MODELING OF CONCENTRATIONS OF BACTERIA IN WATER RECREATION

M.L. Patat*, Ana P. Comino**, Marcelo Scagliola**, Lila Ricci*

*Departamento de Matemática, FCEyN, UNMdP, Funes 3350, CP 7600, Mar del Plata, Argentina. (E-mail: mlpatat@mdp.edu.ar)

**Waters Laboratory, Quality Division, Mar del Plata Sanitary Works, OSSE, Brandsen 6650, CP 7600, Mar del Plata, Argentina (email: laboratorio@osmgp.gov.ar)

Abstract

There are several criteria for classifying the quality of recreational waters. The World Health Organization's Guidelines for Safe Recreational Water Environments have chosen to base criteria for recreational water compliance upon percentage compliance levels, typically 95% compliance levels (i.e., 95% of the sample measurements, taken in the bathing zone, must lie below a specific value in order to meet the standard),. Other criterion stipulates that the geometric mean of several individual samples must meet the corresponding geometric mean standard.

Analyses from a large data base with historical measurements obtained from the city of Mar del Plata, previous implementation of current disinfection by chlorination and outfall construction, with reference to the literature (Díaz, DA, 2005, "Límites de alerta y de acción para el monitoreo microbiológico ambiental", Pharmaceutical Technology, V 75, pp 134-141) showed that the data do not display a normal distribution. Consequently transformations are often applied so that normality is reestablished, and unbiased estimates of the percentiles can be calculated. One objective was to analyze the possibility of estimating percentiles from the true distribution of data in order to preserve the original scale. A set of distributions that provides broad possibilities is the Tweedie family (Tweedie, M., 1984. "An index which distinguishes between some important exponential families." In Ghosh, J.K.; Roy, J. Statistics: Applications and New Directions. Proceedings of the Indian Statistical Institute Golden Jubilee International Conference. Calcutta: Indian Statistical Institute. pp. 579–604) that is a subset of the exponential family and contains the normal, Poisson, gamma, and inverse Gaussian distributions as particular cases. This family allows modeling data with skewed distributions, choosing optimal values for the parameters from an infinite range of possible values. Based on an analysis of the real bacteria data, it was selected an optimal parameter that was very close to that of a gamma distribution. It was generated by Monte Carlo simulations with the distribution obtained. Finally, to illustrate the proposal it was applied both techniques to real data set.

Keywords: recreational water quality, family Tweedie, 95 percentile.

INTRODUCTION

There are several criteria for classifying the quality of recreational waters. One of them is the compliance system 95%, ie measuring the concentration of bacteria in the samples and take the 95 percentile, this value must be below a specific value to meet standard levels; other is directly in reference value taken as the geometric mean. As the concentration of bacteria is a count of the same in the samples, the distribution of these data does not behave as a normal distribution but follows a skewed distribution (Figure 1). Generally, transformations are often applied so that the

normality assumption is not altered, and thus calculate unbiased estimates of the percentiles. Another alternative is to calculate nonparametric percentiles as Hazen, Blom, Tukey or Weibull (see[10]).
Our interest is to estimate the percentiles from the original data distribution. A very useful family of distributions to model data is the family Tweedie model that allows asymmetric maintaining continuous data scale, this is described by Tweedie [8]. It includes as special cases the normal distribution (p = 0), Poisson (p = 1), gamma (p = 2) and inverse Gaussian (p = 3). Tweedie families are an important tool in statistics and can be used as error distributions in Generalized Linear Models whose description can be read in McCullagh and Nelder [6] and Jorgensen [4].
From the set of bacteria observed, we calculated the value of the parameter of this family of distributions. Simulations were made to compare the estimates of the percentiles calculated from Tweedie distribution with percentiles calculated by transforming the data and the nonparametric percentiles. We built a table comparing the mean square error of each percentile observed that estimated from Tweedie distribution is most appropriate.

FAMILY TWEEDIE

The richness of these models comes from that, given a number $p \in R - (0,1)$, there is always a random variable belonging to the family Tweedie, whose density is given by

$$p_p(y,\theta,\lambda) = c_p exp\left(\lambda\left(y\theta - \kappa\theta - \kappa_p(\theta)\right)\right)$$

with

$$\kappa_p(\theta) = \frac{1}{2-p}\left((1-p)\theta\right)^{\frac{p-2}{p-1}}$$

$\theta \in R^-$ is the parameter of position and $\emptyset > 0$ dispersion parameter. The function $c_p(y,\theta)$ is obtained by applying the Fourier inversion formula (Feller, [3], p. 581) and $p \geqslant 2$ is

$$c_p(y,\lambda) = \frac{1}{\pi\lambda y}\sum_{k=1}^{\infty}\frac{\Gamma(1+\alpha k)}{k!}\lambda^k\kappa_p^k\left(-\frac{1}{\lambda y}\right)\sin(-k\pi\alpha) \qquad (1)$$

For v.a. Y with distribution in the family Tweedie will use the notation $Y \sim Tw_p(\theta,\phi)$.

