Water Qual Expo Health (2015) 7:227–231
DOI 10.1007/s12403-014-0143-5

ORIGINAL PAPER

Estimating Percentiles of Bacteriological Counts of Recreational Water Quality Using Tweedie Models

Maria Laura Patat · Lila Ricci · Ana Paula Comino · Marcelo Scagliola

Received: 3 December 2013 / Revised: 6 May 2014 / Accepted: 14 August 2014 / Published online: 18 September 2014

Abstract There are general guidelines and standards for measuring the microbial quality of water to prevent the incidence of disease outbreaks. Many agencies have chosen the 95th percentile; one can assess the recreational water quality, depending if the percentile value exceeds the guideline value or not. It is well known that this kind of data do not display a normal distribution and several alternatives have been proposed and are in use for estimating the percentile. A review of existing methods is given, that includes non parametric estimators as Hazen, Blom, Tukey and Weibull. We also describe transformations such as logarithmic and Box–Cox, that generate near normal data, after obtaining the normal percentile the inverse transformation is applied to obtain estimators in the original scale. A new methodology is proposed, consisting in finding the Tweedie distribution that better fits the observed data; this family has nonnegative support and can have a discrete mass at zero, making it useful to model skewed data that are a mixture of zeros and positive values. It allows working with parametric models in the original scale. We performed a Monte Carlo simulation to compare the performance of all the percentiles described above. As a result we noted that the percentile calculated from Tweedie distribution has lower mean square error than the others, which makes it the more precise estimator. All these techniques were applied to four data sets and, in all cases the Tweedie estimator was closer to the observed values than non parametric and anti transformed estimators.

M. L. Patat (✉)· L. Ricci
Departamento de Matemática, Universidad Nacional de Mar del Plata, Funes 3350, 7600 Mar del Plata, Argentina
e-mail: mlpatat@gmail.com

A. P. Comino · M. Scagliola
Department of Water, Quality Division, OSSE, Brandsen 6650, 7600 Mar del Plata, Argentina

Keywords Recreational water quality · 95th percentile · Tweedie family · Health protection to water quality

Introduction

The recreational water quality is related to the presence of microorganisms in the water such as fecal coliforms, streptococci, total coliforms and enterococci. There are general guidelines and standards for measuring the microbial quality of water to prevent the incidence of disease outbreaks. These values are derived from studies which link the exposure, the water quality and the diseases related with the presence of microorganisms.

Many agencies have chosen the 95th percentile to measure the quality of recreational waters. One can assess the recreational water quality comparing the observed percentile values with guideline values.

The theoretical 95th percentile is a value such that the probability that the variable is less than it is equal to 0.95, and the observed percentile is the value that leaves 95 % of the observations below it.

As the distribution of the bacteria count has a marked asymmetry, in practice, the percentiles are calculated using log-normal method, that is, logarithmic transformation is applied to the data so that they acquire approximate normal distribution. Percentile obtained in this way is called parametric percentile. A limitation of this method is that we lose the original scale of the data and the inverse transformation has to be applied.

A broader approach consists in applying the Box and Cox (1964) power transformation, that contains the logarithmic one as a particular case. After a transformation, one is often interested in inference on the original scale. Taylor (1985) defines a measure of location on the original scale applying

the inverse Box–Cox to the center of the transformed data to symmetry.

A frequently applied alternative strategy, is to calculate non-parametric percentiles. Some of them are due to Hazen, Blom, Tukey and Weibull (Hunter 2002). The disadvantage of non parametric methods is that they generally ignore relevant information on data, obtaining less accurate estimates.

We propose in this article to apply a Tweedie model (Tweedie 1984), that gives a parametrical percentile estimate, and needs no transformations. This stochastic family allows modeling positive data with skewed distributions, by choosing optimal values for the parameter p from an infinite range of possible values. Gamma, normal, Poisson and inverse Gaussian distributions are particular cases. In this way, we respected the original scale of the data and estimate the 95th percentile from the Tweedie distribution, which better fit the actual data.

We give a review of existing methods for calculating the percentile estimate in "Methods for Calculating the 95th Percentile of a Data-Set: Literature Review" section. In "Proposed Methodology: Estimating the Percentile from a Tweedie Model" section we introduce Tweedie models and define the percentile estimator based on them. "Simulation Study" section shows a simulation study that compares the estimators described above and in "Application to Real Data" section we obtain all these estimators, for real data sets from the beaches of Mar del Plata. Finally a discussion is presented in "Discussion and conclusions" section.

