Feasibility Study of a Simple and Low-Cost Device for Monitoring Trihalomethanes Presence in Water Supply Systems Based on Statistical Models

Almudena Rivadeneyra 1,*, Maria Jesús García-Ruiz 2,3, Fernando Delgado-Ramos 3, Alejandro González-Martínez 3, Francisco Osorio 2,3 and Ovidio Rabaza 4

1 Department of Electronics and Computer Technology, University of Granada, Granada 18012, Spain
2 Department of Civil Engineering, Section of Environmental Technology, University of Granada, Granada 18012, Spain; E-Mails: mjgruiz@ugr.es (M.J.G.-R.); fosorio@ugr.es (F.O.)
3 Institute of Water Research, University of Granada, Granada 18012, Spain; E-Mails: fdelgado@ugr.es (F.D.-R.); agon@ugr.es (A.G.-M.)
4 Department of Civil Engineering, Section of Electrical Engineering, University of Granada, Granada 18012, Spain; E-Mail: ovidio@ugr.es

* Author to whom correspondence should be addressed; E-Mail: arivadeneyra@ugr.es; Tel.: +34-958-249-517; Fax: +34-958-246-138.

External Editor: Yung-Tse Hung

Received: 28 July 2014; in revised form: 6 November 2014 / Accepted: 17 November 2014 / Published: 25 November 2014

Abstract: This paper describes a new method for predicting trihalomethanes (THMs) presence in networks of water supply systems, using a low-cost device that permits a fast monitoring of concentrations without need of complex analysis made in laboratories. This method, based on statistical models, allows the estimation of THM concentration by monitoring parameters whose determination is direct and easy, and therefore, THM presence can be carried out in real-time. These parameters values are introduced in a multiple regression model resulting in the concentration of THMs levels. This model has taken into account parameters compulsory in water quality analysis and it has been shown that six parameters are enough to determine accurately THM concentration. In addition, the feasibility of a low-cost device that directly gives THM concentration is demonstrated. This device can be easily designed to be transported to different points of the water supply network where it is intended to make control campaigns.
1. Introduction

The use of chlorine revolutionized water treatment at the beginning of the 20th century; in fact the chlorination is currently the most extended system for disinfection throughout the world. 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 (THMs) [4]. The THMs 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].

In addition to this, compliance with the European Council Directive 98/83/CE [6] and the deeper understanding in the society about environmental conservation and preservation has brought as a result the establishment of strict requirements for wastewater treatment plants as well as drinking water plants. These regulations pursue to diminish the negative impact on ecosystems and human health. As a consequence of these huge and fast changes, water and wastewater treatment have had to make great effort to achieve the new requisites. Among these new requirements introduced by the Directive 98/83/CE, the obligation to control the presence of THMs in water supply systems is included. Because of the impact on human health and changes in existing supply facilities that this requires, without a doubt, its monitoring and control will be of great importance.

DBP monitoring is usually time consuming and involves expensive techniques such as gas chromatography analysis. Hence, particular interest has grown on the development of models to estimate the formation of DBP, which may be an alternative for the monitoring of DBPs in the field. Conversely, the models can also be very useful in verifying key operational and water quality parameters, which may help to explain the DBP formation potential. So another possibility is to predict THMs concentration indirectly by mathematical models based on other materials and substances present in the water, that are much easier to measure. Such models can be applied in areas where it is difficult to perform the analyses required [7,8]. Indeed, the analysis of disinfection by-products is expensive and developing a simple tool that will allow knowing the concentrations of THMs, particularly in small or poorly resourced water supplies, is a worthy intention.

Therefore, the main objective of this work consisted in the search of a simple system that will allow us to quickly estimate the amount of THMs present in water without requiring their analyses in the laboratory. For this purpose, we have developed a statistical model based on multiple regression from real water samples and confirmed its suitability for the studied water. This kind of model has been chosen due to the fact that, according to Kulkarni and Chellam [9], DBP mass concentrations are typically modeled empirically by linearly regressing each of the water quality parameters influencing DBP formation. In addition, log-linear power functions are extensively employed to model THM and Haloacetic acid (HAA) formation [10–14].

