

## A model for predicting THM presence in networks of water supply systems

F. Osorio<sup>1</sup>, D. Ribes<sup>2</sup>, A. González-Martínez<sup>1</sup>,  
J. M. Poyatos<sup>1</sup> & P. García<sup>3</sup>

<sup>1</sup>*Department of Civil Engineering, Institute of Water,  
University of Granada, Spain*

<sup>2</sup>*AGBAR Group (Water Works Holding), Spain*

<sup>3</sup>*Department of Statistic and O.R., University of Granada, Spain*

### Abstract

The European normative, implemented in the Spanish national legislation in 2003, reduced the acceptable maximum concentration of THMs in drinking water, from 150 to 100  $\mu\text{L}$ . This implies that many existent facilities could not achieve this requirement. Generally speaking, the problems related to THMs in the area studied occur in certain periods during the year, mainly in summer. The present concentration of trihalomethanes (THM) in networks samples of water supply systems was predicted based on a set of easy parameters measured continuously at the site, which evidently did not include THM concentration. Consequently, when these values started being high, it was possible to take necessary measures (e.g. to modify the water source, reduce chlorination, etc.) to prevent higher values which are not permitted by law. The analysis of disinfection by-products is expensive and to develop a simple tool that will allow measuring the concentrations of trihalomethanes, particularly in small, poorly resourced water supplies is therefore a worthy intention. Other studied aspects have been the season of sampling or the sampling point. Data of drinking water supply systems coming from four regions in Spain have been used. In total 893 complete analyses have been managed. For obtaining the model, a statistical analysis of multiple regression and diverse Anova tests have been carried out.

*Keywords: DBP, drinking water, model, THM.*



## 1 Introduction

The use of chlorine revolutionized water treatment at the beginning of the 20th Century. However, recent research has shown that chlorine, when added to the water supply, can produce dangerous compounds which may cause cancer [1–3]. More specifically, chlorine reacts with natural organic carbon in the water to form trihalomethanes [4]. This is a disinfection by-product (DBP) that is produced by the reaction of the naturally occurring organic compound in the water and the chlorine added as a disinfection agent [5]. Nevertheless, although there are a wide variety of organochlorine and organobromine by-products (THMs, halogenic acetic acids, haloacetonitriles, etc.) with mutagenous and carcinogenic properties, the most widely studied are trihalomethanes, which are regarded as chemical pollutants and are used as indicators of the total level of by-products produced by water chlorination [6]. The term, total trihalomethanes (TTHM), refers to the sum of various chlorination by-products: chloroform ( $\text{CHCl}_3$ ), bromodichloromethane ( $\text{CHCl}_2\text{Br}$ ), chlorodibromomethane ( $\text{CHClBr}_2$ ) and bromoform ( $\text{CHBr}_3$ ).

One field of study directly related to THMs is their modeling. The objective is to predict their concentration, based on other materials and substances present in the water. Such models can be applied at locations where it is difficult to perform the necessary analyses [7]. Indeed the analysis of disinfection by-products is expensive and to develop a simple tool that will measure the concentrations of trihalomethanes (THMs), particularly in small, poorly resourced water supplies is therefore a worthy intention. According to Kulkarni and Chellam [8], DBP mass concentrations are typically modeled empirically by linearly regressing each of the water quality parameters influencing DBP formation. Log-linear power functions are extensively employed to model THM and HAA formation e.g. [9]. For our study, multiple regression model analysis was used.

## 2 Materials and methods

### 2.1 Characterization of available data

The analyses upon which this study is based were composed of a total of 87 parameters, which should be included in the complete analyses described in current regulations (Council Directive 98/83/EC). The analyses are of real samples taken in the field in different populations and at two sampling points: (i) the water storage deposit; (ii) the water distribution network. The sampling point was one of the variables studied.

A total of 893 analyses were performed in populations in four regional communities in Spain. Three of these communities are located on the Mediterranean coastline, where the drinking water has the highest levels of THMs. Table 1 shows the main characteristics of the modeled analytical data.

It is important to note that, with a practical point of view, we have used a lot of real data from facilities and only parameters included in all of them were taken into account for the model. In this line some potential important



parameters couldn't be taken into account because they are not even included in the complete analysis described in the European Directive 98/83/CE. Those complete analyses were the data used in our study and therefore the basis for selection of the independent variables.

Table 1: Summary of variables measured in the network water samples.

