drinking water disinfection by-products: an Australian perspective. Water Res 32:1522–1528

Sohn J, Amy G, Cho J, Lee Y, Yoon Y (2004) Disinfectant decay and disinfection by-products formation model development: chlorination and ozonation by-products. Water Res 38:2461–2478

Ueno H, Moto T, Sayato Y, Nakamuro K (1996) Disinfection by-products in the chlorination of organic nitrogen compounds: by-products from kynurenine. Chemosphere 33:1425–1433

Villanueva CM, Cantor KP, Grimalt JO, Malats N, Silverman D, Tardon A, Garcia-Closas R, Serra C, Carrato A, Castaño-Vinyals G, Marcos R, Rothman N, Real FX, Dosemeci M, Kogevinas M (2007) Bladder cancer and exposure to water disinfection by-products through ingestion, bathing, showering, and swimming in pools. Am J Epidemiol 165:148–156

von Gunten U, Driedger A, Gallard H, Salhi E (2001) By-prodcuts formation during drinking water disinfection: a tool to assess disinfection efficiency. Water Res 35:2095–2099

Westerhoff P, Chao P, Mash H (2004) Reactivity of natural organic matter with aqueous chlorine and bromine. Water Res 38:1502–1513

Xue S, Zhao QL, Wei LL, Jia T (2008) Effect of bromide ion on isolated fractions of dissolved organic matter in secondary effluent during chlorination. J Hazard Mater 157:25–33

Results and Discussion

- Characterization of DOM fractions using Fourier transform infrared spectroscopy (FT-IR) and elemental analysis´

Elemental analysis of these DOM fractions provided information about their H/C and C/N atomic ratios (see Table 2, Article 3). The colloidal fraction had higher nitrogen and sulphur percentages and the lowest C/N ratio. HPOF fraction was characterized by the highest carbon percentages, lower H/C ratio and the highest C/N ratio. These results were consistent with previously published results, allowing the comparison of our results with others for different geographical locations. These results confirmed that the colloidal fraction contained the higher amounts of protein structures and that, in contrast, the HPOF fraction was characterized by highly condensed aromatic structures. The transphilic fraction had lower amounts of protein structures than the colloidal fraction and lower amounts of condensed aromatics than the HPOF fraction.

Our conclusions were similar from the FTIR characterisation of DOM fractions. The IR colloidal fraction was characterized by the observed absorption band located at 1,050 cm^{-1}, due to CO groups of N-acetylglucosamine (Croué, 2004), which are formed by the oxidation of carbohydrates with amino groups in bacterial cell wall structures (Leenheer, 2004). Hydrophobic and transphilic fractions had not exhibit a strong intensity IR band near 1,720 cm^{-1}, which was suggestive of greater abundance of carbonyl groups.

- Screening Placket-Burman statistical design for the evaluation of DOM fraction concentration, chlorine dose, temperature, pH, and bromide ion concentration as factors

After the characterization of DOM fractions, a laboratory experiment was performed by simulating the water disinfection process. To identify the most important factors for THMs formation, an experimental design, based on factorial analysis, was used. More specifically, two popular designs were selected: 1) Placket-Burman screening factor design with selected five factors (DOM fraction, chlorine dose, temperature, pH, and bromide ion concentration), and 2) Box-Behnken design for a comprehensive analysis of the effects of three pre-selected factors (DOM fraction concentration, chlorine dose, and temperature).

Table 1 of Article 3 contains more details regarding the experimental conditions and values of factors used in the preparation of the laboratory disinfection experiments.

The preliminary screening (i.e., Placket-Burman design analysis) was performed with eight mixtures plus three experiments (the center point of the design), which were performed using only HPOF and colloid fractions due to the limited amounts of transphilic DOM fraction. Table 3 (Article 3) shows the analysis regarding the effects of five factors on four individual THMs and on their total sum.

The results suggested that high concentrations of colloidal and hydrophobic DOM fractions generated greater quantities of THMs. This finding was consistent with previous research where DOM fraction was considered as main precursor for THMs formation (Leenheer 2004).

The effect of pH on the colloidal DOM fraction disinfection was found to be stronger than such on the HPOF. More specifically, $CHCl_3$ and $CHBrCl_2$ concentrations increased above pH 7, when both DOM fractions reacted with chlorine. In contrast, a systematic reduction of the formation of multi-brominated THMs species was observed, more likely due to the increase of hypochlorite ion concentration at increasing levels of OH^- concentrations. This reaction mechanism, as also confirmed by Nikolaou (2004), explained the formation of multi-chlorinated THMs. At pH values above 7, a disproportionate reaction of hypobromite to bromate and bromide ions (neither of which reacts with organic matter) reduced the formation of multi-brominated THMs.

Chlorine dose was found to be important for the formation of $CHCl_3$ and $CHBrCl_2$. It additionally increases the total THMs concentrations, because the higher concentration of Cl_2 produces higher concentrations of hypochlorite ion.

Temperature was found to be a significant factor during the colloidal fraction disinfection. Increase of $CHCl_3$ and mixed bromo-chloromethanes concentrations were observed at high temperature levels. However, the effect of temperature is not linear, because it has a positive effect on THMs formation up to 38-40 degrees of Celsius and above 40 degrees evaporation processes effectively take place.

The results from the Placket-Burman experimental design revealed the importance of bromide ion concentrations for the disinfection process. Higher bromide concentrations were found to generate higher concentrations of multi-brominated THMs and lower concentrations of multi-chlorinated THMs. Such results could be explained with the increase of OBr^-, when higher concentrations of Br^- were present in the water system.

- *Detailed Box-Behnken factor design for evaluation of DOM fraction concentration, chlorine dose and temperature as factors affecting THMs formation*

The Box-Behnken design analyses were highly useful in revealing the specific effects of DOM concentration, chlorine dose and temperature on the THMs formation. Table 1 (Article 3) shows the chosen levels of these three factors and Table 4 - the amounts of generated THMs during the experiment.

The pie plots of Figure 3a display the contributions (in percentage) of the three DOM fractions for the production of four THMs. The results suggested that the chlorination of colloidal fraction likely explain the formation of approximately a half of the total amount of chloroform. In contrast, most brominated THMs were formed during the chlorination of HPOF and transphilic DOM fractions.

Box-Behnken design was also useful for the calculation of (ANOVA) factor effects on the formation of particular THMs. The linear models, along with the corresponding response surfaces, were visualized (see Figure SI-1, Article 3) to better represent the process of THMs formation. In general**,** the results confirmed that THMs formation mechanisms are complex. Not only the three investigated factors (e.g. organic matter fraction, chlorine dose and temperature) but also their interactions had a significant effect on the THMs formation. These results were summarized as follows:

- High amount of chlorine doses favored the formation of the four THMs, independent of the DOM precursor fraction.

- The role of the factor temperature was not conclusive. Although it had a strong independent positive effect in numerous cases, its interactions (particularly, with chlorine doses) had a negative effect on the formation for specific THMs. Therefore, it was not possible to generalize, because the effect depends on the specific settings regarding the formation of particular THMs.

- Colloidal fraction disinfection favored an increase of chloroform and bromoform.

- High amounts of HPOF fraction increased the formation of $CHCl_3$, $CHBrCl_2$ and $CHBr_2Cl$. In interaction with the other two factors, HPOF played an important effect on the $CHBr_3$ formation, when high concentrations of bromide ions were present in the water system.

- High concentrations of transphilic fraction were found to play a moderate role on $CHCl_3$ formation. In contrast, the interaction of transphilic fraction with Cl_2 and temperature had a strong impact on the formation of brominated trihalomethanes.

- TTHM formation was characterized as complex, where different factors (such as type of DOM fraction, temperature and chlorine doses) and their interactions jointly determine the process.

To summarize, the main conclusions in Article 3 were that transphilic fraction (characterized by multiple functional groups) participated in the formation of all THMs. In particular, colloidal and HPOF fractions played an important role. During chlorination, colloidal fraction predominantly generates chloroform and probably contributed for the formation of other disinfection by-products, including halonitromethanes and haloacetonitriles[1]. The HPOF fraction generates predominantly multibrominated THMs. In regards to the predominant formation of brominated THMs in the SJD DWTP, it would be beneficial to reduce their formation. For the optimal operational management, lower concentrations of brominated THMs can be achieved by reducing the amounts of HPOF and transphilic concentrations during the flocculation process in the DWTP while keeping temperature low.

3.2 Chemometrics modeling of UV spectral and physico-chemical data in finished drinking and wastewater water

This second block includes two articles with chemometrics models of spectral data and physico-chemical parameters of water with the objective to better monitor water quality in the distribution system of Barcelona and in the wastewater treatment plant nearby Girona (Catalonia).

Each raw water source is characterized by its particular natural organic matter composition (NOM). As a function of its content, NOM presents particular UVVIS spectral features or fingerprints, which are usually associated with its geographical origin. The studies in this block examined whether there is a possibility of monitoring the water quality using spectral data and chemometrics analysis.

The first paper develops models which should be able to predict different source apportionments in drinking water mixtures. These chemometrics models had the properties to predict up to five real drinking water types in mixtures by using their spectral profiles and physicochemical parameters.

[1] These by-products are beyond the scope of this work.

The second paper explores the relationships among the different wastewater treatment plant operational parameters, including spectral data, describing water quality. The tests were performed in laboratory conditions with synthetic water mixtures and in the wastewater treatment plant (WWTP) with real water samples.

The two papers comply with the main objectives of the Thesis in the following aspects:

- Development of chemometric regression models able to predict water source apportionments of water blends from up to five different water sources, using their UV spectral profiles and some other physicochemical parameters.
- Evaluation of chemometrics methods to improve online monitoring and control of wastewater treatment plant management using different techniques and routines for continuous water quality monitoring. Selection of a reduced number of UV spectral channels (wavelengths) for to monitor online WWTP operational processes.

3.2.1 Article 4 – Platikanov, S., Garcia, V., Landeros, E., Devesa, R., Matía, L., Tauler, R., *Determination of water supply sources in the Barcelona distribution system by UV spectrophotometry and PLS.* Water Science and Technology- Water Supply 11 (2011) 45-54.

Introduction

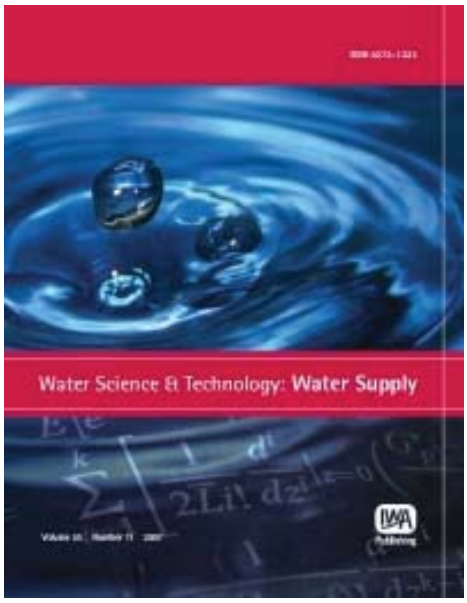
Barcelona metropolitan area has a large supplied drinking water distribution system (WDS) with five different origins, namely different treatment plants for finished drinking water, using raw water from three main water sources (two rivers, Llobregat and Ter, and the Mediterranean sea). Specific knowledge about water source apportionments at different WDS locations at any point in time would be highly useful for the operators in order to improve the global system management.

Because the three main raw water sources for Barcelona metropolitan area are the Ter and the Llobregat rivers (characterized by very different hydrological and biogeochemical processes) and the Mediterranean Sea water, it is generally expected

that there are different NOM levels and features with different physicochemical composition. In Article 4, NOM UV spectral profiles were used as a diagnostic marker (fingerprints) for the raw and finished water sources. When the information in the UV spectral profile was not sufficient to distinguish various closely related finished drinking waters, additional information from the physicochemical parameters was used. The proposed methodology is based on the application of chemometric analysis using partial least squares regression (PLS) on data from UV spectroscopic analysis and physico-chemical water quality analysis for the purpose to obtain the source apportionment of measured water mixture samples.

This feasibility study implemented the analysis of water blends prepared in laboratory in two different case studies. The first case study identified the main sources and their apportionment in mixtures of tap waters (finished drinking water) collected from four districts in the Barcelona area. Three out of four districts were supplied with drinking water mostly using the Ter River water source and the same water treatment plant. The fourth district was mostly supplied with water originating primarily from the Llobregat River. The second case study implemented a more complex design, where water blends were prepared using water samples from five different water treatment plants in the metropolitan area of Barcelona.

The two case studies had the following common characteristics. A Box-Behnken experimental design strategy was applied to obtain representative mixture samples from different sources in the concentration range of 0-100%. In both cases, UV spectra were measured in the range of 190–230 nm using a laboratory diode array spectrophotometer (Agillent Model 8453). A main difference between the studies was that water samples in the second study were measured in laboratory settings, which later were augmented with additional laboratory elemental analysis data. Hence, the feasibility of the apportionment of different water sources was tested combining UV spectra and physicochemical parameters, which were simultaneously processed (data fusion) by PLS regression method.



Determination of water supply sources in the Barcelona distribution system by UV spectrophotometry and PLS

S. Platikanov, V. Garcia, E. Landeros, R. Devesa, L. Matía and R. Tauler

ABSTRACT

A new method for the water source apportionment of the Barcelona (Spain) water distribution system is proposed. The method is based on the combined use of UV spectrophotometric measurements in the wavelength from 190–230 nm, and multivariate data analysis using the Partial Least Squares (PLS) chemometric method. From the differences in the organic matter content of the different water sources and of their corresponding UV spectral features, PLS was able to determine the relative amounts of the two main river water sources in samples of tap water from the different locations of the Barcelona city water distribution system. The extension of the method to determine the relative amounts in water blends, prepared from samples from five water treatment plant sources of the same city's distribution system, required the combined use of some other parameters. In particular, the distinction and apportion of the water coming from a desalinisation plant could be successfully achieved once concentrations of Boron were included in the analysis.

Key words | PLS, source apportionment, UV, water distribution system

S. Platikanov (corresponding author)
R. Tauler
Department of Environmental Chemistry,
IDAEA-CSIC,
Jordi Girona,
18-26, 08026 Barcelona,
Spain
E-mail: *Roma.Tauler@idaea.csic.es*

V. Garcia
R. Devesa
L. Matía
Aigües de Barcelona (Agbar),
Laboratory. General Batet,
5-7, 08028 Barcelona,
Spain

E. Landeros
CETaqua, Water Technology Center,
Passeig dels Tillers 3,
08034 Barcelona

INTRODUCTION

Quality control of drinking water is a major concern everywhere. Natural organic matter (NOM) is a characteristic feature of every type of water source. The presence of NOM in drinking water is of particular importance since it affects aesthetic water qualities such as taste, colour and odour (Spellman 2007). Every different type of NOM has particular features associated with its geographical origin, vegetation, soil, etc. (Schäfer 2001). For example, aquatic algae deliver dissolved organic compounds with high amounts of nitrogen and low contents of aromatic carbon and phenolic groups. On the other hand, terrestrial derived dissolved organic matter (DOC) has relatively low nitrogen content but a large amount of aromatic and phenolic groups. The contribution of each organic matter source is also seasonally dependent (Sharp *et al.* 2006), and the hydrological and biogeochemical processes involved can alter the chemical composition and physical structures of the NOM. However, NOM is a complex mixture of heterogeneous chemical compounds and it is not possible at present to fully describe its chemical structure in the natural environment (Croue 2004). NOM profiles can be used as a diagnostic marker or as a typical fingerprint based on its structure and features. The more frequently used analytical technique for the direct investigation of NOM in water systems is UV/VIS spectroscopy (Thomas & Burgess 2007).

The drinking water in the distribution network system (WDS) of Barcelona is distributed by the AGBAR (Aigües de Barcelona) group of companies. The water in the WDS originates from more than one water source and is processed in different treatment plants. The two main water sources are, at present, Ter and Llobregat river systems (see Figure 1), together with local groundwater sources. Additionally, a large water desalinisation plant has recently been built (Gueguen *et al.* 2008) to provide drinking water from the

doi: 10.2166/ws.2011.007

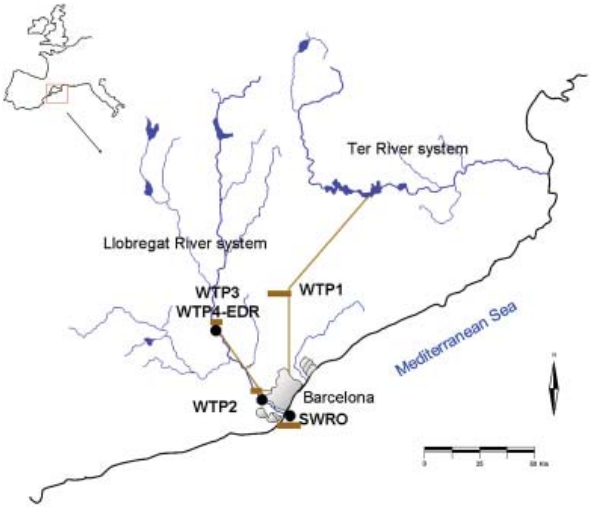


Figure 1 | Water distribution system and water treatment plants (WTP) in Barcelona geographical area. WTP1 is fed with water from Ter River, WTP2, WTP3 and WTP4-EDRare fed with water from Llobregat River and SWRO is fed with water from Mediterranean Sea near Barcelona.

Mediterranean Sea. These different raw water sources show great physicochemical differences in the water quality before and after treatment processes. Another possible cause of variation in the water quality is due to the different treatment processes implemented in the five treatment plants present in the area (see Figure 1). The plant disinfection processes comprise conventional disinfection procedures, electrodialysis reversal and seawater desalinisation procedures. Hereafter the plants will be encoded as:

WTP1: Cardedeu water treatment plant using conventional disinfection procedures to treat water from Ter River;

WTP2: Sant Joan Despí water treatment plant using conventional disinfection procedures to treat water from Llobregat River;

WTP3: Abrera water treatment plant using conventional disinfection procedures to treat water from Llobregat River;

WTP4-EDR: Abrera water treatment plant using electrodialysis reversal treatment of water from Llobregat River;

SWRO: El Prat reverse osmosis seawater plant treats water from the Mediterranean Sea, close to the Barcelona area.

See Figure 1 for a map of the different locations of these water plants. The finished water obtained at the exit of each treatment plant can vary significantly due to its NOM content (also because of its natural origin, Llobregat, Ter or Mediterranean Sea) and to its mineral/elemental content (also because of its different natural origin and/or because of the plant treatment procedures implemented).

The identification of the source of drinking water in any supply location of the water distribution system (WDS) is a challenge for the operational maintenance, repair and management of the system. Knowledge about water source apportionments (% of each water source in mixtures), at a specific location of the WDS, is of interest for a better understanding and detection of possible sources in leak accidents and in establishing legal property rights. Also, it is very important to know the sources of the waters and percentages of the blends in order to understand the origin of the problem and how to act when a quality problem appears in the distribution system.

A proper inspection has a significant impact on the operation and maintenance cost and on the effectiveness of the systems. A method for water source determination has been proposed based on the solid-phase microextraction GC/MS spectroscopic assay of the chlorination by-products (Dufresne *et al.*, http://www.thermo.com/eThermo/CMA/PDFs/Articles/articlesFile_12035.pdf.). It was associated with a usage of specific software performing pattern recognition on the MS profiles of different water sources. However, there is a need to develop inexpensive reliable methods for the *in situ* identification and determination of water sources and of their apportionment at each point of the WDS.

UV spectrophotometry is one of the most appealing approaches for broader analytical analyses. This spectrophotometric method is frequently used for the investigation of NOM content in drinking water. Usually organic matter content is measured at UV254 nm (USEPA Method 415.3), but also UV/VIS measurements can be expanded to the entire spectral range from 190–1100 nm. UV spectra of complex water blends will result in extensive spectral overlap due to the high number of absorbing components in this wavelength range in natural water samples. UV spectrophotometric measures can be coupled with advanced data analysis chemometric techniques to improve their resolution power. Chemometrics is at present a well established field in

chemical data analysis (Brown 2000) and has also recently been recognized in the analysis of water science data (Nollet 2007). There is a large number of chemometric techniques suitable for the study of complex multivariate water data sets, like principal component analysis (PCA) (Simeonova & Simeonov 2006) or PCA based regression methods (Wentzell & Lohnes 1999; Narasimhan & Shah 2008) which were shown to achieve very good predictive abilities in spectral regression problems, and partial least squares regression (PLS) (Platikanov *et al.* 2007) and several other methods not well known outside chemometrics that may play an important role. In summary, the main goal of this study was to establish PLS models for water source apportionment in blends on the basis of their NOM spectral profiles and physicochemical parameters.

METHODS

Methodology

Multivariate regression techniques reveal the relationship between two different data blocks (matrices) of chemical data. The main goal of this study is to build multivariate linear models able to describe, explain and predict the apportionment of the different water sources inside the Barcelona water distribution system and in house tap water blends (**y** block of variables) obtained from them as a function of their UV spectra together with other possible physicochemical parameters like the concentrations of the chemical elements present in the water samples (**X** block of variables). This involves finding an adequate mathematical model or function that relates the variables in these two data blocks, i.e. $\mathbf{y} = f(\mathbf{X})$, where the **y** block gives the apportionment of each source in each water blend and the X block gives the UV spectra and some additional physicochemical parameter, like the elemental concentrations of the blends. This modeling follows a two-step approach. The first step is the model calibration. The calibration step is followed by a second prediction step in which this model is used to estimate unknown apportions of blend samples from their UV spectra. Once experimental data were properly arranged in data matrices, they were mean centered (considering only UV spectra as predictors) or autoscaled (column mean centering and scaling of UV spectra and one additional physicochemical parameter at a time). Group scaling could have been considered as an alternative pre-processing technique. This would be appropriate when the full wavelength spectra and only one or two element concentrations were used as predictors. In this study however, only wavelengths where the absorption was high (between 190–230 nm) were selected for data analysis and noisy wavelengths were excluded. This assured reducing of the possible noise amplification in the scaling operation. This preliminary data treatment eliminated offsets, changes in measurement units and focused the analysis on proper modeling of observed variances in measured variables. This data pretreatment is frequently used in multivariate data analysis (Massart *et al.* 1998).

In particular, the Partial Least Squares (Geladi & Kowalski 1986; Massart *et al.* 1998) multivariate linear regression (PLS) method has been used and evaluated for the modeling of the water sources apportionment in their blends. PLS attempts to maximize the covariance between **X** and **y** data blocks. PLS searches for a common factor subspace most congruent to both data blocks, and its predictions are usually better than using other multilinear regression methods such as the ordinary multilinear regression method (MLR), especially if the **X** variables are highly correlated, like in the spectroscopic measurements used in this work. PLS transforms the high number of original variables (spectra) into a smaller number of orthogonal variables called "components", "factors" or "latent variables", which are linear combinations of the original variables. The first latent variables contain useful information about the major covariance sources between the two blocks of variables, whereas the last ones basically represent the uncorrelated variance and noise, which has to be discarded and is not considered in the modeling. A new matrix of weights (reflecting the covariance structure between the **X** predictors and **y** response variables) is obtained from PLS analysis, which provides rich information about the nature of the different covariance sources. The selection of the optimal number of components (the number of latent variables) in PLS is performed using internal cross validation (leaving out one sample at a time) and external validation for optimal prediction of **y** values in new samples not used in the calibration step. In this work, for method/model validation, an external set of samples (water

blends with known source apportions), not participating in the model calibration, was used.

For brevity, in this study the relative errors of water source apportionment in percentage are reported for both calibration and prediction steps.

Relative errors of concentrations in percentage, for both calibration and prediction steps, are calculated as follows:

$$Rel.\ error\ in\ \% = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{\sum_{i=1}^{n}(y_i)^2}} \times 100;$$

Preparation of water samples, instrumentation, chemical analysis and software

This work consisted of two studies. The first one was intended to identify the sources/apportionment of tap water collected from four of Barcelona's districts. UV spectra were used as a data set of water blends. Three of these city districts, Fondo, Gracia and Horta (D1–D3), were supplied with drinking water treated in WTP1 (see Figure 1), representing the original Ter River water source. The fourth district, Les Corts (D4), is supplied with water, treated in WTP2 and WTP3, originating from Llobregat River water. WTP4-EDR and SWRO plants were not involved in this first study.

In a second study, water samples were provided directly from the 5 water treatment plants. The feasibility of the apportionment of the five water sources was studied from blends of these water samples, also using their UV spectra, elemental analysis and chemometrics.

In both cases, water blends were prepared according to a three-level experimental design based on a Box-Behnken design. In the first experiment, 24 samples were prepared according to a Box-Behnken statistical experimental design (Box & Behnken 1960) and after their random sampling two new data subsets were selected, one for calibration and the other for validation. The calibration data set comprised 18 samples, and the remaining 6 samples were selected for external validation. In the second experiment, new water blends were prepared using a similar experimental design to the one just described but prepared using 22 water blends with 16 of them used for the calibration step and the other 6 used for the external model validation.

Apportioning of water sources in the water blends were done in the range of 0–100% with predominance in the middle range values such as 11, 20 and 50% of each water source in the blend. In this study apportionments were never predicted below 0 or over 100%. If this was the case, an alternative regression method like PLS logistic regression (Bastien *et al.* 2005) could have been used.

UV spectra were recorded in the range of 190–1100 nm in a 1 cm quartz cell, using an Agilent HP8453 diode array spectrophotometer. Wavelengths range used was between 190–230 nm.

In the first experiment two replications were performed during a month. Both times, water samples were collected during the morning and after allowing the water to run out from the tap for some minutes.

In the second experiment water samples were prepared externally by the technician staff working in the Barcelona WDS. They were delivered to the laboratory the same day of production and they were appropriately mixed. A preliminary study of water samples' evolution over time did not show any significant changes in its spectrophotometric properties within a week. This was expected, since water remains in the distribution systems for one week or less and, therefore, the experimental conditions used in this work can be considered appropriate for emulating what occurs in the Barcelona distribution system.

Aluminium, Barium, Boron, Calcium, Copper, Chromium, Strontium, Iron, Magnesium, Manganese, Nickel, Potassium, Silicon, Sodium and Zinc concentrations were determined by Inductively Coupled Plasma Optical Emission Spectrometry, ICP-OES, (Perkin Elmer Optima 4300 DV). Bicarbonate was analyzed by a robotic titrosampler with conductivity module 855 and 856. Chloride, Nitrate and Sulfate concentrations were estimated by Ionic Chromatography (Dionex ICS-2000). All these analytical determinations (except bicarbonates, which are not included in EC Drinking Water Directive 98/83/EC), were ISO17025 accredited.

The variables in $\mathbf{X}$ and $\mathbf{y_i}$ data sets were initially arranged using EXCEL (Microsoft, Redmont, WA, USA) and subsequently transferred to the MATLAB computer workspace environment (MATLAB version 6.5, The Mathworks, Natick

MA, USA). Chemometrics modeling was performed using PLS Toolbox 4.2 (Eigenvector Research, Manson, WA, USA). Statistical experimental design was carried out by using the software Unscrambler 9.8 (CAMO PROCESS AS, Oslo, Norway).

RESULTS AND DISCUSSION

UV spectrophotometric analysis of tap water samples

Before multivariate modeling of water blend spectra, the analysis of the spectra of different Barcelona district water samples led to the preliminary conclusion that D1–D3 district waters had practically the same UV spectra profile and that they had the same original source of water (from Ter River) as a result of having a similar organic matter content. Figure 2 shows the spectral difference between the two main sources of water for Barcelona, i.e. Ter River (D1 blue lines) and Llobregat River (D4 red lines). The sample spectra were obtained from two district water locations, each one replicated after 12 hours (solid and dashed lines). The replicates in the plot cannot be clearly distinguished, because they were highly overlapped due to the lack of variation in their chemical content for this period. The UV spectra of the tap water from the D4 district were clearly different from the UV spectra of the other tap water samples, leading to the conclusion that the D4 water sample was the only one coming from Llobregat River. This fact was confirmed by the technical staff of the water distribution company.

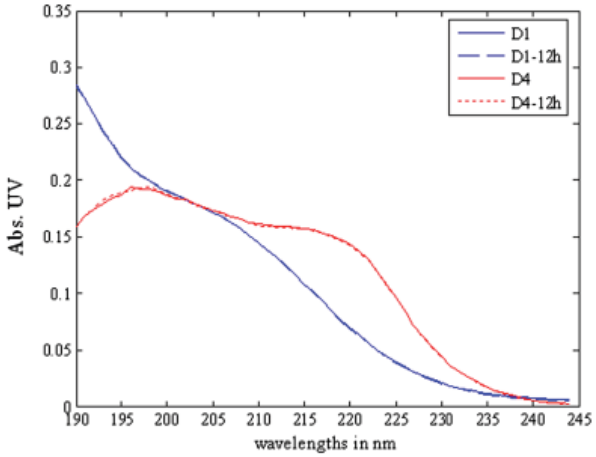


Figure 2 | UV normalized spectra of two tap water samples from two different districts of Barcelona, representative of the two different water sources: Ter River (D1 blue solid and dashed lines) and Llobregat River (D4 red solid and dashed lines). Replicated sample spectra (dashed lines) were recorded after 12 hours.

This preliminary study led to the conclusion that the two main water sources (Ter and Llobregat water river sources) could be simply distinguished specrophotometrically by means of their UV spectra. Consequently, the apportionments of the different district water sources from D1 to D3 were summed up and recalculated as a single Ter River source against the amount of D4 (Llobregat River sample). PLS modeling was then performed using a calibration subset of 18 blends. In the case of the modeling of the Llobregat River water apportionment, two latent variables were selected (after cross-validation) to explain almost 99% of the variance/information in the data and leading to very low prediction errors (6% prediction error in the calibration). This calibration model was then used to predict the Llobregat River source content of the external subset of the 6 blends selected for validation, resulting in only 7% prediction errors. Similarly, low prediction errors were also achieved for Ter water source apportionment in the blends by PLS modeling. These low errors obtained in the validation were achieved because of the reproducible intrinsic patterns in the organic matter of the different sampling locations in the distribution system. The model that was based on the 2 latent variables was reliable enough and offered good prediction properties.

To summarize, the separation between the two main river sources was possible in this first study due to their UV absorption spectra and because of their distinct organic matter patterns, which differ considerably between the two river water sources. The next, more difficult, step was to try to distinguish among the five water treatment plant sources (see Figure 1) and the relative apportionment of their blends.

UV spectrophotometric analysis plus elemental and mineral content analysis of AGBAR water blend samples

Like in the first experiment, UV spectra of the five different water source samples were first registered. Figure 3 shows the normalized spectra of the five water sources provided by the company. Very large similarities among some of the spectra

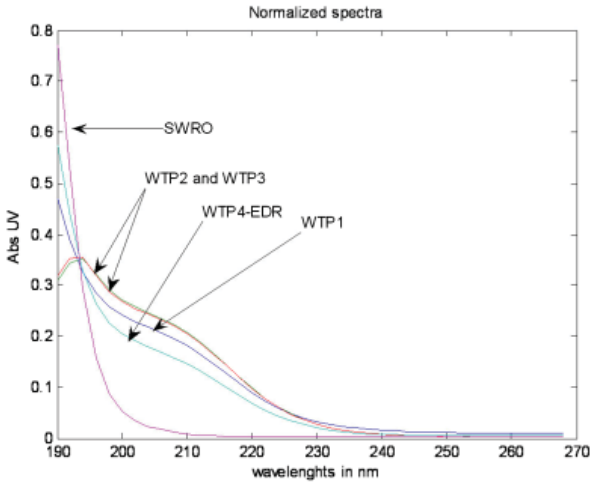


Figure 3 | UV normalized spectra of five samples of water from the five different water treatment plants, which are the main sources of the Barcelona water distribution system.

can be noticed, especially between the spectra of WTP2 and WTP3 water sources.

The highest pairwise correlation coefficient was found to be of 0.999 between WTP2 and WTP3 water sources. These two water sources come from Llobregat River and the two treatment plants are employing similar conventional water treatment procedures. Therefore, samples from these two water sources are expected to have very similar organic matter content and give analogous UV absorption spectra. On the contrary, the sample from WTP4-EDR water source (a plant which implements reverse electrodialysis filtering as water treatment) gave an absorption spectrum rather different to the previous two absorption spectra mentioned, despite their common origin from the Llobregat River. This means that the reverse electrodialysis process had changed significantly the organic matter content of the Llobregat water, leading to water with a different UV spectrum. Moreover, the WTP4-EDR water spectrum showed a similar spectrum with a high pairwise correlation coefficient 0.98 to the WTP1 river source spectrum. On the other hand, SWRO water (from the desalinization plant) spectrum did not show any characteristic spectral feature in the UV region. It only absorbed UV light in a very narrow range from 190–205 nm, confirming that this type of water had very low organic matter content and that modeling this source on the sole basis of UV absorption will probably not be possible. Although the pairwise correlation coefficients of its spectrum with the spectra of the other water sources was very low, ranging from 0.6 to 0.88 (because it had nearly no absorption, whereas the others did), it only had a dilution effect; moreover, it was very difficult to distinguish it from the others and the prediction of its apportionment on the sole basis of its UV absorption was not possible.

Therefore, other physicochemical properties, such as conductivity and mineral content, were considered to describe the five water sources and their relative content. The results obtained from this mineral content analysis, together with the spectrophotometric data, were arranged in the same table data matrix (independent variables **X**) to be subjected to PLS modeling. Elemental and mineral analysis comprised 21 parameters routinely monitored by the water company quality control laboratory. Table 2 gives the results of these analytical determinations and their basic statistics.

Some of the parameters analyzed did not show any significant feature for the characterization of the water blend samples analyzed and were therefore not useful for the purpose of this study. The parameters showing significant changes in their values were incorporated one by one into the **X** data table, together with the spectral measurements, and were then submitted to new PLS modeling. It was obvious that WTP2 water showed many of the maximum parameter changes, which was in contrast to the SWRO water (from the desalinization plant), which showed minimum changes in most of its parameter values. Finally, only the boron concentration was found to be at its maximum concentration in SWRO water samples. This can be explained by its original sea water source and the low effect the water treatment had on this parameter.

Table 1 | Pairwise correlation coefficients among the 5 AGBAR water treatment plant spectra

	WTP1	WTP2	WTP3	WTP4-EDR	SWRO
WTP1	1	0.97	0.97	0.98	0.77
WTP2		1	0.999	0.9	0.6
WTP3			1	0.91	0.61
WTP4-EDR				1	0.88
SWRO					1

Table 2 | Concentration values and parameters analyzed in water samples from the 5 water treatment plants operating in the barcelona water distribution system

	Conductivity 20°C	Bicarbonate	Chloride	Sulfate	Nitrate	Sodium	Potassium	Calcium	Magnesium
WTP1	382	145[a]	38[a]	50.6[a]	4.15[a]	15.4[a]	5[a]	49.5[a]	9.1[a]
WTP2	1194	233	214	137	11.5	111	19.8	102.3	25.3
WTP3	1100	217	189	132	10.7	96.9	18.7	99.4	23.1
WTP4-EDR	365	101	55	24.4	2.71	49.2	6.5	17.3	4.8
SWRO	311	8.1	97	<5[c]	<0.5	54.5	<5	<5	1.1
Min	311	8.1	38	<5	<0.5	15.4	<5	<5	1.1
Max	1194	233	214	137	11.5	111	19.8	102.3	25.3
Average	670.4	140.82	118.6	69.8	5.912	65.4	11	54.7	12.68
	Boron	**Strontium**	**Iron**	**Manganese**	**Nickel**	**Barium**	**Silicon**	**Aluminium**	
WTP1	27[b]	0.45[b]	111[b]	6[b]	11[b]	31[b]	1.7[b]	34[b]	
WTP2	82	1.48	5	<1	6	55	2.19	<24	
WTP3	45	1.37	7	<1	<5	53	2.15	<24	
WTP4-EDR	34	0.24	<5	<1	6	10	1.92	<24	
SWRO	742	0.05	<5	<1	<5	9	0.5	<24	
Min	27	0.05	<5	<1	<5	9	0.5	<24	
Max	742	1.48	111	6	11	55	2.19	34	
Average	186	0.718	26.6	2	6.6	31.6	1.692	26	

[a]values recorded in mg/L
[b]values recorded in μg/L
[c]values below the limit of quantitation

Figure 4 depicts the prediction errors finally obtained for calibration and validation when spectral data and one of the additional physicochemical parameters at a time were considered. PLS analysis of UV spectra, together with some of the physicochemical parameters considered, gave satisfactory predictions. In particular, the best parameters proved to be Barium (Ba) and Boron (B). Prediction errors for the external validation water samples were from 10–12%. Results obtained in the modeling of the WTP4-EDR samples showed that Boron was very important to correctly predicting the apportionment of this source, giving just a 5% error in its prediction for the validation samples. Very good results were also found in the modeling and prediction of the desalinization plant (SWRO) water source, showing again that the concentration of Boron was the most important parameter to consider (prediction errors below 10%). However, less satisfactory results were achieved for the differentiation of WTP2 and WTP3 sources, both in the calibration and in the validation step.

Figure 5 shows the results obtained for the prediction of WTP2 and WTP3 water sources when they were considered individually, as separate sources, and jointly, as coming from the same source, e.g. Llobregat River. When they were considered separately, the predictions of their relative contents were wrong; the results yielded high prediction errors both in calibration and in validation, regardless of the physicochemical parameter added to the spectral data. In contrast, when WTP2 and WTP3 were considered jointly, i.e., when they were considered from the same source, low prediction errors were then achieved. Prediction errors for the external validation samples were usually below 10%. Again, Boron was the most important parameter included in the model.

Table 3 summarizes one example of the results obtained in the complete modeling of the whole experimental data set. Boron was confirmed as the most important parameter in the data analysis, together with UV spectral data, for water source apportionment. Also, a possible good combination of parameters to be considered together with spectral data was Boron and conductivity. The addition of conductivity increased the reliability and robustness of the PLS models and made them more versatile for the different conditions in the Barcelona water distribution system.

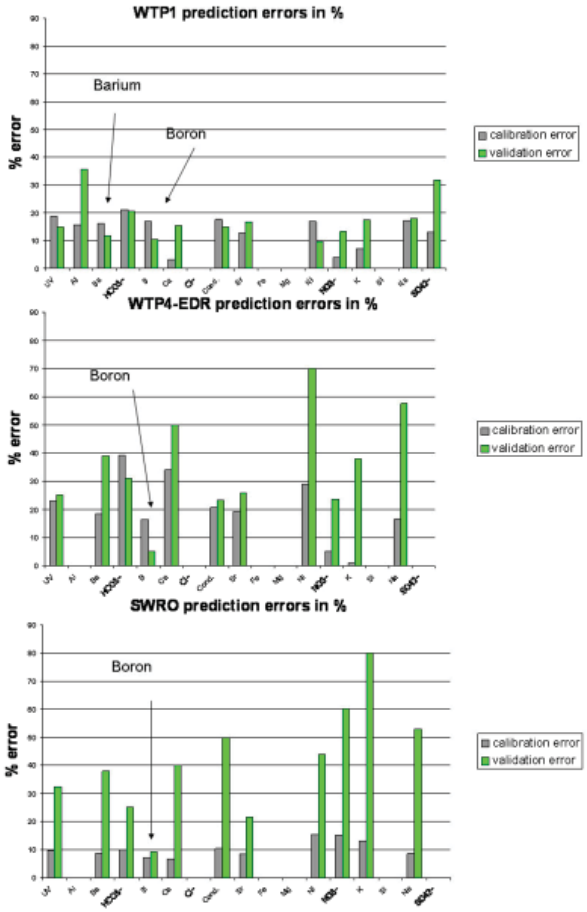


Figure 4 | Summary of PLS prediction errors for calibration and external validation samples using spectral data plus one additional physicochemical parameter at a time for the prediction of WTP1, WTP4-EDR and SWRO water source apportionments. The lowest errors and the corresponding parameters are highlighted with arrows.