As mentioned before, measurement of DBP concentration in drinking water usually requires tedious techniques which are time consuming and significantly expensive. There have been many efforts in
development predictive models for DBPs. Initial models included univariate analysis where DBPs were correlated with Total Organic Carbon (TOC) content in raw waters [15]. Other works have looked for relationships between precursor and operational indicators and DBPs as well as between different species of DBPs [16–20]. There have been also developed multivariate models to correlate DBPs with different combinations of explanatory variables, such as water quality and parameters related to disinfection. Furthermore, predictive models for DBPs are based on data from field and laboratory-scaled studies. These data are collected at different sampling points. Most of these models have been based on empirical relationships among different parameters and DBPs concentrations, whilst a limited number of studies based on kinetic relationships have been performed [21–26]. These models have been usually developed using multivariate regression techniques while some studies are based on first- and second-order kinetic models and the coefficients have been estimated using multivariate regression analyses.

In relation to measuring instruments, electronic tongues and noses have emerged as a fast, low-cost and easy system for the automated detection and classification of odor, vapors, gases and liquid analysis. The strategy is based on the employment of specific sensors that generates a signal pattern related to either a quality aspect or a specific component when it is exposed to a sample. Although the output of only one sensor is not enough to provide a complete characterization of the sample, the combination of several different devices can give lot of valuable information. Due to the versatility of electronics tongues and noses, there is a lot of research interest in their development for a wide range of applications, such as food industry (milk, wine, beer, fish, etc.) [27–33], environmental analyses, water quality monitoring, etc. [27,34–37]. For instance, Campos et al. [27] show an electronic tongue that is able to analyze COD (Chemical Oxygen Demand), BOD (Biochemical Oxygen Demand), ammonia (NH₄–N), orthophosphate (PO₄–P), sulphate (SO₄–S), acetic acid (HCA) and alkalinity (Alk) from an array of metallic electrodes and using a partial least squares (PLS) analysis. Another example of electronic tongue for water analysis is proposed by Kundu et al. [38]. In that work a system for water sample authentication is presented based on a back-propagation learning based neuronal network (BPNN). Finally, not only chemical and physical properties can be determined, but also biological parameters, for example, Nayak et al. [35] present an electronic nose for *Escherichia coli* (*E. coli*) applying a pattern recognition.

In particular, we propose a device with different integrated sensors that measure several simple parameters on-site. The values obtained from these measures are automatically introduced into a mathematical model based on multiple regression, obtaining the value of THMs concentration estimated by the model. As on-site measurements are made in continuous, data acquisition procedure can be programmed with the model with the desired frequency.

The device is designed to be easily transported to different points of the water supply system, where it is intended to carry out a control campaign. In case of detection of high concentrations of THMs, it could be reasonable taking samples and to make an exhaustive analysis in the laboratory, but having the certainty that there is actually a problem.

It is also important to keep in mind that the problems with the THMs in the water supply systems are usually seasonal, having the higher risk of concentration in summer. This means that the device designed is useful to transport it from one place to another, as indicated above, but it is also interesting if it is installed for a prolonged period at a specific point in the network, in order to alert us in case of
an upward trend of the THMs concentration, and thus avoiding costly campaigns of laboratory analysis or measures to control them.

2. Materials and Methods

2.1. Characterization of 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 [6]. The analyses were performed on real samples collected from different water distribution networks. A total of 892 analyses were performed on populations in a regional community in Spain. This regional community is located on the Mediterranean coastline, where the drinking water has high levels of THMs. These data have been provided by the AGBAR Corporation (Aguas de Barcelona, Barcelona, Spain), a multinational holding company of water management.

From these 892 analyses and 87 different parameters, only those parameters included in all of the analyses were finally taken into account for the model. In addition, some potentially important parameters could not be taken into account because they are not included in the complete analysis described in the European Council Directive 98/83/CE.

Multiple regressions and analysis of variance were performed to model the dependency of total THMs values with these 87 parameters, but finally, (as described in Sections 2.2 and 3), seven parameters showed the most significant relation: Total Organic Carbon (TOC), Combined Residual Chlorine (CRC), Free Residual Chlorine, Bicarbonates, Conductivity, Chloride and Temperature (Table 1).

### Table 1. Summary of most significant parameters measured in the 892 network water samples.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Average</th>
<th>Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total Organic Carbon (mg/L)</td>
<td>2.10</td>
<td>0.66</td>
</tr>
<tr>
<td>Combined Residual Chlorine (mg/L)</td>
<td>0.80</td>
<td>0.31</td>
</tr>
<tr>
<td>Free Residual Chlorine (mg/L)</td>
<td>0.61</td>
<td>0.29</td>
</tr>
<tr>
<td>Bicarbonates (mg/L)</td>
<td>229.94</td>
<td>76.40</td>
</tr>
<tr>
<td>Conductivity (μS/cm)</td>
<td>1172.76</td>
<td>615.89</td>
</tr>
<tr>
<td>Chloride (mg/L)</td>
<td>221.56</td>
<td>153.67</td>
</tr>
<tr>
<td>Temperature (°C)</td>
<td>19.5</td>
<td>4.2</td>
</tr>
<tr>
<td>Total THM (μg/L)</td>
<td>105.80</td>
<td>63.61</td>
</tr>
</tbody>
</table>

Note: Analytical methods according to Annexes I and III of the Council Directive 98/83/CE (3 November 1998), the regulation where health criteria of the quality of water for human consumption are established.