		Region				
		Region 1	Region 2	Region 3	Region 4	Total
	<b>Samples number</b>	<b>342</b>	<b>112</b>	<b>209</b>	<b>230</b>	<b>893</b>
Total organic carbon (TOC), mg/L	Average	2,10	3,88	1,41	,93	1,86
	Standard Deviation	,66	2,12	,47	,85	1,33
Combined residual chlorine, mg/L	Average	,80	,30	,14	,04	,39
	Standard Deviation	,31	,22	,05	,05	,40
Free residual chlorine, mg/L	Average	,61	,37	,79	,42	,57
	Standard Deviation	,29	,38	,33	,45	,39
Bicarbonates, mg/L	Average	229,94	134,80	144,27	254,53	205,48
	Standard Deviation	76,40	79,12	36,79	68,03	82,43
Conductivity, $\mu\text{S}/\text{cm}$	Average	1172,76	373,40	1052,83	781,76	943,73
	Standard Deviation	615,89	171,56	261,56	496,93	546,37
Chloride, mg/L	Average	221,56	40,42	147,49	119,02	155,15
	Standard Deviation	153,67	21,31	78,98	154,57	142,71
Total THM, $\mu\text{g}/\text{L}$	Average	105,80	78,56	74,68	26,97	74,76
	Standard Deviation	63,61	47,59	29,44	35,45	57,55

## 2.2 Statistical methods

Multiple regression was performed to model the dependency of THMs values as well as One-way Analysis of Variance (ANOVA). SPSS v.15.0 software was used to manage and analyze data from the considered samples.

The inclusion/exclusion of variables in the model proposed was obtained by means of Forward Selection. In this procedure, in the F-test, variables that fit this criterion and which had the highest correlation with the dependent variable were added to the model. The F of the entry for each variable was calculated by the maximum likelihood method. Also considered in this model were diagnoses of collinearity and non-correlation by means of the tolerance level of each variable (and their corresponding inflation factors) and the Durbin-Watson test, respectively.

The statistical analyses considered the transformation of the response variable, namely, the square root of the trihalomethane value since the goodness-of-fit with this variable was much better than that obtained for the observed THM values. This treatment is common in data analysis since because of their

nature, certain variables have wider ranges of variation that underlie the fact that the presence/absence of certain values of independent variables has a multiplier effect on the dependent variable.

### 3 Discussion and results

#### 3.1 THMs by season of the year

By considering precipitation levels over recent years and dividing each time period into a dry season and a rainy season, it was possible to study the influence of the *climate factor* on the presence of THMs in the water of the populations analyzed.

Various studies verified that the greatest statistical significance was obtained by considering the dry season to be August, September, and October, and the rainy season to be the rest of the year. Figure 1 is a graph of confidence intervals (95%) for the means by regions in order to deduce the conclusions of the study.

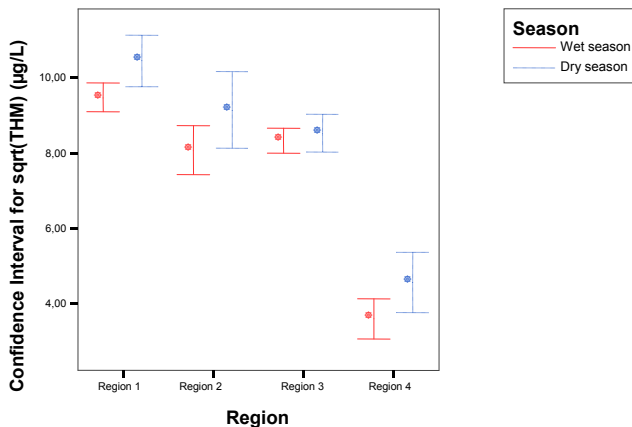


Figure 1: Confidence intervals (95%) for the means by regions.

Two-way ANOVA analysis showed that there were statistically significant differences for region ( $F_{3,887}=188.17$ ,  $p<0.001$ ) and season ( $F_{1,887}=14.075$ ,  $p<0.001$ ). Consequently, as expected, the mean THMs concentration in the dry season tended to be lower than the concentration in the rainy season.

#### 3.2 Obtaining a THMs prediction model based on easily controlled parameters

##### 3.2.1 Joint regression model for all the regional communities

As explained before, it is important to clear that, with a practical point of view, we have used a lot of real data from facilities and only parameters included in all of them were taken into account for the model. In this line some potential

important parameters couldn't be taken into account because they are not even included in the complete analysis described in the European Directive 98/83/CE. Those complete analyses were the data used in our study and therefore the basis for selection of the independent variables.

Bivariate correlations between the different variables showed that the most striking correlations were those between THM and the following variables: TOC, Conductivity, Combined Residual Chlorine, pH, Bicarbonates and Temperature.

After several previous studies, the models analyzed were composed of the following variables:

1. TOC (Total organic carbon)
2. TOC and ln (Conductivity)
3. TOC, ln (Conductivity) and Combined Residual Chlorine
4. TOC, ln (Conductivity), Combined Residual Chlorine, and pH
5. TOC, ln (Conductivity), Combined Residual Chlorine, pH and ln (Bicarbonates)
6. TOC, ln (Conductivity), Combined Residual Chlorine, pH, ln (Bicarbonates) and Temperature.