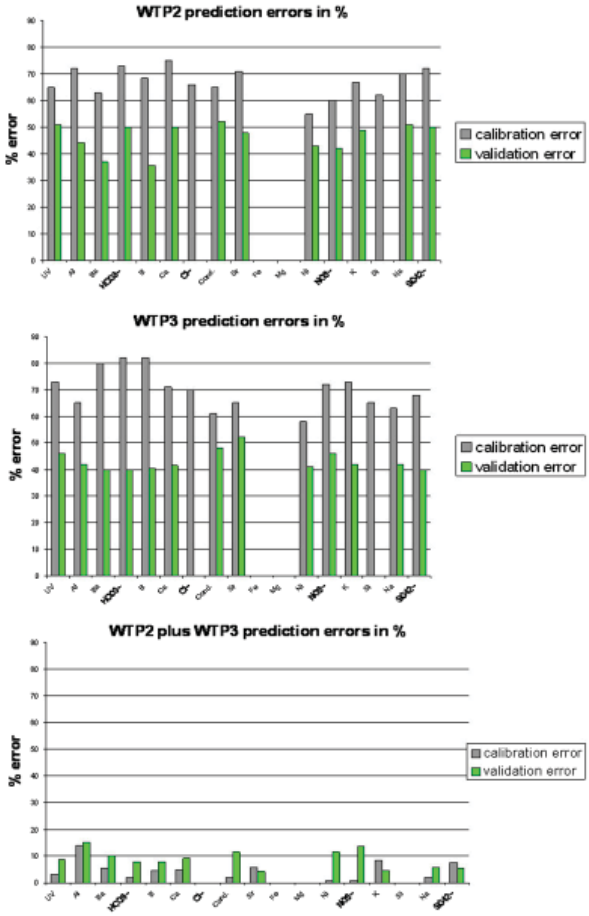


Figure 5 | Summary of PLS prediction errors for calibration and external validation samples using spectral data plus one additional physicochemical parameter at a time for the prediction of WTP2, WTP3, and WTP2 + WTP3 water sources.

Special attention has been paid to the prediction of SWRO apportionments. This was shown not to be possible using only UV spectra. Including boron and conductivity greatly improved the prediction of SWRO water content for the water blends investigated. Indeed, boron was found at its highest concentrations in SWRO water source samples. An example of SWRO modeling when UV spectra, Boron and conductivity were simultaneously considered as parameters is given in detail in Figure 6. Five latent variables were found to explain more than 99% of the **y** block variance, assuring as low prediction errors as 5.3% for the calibration samples and 10.9% for the external validation samples. Graphically, these predictions were compared to the actual concentrations in Figure 6.

Future research should be focused on the development of an experimental system for on-line monitoring of the water distribution system based on the combined use of UV, some rapid elemental analysis for boron, and chemometrics.

CONCLUSIONS

The following were the main conclusions derived from the present work:

- UV spectrophotometric analysis coupled with chemometrics has been shown to be a powerful tool for the differentiation of different raw water sources having

Table 3 | Summary of PLSR prediction errors in % for the determination of the apportionment of the five water treatment plant sources (data block y) when UV (alone), UV plus Boron concentration and UV plus Boron and conductivity were added to the spectral data (data block X)

Calibration	UVspectra	UV + Boron	UV + Boron + Cond
WTP1	18.9	17	16.5
WTP2	56	68.5	66
WTP3	62	82	81
WTP4-EDR	23.2	16.4	11
SWRO	9.7	7.2	5.3
WTP2 + WTP3	3.2	4.6	4.6
Validation	**UVspectra**	**UV + Boron**	**UV + Boron + Cond**
WTP1	14.9	10.6	9.1
WTP2	39	35.6	50
WTP3	44	40.5	41
WTP4-EDR	25.1	5.2	4.9
SWRO	32.5	9.1	10.9
WTP2 + WTP3	8.5	7.8	6.8

different natural organic matter contents. In the case of tap water from the Barcelona distribution system, it was possible to distinguish between water sources from two different rivers (Ter and Llobregat) in their blends and to estimate their relative apportionments with prediction errors of around 7%.

- The compositions of water blends from 5 different Barcelona water treatment plants were not predicted well on the basis of UV analysis only, due to overlapping spectra and a lack of natural organic matter content in some of these water sources. Especially difficult when predicting the composition of water blends was the estimation of the individual water apportionments from electrodialysis and desalinization plants. In contrast, the clear differences among the nature and amounts of organic matter in Llobregat and Ter rivers allowed successful prediction of their apportionments.
- The addition of information from elemental/mineral analysis of the different water sources, apart from their UV spectra, allowed successful prediction of all the water sources in the distribution system.
- Boron concentration was the most important parameter, apart from UV spectra, for the correct prediction of water sources from the desalinization plant. The prediction errors were below 10% for the external validation samples. Therefore, Boron can be considered as a very important variable in cases where a desalination plant is part of the distribution system.
- Two conventional water treatment plants, situated in the same river system and located near each other, were impossible to differentiate either on the basis of UV spectra profile analysis or their mineral content. However, their joint contribution could be estimated as a sum of apportionments, i.e. as coming from the same source.

ACKNOWLEDGEMENTS

The methodology described in this article has been developed within the research project QUALITAX, leaded by CETaqua Water Technologic Centre with the financial support of

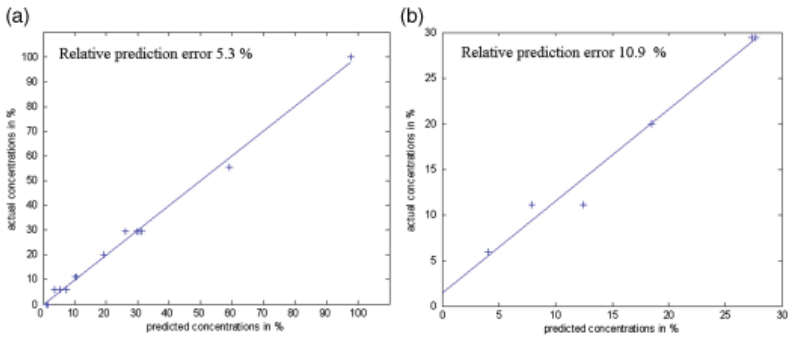


Figure 6 | Prediction of SWRO water source apportionments: (a) PLS predicted versus actual calibration SWRO water source apportionment values; (b) PLS predicted versus actual external validation SWRO source apportionment value.

AGBAR. The project has been carried out by Aigües de Barcelona (AGBAR Group) and the Department of Environmental Chemistry, IDAEA-CSIC.

REFERENCES

Bastien, P., Vinzi, V. & Tenenhaus, M. 2005 PLS generalized linear regression. *Comput. Statist. Data Anal.* **48**, 17–46.

Box, G. & Behnken, D. 1960 Some new three level designs for the study of quantitative variables. *Technometrics* **2**, 455–475.

Brown, S. D. 2000 Chemometrics. In *Encyclopedia of Analytical Chemistry*, R. A. Myers, Ed., Wiley Interscience, 9669–9671.

Croue, J. P. 2004 Isolation of humic and non-humic nom fractions: structural characterization. *Environ. Monit. Assess.* **92**, 193–207.

Dufresne, C., Blount, B., Harrison, S. J., Bukowski, N., Lin, L. & Blackburn, M. *Automated sourcing of potable waters using SPME GC/MS coupled with on-line chromatographic pattern matching.* http://www.thermo.com/eThermo/CMA/PDFs/Articles/articlesFile_12035.pdf.

Geladi, P. & Kowalski, B. R. 1986 Partial least-squares regression: a tutorial. *Anal. Chim. Acta* **185**, 1–17.

Gueguen, F., Sanz, A., Langlais, C., Bonnélye, V., Créme, G., Del Campo, I. & Beltran, F. 2008 Barcelona Desalination Plant: 200.000 m3/day of Drinking Water. Proceedings of the IWA World Water Congress and Exhibition. Vienna (Austria).

Massart, D. L., Vandeginste, B. G. M., Buydens, L. M. C., de Jong, S., Lewi, P. J. & Smeyers Verbeke, J. 1998 *Handbook of Chemometrics and Qualimetrics*. Elsevier Science, Amsterdam.

Narasimhan, S. & Shah, S. 2008 Model identification and error covariance matrix estimation from noisy data using PCA. *Control Eng. Pract.* **16**, 146–155.

Nollet, L. 2007 *Handbook of water analysis*. CRC Press, Taylor and Francis.

Platikanov, S., Puig, X., Martin, J. & Tauler, R. 2007 Chemometric modeling and prediction of trihalomethane formation in Barcelona's water works plant. *Water Research* **41**, 3394–3406.

Schäfer, A. 2001 *Natural Organic Matter Removal using Membranes: Principles, Performance and Cost*. Technomic, Lancaster.

Sharp, E., Parsons, S. & Jefferson, B. 2006 Seasonal variations in natural organic matter and its impact on coagulation in water treatment. *Sci. Total Environ.* **363**, 183–194.

Simeonova, P. & Simeonov, V. 2006 Chemometrics to evaluate the quality of water sources for human consumption. *Microchimica Acta* **156**, 315–320.

Spellman, F. 2007 *The science of water. Concepts and Applications*. CRC Press, Taylor and Francis.

Thomas, O. & Burgess, C. 2007 *UV-visible Spectrophotometry of Water and Wastewater*, Vol. 27. In: Techniques and Instrumentation in Analytical Chemistry series, Elsevier, Amsterdam.

USEPA Method 415.3 2005 Determination of Total Organic Carbon and Specific UV Absorbance at 254nm in Source Water and Drinking Water.

Wentzell, P. & Lohnes, M. 1999 Maximum likelihood principal component analysis with correlated measurement errors: theoretical and practical considerations. *Chem. Intell. Lab. Sys.* **45**, 65–85.

Results and Discussion

- UV spectrophotometric analysis and chemometrics for source apportionment of tap water samples

The spectral analysis of the four Barcelona district water samples confirmed that three of districts waters had practically the equivalent UV spectra profile. These three districts were found to share water from the same original river source, the Ter River. UV spectra of tap water from the fourth district were found to be distinctive in their shapes (see Figure 2, Article 4) and the location of the district evidenced their different source, namely the Llobregat River.

Therefore, two main water sources of river water (Ter and Llobregat rivers) were distinguished using their distinctive UV spectral features that reflect particular organic matter content. To properly perform the source apportionment study of these two river origins from their blends, water apportions from the three districts in the experimental design were considered together. Therefore, in the composition of these synthetic mixtures, Ter River water source was combined with Llobregat River water source. PLS modelling was performed using a calibration subset of 18 blends, which was followed by validation using 6 new (external) blends. To model Llobregat River water apportionment with PLS, two latent variables were found to explain approximately 99% of the variance (information), having only 7% of prediction error in the external samples. Ter water source apportionment was also predicted in blends by using PLS. The main conclusion of this preliminary study was that the two main river sources were possible to be distinguished in blends using their UV spectral profile as a consequence of their distinct organic matter patterns and of the appropriate chemometrics data analysis.

This preliminary study suggested that there is a possibility to distinguish and perform source apportionment studies of water sources from the different water treatment plants in the Barcelona area, which use different water treatment procedures for river water and sea water.

- *UV spectrophotometric and elemental analyses and chemometrics for source apportionment of mixtures of water samples from different origins*

UV spectra of the five different water sources were compared after normalization (see Figure 3, Article 4). WTP2 and WTP3 water sources gave similar UV spectra and had a correlation coefficient close to 1 (0.999). This result indicates that both sources had very similar organic matter content. The origin of the two water sources was the Llobregat River, which is treated with conventional procedures in two different plants. In contrast, the use of an electrodialysis reversal filtering system significantly changed the spectral profile of the corresponding water source samples (WTP4-EDR), despite of their common water origin (i.e., both from the Llobregat River). The reverse-osmotized water from the sea water desalinization plant (SWRO) had significantly lower organic matter content, explaining why its UV spectrum has no specific absorption band. This water source will produce a dilution effect and the prediction of its contribution should be difficult if no additional information is given. For this reason, information regarding the mineral content of the five water sources was incorporated in this study.

The mineral content analysis, along with the spectrophotometric data, were organized in a table (data matrix) and later employed in the PLS modelling of water mixtures from the five water sources. Table 2 (Article 4) gives the results of the analytical determinations and the basic statistics regarding 21 parameters, which were routinely monitored by the water company quality control laboratory. Some of these parameters did not vary whatever water blend samples were, and therefore they were discarded for further analysis. The parameters with significant variation were incorporated into the data table on an individual basis (one-by-one), along with spectral measurements, and later submitted to an updated PLS modelling.

Descriptive statistics of such parameters was thoroughly examined. It suggested that conventionally treated water (WTP2) from the Llobregat river was the one that changed more (i.e., the maximum parameter changes), in contrast to desalinated sea water (from the desalinization plant), showing minimum changes in most parameter values. Boron ion concentration was found to be a characteristic feature of desalinated SWRO water sample.

PLS analysis of UV spectra data, along with some of the considered physicochemical parameters data, allowed us to obtain relatively accurate predictions of water source apportionments (see Figure 4, Article 4). More specifically, the best mineral parameters to complement spectral data were found to be Barium (Ba) and Boron (B) concentrations. Using them, prediction errors of 10-12% for the external validation water samples were obtained. Boron was found to be a very important element for the apportionments of reverse-osmotized sea water and electrodialysis reversal filtered river water (SWRO and WTP4-EDR), with lower than 10% errors in prediction for external validation samples.

It was impossible to apportion individually the water sources from the two conventionally treated water samples from Llobregat River (WTP2 and WTP3 sources). The reason was that there was not enough spectral (organic matter) and mineral composition differences between them to be successfully modelled using PLS. Therefore, these two type of water samples were classified as belonging to the same source in their mixture blends and hence aggregated before PLS modelling.

When WTP2 and WTP3 water samples were attributed to the same water source, low prediction errors (below 10%) were obtained for their prediction in external validation. Similarly to other samples, boron concentration was found to be a relevant parameter for the PLS model when water samples from SWRO and WTP4-EDR were included.

Taken together, these results suggested that boron was a very important parameter to be included in the data analysis, along with UV spectral data, for water sources apportionment. Boron concentrations together with conductivity provided an excellent parameters combination for the optimal PLS modelling. It was found that the addition of conductivity parameter increased the reliability and robustness of the PLS models (especially for the desalinated sea water, SWRO). Furthermore, it made them more versatile for the different conditions which can be encountered in the Barcelona water distribution system. The above findings are presented in Figure 4, Figure 5 and Table 5 of Article 4.

In summary, UV spectrophotometric analysis coupled with chemometrics is a powerful strategy for differentiating raw water sources with different natural organic matter content. In the case study of the Barcelona distribution system, it was possible to distinguish and perform accurate source apportionments of up to five different water sources in blends. Differentiation in blend compositions from different water sources

was made possible when additional information about certain elements was introduced to the UV spectral data. Apart from UV spectra, boron concentration and conductivity were the most important parameters for the correct prediction of water sources.

3.2.2 Article 5 – Platikanov, S., Rodriguez-Mozaz, S., Huerta, B., Barcelo, D., Cros, J., Batlle, M., Poch, G., Tauler, R. *Chemometrics quality assessment of wastewater treatment plant effluents using physicochemical parameters and UV absorption measurements.* Journal of Environmental management 140 (2014) 33-44.

Introduction

Traditional laboratory methods exist for the characterization of wastewater quality parameters such as total carbon (TC), inorganic carbon (IC), total organic carbon (TOC), non-purgeable organic carbon (NPOC), pH, alkalinity, conductivity, chloride, sulfate, nitrate and fluoride, and some toxic pollutants. Although such methods are accurate and precise, they may be time-consuming and expensive when analysing large sample sets for surveying, routine monitoring and similar purposes.

UV-Visible spectrophotometry (UVVIS) is highly useful for on-line and in-line measurements in wastewater quality monitoring. UVVIS monitoring can significantly improve the correct operation of the water treatment systems. To the extent that many organic compounds and a few soluble minerals (such as nitrates) absorb in the UV region, this fast and simple method is able to follow wastewater quality based on organic matrix fingerprinting.

Multivariate calibration methods, such as principal component analysis (PCA) and partial least squares (PLS), give further advantage to UVVIS in successfully analyzing overlapped spectra of multiple compounds. A significant advantage of applying multivariate calibration methods is that they simplify the sample preparation by avoiding preliminary separation steps in complex sample matrices.

This article includes two case studies, namely: a) exploration of the relationship among different parameters in monitoring wastewater quality, and b) modelling (prediction) of specific target water quality parameters. In both studies, chemometric approaches were applied on matrices containing concentration data for physicochemical parameters and UVVIS spectral data.

In the first case study, we investigate the temporal/seasonal variability of physicochemical water quality parameters and spectral data in data sets generated at a real WWTP. Different PCA models were calculated and compared using data sets obtained using conventional laboratory methods or instrumental systems for online water quality monitoring. These datasets have been collected in different time intervals during monitoring campaigns. Specific to this study is the comparison and evaluation of the results obtained following the application of a recently developed experimental sensor system, which enable the simultaneous recoding of physicochemical parameters and selected UV bands. A comparison between PCA results obtained using this new proposed experimental system and the data generated by conventional laboratory standard methods was performed.

In the second case study, we investigate the possibility to derive a chemometric prediction of concentration of four target water quality parameters in experimentally designed synthetic mixtures. Target parameters for prediction were phenol, nitrate, dodecylbezosylphonate and DOM. Different concentrations from the selected target parameters were mixed in laboratory according to a Box-Behnken experimental design. Next, a predictive PLS2 model was developed for the simultaneous prediction of such parameters, modelling the obtained UVVIS spectral data in the mixtures. More specifically, a methodology for the selection of the most important spectral wavelengths in PLS models is proposed. Using the variable-importance-in-projection (VIP) scores obtained from PLS modelling parameters, full UVVIS spectra were reduced to few spectral bands and, a new PLS model was recalculated. VIP scores technique was also applied to actual WWTP data to predict nitrate concentrations and Total Organic Carbon (TOC) in real samples. Predicted values of these two parameters were also compared to the corresponding reference values that were obtained by standard laboratory methods with satisfactory results.

Journal of Environmental Management 140 (2014) 33–44

Contents lists available at ScienceDirect

Journal of Environmental Management

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

Chemometrics quality assessment of wastewater treatment plant effluents using physicochemical parameters and UV absorption measurements

S. Platikanov [a], S. Rodriguez-Mozaz [b], B. Huerta [b], D. Barceló [a,b], J. Cros [c], M. Batle [c], G. Poch [d], R. Tauler [a,*]

[a] IDAEA-CSIC, Jordi Girona 18-26, 08034 Barcelona, Spain
[b] Catalan Institute for Water Research (ICRA) H2O Building, Scientific and Technological Park of the University of Girona, Emili Grahit 101, E-17003 Girona, Spain
[c] Adasa Sistemas S.A.U., Barcelona, Spain
[d] TRARGISA, Girona, Spain

ARTICLE INFO

Article history:
Received 10 December 2013
Received in revised form
15 March 2014
Accepted 17 March 2014
Available online

Keywords:
Chemometrics
PCA
PLS
Wastewater
Water quality
UVVIS

ABSTRACT

Chemometric techniques like Principal Component Analysis (PCA) and Partial Least Squares Regression (PLS) are used to explore, analyze and model relationships among different water quality parameters in wastewater treatment plants (WWTP). Different data sets generated by laboratory analysis and by an automatic multi-parametric monitoring system with a new designed optical device have been investigated for temporal variations on water quality parameters measured in the water influent and effluent of a WWTP over different time scales. The obtained results allowed the discovery of the more important relationships among the monitored parameters and of their cyclic dependence on time (daily, monthly and annual cycles) and on different plant management procedures.

This study intended also the modeling and prediction of concentrations of several water components and parameters, especially relevant for water quality assessment, such as Dissolved Organic Matter (DOM), Total Organic Carbon (TOC) nitrate, detergent, and phenol concentrations. PLS models were built to correlate target concentrations of these constituents with UV spectra measured in samples collected at (1) laboratory conditions (in synthetic water mixtures); and at (2) WWTP conditions (in real water samples from the plant). Using synthetic water mixtures, specific wavelengths were selected with the aim to establish simple and reliable prediction models, which gave good relative predictions with errors of around 3–4% for nitrates, detergent and phenols concentrations and of around 15% for the DOM in external validation. In the case of nitrate and TOC concentrations modeling in real water samples from the effluent of the WWTP using the reduced spectral data set, results were also promising with low prediction errors (less than 20%).

1. Introduction

Continuous monitoring of wastewater treatment plants (WWTP) is carried out to guarantee that the operational process acts in accordance with the legislative requirements on water quality for safety, environmental protection and efficient use of water resources. The number of quality and operational parameters/variables continuously measured is high, and requires a systematic approach to analyze the entire WWTP process and extract valuable information (Rosen and Lennox, 2001).

There are many methods available for the determination of water quality parameters in a WWTP such as total carbon, inorganic carbon, total organic carbon (TOC), non-purgeable organic carbon (NPOC), pH, alkalinity, conductivity, chloride, sulphate, nitrate and others (APHA, 1998). Although accurate and precise, these procedures are time consuming and thus expensive when analyzing a large number of samples during routine monitoring. In addition, some of these procedures require long analysis time and therefore the quality values then generated cannot help WWTPs managers to address the potential problems occurring in the plant in real time.

* Corresponding author. Tel.: +34 934006140; fax: +34 932045904.
E-mail address: rtaqam@idaea.csic.es (R. Tauler).

http://dx.doi.org/10.1016/j.jenvman.2014.03.006
0301-4797/

In recent years, UV–Vis spectrophotometry has emerged as a possible and adequate technique for batch and continuous online water quality monitoring. This technology takes advantage of the fact that most organic compounds and a number of soluble mineral compounds (such as nitrates) absorb in the UV-VIS region. UV spectrophotometry can thus be useful in this field for the proper operation of water treatment systems (Thomas and Burgess, 2007). It is a fast, non-destructive, inexpensive and simple analytical technology that has been already proposed for wastewater quality monitoring and organic matter characterization (Clement and Yang, 2001; Vaillant et al., 2002), despite of its major drawback, the lack of selectivity.

On the other hand, multivariate analysis is a powerful tool for extraction of valuable information from large multivariate data collected in WWTP monitoring programs (Rosen and Lennox, 2001). Multivariate data analysis methods, such as Principal Component Analysis (PCA, Jolliffe, 1986), are techniques which are able to visualize and interpret relationships existing among observations/samples and operational WWTP parameters/variables in large and complex databases (Oulali et al., 2009). Generally, water quality monitoring data sets have a multi-dimensional structure, with spatial and temporal variations of multiple variables and operational conditions (Singh et al., 2006). Multivariate data analysis methods are required to explore and extract relationships and information hidden in the data sets. Multivariate regression methods such as Partial Least Squares (PLS, Wold et al., 1983) have been successfully applied to the analysis of multicomponent systems with overlapped and multicollinear spectra (Jørgensen et al., 2004) such as UV-VIS spectra. The combination of UV-VIS spectroscopy and multivariate data analysis is proposed an adequate technique for routine monitoring of water quality inside WWTPs.

Some experimental systems for on-line and in-line wastewater quality spectrometric monitoring have already been proposed in the literature or even in the market. It is worth to mention the experimental system proposed by Thomas et al. (1997) comprising direct UV-VIS exploitation of spectra using a sequential injection analysis system, based on chemical reaction to estimate sequentially a number of parameters (e.g. TOC, COD, BOD, TSS, global N). Other major innovations have been the implementations of submersible UV-VIS spectrometers (like ScanLyser by s:can Messtechnik GmbH, Austria). This instrument utilizes the whole UV/VIS range between 200 and 750 nm (Langergraber et al., 2004) and allows the monitoring of many parameters simultaneously, among them nitrate, organic matter and suspended solids. The estimation of these parameters is performed using a chemometrics calibration modeling, which is integrated inside the instrument (Langergraber et al., 2003). The spectrometer is installed directly in the water flow with the advantages of allowing continuous measurement in real-time. The main disadvantages are its high price and the need of continuous recalibrations in order to achieve good accurate predictions.

This paper presents two surveys where a chemometric approach is applied to data obtained from different case studies:

In a first survey, changes of water quality parameters in a real WWTP were explored by PCA over different time scales campaigns. In addition, results obtained using conventional laboratory methods were compared with those obtained with an automatic multi-parametric monitoring system (AMS) and with this system coupled with a multi-LED optical device recording spectra at particular ultraviolet wavelengths (AMS-LED).

The second survey includes the chemometrics modeling of four target water quality parameters such as phenol, nitrate, dodecylbezosylphonate and DOM concentrations in synthetic mixtures prepared in laboratory. A predictive PLS model was developed based on the UV–VIS spectral data for the simultaneous determination of all four parameters above mentioned. In particular, the most important spectral wavelengths were selected based on Variable Importance in Projection scores (VIP, Chong and Jun, 2005), obtained in a PLS model, and were able to predict the target analyte concentrations with the lowest errors. VIP scores were also applied in a real WWTP scenario to predict concentrations of Nitrate (NO_3) and Total Organic Carbon (TOC) in real samples. Predicted values of these parameters were also compared with the corresponding reference values obtained by standard laboratory methods.

In summary, the global aims of these two surveys are the following: i) evaluation of the existing possibilities (techniques and routines) for continuous monitoring of water quality at a WWTP, ii) gaining knowledge about type and variation sources on water quality parameters in WWTP effluents and iii) possible use of a reduced number of UV spectral channels (wavelengths) to monitor WWTP operational processes.

2. Materials and methods

2.1. Description of the wastewater treatment plant

Monitoring campaigns for data collection of water quality parameters were conducted in the municipal conventional wastewater treatment plant (TRAGISA) located nearby the town of Girona, Northern Spain (45.000–55.000 m3/day flow plant capacity), which collects urban wastewater from this town of approximately 10.000 inhabitants and its surroundings.

2.2. Data acquisition and organization

Table 1 presents the water data sets investigated in this work, focusing on time scales of monitoring campaigns, sampling frequencies and applied analytical techniques. Also, locations where samples were taken from, and number of quality parameters included in every analysis are given. Data sets A and B from Table 1 are first presented and discussed separately in the paper. Water data sets C1–C3 and D1-D3 refer to data generated over 24 h and over 7 days of monitoring using either a classical UV-VIS spectrophotometer, the APHA standard laboratory methods for analysis, or the analytical multi-parametric Monitoring System (AMS) with an UV–VIS absorption spectral LED device. Table 2 shows all parameters measured and monitored in the WWTP either by the application of APHA standard laboratory methods or by the AMS with and without an UV multi-LED optical device.

2.2.1. Monitoring of conventional water quality parameters in the WWTP

Routinely, different water quality parameters from influent and effluent water are measured daily in the WWTP laboratory. All analytical methods applied in WWTP are in accordance with Standard Methods for Examination of Water and Wastewater (APHA, 1998). The parameters routinely monitored in influent and effluent water were: pH, conductivity, soluble solids, chemical oxygen demand (COD), biological oxygen demand in 5 days (BOD_5), phosphates (PO_4–P), total nitrogen (NTK), nitrates (NO_{3-}). Additional parameters (see Table 3, Data sets A, C2 and D2) were measured in some particular cases. In all cases, water samples were collected and kept at 4 °C prior the laboratory analyses.

Data set A contains information about water quality parameters obtained in the WWTP routine daily monitoring, both in influent and effluent wastewater. The complete data set covered 138 samples-days taken throughout the entire 2010 year (including days from all four seasons). These parameters and their

Table 1
List of data sets used for multivariate analysis in this paper.

Data set	Monitoring campaign	Sampling	Analytical technique[a]	Number of parameters	Location	Chemometric method
A	One year, 138 samples	Daily	APHA methods	18	Influent -effluent	PCA
B	One month, 1738 samples	Every 20 min	AMS	9	Effluent	PCA
C1	One week, 7 samples	Daily	UV-VIS spectroscopy	190-1100 nm	Effluent	PCA
C2	One week, 7 samples	Daily	APHA methods	11	Effluent	PCA
C3	One week, 7 samples	Daily	AMS-LED	14	Effluent	PCA and PLS
D1	Twenty-four hours, 24 samples	Every hour	UV-VIS spectroscopy	190-1100 nm	Effluent	PCA
D2	Twenty-four hours, 24 samples	Every hour	APHA methods	12	Effluent	PCA
D3	Twenty-four hours, 24 samples	Every hour	AMS-LED	13	Effluent	PCA and PLS
E	Batch, 49 samples	Exp. design	UV-VIS spectroscopy	190-1100 nm	Mixtures	PLS

[a] APHA (Standard Methods for the Examination of Water and Wastewater, APHA, 1998) analytical techniques implemented in WWTP; AMS Analytical multi-parametric Monitoring System of (ADASA S.A. Inc. Barcelona, Spain) UV-VIS spectral measurements by the Agilent HP8453 spectrophotometer; AMS-LED Analytical multi-parametric Monitoring System with an ultraviolet optical multi-LED device measuring at six specific wavelengths (ADASA S.A. Inc, Barcelona, Spain).

corresponding descriptive statistics are summarized in Table 3, Data set A.

Data sets C2 and D2 were obtained over shorter-scale monitoring campaigns just in the wastewater effluent. Daily samples were taken along a week for data set C2 (descriptive statistics shown in Table 3, Data set C2), whereas for data set D2, a sample per hour was collected over a day (descriptive statistics shown in Table 3, Data set D2).

In all cases, water samples were collected and kept at 4 °C prior to laboratory analyses. The parameters routinely monitored in this study were: pH, conductivity, soluble solids, chemical oxygen demand (COD), biological oxygen demand in 5 days (BOD_5), phosphates (PO_4–P), total nitrogen (NTK), nitrates (NO_{3-}).

2.2.2. *Monitoring of water quality parameters using an automatic multi-parameter system (AMS) and an UV-VIS absorption multi-LED optical device (AMS-LED)*

The automatic multi-parametric monitoring system (AMS; ADASA S.A. company, Barcelona), was developed for online water quality monitoring at the WWTP effluent. The analytical part of the system is composed by 2 equipments with electric and hydraulic elements (pump, sample intake system and auto cleaning elements) necessary to ensure that the system can work properly with no human operation during long periods of time.

AMS instrument (aquaTest-MO equipment, developed by ADASA S.A. company) monitored water quality data continuously for 9 parameters including temperature, pH, redox potential, conductivity, dissolved oxygen, biological oxygen demand, turbidity, absorption at 254 and 365 nm (254 nm is used to measure the optical absorption of organic matter; 365 nm is used to compensate the optical absorption due to turbidity). Measurements were taken every 20 min during one month (data set B of Table 1). Table 3 B reports the descriptive statistics of the above mentioned parameters.

As an upgrade of the AMS, special interest was paid to the development and performance of a recently developed UV-VIS multi-LED optical device coupled to the AMS (data sets C3 and D3). This new optical device was able to record spectral data at 240, 250, 254, 260, 275 and 365 nm wavelengths. Table 3 data sets C3 and D3 report the descriptive statistics about parameters monitored by AMS-LED over the two shorter scale monitoring campaigns.

2.2.3. *Spectrophotometric analysis of synthetic mixtures of water constituents*

A set of synthetic water samples containing several constituents were synthetically prepared in the laboratory using chemical-grade pure compounds. Investigated relative concentrations of these water mixtures were designed to represent a range of concentration values close to those measured in real plant conditions. Humic acid (Aldrich) concentrations were added to simulate a possible source of dissolved organic matter (DOM). Nitrate standard solution was used to prepare samples with nitrates ions. Phenol (99% pure, Fluka) and kaolin (Aldrich) were used to simulate phenol concentrations and soluble solids in water. Sodium dodecylbenzene sulfonate (pure DBS, Aldrich) was used to give known concentrations of surfactant/detergent in the mixtures. Samples from pure target compounds and of their mixture combinations of 2, 3 and 4 components were prepared using two Box-Behnken experimental designs, BBD (Box and Draper in 1987 did show that BBD is highly effective in 3-levels, two-to several factor experiments, giving relatively low number of samples).

A total number of 49 samples (40 were used for calibration and 9 were used for external validation) of one-to-four components in the concentration range 0–10 mg/L were prepared. Kaolin

Table 2
List of parameters measured and monitored in this study[a].

Parameters	APHA methods[b]	AMS[b]	AMS-LED[b]	UV-VIS[b]
Temperature		x	x	
pH	x	x	x	
Redox potential		x	x	
Conductivity	x	x	x	
Dissolved oxygen		x	x	
SAC		x	x	
Absorption at 240 nm			x	
Absorption at 250 nm			x	
Absorption at 254 nm		x	x	
Absorption at 260 nm			x	
Absorption at 275 nm			x	
Absorption at 365 nm		x	x	
Absorption 190–1100 nm				x
BOD	x		x	
Turbidity		x		
Suspended solids	x			
COD	x			
PO_4–P	x			
NTK	x			
NO_3	x			
NH_3	x			
Water flow influent	x			
Water flow effluent	x		x	
TOC	x			

[a] The number of parameters in Table 2 does not correspond to the number of parameters reported in Table 1, since some of them were measured in the influent and other in the effluent for a specific data set.

[b] APHA (Standard Methods for the Examination of Water and Wastewater, APHA, 1998) analytical techniques implemented in WWTP; AMS Analytical multi-parametric Monitoring System of (ADASA S.A. Inc. Barcelona, Spain) UV-VIS spectral measurements by the Agilent HP8453 spectrophotometer; AMS-LED Analytical multi-parametric Monitoring System with an ultraviolet optical multi-LED device measuring at six specific wavelengths (ADASA S.A. Inc, Barcelona, Spain).

Table 3
Descriptive statistics of parameters monitored in Girona WWTP[a]. Tables C1 and D1 report only spectral data and are not presented here.

Data set A. Physicochemical parameters APHA[1], one year	Abbreviation	Mean	Min	Max	Std
pH in the outlet	pHout	7.3	7.0	8.5	0.2
Conductivity in the outlet μS/cm	CONDout	945.3	613.0	1184.0	110.9
Biological oxygen demand 5 days in the outlet mg/L	BOD5out	4.1	2.0	17.4	1.8
Chemical oxygen demand in the outlet mg/L	CODout	29.5	8.2	88.0	9.2
Suspended solids in the outlet mg/L	SSout	4.9	1.6	28.0	3.1
Total nitrogen (Khejdal method) in the outlet mg/L	NTKout	6.7	1.5	21.5	3.1
Nitrates in the outlet mg/L	NO3out	3.9	0.8	8.5	1.5
Total phosphates in the outlet mg/L	Ptotout	0.2	0.1	1.5	0.2
Water flow in the outlet m3/h	Flowout	37.900	32.600	57.400	4.000
pH in the inlet	pHin	7.6	7.3	7.9	0.1
Conductivity in the inlet μS/cm	CONDin	1162.3	118.0	2180.0	195.6
Biological oxygen demand 5 days in the inlet mg/L	BOD5in	206.1	18.0	369.0	57.4
Chemical oxygen demand in the inlet mg/L	CODin	513.1	6.4	1049.0	153.1
Suspended solids in the inlet mg/L	SSin	264.0	74.0	660.0	109.1
Total nitrogen (Khejdal method) in the inlet mg/L	NTKin	49.6	23.1	94.9	12.5
Ammonia in the inlet mg/L	NH4in	33.3	9.0	71.0	8.9
Nitrates in the inlet mg/L	NO3in	1.2	0.1	3.3	0.7
Total phosphates in the inlet mg/L	Ptotin	7.3	3.3	13.8	2.1

Data set B. Parameters AMS, one month	Abbreviation	Mean	Min	Max	Std
pH	pH	6.9	6.7	7.0	0.0
Conductivity μS/cm	COND	930.6	647.8	1014.7	78.1
Dissolved oxygen mg/L	O_2	3.9	2.0	5.3	0.6
Redoxy potencial mV	RedOx	210.9	150.5	272.1	28.6
Temperature in C°	TEMP	19.5	18.0	22.4	0.8
Turbidity NFU	TURB	2.9	0.0	12.0	3.1
Specific absorption coefficient	SAC	20.5	0.1	30.8	8.6
Absorption at 254 nm	Abs 254	20.8	1.5	29.3	7.7
Absorption at 365 nm	Comp 365	4.2	1.9	5.7	0.9

Data set C2. Parameters APHA, week	Abbreviation	Mean	Min	Max	Std
pH	pH	7.2	7.0	7.6	0.2
Conductivity μS/cm	COND	1055.4	952.0	1125.0	58.4
Chemical oxygen demand mg/L	COD	24.1	11.0	29.0	6.3
Biological oxygen demand 5 days in the inlet mg/L	BOD5	6.1	5.0	12.0	2.6
Volatiles suspended solids mg/L	SSV	3.0	2.0	5.0	1.2
Suspended solids in the outlet mg/L	SS	3.6	2.0	6.0	1.5
Total nitrogen mg/L	NT	6.9	4.0	9.1	1.9
Nitrates mg/L	NO3	5.4	2.8	7.2	1.6
Phosphates mg/L	PO4	2.5	0.9	5.4	1.8
Total organic carbon mg/L	TOC	14.5	10.6	24.6	5.6
Water flow in the inlet m3/h	Flowin	2002.9	1912.0	2099.0	59.4

Table 3 (*continued*)

Data set C3. Parameters AMS-LED, week	Abbreviation	Mean	Min	Max	Std
pH	pH	6.9	6.9	7.0	0.1
Conductivity μS/cm	COND	906.0	820.0	941.0	41.0
Temperature in C°	TEMP	24.4	23.7	24.7	0.4
Redoxy potential mV	RedOx	241.0	234.0	245.0	4.0
Dissolved oxygen mg/L	O2	3.7	3.2	4.1	0.3
Biological oxygen demand mg/L	BOD	11.1	6.3	15.9	2.8
Biological oxygen demand mg/L	BOD_2	11.1	7.4	15.2	2.6
Specific absorption coefficient	SAC	27.2	25.6	28.6	1.1
Absorption at 240 nm	A240	25.1	23.7	26.1	1.0
Absorption at 250 nm	A250	17.8	16.3	18.7	1.1
Absorption at 260 nm	A260	31.9	30.8	32.5	0.7
Absorption at 275 nm	A275	29.3	28.3	29.9	0.7
Absorption at 254 nm	A254	24.8	23.7	25.8	0.8
Absorption at 365 nm	A365	2.8	2.5	3.0	0.2
Water flow in the outlet m3/h	Flow	1.4	1.4	1.5	0.0

Data set D2. Parameters APHA, 24 h	Abbreviation	Mean	Min	Max	Std
pH	pH	7.3	7.2	7.7	0.1
Conductivity μS/cm	COND	996.4	982.0	1017.0	9.8
Chemical oxygen demand mg/L	COD	27.0	22.0	52.0	6.0
Biological oxygen demand 5 days in the inlet mg/L	BOD5	6.4	5.0	16.0	2.2
Suspended solids in the outlet mg/L	SS	3.9	3.0	9.0	1.4
Volatiles suspended solids mg/L	SSV	3.1	0.0	8.0	1.6
Total nitrogen mg/L	NT	9.1	8.4	9.9	0.6
Nitrates mg/L	NO3	7.3	6.5	8.0	0.5
Phosphates mg/L	PO4	1.0	0.5	5.6	1.0
Total organic carbon mg/L	TOC	11.7	10.0	17.7	1.4
Water flow in the inlet m3/h	Flowin	1655.8	831.0	2249.0	406.8
Water flow in the outlet m3/h	Flowout	1619.8	795.0	2213.0	406.8

Data set D3. Parameters AMS-LED, 24 h	Abbreviation	Mean	Min	Max	Std
pH	pH	6.9	6.9	7.0	0.0
Conductivity μS/cm	COND	943.0	924.0	955.0	8.0
Temperature in C°	TEMP	24.4	23.7	26.9	0.8
Redoxy potential mV	RedOx	233.0	217.0	279.0	12.0
Dissolved oxygen mg/L	O2	3.6	3.2	4.3	0.3
Biological oxygen demand mg/L	BOD	13.2	9.9	16.1	2.1
Specific absorption coefficient	SAC	25.8	24.8	26.7	0.5
Absorption at 240 nm	A240	24.3	23.5	25.1	0.5
Absorption at 250 nm	A250	16.8	15.9	17.9	0.6
Absorption at 260 nm	A260	31.0	30.3	32.2	0.5
Absorption at 275 nm	A275	28.6	27.8	29.3	0.4
Absorption at 254 nm	A254	23.7	22.7	24.5	0.5
Absorption at 365 nm	A365	2.8	2.4	3.0	0.2

[a] Tables C1 and D1 report spectral data.

concentration varied in the range between 0 and 15 mg. Once prepared, every sample was measured using a conventional UVVIS diode array spectrophotometer (Agilent HP8453) and their spectrums were recorded and use to build data set E (see Table 1). UV spectra were recorded in the wavelength range between 190 and 350 nm where strong absorption of the target compounds was measured.