Mean and dispersion of Y are $E(Y) = \mu = \left((1-p)\theta\right)^{\frac{1}{1-p}}$ and $Var(Y) = \frac{\mu^p}{\emptyset} = \frac{1}{\emptyset}V(\mu)$.
Which $V(\mu) = \mu^p$ is called variance function uniquely characterizes the variable. A detailed discussion of these models can be found in [5].

There is a close relationship between families and the extreme stable distributions, theoretically important because it limits distributions as evidenced by Feller in [3].

Another fundamental property is the change of scale invariance. This means that, if Y belongs to a family then for any positive real number c, is cY also belongs to a family of this class. More specifically, it can be proved that if Y is an rv with distribution then cY has a distribution $Tw_p(c\theta, c^{2-p}\emptyset)$. (for a demonstration of this property see [5]).

In practical applications are often required of such models, particularly when working

with positive continuous data.
It is however clear that the expression (1) does not have a simple digital processing. This is perhaps the main reason that limits the use of these models in real data applications. A method to assess their density was developed by Dunn and Smith [2] and is implemented in the R language package Tweedie [7].

Outside the interval (0.1), for each real value of p has a family and given a set of data is possible by calculating a profile likelihood, find the value of the parameter p such that the corresponding Tweedie family is the most follows the same distribution as detailed in [2]. This method provides a choice of representations very "tailored" for the distribution of observations in each situation by choosing the optimal value of p is implemented in the R package [1].

METHODS FOR CALCULATING THE 95 PERCENTILE OF A SET OF DATA

For normally distributed data, the percentile 95 can be easily calculated from the mean (m) and standard deviation (s) of the data using the following formula: P = m + sz (P: Parametric percentile) where z is the quantile corresponding to the standard normal distribution and its value is 1.6449. This method is described in [9]. Bacterial count does not follow a normal distribution, however, a marked asymmetry, which is why it is often applied logarithmic transformation to be considered as a normal. In this way we estimate the percentile but above the transformed data, P '= m + sz where m y s now are average and deviation of the transformed data, respectively. Then the percentile of the data in the original scale is: $Pn = 10^{P'}$. Another way to get the percentiles is based on nonparametric statistics. There is no single way to calculate them [10]. There are four formulas that allow the calculation:

*Hazen: r = 0.5+P*n/100*

Blom: r = 3/8+P(n+1/4)/100*

Tukey: r = 1/3+P(n+1/3)/100*

Weibull: r = P(n+1)/100*

where r represents the place of the percentile data from lower to higher, P is the percentile value, in this case 95, n is the sample size.

Once the value of r, the percentile is calculated as follows:

$P\ (percentile) = (1 - rf) * X_{ri} + rf * X_{ri+1}$

where ri is the integer part of r rf the fractional part of r [11]. We propose to estimate the percentile from the distribution Tweedie. The expression of this distribution is not a simple numerical treatment. A numerical method for evaluating their densities was developed by Dunn and Smith [2] and is implemented in the R language package Tweedie [7].

Using the statistical package, from the bacteria count data, we estimate the parameters of the distribution and estimate the 95 percentile value. Figure 1 shows the power parameter of this distribution for real data of bacteria count in recreational waters.

SIMULATION STUDY

We perform a Monte-Carlo simulation to compare the mean square errors of the percentiles obtained by the proposed methods. We used the R statistical package, version 2.12.2, [7]. 1000 iterations were generated samples of size 100 with Tweedie distribution with the following parameters $n = 100$, $p = 2.5$, $\mu = 840$ y $\Phi = 0.01$. These values were chosen according to the behavior of the count of fecal bacteria (fecal) in the set of observed data.

Method	Mean	Variance	MSE
Parametric (Pn)	1765.50	17430.86	20866.84
Hazen	1707.44	26060.15	26060.47
Blom	1715.97	26852.38	26934.99
Tukey	1718.81	27151.73	27294.08
Weibull	1741.55	30181.39	31383.30
Tweedie	1703.16	17033.40	17047.23
Theoretical	1706.88	----------	----------

Table 1. These values were chosen according to the behavior of the count of fecal bacteria (fecal) in the set of observed data. From these results obtained show that the 95 percentile estimate using Tweedie distribution with p = 2.5, is the estimator with smaller mean square error.