Methods for Calculating the 95th Percentile of a Data-Set: Literature Review

Non-parametric Percentile

Several methods of estimating percentiles employ non-parametric statistics (Ellis 1989), we will describe Hazen, Blom, Tukey and Weibull methods. In all of them, percentile estimators can be calculated by a two-step non-parametric procedure: the first step consists in obtaining a number r, defined in each case by:

$$\text{Hazen: } r_H = 0.5 + 0.95n \tag{1}$$

$$\text{Blom: } r_B = 3/8 + 0.95(n + 0.25) \tag{2}$$

$$\text{Tukey: } r_T = 1/3 + 0.95(n + 1/3) \tag{3}$$

$$\text{Weibull: } r_W = 0.95(n + 1) \tag{4}$$

where n is the sample size.

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

$$P_* = (1 - rf_*) * X_{ri_*} + rf_* * X_{ri_*+1} \tag{5}$$

where X is the original variable, * is H, B, T or W, respectively, the subscript ri indicates the integer portion of r and rf indicates the fractional part of r.

Estimated Percentile from Anti-logarithmic Transformation

For normally distributed data, the 95th percentile can be easily calculated from the mean (m) and standard deviation (s) of the data using the formula $P = m + sz$ (P: parametric percentile) where $z = 1.6449$ is the quantile corresponding to the standard normal distribution.

But bacterial count does not follow a normal distribution and logarithmic transformation is often used to approach normality. Thus, we estimate the percentile from the transformed data with $P' = m' + s'z$, where m' and s' are mean and standard deviation of the logarithm of the data respectively, and where z is the same as above. Then, the estimated percentile back in the original scale is obtained via the inverse transformation: $P_{log} = 10^{P'}$. This approach is outlined in Bartram and Rees (2000).

An important limitation is that no always a logarithm transformation gives normal data.

Estimated Percentile from Inverse Box–Cox Transformation

A more general approach is given by Box–Cox transformations (see Box and Cox 1964) defined as:

$$Y = \begin{cases} \frac{(X^\lambda - 1)}{\lambda} & \text{when } \lambda \neq 0 \\ ln\,(X) & \text{when } \lambda = 0 \end{cases} \tag{6}$$

being X a positive random variable. It can be proved that there exists an optimal value λ such that the transformed variable Y has the more accurate approximation to a normal distribution with mean μ and variance σ^2. Note that the logarithmic transformation is a particular case, for $\lambda = 0$.

The distribution of the anti-transformed data belongs to the power Normal (PN) family and a detailed description of these variables can be found in Freeman and Modarres (2006). These authors also consider the quantile functions that can be applied in statistical modeling when interest focuses particularly on the extreme observations in the tails of the data (Modarres et al. 2002), as is in our case.

The quantile function of $PN\left(\lambda, \mu, \sigma^2\right)$ is given by

$$P_{BC}^{\lambda}(p) = \begin{cases} \left(\lambda\left(\sigma\Phi^{-1}(V(p)) + \mu\right) + 1\right)^{1/\lambda} & \lambda > 0 \\ exp\left(\mu + \sigma\Phi^{-1}(p)\right) & \lambda = 0 \\ \left(\lambda\left(\sigma\Phi^{-1}(p) + \mu\right) + 1\right)^{1/\lambda} & \lambda < 0 \end{cases} \tag{7}$$

where Φ is the standard normal cumulative distribution and $V(p) = 1 - (1 - p)\,\Phi(T)$, for $0 < p < 1$, being

$T = \frac{1}{\lambda\sigma} + \frac{\mu}{\sigma}$ the truncation point. It is well known that the estimator $\hat{P}^{\lambda}_{BC}(p)$ has asymptotic normality, where $\hat{\mu}$ and $\hat{\sigma}2$, are the maximum likelihood estimators (*MLE*s) of the mean and variance on the normal scale. When p is replaced by 0.95, an estimator of the 95th percentile is obtained, and in the particular case of $\lambda = 0$ we obtain the same estimator as in "Estimated Percentile From Anti-logarithmic Transformation" section.