2.3. Chemometric methods and figures of merit

Experimental data sets were properly arranged in data matrices and autoscaled (column mean centered and scaled) to eliminate offsets and changes in measurement scales. PCA (Jolliffe, 1986) was applied for exploratory analysis of water quality parameters of WWTP. This chemometrics method extracts useful information about the latent (hidden) structures of a particular data set. It transforms a large number of correlated original measured variables (in our case water quality parameters or UV spectral data) into a smaller number of uncorrelated, orthogonal variables explaining

maximum variance, called principal components (PCs). Two types of plots are obtained from the application of PCA. Loadings plots which describe and map the relationships between water quality parameters and the extracted principal components, and scores plots, which describe and map the samples (different time observations, minutes/hours/days of monitoring properties) in the new axes defined by the principal components, allowing the easier investigation of their relationship.

In order to predict the concentrations of the selected target compounds/parameters in wastewater, two PLS regression algorithms have been applied in this work (Wold et al., 1983). First, PLS2 was used for the case of building a single model between the concentrations of a set of wastewater constituents (like organic matter, phenols, NO_3, DBS) with the collected UV spectra in the laboratory. Second, multiple PLS1 models were used to predict NO_3 and TOC concentrations in a real WWTP scenario, using the AMS-LED spectral data.

PLS2 was preferred in the first case over PLS1, because it allowed simultaneous calibration of all target analyte concentrations with the lowest possible prediction errors. The use of Variable Importance in Projection (VIP) scores is proposed as useful tool for interpreting the more relevant variables in PLS models (Chong and Jun, 2005). Individual VIP scores from this PLS2 model can be compared visually in order to select a few more important wavelengths, useful for the simultaneous prediction of all target parameter concentrations.

In all cases, the same strategy was used, starting with model calibration, followed by its internal cross-validation (leaving-one out) and ended by its external validation (with samples not included in the calibration). Selection of optimal number of components in PCA or of latent variables in PLS1 and PLS2, has been using the lowest prediction error in cross validation (leaving-out-one sample at a time) and in external validation. The model giving the lowest relative prediction errors in external validation prevails and it is finally chosen.

Quality assessment of the obtained results is discussed by comparison of predicted values versus measured values, both for calibration and for validation data sets. To evaluate numerically the quality of the obtained results, the coefficient of determination of model fitting or R^2 and the Root Mean Squared Error of Calibration or of external Prediction (RMSEC and RMSEP) were used. RMSEC is calculated as follows:

$$\mathrm{RMSEC} = \sqrt{\frac{\sum_{i=1}^{n} (\hat{y}_i - y_i)^2}{n}},$$

Where the $\hat{y}_i$ are the values of the model predicted concentrations and y_i are the actual values for calibration samples and n is the number of samples. RMSEC is a measure of how well the model fits experimental concentrations. RMSEP is calculated exactly as RMSEC except that the estimates are now the values from external validation samples. RMSEP is a measure of how well the model will make predictions. Moreover, in this work the calculation of relative prediction errors of concentrations in percentage is also given, for both calibration and prediction steps and they are calculated as follows:

$$\text{Rel. error in\%} = \sqrt{\frac{\sum_{i=1}^{n} (\hat{y}_i - y_i)^2}{\sum_{i=1}^{n} (y_i)^2}} \times 100;$$

Initial data preparation and data arrangement of the different data sets were performed using EXCEL (Microsoft, Redmon, WA, USA). All calculations for chemometric analysis were performed using MATLAB computer and visualization environment and using Statistical Toolbox of MATLAB 6.5 (The Mathworks, Natick MA, USA) and PLS Toolbox 5.8 (Eigenvector Research, Manson, WA, USA). Statistical experimental design was carried out by using the software Unscrambler 9.8 (CAMO PROCESS AS, Oslo, Norway).

3. Results and discussion

3.1. PCA results of the one year WWTP data set (dataset A in Table 3)

PCA on data set A after data autoscaling identified 6 components (PC1 25.6%; PC2 13.2%; PC3 9.2%; PC4 7.6%; PC5 6.5%; PC6 6.2%), which explained 68% of the total data variance. This model is rather complex with a high number of PCs, probably due to the fact that measured variables at the influent and effluent were rather independent before and after the treatment process.

Fig. 1a shows PCA scores plot for PC1 and Fig. 1b shows PC1 vs PC2 loadings plot. PC1 scores and loadings figured out the seasonal (the smooth sigmoid trend over the entire time under investigation, e.g. one year) and weekly (short-termed cyclic fluctuations) water quality trends due to the WWTP activity (see Fig. 1a). Positive scores on 1st PC are observed for samples from spring, autumn and winter seasons, while in summer negative scores on 1st PC are observed. Almost all physical–chemical parameters under investigation at WWTP show positive loadings on PC1 (see Fig. 1b), which means that their concentrations were found above the average values during spring, autumn and winter seasons and below the average values for the summer period. All these suggest that WWTP is characterized with the higher plant activity in spring, autumn and winter, following a considerable reduction of the plant activity in summer, probably due to the reduced urban activity and to the lower water flow. This fact has been confirmed by the experienced technical staff of WWTP. In addition, when scores were examined in more detail, changes in water quality among different days were also detected, e.g. many of weekend samples were discriminated from working days samples (not shown in Fig. 1 for brevity),

On PC1vsPC2 loadings plot (see Fig. 1b), the effect of influent-effluent variation in water quality is shown, from which the role of WWTP performance in the wastewater treatment can be investigated. Many of the parameters monitored in the influent are distributed at the bottom-right corner of the plot (negative loadings on 2nd PC) and are distinguished from parameters monitored in the effluent (positive loadings on 2nd PC) at the upper-right corner. This means that it is observed an inverse correlation between water quality parameters measured in the plant influent and in the plant effluent. The effluent concentrations have positive scores, meaning that above average concentrations in the effluent result in higher values for the 2nd PC while below average influent concentrations result in negative scores. The direction from negative loadings to positive loadings on PC2 figures out the efficiency of plant procedures.

Water quality parameters with high loadings on PC2 will assume low reduction in their concentration, respectively a low plant performance. On the contrary water quality parameters with low loadings for PC2 will assume higher reduction in their concentrations, respectively, a high plant performance.

Some exceptions can be detected for nitrate concentrations - smaller PC2 loadings for the nitrate concentrations in the effluent than in the influent. This, however, confirms the fact that nitrate concentrations increased in the effluent due to WWTP activity (Thomas and Burgess, 2007). Concentrations of all other water quality parameters, like SS, COD, BOD_5, total nitrogen and others, were reduced due to WWTP activity.

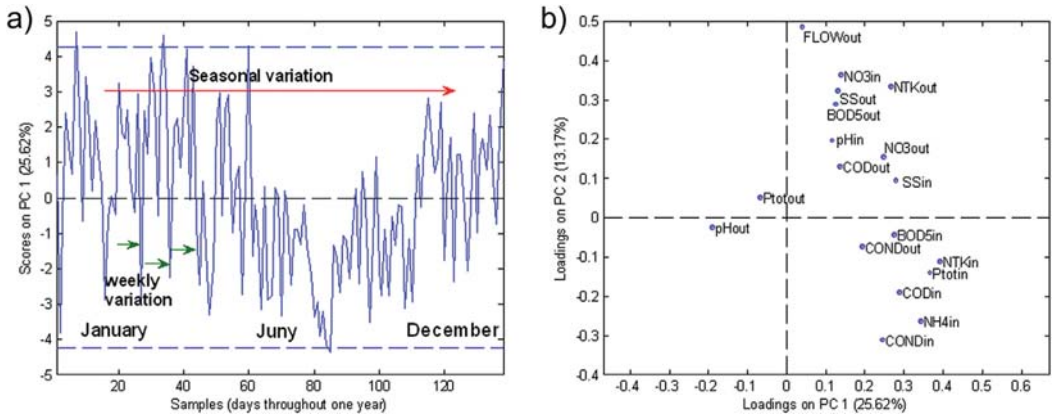


Fig. 1. PCA results of data set A generated in the Girona WWTP during one year: a) PC1 sample scores plot (days of observation); b) PC1 vs PV2 loadings plot of the eighteen physicochemical parameters measured by APHA methods.

3.2. PCA results of the one month WWTP data set (dataset B in Table 3)

A thorough exploratory analysis on the variation of quality in wastewater effluent was performed with data collected every 20 min by the AMS during one month of continuous monitoring. PCA was also carried out after data (see Tables 1 and 2). PCA model with 3 principal components explained around 88% of the total variance in the data (PC1 45.9%; PC2 27.4% and PC3 11.5%, all components with eigenvalues larger than one), This PCA model gave better features (only three principal components explaining 88% of data variance) than this obtained in previous analysis (six principal components explaining 68% of variance) of conventional plant parameters (dataset A in Table 3). This fact is probably due to the higher correlation that exist among measured variables, resulting in a PCA model that explains better the changes in the water quality when data collection was subjected to every 20 min (AMS system) in contras to data generated once a day (data collected by APHA methods).

Fig. 2 shows PCA scores and loadings for the model with 3 PCs. Interpretation of scores and loadings on 1st PC are given in Fig.s. 2a and b. They revealed notorious diurnal, short-termed cyclic fluctuations of water quality. Moreover, a larger fluctuation of the trend line (Fig. 2a at the right corner) at the end of the month is observed (strong negative scores on PC1). This change of the trend line in water quality was presumably due to the joint effect of Easter holiday vacation (decrease of the urban activity) and several episodes of rainfalls occurred during these days in Girona area. The corresponding loadings plot (Fig. 2b) confirms a notable increase of dissolved oxygen concentration and of redox potential values (i.e. strong negative loadings on PC1 on contrary to the other parameters loadings) in the water effluent due to the reduced urban activity for the holidays plus the accumulation of fresh rain water, richer in dissolved oxygen.

On the scores plot of PC2 in Fig. 2c (27.5% of the explained variance), variations of water quality due to morning-midnight fluctuations are displayed. Positive scores are monitored at mornings and, on the contrary negative scores are at midnight hours throughout this investigation. In Fig. 2d, water temperature, pH, red-oxy potential and dissolved oxygen have the highest positive loading values on PC2. The last two parameters have higher concentration values during the morning than at midnight hours. Third PC (11.5% of the explained variance) also illustrates the operational pattern of the WWTP process with the cyclic morning-afternoon recurrence (see Fig. 2e). Morning samples are monitored with positive scores in contrary to afternoon samples with negative scores. The explanation for this trend can be found on the loadings plot of Fig. 2f. It shows that the turbidity parameter had the major positive contribution on this PC3. Water with higher turbidity values was released in the mornings (also with positive sample scores on Fig 2e). All this reflects WWTP operational procedures where large volumes of treated water are released from the WWTP in mornings, i.e. when new wastewater also entered the plant. On the contrary, water turbulence decreased during the day, when the water incoming-outgoing process turned steady in the afternoons (negative sample scores on Fig. 2e).

3.3. PCA results of the seven days WWTP data sets (datasets C1, C2 and C3 in Table 3)

Fig. 3 shows the achieved PCA results in the analysis of the three data sets collected simultaneously by the UVVIS laboratory instrument (data set C1), from the standard laboratory methods (data set C2) and by the AMS-LED (data set C3), over the same week of monitoring. Samples were collected once a day at 11 h A.M. The sampling time was adjusted considering residential time of water in the WWTP in order to get representative samples and to make possible their comparison.

PCA model of spectral data (data set C1) collected by a laboratory UVVIS instrument explained more than 98% of the variance with 2 PCs. The other two PCA models (one for physicochemical data set C2 generated in the laboratory and another for data set C3 collected by AMS-LED instrument) explained around than 60–70% of variance using first two PCs. Fig. 3a, c and e show score plots for the first two PCs and Figs. 3b, and f show the corresponding loading plots.

The five weekdays were well discriminated from weekend days on the three score plots. The scores for the weekend show negative scores on PC1 in Fig. 3a and c. In Fig. 3e the combination of PC1 vs PC2 distinguish similarly weekends from the weekdays. The analysis of the loadings plot of spectral data (data set C1) revealed a shoulder at 220 nm and a band maximum at 226 nm (see Fig. 3b). The absorption at 220 nm is assigned to organic matter and nitrates (APHA, 1998) and the peak at 226 nm is assigned to detergents, since many of them do absorb at this wavelength (Thomas and Burgess, 2007). Therefore, the relationship between concentrations of organic matter and detergents is of major importance for the description of changes during weekdays.

Analysis of loadings plot of laboratory data (data set C2, see Fig. 3d) reveals the importance of pH (strong negative loadings on PC1) and nitrates (strong positive loadings on PC1) parameters,

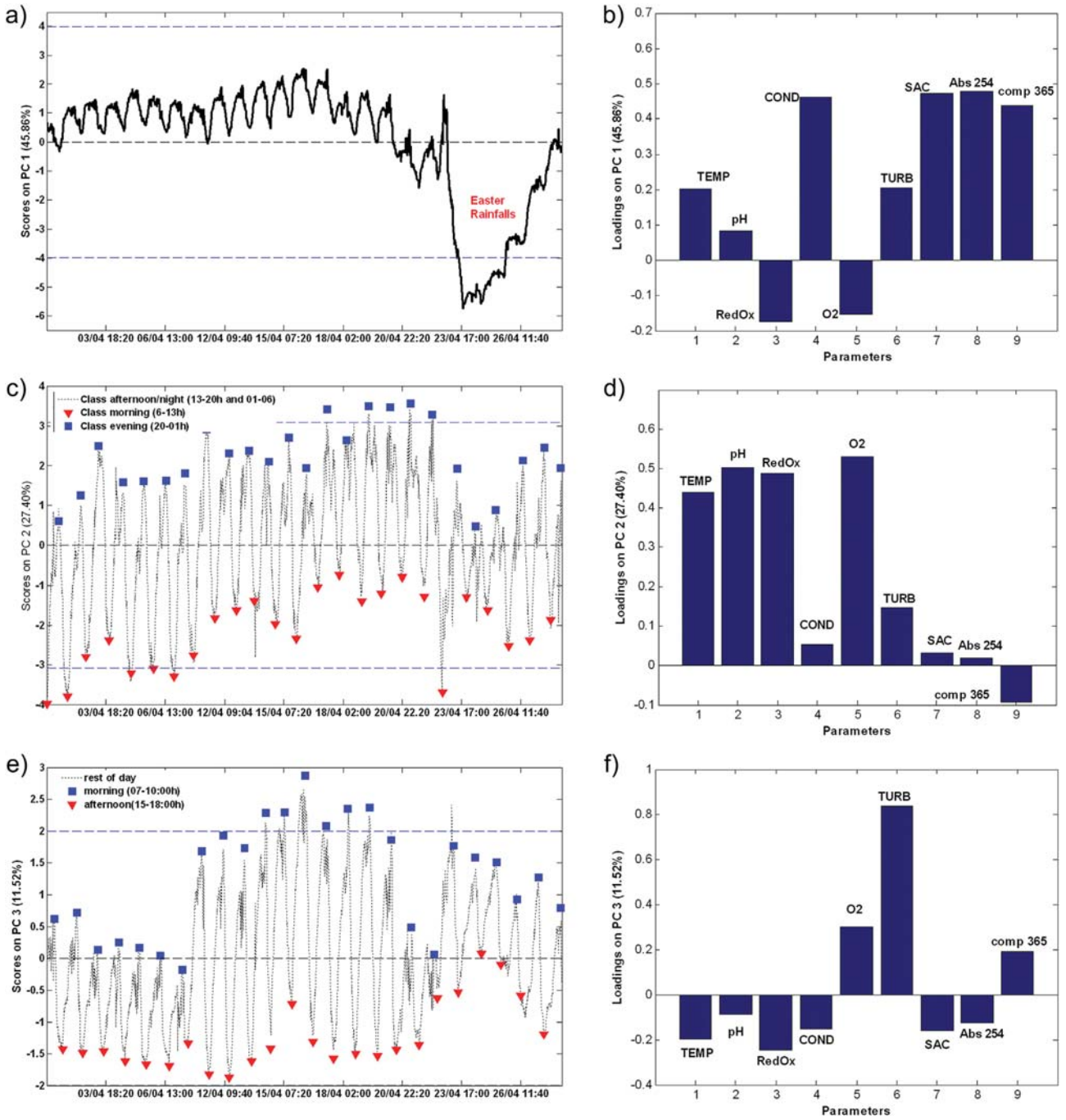


Fig. 2. PCA results of data set B collected by the Analytical Multi-parametric Monitoring System (AMS) in the WWTP Girona during one month: a) PC1 scores plot for samples taken every 20 min; b) PC1 loadings plot for the nine physicochemical parameters recorded by AMS. c) PC2 scores plot for samples taken every 20 min; d) PC2 loadings plot for the nine physicochemical parameters recorded by AMS. e) PC3 scores plot for samples taken every 20 min); f) PC3 loadings plot for the nine physicochemical parameters recorded by AMS. The blues squares and the red triangles at Fig. 2c and e were associated with extreme samples from morning, afternoon and midnight time zones. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

which correlate inversely. Therefore, the five weekdays (positive scores on PC1) were characterized with high concentrations of nitrates in contrast to the weekend (negative scores on PC1), when pH increased probably due to higher concentrations of detergents in the wastewater from the intensive household activity in Girona town.

Fig. 3e shows that PCA model of data set C3 from the AMS-LED instrument, differentiates Tuesday to Friday (all positive scores on PC1) plant treatment from the rest of days (all negative scores on PC1). This fact can be deduced from the analysis of the corresponding loadings plot (Fig. 3f). The importance of the absorption selected variables between 240 and 275 nm, as well as BOD, is

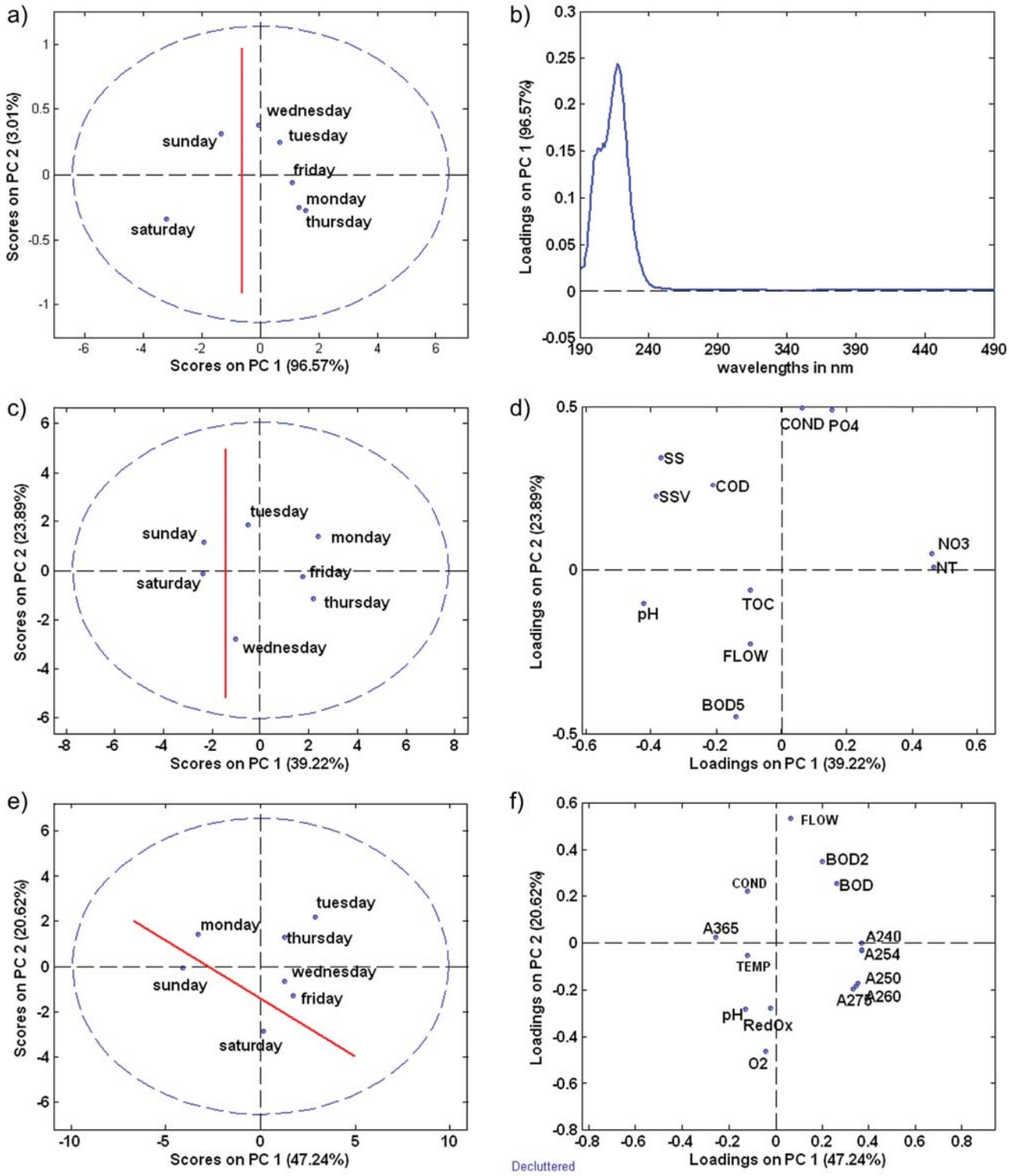


Fig. 3. PCA results of data sets C1, C2 and C3 (seven days of monitoring using three different analytical techniques): a) PC1 scores plot for samples analyzed by UV-VIS spectrophotometer during seven days (spectral data set C1); b) PC1 loadings plot for spectral region between 190 and 490 nm (spectral data set C1); c) PC1 scores plot for samples analyzed by APHA methods during seven days (data set C2); d) PC1 loadings plot for eleven physicochemical parameters obtained by APHA methods (data set C2). The red cutoff lines are hand written to distinguish better. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

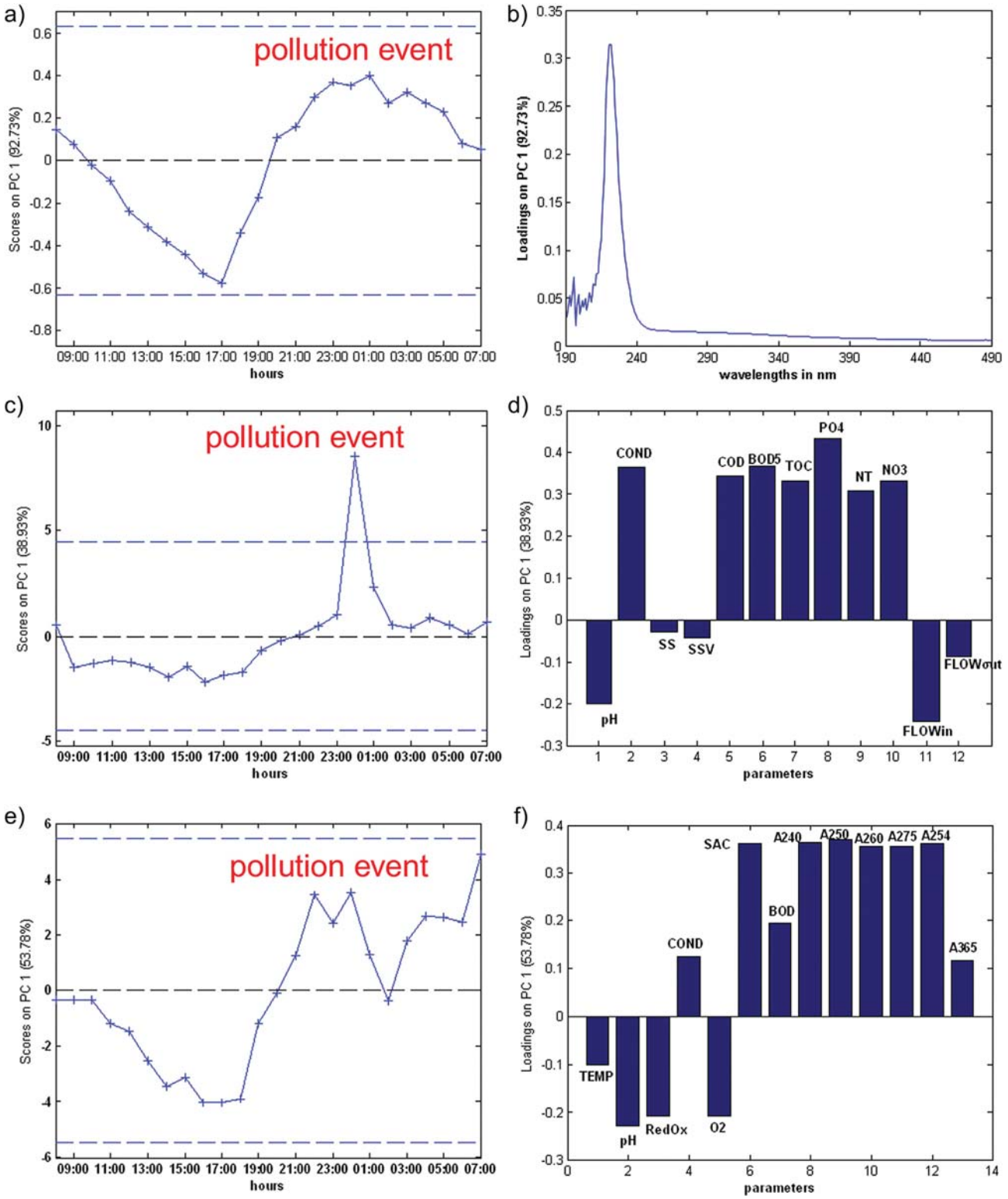


Fig. 4. PCA results of data sets D1, D2 and D3 (twenty-four hours monitoring by three different analytical techniques): a) PC1 scores plot of samples analyzed by UV-VIS spectrophotometer during seven days (spectral data set D1); b) PC1 loadings plot for spectral region between 190 and 490 nm (spectral data set D1); c) PC1 scores plot for samples analyzed by APHA methods during twenty-four hours (data set D2); d) PC1 loadings plot for twelve physicochemical parameters obtained by APHA methods (data set D2); e) PC1 scores plot for samples (hours) analyzed by AMS-LED during 24 h (data set D3); f) PC1 loadings plot for thirteen physicochemical parameters collected by AMS-LED (data set D3).

displayed on the positive side of PC1 (right half of the plot), which should be related to the absorption of higher organic matter concentrations. Tuesday to Friday are characterized by higher absorbance values and higher concentrations of organic matter. Generally, higher content of organic matter is considered to be an indication of lower water quality and therefore, the results achieved directly from PCA modeling of optical data can provide valuable information about water quality. In contrast, weekends are characterized with higher values of parameters like pH, dissolved oxygen and redox potential (down left corner of the plot). These parameters, with high influence on the water quality, show that the quality of the effluent water during the weekend is better than in week days (with higher concentrations of dissolved oxygen and lower amounts of organic matter) even trough the quality of the effluent during the weekend had being dominated by the household activity with higher detergent concentration.

However all discovered trends (in one week and one day monitoring of WWTP activity by analysing data from APHA methods, UVVIS and AMS-LED) were rather similar -proving the efficiency of the AMS-LED system. These have been only preliminary results and they should be performed on a longer timescale studies to prove validity of the derived results.

3.4. PCA results of the twenty-four hours WWTP data sets (data sets D1, D2 and D3 in Table 3)

In Fig. 4, PCA scores and loadings for the twenty-four hours (twenty-four samples) monitoring UV-VIS full spectral data set D1 are given (see Fig. 4a and b). PCA results for data set D2 generated by the standard laboratory methods in WWTP are given in Figs. 4c and d and PCA results for data set D3 collected by the AMS-LED system are given in Figs. 4e and f. PCA models displayed in the three cases a continuous daily-night sigmoid trend in the water quality of the effluent. Fig. 4a shows a smooth trend line with time hours. Nighttime hours have positive scores on the plot in contrast to the morning-afternoon diurnal hours, with negative scores. There is an inflection point with a sharp change of water quality at 17.00 h in the afternoon. The maximum value of this trend line can be found around 00.00 h at midnight. Fig. 4b gives the corresponding loadings and it shows the spectral bands with highest positive contributions to the first PC of the model. The maximum absorption value was found around 220 nm. Nitrates and dissolved organic matter are most probably the responsible for this strong absorption at this wavelength (APHA, 1998). It can be concluded that the sharp change of the trend line is due to the rapid increase of concentrations of nitrate and organic matter at this time.

Fig. 4c gives the PCA scores plot for physicochemical data (data set D2) generated in WWTP laboratory. The scores trend line is also dominated by a sharp peak detected at 00.00 h midnight. The rest of scores followed the previously mentioned daily-night trend line for the absorption data (data set D1), with positive scores for the evening-night hours and negative scores for the daylight time. Investigation of Fig. 4d made possible the determination of parameters with highest loadings on PC1. They were COD, BOD_5, TOC and PO_4. All these parameters identified the presence of a pollution event, with very high amounts of organic matter occurring at midnight.

Fig. 4e shows PCA results of the physicochemical-spectral data set D3 (collected by the AMS-LED). Like previously for data set D2, the same trend line for water quality variation is monitored in the PC1 scores plot. Positive scores at night hours and negative scores at morning-afternoon hours of monitoring are observed. The pollution incident appeared again as a peak in the water quality trend line at midnight. Exploration of the corresponding loadings in Fig. 4f revealed that the most influential variables were the specific absorbance coefficient (SAC), with all six measured wavelengths, together with the BOD parameter. Thus, absorption readings of these 6 wavelengths were important to detect the pollution event.

3.5. PLS prediction of four target water quality parameters in synthetics mixtures using UV absorption at selected wavelengths

PLS2 model of the mean-centered spectral data (Esbensen, 2002) with five LVs was able to explain more than 90% of the total **Y** variance. Fig. 5 shows VIP scores obtained by PLS2 modeling of the four target compounds (nitrates, organic matter, detergent and phenols) in the presence of kaolin and in the spectral region 190–350 nm. Three different spectral regions can be distinguished for optimal PLS model performance: 190–210, 220–260 and 270–276 nm since all wavelengths in these regions show absorption values with significant VIPs above the threshold value of 1 (Chong and Jun, 2005). Using this wavelength range selection, amount of data were reduced considerably. A visual preselecting criterion was then performed searching for the most characteristic peaks and for some inflection points to further restrict the spectral wavelengths to a lower number. Spectral region between 190 and 200 nm was not finally considered because the experiment needed to be done in a vacuum chamber (Anderson et al., 2004). Our final selection included wavelengths 201, 205 (nitrate absorption), 226 (detergents), 237, 254 and 285 nm (dissolved organic matter), 270 and 276 nm (phenols). PLS2 model was then recalibrated using only these eight preselected wavelengths.

The new PLS2 model with five LVs was able to explain above 90% of the total variance of all target analyte concentrations. Fig. 6 shows predicted versus actual concentrations for validation samples finally achieved with this PLS2 model. Results demonstrated the presence of a very strong correlation between the actual and predicted values of the target compounds and with good quality parameters (see in Fig. 6. R^2, RMSEC, RMSECV and bias values). In external validation, largest relative prediction errors were found for DOM concentrations (26%). For the rest of target compounds predicted errors varied between 3 and 4%. These figures of merit confirmed the good predictive abilities of the achieved PLS2 model and confirmed therefore the reliability of the preselection method based on VIP scores and additional visual inspection. Possible use of

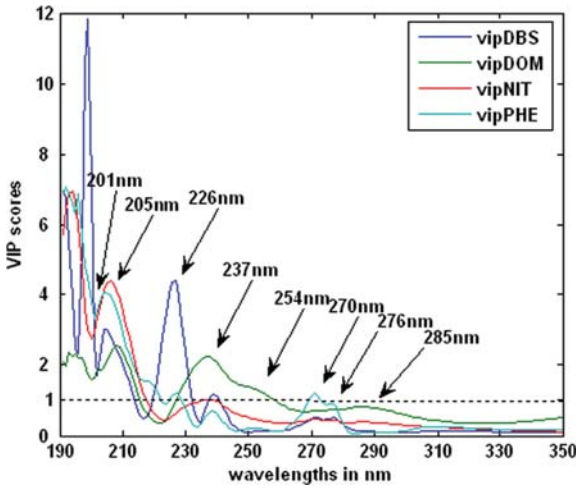


Fig. 5. PLS2-VIP (variable importance in projection) scores in the simultaneous analysis of four target compounds (DBS detergent, dissolved organic matter DOM, nitrates NIT and phenols PHE) in different mixtures (data set E). Wavelengths with VIP scores above the threshold value of one (dotted line) were considered to be the most significant in PLS2 overall modeling. Selected wavelengths in this work are highlighted with arrows and with their nominal values.

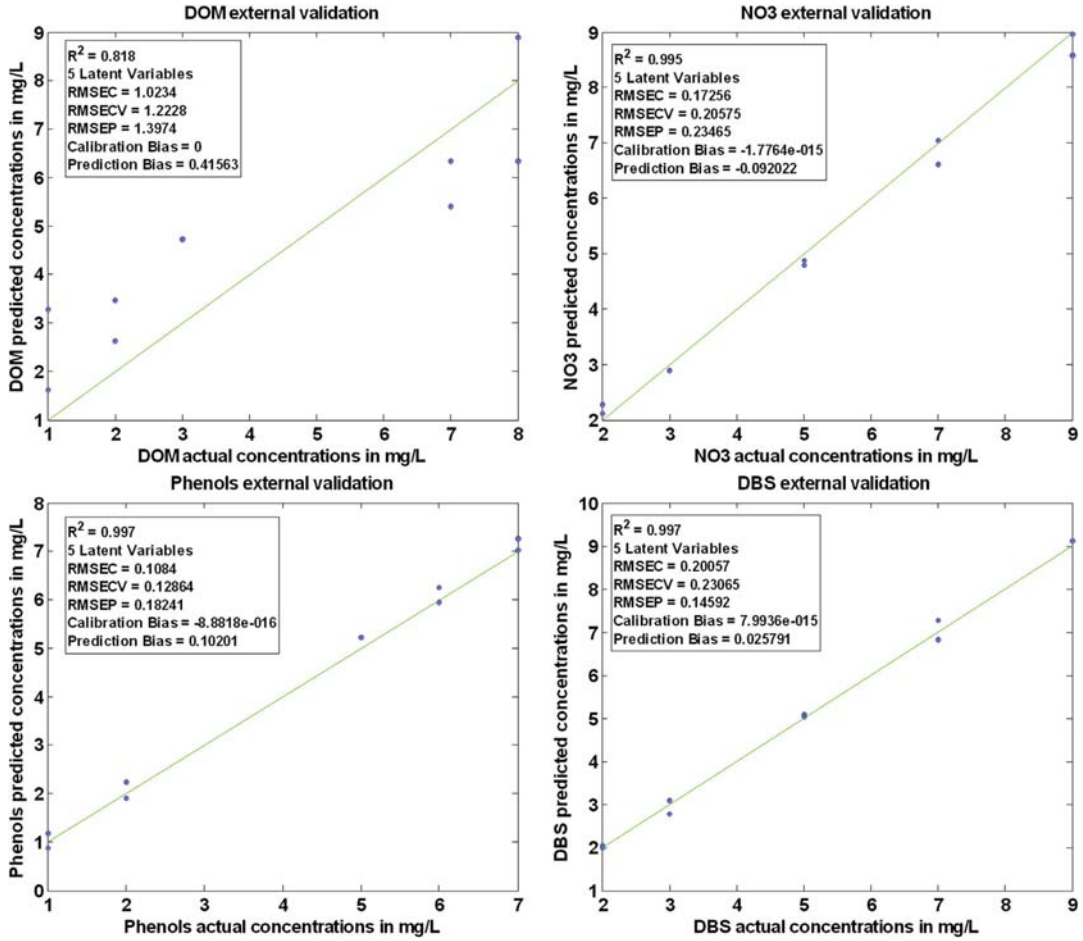


Fig. 6. PLS2 prediction results using eight wavelengths selected from highest VIP score vales (Fig. 5) in the analysis of samples from data set E for the four target compound (DBS detergent, dissolved organic matter DOM, nitrates NO_3 and phenols).

more wavelengths or PLS1 modeling could help to improve DOM prediction results.

3.6. PLS prediction of NO_3 and TOC water quality parameters in real samples using AMS-LED spectral data

Additional individual multivariate PLS1 models were built using only the spectral part of data sets C3 and D3 collected by AMS-LED to quantify individually NO_3 and TOC during the 24 h and 7 days real samples monitoring campaigns. Models developed for the 24 h data were externally validated using randomly selected 18 samples for calibration and 6 samples for external validation. PLS1 models developed for the seven day monitoring was only internally cross validated, due to the lack of enough number of samples. The best model was found for the prediction of NO_3 in the 24 h monitoring, explaining 70% of NO_3 concentration values using 94% of the spectral variance (1st LV) captured by the six wavelengths (**X** data). When the model was applied to external validation data subset (6 h), a relative prediction error of 3.5% was obtained. Model developed for the 7 days data set (spectral part of data set C3) reported a low RMSECV value, which confirmed the possibility of modeling NO_3 over a larger time. Considering the laboratory study with synthetic samples and previous literature results (Thomas et al., 1990), NO_3 shows maximum absorbance at 205 nm. Modeling of NO_3 concentrations using the selected 6 wavelengths (240, 250, 254, 260, 275 and 365 nm) should be based mostly only on spurious correlation with other compounds. This fact has been already reported by other authors (Dahlén et al., 2000) for other compounds modeled in the wastewater.

