

Application of Generalized Linear Models to Data Analysis in Drinking Water Treatment

Aplicación de modelos lineales generalizados al análisis de datos en el
tratamiento de agua potable

ZULMA CEPEDA*
EDILBERTO CEPEDA C.**

Abstract

In drinking water treatment we found appropriate linear models to explain *dose and concentration of sulphate, dose of lime, lime application probability* and *polymer application probability*, including as explanatory variables some physical and chemical properties of raw water.

Keywords: Water treatment, Generalized Linear Models.

Resumen

En el tratamiento para obtener agua potable encontramos modelos lineales para explicar *dosis y concentración de sulfato, dosis de cal, probabilidades de aplicación de cal y de polímero*, incluyendo como variables explicativas algunas propiedades físicas y químicas del agua cruda.

Palabras Claves: Tratamiento de agua, modelos lineales generalizados.

1. Introduction

This study was developed in a water treatment plant located in Girardot (Colombia), whose source is the Magdalena River. The treatment of drinking water comprises the aeration, coagulation, sedimentation, filtration and disinfection of raw water. To remove undesirable substances, mechanical and chemical procedures are used. Colloidal particles are separated from water by means of a

*Universidad Industrial de Santander, Escuela de Ingeniería Química, E-mail: zcepeda@universia.com

**Universidad Nacional de Colombia, Departamento de Estadística, E-mail: ecepedac@unal.edu.co

chemical coagulation process consisting in the charge destabilization of the suspended particles by adding a coagulate substance, usually aluminum sulfate. The chemical coagulation is performed for several reasons, mainly to remove organic and inorganic turbidity; color, true and apparent; harmful bacteria and other pathogens; plankton and hardness. As a common practice, aluminum sulfate is applied in a quantity experimentally determined by jars test.

Since raw water is taken from a river, frequent measurements of variables such as *turbidity level*, *color* and *alkalinity* are needed to carry out the jars test to determine the dose of coagulate. Time and operation costs make it desirable to use a statistical model to establish optimal doses of sulfate, lime and polymer to be applied to water, based on *turbidity*, *color*, *temperature*, *pH*, *alkalinity* and *hardness*.

The raw water variables, object of this study, present the following variation intervals: *turbidity* 200 – 1200 NTU, *color* 50 – 250 UC, *pH* 7.1 – 8.33, *alkalinity* 43 – 55 mg/l CaCO₃, *hardness* 43 – 50 and *temperature* 24 – 28°C. The results and conclusions found in this paper are restricted to those conditions, though the methodology can be used to optimize the treatment process in other water treatment plants.

Linear models were initially used to describe the behavior of aluminum sulfate optimal dose, its optimal concentration of application, and lime optimal dose, as a function of the water properties (McCullagh & Nelder 1996). Finally, logistic models were used to show how the application of lime and polymer are related to the raw water properties (Agresti, A. 1990).

2. Physical, Chemical and Dosage Variables

2.1. Physical Variables

Turbidity is the optical effect caused by the dispersion and interference of a bright ray through the water having colloidal particles in suspension. These particles are formed by organic or mineral matter in untreated waters, by aluminum hydroxide in treated waters, by iron oxide in corrosive waters or by microorganisms and other substances. The turbidity units are equivalent to SiO₃ p.p.m. and receive the name of turbidity nephelometric units (NTU). Drinkable water has 1 NTU and it is considered not drinkable if turbidity is higher than 5 NTU.

Color is due to solution substances. The universal method for color determination is a visual comparison of a sample contained in Nessler pipe of 50 ml, with permanent patrons. Those made by chlorine-potassium platinate and cobalt chloride which are contained in identical pipes. In treated waters the desirable value is 5 UC (color units). The permissible maximum value is 15 UC.

Temperature in treated water should be pleasant to the senses and give a sensation of fresh water. According to the french regulation, the water temperature should lie between 9 and 12 °C. A temperature higher than 15 °C favors microorganisms development in the channels and intensifies odors and flavors.