CONCLUSIONS

In this paper, we suggest a procedure for estimating the 95 percentile as well as being simple, no need to transform the scale of the data. We believe that working with data in the original scale is more convenient than estimates obtained appear to be unbiased and have more desirable asymptotic properties. Subsequent tests of hypotheses that arise are therefore more powerful.

The family of Tweedie distribution model allows data sets with asymmetries such as the bacteria count in recreational waters, while respecting the scale of the data source. Our contribution is then to estimate the percentile assuming the data follow the distribution Tweedie. When comparing the mean square errors of different percentile estimates, we see that the best behaved is obtained from Tweedie family.

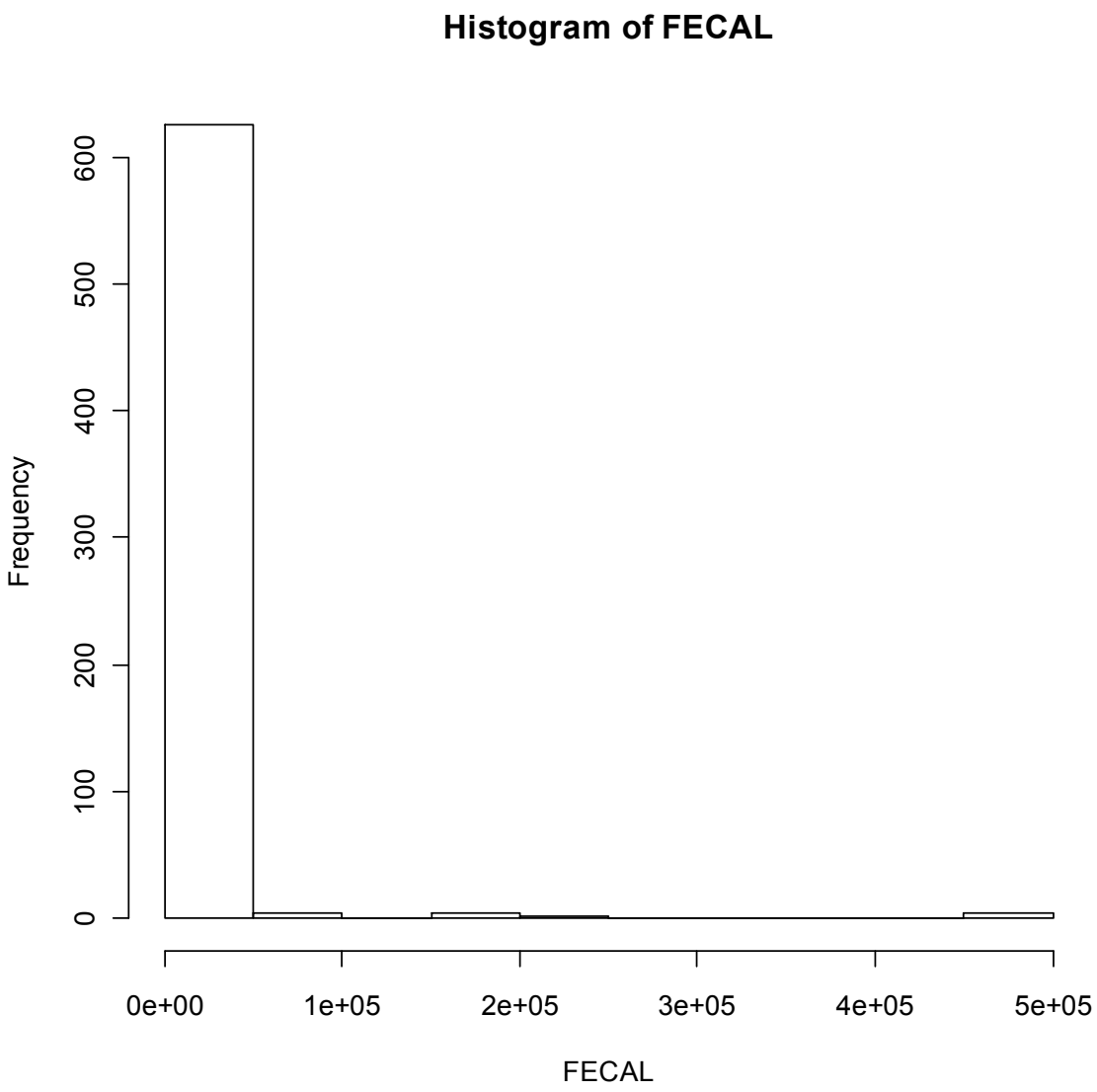

Figure 1. Histogram of Fecal bacteria counts.