When the spectral data set C3 from 7 days monitoring were concatenated to the spectral data set D3 of 24 h monitoring (the new data set has 31 samples), the range of variation in nitrate concentration increased from 3 to 8 mg/L and therefore this new model was more robust. Relative prediction errors in the external validation were found to be 12.5%. It should be noted however, that in order to have more accurate predictions, more observations are needed and a broader concentration range should be considered.

Prediction results worsened when the model was built on data considering only first twelve hours in the calibration set and the model was then applied over the next 12 h data. This happened because of the existing continuous systematic changes in water quality as demonstrated from its constant fluctuations. In this study, the situation was avoided by a random selection of samples. However, PLS global models have to be calibrated on data covering up a minimum period of one full-range cyclic fluctuation in the WWTP or in case of larger monitoring campaigns (with multiple cyclic reoccurrences) using a larger time period covering multiple fluctuation cycles and after applying a proper block cross-validation procedures. Local models can be constructed for defined short-time

monitoring data only after investigation of systematic changes in water quality. Other multivariate techniques like recursive PCA or other specific PLS algorithms for adaptive data modeling (Dayal and MacGregor, 1997; Rosen and Lennox, 2001) have been proposed to overcome problems in large time scales studies.

Worse results were obtained in the estimation of TOC concentrations when PLS models were externally validated using the 24-h spectral values of data set D_3, or even when these data sets were analyzed simultaneously with the data from the 7-days monitoring (spectral part of data set C_3). Relative prediction errors were in this case between 10% and 22% in external validation and PLS models never explained more than 30% of the TOC concentrations variance. This can be also attributed to the relatively small variability of TOC concentrations in the calibration data set (TOC concentrations varied only between 10 and 12 mg/L during the investigated 24-h of monitoring) compared to the other data sets. This small variance was not sufficient to represent the complexity and nature of organic matter and its relationship with spectral feature changes (Thomas and Burgess, 2007). Strong matrix effects due to water turbidity changes and the presence of cross sensitivities due to absorption at the selected wavelengths of many other background compounds (Langergraber et al., 2003), could be among the possible explanations for the model limitations encountered in this work.

4. Conclusions

Chemometric data analysis tools like Principal Component Analysis (PCA) and Partial Least Squares Regression (PLS) are suggested to be efficient tools to explore, analyze and model relationships among different water quality parameters in wastewater treatment plants (WWTP). In this work PCA and PLS methods have been applied to the analysis of different type of physicochemical parameters from standard laboratory methods and from automatic multi-parameter monitoring systems including UV absorption spectral data. These multivariate data analysis tools could be useful to monitor the performance of large- or short-scale continuous processes in WWTP. Temporal (seasonal, weekdays-weekend, diurnal urban activity) and spatial changes (influent-effluent changes) on water quality can be easily detected by multivariate analysis of WWTP monitoring data as well as possible disturbances like rainfall episodes and accidental water pollution events.

UV spectrophotometry is shown to be an efficient and useful complementary technique for water quality monitoring which provides rich multivariate (muliwavelength) information about WWTP processes. In this work, a new ultraviolet multi-LED optical device (based on spectral recording of six wavelengths) is evaluated as an useful diagnostic tool to detect unusual water quality disturbances in cooperation with other conventional standard laboratory methods. The results of this work have emphasized the usefulness of absorbance values at 240, 250, 254, 260, 275 and 365 nm for a correct of water quality assessment in WWTP effluents.

UV–VIS spectrophotometry of synthetic water mixtures in laboratory conditions at the concentrations encountered in real samples has shown that it is possible to perform the simultaneous estimation of concentrations of some common water quality parameters, such as concentrations of dissolved organic matter, nitrates, detergents and phenols using full range UVVIS spectra. VIP scores technique together with PLS2 allowed optimal selection of a reduced set of wavelengths without a significative lost of predictive power. Eight wavelengths were finally selected with relative concentration prediction errors in external validation samples of around 2–3% for nitrates, detergent and phenols, and of around 15% for dissolved organic matter in laboratory experiments. When the method was applied to WWTP effluents, NO_3 concentration values were highly correlated with laboratory reference values, with prediction errors of around 12.5%. TOC concentrations could also be predicted in WWTP effluent with relative errors between 10 and 22% depending on the investigated period. However, some uncertainties and model limitations should still be considered, such as reference concentration ranges, water matrix effects, cross sensitivity and the need for longer-term monitoring campaigns. More research is still needed to clarify these questions and to evaluate the full potential of UV–VIS optical devices for water quality assessment of WWTP processes.

Acknowledgements

The AMS-LED technology was developed with the financial help of RD09-1-0001 SAFENATUR Project, granted by ACC1Ó, the Catalan Agency for enterprise innovation and competitiveness (http://www.acc10.cat/).

S. Platikanov also acknowledges the Ministerio de Economía y Competividad, Spain (grant CTQ2012-38616-C02-01) for his PhD contract.

This work has been co-financed by European Union through the European Regional Development Fund (ERDF) and partly supported by the Generalitat de Catalunya (Consolidated Research Group: Water and Soil Quality Unit 2009-SGR-965).

References

Anderson, R.J., Bendell, D.J., Groundwater, P.W., 2004. Organic Spectroscopic Analysis. The RSC, Cambridge, UK.
APHA, AWWA, WEF, 1998. Standard Methods for the Examination of Water and Wastewater, twentieth ed. American Public Health Association, Washington, DC.
Box, G.E.P., Draper, N.R., 1987. Empirical Model-building and Response Surfaces, A Wiley-Interscience Publication, first ed. Canada John Wiley and Sons. 34–57, 304–381, 423–474.
Chong, I., Jun, C., 2005. Performance of some variable selection methods when multicollinearity is present. Chemom. Intelligent Laboratory Syst. 78, 103–112.
Clement, R., Yang, P., 2001. Environmental analysis. Anal. Chem. 73, 2761–2790.
Dahlén, J., Karlsson, S., Bäckström, M., Hagberg, J., Pettersson, H., 2000. Determination of nitrate and other water quality parameters in groundwater from UVVIS spectra employing partial least squares regression. Chemosphere 40, 71–77.
Dayal, B.S., MacGregor, J.F., 1997. Recursive exponentially weighted PLS and its application to adaptive control and prediction. J. Proc. Cont. 7 (3), 169–179.
Esbensen, K., 2002. Multivariate Data Analysis – in Practice, fifth ed. CAMO Process AS, Oslo.
Jolliffe, I.T., 1986. Principal Component Analysis. Springer, New York.
Jørgensen, K., Segtnan, V., Thyholt, K., Næs, T., 2004. A comparison of methods for analysing regression models with both spectral and designed variables. J. Chemom. 18, 451–464.
Langergraber, G., Fleischmann, N., Hofstädter, F., 2003. A multivariate calibration procedure for UV/VIS spectrometric quantification of organic matter and nitrate in wastewater. Water Sci. Technol. 47 (2), 63–71.
Langergraber, G., Gupta, J., Pressl, A., Hofstaedter, F., Lettl, W., Weingartner, A., Fleischmann, N., 2004. On-line monitoring for control of a pilot-scale sequencing batch reactor using a submersible UV/VIS spectrometer. Water Sci. Technol. 50 (10), 73–80.
Ouali, A., Azri, Ch, Medhioub, Kh, Ghrabi, A., 2009. Descriptive and multivariable analysis of the physico-chemical and biological parameters of Sfax wastewater treatment plant. Desalination 246, 496–505.
Rosen, C., Lennox, J.A., 2001. Multivariate and multiscale monitoring of wastewater treatment operation. Water Res. 35 (14), 3402–3410.
Singh, K., Malik, A., Singh, V.K., Basant, N., Sinha, S., 2006. Multi-way modeling of hydro-chemical data of an alluvial river system—A case study. Anal. Chim. Acta 571, 248–259.
Thomas, O., Burgess, C., 2007. UV-visible Spectrophotometry of Water and Wastewater. Elsevier.
Thomas, O., Gallot, S., Mazas, N., 1990. Ultraviolet multiwavelength absorptiometry (UVMA) for the examination of natural waters and wastewaters. Part II. Determination of nitrate. Fresenius J. Anal. Chem. 338, 238–240.
Thomas, O., Theraulaz, F., Cerda, V., Constant, D., Quevauviller, P., 1997. Water quality monitoring. Trend. Anal. Chem. 17 (7), 419–424.
Vaillant, S., Pouet, M., Thomas, O., 2002. Basic handling of UV spectra for urban water quality monitoring. Urban Wat. 4, 273–281.
Wold, S., Albano, C., Dunn, W.J., Esbensen, K., Hellberg, S., Johansson, E., Sjöström, M., 1983. Pattern recognition: finding and using regularities in multivariate data. In: Martens, H., Russworm, H. (Eds.), Food Research and Data Analysis. Applied Science Publishers, London, pp. 147–188.

Results and Discussion

- PCA results of one WWTP physicochemical data (dataset B in Table 3)

The analysed dataset contained information about 18 water quality parameters, routinely measured over 138 days during one year in TRARGISA WWTP using standard laboratory methods. These 18 parameters characterized the water quality in the influent and in the effluent of the TRARGISA WWTP.

PCA model of this dataset was built with 6 components, explaining about 68% of the total data variance. Such a high number of PCs suggested complex behavior of these parameters in relation to the dynamics of the influent-effluent water quality relationships and of the treatment processes.

The analysis of 1) the scores plot for PC1 (Fig 1a, Article 5) and 2) the PC1 vs PC2 loadings plot (Fig. 1b, Article 5) revealed the seasonal variation of the water quality. The water quality trend followed a smooth sigmoid shape over the entire period under investigation (i.e., one year). Positive scores on PC1 were associated with samples from spring, autumn and winter seasons and negative scores - for samples from summer. Additionally, positive loadings on PC1 were observed for almost all physical-chemical parameters (see Figure 1b), which suggested that their concentrations were above the average during spring, autumn and winter seasons. In contrast, concentration values of these parameters below the average were found for the summer period. These findings could be explained due to the higher plant activity in spring, autumn and winter and due to the reduced plant activity during summer. Reduced urban activity in summer compared to the other seasons resulted in reduced wastewater inflow during summer season. A more detailed examination of the score plots revealed that the water quality variation over weekends and week days also exhibited short-term fluctuations in the water quality trend line (see Figure 1).

The comprehensive analysis of PC1 and PC2 loadings plot (see Figure 1b) further revealed the effect of influent-effluent variation in water quality due to the role of WWTP. The parameters monitored in the influent were distinguished from the parameters monitored in the effluent. An inverse correlation was detected between water quality parameters measured at the plant influent and such measured at the plant effluent. Therefore, the plant treatment effect could be detected from this plot.

PCA loadings plot was consistent with previous literature findings regarding the increase of nitrate concentrations in the effluent due to WWTP activity (Thomas and Burgess, 2007).

- PCA results of the one month WWTP data set (dataset B in Table 3)

Another dataset had water quality physico-chemical parameters collected by an automatic monitoring station (AMS) which was installed at the water treatment plant exit. PCA of the water quality data in the effluent collected during one month period, every 20 minutes, was performed (see Tables 1 and 2, Article 5). Using three principal components, 88% of the total variance was explained, where PC1 accounted for 45.9% of the total variance. Based on the model performance, PCA explained a higher data variance (i.e., 88% using 3PCs) in comparison to the previous model (i.e., 68% using 6PCs with data generated by standard laboratory methods).

The visualization of scores and loadings for the three PCs is displayed on Figure 2 (Article 5). PC1 was dominated by the notorious diurnal, short-termed cyclic fluctuations of water quality. The effect of Easter holiday vacations (with a decrease of the urban activity) and several episodes of rainfalls in Girona area were detected in the water quality trend over the sampled period. This result can be explained by the observed increase of dissolved oxygen concentration and redox potential values in the water effluent (due to rainfalls in that area), as well as by the significant decrease in concentrations of the main pollution parameters such as COD and TOC (attributable to the reduced urban activity during holidays).

PC2 captured morning-midnight fluctuations trend. It was detected that such fluctuations were largely explained by the variations in the three water quality parameters, namely water temperature, pH, red-oxy potential and dissolved oxygen. The last two parameters showed higher concentration values in the morning as compared to the midnight hours.

PC3 accounted for the operational pattern with cyclic morning-afternoon recurrence and was correlated with WWTP processing activities. The turbidity parameter was found to be the most influential for this component. Furthermore, the results indicated that turbidity values were higher in the morning as compared to the rest of the day. WWTP operational procedures, which include the release of large volumes of treated water, were usually implemented in the morning and hence new wastewater entered the plant at this point in time, increasing water turbidity. In contrast, the water turbulence decreased during the day, when the water incoming-outgoing process was more stable.

- PCA results of seven days WWTP data sets. Comparison of the obtained results from data sets generated by a new automatic multiparameteric station with optical probe and two standard laboratory techniques.

The analysis included a comparison between PCA results obtained using three datasets simultaneously collected over the same week, namely 1) by the UVVIS laboratory instrument (Dataset C1, Table 1), 2) using standard laboratory methods (Dataset C2), and 3) by the AMS-LED (Dataset C3).

The obtained three PCA models using these data sets were found to be informative, explaining above 60-98% of the variance of the corresponding dataset. PCA modelling of laboratory UVVIS instrumental data explained more than 98% of the variance with two PCs. The PCA model using physicochemical data generated in laboratory after the application of APHA methods was able to explain 63% of the total variance, and the PCA model using the dataset obtained with AMS-LED system - 67% of the variance using first two PCs. The results of scores and loadings in these three cases using two PCs are visualized in Figure 3. The comparative analysis of the three obtained plots suggested that the five working weekdays could be clearly distinguished from the weekend.

The examination of the loadings plot of spectral data (Dataset C1) revealed a shoulder at 220nm and a band maximum at 226nm (see Figure 3b) as the most important spectral regions. The absorption at 220nm was attributed to organic matter and nitrates (APHA, 1998) and the peak at 226nm - to detergents according to the existing literature (Thomas and Burgess, 2007). Therefore, the relationship between concentrations of organic matter and detergents is highly relevant in order to describe the changes of water quality during weekdays.

The analysis of loadings plot of laboratory data (Dataset C2, see Figure 3d) suggested that pH and nitrate parameters, which presented an inverse correlation in the data, were the most influential parameters. Weekdays were characterized by high concentrations of nitrates and were different from weekends, when the intensive household activity resulted in an increase of pH.

Similar conclusions have been derived from the analysis of PCA results obtained using the AMS-LED instrument. Weekends were distinguished from weekdays largely because of the inverse correlation between the two groups of parameters. Whereas water quality during the weekend was characterized by higher concentrations of dissolved

oxygen and redox potential and higher pH values, weekdays were characterized by higher UV absorption values at 240-275 nm and BOD parameters, suggesting a higher organic contribution.

By comparing the three techniques, we concluded that high discharges of organic matter (probably with an industrial origin) were common for weekdays; in contrast, water quality was dominated by compounds with urban or household origin during weekends. In general, the quality of effluent water was better during weekends as compared to weekdays.

- PCA results of the twenty-four hours WWTP data sets. Comparison of the obtained results from data sets generated by a new automatic multiparameteric station with optical probe and two standard laboratory techniques.

In this work, we compared PCA results from three datasets simultaneously collected over a period of twenty-four hours using: 1) UVVIS laboratory instrument (data set D1 of Table 1 and 2) standard laboratory methods (data set D2), and 3) AMS-LED (data set D3).

PCA scores using a UVVIS laboratory instrument (Figure 4a, Article 5) were similar to PCA scores using data generated using standard laboratory methods (Figure 4c) and data collected using the new AMS-LED system (Figure 4e). In the three cases, PCA models captured a continuous daily-night sigmoid water quality trend in the effluent. In the three score plots, the night time hours had positive scores and, in contrast, the morning-afternoon hours - negative scores. An inflection point was detected where there was a significant change of the water quality (at 17.00h in the afternoon monitored with the UVVIS instrument). Such an inflection point was also found when using the other two score plots (i.e., APHA, or standard laboratory methods, and AMS-LED data sets). Additionally, the analysis revealed a peak with a maximum value of the trend lines around 00.00h (midnight), which is present on the three score plots. The analysis of the UV-VIS spectral data set found that the maximum absorption value was at 220nm. Such a strong absorption at wavelength can be most likely attributed to nitrates and dissolved organic matter (APHA, 1998). In general, the sharp change of the trend line could be associated with the rapid increase of concentrations of nitrate and organic matter at this midnight time.

As displayed on Figure 4d, the corresponding loadings plot for APHA data set suggested that COD, BOD_5, TOC and PO_4 are the most influential parameters captured in PC1. This result indicates the presence of a pollution event occurring at midnight, when organic matter was significantly higher.

The analysis of the scores plot of Figure 4e for the data set collected by the AMS-LED system additionally detected such pollution event at midnight, as suggested by the specific absorbance coefficient (SAC). All six measured wavelengths and the BOD parameter confirmed an increased level of organic matter discharge at midnight. Therefore, the interpretation of LED absorptions of the six wavelengths was found to be particularly useful in immediately detecting such pollution events.

In general, we concluded that, independently of the methods applied (i.e., APHA methods, UVVIS or AMS-LED new system), the one-day and weekly trends in water quality monitoring of WWTP activity were similar. This result confirmed that the newly proposed AMS-LED system was efficient and reliable. A longer time-scale investigation however would be required to generalize our results and verify the potential use of the AMS-LED system in quality control.

- PLS prediction of four target water quality parameters in synthetics mixtures by UV absorption at selected wavelengths

In Article 5, PLS2 models were additionally developed using meancentered UV spectral data with five latent variables (LVs), explaining more than 90% of the total variance. Following the performance of a PLS2 model for the four selected target compounds (namely, nitrates, organic matter, detergent, and phenols) and in the presence of kaolin (interference), VIP scores were used to select the most important wavelengths.

Figure 5 (Article 5) displays the VIP scores in the prediction of the parameters. VIPs were considered significant when they were above the threshold value of 1. Significant VIPs were found in the three spectral ranges: a) 190-210, b) 220-260, and c) 270-276 nm. The analysis included a visual preselection, when the most characteristic peaks and spectral inflection points were detected. Spectral region between 190-200nm was disregarded, because VIP scores are considered to be unreliable for such low UV wavelengths. The final selection included the following ranges: 201, 205 (nitrate absorption), 226 (detergent absorption), 237, 254 and 285nm (dissolved organic matter absorption), 270 and 276 nm (phenol absorption wavelengths.

Following this preselection, a new PLS2 model was recalibrated using eight wavelengths. The new recalculated PLS2 models were able to explain more than 90% of the total variance of all target concentrations.

The comparison between predicted and actual concentrations in external validation suggested that they were highly correlated. Newly recalculated PLS2 models had good predictive properties (see in Figure 6 the following model parameters: R^2, RMSEC, RMSECV and bias values). By doing so, the number of wavelengths used to estimate the models was significantly reduced without losing prediction power. More specifically, DOM concentrations were predicted with 26% relative error in external validation; nitrates, detergent and phenol concentrations - with 3 and 4% of relative prediction error in external validations. These figures of merit confirmed that PLS2 model had good predictive properties and that the preselection method using visual inspection of VIP scores additional was a reliable approach.

- PLS prediction of NO_3 and TOC water quality parameters in real samples using AMS-LED spectral data

In Article 5, multivariate PLS1 models were built to quantify individually NO_3 and TOC using the spectral part of data sets collected by AMS-LED (see datasets C3 and D3 of Table 1).

The prediction models of NO_3 performed better than the models of TOC. The best prediction model was obtained for NO_3 using 24-hours monitoring data. The model could predict up to 70% of NO_3 concentration changes, when using the first latent variable and six wavelengths. Relative prediction error in external validation was found to be 3.5%. Alternative data sets rearrangements were tested prior to modelling and, in such recalculated models, the maximum prediction error in external validation could not exceed 12.5%, suggesting that NO_3 prediction was feasible using the AMS-LED system. Furthermore, it was confirmed that optimal modelling of NO_3 concentrations could be achieved using only six AMS-LED wavelengths (240, 250, 254, 260, 275 and 365nm).

Predicting TOC concentrations in external validation samples was not as good as for NO_3. Relative prediction errors were in the range between 10% and 22% in the external validations. Additionally, PLS models had never explained more than 30% of TOC concentrations variance. This result can be attributed to the relatively small variability of the reference TOC concentrations in the calibration data set (i.e., TOC

concentrations varied only between 10-12 mg/L in 24 hours). Additionally, there were strong matrix effects related to water turbidity changes and cross sensitivities associated with the absorption of many other background compounds at the selected wavelengths, probably explaining these model limitations.

In situ predictions of TOC and NO_3 were not as accurate as expected. Results could have significant implications for future work. Perhaps the most important one is that more observations should be required and a broader concentration range of the target compounds should be considered for optimal PLS modelling. Furthermore, data sets should be carefully split for calibration and external validation in order to avoid overfitting problems. Continuous systematic changes in water quality further make the selection of variables for analysis and updating difficult. In our case, the investigated data sets were relatively small and hence such difficulties were overcome by a random selection of samples. In general, PLS models have to be calibrated using data that covers at least a full-range of cyclic fluctuations in the WWTP. In case of larger monitoring campaigns (with multiple cyclic reoccurrences), a larger time period covering multiple fluctuation cycles should be used.

3.3 Chemometrics methods applied to water taste related data in sensory science

This block includes an article that deals with the application of multivariate data analysis on sensory data.

The paper complies with the main objective of the Thesis in the following aspects:

a) Discovery of the most influential physicochemical water parameters associated with the panellists' taste liking of mineral bottled and tap waters by means of chemometrics methods.

3.3.1 Article 6 – Platikanov, S., Garcia, V., Fonseca, I., Rullan, E., Devesa, R., Tauler, R., *Influence of minerals on the taste of bottled and tap water: A chemometric approach.* Water Research 47 (2013) 693-704.

Introduction

The Barcelona Metropolitan Area (BMA) is supplied with drinking water primarily from two surface water resources – the Llobregat River and the Ter River. As a consequence of the Mediterranean climate, such water resources are cyclically exposed to serious droughts, particularly when the raw water cannot meet the demand in the area. In terms of operational management, the conditions are even worst when there is the pollution effect of mining and industrial discharges alongside the two river basins, seriously affecting the incoming quality of the fresh water.

Membrane technologies and desalination of sea water for drinking water have been suggested to provide new alternative resources which may guarantee the demand for cleaner water. Although these technologies are very efficient in removing undesirable chemical species, organic matter and pathogens, a permanent remineralization is required to improve the organoleptic properties of the supplied drinking water, thus sustaining and minimizing the corrosive effect of many minerals.

The quality of drinking water is perceived by the general public. For instance, it is well known that the taste of water depends on the chemical composition of the salt content, where cations and anion contribute to a different extent and interact by

synergism and antagonism. Water taste preferences in the BMA distribution network were investigated for various years. However, our knowledge regarding the effect of particular chemical species on the overall public satisfaction is largely limited. Therefore, panels testing procedures are able to provide valuable information regarding water taste.

This article is based on two independent case studies, which propose PCA and PLS methods to discover the most influential physicochemical parameters associated with the overall satisfaction of water taste. Two data sets were used in the analysis. The first data set included 20 water samples, namely 11 commercial bottled-mineral water samples and 9 artificially generated blends on the basis of bottled-mineral water samples with the objective to obtain water samples with a desired chemical profile. The second data set included 25 samples: 13 samples of drinking tap waters and 12 samples of bottled waters. In both tests, water samples were independently and blindly tasted by a set of trained panelists in several sessions. The panelists were asked to express their overall liking per water sample using score values in the range of 0 (worst flavor) and 10 (best flavor).

Both sets of water samples were analyzed in laboratory for thirteen different physicochemical parameters and the residual chlorine (in the case of tap waters). Among the investigated parameters were: sodium, potassium, calcium, magnesium, silica, conductivity at 20 ºC, pH, bicarbonate, chlorides, nitrates, sulphates, and free residual chlorine.

In both studies, the analysis suggested that, on average, panelists' overall liking was correlated with the physicochemical properties of water samples. The chemometrics methods included in Article 6 were the traditional analysis of variance (ANOVA), Principal Component Analysis (PCA), and Partial Least Squares regression (PLS). More specifically, the analysis served the following purposes. Two-way ANOVA was employed to evaluate the main effects on panelists' preferences of the two considered factors, namely the water type and the panelists, as well as their possible interaction. PCA was applied to the data containing the physicochemical parameters and the panelists' mean liking, providing simplified models that were able to explain a main part of the total data variance. To examine the main features underlying the panelists' ratings, PLS was applied to the panelists' mean liking (y variable) and the physicochemical data (**X** variables).

A rating test was used to assess the overall liking of water samples. Panelists were asked to rate the flavor of samples on a 0-10 scale, where 0 stands for an extremely bad flavor and 10 – for an excellent flavor.

Because the two sensory tests were actually rating tests, it was required to decide which statistical parameter could represent the panelists overall liking. Two statistical parameters were evaluated, namely the adjusted mean and the median. All calculations and figures visualization regarding PCA and PLS analyses were compared. A significant difference between results obtained using any of the two parameters was not found. The adjusted mean was therefore selected and used in further analysis as a statistical parameter representing panelists' mean liking.

WATER RESEARCH 47 (2013) 693–704

Available online at www.sciencedirect.com

SciVerse ScienceDirect

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

Influence of minerals on the taste of bottled and tap water: A chemometric approach

Stefan Platikanov [a,1], Veronica Garcia [b], Ignacio Fonseca [c], Elena Rullán [c], Ricard Devesa [b], Roma Tauler [a,*]

[a] Department of Environmental Chemistry, IDAEA-CSIC, Jordi Girona, 18-26, 08026 Barcelona, Spain
[b] Aigües de Barcelona (Agbar) Laboratory, General Batet, 5-7, 08028 Barcelona, Spain
[c] CETAQUA, Water Technological Center, Ctra. Esplugues, 75, 08940 Cornellà de Llobregat, Barcelona, Spain

ARTICLE INFO

Article history:
Received 13 June 2012
Received in revised form
10 October 2012
Accepted 24 October 2012
Available online 9 November 2012

Keywords:
Taste
Flavor
Bottled water
Tap
Sensory analysis
PLS
PCA

ABSTRACT

Chemometric analysis was performed on two sets of sensory data obtained from two separate studies. Twenty commercially-available bottled mineral water samples (from the first study) and twenty-five drinking tap and bottled water samples (from the second study) were blind tasted by trained panelists. The panelists expressed their overall liking of the water samples by rating from 0 (worst flavor) to 10 (best flavor). The mean overall score was compared to the physicochemical properties of the samples. Thirteen different physicochemical parameters were considered in both studies and, additionally, residual chlorine levels were assessed in the second study. Principal component analysis performed on the physicochemical parameters and the panelists' mean scores generated models that explain most of the total data variance. Moreover, partial least squares regression of the panelists' sensory evaluations of the physicochemical data helped elucidate the main features underlying the panelists' ratings. The preferred bottled and tap water samples were associated with moderate (relatively to the parameters mean values) contents of total dissolved solids and with relatively high concentrations of HCO_3^-, SO_4^{2-}, Ca^{2+} and Mg^{2+} as well as with relatively high pH values. High concentrations of Na^+, K^+ and Cl^- were scored low by many of the panelists, while residual chlorine did not affect the ratings, but did enable the panel to distinguish between bottled mineral water and tap water samples.

1. Introduction

It is well known that the taste of water depends on the chemical composition of the salt content, with both cations and anions contributing in different ways and interacting through synergism and antagonism (Burlingame et al., 2007). In addition to the dissolved inorganic salts (total dissolved solids, TDS), some volatile organic compounds can be detected through retro-nasal mechanisms when drinking water (Dietrich, 2009). Therefore, the global perception of water is considered more of a flavor than a taste (Dietrich, 2006).

The mineral and chemical contents of bottled natural mineral water are determined by the composition of the rocks from which it is extracted and by geochemical processes (van der Aa, 2003). Moreover, potable tap water is also characterized by its specific chemical (mineral and organic) content (Meng and Suffet, 1997) in relation to the incoming raw water and the disinfection procedures implemented. The latter may

* Corresponding author.
E-mail addresses: splqam@iiqab.csic.es (S. Platikanov), Roma.Tauler@idaea.csic.es (R. Tauler).
[1] Tel.: +34 645257566.
0043-1354/$ – see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.watres.2012.10.040

add other chemical compounds at significant concentrations (such as residual chlorine), which also contribute to the final flavor of drinking water.

Traditionally, water distribution companies have tried to improve the flavor of water through conventional water treatment procedures. Their efforts have focused on producing water with low organic matter content and minimum levels of residual disinfectants that are just enough to sanitize water. With the introduction of membrane technologies, it is now possible to significantly reduce organic matter levels and water mineralization (Bruchet and Lainé, 2005), thus improving flavor. The remineralization stage, which is commonly applied to treated water, also plays an important role in final taste (Devesa et al., 2010).

The assessment of improvements in the flavor of drinking (tap) and bottled water needs sensory experiments in which water samples are systematically presented to trained panelists or random consumers, who test them in various ways (Naes and Risvik, 1996). Recent studies by Teillet et al. (2010a,b) confirmed that water flavor assessments are complex and should be performed very carefully. Using a free sorting task technique, these authors concluded that consumer acceptance of water was mostly driven by the mineral content and that medium-mineralized water was preferred by regular consumers; however, no conclusions were reached about what specific components were responsible for their preferences. The use of a trained panel of assessors is appropriate for a reliable sensory description and assessment of unknown water sample preferences.

Carefully-designed experiments and analysis of variance (ANOVA) are the starting steps in sensory studies (Hibbert, 2009). Data sets are usually multivariate and multiway, having as dimensions the number of samples multiplied by the number of sensory attributes, multiplied by the number of panelists. Usually, sensory data are noisy (between people and along time for the same person) (Hibbert, 2009). The application of ANOVA is extremely useful in studying the sources of data variance. It helps to test the panelist's performance and statistically evaluate all factors that influence responses to the sensory attributes, i.e., the physicochemical parameters in our case studies (Naes and Risvik, 1996). Since these responses can be influenced by two or more factors as well as their interactions, the proper methodology in this case is a two-way ANOVA or, when more than two factors are simultaneously tested, an N-way ANOVA (Peña Sánchez, 1994).

Principal component analysis (PCA) has already been shown to be useful for correlating chemical and sensory data in drinking water samples from a distribution system (Meng and Suffet, 1997). PCA (Jolliffe, 2002) has been applied to reveal the most important patterns in the physicochemical parameters that correlate with the panelists' ratings. Mallevialle and Suffet (1987) introduced background information on chemical/sensory correlation methods and showed general results from correlation studies. Suffet et al. (1989) used a similar statistical method, factorial correspondence analysis, to link chemical and sensory data. Generally, the choice of method depends on the type of sensory data, e.g., whether the data are of a ranking or rating (giving evaluation scores) type, a categorical or continuous character, or even concatenated in multiblocks (Stanimirova et al., 2011).

Partial least squares regression (PLS) is a well-known and useful tool in consumer preference analysis (Lengard and Kermit, 2006). The complex relationship between the sensory panel rates and the physicochemical parameters of the mineral and tap water will be discovered using PLS (Geladi and Kowalski, 1986). Recently, the variable importance in projection (VIP) scores has been proposed as useful tool for interpreting PLS models (built from several latent variables) (Chong and Jun, 2005). The interpretation of VIP scores can be employed to evaluate the importance of each water physicochemical parameter in the final PLS projection.

This paper aims to discover the most influential physicochemical parameters associated with the overall score of water flavor in two separate sensory studies performed using selected bottled mineral and tap water samples with different mineral contents and origins.

2. Material and methods

2.1. *Water samples*

In the first study (A), 11 bottled mineral water samples, commercially available in Spain, were collected. In addition, 9 new samples were obtained by blending two water samples or by diluting with purified water. Therefore, this study was performed with 20 water samples. The high number of samples considered, and the process of blending and diluting were decided in order to cover a broad range of TDS and mineral composition, as well as the percentages of the different cations and anions. Therefore, the study included water samples with very different mineralization levels and percentages of sodium, calcium, magnesium, chloride, bicarbonate and sulfate, i.e., the most relevant species driving the taste of water.

In the second study (B), 25 water samples were considered: 12 bottled waters selected from the first study and another 13 tap water samples from different resources and networks in Catalonia, northwest Spain. The tap water samples were selected with the same intent as in the first study: covering a wide range of mineralization and chemical composition types.

The results of the physicochemical compositions of the water samples used in the studies A and B are presented in Tables 1 and 2, respectively. Analytical analyses of the samples were made in the accredited laboratory of the Aigües de Barcelona Company. Water blends and dilutions were allowed for 48 h of equilibration before analysis. Sodium, potassium, calcium, magnesium and silica levels were analyzed by inductively coupled plasma optical emission spectrometry (ICP-OES) (Perkin Elmer Optima 4300 DV). Conductivity at 20 °C, pH and bicarbonate levels were determined by a robotic titrosampler (Metröhm modules 855 and 856). Chloride, nitrate and sulfate concentrations were analyzed by ionic chromatography (Dionex ICS-2000). TDS (dry residue at 180 °C) levels were measured by gravimetry. Free residual chlorine was analyzed by the classical DPD colorimetric method. All these analytical determinations, except for the bicarbonates and TDS that are not included in

Table 1 – Identification and mineral contents of the bottled water samples used in the present study.

Samples	Labels[a]	Conductivity (μS/cm)	TDS (mg/L)	Cl^- (mg/L)	SO_4^{2-} (mg/L)	NO_3^- (mg/L)	HCO_3^- (mg/L)	Ca^{2+} (mg/L)	Mg^{2+} (mg/L)	Na^+ (mg/L)	K^+ (mg/L)	pH	Si (mg/L)	Mean liking
1	LM·min$_1$[b]	30	25	3	3	3	6.3	1.7	0.5	3.1	0.4	6	6.9	5
2	LM·min$_2$	38	26	0.67	1.26	3.5	18	5	1	1.3	0.3	6.7	8.9	5.5
3	$NaHCO_3$·min$_1$	278	193	17	11.6	0.36	145.42	12.4	0.8	51.5	1.5	7.3	16	5.9
4	$NaHCO_3$·min$_2$	539.2	347	27.4	21.3	0.7	285.1	22.8	1.6	87.7	4	7.9	32.2	5.4
5	$Ca(HCO_3)_2$·min$_1$	201.4	121	4.2	12.2	0.85	124.2	21.3	12.6	2.7	0.5	8.1	5	6.2
6	$Ca(HCO_3)_2$·min$_2$	294	213	16.2	16.4	3.1	149	38.5	9.7	13.2	1.1	7.8	4.2	6.6
7	$Ca(HCO_3)_2$·min$_3$	375	262	7.8	21.9	1.9	285	56.9	25.5	5.3	1.1	8	7.5	5.7
8	$Ca(HCO_3)_2$·min$_4$	446	278	2.8	20.1	2.9	307	75.6	19.4	1.6	0.4	7.6	0.7	6.5
9	$Ca(HCO_3)_2$·min$_5$	602	355	4.14	12.7	1.9	415	77.7	40.3	1.6	0.4	7.5	3.2	4.9
10	$Ca(HCO_3)_2$/ $CaSO_4$·min$_1$	220	205	2.96	57.9	0.46	74	36.3	7.8	0.6	0.4	7.5	1.7	6.6
11	$Ca(HCO_3)_2$/ $CaSO_4$·min$_2$	388.4	287	5.5	109	1.1	115.9	70.4	16.1	1.1	0.8	7.9	4.8	6.5
12	$Ca(HCO_3)_2$/ $CaSO_4$·min$_3$	993	844	7.9	328.9	4.3	399	203.8	43.1	5	1.9	7.5	9.67	5.7
13	$CaSO_4$·min$_1$	220	189	2.7	86.8	0.16	32.5	35.1	6	0.5	0.3	7.6	0.53	6.5
14	$CaSO_4$·min$_2$	395.5	314	2.9	166.7	0.32	63.98	68.9	11.9	1.1	0.6	7.8	1.8	6.8
15	$CaSO_4$·min$_3$	786	662	3.7	385.8	1.1	122.9	162.9	27.6	2.3	1.1	7.3	3	6.3
16	$CaSO_4$·min$_4$	1157	1083	5.7	596	1.5	188	241.6	41	3.8	1.5	7.61	5.2	5.4
17	NaCl·min$_1$	274	165	74.3	1.2	3.5	25.9	6.3	1	121.6	0.4	6.9	10.9	4.9
18	NaCl·min$_2$	635.3	389	193.86	1.45	3.94	20.84	6.2	0.9	31.3	0.4	7.01	9.6	4.3
19	NaCl/$NaHCO_3$· min$_1$	289	195	49.6	6.49	1.8	85.6	9.6	1	52.8	0.9	7.2	13	5.6
20	NaCl/$NaHCO_3$· min$_2$	591	373	112.2	12.1	2.49	164.4	12.2	1.3	116	1.7	7.2	20.8	4.6
21	Mean	437.6	326.3	27.23	93.64	1.94	151.4	58.3	13.5	25.2	0.99	7.4	8.3	5.7
22	Maximum	1157	1083	193.86	596	4.3	415	241.6	43.1	121.6	4	8.1	32.2	6.8
23	Minimum	30	25	0.67	1.2	0.16	6.3	1.7	0.5	0.5	0.3	6	0.53	4.3

a The water type was identified by the most characteristic mineral contents.
b Labels LM1 and LM2 identify two water samples with very low TDS concentrations.

the EC Drinking Water Directive (98/83/EC), are ISO17025-accredited by ENAC, the accreditation body in Spain.

The same batch was used for each of the mineral water samples tested. Bottles were stored at room temperature and in a dark place free of odors. Tap water samples were taken in established sampling points of the routine monitoring supply systems. Samples were taken in glass amber bottles, previously cleaned with a procedure appropriate for sensory analysis. Bottles were refrigerated until a few hours before the tasting sessions. The water samples were not dechlorinated and no other treatment was applied.

Tables 1 and 2 show the collected mineral and tap water samples and their mineral composition described as physicochemical parameters.

2.2. *Sensory analysis*

Both studies were carried out by trained tasters from the AGBAR panel (Devesa et al., 2004), which is trained according to the flavor profile analysis (FPA) method (Devesa et al., 2007). The sensory test used in this study, the rating test, is not of a descriptive nature (the taste-and-odor wheel is not used) and therefore, it does not require a high degree of training. Nonetheless, the use of a trained panel gives additional value to the study.

Samples were always presented to the panelists as coded and without any information that could influence their appreciation. The tasting took place in a room specifically intended for this purpose, comfortable and free from interfering odors. Water samples were served at 25 °C in 200 mL transparent glasses, which were filled up to one-third of their volume. Panelists were allowed to spit out remaining water after its testing. Members worked individually and no discussion took place after the session. The water was presented in a randomized series of five samples. Testing samples were designed to be TDS balanced, that is, they were designed to contain waters with different degrees of mineralization. The order of presentation of the samples in each set was randomized for each participant. Two series were presented in each session. Such a small number of samples was chosen for each session to avoid any risk of fatigue. In addition, for the same reason, a rest period was allowed between the two series. A blank sample consisting of mineral water defined previously as "neutral" by the panel was used compulsorily between consecutive tastings.

In the first study, 17 trained panelists composed by 6 males and 11 females, aged 20–57 years were used. In the second study, 13 trained panelists composed by 5 males and, 8 females, aged 21–57 years, were used. In both studies, each panelist had to taste each water sample in two different sessions. In each session, the same water sample was presented twice (duplicates) to each panelist, which meant that each experiment gave four results per panelist. This design meant that 8 tasting sessions were required by the first experiment and 10 by the second experiment.