Of these six possibilities, the best was option 6 (TOC, conductivity, combined residual chlorine, pH, bicarbonates, and temperature) though the difference in goodness-of-fit was very similar in options 2 to 6.

Table 2 shows that, statistically speaking, the most significant variable was TOC, followed by ln (conductivity), and then, CRC, combined residual chlorine. On another order of magnitude, with a similar significance but at some distance from the first three are pH, ln (bicarbonates), and temperature, in that order. Moreover, the sign of ln (bicarbonates) is different from the rest. This means that the sense of its influence, implying its presence or absence, is the inverse of the other variables.

As a result, the obtained model was:

$$\text{sqrt(THM)} = -28.826 + 1.583 \cdot \text{TOC} + 2.713 \cdot \ln(\text{conductivity}) - 1.307 \cdot \ln(\text{bicarbonate}) + 3.744 \cdot \text{CRC} + 2.427 \cdot \text{pH} + 0.102 \cdot \text{Temperature}.$$

Table 2: Statistical summary of coefficients for the variables in the regression model.

	Non standardized coefficients		Standardized coefficients	t	Sig.
Dependent variable: sqrt(THM)	B	Error tip.	Beta		
Total organic carbon (TOC)	1,583	0,071	0,533	22,229	0,000
ln_Conductiv	2,713	0,131	0,459	20,736	0,000
ln_Bicarbonat	-1,307	0,192	-0,153	-6,804	0,000
Combined residual chlorine	3,744	0,253	0,399	14,773	0,000
pH	2,427	0,270	0,206	8,992	0,000
Temperature	0,102	0,021	0,115	4,783	0,000



In this model, the total THMs values were transformed by means of the square root (sq THMs). Chlorides were excluded since it was found that this variable is determined almost exactly by conductivity, which would have overspecified the model. Furthermore, the variables, conductivity and bicarbonates, were transformed by logarithm since the coefficients of the model were significantly low.

Figure 2 shows the fit of the observed data and the data predicted by the model for each region. The figure shows the goodness-of-fit proposed for each regional community. The prediction bands of each fit were acceptable. The worst prediction was obtained in the case of Region 4 because of the 0 THMs values. The data for Region 3 was fairly homogeneous, but compatible with the prediction of the model.

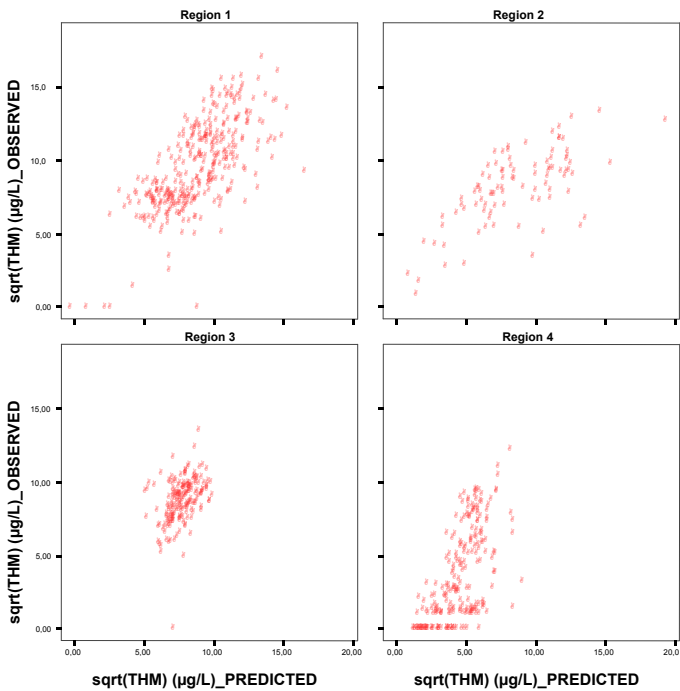


Figure 2: Bands of values of THM\_PREDICTED Vs. THM\_OBSERVED by regions.

Figure 3 shows the fit of the observed data and the data predicted by the model for each sampling point (i.e. water storage tank/water distribution network). The "storage tanks" are storage throughout the distribution zone, not post-treatment storage.

As can be observed, the fits have very similar prediction bands. For this reason, it is not informative to provide separate fits for each sampling point (water storage tank/water distribution network).