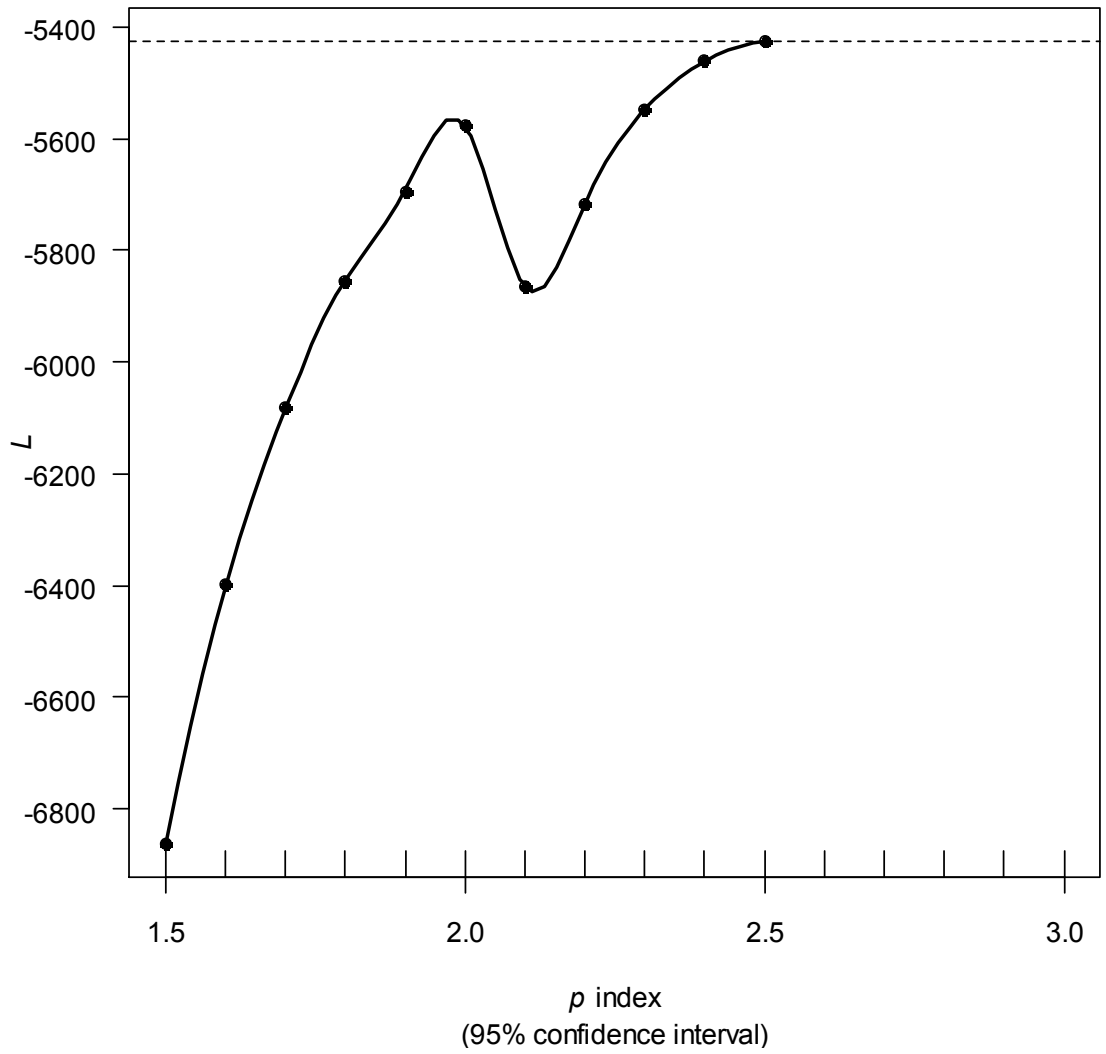

Figure 2. Maximum likelihood estimation of the Tweedie index parameter *power* of FECAL (p = 2.5.)