Table 2 – Identification and mineral contents of the bottled and tap water samples used in the present study.

Samples	Labels[a]	Conductivity (μS/cm)	TDS (mg/L)	Cl^- (mg/L)	SO_4^{2-} (mg/L)	NO_3^- (mg/L)	HCO_3^- (mg/L)	Ca^{2+} (mg/L)	Mg^{2+} (mg/L)	Na^+ (mg/L)	K^+ (mg/L)	pH	Si (mg/L)	Cl_2 (mg/L)	Mean liking
1	LM·min_1 (low mineralization)[b]	38	26	0.7	1.3	3.5	18	5	1	1.3	0.3	6.7	8.9	0[d]	6.1
2	$NaHCO_3$·min_1	278	193	17	11.6	0.4	145.4	12.4	0.8	51.5	1.5	7.3	16	0	6.4
3	$NaHCO_3$·min_2	539	347	27.4	21.3	0.7	285.1	22.8	1.6	87.7	4	7.9	32.2	0	5.5
4	$Ca(HCO_3)_2$·min_1	201	121	4.2	12.2	0.9	124.2	21.3	12.6	2.7	0.5	8.1	5	0	6.9
5	$Ca(HCO_3)_2$·min_2	375	262	7.8	21.9	1.9	285	56.9	25.5	5.3	1.1	8	7.5	0	6.8
6	$Ca(HCO_3)_2/CaSO_4$·min_1	388	287	5.5	109	1.1	115.9	70.4	16.1	1.1	0.8	7.9	4.8	0	6.5
7	$Ca(HCO_3)_2/CaSO_4$·min_2	993	844	7.9	328.9	4.3	399	203.8	43.1	5	1.9	7.5	9.7	0	6.4
8	$CaSO_4$·min_1	220	189	2.7	86.8	0.2	32.5	35.1	6	0.5	0.3	7.6	0.5	0	6.7
9	$CaSO_4$·min_2	786	662	3.7	385.8	1.1	122.9	162.9	27.6	2.3	1.1	7.3	3	0	5.5
10	NaCl·min_1	274	165	74.3	1.2	3.5	25.9	6.3	1	121.6	0.4	6.9	10.9	0	6.1
11	NaCl·min_2	635	389	193.9	1.5	3.9	20.8	6.2	0.9	31.3	0.4	7	9.6	0	4.9
12	$NaCl/NaHCO_3$·min_1	289	195	49.6	6.5	1.8	85.6	9.6	1	52.8	0.9	7.2	13	0	6.1
13	$NaCl/NaHCO_3$·min_2	591	373	112.2	12.1	2.5	164.4	12.2	1.3	116	1.7	7.2	20.8	0	5.5
14	LM·tap_1	50	45	6.2	4.9	0.8	22.9	5.6	0.8	4.7	0.7	7.2	7.3	0.54	5.3
15	$Ca(HCO_3)_2$·tap_1	208	128	0.1	3.8	1.7	142.1	42.7	0.2	6.2	1.4	7.5	0.2	0.59	6.1
16	$Ca(HCO_3)_2$·tap_2	500	346	20.3	17.9	8.4	307	116.7	6.3	8.6	1	7.6	12.2	0.79	5.3
17	$Ca(HCO_3)_2/CaSO_4$·tap_1	988	801	25	311	21	337.5	189.4	52.1	10.9	2.1	7.5	10.4	0.56	5.2
18	$Ca(HCO_3)_2/CaSO_4$·tap_2	1075	893	32.8	405	14.3	283.9	200	44.8	20.6	2.4	7.8	8.3	0.45	4.5
19	$NaHCO_3/NaCl/Ca(HCO_3)_2$·tap_1	660	429	107.9	74.8	6.3	158.4	64.7	13	58.7	11.3	7.3	3	0.64	4.8
20	$NaHCO_3/Ca(HCO_3)_2$·tap_1	282	183	17.4	47	4.6	116	38.3	7.8	9.7	2.5	7.6	3.4	0.66	5.9
21	$NaHCO_3/Ca(HCO_3)_2$·tap_2	418	281	30.2	54.3	8.2	164.7	59.7	12.3	18.2	3.1	7.8	4.4	0.62	5.5
22	HM·tap_1 (high mineralization)[c]	1392	895	288	149	12	241.6	102.4	29	138.9	32.2	7.9	4.6	0.28	3.6
23	HM·tap_2	1459	1009	302.5	183.6	10.8	171.3	110.9	33.3	165.1	31.5	7.7	4.7	0.52	3.6
24	HM·tap_3	1700	1346	139.8	401	27.1	362.8	272.9	80.7	140	10.5	7.4	10.4	0.33	3.7
25	HM·tap_4	2617	1983	506.5	608.4	31.4	334.2	290.1	85.6	286.8	10.1	7.2	12.4	0.63	2.6
26	Mean	678.2	495.7	79.3	130.4	6.9	178.7	84.7	20.2	53.9	4.9	7.5	8.9	0.26	5.4
27	Maximum	2617	1983	506.5	608.4	31.4	399	290.1	85.6	286.8	32.2	8.1	32.2	0.79	6.9
28	Minimum	38	26	0.1	1.2	0.2	18	5	0.2	0.5	0.3	6.7	0.2	0	2.6

a The water type was identified by the most characteristic mineral contents.
b Labels LM·min_1 and LM·tap_1 identify bottled mineral and tap water samples with very low TDS concentrations.
c The HM·tap label identifies tap water samples with very high TDS concentrations.
d Chlorine concentrations below the detection limit (<0.15 mg/L).

The rating test was used to assess the overall liking of the water samples. Panelists were asked to rate the flavor of samples on a 0–10 scale (0 corresponding to an extremely bad flavor and 10 to an excellent one).

In both studies, two parameters were used and evaluated for chemometric analysis: the adjusted mean (Kermadec et al., 1997) and the median of the global panel's or the individual panelist's score. It was found that there were no significant changes in the results when using either of the two parameters. In this paper, the adjusted mean was used as a statistical parameter for representing the mean scores.

2.3. *Data organization and software*

In both experiments, all samples were analyzed in our laboratory according to the standard methods of analysis of physicochemical parameters, which usually represent the main mineral content of any kind of water, either mineral or tap. Different data tables were generated in both experiments according to the subsequent chemometric analysis.

In the first study, several data sets were considered. The first one was a data matrix, $\mathbf{X}_1[20, 12]$, which had 20 water samples (as rows) against 12 physicochemical parameters (as columns). The global panel's mean score rates were given as a data vector, $\mathbf{y}_1(20, 1)$, after 4 testing sessions. Moreover, the 17 individual panelist's mean score rates gave the corresponding vectors (after 4 sessions) for each water sample, which were arranged in a data matrix of dimensions $\mathbf{Y}_1[20, 17]$. Chemometric analysis was performed by PCA on the two row-wise augmented data matrices, $[\mathbf{X}_1, \mathbf{y}_1]$ and $[\mathbf{X}_1, \mathbf{Y}_1]$, and PLS regression on $\mathbf{y}_1 = f(\mathbf{X}_1)$.

Analogously, in the second study, data were organized in a similar way. The 13 physicochemical parameters (the residual chlorine in tap water samples was now included) for 25 water samples were arranged in the data matrix of dimensions $\mathbf{X}_2[25, 13]$. The global panel mean score vector was now $\mathbf{y}_2(25, 1)$ and the 13 individual panelist's mean score vectors (after all 4 sessions) were collected in the data matrix of dimensions $\mathbf{Y}_2[25, 13]$. The chemometric part of the experiment consisted of PCA on the row-wise augmented data matrices $[\mathbf{X}_2, \mathbf{y}_2]$ and $[\mathbf{X}_2, \mathbf{Y}_2]$, and PLS regression on $\mathbf{y}_2 = f(\mathbf{X}_2)$.

Two-way ANOVA was used to investigate the main effects on preferences of the two considered factors, the water type and the panelists, and their possible interaction. The hypotheses being tested were whether:

- preferences for the selected water samples differed enough (they were a significant factor) based on their physicochemical properties defined by their water type.
- there were statistically significant differences among the panelists in their evaluation of water samples.
- there was a synergistic effect of the two factors (i.e., physicochemical features and the panelists' subjective ratings).

The 4 replicates obtained during the sessions for the same water sample and panelist were considered for calculating the experimental error in the analysis.

All calculations for two-way ANOVA, PCA and PLS were performed using PLS Toolbox 5.8 (Eigenvector Research, Manson, WA, USA) and MATLAB 6.5 with Statistical Toolbox (MathWorks Inc., Natick, MA, USA).

3. Results

3.1. *Descriptive statistics of the physicochemical parameters of the two sets of water samples*

Pairwise correlation coefficients between the 12 physicochemical parameters in the first study and between the 13 in the second one were calculated and are shown as correlation maps in Fig. 1. In both experiments, either relatively high positive (intense red colors) or relatively high negative (intense blue colors) correlations were observed (see Fig. 1) for most of the physicochemical parameters. In the first experiment (Fig. 1a), a cluster of parameters with high correlations were distinguished. Mg^{2+}, Ca^{2+} and SO_4^{2-} had very strong positive correlations among them and formed one of these clusters. Negative correlations were found between the

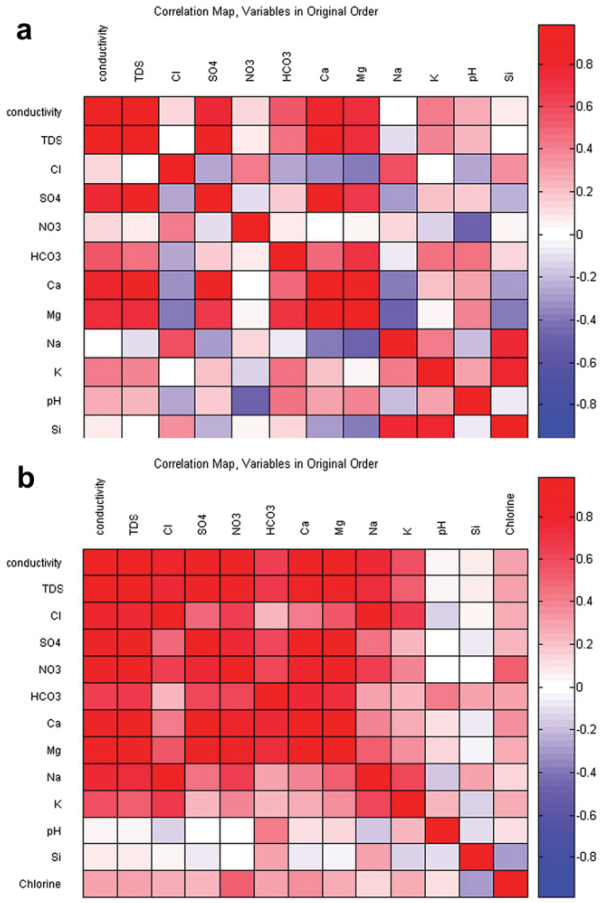


Fig. 1 – Pairwise correlation coefficient map between all considered physicochemical parameters: (a) when 20 bottled water samples were tested; (b) when 13 bottled and 12 tap water samples were tested. The more intense the red colors are, the higher positive correlations are; the more intense blue colors are, the higher negative (inverse) correlations are. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 3 – Two-way ANOVA to evaluate the effects of the two factors considered (water type and panelists) and their interaction on the trained panelists' evaluation of water samples.

Source of variance	SS[a]	df	MS	F	Prob > F
Data for 20 bottled water samples					
Water type	712.25	19	37.487	27.07	0
Panelist	541.7	16	33.856	24.45	0
Interaction	1209.39	304	3.978	2.87	0
Error (replicates)	1412.63	1020	1.385		
Total	3875.96	1359			
Data for 25 bottled and tap water samples					
Water type	1545.65	24	64.402	49.93	0
Panelist	165.92	12	13.826	10.72	0
Interaction	706.09	188	2.451	1.9	0
Error (replicates)	1257.5	975	1.289		
Total	3675.17	1299			

a SS, sum of squares; df, degree of freedom; MS, mean square; *F*, *F*-statistic; Prob > *F*, *p*-value associated with the *F*-statistic.

parameters in this first group and NO_3^-, Si, Na^+ and Cl^-, which formed another cluster. Parameters like TDS, conductivity, bicarbonates and pH showed positive or negative correlations with some of the parameters belonging to one of these clusters.

In the second experiment (Fig. 1b), strong positive pairwise correlations between most of the parameters were detected. Overall, the incorporation of the new tap water samples changed the correlation structure observed with the data from the first study. Only pH and silica parameters did not show a clear trend in relation to the rest of the parameters. The highly correlated structure of the analyzed data sets suggests that the application of multivariate data analysis methods, like PCA or PLS, can be used to discern major patterns in the panelists' preferences.

3.2. *Analysis of variance results*

Table 3 presents the results of the two-way ANOVA tests for the two data sets. The four-session replications of experimental settings were considered when determining experimental reliability and error estimations. The results confirmed (see Section 2.3 *Data organization and software*) that the water type and the panelist as factors, as well as their joint effect (interaction), statistically affected the evaluations at the 1% significance level. Therefore, water samples, in function of their physicochemical content, did significantly affect a panelist's rating. Furthermore, it can be concluded that the panel itself contained groups of individuals with different preferences of specific water types. Since both factors were significant, the recorded ratings also reflected the different panelists' personal preferences in relation to the specific mineral content of the water sample. The interaction variable was also significant, additionally supporting the presence of synergistic effects between these two factors and reflecting the high complexity of the sensory data in both studies. At this

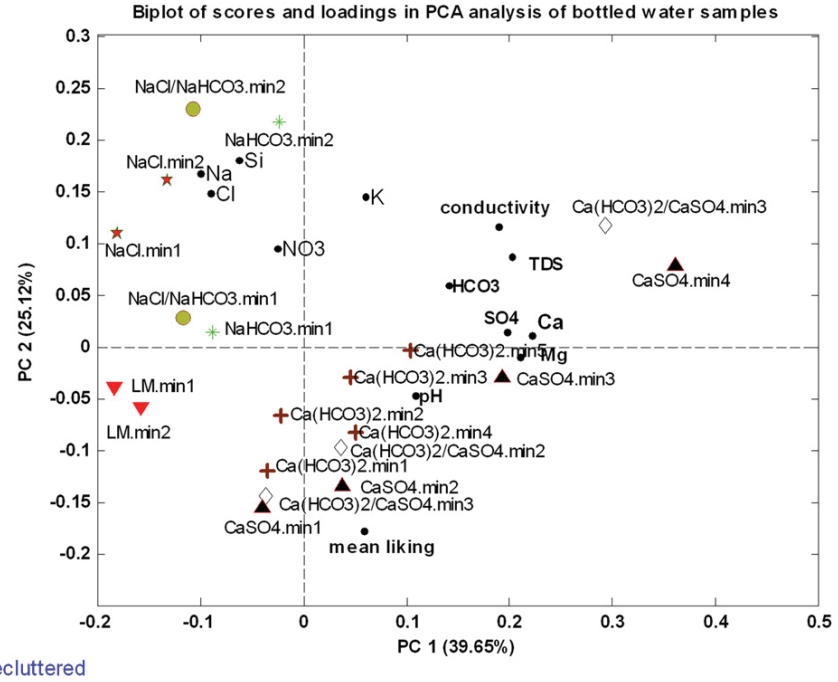


Fig. 2 – PCA results of the autoscaled [X_1, y_1] augmented data matrix: PC1 vs PC2 biplot showing only the PCA scores of bottled samples, as well as the PCA loadings of the physicochemical parameters and the panelists' global mean score. Water sample (with different symbols) score labels are given in Table 1. Water samples with similar physicochemical profiles are presented using same color markers according to Table 1 labels. Physicochemical and parameter loadings (black dots) are identified by their chemical names, except for the panelists' mean score (mean liking loadings at the bottom right quadrant of the plot). Axis scales dimensions are between −1 and 1, due to data autoscaling.

stage, it was not possible to distinguish the specific physicochemical features that explained the preference for a particular water sample.

3.3. Multivariate analysis

3.3.1. Exploratory PCA of the physicochemical parameters and the panelist's mean score on 20 bottled water samples: PCA of the [$\mathbf{X}_1$, $\mathbf{y}_1$] augmented data matrix

PCA on the autoscaled [$\mathbf{X}_1$, $\mathbf{y}_1$] augmented data matrix with a model of 2 principal components (PCs) explained more than 65% of the variance (information) as the two first principal components already captured 40% and 25% of the variance, respectively.

Fig. 2 shows the PC1 vs PC2 biplot for the 20 bottled water samples using the descriptive physicochemical parameters ($\mathbf{X}_1$) and the global panel's mean score vector ($\mathbf{y}_1$). Analysis of this biplot revealed that on PC1, water samples were distributed from the left to the right following the trend of the lowest to the highest TDS values. TDS loadings were also large and with positive signs on this PC1. Medium-mineralized samples were located between these two extremes. The panelists mean score vector loading was also located in the intermediate region, where water samples with intermediate TDS values (around 200–400 mg/L) occurred. This indicates that the panelists disliked waters with low (30–40 mg/L or lower) and high (above 800 mg/L) TDS levels. The distribution of loadings of the parameters on PC2 showed the panel's preference for water samples with moderate concentrations of calcium and relatively higher pH values of around 7.5–8.1. Positive correlations between the panelists' mean score and higher pH values as well as higher Mg^{2+}, Ca^{2+}, HCO_3^- and SO_4^{2-} concentrations are shown in this biplot (high concentrations/values for the parameter under investigation). The most liked water samples contained $CaSO_4$ and $Ca(HCO_3)_2/CaSO_4$. Conversely, the panelists' mean score vector showed inverse correlations with high K^+, Si, Na^+ and Cl^- concentrations, showing that the corresponding water samples were rated very low.

3.3.2. Exploratory PCA of the physicochemical parameters and the panelists' mean score on 13 bottled and 12 tap water samples: PCA of the [$\mathbf{X}_2$, $\mathbf{y}_2$] augmented data matrix

PCA analysis on the autoscaled [$\mathbf{X}_2$, $\mathbf{y}_2$] augmented data matrix with 3 principal components explained above 80% of the variance (information), with the first PC already capturing more than 57% of the variance and the second and third explaining 13% and 10% of the variance, respectively. A closer view of the first two components is presented in Fig. 3.

The Fig. 3 biplot reveals again that the tap water samples with the highest level of mineralization (expressed as the highest concentrations of TDS) gave large positive PC1 scores on the very right side of the biplot. On the contrary, water samples with low to moderate concentrations of TDS were located on the left side of this biplot. The whole set of monitored physicochemical parameters gave positive PC1 loadings

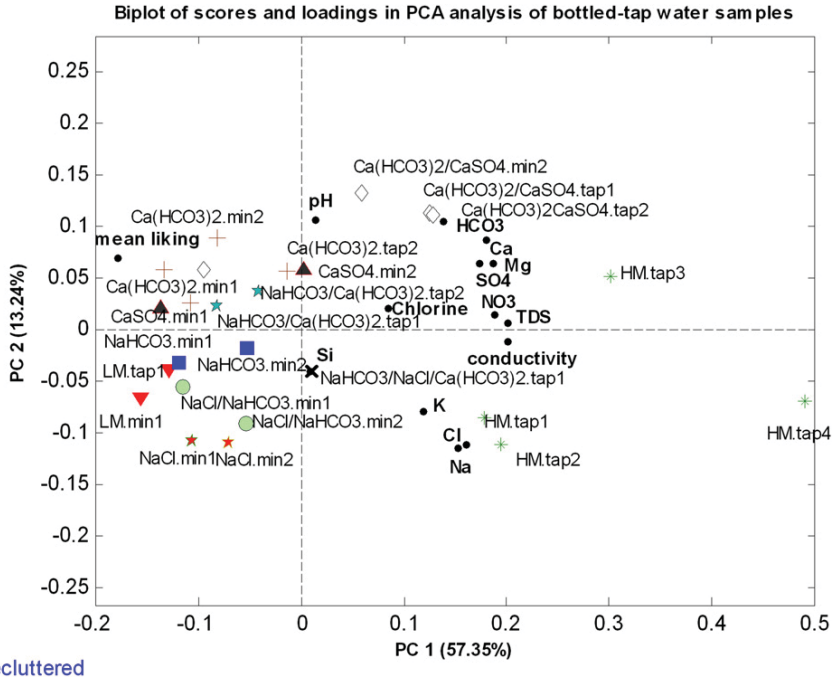


Fig. 3 – PCA of the autoscaled [$\mathbf{X}_2$, $\mathbf{y}_2$] augmented data matrix: PC1 vs PC2 biplot of the PCA scores of the 12 bottled and 12 tap water samples, as well as the PCA loadings of the physicochemical parameters and the panelists' mean score. Water samples with similar physicochemical profiles are presented using same color markers according to Table 2 labels. Physicochemical and parameter loadings (black dots) are identified by their chemical names, except for the panelists' mean score (mean liking loadings at the upper left quadrant of the plot). Axis scales dimensions are between −1 and 1, due to data autoscaling.

(right side of the plot) in contrast to the panelists' mean liking vector, which gave negative PC1 loadings (left side of the plot). This means that the panelists' mean score loadings were now on the side of low-to-moderate TDS concentrations (up to 400 mg/L), therefore indicating that the panelists did not like water samples with extremely high concentrations of minerals (now including tap water). Parameter loadings on PC2 showed very similar patterns in the panelists' scores to those found in the first experiment. Water samples with relatively higher Mg^{2+}, Ca^{2+}, HCO_3^- and SO_4^{2-} concentrations (relatively high compared to the parameter's mean values) were preferred more than those with higher concentrations of K^+, Si, Na^+ and Cl^-, without taking into account their origin (bottled or tap).

Interesting information is revealed when PC3 scores and loadings are also displayed. Fig. 4 gives the PC1 vs PC3 biplot for the 13 bottled (red triangles) and 12 tap (blue squares) water sample scores together with the physicochemical parameters and the global mean score loadings. In this case, all bottled mineral water samples were encoded as one single class and all tap water samples as another. The outlined class borders are drawn as red (bottled mineral) and blue (tap) lines. The two sample groups could be separated in this PC1–PC3 plot. Bottled water samples (in red) were located more on the more positive side of PC3 and the more negative side of PC1, while tap water samples were located more on the negative side of PC3 and the more positive side of PC1. The panelists' mean liking gave positive PC3 loadings, consistent with the panel's preference for bottled mineral over tap water samples. This result was also probably due to the presence of residual chlorine (chlorine also produced negative loadings on PC3) in the tap water samples. This clearly highlights the abilities of the highly trained panel since a recent study (Teillet et al., 2010a) showed that untrained consumers (two-thirds of them) could not discriminate between tap and bottled water samples when chlorine was allowed to disappear by natural aeration.

3.3.3. *PCA exploratory analysis of physicochemical parameters and individual panelist's mean score vectors for both studies: PCA of [$\mathbf{X}_1$, $\mathbf{Y}_1$] and [$\mathbf{X}_2$, $\mathbf{Y}_2$] augmented data matrices*

A new PCA was performed on [$\mathbf{X}_1$, $\mathbf{Y}_1$] and [$\mathbf{X}_2$, $\mathbf{Y}_2$] augmented data matrices. In the analysis of [$\mathbf{X}_1$, $\mathbf{Y}_1$] by PCA, the first two components captured 35 and 26% of the variance, respectively, whereas in the analysis of [$\mathbf{X}_2$, $\mathbf{Y}_2$] by PCA, the first two major components captured 59 and 13% of the variance, respectively.

Fig. 5a gives PC1 vs PC2 loadings obtained in the analysis of the [$\mathbf{X}_1$, $\mathbf{Y}_1$] augmented data matrix (only bottled water samples), considering all the panelists' responses individually. PC1 shows that groups of panelists exhibited preferences for and discrimination against specific water samples, with those containing relatively higher concentrations of Mg^{2+}, Ca^{2+}, HCO_3^- and SO_4^{2-} on one side and others comprising relatively higher contents of K^+, Si, Na^+ and Cl^- on the other. Outstandingly, there was a group of panelists (7, 8, 10 and 14 in Fig. 5a on the negative side of PC1) who liked water with higher amounts of sodium and chloride anions and lower pH values (below 7). The rest of the panelists were distributed on the positive side on PC1. Another group of panelists (3, 5, 6, 11 and 13 in Fig. 5a) rated very positively water samples rich in Mg^{2+}, Ca^{2+} and SO_4^{2-}. Finally, a third group of panelists (4, 9, 15, 16 and 17 in Fig. 5a) liked water samples with intermediate concentrations of Mg^{2+}, Ca^{2+}, HCO_3^- and SO_4^{2-}. Panelists from

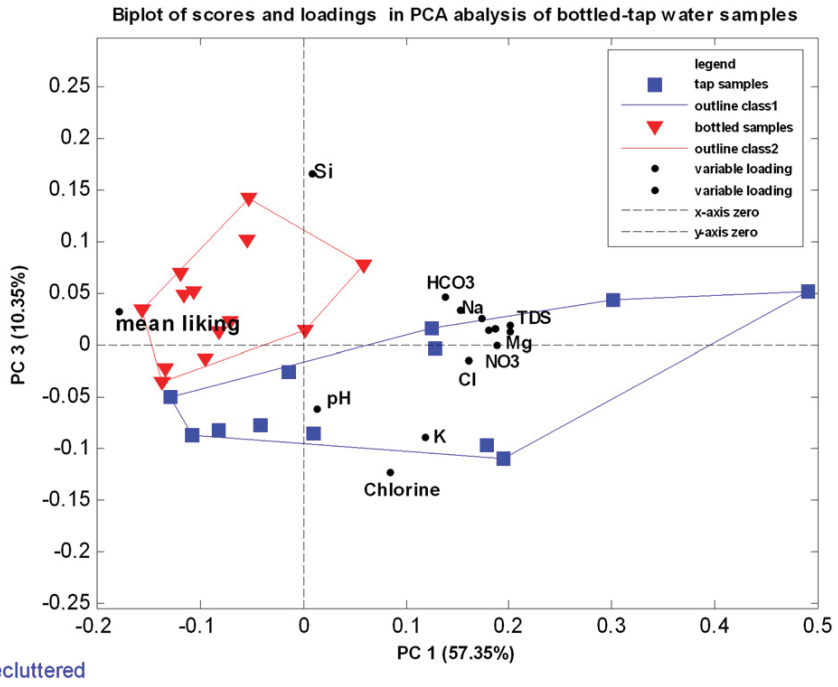


Fig. 4 – PCA of the autoscaled [X_2, y_2] augmented data matrix: PC1 vs PC3 biplot of the same parameters (variables) and samples. Outline borderlines are drawn for the two types of water samples (see Fig. 4 legend and Table 2). Axis scales dimensions are between −1 and 1, due to data autoscaling.

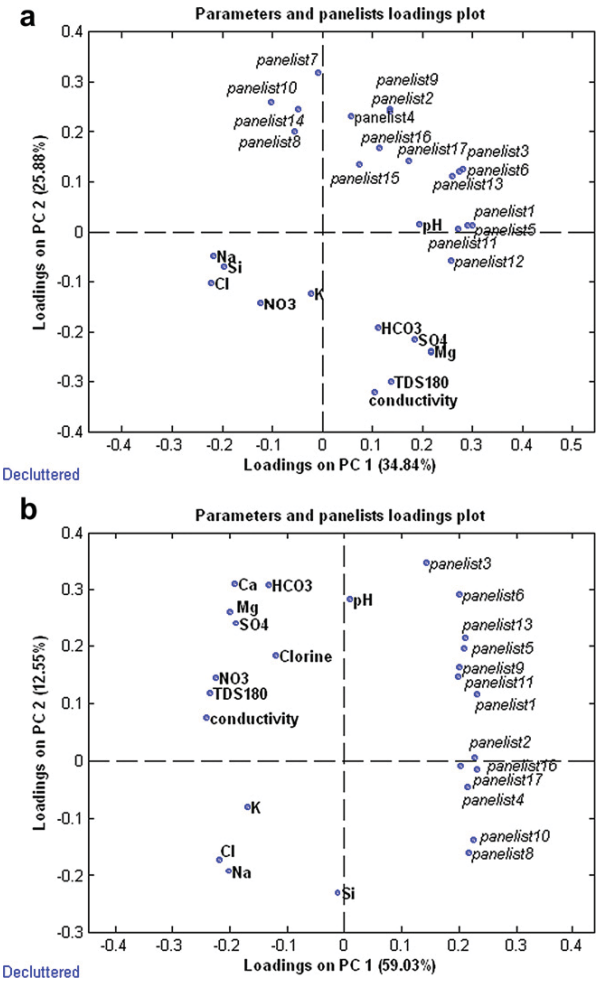


Fig. 5 – PC1 vs PC2 loadings plot of the autoscaled [X_1, Y_1] and [X_2, Y_2] augmented data matrices containing the measurement of physicochemical parameters and individual mean score vectors after 4 sessions in: (a) analysis of bottled water samples and (b) simultaneous analysis of bottled and tap water samples. Axis scales dimensions are between −1 and 1, due to data autoscaling.

the second and third groups also rated positively water samples with higher pH (above 7.5).

Fig. 5b gives PC1 vs PC2 loadings obtained in the simultaneous analysis of both the bottled and tap water samples, using the [$\mathbf{X}_2$, $\mathbf{Y}_2$] augmented data matrix and considering all the panelists' responses individually. Whereas loadings for individual panelists were on the right side (positive PC1 loadings) of the plot, all physicochemical parameters were now on the negative side on PC1. This again indicates the individual panelist's dislike of water with higher contents of all salts and TDS. Similar to previous results, PC2 suggests that some differences occurred among the panelists in terms of their preferences for samples with varying mineral contents. PCs separate panelists preferring water samples with relatively higher contents of K^+, Si, Na^+ and Cl^- from those preferring higher concentrations of calcium and magnesium (Mg^{2+}, Ca^{2+}, HCO_3^- and SO_4^{2-}). It should be noted that the same panelists were used in both studies for mineral and tap water samples. It is also possible to define a third group of panelists that liked water samples with intermediate levels of minerals, as can be seen around the 'zero' region on PC2.

3.3.4. *PLS regression of the panel's mean score vector on the physicochemical parameters*

PLS modeling $\mathbf{y}_1 = f(\mathbf{X}_1)$ resulted in a two-latent variable model, which also captured a large part of the information (around 72% of the $\mathbf{y}_1$ variance).

Fig. 6a displays the plot of PLS LV1 (the first latent variable or component) weight loadings of physicochemical parameters ($\mathbf{X}$ data matrix) on prediction of panelists' mean liking ($\mathbf{y}$ data vector). Fig. 6b shows the same for LV2. LV1 explained 25% of the $\mathbf{X}$ data variance and more than 64% of the $\mathbf{y}$ data variance and has positive weight loadings for pH, Mg^{2+}, Ca^{2+} and SO_4^{2-} and negative loadings for NO_3^-, Si, Na^+ and Cl^-. LV2 (which explained 34% of $\mathbf{X}$ variance and only 7.5% of the $\mathbf{y}$ variance) showed low positive weight loadings for Na^+, K^+ and high positive weight loadings for pH. All the rest of the parameters had negative weights on LV2 thus confirming the results already obtained by PCA in the previous section. A larger group of panelists liked water samples with higher levels of pH, Mg^{2+}, Ca^{2+} and SO_4^{2-}, while only a small group preferred water samples with higher concentrations of Na^+ and K^+.

Fig. 6c gives the VIP scores plot of the physicochemical parameters in the PLS model with two latent variables. Parameters with VIP scores above the threshold of one (red line) were considered to be the most significant in forming the panelists' decision. This plot demonstrates that Cl^-, NO_3^-, Na^+, pH and Si were in fact the parameters with the highest global importance in the construction of the final model. This could be because of the negative responses of most of the panelists toward water with higher contents of these parameters.

PLS modeling $\mathbf{y}_2 = f(\mathbf{X}_2)$ also resulted in a two-latent variable model that captured a large part of the $\mathbf{y}$ variance (around 89% of the $\mathbf{y}_2$ variance).

LV1 loading weights (explaining 56% of the $\mathbf{X}$ variance and 77% of the $\mathbf{y}$ variance) were positive only for the pH parameter, all other parameters showing negative weights (Fig. 6d). LV2 (explaining 13% of the $\mathbf{X}$ variance and only 12% of the $\mathbf{y}$ variance) showed only low positive weights for Na^+ and K^+ and a high positive one for pH (Fig. 6e). All these highlight the fact that panelists liked water samples with higher pH values and disliked the increase in global mineralization, e.g., an increase in TDS values that is too high. Similar to previous PLS results for the analysis of only bottled mineral water, LV2 indicated that some panelists preferred water samples richer in Na^+ and K^+. The analysis of VIP scores (Fig. 6f) also showed quite similar features in the preferences of most of the panelists. Again, their responses were related to features that elicited disliking of the water sample more than those eliciting a preference.

4. Discussion

Some references can be found in the literature about the influence of TDS on water liking. There is not a clear

agreement between classifications based on mineralization levels. Terms like "high", "intermediate", "moderate" or "medium" referred to TDS concentration depend on regional considerations, or on what were the water types used in a given study (van der Aa, 2003; Burlingame et al., 2007; Whelton et al., 2007).

Most of the national and international regulations, including the World Health Organization, the European Union and the USEPA refer to TDS as an esthetic quality parameter and give maximum guide values of 1000 mg/L (WHO, 2011), of 1600 mg/L – corresponding to 2500 μS at 25 °C (EU Directive, 1998), and of 500 mg/L (USEPA). A general agreement exists on that waters with high mineralization are not liked, and it seems that medium mineralized waters are best valued, but there is no conclusive data in relation to this. Bruvold and Daniels (1990) pointed out that the most appreciated waters were with TDS values up to 450 mg/L. McGuire et al. (2007) indicate that a consumer panel preferred TDS values of

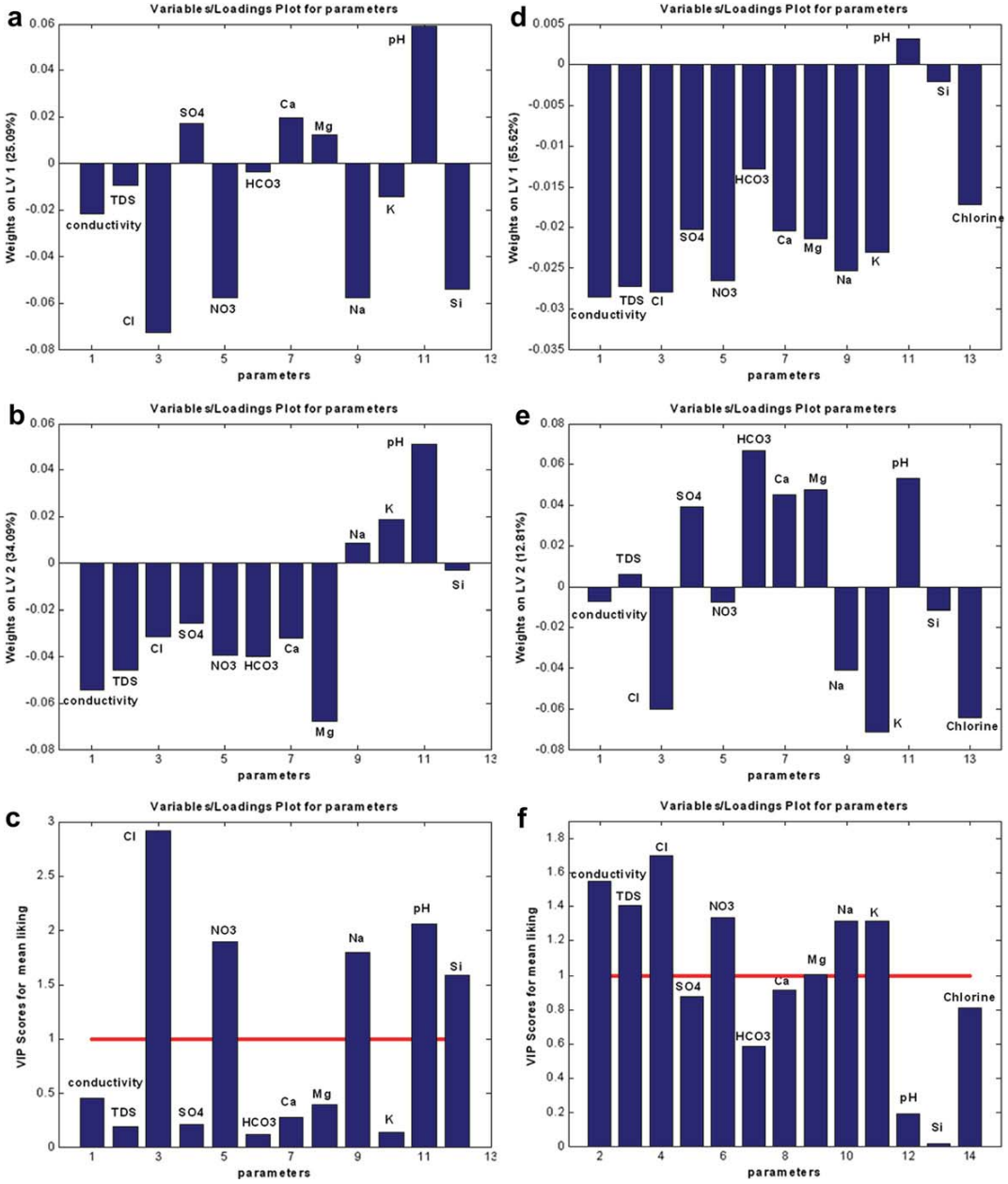


Fig. 6 – PLS results for the analysis of bottled water samples only (a, b and c) and the simultaneous analysis of bottled and tap water samples (d, e and f). Weight loading plots (a, b, d and e) and VIP (variable importance in projection) scores (c and f) are also shown. Parameters with VIP scores above the threshold value of one (red line) were considered to be the most significant in the overall panelists' score. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

450 mg/L instead of 650 mg/L although the liking scores were globally small. Two works by Devesa et al. (2007), and García and Devesa (2009) about blending of a high mineralization resource of about 1000–1100 mg/L with membrane treated waters showed a gradual improvement when TDS decreased, and the best results were reached at 270–350 mg/L. In a recent study, Teillet et al. (2010a) pointed out a TDS of 300–350 mg/L as preferred values.

The present study confirmed that high mineralized waters over 800 mg/L show low acceptance. The study with bottled waters showed that the preferred waters were with TDS around 200–400 mg/L, and panelists disliked also very low mineralized waters of about 30 mg/L. This difference between medium and low TDS was not deduced from the second study with bottled plus tap waters, which just showed that high mineralized waters (above 800 mg/L) received the lowest rates.

Papers about the role of the different cations and anions on taste are rare (Suffet et al., 1995; Burlingame et al., 2007; Whelton et al., 2007). A lot of information is given about detection thresholds and about the behavior of individual salts (anion–cation), but in natural waters synergistic and antagonistic effects take place between the different ions and they define the taste of given water sample. Therefore, there are discrepancies about the positive or negative effect of some species. The present work shows what is the contribution of some species at significant concentrations: positive for Ca^{2+}, SO_4^{2-}, HCO_3^- (and, consequently, high pH), and negative for Na^+ and Cl^-. Other chemical species found to have significant correlations with overall liking, K^+ and NO_3^-, have been found to be at concentrations too low to be relevant to influence the water taste. The correlation between K^+ and NO_3^- concentrations can be considered to be spurious since in fact, what happens is that both are correlated to the Na^+ concentration, which appeared as the major parameter for the negative panelist responses. The levels of K^+ and NO_3^- in the selected water samples are so small that they cannot be responsible for the water taste. However when data were autoscaled prior to PCA or PLS analyses, concentrations of minor constituents are then in a similar metrics to major constituents. Since K^+ and NO_3^-, both have positive correlations with Na^+, they both show then apparent negative taste qualities too.