2.2. Statistical Method

Multiple regression was performed to model the dependency of THMs values as well as One-way Analysis of Variance (ANOVA). The software SPSS (Version 15.0; SPSS Inc. Chicago, IL, USA) was used to manage and analyze data from the considered samples.
There are several examples where models based on multiple regressions were performed to predict Trihalomethanes formation [39–42]. It is true that the characteristics in the THMs formation makes it difficult to develop an equation that can be applied universally, but it is completely feasible to quickly obtain an equation that can be applied to water supply systems with similar environmental conditions. If the environmental conditions change, the same procedure can be applied, being this procedure fast, and the model easily adaptable.

Obviously, the relationships between the THMs levels and factors analyzed (total organic carbon, conductivity, combined residual chlorine, pH, temperature, bicarbonates) are not linear; therefore it was necessary to use statistical tools that allow us to correct the nonlinearity between variables.

One of these tools is known like “the bulging rule of Mosteller and Tukey” [43] that help to find existent linear relations between functions whose variables were used in the statistical analysis:

\[ F(y_i) = a_0 + a_1 \cdot G_1(x_{1i}) + \cdots + a_p \cdot G_p(x_{pi}) + \epsilon_i; \ i = 1 \cdots N \]  

where \( y_i \) is the expected value directly related with the THMs concentration, \( a_j \) are the coefficients as result of the regression analysis, \( x_{ij} \) are the observed values of the independent variables and \( \epsilon_i \) is the error of the model. These independent variables are used to decide in which of a set of pre-established groups should be classified each sample.

With the data \( (y_i, x_{1i}, x_{2i}, \ldots, x_{pi}) \), \( 1 < i < n \), we used an algorithm that allows us to find the transformations \( F \) \( y \) \( G \), such that the empirical correlation of the transformed data \( (F(y_j), G_1(x_{1i}), \ldots, G_p(x_{pi})) \) is approximately maximized. The procedure was made by using the above mentioned software and others algorithms like ACE (alternating conditional expectations), AVAS (additive and variance stabilization) and the Box-Cox transformations technique, obtaining similar results.

In any case, the procedure provides the coefficients \( a_j \) and \( \epsilon_i \) of Equation (1) with the data acquired from local samples, allowing the recalibration of the model when conditions change, in other words, shift the values of the equation coefficients but not the transformations \( F \) \( y \) \( G \), therefore, there is only a general model where the values of the coefficients depend on the local characteristics of a region or group of water supply systems.

2.3. Device Description

The algorithm proposed as a result of the multiple regression analysis could be easily integrated in autonomous systems for monitoring the presence of total trihalomethanes in water. To do that, we have designed an electronic platform that integrates all the required sensors as well as the microcontroller with the proper matching networks to directly and fast obtain the value of THMs in water.

According to the algorithm of determination of THM presented in the following section, it is necessary to include six different sensors to detect the following properties:

- pH
- TOC
- Cl⁻
- Bicarbonate
- Conductivity
- Temperature
In addition to these sensors, the platform must include all the required circuitry to match the output signal of the sensors to the microcontroller. There are many different transduction effects and each of them results in a different way to manipulate the signal. The manipulated signal has also to be between the operating ranges of the microcontroller. All this additional circuitry is referred as matching networks. Note that a block to provide power to the platform is also required.

Moreover, it is necessary to add mechanisms to acquire data. For this purpose a display to show results, a memory to store data and a user interface to be connected to a computer, are also included.

There are several of these sensors available in the market, so the development of the proposed platform should be immediate. Further research work can be done by design an integrated platform [44]. A first approach could be a TOC analyzer similar to the one described by Su et al. [45]. This analyzer is based on a Bulk Acoustic Wave (BAW) impedance sensor, where the Total Inorganic Carbon (TIC) value is directly measured and TOC value is obtained by oxidation of the sample. Another option could be the sensor one [46] based on a capacitance membrane.

In order to measure the pH, we propose a system similar to Fraser et al. [47] which provides directly a color barcode of the water pH. Another option is presented by Diamond et al. [48] with a pH sensor and the complete needed circuitry. Regarding the bicarbonate value, this sensor would be based on the patent of Benco et al. [49]. This sensor requires the measurement of the pH to determine bicarbonates, but this problem is solved due to the previous explained sensor.