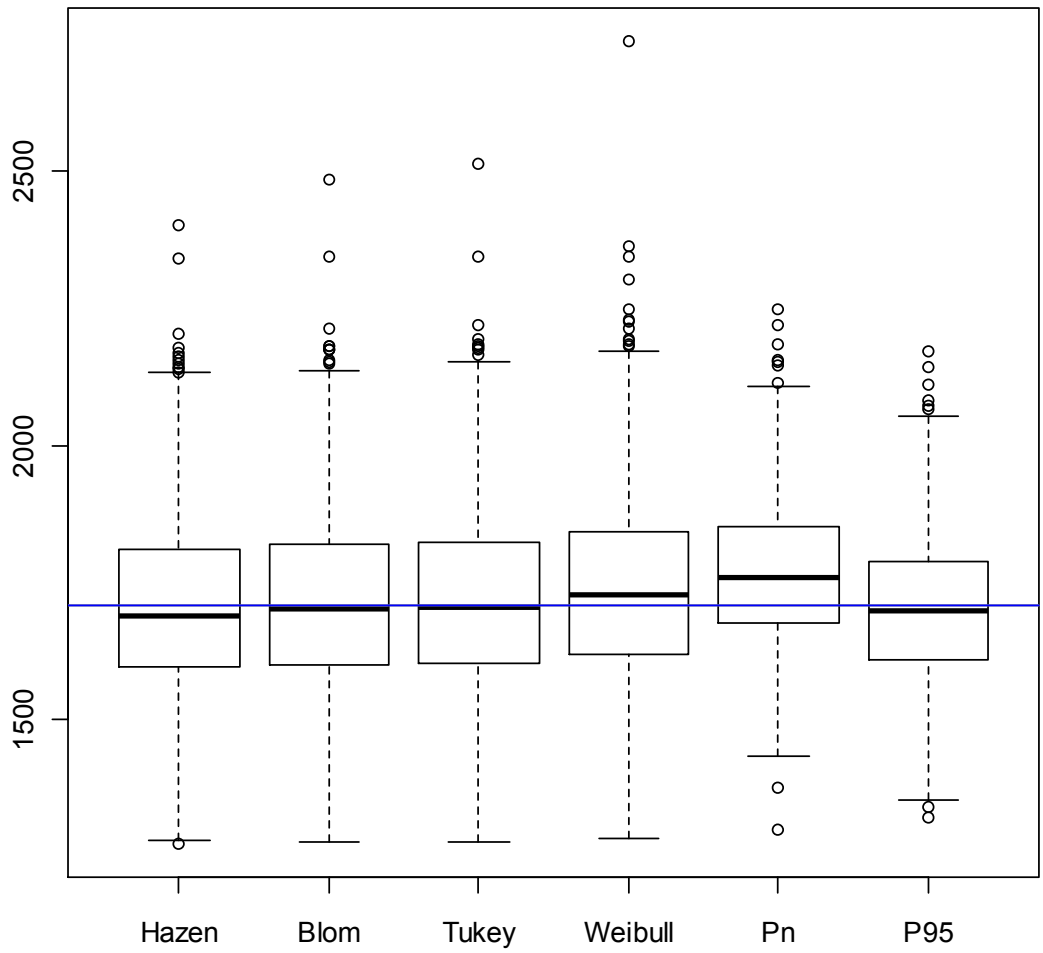

Figure 3. Comparison of methods of estimating percentiles. The highlighted line indicates the theoretical percentile value (Pt = 1706.88).
(Pn: transforming the data obtained percentile, P95: percentile obtained through family Tweedie)

REFERENCES

[1] Dunn, P. 2004. Tweedie exponential family models. R package version1.02. http://www.r-project.org/.

[2] Dunn, P.K. y Smyth, G.K. 2005. Series evaluation of Tweedie exponential dispersion model densities. Statistics and Computing 15(4): 267-280.
[3] Feller, W. 1978. Introducción a la Teoría de Probabilidades y sus Aplicaciones. Volumen II. Ed. Limusa.
[4] Jørgensen, B. 1992. The theory of exponential dispersion models and analysis of deviance. Monografías de Matemática no 51. IMPA, Rio de Janeiro, Brasil.
[5] Jørgensen, B. 1997. The Theory of Dispersion Models. Chapman & Hall.
[6] McCullagh, P. y Nelder, J. 1989. Generalized Linear Models. Chapman & Hall.
[7] R Development Core Team, 2006. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, http://www.r-project.org/.
[8] Tweedie, M. 1984. An index which distinguishes between some important exponential families. Statistics: Applications and new directions. Proceedings of the Indian Statistical Institute Golden Jubilee International Conference 579-604. Calcuta.
[9] Bartram, J. and Rees, 2000. Monitoring Bathing Waters. London: E & FN Spon.
[10] Ellis, J.C., 1989. Handbook on the Design and Interpretation of Monitoring Programmes. Report NS 29: Medmenham, England: WRc Environment, Water. Research Centre.
[11] G McBride and G Payne, NIWA, Hamilton, NZ, September 2002 & May 2009.