The role of Mg^{2+}, which strongly depends on the anion of its salt, is still unclear. In the present work, $Mg(HCO_3)_2$ and $MgSO4_2$ (probably with a positive or neutral effect on taste) predominated over $MgCl_2$ (negative effect). On the other hand, the presence of Mg^{2+} salts is highly correlated with the presence of Ca^{2+} salts.

The influence of silica on the water taste is not clear in the literature. This study showed its positive correlation with Na^+, therefore explaining its apparent negative effect on taste ratings.

In the analysis of the second experiment with tap and bottled water samples (see Fig. 3) no difference was found between them (unchlorinated and chlorinated water samples). This fact is in agreement with previous results reported by Weisenthal et al. (2007), based on the Weber–Fechner curves/FPA method and on the use of sodium chloride to simulate TDS. Results showed that chlorine had no significant effect (antagonistic or synergistic) on the perception/assessment of the taste of drinking waters. Also, Teillet et al. (2010a) pointed out that two-thirds of the untrained panel were unable to discriminate between tap and bottled water samples once chlorine was allowed to volatilize by natural aeration. However, when PC3 (10% of the explained variance) is taken into account (see Fig. 4), a rather well separated group is formed for tap waters, probably due to the still presence of some residual chlorine in these water samples This fact would confirm the sensitivity of the trained panel and of PCA to detect this minor but still significant contribution, not detected in PC1 vs PC2 plot.

5. Conclusions

Some conclusions about the overall panelists' preferences for bottled mineral and tap water are:

1. The most important factor that influenced panelists' preferences was the overall level of mineralization (TDS). In both studies (one restricted to bottled waters and another one to bottled and tap waters), none of the panelists liked water samples with high levels of TDS, i.e., with more than 800 mg/L. The study with bottled waters showed that panelists preferred waters with TDS values around 200–400 mg/L, and that panelists disliked very low mineralized waters, with TDS values around 30 mg/L. This difference between medium and low TDS was not obtained in the second study, which only confirmed that high mineralized waters (above 800 mg/L) received the lowest rates.
2. For moderate levels of TDS, the physicochemical content appeared to be an influential factor. The contribution of several chemical species to the taste of water was determined. The preferred water samples had relatively high pH values (pH around 7.5–8.1), and relatively high concentrations of Ca^{2+}, Mg^{2+}, SO_4^{2-} and HCO_3^- ions. In general, panelists disliked water samples with high concentrations of K^+, Na^+, Cl^-, NO_3^- and Si, or with low pH (below 7) values. The present work shows the positive contribution to the water taste of Ca^{2+}, SO_4^{2-} and HCO_3^- (and consequently high pH), and the negative one of Na^+ and Cl^- at significant concentrations considering the literature information (Burlingame et al., 2007).
3. It was not clear whether low levels of Mg^{2+} influenced the taste ratings. In the present work, $Mg(HCO_3)_2$ and $MgSO_4$ (probably with a positive or neutral effect on taste) predominated over $MgCl_2$ (negative effect). On the other hand, this cation shows a strong positive correlation with Ca^{2+}, known to be mainly responsible for water hardness. Meanwhile, low levels of K^+ and NO_3^- are unlikely to influence taste ratings, apart from its high positive correlation with Na^+. The role of Si in water taste is not clear in the literature. This study showed their positive correlation with Na^+, therefore explaining their possible negative influence on taste ratings. More research is still needed on this.
4. The presence of residual chlorine did not significantly influence the panelists' ratings. However, this parameter allowed them to discriminate between bottled and tap water samples. This fact would confirm the sensitivity of

a trained panel and the usefulness of Principal component analyses to detect the minor but still significant contribution of residual chlorine.

5. The application of chemometric techniques allowed discrimination among groups of panelists according to their water type preferences. The application of PCA and PLS to physicochemical and sensory data in well-designed experiments is a useful approach for determining taste features of bottled and tap water samples.

Acknowledgments

The authors would like to thank the SOST-CO2 (CEN-2008-1027) and CTQ2009-11572 research projects from the CDTI, as well as the Spanish Ministerio de Economía y Competitividad for funding. They are also grateful to Philippe Piriou (CIRSEE, Paris) for his advice and information, and to the panelists for their collaboration in the sensory sessions.

REFERENCES

Bruchet, A., Lainé, J.M., 2005. Efficiency of membrane processes for taste and odor removal. Water Science and Technology 51 (6–7), 257–265.

Bruvold, W.H., Daniels, J.I., 1990. Standards for mineral content in drinking water. Journal of American Water Works Association 82 (2), 59.

Burlingame, G.A., Dietrich, A.M., Whelton, A.J., 2007. Understanding the basics of tap water taste. Journal of the American Water Works Association 99 (5), 100–110.

Chong, I., Jun, C., 2005. Performance of some variable selection methods when multicollinearity is present. Chemometrics and Intelligent Laboratory Systems 78, 103–112.

Devesa, R., Fabrellas, C., Cardeñoso, R., Matia, L., Ventura, F., Salvatella, N., 2004. The panel of Aigües de Barcelona: 15 years of history. Water Science and Technology 49 (9), 145–151.

Devesa, R., Cardeñoso, R., Matía, L., 2007. Contribution of the FPA tasting panel to decision making about drinking water treatment facilities. Water Science and Technology 55 (5), 127–135.

Devesa, R., Garcia, V., Matía, L., 2010. Water flavour improvement by membrane (RO and EDR) treatment. Desalination 250, 113–117.

Dietrich, A.M., 2006. Aesthetic issues for drinking water. Journal of Water and Health 4 (Suppl. 1), 11–16.

Dietrich, A.M., 2009. The sense of smell: contributions of orthonasal and retronasal perception applied to metallic flavor of drinking water. Journal of Water Supply: Research and Technology - AQUA 58 (8), 562–570.

EC (European Community), 1998. Council Directive 98/83/EC of 3 November, on the Quality of Water Intended for Human Consumption.

Geladi, P., Kowalski, B., 1986. Partial least-squares regression: a tutorial. Analytica Chimica Acta 185, 1–17.

García, V., Devesa, R., 2009. Supply of blends of desalinated seawater: effects on the flavour. Water Science & Technology: Water Supply 9 (1), 75–80.

Hibbert, D.B., 2009. Chemometric analysis of sensory data. In: Brown, S., Tauler, R., Walçzak, B. (Eds.), Comprehensive Chemometrics. Elsevier, Amsterdam, pp. 378–424.

Jolliffe, I.T., 2002. Principal Component Analysis, second ed. Springer Verlag, Berlin, Germany.

Kermadec, F., Durand, J., Sabatier, R., 1997. Comparison between linear and nonlinear PLS methods to explain overall preferences from sensory characteristics. Food Quality and Preference 8, 395–402.

Lengard, V., Kermit, M., 2006. 3-way and 3 block PLS regressions in consumer preference analysis. Food Quality and Preference 17, 234–242.

Mallevialle, J., Suffet, I., 1987. Identification and Treatment of Tastes and Odors in Drinking Water. American Water Works Association Research Foundation, Denver, CO (Appendix E).

McGuire, M.J., Loveland, J., Means, E.G., Garvey, J., 2007. Use of flavour profile and consumer panels to determine differences between local supplies and desalinated seawater. Water Science and Technology 55, 275–282.

Meng, A.K., Suffet, I., 1997. A procedure for correlation of chemical and sensory data in drinking water samples by Principal component factor analysis. Environmental Science and Technology 31, 337–345.

Naes, T., Risvik, E., 1996. Multivariate Analysis of Data in Sensory Science. Elsevier, Amsterdam, The Netherlands.

Peña Sánchez de Rivera, D., 1994. Estadística, Modelos y Métodos, second ed., vol. 1 y 2. Alianza Universidad Textos.

Stanimirova, I., Boucon, C., Walczak, B., 2011. Relating gas chromatographic profiles to sensory measurements describing the end products of the Maillard reaction. Talanta 83 (4), 1239–1246.

Suffet, I., Meng, A., Khiari, D., Brenner, L., Mallevialle, J., Anselme, C., Bordet, J., 1989. Proceedings of the AWWA Annual Conference Proceedings, No. 20036TT, Los Angeles, CA. AWWA: Denver, CO.

Suffet, I.H., Mallevialle, J., Kawcynski, E. (Eds.), 1995. Advances in Taste-and-Odour Treatment and Control. American Water Association Research Foundation – Lyonnaise des Eau, Denver, CO, pp. 247–280 (Chapter 6) by L. Matia.

Teillet, E., Urbano, C., Cordelle, S., Schlich, P., 2010a. Consumer perception and preference of bottled and tap water. Journal of Sensory Studies 25, 463–480.

Teillet, E., Schlich, P., Urbano, C., Cordelle, S., Guichard, E., 2010b. Sensory methodologies and the taste of water. Food Quality and Preference 21 (8), 967–976.

USEPA. http://water.epa.gov/drink/contaminants/index.cfm#Secondary. 03.10.12.

van der Aa, M., 2003. Classification of mineral water types and comparison with drinking water standards. Environmental Geology 44, 554–563.

Weisenthal, K.E., McGuire, M., Suffet, I.H., 2007. Characteristics of salt taste and free chlorine or chloramines in drinking water. Water Science and Technology 55 (5), 293–300.

WHO (World Health Organization), 2011. Guidelines Ford Drinking-Water Quality, fourth ed. WHO, Geneva, Switzerland (Chapter 10).

Whelton, A.J., Dietrich, A.M., Burlingame, G.A., Johnson, M., Duncan, S.E., 2007. Minerals in drinking waters: impacts on taste and importance to consumer health. Water Science and Technology 55 (5), 283–291.

Results and Discussion

- Descriptive statistics of the physicochemical parameters of the two sets of water samples

Pairwise correlation coefficients between the 12 physicochemical parameters in the first data set, as well as the correlation between the 13 of parameters the second data set, were calculated. These correlation coefficients were visualized in the form of a correlation map, where intense red color represents high positive correlations and intense blue color - high negative correlation (see Figure 1, Article 6). The physicochemical parameters from the data set containing 20 bottled waters were found to present high positive or negative correlations and generally these parameters formed two clusters. For example, it was detected a cluster of parameters with high positive correlations between Mg^{2+}, Ca^{2+} and SO_4^{2-} (see Figure 1a, Article 6). Another cluster of physicochemical parameters included positively correlated, $NO3^-$, Si, Na^+ and Cl^-. The parameters from the two clusters were negatively correlated between them.

In the second data set, strong positive pairwise correlations between most of the parameters were observed (see Figure 1b). This result is attributed to the nature of the tap water samples included in this study. pH and silica were the only parameters, which had a distinct behavior.

- Two-way Analysis of variance results

Two-way ANOVA tests for the two data sets indicated that water type and panelist as factors, as well as their joint effect (interaction), were statistically significant at the 1% significance level. This result suggested that water samples with their mineralization had a strong effect on the panelists' overall rating. The panel structure also seemed to be a significant factor, because it contained groups of panelists with different preferences for specific water types. Interaction between factors was also significant, suggesting that synergistic effects between these two factors could present.

- Results from PCA of the physicochemical parameters and the panelists' mean score for 20 bottled water samples.

PCA of the autoscaled data matrix resulted in a model with 2 principal components (PCs). This amount of explained variance was found to be moderate, i.e., more than 65%. PC1 explained the larger portion of the total variance (i.e., 40%), whereas PC2 - 25%. Analysis of the PC1-PC2 biplot (water type scores and parameters

loadings plotted together) revealed that the largest effect on PC1 had the overall mineralization (see Figure 2, Article 6). The scores for water samples are distributed as a function of TDS values (i.e., from the left to the right, we have lowest to highest TDS values). TDS loading on PC1 corroborated this result. It was further detected that the panelists' mean score vector loading was largely concentrated in the intermediate area, suggesting that panelists' preferences for water samples were toward intermediate TDS values (around 200-400 mg/L). Therefore, we conclude that panelists disliked waters with relatively low and high TDS levels (i.e., 30-40 mg/L or lower and above 800 mg/L, respectively)**.**

Further analysis of the distribution of parameter loadings on PC2 suggested that panelists exhibit a preference towards water samples with concentrations above their average for Ca^{2+} , Mg^{2+} Ca^{2+}, HCO_3^- and SO_4^{2-} concentrations (see more the biplot). The most liked water samples had $CaSO_4$ and $Ca(HCO_3)_2/CaSO_4$ profiles. In contrast, water samples with concentrations above their average for K^+, Si, Na^+ and Cl^- concentrations had been ranked low.

- *Results from PCA of the physicochemical parameters and the panelists' mean score of 13 bottled and 12 tap water samples.*

PCA of the autoscaled data matrix established a model with 3 principal components, explaining more than 80% of the total variance. The first PC accounted for more than 57% of the variance, whereas the second and third PC - 13% and 10%, respectively. The first PC captured the effect of overall mineralization in the data. The biplot (see Figure 3, Article 6) represents the distribution of the water samples scores from lowly mineralized (left) to highly mineralized waters (right). The physicochemical parameters loadings were found on right side of the plot, whereas the panelists' mean liking vector was on the left side. Therefore, we conclude that panelists' preferences were for low-to-moderate TDS concentrations and hence suggest a possibility to narrow the upper limit of overall mineralization, or TDS values below 400 mg/L. In the analysis of tap-bottled water, PC2 corroborated previously reported results that panelists positively rated water samples with relatively high Mg^{2+}, Ca^{2+}, HCO_3^- and SO_4^{2-} concentrations, as compared to the average value. In contrast, water samples (bottled or tap) with higher concentrations of K^+, Si, Na^+ and Cl^- were rated low.

The interpretation of PC3 provided interesting implications. More specifically, this principal component captured the effect of residual chlorine on the panelists' preferences. The relation bottled-tap waters was represented in the distribution on the biplot for the pair of species Si- residual chlorine (see Figure 4, Article 6). This pair characterized the main difference between the two groups. The panelists exhibited a weak preference towards bottled mineral water samples. The lower concentrations of Si in water samples and the lack of literature regarding its effect on taste were suggestive that this result could be most likely explained by the presence of residual chlorine in the tap water samples.

- *Results from PCA of physicochemical parameters and individual panelist's mean score vectors for both studies*

In Article 6, two new PCA models were performed using augmented data matrices with physicochemical parameters in addition to individual panelists mean scores. In the analysis of bottled water samples, PCA with the first two components captured 35% (PC1) and 26% of the variance (PC2), respectively. In the analysis of bottled-tap water samples, PCA with the first two components captured 59% and 13% of the variance, respectively.

Both biplots PC1-PC2 confirmed again that panelists disliked water with higher contents of all salts and TDS (see Figure 5, Article 6). Similarly to previously reported results, PC2 suggested that panelists exhibited divergent preferences for water samples with different levels of mineral contents. The results confirmed that there were three groups of panelists formed on the basis of their preferences for mineral content. The largest group of panelists preferred water samples with relatively higher contents of Mg^{2+}, Ca^{2+}, HCO_3^- and SO_4^{2-}. The second group liked water samples with moderate levels of all minerals. The third, smaller group of panelists expressed a preference towards water samples with relatively higher contents of K^+, Si, Na^+ and Cl^-. The analysis also showed that many panelists were very consistent in their evaluations, collected in taste rating sessions with more than two-year time difference.

- *Results from PLS regression of the panel's mean score vector on the physicochemical parameters for both data sets (only bottled waters and bottled-tap waters)*

In both regressions examples, PLS modelling was used to obtain two-latent variables (LVs) models, which were able to explain approximately 72% and 89% of the

variance in panelists' liking. Figure 6 (Article) 6 contains the contribution (i.e., weights) of the physicochemical parameters for the first two LVs for both PLS models. The model using bottled water samples data set revealed that pH, Mg^{2+} Ca^{2+-} and SO_4^2 had the strongest positive contribution (highest weights loadings values) on the first LV. In contrast, LV2 was determined by positive weight loadings for Na^+ and K^+. Therefore, we corroborate previously reported results from PCA that there were different groups inside the panel, formed on the basis of panelists' preferences towards hard- or soft-salty waters. In comparison, the model using bottled-tap water samples data set revealed that first LV was largely dominated by the strong negative contribution of almost all physicochemical parameters and that the relationship hard- (positive contribution for pH, Mg^{2+} Ca^{2+-} and SO_4^2-) versus soft-salty (negative contribution for Cl^-,Na^+ and K^+) water samples was explained by the second LV. This result confirmed that the increase of the mineralization by incorporation of tap water samples with very high TDS concentrations affected panelists' perception regarding the overall mineralization.

Finally, the analysis of VIP scores in both models displayed similar patterns in the preferences of most panelists. Panelist could better identify which properties of water samples they particularly disliked, since the most relevant parameters according to the PLS VIP scores were TDS (high mineralization), Na and K (salty taste).

Chapter 4

General Conclusions

Conclusions related to the of water quality analysis and monitoring

- It was confirmed that the formation of trihalomethanes depended strongly on ambient and operational water treatment plant parameters, involved in the disinfection process. The most significant ambient parameters were temperature, water ultraviolet absorbance at 254nm, total organic carbon concentrations, water conductivity, oxidability and turbidity, monitored in the raw river water and along disinfection process stages. The most important operational parameters during disinfection for the THMs formation resulted: the chlorine doses and the age of carbon filters.

- It was confirmed that the trihalomethanes formation took place during the pre-chlorination and the post-chlorination stages at the San Joan Despi DWTP. The process of active carbon filtration and the addition with underground water quantities significantly reduced the trihalomethanes formation.

- It was discovered also, that the formation of trihalomethanes has a seasonal dependence in the Sant Joan Despí DWTP (Barcelona, Spain). The spring appeared the season with largest amounts of THMs formed with predominance of brominated species over chloroform.

- Dissolved organic matter was confirmed as very important for the speciation of THMs during formation. The colloidal fraction of DOM was responsible for about 20-30 % of the formation of each individual THM and about a quarter of the total THMs formation. The hydrophobic fraction was found as the most relevant for the formation of brominated THMs. The transphilic fraction had importance for the formation of chloroform formation. Both, hydrophobic and transphilic fractions were found important in the formation of about 80% of the total THMs.

- This Thesis confirmed the importance of the bromide anions concentration for the formation of brominated trihalomethanes during disinfection. Despite the type of DOM fraction, just a small quantity of bromide anions in the water would alter the formation of brominated trihalomethanes.

- This Thesis proved that UV spectrophotometry coupled with chemometrics resulted as a powerful analytical tool for mixture analysis of water blends. This hyphenated approach allowed for differentiation of various water sources with specific natural organic matter content in blends. Also, it permitted to estimate very accurately the concentrations of several water quality parameters, such as dissolved organic matter, nitrates, detergents and phenols.

- In this Thesis, UV spectrophotometry was used individually or as a complementary source of information to physicochemical parameters for better modeling and data monitoring of water quality.

- Temporal and spatial dynamics (seasonal, diurnal urban activity and influent-effluent changes) was monitored in the wastewater quality, modeling spectral and physicochemical data. Unusual disturbances like rainfall episodes and accidental water pollution events were being possible to detect also.

- It was proved that the overall mineralization of bottled and tap waters was the strongest factor that influenced on the panelists' liking. Overall water mineralization around 200- 400 mg/L, expressed as total dissolved solids, was liked more than waters with mineralization above 800 mg/L and below 50 mg/L. The largest number of panelists liked bottled and tap waters with relatively high concentrations of Ca^{2+}, Mg^{2+}, SO_4^{2-} and HCO_3^- ions. It was discovered that a smaller number of panelists with preferences for high concentrations of K^+, Na^+, Cl^-, NO_3^- and Si existed also.

- It was discovered that the residual chlorine in tap waters did not significantly influence the panelists' liking, however it was tasted to discriminate between bottled and tap water samples.

Conclusions related to the application of chemometric methods for water quality data analysis and monitoring

- Different chemometrics methods were found very appropriate for application in the analysis of problems related to the water quality. In this Thesis, we have demonstrated the effectiveness of methods from the multivariate exploratory analysis such as Principal Component Analysis (PCA) and from the multivariate calibration such as Multilinear Regression (MLR), Partial Least Square Regression (PLS), Support Vector Machine Regression (SVM), kernel Radial Basis Function Partial Least Square Regression (K-PLS) and others. Experimental design with response surface strategy were found also very useful as an economical approach to discover information about the important factors affecting THMs formation in a laboratory study, included in this Thesis.

- The Thesis confirmed the efficiency of different visualization techniques and tools from PCA and PLS such as visualization of scores, loadings and weights loading plots. Such visualization made possible to discover the major sources (factors) of variance in the investigated cases. The score plots reveled various temporal (seasonal, diurnal) dynamics in the THMs formation; urban activity for the quality of the waste water and for example - diverse grouping of panelists according to their water taste preference. Spatial information about influent-effluent dynamic changes in wastewater quality was also discovered, analyzing score plots. Unusual disturbances, like rainfall episodes and accidental water pollution events, were diagnosed also, proving the usefulness of selected methodology for future quality control analysis. Vizualization of the loadings and weights loadings plots of PCA and PLS made possible to underline the importance of various ambient and operational parameters for the THMs formation in the drinking water treatment plant process. These type of plots were very useful to discover variables with strongest influence for the water quality dynamics in the waste water and also to feature the most important mineral components affecting the taste liking of bottled and tap water.

- In this Thesis, it was tested and confirmed its usefulness of new visualisation techniques from K-PLS and SVM regression analyses in order to reveal complex nonlinear relationships among several ambient and operational parameters monitored at the DWTP and having a strong impact on the THMs formation.

- Variable Importance in Projection of the scores technique was proved as very useful for optimal selection of a reduced set of different parameters /variables (physicochemical parameters, spectral wavelengths, mineral ions) in the global analysis of the parameters' importance or in case of prediction - without a significate loss of predictive power.

- The most accurate prediction errors were achieved using MLR, PLS and K-PLS regression techniques and they are recommended for routine use in the modeling of target compounds in problems related to the water quality.

- Factorial analysis strategy, applied to understand the main factors for the THMs formation in the water disinfection, appeared as a very useful approach. The strategy implemented a fast screening (Placket-Burman) and a detailed experimental design (Box-Behnken) and allowed to describe the most influential parameters in the formation of trihalomethanes as evaluated their effects and interactions.

Chapter 5

Summary in Spanish (Resumen de la Tesis)

5.1 Resumen

En esta Tesis, se propone la utilización de diferentes métodos quimiométricos para la exploración, el análisis y la interpretación de la información presente en datos experimentales obtenidos en la determinación de la calidad de aguas potables y de aguas residuales. Los conjuntos de datos analizados fueron obtenidos en (a) procesos de desinfección de agua potable, (b) procesos de tratamiento de agua residual, (c) análisis sensoriales del gusto del agua mediante panelistas, y (d) experimentos en laboratorio que simulan procesos de desinfección.

1) Un primer objetivo de esta Tesis es el de mejorar el conocimiento con respecto a la formación de trihalometanos (subproductos de la desinfección en la *estación de tratamiento de agua potable* de Sant Joan Despí, ETAP-SJD) y los principales factores que afectan a su formación.

2) Otro objetivo de la Tesis consiste en el desarrollo de un método quimiométrico que permita la diferenciación y cuantificación de las fuentes (orígenes) diferentes de agua potable en la red de distribución (WDS) de Barcelona utilizando la espectroscopia de absorción en el ultravioleta y la medición de parámetros físico-químicos.

3) Por otra parte, en la Tesis se propone una tecnología basada en la aplicación de métodos quimiométricos que facilite el control de la calidad del agua en la estación depuradora de aguas residuales (EDAR) cerca de la ciudad de Girona.4) Finalmente, se estudia también el problema de la evaluación del gusto del agua embotellada o del grifo en función de su contenido de minerales con modelos quimiométricos utilizando panelistas entrenados.

En los cuatro casos, los métodos quimiométricos se han aplicado a matrices de datos multi-paramétricos generadas por distintas técnicas instrumentales como, por ejemplo, la espectroscopia en el UV-VIS, la cromatografía de gases con detector de captura de electrones (GC-ECD) y el espectrofotómetro de emisión acoplado inductivamente a plasma (ICP-OES). Además, se obtuvieron otros conjuntos de datos mediante la aplicación de métodos estándar de laboratorio para la estimación de los diferentes parámetros físico-químicos, o de datos multi-paramétricos a partir del Sistema de Gestión de Información de Laboratorio (LIMS) de AGBAR. Finalmente, fueron adquiridos también otros conjuntos de datos mediante una estación semi-

automática multi-paramétrica, de monitorización en línea y mediante un diseño experimental de experimentos sensoriales.

Los métodos quimiométricos empleados en esta Tesis incluyen el Análisis de Componentes Principales (PCA), la Regresión Lineal Múltiple (MLR), la Regresión de Componentes Principales (PCR) y el método de regresión de Mínimos Cuadrados Parciales (PLS). Se han comparado estos diferentes métodos de regresión lineal con los métodos de regresión no lineal, tales como el procedimiento 'kernel' de Mínimos Cuadrados Parciales (K-PLS) o el método de regresión basado en Máquinas de Soporte Vectores de Soporte (Support Vector Machine Regression, SVR).

Los resultados más significativos de esta Tesis han sido:

a) la identificación de un conjunto de parámetros fisicoquímicos ambientales y del proceso de desinfección del agua que tienen relevancia en la formación de trihalometanos;

b) el desarrollo de modelos quimiométricos para la estimación de la estimación de los cambios de concentración de las cuatro especies de THM y de su suma en total en la planta de tratamiento de aguas ETAP-SJD con errores de predicción bajos;

c) la identificación de que las especies HCO_3^-, SO_4^{2-}, Ca^{2+}, Mg^{2+} (a una concentración de mineralización total moderada) son los parámetros físico-químicos más relevantes para el gusto del agua ;

d) la detección de las variaciones temporales más importantes en la calidad del agua residual tratada en la planta de tratamiento de aguas EDAR de Trargisa en Girona, con una frecuencia diaria, mensual y estacional;

e) el desarrollo de modelos de detección y de predicción de las fuentes u origines del agua potable dentro de la red de distribución de aguas, WDS, de la ciudad de Barcelona.

Además, se han probado y evaluado diferentes técnicas quimiométricas para la visualización e interpretación de los datos de análisis de la calidad del agua. La capacidad de predicción de los métodos de regresión lineal y no lineal ha sido comparada en relación el desarrollo de modelos empíricos de predicción de los parámetros de calidad del agua. A partir de la utilización de los procedimientos quimiométricos propuestos, los errores de predicción de Las concentraciones de THM en agua potable, nitratos, fenoles, materia orgánica en agua residual, los origines de

agua potable en el sistema de distribución de agua y las puntuaciones de los panelistas al gusto de agua han sido bajas.

En conclusión, esta Tesis muestra las ventajas del uso de métodos quimiométricos de análisis multivariante de datos en la evaluación de la calidad del agua en diferentes contextos (plantas de tratamiento, potabilización, redes de distribución, aguas de bebida,…). Se demuestra por lo tanto, que la utilización de métodos de quimiométricos representa un avance significativo comparado con los métodos de control basados en estadística univariante, los cuales requieren muestreos más caros y una cantidad elevada de tiempo.

Los resultados de la investigación realizada en esta Tesis se muestran en un conjunto de seis artículos publicados en revistas de elevado impacto internacional en el ámbito de calidad de agua.

5.2 Objetivos de la Tesis

El principal objetivo de esta Tesis incluye la aplicación y la promoción de los métodos quimiométricos para el análisis de la calidad del agua potable y del control de la calidad del agua residual. Esta Tesis está basada en seis diferentes estudios: cinco estudios relacionados a la gestión del agua potable y un estudio relacionado a la monitorización del agua residual. Se han generado y analizado mediante métodos quimiométricos diferentes conjuntos de datos en varios estudios, relacionados con la calidad del agua en las plantas de tratamientos de aguas potables y residuales en Cataluña.

Según el ámbito de investigación, este objetivo principal se subdivide de los siguientes apartados:

Objetivos relacionados con el análisis y monitorización de la calidad del agua

- Desarrollo de modelos de regresión que permitan predecir la formación de trihalometanos (THMs) en la salida de la planta potabilizadora de agua de Sant Joan Despí (ETAP), basados en modelos lineales y no lineales a partir de parámetros de calidad del agua que caracterizan el proceso de desinfección
- Evaluación e interpretación de los parámetros más importantes para la formación de THMs basándose a técnicas quimiométricas, que faciliten el seguimiento y control de la calidad del agua;
- Valoración de la importancia de la materia orgánica (NOM) en la formación de THMs durante desinfección simulada en laboratorio a partir de experimentos diseñados estadísticamente;
- Desarrollo modelos quimiométricos para la diferenciación y cuantificación de los cinco orígenes diferentes de agua potable que alimentan la red de distribución de Barcelona utilizando espectroscopia de absorción en ultravioleta y parámetros físico-químicos.
- Identificación de los parámetros físico-químicos del agua más influyentes en el gusto de agua embotellada o de la red de distribución a partir de modelos

quimiométricos de predicción utilizando las evaluaciones de panelistas entrenados;

- Evaluación de los métodos quimiométricos para mejorar la monitorización y el control de la estación depuradora de aguas residuales (EDAR) de Girona (TRARGISA), basándose en diferentes técnicas y rutinas empleadas para controlar en línea periódicamente la calidad del agua;
- Selección de un número reducido de canales espectrales en ultravioleta (longitudes de onda) para mejorar la monitorización en línea de los procesos operativos en EDAR, a partir de medidas de espectrometría de absorción.

Objetivos relacionados con la aplicación de los métodos quimiométricos

- Desarrollo, aplicación y validación de los métodos quimiométricos de regresión lineal y no lineal en el análisis de los datos de calidad del agua en diferentes procesos de tratamiento de agua, basándose en parámetros medidos in situ o en el laboratorio;
- Comparación de las capacidades predictivas de los métodos de regresión lineal y no lineal en la formación de THM;
- Identificación y evaluación de las técnicas quimiométricas y de sus herramientas de visualización más eficaces para seleccionar las variables más importantes (parámetros) en los modelos lineales;
- Aplicación y evaluación de las técnicas de visualización de variables más importantes en modelos no lineales (K-PLS y SVR) y su comparación posterior con las obtenidas con los modelos lineales;
- Aplicación de las técnicas de diseño experimental (DOE) con el objetivo de obtener conjuntos de datos de calibración y realizar la evaluación de los factores más importantes;
- Evaluación del método de preselección de los "variable más importantes en proyección" (Variable Importance in Projection VIP) a partir de los modelos PLS y de los datos de calidad del agua que permitan la selección de un número reducido de variables que preserven el poder predictivo de los modelos.

Estructura de la tesis

Esta tesis se divide en dos partes principales. La primera parte incluye la introducción general sobre los problemas asociados a la evaluación de la calidad del agua en los sistemas estudiados en esta Tesis. Entre los problemas estudiados están la formación de trihalometanos, los factores que influyen en el gusto del agua, y los métodos de monitorización de la calidad de las aguas potables y de las aguas residuales. También se realiza una introducción breve de los métodos aplicados quimiometría utilizados en esta Tesis. La segunda parte de la Tesis contiene los artículos científicos publicados, junto con una introducción y discusión breves de los resultados obtenidos. Esta segunda parte acaba con un resumen breve de la tesis y con las referencias bibliográficas.

La tesis consta de los siguientes seis capítulos:

- En el Capítulo 1, se presentan los objetivos de la Tesis, se detalla la estructura de la tesis y se incluye la lista de publicaciones en relación de esta Tesis.
- En el Capítulo 2, se detallan brevemente los problemas principales relacionados con la evaluación de la calidad del agua y se describen los métodos quimiométricos aplicados en esta Tesis. Se presenta información sobre la formación de trihalometanos y la legislación europea sobre sus niveles regulados; se estudia la importancia de la materia orgánica en la formación de trihalometanos; se describen los estudios epidemiológicos con respecto al riesgo sanitario de THM; se analizan las fuentes (origines) de agua potable de la red de distribución de Barcelona. Se presenta información y ejemplos de una planta ETAP clásica y de una planta EDAR. Se describen los principales parámetros de calidad del agua, controlados regularmente en este tipo de instalaciones. Se incluye una breve discusión sobre sistemas automáticos para monitorización de la calidad del agua en línea basados en sensores de ultravioleta. Se discuten los aspectos organolépticos y del gusto del agua. Este capítulo concluye con una revisión de los métodos quimiométricos aplicados en esta Tesis.
- En el Capítulo 3, se discuten los resultados obtenidos de los estudios incluidos en esta tesis. Este capítulo se divide en tres bloques de la siguiente manera:

a) En el primer bloque, se presentan una breve introducción y discusión posterior de los tres artículos que investigan la formación de THM en experimentos diseñados en laboratorio y en el control de la calidad del agua de ETAP-SJD, basándose de métodos quimiométricos.
b) En el segundo bloque, se presentan dos artículos con estudios relacionados con la modelización y el control de la calidad del agua potable o residual. En el primer artículo, se realiza un estudio relacionado con la diferenciación y cuantificación de los cinco orígenes o fuentes diferentes de agua potable que alimentan la red de distribución de Barcelona a partir de espectroscopia de absorción en ultravioleta y medición de parámetros físico-químicos. El segundo artículo describe el desarrollo de metodología quimiométrica, para mejorar el control de la calidad del agua en una planta de tratamiento de aguas residuales (EDAR), a partir de los parámetros de calidad de aguas residuales obtenidos con varias técnicas de monitorización.
c) En el tercer bloque, se presentan los resultados de un estudio sensorial con diferentes muestras de agua embotellada y de la red pública de distribución. El principal objetivo de este estudio consistía en correlacionar los parámetros físico-químicos medidos en las aguas con las evaluaciones de panelistas entrenados del gusto de estas aguas.

- En el capítulo 4, se presentan las conclusiones generales más importantes de esta Tesis.
- En el capítulo 5, se incorpora un resumen en castellano de los trabajos realizados en esta Tesis.
- En el último capítulo 6, se presentan las referencias empleadas en esta Tesis.

Lista de artículos científicos incluidos en esta Tesis

1. **Article 1** – Platikanov, S., Puig, X., Martin, J. and R. Tauler. *Chemometric modeling and prediction of trihalomethane formation in Barcelona's water works plant.* **Water Research** 41 (2007) 3394-3406.

2. **Article 2** – Platikanov, S., Martin, J. and R. Tauler. *Linear and non-linear chemometric modeling of THM formation in Barcelona's water treatment plant.* **Science of Total Environment** 432 (2012) 365-374.

3. **Article 3** – Platikanov, S., Tauler, R., Rodriguez, P., Antunes, M., Pereira, D. and J. Esteves da Silva. *Factorial Analysis of the trihalomethane formation in the reaction of colloidal, hydrophobic and transphilic fractions of DOM with free chlorine.* **Environmental Science and Pollution Research** 17 (2010) 1389-1400.

4. **Article 4** – Platikanov, S., Garcia, V., Landeros, E., Devesa, R., Matía, L., Tauler, R., *Determination of water supply sources in the Barcelona distribution system by UV spectrophotometry and PLS.* **Water Science and Technology-Water Supply** 11 (2011) 45-54.

5. **Article 5** – Platikanov, S., Rodriguez-Mozaz, S., Huerta, B., Barcelo, D., Cros, J., Battle, M., Poch, G., Tauler, R. *Chemometrics quality assessment of wastewater treatment plant effluents using physicochemical parameters and UV absorption measurements.* **Journal of Environmental Management** 140 (2014) 33-44.

6. **Article 6** – Platikanov, S., Garcia, V., Fonseca, I., Rullan, E., Devesa, R., Tauler, R., *Influence of minerals on the taste of bottled and tap water: A chemometric approach.* **Water Research** 47 (2013) 693-704.

5.3 Introducción

El Área Metropolitana de Barcelona (BMA) se caracteriza por tener un ciclo urbano complejo del agua (Marín et al., 2012). Los recursos de agua cruda para la potabilización incluyen agua superficial (del río Ter), agua salobre (del río Llobregat), agua subterránea (de los acuíferos del Llobregat y el Besòs), y agua de mar (del Mar Mediterráneo). Estés fuentes diferentes de agua cruda se caracterizan por una gran variabilidad en la calidad, y por tener diferentes factores y niveles de estrés.

Aigües de Barcelona (AGBAR) es responsable del suministro de agua potable (Paraira and West, 2015) a más de 3 millones de habitantes del Área Metropolitana de Barcelona (BMA). El suministro de agua potable a una población tan grande y a todos los hogares bajo el cumplimiento de las estrictas normas regulatorias es un gran desafío desde el punto de vista de una gestión operativa eficaz.

La lista de los principales problemas de la calidad del agua encontrados en la BMA incluye:

• La contaminación del agua de los ríos por fuentes industriales y agrícolas;
La escasez de agua cruda principalmente debido a la sobreexplotación de las fuentes principales de agua, es decir, de los dos ríos cercanos de Barcelona (Llobregat y Ter);
• Formación de subproductos de la desinfección no deseados (DBPs) en el agua potable;
• Formación, distribución y el comportamiento de estos subproductos de desinfección a lo largo de la red de distribución de agua;
• Mejora del gusto del agua;

El agua cruda debe ser tratada para intentar eliminar los contaminantes y los patógenos, antes de su distribución y uso como agua potable. El diseño de un proceso adecuado de tratamiento del agua potable está impulsado por la adecuación de la calidad del agua de los diferentes origines del agua cruda. Hoy en día, las instalaciones de tratamiento de agua en el BMA emplean diversas tecnologías de tratamiento tales como: (a) desinfección con cloro o dióxido de cloro, ozonización, filtración con cartuchos de carbón activado granular (GAC); (b) ultra-filtración; (c) ósmosis inversa para el agua salobre y para el agua de mar (RO); (d) tratamiento del agua mediante operaciones de electrodiálisis reversible (EDR), y (e) la re-mineralización del agua.

Los trihalometanos (THM) son un conjunto de compuestos formados como subproductos de la desinfección, por la interacción química entre los compuestos orgánicos presentes en las aguas superficiales con los oxidantes como el cloro y dióxido de cloro (Rook, 1974; Richardson, 2002) utilizados para su desinfección. En particular, los compuestos investigados en esta tesis son el cloroformo ($CHCl_3$), el bromodiclorometano ($CHCl_2Br$), el clorodibromometano ($CHBr_2Cl$) y el bromoformo ($CHBr_3$). La presencia de estos compuestos en las aguas domésticas se considera nocivo debido a sus efectos negativos sobre la salud humana (McGeehin et al, 1993; Simpson y Hayes, 1998). Debido a que la desinfección es generalmente necesaria para garantizar la calidad del agua potable, es importante tener procedimientos de control y monitorización de la calidad del agua para poder predecir y evitar la formación de concentraciones altas de trihalometanos y mantenerlos a niveles por debajo de los límites recomendados para usos sanitarios.