2.2. Chemical Variables

The chemical variables refer to mineral elements or organic compounds that exist in water and may affect its quality according to established norms. The chemical properties generally considered are: *pH*, *alkalinity*, *hardness*, *electrical conductivity*, *acidity*, *CO₂*, *residual chlorine*, *calcium contents*, *magnesium*, *iron*, *chlorides*, *sulfates*, *nitrogenous compounds* and *toxic substances contents* (Rodier 1981). Technical restrictions of the firm where the study was carried out limited the variables to *pH*, *alkalinity*, and *water hardness*.

pH (pH) is an indicator of the concentration of hydrogen ion in solution, and is measured in mol per liter. *pH* indicates the intensity of acidity or alkalinity of a substance and is defined by $\text{pH} = -\log [\text{H}^+]$. In treated waters *pH* values lie between 7 and 8.5. However values between 6.5 and 9 are considered acceptable.

Alkalinity (ALC) is the measure of the power of one solution to neutralize the hydrogen ions and is expressed in terms of one equivalent quantity of calcium carbonate. In natural waters the total alkalinity represents the contents of hydroxides (OH), carbonates (CO₃) and bicarbonates (HCO₃). In general terms, treated water should not have too low values to produce corrosion neither too high to produce incrustations.

Water hardness (DUR) is due to the presence of calcium and manganese salts. Although these salts do not affect the water sanitary qualities, precipitate insoluble compounds allowing to form incrustations in distribution and heating systems. The *water hardness* for domestic consumption should be between 30 and 50 mg/l of CaCO₃. *Water hardness* only is included as explanatory variable in the linear model explaining the *aluminum sulfate optimal concentration*.

2.3. Dosage Variables

Dosage variables are the optimal doses of aluminum sulfate, polymer and lime necessary to get drinking water (American Public Health Association 1985).

Aluminum sulfate optimal dose. To calculate the required dose of coagulate, laboratory tests should be made using the jars test. To measure the *aluminum sulfate optimal dose*, an equipment with a mechanic agitator with 4 to 6 palettes, which operates at velocities from 0 to 100 r.p.m., a floc illuminator placed in the agitator base and precipitation glasses of 2000 ml of capacity are used.

The *aluminum sulfate optimal dose* is obtained by adding coagulate to water in progressive doses in each precipitation glass. The samples are injected in the mechanical agitator jars setted on 100 r.p.m. If water alkalization is required, it should be made before coagulation test.

After 30 seconds to 2 minutes of quick mixture, starting from coagulate application, speed is diminished to the velocity selected as the optimal flocculation condition, and then, the samples are flocculated for five to twenty minutes. After this, agitation is suspended, the palettes are removed and water is sedimented during five to ten minutes.

Samples are taken from each of the jars and *turbidity* is measured. The *aluminum sulfate optimal dose* will be obtained from the jar where the smallest *turbidity* and *color* values were found. This dose, denoted by D_{SUL} , will be taken as the quantitative response for the linear model analysis.

Figure 1 shows the behavior of residual *turbidity* as a function of *aluminum sulfate dose*. The lowest admissible value of *turbidity* is got by the application of 40 mg/l of *sulfate*. This quantity was the minimum necessary so that the conditions of turbidity of water become acceptable for human consumption and it is considered, for initial water conditions, the optimal dose to apply as D_{SUL} value. The initial conditions of water used for the data determination, were the following: *turbidity* 200 NTU, *color* 100UC, *pH* 8.03, *alkalinity* 40 mg/l CaCO₃ and *temperature* 24°C.