Moreover, a chlorine sensor is also required. This device could be similar to the one described by Wang et al. [50], which is a resistance sensor, or similar to the one introduced by Chou et al. [51]. A conductivity sensor is also required, one example of this kind of transducer is shown by Hilhorst [52] where conductivity is related to changes in the electrical permittivity. Finally, a thermometer is also needed; there are several examples in the literature for this sensor [53]. Therefore, all these sensors are commercially available and also can be custom-manufactured.

As it has been mentioned above, each of these sensors requires a particular matching network to adapt its output signal to the input signal in the microcontroller. Furthermore, it is necessary to optimally program the microcontroller to calculate from the input signals, the corresponded values of each parameter and finally determine the THM value of the sample following Equation (1).

Finally, as it has been mentioned before, it is necessary to include a memory to store data, an interface to extract these data and/or a display to visualize the current measure to functionalize the electronic equipment.

According to the European Council Directive 98/83/CE, these data are requested 4–10 times/year depending on the volume of water distributed or produced each day within a supply zone; thus, this value would fix the minimum frequency to register these variables. Anyway, in order to have a more representative measure, it would be recommended to take at least three water samples but users could monitor as many times as they consider necessary taking into account the ease of measuring and calculations. Furthermore, there is no restriction on a specific point in the water network because this device is portable and operators can analyze water where they need it. Moreover, this feasibility is a key feature to detect anomalies in the network.
3. Results and Discussion

Multiple regressions and analysis of variance were performed to model the dependency of total THMs values with 87 parameters included in the Annexes I and III of the Council Directive 98/83/CE [6] but finally, seven parameters showed the most significant relation: Total Organic Carbon (TOC), Combined Residual Chlorine (CRC), Free Residual Chlorine, Bicarbonates, Conductivity, Chloride and Temperature.

For obtaining the statistical model, a transformation of the dependent variable (THMs) was proposed in order to improve the fit of the model. The first obtained model indicated that the fit was not appropriate enough ($r = 0.686$). However, the residual contains a certain functional dependence, which can be resolved by proposing a transformation of the dependent variable (THMs). This transformation consisted of a quadratic function.

Therefore, a model with transformed values provided a better fit ($r = 0.923$) with the next adjusted model:

$$\sqrt{\text{THM}} = -28.826 + 1.583 \cdot \text{TOC} + 2.713 \cdot \ln(\sigma) - 1.307 \cdot \ln(B_c) + 2.427 \cdot \text{pH} + 3.744 \cdot \text{CRC} + 0.102 \cdot T$$

$$\varepsilon(\sqrt{\text{THM}}) = 1.8$$

where TOC is the Total concentration of Organic Carbon, $B_c$ is the concentration of Bicarbonate, $\sigma$ is the conductivity, CRC corresponds to Combined Residual Chlorine, $T$ is the temperature at Celsius degree and $\varepsilon$ is the mean absolute error. Similar studies [4,54] have shown values of the correlation coefficient $r$ in the same order of magnitude than the one obtained in our model.

In this model, the total THMs values were transformed by means of the square root. Bivariate correlations between the different variables show that one of the most striking correlations is that between chlorides and conductivity ($r = 0.945$); therefore, both parameters should not be simultaneously included in the model to avoid multicollinearity problems, for this reason, chlorides were excluded since it was found that this variable is determined almost exactly by conductivity.

Furthermore, the variables, conductivity and concentrations of bicarbonates, were transformed by logarithms since the coefficients of the model were significantly low.

As a result, the different models studied were composed of the following variables, which showed the most significant relation to THMs:

1. TOC (Total Organic Carbon)
2. TOC and ln($\sigma$)
3. TOC, ln($\sigma$) and Combined Residual Chlorine (CRC)
4. TOC, ln($\sigma$), CRC, and pH
5. TOC, ln($\sigma$), CRC, pH and ln($B_c$)
6. TOC, ln($\sigma$), CRC, pH, ln($B_c$) 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, increasing the $r$ values from 0.742 to 0.923.
A statistical summary of coefficients for the variables in the regression model is shown in Table 2, where \(a_j\) are the coefficients of the linear equation and in brackets their associated standard error, \(\beta_j\) are the associated standardized coefficient that show which of the independent variables have a greater effect on the dependent variable and finally, values \(t\) and associated significance indicated that the former coefficients were significantly different from zero, that is to say, all of them are statistically significant for the model. This Table shows that, statistically speaking, the most important predictor was TOC, followed by ln(conductivity), and then CRC. 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.