Después del tratamiento de desinfección, el agua se distribuye a los clientes a través de un sistema de tuberías a presión, bombas, válvulas y tanques de almacenamiento, que forman parte del sistema de distribución de agua (WDS) de BMA. Los dos ríos (el río Llobregat y el Ter), el agua subterránea y el agua de mar se caracterizan por tener diferentes niveles de calidad del agua, los cuales, junto con los diferentes procedimientos de tratamiento empleados en las cinco plantas potabilizadoras explica por qué el agua potable suministrada es tan diferente en calidad y características organolépticas dentro de la misma BMA. En la red de distribución, se producen varios procesos de mezcla, que han sido implementados para homogeneizar la calidad del agua antes de su consumo y estandarizar las características estéticas de agua y asegurar el suministro constante de agua potable en la BMA (Valero y Arbós, 2010). Como elemento importante de la gestión eficiente de esta distribución del agua, se considera la posibilidad de identificar los origines de agua en mezcla en el interior del WDS. Este análisis es crítico para el funcionamiento adecuado de la WDS, así para eliminar la posible corrosión de las tuberías (Lahav et al., 2009). Por lo tanto, se necesita un método que permita distinguir los diferentes orígenes de agua y que facilite la identificación de posible contaminación accidentaria, así como la reducción de las quejas de los consumidores relacionadas con los problemas organolépticos.

Los procedimientos físico-químicos tradicionales empleados en el tratamiento del agua en una ETAP sirven para desinfectar el agua y eliminar los patógenos. Las nuevas tecnologías se implementan permanente con el objetivo de mejorar aún más la

calidad del agua cruda., Los procedimientos de filtración con membranas (EDR y RO) son métodos que permiten eliminar casi todos los contaminantes y la materia orgánica. Como inconveniente de estos procedimientos, se produce sin embargo también un efecto negativo sobre las propiedades organolépticas del agua (alteración del gusto del agua). El agua potable no sólo debe cumplir con las normas sanitarias, sino que también debe tener en cuenta las preferencias de los consumidores, incluyendo el gusto, el olor y el color del agua suministrada. Normalmente, se requieren procedimientos de re-mineralización (Vingerhoeds et al., 2016). El agua re-mineralizada ofrece a los consumidores la concentración adecuada de sales principales y también reduce la probabilidad de corrosión de las tuberías.

El sabor del agua depende considerable de la composición química de los minerales disueltos. Tanto los cationes como los aniones disueltos en el agua contribuyen a la formación del gusto del agua. También pueden interactuar entre ellos a través de efecto de sinergismo y antagonismo (Burlingame et al., 2007). Además de las sales inorgánicas disueltas (es decir, los sólidos disueltos totales, TDS), los compuestos orgánicos volátiles también afectan el sabor del agua. El gusto del agua se puede correlacionar con el contenido mineral del agua a partir de experimentos sensoriales con la participación de panelistas. El panel incluye un grupo de personas, entrenadas y familiarizadas con el análisis organoléptico de aguas naturales y de consumo. Para este propósito se ha trabajado con el método de análisis del perfil olfato-gustativo (FPA, Flavour Profile Analysis) modificado en AGBAR (Devesa et al., 2004), que consiste en una descripción individual de las características de olor y gusto de una muestra de agua mediante una serie de descriptores e indicando sus intensidades. Después de los experimentos sensoriales, el estudio continúa con un análisis estadístico de los datos recogidos (Naes and Risvik, 1996)

Después del su consumo, el agua usada se transporta al sistema de alcantarillas de recogida de aguas residuales. La calidad del agua se ha deteriorado gravemente y se requiere un tratamiento de las aguas residuales antes que el agua pse vierta de nuevo a los hábitats naturales, como a los ríos o al mar Mediterráneo. En este caso, las plantas de tratamiento de aguas residuales juegan un papel importante en el ciclo urbano del agua mediante la aplicación de diversos procesos biológicos, físicos y químicos que permiten eliminar los desechos del influente y restaurar la calidad del agua. El uso de aguas residuales que tienen una buena calidad en agricultura puede ayudar a la conservación del agua cruda (de mayor calidad). Los procesos de tratamiento de aguas

residuales presentan un gran desafío en términos de implementación y control. Se reconoce que los procesos de tratamiento de aguas residuales son dinámicos y complejos, debido a las variaciones de la calidad de aguas residuales municipales en el influente. Las aguas residuales en el influente se caracterizan por tener un flujo dinámico y una composición química muy diversa, relacionada con las actividades industriales o urbanas, episodios de lluvia y otros. La monitorización y el control permanente de la calidad de las aguas residuales forman parte importante de la gestión eficiente del ciclo urbano del agua, y favorecen la protección del medio ambiente, y así permitir un crecimiento sostenible. Para cumplir con las normas sanitarias y legislativas, se requiere la monitorización y el control permanente de la calidad del agua cruda, durante los procesos de tratamiento en ETAP, de la calidad del agua en el transporte en WDS, y en los procesos de tratamiento en EDAR. Esta monitorización y control de la calidad del agua se realizan a partir del seguimiento permanente de un gran número de contaminantes ambientales, de los compuestos DBPs, y de varios parámetros físico-químicos de calidad del agua. Las técnicas más comunes para la monitorización y control de la calidad de las aguas incluyen: (a) los métodos estándar de análisis en laboratorio, (b) la monitorización con sensores y (c) diversas técnicas de análisis instrumental. Los parámetros físico-químicos de calidad del agua, tales como la temperatura del agua, la turbidez, la salinidad, el TOC, la conductividad o el pH, se pueden medir empleando sensores de detección o mediante la aplicación de las metodologías estándar de análisis (APHA, 1995). Estos tipos de monitorización por lo general no son caros y no exigen personal cualificado. Por el contrario, el análisis de compuestos tóxicos de contaminación, tales como pesticidas, fármacos, disruptores endocrinos, DBPs y otros, requieren el empleo de técnicas instrumentales como por ejemplo LC-MS / MS cromatografía líquida con detección por espectrometría de masas en tándem (Kuster et al., 2008) o LC-TOF-MS (cromatografía líquida con detección por espectrometría de masas de tiempo de vuelo, Martinez Bueno et al., 2007), que pueden detectar concentraciones muy bajas (en pg/L). Sin embargo, estos tipos de análisis son caros y requieren personal cualificado para llevar a cabo dichos los experimentos.

Debido a que la mayoría de los compuestos orgánicos en el agua cruda, potable y residual, absorben la radiación ultravioleta (UV), la espectroscopia de UV se ha convertido en una alternativa de bajo precio para la monitorización de la calidad de las aguas (Langergraber et al., 2004). Además, los sensores UV tienen otra ventaja - que pueden generar una gran cantidad de datos en un período corto de tiempo (Rieger et al.,

2006). Los datos de calidad del agua, registrados con sensores UV, se pueden analizar mediante métodos de análisis multivariante de datos. La estrategia de registrar conjuntamente datos de parámetros fisicoquímicos y de espectros de UV puede ser muy útil para obtener información sobre la evolución de la calidad del agua a lo largo del tiempo; facilitar la determinación de la variabilidad espacial (es decir, comparar entre puntos de muestreo) de la calidad del agua en las plantas ETAP, EDAR o en la WDS. Por otra parte, el análisis multivariante de la calidad del agua se puede aplicar para la detección de eventos en tiempo real y en sistemas de alerta temprana. Además, los modelos multivariante en tiempo real pueden utilizar para predecir instantáneamente la calidad del agua en diferentes puntos de observación donde se encuentra el sensor o una estación automática de monitorización multi-paramétrica.

Esta Tesis es coherente con la política medioambiental de la UE, que intenta evitar la contaminación del agua y la preservación de los recursos hídricos. Esta tesis está relacionada directamente con dos directivas de la Unión Europea: a) la Directiva 98/83/CE relacionada con la calidad del agua potable, y b) la Directiva 91/2271/ CEE relacionada con el tratamiento de las aguas residuales urbanas.

a) La directiva 98/83/CE está relacionada con la calidad del agua potable destinada al consumo humano. Esta directiva tiene como objetivo establecer nuevas normas de calidad para muchos parámetros microbiológicos, tóxicos y organolépticos. Entre los parámetros investigados en esta Tesis están los trihalometanos (subproductos de la desinfección en ETAP). La Directiva 98/83/CE regula la concentración máxima de la suma de todos THMs, que debe ser inferior a 100 mg/l.

b) La Directiva 91/2271/CEE está relacionada con el tratamiento de las aguas residuales urbanas. Esta directiva procura una reducción de la contaminación de las aguas superficiales, por ejemplo, mediante la mejora de la calidad de las aguas residuales urbanas vertidas al medio ambiente después de su tratamiento en EDAR.

En general, cada componente del ciclo urbano del agua está diseñado para generar mejoras sociales y beneficios económicos, su gestión de manera eficiente es un gran desafío. Para conseguir un ciclo urbano del agua sostenible, lo más importante es el mantenimiento adecuado de todos los procesos de tratamiento de agua, a partir de la aplicación de sistemas eficientes de monitorización y control de la calidad del agua. Con el fin de posibilitar el desarrollo de soluciones sostenibles, en primer lugar hay que entender los desafíos y los problemas encontrados en el funcionamiento del círculo urbano del agua

5.4 Resultados

Esta tesis incluye seis artículos que presentan diferentes estudios basándose en la aplicación de diversos métodos quimiométricos en la investigación de varios problemas relacionados con la calidad del agua en su círculo urbano. Los artículos están organizados en tres bloques de acuerdo con los objetivos principales.

Modelización de la formación de trihalometanos en una ETAP y en condiciones de laboratorio basándose en métodos quimiométricos

La formación de trihalometanos (THMs) en la estación de tratamiento de agua potable de Sant Joan Despí (ETAP-SJD) está asociado con los procedimientos de desinfección, empleados para garantizar la calidad del agua potable. Estos procedimientos pueden favorecer la formación de altas concentraciones de THMs y pueden incumplir con los límites legislativos de la UE. Por eso es tan importante identificar cuáles son los principales factores que afectan a la formación de trihalometanos y predecir sus concentraciones.

En este primer bloque, se presentan tres artículos relacionados con la investigación del problema de formación de THMs.

Los resultados de la modelización quimiométrica de la formación de los subproductos de desinfección THMs ($CHCl_3$, $CHBr_3$, $CHCl_2Br$, y $CHBr_2Cl$), generados en la ETAP-SJD se presentan en los dos primeros artículos. Ambos trabajos investigan las relaciones lineales y no lineales entre los THMs en el agua potable y diversos parámetros de operación y funcionamiento, controlados en diferentes procesos y que han sido monitorizados en diferentes ubicaciones del tratamiento. Diversos métodos quimiométricos, como el análisis de componentes principales (PCA), la regresión lineal múltiple (MLR), la regresión de componentes principales (PCR) y la regresión de mínimos cuadrados parciales (PLS), la regresión Kernel de mínimos cuadrados parciales (K-PLS) y la regresión con máquinas de vectores de soporte (SVR), han sido aplicados y comparados a la hora de predecir y visualizar las concentraciones de los trihalometanos y de los parámetros importantes que son responsables de su formación.

El tercer artículo presenta una metodología para la interpretación de la formación y especiación de los THM, durante la reacción entre la materia orgánica

natural (NOM) y el cloro en experimentos de desinfección simulada en laboratorio y utilizando procedimientos quimiométricos. Se han aplicado dos tipos de diseños experimentales con el objetivo de evaluar y describir algunos de los principales factores influentes en la reacción de formación de trihalometanos.

En el primer artículo la correlación entre veintitrés parámetros del funcionamiento de la planta potabilizadora ETAP-SJD, medidos en diferentes puntos del proceso de desinfección durante un año, se ha analizado en relación con la formación de los trihalometanos medidos de forma independiente en el laboratorio por cromatografía d gases. Se han aplicado y comparado el método de análisis de componentes principales (PCA), la regresión lineal múltiple (MLR), la regresión de componentes principales (PCR) y la regresión de mínimos cuadrados parciales (PLS) en la modelización de la formación de los trihalometanos medidos en el agua potable a la salida de la planta. Los resultados, obtenidos con PCA, la interpretación de la significancia estadística de los coeficientes de regresión lineal obtenidos (MLR), y la evaluación de los variables latentes (en PCR y PLS) revelaron que los parámetros más importantes para la formación de trihalometanos eran: la temperatura del agua, la concentración del carbono orgánico total, las concentraciones de cloro disuelto, la absorbancia ultravioleta, la turbidez, así como otros parámetros como las cantidades añadidas del agua subterránea y la edad de los filtros de los cartuchos de carbón activado granular utilizados. En general, los métodos MLR y PLSR consiguieron resultados predictivos con errores de predicción similares. Se han obtenido las mejores predicciones (errores de predicción más bajos) para la suma de todos los trihalometanos, THMs. Los errores relativos en su predicción fueron de 12% (en la calibración) y de 16% (en la validación externa) respectivamente. Entre los trihalometanos individuales, las concentraciones de $CHBr_3$ fueron predichas con errores relativos altos, entre 21-25% (en la calibración) y 29-31% (en la validación externa), seguidos por $CHCl_2Br$ con 23-26% y 25-27%. Las mejores predicciones fueron obtenidas para las concentraciones de $CHBr_2Cl$ con errores relativos entre 17% y 21%, y para las concentraciones de $CHCl_3$ con 21-24% y 23-25%.

El segundo artículo ha sido continuación del trabajo hecho en la investigación previa con un la incorporación de medidas sobre parámetros nuevos. Los 23 parámetros de funcionamiento medidos en ETAP-SJD en el primer estudio, fueron obtenidos en las últimas etapas del proceso de desinfección (después de las etapas de filtración con arena), los cuales eran controlados desde dentro de la planta. Varios parámetros

importantes de la ETAP–SJD no eran incluidos en este primer análisis. Se sospechó por lo tanto, que estos parámetros podían ser críticos para mejorar la capacidad de predicción de los modelos. La revisión de la literatura sugirió que algunos parámetros de calidad medidos en ETAP-SJD, como los del agua cruda y los del pre-tratamiento inicial con cloro, podían ser importantes para el proceso de formación de trihalometanos. Por eso en este segundo estudio, se incorporaron dieciocho nuevos parámetros medidos en el agua cruda y en las etapas de precloración (las dos primeras etapas omitidas en el primer estudio). Se hizo un nuevo análisis del conjunto de datos con el objetivo de mejorar los resultados de la predicción de THMs del estudio anterior. La nueva investigación analizó las correlaciones entre los cuarenta y uno parámetros de la ETAP-SJD durante un período de un año.

Como novedad de este segundo estudio ha sido también la aplicación de técnicas de regresión no lineales, tales como la regresión de Kernel por mínimos cuadrados parciales (K-PLS) utilizando las funciones de base radial y la regresión con máquinas de soporte vectorial (SVR). EL objetivo fue la modelización de las posibles relaciones no lineales entre los parámetros de funcionamiento y del agua cruda y las concentraciones de THMs en el agua potable en la salida de la planta.

La modelización lineal basándose en los 41 parámetros de ETAP-SJD resultó mejor con errores de predicción entre 2-4% más bajos en la validación externa en comparación a cuando se utilizaron 23 parámetros (en el primer estudio). La incorporación de los nuevos 18 parámetros resultaba pues importante para la mejora en la predicción global de THMs.

Los mejores resultados se han logrado sin embargo con la aplicación de K-PLSR y SVR. Los errores relativos de predicción fueron entre 6% y 10% más bajos en comparación con los obtenidos aplicando métodos de regresión lineal. Se encontró que la mejor predicción se obtuvo para la suma de todos THMs con errores relativos de predicción de 13,6%. En este segundo estudio la predicción del cloroformo y del bromoformo se ha mejoraron con errores de 8-10% más bajos en comparación con los errores obtenidos en el primer estudio.

También en este estudio, se han evaluado las posibilidades de interpretar los resultados de las técnicas de regresión no lineal. . En este estudio se conseguía la visualización e interpretación de las interacciones entre los parámetros en los modelos de K-PLSR y SVR. Las observaciones demostraron que es posible discutir las relaciones no lineales entre los parámetros de monitorización en ETAP-SJD. Entre los

nuevos parámetros más importantes para la formación de los THMs destacaron la temperatura del agua cruda, la oxidabilidad, la turbidez, los cloruros y el consumo de cloro al principio de la precloración. También se han confirmado como influentes todos los parámetros destacados en el primer estudio.

Finalmente, en el tercer artículo se han estudiado los factores que influyen de forma más importante en la formación de trihalometanos entre diferentes fracciones (parte coloidal, parte hidrofóbica y parte transphilica) de la materia orgánica disuelta (DOM) a partir de la reacción con cloro en la desinfección simulada en el laboratorio de soluciones acuosas. La DOM fue fraccionada utilizando agua de embalse de Caldeirão (Portugal). Esta investigación implicó el diseño experimental factorial por el procedimiento Plaquet-Burman con cinco factores (la concentración de DOM, la dosis de cloro, la temperatura, el pH y la concentración de bromuros) y un segundo diseño experimental factorial Box-Behnken para un análisis detallado de tres de los factores más importantes (la concentración de DOM, la dosis de cloro y la temperatura). Los resultados mostraron que la fracción coloidal tiene una contribución relativamente baja a la formación de THM. La fracción transphilica es importante para la formación aproximadamente del 50% de del cloroformo. La fracción hidrofóbica fue la fracción más importante para la formación de los THM bromados. Cuando se desinfectaron soluciones acuosas con fracción coloidal y fracción hidrofóbica, la mayor concentración de la fracción de DOM generó mayores concentraciones de THM. El aumento del pH produjo mayores niveles de cloroformo y reducción de los del bromoformo; dosis altas de cloro y temperaturas altas produjeron un aumento de la formación de THM en total, y más específicamente de los THMs clorados. Por otra parte la mayor concentración de bromuros generó mayores concentraciones de los THM bromados. Se aplicaron modelos lineales de mezcla y se obtuvieron gráficos de superficie de respuesta para las cuatro concentraciones de THM y para su suma en total, en función de la concentración de DOM, de la dosis de cloro, y de la temperatura. En general, los resultados indicaron que los modelos de formación de THM son muy complejos debido a los efectos de los factores individuales y de las interacciones significativas entre los factores.

Lista de artículos científicos incluidos en este bloque:

1. **Article 1** – Platikanov, S., Puig, X., Martin, J. and R. Tauler. *Chemometric modeling and prediction of trihalomethane formation in Barcelona's water works plant.* **Water Research** 41 (2007) 3394-3406.

2. **Article 2** – Platikanov, S., Martin, J. and R. Tauler. *Linear and non-linear chemometric modeling of THM formation in Barcelona's water treatment plant.* **Science of Total Environment** 432 (2012) 365-374.

3. **Article 3** – Platikanov, S., Tauler, R., Rodriguez, P., Antunes, M., Pereira, D. and J. Esteves da Silva. *Factorial Analysis of the trihalomethane formation in the reaction of colloidal, hydrophobic and transphilic fractions of DOM with free chlorine.* **Environmental Science and Pollution Research** 17 (2010) 1389-1400.

Modelización quimiométrica de datos espectrales de ultravioleta y de parámetros físico-químicos en el agua potable y en el agua residual

El segundo bloque incluye dos artículos con estudios de modelización quimiométrica utilizando datos espectrales de ultravioleta (UV) y parámetros físico-químicos del agua potable o residual con el objetivo de mejorar la monitorización de la calidad del agua en la red de distribución del agua de Barcelona (WDS) y en la salida de la planta de tratamiento de aguas residuales (EDAR) cerca de la ciudad de Girona (Cataluña).

Las fuentes principales de agua cruda (superficial, subterránea o del mar), agua tratada (diferentes instalaciones y proceso de desinfección) o agua residual (urbana o industrial) se caracterizan por su materia orgánica natural (NOM) y presentan una composición química específica. En función de esta composición de NOM, las fuentes principales del agua cruda presentan características muy particulares o huellas (fingerprint) en sus espectros de UV, porque en general están asociados con su origen geográfico. En los estudios incluidos en los dos artículos se ha investigado la posibilidad de monitorizar y predecir la calidad del agua aplicando quimiometría a datos espectrales.

En el primer trabajo se han desarrollado modelos quimiométricos capaces de predecir diferentes fuentes u orígenes del agua potable (en función del su origen

geográfico o del tratamiento empleado en ETAP) en sus mezclas, situación típica de la red de distribución de Barcelona.

El área metropolitana de Barcelona tiene una red de distribución (WDS) del agua potable muy larga y sofisticada, en continua reconstrucción. Hasta cinco diferentes orígenes de agua de suministro se pueden distinguir en esta red - diferentes plantas de tratamiento de agua potable tratando agua cruda procedente de los dos ríos (el Llobregat y el Ter), de acuíferos y del mar Mediterráneo. El conocimiento de las fuentes de agua en diferentes lugares de la WDS a lo largo del tiempo facilita a los operadores mejorar la gestión global del sistema de distribución.

Los métodos empleados se han basado en el uso combinado de datos espectrales de UV entre 190 y 270 nm, y en la aplicación de la regresión de mínimos cuadrados parciales (PLS). A partir de las diferencias en la composición de la materia orgánica de las diferentes fuentes de agua y de sus características espectrales, el modelo PLS fue capaz de determinar las cantidades relativas (en mezclas diseñadas) de los dos principales origines de agua cruda - los ríos el Llobregat y el Ter, con muestras tomadas en diferentes lugares (agua de la red de distribución, del grifo) de WDS de Barcelona. A continuación, se ha ampliado este estudio al desarrollo de un nuevo método con el objetivo de determinar las cantidades relativas de las aguas potables presentes en mezclas diseñadas sintéticamente, a partir de agua de las cinco plantas de tratamiento de agua potables de Barcelona. Durante esta modelización se ha requerido información adicional de otros parámetros que se ha incorporado al modelo PLS. En particular, la determinación y cuantificación del origen del agua potable procedente de una planta de desalinización, se ha logrado cuando las concentraciones de boro ese han incluidos en el análisis.

En el segundo artículo de este bloque se estudian las correlaciones entre diferentes parámetros físico-químicos de funcionamiento y los datos espectrales de UV en relación con la calidad del agua residual en la entrada y la salida de una planta de tratamiento de aguas residuales real. También se realizaron varios experimentos diseñados en laboratorio con mezclas sintéticas de agua.

Los métodos quimiométricos tales como el análisis de componentes principales (PCA) y la regresión por mínimos cuadrados parciales (PLS) se han aplicado para explorar y analizar los procesos de tratamiento, para comparar y evaluar las técnicas de monitorización de los parámetros de calidad del agua en una planta de tratamiento de aguas residuales (EDAR). Diferentes conjuntos de datos (obtenidos del análisis de

laboratorio rutinario y por un sistema automático de seguimiento o monitorización multi-paramétrica con un nuevo dispositivo espectral) han sido investigados con los métodos quimiométricos. Se han detectado variaciones en la de calidad de agua monitorizada en el influente y en el efluente de la planta depuradora en función del tratamiento. También, los resultados obtenidos han permitido la investigación de las relaciones más importantes entre los parámetros monitorizados y de su dependencia cíclica en el tiempo (variación diaria, mensual y anual). En este estudio también se ha investigado la modelización y predicción de las concentraciones de varios de los parámetros fisicoquímicos del agua que son relevantes en la evaluación de la calidad del agua. Los parámetros de modelización han sido la materia orgánica disuelta (DOM), el carbono orgánico total (TOC), las concentraciones de nitratos, de detergentes, y del fenol. Estos modelos PLS han sido desarrollados para correlacionar las concentraciones de un determinado parámetro en función de los espectros de UV, medidos en muestras recogidas en: (1) en el laboratorio a partir de mezclas de agua sintética y pre-diseñadas); y (2) in-situ en una EDAR, con muestras reales de agua residual medidas en la planta. En el estudio con las mezclas de agua sintética pre-diseñadas, se seleccionaron longitudes de onda específicas que permitieran establecer modelos simples y fiables de predicción. De esta manera, se obtuvieron resultados de predicción con errores relativos de 3% a 4% para las concentraciones de los nitratos, de los detergentes y del fenol. El error relativo en la predicción de DOM era 15% en la validación externa. La predicción de los nitratos y de la TOC en muestras reales medidas en el efluente de la planta ha sido bastantes buenas con errores de predicción bajos (menos de 20%).

Lista de artículos científicos incluidos en este bloque:

1. **Article 4** – Platikanov, S., Garcia, V., Landeros, E., Devesa, R., Matía, L., Tauler, R., *Determination of water supply sources in the Barcelona distribution system by UV spectrophotometry and PLS.* **Water Science and Technology-Water Supply** 11 (2011) 45-54.

2. **Article 5** – Platikanov, S., Rodriguez-Mozaz, S., Huerta, B., Barcelo, D., Cros, J., Batlle, M., Poch, G., Tauler, R. *Chemometrics quality assessment of wastewater treatment plant effluents using physicochemical parameters and UV*

absorption measurements. **Journal of Environmental management** 140 (2014) 33-44.

Aplicación de métodos quimiométricos a datos relacionados con el gusto del agua y el análisis sensorial

El Área Metropolitana de Barcelona (BMA) suministra agua potable principalmente procedente de los dos ríos -el río Llobregat y el rio Ter. Como consecuencia del clima local, los recursos hídricos están expuestos a sequías cíclicas. En términos de la gestión del agua potable y de su suministro, tiene también un efecto negativo la contaminación del agua debido a la industria minera y a los vertidos industriales en las dos cuencas hidrográficas, afectando gravemente a la calidad del agua cruda en las entradas de ETAP. Las tecnologías de filtración por membranas y de desalación de agua de procedente del mar permiten mejorar la calidad y cumplir con la demanda del agua potable, como nuevos recursos alternativos. Dichas tecnologías son muy eficientes en la eliminación de compuestos tóxicos, la materia orgánica y los patógenos, pero su uso requiere una re-mineralización para mejorar las propiedades organolépticas del agua potable y para minimizar el efecto corrosivo de muchas sales.

En general la calidad del agua potable es apreciada por la gente. Por ejemplo, se sabe que el gusto del agua depende de la composición química (contenido de sales) y que cationes y los aniones contribuyen en manera diferente a partir de mecanismos de interacción por sinergismo y antagonismo. Las preferencias del gusto del agua en la red de distribución del agua de la BMA han sido investigadas durante varios años. Sin embargo, se requiere mejorar el conocimiento sobre los efectos de los compuestos químicos en la satisfacción de los gustos de la gente. Los experimentos del análisis sensorial empleando panelistas entrenados pueden proporcionar información valiosa para entender mejor el gusto del agua.

En este artículo, se realizó un análisis quimiométrico de dos conjuntos de datos sensoriales obtenidos en estudios separados. Han sido examinadas veinte aguas embotelladas por panelistas entrenados en un primer estudio, veinte aguas embotelladas y veinticinco aguas embotelladas o de la red de distribución en el segundo estudio. . Los panelistas han expresado sus preferencias del gusto de cada muestra del agua calificando de 0 (el peor gusto) a 10 (el mejor gusto del agua). La puntuación media de todos los panelistas se correlaciono con las propiedades físico-químicas de las mismas

muestras. Trece diferentes parámetros físico-químicos fueron analizados en ambos estudios. En el segundo estudio se añadieron los niveles de cloro residual de aguas del sistema de distribución. Se aplicó el Análisis de componentes principales (PCA) a los parámetros físico-químicos y a la puntuación promedio de todos los panelistas. Los modelos PCA desarrollados han explicado la mayor parte de la varianza (información) presente en los datos experimentales. Por otra parte, se ha aplicado la regresión de mínimos cuadrados parciales (PLS) para correlacionar las puntuaciones promedio de los panelistas con los datos físico-químicos, con el objetivo de explicar cuales son los motivos de estos evaluaciones de los panelistas. Los resultados obtenidos en los dos estudios surgirieron que las muestras preferidas de agua se caracterizaban con un contenido de minerales de concentración moderada y con unas concentraciones de HCO_3^-, SO_4^{2-}, Ca^{2+} y Mg^{2+} relativamente altas y con valores de pH relativamente alto. Las altas concentraciones de Na^+, K^+ y Cl^- se puntuaron con valores bajos por parte de muchos de los panelistas, mientras el cloro residual no afectó a las puntuaciones. Se ha comprobado que la presencia de cloro residual ha permitido a los panelistas distinguir entre muestras de aguas embotelladas y aguas del grifo.

Lista de artículos científicos incluidos en este bloque:

1. **Article 6** – Platikanov, S., Garcia, V., Fonseca, I., Rullan, E., Devesa, R., Tauler, R., *Influence of minerals on the taste of bottled and tap water: A chemometric approach.* **Water Research** 47 (2013) 693-704.

5.5 Conclusiones

Conclusiones relacionadas con el análisis y monitorización de la calidad del agua

- Se ha confirmado que en el caso de la planta de potabilización de aguas de Sant Joan Despí, la formación de trihalometanos dependía de los parámetros de funcionamiento de la planta de tratamiento en el proceso de desinfección y de algunos parámetros ambientales típicos que afectan a la calidad de agua del río. Los parámetros ambientales más importantes han sido la temperatura del agua, la absorbancia ultravioleta a 254 nm (que refleja el contenido en materia orgánica), las concentraciones del carbono orgánico total, la conductividad del agua, la oxidabilidad y la turbidez. Estos parámetros son controlados en el agua original del río y a lo largo del proceso de desinfección en la planta. Los parámetros de funcionamiento más importantes para la formación THMs durante la desinfección han sido: la dosis del cloro y la edad (tiempo de uso) de los filtros de carbón.

- Se ha confirmado que la formación de trihalometanos en la planta de San Juan Despí empieza durante la etapa de pre-cloración y continúa en las etapas de post-cloración. El proceso de filtración con carbón activo y la adición temporal de determinadas cantidades de agua subterránea reduce significativamente la formación de trihalometanos.

- Se ha detectado que la formación de trihalometanos está caracterizada por una variación estacional apreciable. La estación con más formación de trihalometanos ha resultado ser la primavera, con predominio de las especies bromadas de los THMs y con disminución del cloroformo.

- Se ha confirmado, que la materia orgánica disuelta (DOM) ha sido un factor muy importante para la especiación de los THMs durante su formación. La fracción coloidal de DOM generó aproximadamente 20-30% de cada THMs y aproximadamente 25% de los trihalometanos formados en total. La

fracción hidrofóbica ha sido la más relevante en la formación de los trihalometanos bromados. En cambio la fracción transphilica?? resulta más importante en la formación del cloroformo. Ambas fracciones, resultan ser responsables de la formación de alrededor el 80% de los THMs totales.

- Se ha confirmado la importancia de la presencia inicial en el agua de bromuro en la formación de los trihalometanos bromados durante la desinfección. Cantidades pequeñas de bromuros en el agua de entrada del rio alterara significativamente la formación de trihalometanos bromados.

- En esta tesis se ha demostrado que la espectrofotometría ultravioleta visible acoplada con los métodos quimiométricos puede ser una herramienta analítica poderosa para el análisis de la composición de mezclas de agua de diferente funete7origen. Se pudo diferenciar hasta cinco diversas fuentes/orígenes del agua potable para Barcelona, basándose en el contenido de materia orgánica natural, detectable a partir de espectrofotometría UV-VIS. Además, se pudo predecir las concentraciones de varios parámetros de calidad del agua, tales como materia orgánica disuelta, nitratos, detergentes y fenoles.

- Se ha utilizado la espectrofotometría ultravioleta visible, individualmente o de forma complementaria con los parámetros fisicoquímicos, para la modelización y seguimiento de la calidad del agua.

- Se han detectado las variaciones temporales y espaciales (estacionales, de actividad urbana diurna y nocturna, de los cambios de calidad del agua en el influente o en el efluente) en la calidad de las aguas residuales, a partir del análisis quimiométrico conjunto de los datos espectrales y fisicoquímicos. A partir del análisis de estos datos se pueden detectar los eventos no típicos y poco frecuentes, tales como los episodios de precipitación y accidentes de contaminación del agua también.

- Se ha demostrado que la mineralización total de las aguas embotelladas y del grifo ha sido el factor más influyente que determina el gusto por el agua de los panelistas empleados en los experimentos sensoriales. Ha sido más

apreciada la mineralización total del agua, entre 200- 400 mg/L, (expresada en forma de sólidos disueltos totales). No han sido tan apreciadas las aguas con mineralización total por encima de 800 mg/L y por debajo de 50 mg/L. Se ha detectado que un gran número de panelistas que preferían aguas con concentraciones relativamente altas de Ca^{2+}, Mg^{2+}, SO_4^{2-} y HCO_3^-. Se ha detectado también que existe un grupo más reducido de panelistas que preferían las aguas con altas concentraciones de K^+, Na^+ y Cl^-.

- Se ha detectado que el cloro residual en al agua del grifo (en los niveles en que se encuentra y a la temperatura de los ensayos) no influye de forma determinante en el gusto de los panelistas, sin embargo, los panelistas usaban este parámetro para discriminar entre las muestras de agua embotellada y del grifo.

Las conclusiones relacionadas con la aplicación de los métodos quimiometría en el análisis de datos y monitorización de calidad del agua son las siguientes:

- Se ha comprobado la validez de diversos métodos quimiométricos para el análisis de problemas relacionados con la determinación y seguimiento de la calidad del agua. En esta Tesis, se ha demostrado la eficacia de los métodos del análisis exploratorio de datos multivariante, tales como el Análisis de Componentes Principales (PCA), y de los métodos de calibración multivariante., tales como los métodos de análisis de regresión múltiple (MLR), regresión de mínimos cuadrado parciales (PLS), regresión mediante máquinas de soporte vectorial (Support Vector Machine Regressión), regresión Kernel no lineal de Mínimos Cuadrados Parciales (K-PLS). Se ha comprobado la aplicabilidad de la estrategia de diseño experimental mediante superficies de respuesta, que ha resultado ser una herramienta útil y económica en el descubrimiento de los factores más importantes en la formación de trihalometanos.

- Se ha confirmó la eficacia de las diversas técnicas y herramientas de visualización de PCA y PLS a partir de los gráficos de los 'scores', 'loadings' y 'weights'. Esta visualización ha permitido a descubrir las fuentes principales (factores) de varianza de los datos en los casos investigados. El análisis de los 'scores' ha permitido detectar la variabilidad temporal, las variaciones semanales, diurnas o nocturnas en la formación de THM, la actividad urbana en relación con la calidad del agua residual y el comportamiento de los diversos grupos de panelistas en función de su preferencia por el gusto del agua. También se han analizado los mapas de distribución de los scores que permiten diagnosticar la variabilidad relacionada con los cambios dinámicos en la calidad del agua residual en la entrada o en la salida de la planta de trataminto; los eventos no típicos del proceso de tratamiento, tales como los episodios de precipitaciones y de accidentes no deseados de contaminación del agua. Todo esto confirma la utilidad de la metodología quimiométrica utilizada para análisis de control de calidad.

- La visualización de los 'loadings', y 'loadings weights' (pesos de los variables) en PCA y PLS, hizo posible identificar diversos parámetros ambientales y de operación de planta, que han sido importantes en la formación de trihalometanos durante la desinfección del agua potable; permiten determinar cuáles han sido las variables con mayor influencia en la dinámica de la calidad del agua de las aguas residuales y cuáles son los compuestos minerales más determinantes del gusto del agua embotellada o del grifo.

- En esta tesis, se ha confirmado la utilidad de las nuevas técnicas de visualización obtenidas durante la modelización no lineal con K-PLS y SVR, que permiten revelar cuales son las relaciones no lineales entre los varios parámetros ambientales y de operación en la planta potabilizadora.

- El procedimiento de evaluación de la importancia de las variables en proyección (Variable Importance in Projection, VIP) ha resultado útil para la selección óptima de un conjunto reducido de diferentes parámetros físico-

químicos, de longitudes de onda espectrales, y de compuestos minerales en la predicción de la calidad del agua.

- Las técnicas de regresión tales como MLR, PLS y K-PLS han producido estimaciones precisas de los parámetros de calidad investigados. Se recomienda la aplicación rutinaria de estas técnicas para la modelización de los compuestos y parámetros de interés en estudios relacionados con la calidad del agua.

- Se ha comprobado que el análisis factorial basado en el diseño experimental es muy útil para comprender los factores principales que afectan a la formación de THMs durante la desinfección del agua. Los diseños experimentales aplicados en esta Tesis, tales como el Plaquet-Burman y el Box-Behnken, han sido muy útiles para seleccionar y evaluar la importancia de los parámetros influyentes y sus efectos e interacciones en la formación de trihalometanos.

Chapter 6

References

References

1. 140/2003 Real Decreto 140/2003, de 7 de febrero, por el que se establecen los criterios sanitarios de la calidad del agua de consumo humano

2. 2000/60/EC Directive 2000/60/EC of the European Parliament and of the Council of 23 October 2000 establishing a framework for Community action in the field of water policy, L 327/1.

3. 75/440/EEC Council Directive 75/440/EEC of 16 June 1975 concerning the quality required of surface water intended for the abstraction of drinking water in the Member States

4. 76/160/EEC Council Directive of 8 December 1975 concerning the quality of bathing water. Brussels, European Commission, 1975

5. 76/464/EEC Council Directive 76/464/EEC of 4 May 1976 on pollution caused by certain dangerous substances discharged into the aquatic environment of the Community.

6. 80/778/EEC Council Directive 80/778/EEC of 15 July 1980 relating to the quality of water intended for human consumption

7. 91/271/EEC Council Directive 91/271/EEC of 21 May 1991 concerning urban waste-water treatment.

8. 98/83/EC Council Directive 98/83/EC of 3 November 1998 on the quality of water intended for human consumption, L 330/32.

9. Abdullah, A., El-dien Hussona, S. 2013. Predictive model for disifection by-product in Alexandria drinking wáter, northern west of Egypt. . Environ Sci Poll Res 20, 7152-7166.

10. Adin, A., Katzhendler, J., Alkaslassy, D., Rav-Acha, C. 1991. Trihalomethane formation in chlorinated drinking water: a kinetic model. Water Res 25, 797–805.

11. adsorption. J Environ. Sci. Technol. 15, 463-466.

12. Al-Mutaz, I. 1989.Treated wastewaters as a growing water resource for agriculture use. Desalination 73, 27-36.

13. Al-Omari, A., Fayyad, M., Qader, A. 2004. Modeling trihalomethane formation for Jabal Amman water supply in Jordan. Environ Model Assess 9, 245–252.

14. Al-Omari, A., Muhammetoglu, A., Karadirek, E., Jiries, A., Batarseh, M., Topkaya, B., Soyupak, S. 2014. A review on formation and decay kinetics of trihalomethanes in wáter of different qualities. CLEAN-Soil Air Water 42, 1-14.

15. Amine, H., Gomez, E., Halwani, J., Casallas, C., Fenet, H.2012. UV filters, ethylhexyl methoxycinnamate, octocrylene and ethylhexyl dimethyl PABA from untreated wastewater in sediment from eastern Mediterranean river transition and coastal zones. Mar. Pollut. Bull. 64, 2435-2442.

16. Amy, G., Chadik, P., Chowdhury, Z.1987. Developing models for predicting trihalomethanes formation potential kinetics. J Am Water Works Assoc 79, 89–97.

17. Amy, G., Siddiqui, M., Ozekin, K., Zhu, H., Wang, C. 1998. Empirical based models for predicting chlorination and ozonation byproducts: haloacetic acids, chloral hydrate, and bromate. USEPA EPA report CX 819579.

18. Analytical Methods Committee. AMCTB No 55. Experimental design and optimisation (4): Plackett–Burman designs. Anal Methods. 2013;5:1901- 3.