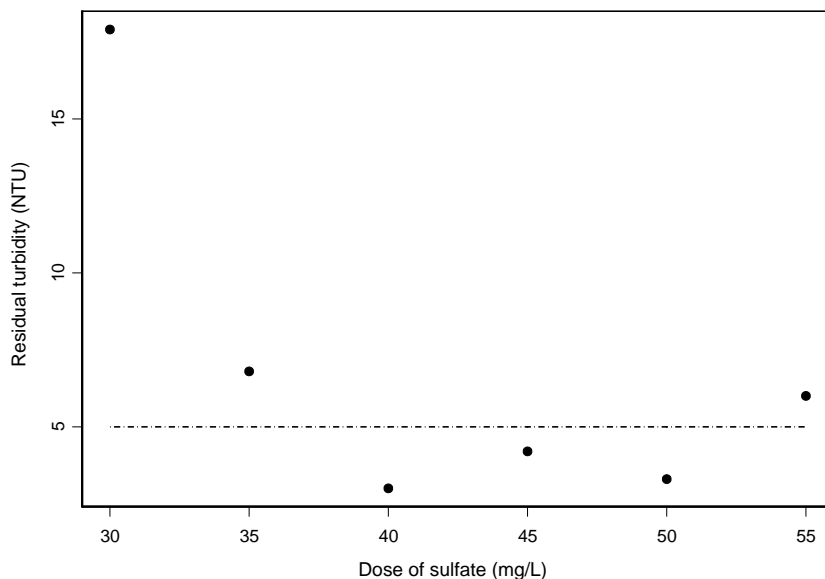


Figure 1: Dose of sulfate effect on residual turbidity.

Polymer optimal dose (D_POL). Once the test to determinate *aluminum sulfate optimal dose* is realized and a polyelectrolite has been chosen, the polyelectrolite optimal dose is obtained by the following test: The aluminum sulfate dose is reduced (for example in 75% of the optimal dose) and a variable quantity of polyelectrolite is added, having into account technical conditions and health entities recommendations, or maintaining constant the polyelectrolite dose and changing the aluminum sulfate quantity to obtain the lowest cost so that the dose of aluminum/polyelectrolite gives the water acceptable conditions for human consumption. If water quality in the plant changes through time, testing should have to proceed to new *turbidity* and/or *color* conditions.

The Willcomb index determination, floc apparition time, *residual turbidity*,

true *residual color* and *residual pH* will be necessary to determine the *polymer optimal dose*, denoted by D_{POL} .

Lime optimal dose (D_{CAL}). To determine the *lime optimal dose* we took a sample of 1000 ml of treated water with both *aluminum sulfate* and *polymer optimal doses*. If the *pH* is smaller than 7, a *lime* solution is applied to get this value. The *lime* quantity needed to get a *pH* equal to 7 is called the *lime optimal dose*, denoted by D_{CAL} .

The *Aluminum sulfate optimal concentration (C_{SUL})* is considered as a dosage variable, and its values are determined experimentally. A jars test is done in the conventional form and the sulfate optimal dose is determined according to the experimental procedure described before.

This test is made using a coagulate solution of 10 g/l (1%) with pH between 4 and 4.5. In five precipitate glasses with the same quantity of sample water, the optimal dose determined in the jars test, with different quantity of coagulate concentrations, such as 10%, 5%, 1%, 0.5% and 0.1% are added.

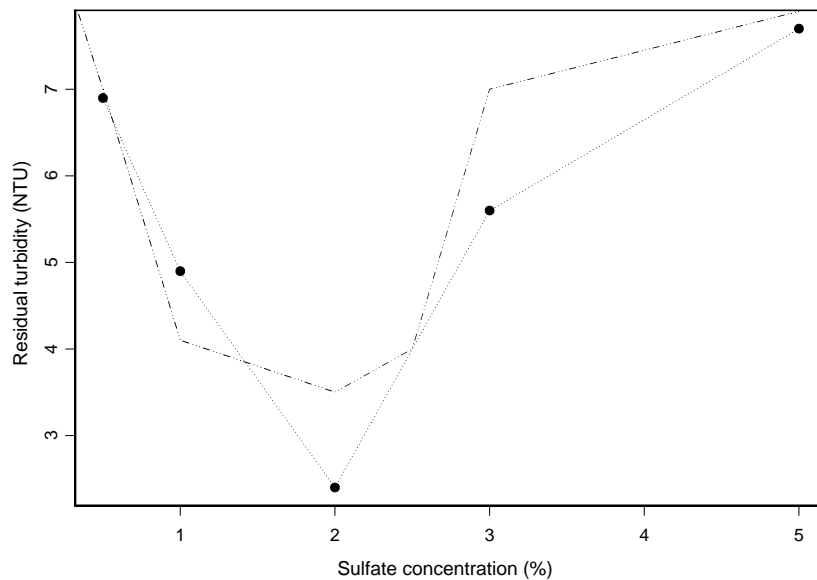


Figure 2: Sulfate concentration effect on turbidity.