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

<table>
<thead>
<tr>
<th>Dependent Variable</th>
<th>(\sqrt{\text{THM}})</th>
<th>(a_j)</th>
<th>(\beta_j)</th>
<th>(t)</th>
<th>Significance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total Organic Carbon (TOC)</td>
<td>1.583 (0.071)</td>
<td>0.533</td>
<td>22.229</td>
<td>0.000</td>
<td></td>
</tr>
<tr>
<td>ln((\sigma))</td>
<td>2.713 (0.131)</td>
<td>0.459</td>
<td>20.736</td>
<td>0.000</td>
<td></td>
</tr>
<tr>
<td>ln((B_c))</td>
<td>−1.307 (0.192)</td>
<td>−0.153</td>
<td>−6.804</td>
<td>0.000</td>
<td></td>
</tr>
<tr>
<td>Combined residual chlorine (CRC)</td>
<td>3.744 (0.253)</td>
<td>0.399</td>
<td>14.773</td>
<td>0.000</td>
<td></td>
</tr>
<tr>
<td>pH</td>
<td>2.427 (0.270)</td>
<td>0.206</td>
<td>8.992</td>
<td>0.000</td>
<td></td>
</tr>
<tr>
<td>Temperature</td>
<td>0.102 (0.021)</td>
<td>0.115</td>
<td>4.783</td>
<td>0.000</td>
<td></td>
</tr>
</tbody>
</table>

The relationship between measured values and predicted values by the model is shown in Figure 1. As can be seen, the prediction band of the fit was acceptable. The data were fairly homogeneous, but compatible with the prediction of the model.

**Figure 1.** Relationship between measured values and predicted values by the model.

Some authors [42,43] also included the factor of season in their models because the seasons affect the THM formation. In our case, a statistical study of a 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. During the dry season, a higher THMs concentration was expected. Nevertheless, regarding to the model, the differences were not statistically significant in the THM prediction and the factor of season was not included.
4. Conclusions

A new multiple regression model for total THMs formation was generated, and the feasibility of a device for its measurements was performed. The mathematical formulation can be integrated in an autonomous system for predicting total THM presence in networks of water supply systems.

The complexity of the trihalomethanes formation reaction makes it difficult to develop a universally applicable formula, but with this model it is completely feasible to obtain quickly an equation that can be applied to water supply systems with similar environment conditions. When environmental conditions change, a recalibration procedure is performed in order to recalculating coefficients, being this process fast and the model easily adaptable.

Regarding to the model, the most significant variables to predict the concentration of total THMs by order of importance were following:
- Total organic carbon (TOC)
- Combined residual chlorine (CRC)
- Conductivity
- pH
- Bicarbonates
- Temperature.

TOC was found to be the most significant parameter, followed by CRC and conductivity. The other parameters (pH, bicarbonates and temperature) were at a greater distance from the others. The seasonality was not included in the model although it is directly related with the temperature.

Chlorides were excluded since it was found that this variable is determined almost exactly by conductivity. This meant that if they had both been included in the model, this could have produced an overestimation of the importance of these variables.

Therefore, the designed electronic device has six different sensors that are commercially available and can also be custom-manufactured. Each of these sensors requires a particular matching network to adapt its output signal to the input signal in the microcontroller, then it is necessary to optimally program the microcontroller to calculate from the input signals, the corresponded values of each parameter and finally determine the total THM value of the sample. Finally, the system includes a memory to store data, an interface to extract these data and/or a display to visualize the current measure to functionalize the electronic equipment.

The device can be measuring in real time, sending an alert signal to the users in advance, allowing them to take appropriate measures before there are excessive concentrations of THM, which are not permitted by law. It is important to keep analyzing directly THM several times per year in order to verify or to recalibrate the model.

Preliminary experimental results show good agreement with the model.

Acknowledgments

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

Almudena Rivadeneyra has been primarily responsible for the compiled information and the development of the device configuration; Maria Jesús García-Ruiz and Alejandro González-Martínez have been responsible for the data treatment and the development of the model; Fernando Delgado-Ramos has been responsible for the technical supervision of the article, with knowledge in water management and planification; Francisco Osorio has been responsible for the technical supervision of the article, with knowledge in water treatment technologies and water networks; Ovidio Rabaza is the general coordinator of the work.

Conflicts of Interest

The authors declare no conflict of interest.

References


© 2014 by the authors; licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution license (http://creativecommons.org/licenses/by/4.0/).