19. APHA, AWWA, WEF 1995. Section 4500-Cl DPD colorimetric method in A. Eaton, L. Clesceri, A. Greenberg (eds.), Standard methods for the examination of water and wastewater. 20th ed. American Public Health Association, Washington, DC

20. APHA, AWWA, WEF, 1998. Standard Methods for the Examination of Water and Wastewater, 20th ed. American Public Health Association, Washington, DC.

21. APHA, AWWA., WEF, 2005. Standard methods for the Examination of Water and Wastewater. 21ª edición . Article 2170 Flavour Profile Analysis. American Public Health Association, Washington, DC.

22. AWWA, American Water Works Association. 2000. Disinfection systems survey committee report (May, 2000), Water Quality Division. J. Am. Water Works. Assoc. 9: 24-43.

23. AWWA, American Water Works Association 2005. History of water. http://watermuseum.tripod.com/id1.html (accessed April 2016).

24. AWWARF, American Water Works Association Research Foundation. 1996. Characterisation and Modeling of Chlorine Decay in Distribution Systems. AWWA, USA.

25. Bach, L., Garbelini, W., Stets, S., Peralta-Zamora, P., Emmel, A. 2015. Experimental design as tool for studying THMs formation parameters during water chlorination. Microchemical journal 123,252-258.

26. Bakeev, K. 2005. Process analytical technology. Blackwell publishing, Oxford, UK.

27. Baxter, C., Zhang, Q., Stanley, S., Shariff, , R., Tupas, R., Stark, H. 2001. Drinking water quality and treatment: the use of artificial neural networks. Can. J. Civil Engin. 28, suppl. 1, 26–35.

28. Baxter, G. 1995. Chlorine disinfection: The industry standard. Water Supply 13, 183-193.

29. Bellar, T., Lichtenberg J., Kroner, R. 1974. The occurrence of organohalides in chlorinated drinking waters. J. Am. Water Works Assoc. 66, 703 –706.

30. Boleda, M., Diaz, A., Marti, I., Martin-Alonso, J., Matia, L., Romero, J., Ventura, F. 2007. A review of taste and odour events in Barcelona's drinking wáter área (1990-2004). Water science and technology 55: 217-221

31. Box, G., Behnken, D. 1960. Some new three level designs for the study of quantitative variables. Technometrics 2, 455-475.

32. Brereton, R.G. 2003. Chemometrics: Data Analysis for the Laboratory and Chemical Plant. Chichester, UK: John Wiley & Sons

33. Bridge, C.J. 2005. Holistic Trihalomethane Mangement of Poorer Quality Surface Waters. PhD Thesis, The University of Birmingham.

34. Burlingame, G.A., Dietrich, A.M., Whelton, A.J., 2007. Understanding the basics of tap water taste. J. Am Water Works Assoc. 99, 100-110.

35. Carlson, M., Hardy, D. 1998. Controlling DBPs with monocholoramine. J AWWA 90, 95-106.

36. Céspedes, R., Lacorte, S., Ginebreda, A. Barceló, D. 2006. Chemical monitoring and occurrence of alkylphenols, alkylphenol ethoxylates, alcohol ethoxylates, phthalates and benzothiazoles in sewage treatment plants and receiving waters along the ter River basin (Catalonia, N. E. Spain). Anal. Bioanal. Chem. 385, 992-1000.

37. Céspedes, R., Lacorte, S., Ginebreda, A. Barceló, D. 2007. Occurrence and fate of alkylphenols and alkylphenol ethoxylates in sewage treatment plants and impact on receiving waters along the Ter River (Catalonia, NE Spain). Environ.Poll. 153, 384-392.

38. CETaqua, Chris Fife-Schaw 2010. Strategies for addressing water shortages: Barcelona – A case study. TECHNEAU project 0183320 under Sixth Framework Programme.

39. Chang, C., Lin,C. 2010. LIBSVM: A Library for Support Vector Machines, National Taiwan University, http://www.csie.ntu.edu.tw/~cjlin/libsvm/

40. Chang, E., Chao, S., Chiang, P., Lee. J. 1996. Effects of chlorination on THM formation in rawwater. Toxicol Environ Chem 56, 211–225.

41. Chang, E.E., Lin, Y.P., Chiang, P.C. 2001. Effects of bromide on the formation of THMs and HAAs. Chemosphere 43, 1029-1034.

42. Chen, B., P. Westerhoff. 2010. Predicting disinfection by-product formation potential in water. Water Research 44, 3755-3762.

43. Chowdhury, S., Champagne, P. 2008. An investigation on parameters for modeling THMs formation. Global NEST Journal 10, 80-91.

44. Chowdhury, S., Champagne, P., McLellan, PJ. 2009. Models for predicting disinfection byproduct (DBP) formation in drinking waters: A chronological review. Sci. Tot. Environ. 407, 4189-4206.

45. Chowdhury, S., Champagne, P., McLellan, PJ. 2010. Factorial analysis of trihalomethanes formation in drinking water. Water Environ Res 82, 556-566.

46. Clark, R., Sivaganesan, M. 1998. Predicting chlorine residuals and formation of TTHMs in drinking water. J Environ Eng 124, 1203–1210.

47. Clark, R., Sivaganesan, M. 2002. Predicting chlorine residuals in drinking water: A Second Order Model. J Water Resour Plan Manage. 128, 152-161.

48. Clark, R., Thurnau, R., Sivaganesan, M., Ringhand, P. 2001. Predicting the formation of chlorinated and brominated by-products. J Environ Eng 2001 127, 493–501.

49. Cristianini, N., Shawe-Taylor, J. 2000. An introduction to support vector machines and other kernel based learning methods. Cambridge University Press. UK.

50. Croué, J.P. 1999.Isolation, Fractionation, Characterization and Reactive Properties of Natural Organic Matter. In: Proceedings of the Australian Water Works Association 18th Federal Convention. Adelaide, Australia.

51. Croué, J.P. 2004. Isolation of humic and non-humic nom fractions: structural characterization. Environ Monit Assess 92, 193–207.

52. Daszykowski, M., Serneels, S., Kaczmarek, K., Van Espen, P., Croux, C., Walczak, B. 2007. TOMACAT: A MATLAB toolbox for multivariate calibration techniques. Chem. Intell. Lab. Sys 85, 269-277.

53. Davis, S., Whittemore, D., Fabryka-Martin, J.1998. Uses of chloride/bromide ratios in studies of potable water. Ground Water 36, 338–50.

54. Dayal, B., MacGregor, J. 1997. Improved PLS algorithms. J. Chemometrics 11, 73-85.

55. Deborde, M., von Gunten, U. 2008. Reactions of chlorine with inorganic and organic compounds during water treatment - kinetics and mechanisms: a critical review. Water Res. 42, 13-51.

56. Devesa, R., García, V., Matía. L. 2010. Water flavour improvement by membrane (RO and EDR) treatment. Desalination 250, 113–117.

57. Devesa, R., Cardeñoso, R., and Matía, L. 2007. Contribution of the FPA tasting panel to decision making about drinking water treatment facilities, Water Sci. Technol. 55, 127–135.

58. Devesa, R., Fabrellas, C., Cardeñoso, R., Matia, L., Ventura, F., Salvatella, N., 2004. The panel of Aigües de Barcelona: 15 years of history. Water Sci. Technol., 49, 145-151.

59. Dietrich, A., 2006. Aesthetic issues for drinking water. J. Water Health 4, suppl.1, 11-16.

60. Dietrich, A.M. 2009. The sense of smell: contributions of orthonasal and retronasal perception applied to metallic flavor of drinking water. J Water Supp Res Technol- AQUA 58, 562-570.

61. Dodds, L., King, W. 2001. Relation between trihalomethane compounds and birth defects. Occup. Environ.Med. 58, 443-446.

62. Draper, N., Smith, H. 1981. Applied regression analysis.2nd edition, Wiley, NY, USA

63. Eikebrokk, B. 2004. NOM increase in Northern European source waters: Impacts on coagulation and filtration processes. Proc. NOM Innovations and Applications, 2-5 March, Victor Harbor, Adelaide, Australia.

64. Elshorbagy, W., Abu-Qdais, H., Elsheamy, M. 2000. Simulation of THMs species in water distribution systems. Water Res 34, 3431–3439.

65. Engerholm, B., Amy, G. 1983. A predictive model for chloroform formation from humic acid. J Am Water Works Assoc 75, 418–423.

66. Esbensen, K., Guyot, D., Westad, F., Houlmoller, L. 2000. Multivariate Data Analysis - in Practice: An Introduction to Multivariate data analysis and experimental design. Camo Press. P.127.

67. Espadaler, I., caixach, J., Om, J., Ventura, F., Cortina, M., Pauné, F, Rivera, J. 1997. Identification of organic pollutants in Ter river and its system of reservoirs supplying water to Barcelona (Catalonia, Spain): A study by GC/MS and FAB/MS. Water Res. 31, 1996-2004

68. Fernandez-Turiel, J., Rodriguez, G., Carnicero, M., valero, F. 2003. Spatial and seasonal variations of wáter quality in a Mediterranean catchment: the Llobregat River(NE Spain). Environ. Geochem. Health 25, 453-474.

69. Fisher, R. 1971. The design of experiments. (9th ed.). Macmillan.

70. Freese, S., Nozaic, D. 2004. Chlorine: Is it really so bad and what are the alternatives? Water South Africa 30, pp. 18-24.

71. Friedman, J. 1991. Multivariate adaptive regression splines. Ann. Statist. 19, 1-67.

72. Friedman, J.H.1984. SMART:User's guide. Technical Report LCS01, Dept. Stat., Stanford University.

73. Gallagher, D., Dietrich, A. 2014. Statistical approaches for analyzing customer complaint data to assess aesthetic episodes in drinking water. J Water Supply Res. Technol. 63, 358-367.

74. Gallard, H., von Gunten, U. 2002. Chlorination of natural organic matter: kinetics of chlorination and of THM formation. Water Res 36, 65–74

75. Gang, D., Segar, R., Clevenger, T., Banerji, S. 2002. Using chlorine demand to predict THM and HAA9 formation. J Am Water Works Assoc 94, 76–86.

76. Garcia, I. 2005. Removal of natural organic matter by enhanced coagulation in Nicaragua. PhD Thesis, Department of chemical engineering and technology, Stockholm University, Sweden.

77. García, V., Fernandez, A., Medina, M., Ferrer, O. Cortina, J., Valero, F., Devesa, R. 2015. Flavour assessment of blends between desalinated and conventionally treated sources. Desalin. Water Treat. 53, 3466-3474.

78. García, V., Fernandez, A., Medina, M., Ferrer, O., Cortina, J., Valero, F., Devesa, R. 2015. Flavour assessment of blends between desalinated and conventionally treated sources. Desalination and wáter treatment 53, 3466-3443.

79. Garcia-Villanova, R., Garcia, C., Gomez, J., Garcia, M., Ardanuy, R. 1997a. Formation, evolution and modelling of trihalomethanes in the drinking water of a town: I. At the municipal treatment facilities. Water Res 31,1299–1308.

80. Garcia-Villanova, R., Garcia, C., Gomez, J., Garcia, M., Ardanuy, R.1997b. Formation, evolution and modelling of trihalomethanes in the drinking water of a town: II. In the water distribution system. Water Res 31, 1405–1413.

81. Geladi, P., Esbensen, K., 1991. Regression on multivariate images: Principal Component Regression for modeling, prediction and visual diagnostic tool. J. Chemometrics, 5, 97-111.

82. Geladi, P., Kowalski, B. 1986. Partial Least Squares Regression: a Tutorial, Anal. Chim. Acta, 185, 1-17.

83. Gemperline, P. 2006.Principal component analysis in : Gemperline (ed), Practical guide to chemometrics. second edition. CRC Press, Boca Raton, USA

84. Ginebreda, A., Barata, C., Barceló, D. 2012. Risk assessment of pollutants in the Llobregat river. In S. Sabater et al. (eds), The Llobregat: The Story of a polluted Mediterranean river. Hdb. Env. Chem. 2012, 21, 236-296. Springer-Verlag, Berlin

85. Golfinopoulos, S., Xilourgidis, K., Kostopoulou, N,, Lekkas, T. 1998. Use of a multiple regression model for predicting trihalomethane formation. Water Res 32, 2821–2829.

86. Golfinopoulos, S., Arhonditsis, G. 2002. Multiple regression models: a methodology for evaluating trihalomethane concentrations in drinking water from raw water characteristics. Chemosphere 47, 1007-1018.

87. Golub, G., Van Loan, C. 1996.Matrrix computations. 3rd edition. Johns Hopkins University Press;

88. Graves, C., Matanoski, G., Tardiff, R. 2001. Weight of evidence for an association between adverse reproductive and developmental effects and exposure to disinfection by-products: a critical review. Regul.Toxicol. Pharmacol. 34, 103-124.

89. Gunn, S. 1998. Support vector machines for classification and regression. Technical report, University of Southampton. UK. http://www.isis.ecs.soton.ac.uk/isystems/kernel/svm.zip

90. Haas, C.N. (1999). Disinfection. In: Letterman, R.D (Ed.). Water Quality and Treatment, 5th Edition, American Water Works Association, Denver, Colorado, pp. 14.1-14.60.

91. Harrington, G., Chowdhurry, Z., Owen, D. 1992. Developing a computer model to simulate DBP formation during water treatment. J Am Water Works Assoc 84(11), 78–87.

92. Health Canada, Guidelines for Canadian Drinking Water Quality, 2012.

93. Hespanhol, I. 1997. Chapter 4. Wastewater as a resource. In R. Helmer and I. Hespanhol, Water pollution control- a guide to the use of water quality management principles.WHO/UNEP, E&FN SPON.

94. Himmelblau, D. 2008. Accounts of experiences in the application of artifiacial neural networks in chemical engineering. Ind, Eng. Chem.res. 47 (16), 5782-5796.

95. Hong, H., Liang, Y., Han, B., Mazumder, A., Wong, M. 2007. Modeling of trihalomethane (THM) formation via chlorination of the water from Dongjiang River (source water for Hong Kong's drinking water). Sci Total Environ 385, 48–54.

96. Ibarluzea, J., Goni, F., Santamaria, J. 1994. Trihalomethanes in water supplies in the San Sebestian area Spain. Bull Environ Contam Toxicol 52, 411–418.

97. Jain, R., Wang, R. 2008. Limitations of maximum likelihood estimation procedures when a majority of the observations are below the limit of detection. Anal Chem. 80, 4767-4772.

98. Johnson J. D. and Jensen J. N. 1986. THM and TOX formation: routes, rates, and precursors. J. AWWA 78, 156-1620.

99. Jolliffe, I.T., 2002. Principal Component Analysis 2nd Ed, Springer Verlag, Berlin, Germany.

100. Kao, LJ., Chiu, C.C.,Lu, C.J.,Yang, J.L. 2013.Integrating of nonlinear independent component analysis and support vector regression for stock price forecasting. Neurocomputing 99, 534-542.

101. Kennard, R., Stone, L. 1969. Computer design of experiments. Technometrics 11, 137-148.

102. King, W., Marrett, L. 1996. Case-control study of bladder cancer and chlorination by products in treated water (Ontario, Canada). Cancer Causes and Controls 7, 596-604.

103. Kitis, M., Karanfil, T., Kilduff, J.E., Wigton, A. 2001. The reactivity of natural organic matter to disinfection by-products formation and its relation to specific ultraviolet absorbance. Water Sci. Technol. 43, 9-16.

104. Kolla, R. 2004. Formation and modeling of disinfect ion by-products in Newfoundland communities. Masters Thesis; Memorial University of Newfoundland, NL, Canada.

105. Krasner, S.W., McGuire, M.J., Jacangelo, J.G. Patania, N.L., Reagan, K.M., Aieta, E.M. 1989. The occurrence of disinfection by-products in US drinking water. J AWWA 81, 41-53.

106. Krasner, S.W., Croué, J.P., Buffle, J, Perdue, E.M. 1996. Three approaches for characterizing NOM. J AWWA 88, 66-79.

107. Krasner, S. 1999. Chemistry of disinfection by-products formation on formation and control of disinfection by-products in drinking water, Chapter 2. American Water Works Association.

108. Kroehler, C. 2014. Potable water quality standards and regulations: a historical and world overview. In T.Younos and A. Grady (eds.), Potable water: Emerging global problems and solutions. Springer International Publishing, New York, pp. 1-36.

109. Kuster, M., López de Alda, M.J., Hernando, M.D., Petrovic, M., Martín-Alonso, J., Barceló, D., 2008. Analysis an occurrence of pharmaceuticals, estrogens, progestogens and polar pesticides in sewage treatment plant effluents, river water and drinking water in the Llobregat river basin (Barcelona, Spain). J. Hydrol. 358, 112–123.

110. Kwak, J. 2005. Application of Taguchi and response surface methodologies for geometric error in surface gringing process. Int. J. Mach Tools Manuf. 45, 327-334.

111. Lahav, O., Salomons, E., Ostfeld, A., 2009. Chemical stability of inline blends of desalinated, surface and ground waters: the need for higher alkalinity values in desalinated water. Desalination 238, 334–345

112. Langergraber, G., Gupta, J.K., Pressl, A., Hofstaedter, F., Lettl, W., Weingartner, A., Fleischmann, N. 2004. Online monitoring for control of a pilot-scale sequencing batch reactor using a submersible UV/VIS spectrometer. Water Sci. Technol. 50, 73–80.

113. Leenheer, J., Rostad, C., Barber, L., Schroeder, R., Anders, R., Davisson, M. 2001. Nature and chlorine reactivity of organic constituents from

reclaimed water in groundwater, Los Angeles County, California. Environ. Sci. Technol. 35, 3869–3876.

114. Leenheer, J.A. 2004. Comprehensive assessment of precursors, diagenesis, and reactivity to water treatment of dissolved and colloidal organic matter. Water Sci. Technol.: Water Supply 4, 1-9.

115. Lekkas, T., Nikolaou, A. 2004. Development of predictive models for the formation of trihalomethanes and haloacetic acids during chlorination of bromide ion rich water. Water Qual Res J Can 39,149–159.

116. Liu, X., Liu, D., et al., 2014. Optimal support vector regression algorithms for multifunctional sensor signal reconstruction. TELKOMNIKA Ind. J. Electr.. Engin. 12, 2762-2768.

117. López-Roldán, R., Platikanov, S., Martín-Alonso, J., tauler, R., González, S., Cortina, J. 2016. Integration of Ultraviolet-Visible spectral and physicochemical data in chemometrics analysis for improved discrimination of water sources and blends for application to the complex drinking water distribution network of Barcelona. J. Clean. Prod. 112, 4789-4798.

118. Luts J., Ojeda, F., Van de Plas, R., De Moor, B., Van Huffel, S., Suykens, J. 2010. A tutorial on support vector machine-based methods for classification problems in chemometrics. Anal Chim Acta 665, 129-145.

119. Måge, E. Menichelli, T. Næs. Preference mapping by po-pls: separating common and unique information in several data blocks. Food Quality and Preference, 24 (2012), pp. 8–16.

120. Malcolm Pirnie Inc. Bay-Delta water quality modeling report No. 15-041. Los Angeles: Metropolitan Water District of Southern California; 1993

121. Malcolm Pirnie Inc. Water Treatment Plant Simulation Program, Version 1.21, User's manual. Washington, D.C.: U.S. Environmental Protection Agency (EPA); 1992.

122. Marcé, R., Honey-Roses, J., Manzano, A., Moragas, L., Catllar, B., Sabater, S. In S. Sabater et al. (eds), The Llobregat: The Story of a polluted Mediterranean river. Hdb. Env. Chem. 2012, 21, 1-26. Springer-Verlag, Berlin

123. Marín D, Juncà S, Massagué A, Cortina J, Fonseca I, Valero F (2012) .Impacts on climate change of three drinking water treatment plants supplying Barcelona Metropolitan Area, Proceedings of the IWA Water, Climate and Energy Congress, 13-18 May 2012, Dublin, Ireland

124. Martens, H., Næs, T. 1991. Multivariate calibration. Wiley. New York, 1991

125. Martens, H., Næs, T. 2001. Multivariate calibration by data compression. In: Near-Infrared Technology in the Agricultural and Food Industries (second edition), (Eds. Phil Williams and Karl Norris) American Association of Cereal Chemists, Minnesota, USA, pp 59-100. ISBN 1-891127-24-1.

126. Martin-Alonso, J. 2006. Managing resources in an European semi-arid environment: combined use of surface and groundwater for drinking water production in the Barcelona metropolitan area. In Stephen Hubbs (ed). Riverbank filtration Hydrology, Springer, Netherlands, pp 281.298.

127. Martin-Alonso, J., Devesa, R., matia, L. 2007.Managing an odour episode in barcelona's water supply: strategies adopted, the causative agent(diacetyl) and determination of its organoleptic properties. Water Sci Technol. 55, 209-216.

128. Martinčič, R., Kuzmanovski, I., Wagner, A., Novič, M. 2015. Development of models for prediction of the antioxidant activity of derivatives of natural compounds. Anal Chim Acta 868, 23-35.

129. Martínez Bueno, M.J., Agüera, A., Gómez, M.J., Hernando, M.D., García-Reyes, J.F., Fernández-Alba, A.R. 2007. Application of Liquid Chromatography/Quadrupole-Linear Ion Trap Mass Spectrometry and Time-of-Flight Mass Spectrometry to the Determination of Pharmaceuticals and Related Contaminants in Wastewater. Anal. Chem. 79, 9372–9384.

130. Mason, R., Gunst, R. 1985. Selecting principal components in regression. Stat. Probal. Lett. 3, 299-301

131. McGeehin, M.A., Reif, J.S., Becher, J.C., Mangione, E.J., 1993. Case-control study of bladder cancer and water disinfection methods in Colorado. Am. J. Epidemiol. 138, 492– 501.

132. Meng, A., Suffet, I. 1997. A procedure for correlation of chemical and sensory data in drinking water samples by principal component analysis. Environ. Sci. Technol. 31, 337-345.

133. Metcalf & Eddy Inc. Wastewater engineering: treatment and reuse. 4th ed. 2003. Revised by G. Tchobanoglous and F. Burton. New York: McGraw-Hill;

134. Mills, C., Bull, R., Cantor, K., Reif, J., Hrudey, S., Huston, P.1998. Health Risks of Drinking Water Chlorination By-products. Chronic Diseases in Canada 19, 91-102.

135. Milot, J., Rodrı´guez, M.J., Serodes, J.B., 2002. Contribution of neural networks for modeling trihalomethanes occurrence in drinking water. J. Water Res. Pl. ASCE 128 (5), 370–376.

136. Minear, R., Morrow, C.M.1983. Raw water bromide in finished drinking water, research report 9. Water Resources Research Center, University of Tennessee

137. Montgomery W. Mathematical modelling of the formation of THMs and HAAs in chlorinated natural waters. Denver , Colorado: Am. Water Works Assoc.; 1993.

138. Moody, J., Darken, C. 1989. Fast learning in networks of locally-tuned processing units. Neural Computation 1, 151-160.

139. Naes, T., Risvik, E., 1996. Multivariate analysis of Data in Sensory Science. Amsterdam: Elsevier

140. Newcombe, M. Drikas, S. Assemi, and R Beckett. 1997. The influence of characterized natural organic material on activated carbon adsorption: I characterization of concentrated reservoir water. Water Res. 31, 963-972.

141. Nieuwenhuijsen, M.J., Toledano, M.B., Eaton, N.E., Fawell, J., Elliot, P. 2000. Chlorination disinfection by-products in water and their association with adverse reproductive outcomes: a review. Occup. Environ. Med. 57, 73-85.

142. Nikolaou, A. 2004. Investigation of the formation of chlorination byproducts in water rich in bromide and organic matter content. J Environ Sci Health A39, 2835–2853.

143. Nikolaou, A., Lekkas, T., Golfinopoulos, S. 2004. Kinetics of the formation and decomposition of chlorination by-products in surface waters. Chem. Eng. J. 100, 139–148.

144. Nokes, C., Fenton, E., Randall, C. 1999. Modelling the formation of brominated trihalomethanes in chlorinated drinking waters. Water Res. 33, 3557 –3568.

145. Noori, R., Karbassi, A., Moghaddamnia, A., han, D., Zokaei-Ashtiani, M., Farokhnia, A., Ghafari Gousheh, M. 2011. Assessment of input variables determination on the SVM model performance using PCA, Gamma test, and forward selection techniques for monthly stream flow prediction . J Hydrology 401, 177-189.

146. Ostace, G., Baeza, J., Guerrero, J., Guisasola, A., Cristea, V., Agachi, P., Lafuente J. 2013. Development and economic assessment of different WWTP control strategies for optimal simultaneous removal of carbon, nitrogen and phosphorus. Comput. Chem. Engin. 53, 164– 177.

147. Otto, M. 1999. Chemometrics: Statistics and Computer Applications in Analytical Chemistry. Germany: Wiley-VCH.

148. Owen, D.M., Amy, G.L., Chowdhury, Z.K. 1993. Characterisation of natural organic matter and its relationship to treatability. J AWWA 85, 72-78.

149. Paraira M., West, M. Online methodology for determining trihalomethane formation potential and predicted network TTHM levels aids in water treatment operations. 2015 Water Quality Technology Conference. November 15-19, Salt Lake City. Utah, USA.

150. Rahardianto, A., Gao., Gabelich, C., Williams, M., Cohen, Y. High recovery membrane desalting of low-salinity brackish water: integration of accelerated precipitation softening with membrane RO. J. Membrane Science 289, 123-137.

151. Raich-Montiu, J., Barios, J., Garcia, V., Medina, M.E., Valero, F., Devesa, R., Cortina, J.L., 2014. Integrating membrane technologies and blending options in water production and distribution systems to improve organoleptic properties. The case of the Barcelona Metropolitan Area. J. Clean. Prod. 69, 250–259.

152. Rathbun, R. 1996a. Speciation of trihalomethane mixtures for the Mississippi, Missouri and Ohio rivers. Sci Total Environ 180, 125–135.

153. Rathbun, R. 1996b. Regression equations for disinfection by-products for the Mississippi, Ohio and Missouri Rivers. Sci Total Environ 191, 235–244.

154. Reckhow, D.A., Singer, P.C. 1985. Mechanism of organic halide formation during fulvic acid chlorination and implication respect to preozonation. In: Water Chlorination Chemistry: Environmental Impact and Health Effect, Vol.5, Lewis Publisher.

155. Reckhow, D.A., Rees, P., Bryan, D. 2004. Watershed sources of disinfection by-product precursors. Water Supply Technol. Water Supply 4, 62-69.

156. Rey-Salgueiro, L., Gosálbel-García, A., Pérez-Lamela, C., Simal-Gándara, J., Falque-Lopez, E., 2013. Training of panelists for the sensory control of bottled natural mineral water in connection with water chemical properties. Food Chem. 141, 625-636.

157. Richardson, S. 2003. Disinfection by-products and other emerging contaminants in drinking water. Trends Anal. Chem. 22, 666-684.

158. Richardson, S., Simmons, J., Rice, G. 2002. Disinfection byproducts: the next generation. Environ. Sci. Technol. 36, 198-205.

159. Richardson, S.D., Plewa, M.J., Wagner, E.D., Schoeny, R., DeMarini, D.M. 2007. Occurrence, genotoxicity, and carcinogenicity of regulated and emerging disinfection by-products in drinking water: A review and roadmap for research. Mutation Research, 636, 178-242.

160. Rieger, L., Langergraber, G., Thomann, M., Fleischmann, N., Siegrist, H. 2004. Spectral in .situ analysis of NO2, No3, COD, DOC, and TSS in the effluent of a WWTP. Water Sci. technol. 50, 143-152.

161. Rodriguez, M., Serodes, J., Morin, M. 2000. Estimation of water utility compliance with trihalomethanes regulations using modelling approach. J Water Supply: Res Technol 49, 57–73.

162. Rodriguez, M., Serodes, J. 2001. Spatial and temporal evolution of trihalomethanes in three water distribution systems. Water Res. 35, 1572-1586.

163. Rodriguez, M., Silva, J., Antunes, M.2007. Factorial analysis of the trihalomethanes formation in water disinfection using chlorine. Anal Chim Acta 595, 266–274.

164. Romero, J., Ventura, F., caixach, J., Rivera, J., Gode, LX., Niñerola J. 1998. Identification of 1,3-Dioxanes and 1,3-Dioxolanes as malodorous compounds at trace levels in river, water, groundwater, and tap water. Environmental Science and Technology 32, 206-216.

165. Rook, J. 1974. Formation of haloforms during chlorination of natural waters. Water Treatment Examination 23, 234 –243.

166. Rosipal R, Trejo L. Kernel Partial Least Squares regression in reproducing Kernel Hilbert Space. J Mach Learn Res 2001;2:97-123.

167. Rubin, D. 1976. Inference and Missing data. Biometrika, 63, 581-592.

168. Rubulis, J., Dejus, S., Meksa, R. 2011. Online Measurement Usage for Predicting Water Age from Tracer Tests to Validate a Hydraulic Model. American Society of Civil Engineers, 1488–1497.

169. Sabater, F., Armengol, J. 1986. Chemical characterization of the Ter River. Limnetica 2, 75-84.

170. Sabater, F., Sabater, S., Armengol, J. 1990.Chemical characteristics of a Mediterranean river as influenced by land uses in the watershed. Water Res. 24, 143-155.

171. Sadiq, R., Rodriguez, M. 2004. Disinfection by-products (DBPs) in drinking water and predictive models for their occurrence: a review. Sci. Tot. Environ. 32, 21-46.

172. Schölkopf, B., Smola, A. 2002.Learning with kernels: Support vector machines, regularization, optimization and beyond. MIT press.

173. Semerjian, L., Dennis, J., Ayoub, G. 2009. Modeling the formation of trihalomethanes in drinking waters of Lebanon.

174. Serodes, J., Rodriguez, M., Li, H., Bouchard, C. 2003. Occurrence of THMs and HAAs in experimental chlorinated waters of the Quebec City area (Canada). Chemosphere 51, 253–263.

175. Simpson, K.L., Hayes, K.P., 1998. Drinking water disinfection by-products: an Australian perspective. Water Res. 32 (5), 1522-1528.

176. Shawe-Taylor, J., Cristianini, N. 2004. Kernel methods for pattern analysis. Cambridge University Press

177. Shu, S., Liu, S., Wang, X., Yu, L., Shu, S., Zhang, D., Meng, M. 2010. Determination and applications of water age in distribution system. IEEE, 1918–1921. doi:10.1109/MACE.2010.5536510

178. Siddiqui, M., Amy, G., Ozekin, K., Westerhoff, P. 1994. Empirically and theoretically based models for predicting brominated ozonated by-products. Ozone: Sci Eng 16, 157–178.

179. Singer, P. 1999. Humic substances as precursors for potentially harmful disinfection by-products. Water Sci. Technol. 40, 25-30.

180. Singh, K., Rai, P., Pandey, P., Sinha, S. 2012. Modeling and optimization of THMs formation potential of surface water (a drinking water source) using Box-Behnken design. Environ Sci Poll Res 19, 113-127.

181. Sohn, J., Amy, G., Cho, J., Lee, Y., Yoon, Y. 2004. Disinfectant decay and disinfection by-products formation model development: chlorination and ozonation by-products.Water Res 38, 2461–2478.

182. Stanimirova, I. 2013. Practical approaches to principal component analysis for simultaneously dealing with missing and censored elements in chemical data . Analytica Chimica Acta, 796 27-37.

183. Suffet, I., Rosenfeld, P. 2007. The anatomy of odour wheels for odours of drinking water, wastewater, compost and the urban environment. Water Sci. technol. 55, 335-344.

184. Sung, W., Matthews, B., O'Day, K., Horrigan, K. 2000. Modeling DBP formation. Am Water Works Assoc 92, 53–63.

185. Talaiekhozani, A., Bagheri, M., Goli, A., Khoozani, M. 2016. An overview of principles of odour production, emission , and control methods in wastewater collection and treatment systmes. J. Environ. Manag. 170, 186-206.

186. Teillet, E., Urbano, C., Cordelle, S., Schlich, P., 2010. Consumer perception and preference of bottled and tap water. J. Sens. Stud. 25, 463-480.

187. Teixeira, M., Nunes, L. 2011 The impact of natural organic matter seasonal variations in drinking water quality. Desalin. Water Treat. 36, 344-353.

188. Thomas, O., Burgess, C. 2007. UV-Visible Spectrophotometry of Water and Wastewater. Elsevier. Netherlands

189. Thurman, E.M., Malcom, R.L. 1981. Isolation of natural organic matter by resin

190. Todeschini, R., Consonni, V., Mauri, A., Pavan, A. 2004. Detecting "bad" regression models: multicriteria fitness functions in regression analysis. Analytica Chimica Acta 515, 199-208.

191. Trussel, R.R., Umphres, M. 1978. The Formation of Trihalomethanes. J AWWA 70, 604-612.

192. Urano, K., Wada, H., Takemassa, T. 1983. Empirical rate equation for trihalomethane formation with chlorination of humic substances in water. Water Res 17,1797–802.

193. USEPA, (US Environmental Protection Agency) 1992. The USEPA Water Treatment Plant model. Office of Ground water and Drinking water, US EPA, Cincinnati, USA

194. USEPA, (US Environmental Protection Agency) 1999. Disinfection profiling and benchmarking guidance manual, United States Environmental Protection Agency, EPA 815-R-99-013.

195. USEPA (US Environmental Protection Agency). 2001. National primary drinking water standards, USEPA; 816-F-01-007, USA

196. USEPA, (US Environmental Protection Agency) 2006. National Primary Drinking Water Regulations: Stage 2 Disinfectants and Disinfection Byproducts Rule: Final Rule. Federal Register, 71(2), January 4, 2006.

197. USEPA, 2015. National Secondary Drinking Water Regulation. http://water.epa.gov/drink/contaminants/index.cfm#Secondary. Accessed September 25.

198. Üstün, B., Melssen, W., Buydensm L. 2007. Visualisation and interpretation of support vector regression models. Anal. Chim. Acta. 595, 299–309.

199. Uyak, V., Toroz, I. 2005. Seasonal variations of trihalomethanes (THMs) in water distribution networks of Istanbul City. Desalination 176, 127–141.

200. Uyak, V., Toroz, I., Meric, S. 2005. Monitoring and modeling of trihalomethanes (THMs) for a water treatment plant in Istanbul. Desalination 176, 91-101.

201. Valero, F., Arbós, R. 2010. Desalination of brackish river water using electrodialysis reversal (EDR). Desalination 253, 170-174.

202. Valero, F., Barcelo, A., Medina, M., Arbós, R. 2013. Barcelona, three years of experience in brackish wáter desalination using edr to improve quality. New O&M procedures to redice low-value work and increase productivity. Desalination and water treatment 51, 1137-1142.

203. van der Aa, M. 2003. Classification of mineral water types and comparison with drinking water standards. Environ. Geol. 44, 554-563.

204. Vapnik V. The Nature of statistical Learning Theory. New York, USA: Springer-Verlag; 1995.

205. Vapnik, V., 1998.Statisticasl Learning Theory. Springer-Verlag, New York, USA.

206. Vasconcelos, J.J., Boulos, P.F., and Clark, R.M. (1997). Kinetics of chlorine decay. J. AWWA 89, 54-65.

207. Ventura, F., Boleda, M.R., Lloret, R., Martin-Alonso, J. 1998. Strategies for the identification of compounds causing odours in water: a study of creosote spills. Water Research 32, 503–509.

208. Ventura, F., Romero, J., Pares, J.1997. Determination of dicyclopentadiene and its derivatives as compounds cousing odors in groundwater suppliers. Environmental Science and Technology 31: 2368-74.

209. Venugopala Rao, P. 2005. Textbook of environmental engineering. Prentice-Hall, India

210. Vieira, P., Coelho, S. T. and Loureiro, D. (2004). Accounting for the influence of initial chlorine concentration, TOC, iron and temperature when modelling chlorine decay in water supply. J Water Supply Res T. – AQUA 53, 453-467.

211. Villanueva, C., Cantor, K., Cordier, S., Jaakkola, J., King, W., Lynch, C. 2004. Disinfection byproducts and bladder cancer: a pooled analysis. Epidemiology 15, 357-367.

212. Villanueva, C., Gagniere, B., Monfort, C., Nieuwenhuijsen, M.J., Cordier, S. 2007. Sources of variability in levels and exposure to trihalomethanes. Environ. Res. 103, 211-220.

213. Vingerhoeds, M.H, Nijenhuis-de Vries, M.A., Ruepert, N., van der Laan, H., Bredie, W.L, Kremer, S., 2016. Sensory quality of drinking water produced by reverse osmosis membrane filtration followed by remineralisation. Water Res. 94, 42-51.

214. von Gunten, U., Driedger, A., Gallard, H., Salhi, E. 2001. By-prodcuts formation during drinking water disinfection: a tool to assess disinfection efficiency. Water Res. 35, 2095–2099.

215. Walczak, B., Massart, D. 1996a The Radial Basis Functions - Partial Least Squares approach as a flexible non-linear regression techniques Anal. Chim. Acta 331, 177-185.

216. Walczak, B., Massart D. 1996b Application of Radial Basis Functions — Partial Least Squares to non-linear pattern recognition problems: diagnosis of process faults. Anal Chim Acta;331:187–193.

217. Walczak, B., Massart, D. 2001a. Dealing with missing data: part I. Chemom. Intell. Lab. Syst. 58, 15-27.

218. Walczak, B., Massart, D. 2001b. Dealing with missing data: part II. Chemom. Intell. Lab. Syst. 58, 29-42.

219. Wang , D., Wang, X., Tomi, Y., Ando, M., Shintani, T. 2006.Modeling the separation performance of nanofiltration membranes for the mixed salts solution. J. Membr. Sci. 280, 734-743.

220. Wang, G.S., Deng, Y.C., Lin, TF. 2007. Cancer risk assessment from trihalomethanes in drinking water. Sci. Tot. Environ 387, 86-95.

221. Warton, B., Heitz, A., Joll, C., Kagi, R. 2006. A new method for calculation of the chlorine demand in natural and treated waters. Water Res. 40, 2877-2884.

222. Westerhoff, P., Debroux, J., Amy, G., Gatel, D., Mary, V., Cavard, J. 2000. Applying DBP models to fullscale plants. Am Water Works Assoc 92, 89-102.

223. Whelton, A., Dietrich, A., Burlingame, G., Schechs, M., Duncan, S. 2007. Minerals in drinking water: Impacts on taste and importance to consumer health. Water Sci. Technol. 55, 283–291.

224. White, G. 2010. Handbook of chlorination and alternative disinfectants, John Wiley & Sons Inc. 5th edition, USA

225. White, G. 1986. The Handbook of Chlorination, 2nd Edition. Van Nostrand Reinhold, New York.

226. WHO, World Health Organisation 2005. Trihalomethanes in Drinking Water: Background Document for Development of WHO Guidelines for Drinking Water Quality. World Health Organisation, 2005.

227. WHO, World Health Organisation, 2011. Guidelines for drinking-water quality, (4th ed), Geneve, Switzerland.

228. Wise,B., Ricker, N., Veltkamp, D., Kowalski, B. 1990. A theoretical basis for the use of principal component models for monitoring multivariate processes. Process control asnd quality, 1, 41-51.

229. WMO, 2013. World Meteorological Organization. Planning of water-quality monitoring systems, Technical report series. World Meteorological Organization, Geneva.

230. Wold, S., M. Sjostrom and L. Erikson, 2001. PLS-regression: A basic tool of chemometrics. Chem. Intell. Lab. Syst. 58, 109-130.