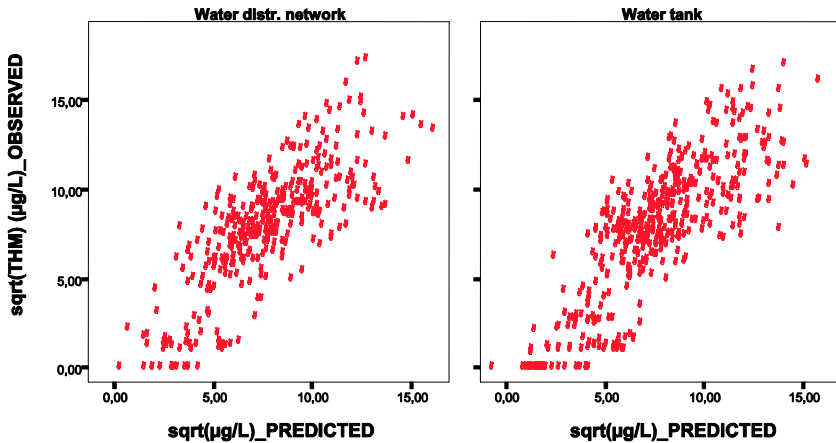


Figure 3: Bands of values of THM\_PREDICTED Vs. THM\_OBSERVED by sampling points.

## 4 Conclusions

By applying the simple model obtained in our research study, it is possible to always be aware of the risk of THM presence. This model alerts users in advance so that they can take appropriate measures before there are excessive concentrations of this substance, which are not permitted by law.

The analysis of the data collected in different populations and times of the year produced the following conclusions:

- The statistical study of the *climate factor* indicated that the greatest differences were between the dry season and the rainy season. The dry season includes the months of August, September, and October, and the rainy season is the rest of the year. These are the months in which a higher THMs concentration was expected. In any case, the differences were not statistically significant.
- A quick and accurate THMs Prediction Model was obtained. The most significant variables to predict the concentration of THMs by order of importance were the following: Total organic carbon, Combined residual chlorine, Conductivity, pH, Bicarbonates, Temperature. The TOC was found to be the most significant parameter, followed by combined residual chlorine and conductivity. The other parameters were at a greater distance from the others.
- A model with all six of these parameters would give the best fit. However, models that included from two to six of these parameters, (more specifically, parameters 2, 3, 4, 5, 6) would have a very similar fit coefficient.



## Acknowledgement

We gratefully acknowledge financial support for this work from AGBAR Group (Aguas de Barcelona Water Works Holding).

## References

- [1] Monarca, S., Zani, C., Richardson, S.D., Thruston, A.D., Moretti, J.M., Feretti D. & Villarini M., A new approach to evaluating the toxicity and genotoxicity of disinfected drinking water. *Water Research*, **38(17)**, pp. 3809-3819, 2004.
- [2] Cantor, K.P., Hoover, R., Hartge, P., Mason, T.J., Silverman, D.T. & Levin, L.I., Drinking water source and risk of bladder cancer: a case-control study. *Water chlorination: chemistry, environmental impact and health effects*. Vol. 5, ed. Jolley, R.L., Bull, R.J., Davis, W.P., Katz, S., Roberts, M.H., Jacobs & V.A., Lewis Publishers, Chelsea, MI., p. 145, 1985.
- [3] Cantor, K.P., Lynch, C.F. & Hildesheim, M., Chlorinated drinking water and risk of glioma: a case-control study in Iowa, USA. *Epidemiology*, **7(4)**(Suppl.), S83, 1996.
- [4] Abdullah, M.D., Yew, C.H. & Ramli, M.S., Formation, modeling and validation of trihalomethanes (THM) in Malaysian drinking water: a case study in the districts of Tampin, Negeri Sembilan and Sabak Bernam, Selangor, Malaysia. *Water Research*, **37(19)**, pp. 4637-4644, 2003.
- [5] Sadiq, R. & Rodriguez, M.J., Disinfection by-products (DBPs) in drinking water and the predictive models for their occurrence: a review. *Science of the Total Environment* **321(1-3)**, pp 21-46, 2004.
- [6] Reuber, M.D., Carcinogenicity of chloroform. *Environmental Health Perspectives*, **31**, pp. 171-182, 1979.
- [7] Nokes, C.J., Fenton, E. & Randall, C.J., Modelling the Formation of Brominated Trihalomethanes in Chlorinated Drinking Waters. *Water Research*, **33(17)**, pp. 3557-3568, 1999.
- [8] Kulkarni, P. & Chellam, S., Disinfection by-product formation following chlorination of drinking water: Artificial neural network models and changes in speciation with treatment. *Science of the Total Environment*, **408**, pp. 4202-4210, 2010.
- [9] Westerhoff, P., Debroux, J., Amy, G.L., Gatel, D., Mary, V. & Cavard, J., Applying DBP models to full-scale plants. *Journal of American Water Works Association*, **92**, pp. 89-102, 2000.
- [10] Council Directive 98/83/EC of 3 November 1998 on the quality of water intended for human consumption.