After mixing, the water is flocculated and deposited. During this process both the floc size (Willcomb index) and the reaction velocity (floc formation initial time) are evaluated. Next, turbidity and residual color in the clarified water are determined. The concentration applied in the jar with the smallest turbidity and residual color is called the *optimal concentration*.

Figure 2 shows the C_{SUL} value for some characteristics of raw water. In this figure it is also shown how the concentration in the application of aluminum

sulfate optimal dose significantly influences the residual turbidity behavior and how, for some values of C_SUL concentration, in each case, the residual turbidity is smaller. This corresponds to the glass of the optimal concentration of aluminium sulfate application. The initial conditions were: *turbidity* 500 NTU, *alkalinity* 58 mg/l CaCO₃, *pH* 7.38, *temperature* 27°C.

3. The Water Treatment Data

In the water treatment plant of Girardot, the measurement of physical, chemical and dose variables are made after a sedimentation process, in which heavy materials are separated from water by gravitation.

Throughout 1997 we measured *turbidity*, *color*, *alkalinity*, *pH*, *temperature* and *sulfate optimal dose*. In order to determine the statistical model to explain the sulfate optimal dose, we selected measures of 40 samples of raw water, considering the variability of raw water data properties. In 20 of these samples we measured the optimal concentration for the sulfate application.

During the same year we carried out 33 experiments to determine the lime optimal dose for water treatment. In each one of these experiments we also measured turbidity, color, alkalinity, pH and temperature of raw water. To determine the probability of lime application and polymer application, 60 observations were taken from the water treatment plant registers.

4. Linear Model to Explain Sulphate Dose

We considered *turbidity* (TURB), *color* (COL), *alkalinity* (ALCA), *pH* (PH) and *temperature* (TEMP) as explanatory variables of the *dose of Sulphate* (D_SUL) variable behavior (dependent variable). The resulting linear model is

$$D_SUL = \beta_0 + \beta_1TURB + \beta_2COL + \beta_3ALCA + \beta_4PH + \beta_5TEMP + \epsilon$$

where $\epsilon \sim N(0, I\sigma^2)$. The determination coefficient is $R^2 = 0.991$. After backward elimination procedure with an elimination criterion of 0.01, the final model is:

$$\hat{D}_SUL = -13.571 + 0.085TURB + 1.592TEMP \quad R^2 = 0.991 \quad (1)$$

The p -values are all smaller than 0.01 and there is not lost of information when COL, ALCA and pH are dropped out from the model. Result (1) shows that the optimal dose of sulphate to be applied in water treatment depends basically on the turbidity and that low temperatures favor the coagulate action. If temperature has low values a less sulphate dose is required.

Table 1 shows the simple correlations between the variables included in the study. The 0.01 and 0.05 bilateral significance levels are indicated by ** and *, respectively. The table shows the highest correlation between sulphate dose and turbidity with a value of 0.978.

Table 1: Simple correlations

	TURB	COL	ALCA	PH	TEMP
D-SUL	0.978**	0.517**	-0.196	0.133	-0.266
TURB	1	0.485**	-0.119	0.156	-0.337*
COL		1	-0.448**	0.048	-0.046
ALCA			1	-0.207	0.181
PH				1	-0.408*

5. Optimal Concentration of Sulphate Dose

In this section we select the best linear model for the explanation of the sulphate concentration behavior. As before, we take an initial model of the form:

$$C_SUL = \beta_0 + \beta_1TURB + \beta_2COL + \beta_3ALCA + \beta_4PH + \beta_5TEMP + \beta_6DUR + \epsilon \quad (2)$$

with $R^2 = 0.530$ and p -values higher than 0.05 for partial hypothesis $\beta_k = 0$, $k = 0, 1, \dots, 6$. After the elimination process we found model (3) with $R^2 = 0.413$.

$$\hat{C}_SUL = -1.597 + 0.445PH \quad (3)$$

This model has intercept p -value=0.133 and slope p -value=0.003. We could consider a zero intercept model, but we prefer the last model, since it is better to predict concentration for a particular value of pH in the interval [7.1,8.33].

6. Linear Model to Explain Dose of Lime

Model (4) were used to explain dose of lime. We found $R^2 = 0.930$ and p -values higher than 0.05 for $H_0 : \beta_i = 0$, $i = 3, 5$.

$$D_CAL = \beta_0 + \beta_1TURB + \beta_2COL + \beta_3ALCA + \beta_4PH + \beta_5TEMP + \epsilon \quad (4)$$

With the BACKWARD elimination process we found model (5), with $R^2 = 0.915$.

$$\hat{D}_CAL = 59.374 + 0.01262TURB + 0.07746COL - 8.831PH \quad (5)$$

All the p -values are smaller than 0.01. Thus, the optimal dose of lime to be applied in water treatment depends basically on *turbidity*, *color* and *pH*.

7. Lime Application Probability

We will determinate the lime application probability in water treatment in the water plant of Girardot, from data set of lime, a dicotomic variable with a

value of one if it is applicated and 0 otherwise and *turbidity*, *color*, *alkalinity*, *pH* and *temperature* as quantitative variables corresponding to physical and chemical properties of raw water that may have influence in the *lime application probability*.

Initially model (6), where $p = Pr(lime = 1|turb, col, alca, ph, temp)$, were considered. For this model $D^*(M) = 2(l_s - l_M) = 15.79$ with 27 degrees of freedom and p -value equal to 96%, with l_s denoting the logarithm of the likelihood function for the saturated model and l_M , the logarithm of the likelihood function for the proposed model.

$$\text{logit}(1 - p) = \beta_0 + \beta_1TURB + \beta_2COL + \beta_3ALCA + \beta_4PH + \beta_5TEMP + \epsilon \quad (6)$$

The result, after the elimination procedure, is model (7), with parameter p -values less than 0.002 and $D^*(M) = 2(l_s - l_M) = 7.91$, with 11 degrees of freedom and p -value equal to 72%.

$$\hat{p} = \hat{P}(lime = 1|int, COL) = \frac{1}{1 + \exp(5.91 - 0.040COL)} \quad (7)$$

If in the model: $\text{logit}P(lime = 0|COL, PH) = \beta_0 + \beta_1COL + \beta_2pH$, we eliminate the intercept, which has a higher p -value than the explanatory variable pH , we obtain model (8). It has parameter p -values less than 0.002 and $D^*(M) = 2(l_s - l_M) = 17.92$, with 30 freedom degrees and p -value equal 96%.

$$\hat{p} = \hat{P}(lime = 1|COL, PH) = \frac{1}{1 + \exp(-0.042COL + 0.813PH)} \quad (8)$$

This model shows that the *probability of lime application* grow up from 0 to 1, when the *color* grow up from 50UC to 250UC.

8. Polymer Application Probability

We will determinate the *polymer application probability* in the water plant of Girardot, from data set of *polymer*, a dicotomic variable: $polymer = 1$ if lime is applicated and 0 otherwise, and *turbidity*, *color*, *alkalinity*, *pH* and *temperature* as quantitative variables corresponding to physical and chemical raw water properties that can have influence in the polymer application probability.

We initially consider the global model with all explanatory variables and apply the same process as before. We found model (9) with parameter p -values less than 0.002 and $D^*(M) = 2(l_s - l_M) = 15.39$, with 20 degrees of freedom and p -value equal to 75%.

$$\hat{p} = \hat{P}(polymer = 1|int, TURB) = \frac{1}{1 + \exp(6.55 - 0.0138TURB)} \quad (9)$$

If, in the model $\text{logit}P(polymer = 1|TURB, pH) = \delta + \beta_1TURB + \beta_2pH$, we eliminate the intercept, with higher p -value than pH , we obtain model (10) with

parameter p -values less than 0.002 and $D * (M) = 2(l_s - l_M) = 14.11$, with 30 degrees of freedom and p -value equal to 96%:

$$\hat{p} = \hat{P}(\text{polymer} = 1 | TURB, pH) = \frac{1}{1 + \exp(-0.0152TURB + 0.925pH)} \quad (10)$$

Figure 3 shows the behavior of *polymer application probability* for model (10) with *turbidity* and *pH* as explanatory variables.

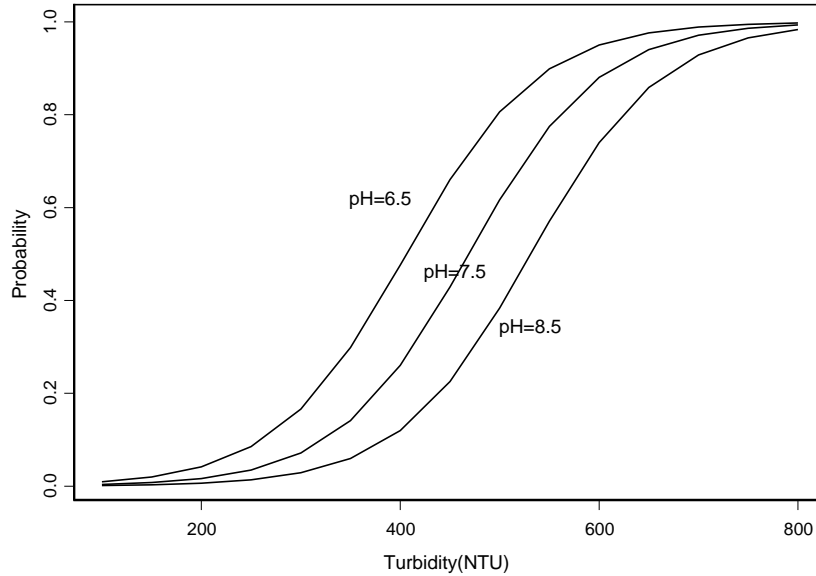


Figure 3: Behavior of polymer application probability, as model (10).

This model shows that the *probability of polymer application* grows up rapidly from 0 to 1, when the *turbidity* grow up from 300unt to 600unt.

9. Conclusions

The best linear models obtained to explain the optimal dose variables are:

$$\hat{D}_{SUL} = -13.571 + 0.085TURB + 1.592TEMP \quad (11)$$

$$\hat{D}_{CAL} = 59.374 + 0.01262TURB + 0.07746COL - 8.831pH \quad (12)$$

These models can be used to predict the optimal dose of *aluminium sulfate* and *lime*, optimizing the drinking water process. *Aluminium sulfate optimal dose* can be predicted if we know the *turbidity* and *temperature* of raw water. *Lime optimal dose* can be predicted if we know *turbidity*, *color* and *pH* of raw water.

In model 11, *temperature* appears to have influence on the *sulfate optimal dose*, since, for instance, larger values in the water temperature interval improve the flocculation process as time for mixing and precipitation gets shorter.

In model 12, *dose of lime* is explained as a linear combination of *turbidity*, *color* and *pH*. From the model, the *optimal dose of lime* decreases when *pH* increases. Higher values of *turbidity* in raw water requires larger sulfate dosages making *pH* to decrease because of the liberation of hydroxide ions in the coagulation process; then, water needs to be neutralized to the *pH* standards for drinking water.

The concentration of *sulfate dose* can be predicted, if we know the *pH* value, through the model $\widehat{C}_{SUL} = -1.597 + 0.445PH$. It can be used to optimize the water drinking treatment, getting precision in the experimental process to obtain the application optimal dose.

The logistic models 6 and 9 show (i) values of *color* where *lime application* is highly probable, (ii) values of *turbidity* where the polymer application is highly probable, and (iii) the *pH* relation in chemical process. This model also can be used to optimize the experimental process and to establish control quality process.

In the treatment plant of Girardot the models have been applied with favorable results and allow: (i) to predict the optimal dose of sulfate and lime and to get a better precision in experimental process. (ii) to establish supervision and control process. (iii) to optimize the treatment process at a lower cost.

This study should be developed in similar water treatment plants and extended to other process including the same or new variables.

Recibido: 16 de Mayo de 2005
Aceptado: 19 de Agosto de 2005

References

- Agresti, A. (1990), *Categorical Data Analysis*, John Wiley, New York.
- Aitkin, M. (1987), 'Modeling Variance Heterogeneity in Normal Regression Using Glim', *Applied Statistics* (36), 332–339.
- American Public Health Association (1985), *Standard Methods for the Examination of Water and Wastewater*.
- McCullagh, P. & Nelder, J. A. (1996), *Generalized Linear Models*, Chapman & Hall, London.
- Rodier, J. (1981), *Análisis de las aguas*, Ediciones Omega, Barcelona.
- Siegfried, W. (1992), *Modelos logísticos*, Universidad de los Andes-preprint, Bogotá.