Proposed Methodology: Estimating the Percentile from a Tweedie Model

Tweedie models form a subclass of the exponential dispersion models. They are defined as exponential dispersion models with unit variance functions of a certain simple form. More precisely, an exponential dispersion model with unit variance functions V is called Tweedie model of order $p \in R - (0, 1)$ if $V(\mu) = \mu^p, \mu \in \Omega$ being Ω the parametric space.

Tweedie models include most of the usual distributions such as normal ($p = 0$), Poisson ($p = 1$), gamma ($p = 2$) and inverse Gaussian ($p = 3$). Their density is given by

$$p_p(y, \theta, \lambda) = c_p(y, \lambda)\, exp\left(\lambda\left(y\theta - \kappa_p(\theta)\right)\right) \tag{8}$$

where $y \in R^+, \theta \in R$ is the position parameter, $\lambda > 0$ the dispersion parameter and the function $\kappa_p(\theta)$ is given by

$$\kappa_p(\theta) = \begin{cases} e^{\theta} & \text{for } p = 1 \\ -log(-\theta) & \text{for } p = 2 \\ \frac{1}{2-p}\left((1-p)\theta\right)^{\frac{p-2}{p-1}} & \text{for } p \notin \{1; 2\} \end{cases} \tag{9}$$

The function $c_p(y, \lambda)$ is obtained using the Fourier inversion formula (Feller 1978, p. 581). If $p > 2$, it is of the form

$$c_p(y, \lambda) = \frac{1}{\pi\lambda y}\sum_{k=1}^{\infty}\frac{\Gamma(1+\alpha k)}{k!}\lambda^k\kappa_p^k\left(-\frac{1}{\lambda y}\right) sin(-k\pi\alpha) \tag{10}$$

For a random variable Y with Tweedie distribution the notation $Y \sim Tw_p(\theta, \phi)$ will be used with

$$\phi = 1/\lambda$$

The mean and variance are given by

$$E(Y) = \mu = \begin{cases} \left((1-p)\theta\right)^{\frac{1}{1-p}} & p \neq 1 \\ e^{\theta} & p = 1 \end{cases}$$

and

$$Var(Y) = \frac{\mu^p}{\phi} = \frac{1}{\phi}V(\mu).$$

A detailed discussion of these models can be found in (Jørgensen 1997). A fundamental property is their scale invariance: if Y belongs to a given family then for any positive real number c, cY also belongs to a family from this class. They are also limiting distributions, in the sense that they have domains of attraction. In practical applications such models are often required for skewed positive continuous data.

However, it is clear that expression (8) is not simple, which may be the main factor limiting the use of these models with real data. A method of obtaining the density was developed by Dunn and Smyth (2005) and it is implemented in the R package (R Development Core Team 2006).

Outside the interval (0,1), each real value of p generates a family. Given a set of observed data, the optimal value for p can be determined via profile likelihood estimation (Dunn 2004). This numerical method provides a selection of representations that are closely "tailored" to data sets with skewed distributions based on the chosen optimal value of p parameter.

Given a data set, we propose the following strategy:

1. Obtain the optimal value of the p parameter via profile likelihood estimation, so $\sim Tw_p(\theta, \phi)$.
2. Calculate the theoretical 95th percentile, P_{T_w} such that $P\left(Y \leq P_{T_w}\right) = 0.95$; with namely

$$0.95 = \int_0^{P_{T_w}} c_p(y, \lambda)\, exp\left(\lambda\left(y\theta - \kappa_p(\theta)\right)\right) dy \tag{11}$$

In this way, we preserve the original scale and estimate the 95th percentile from that Tweedie distribution which better fit the actual data.

Simulation Study

We performed a Monte Carlo simulation to compare the performance of the different percentile estimators. The routines were written in R language and the package "TWEEDIE" was used to generate data (R Development Core Team 2006). We ran 1,000 iterations generating a sample of 100 observations each time, following a Tweedie distribution with parameters $p = 2.5, \mu = 1, \phi = 0.58$. The theoretical 95th percentile [see (11)] was calculated.

Table 1 The 95th percentile estimators obtained using the proposed methods from a simulation study with parameters $p = 2.5, \mu = 1, \phi = 0.58$

	P_{H}	P_{B}	P_{T}	P_{W}	$P_{\log}$	P^{λ}_{BC}	P_{TW}
Mean	2.476	2.491	2.496	2.537	3.104	2.424	2.465
MSE	0.0813	0.0836	0.0846	0.0962	0.5229	0.067	0.0538

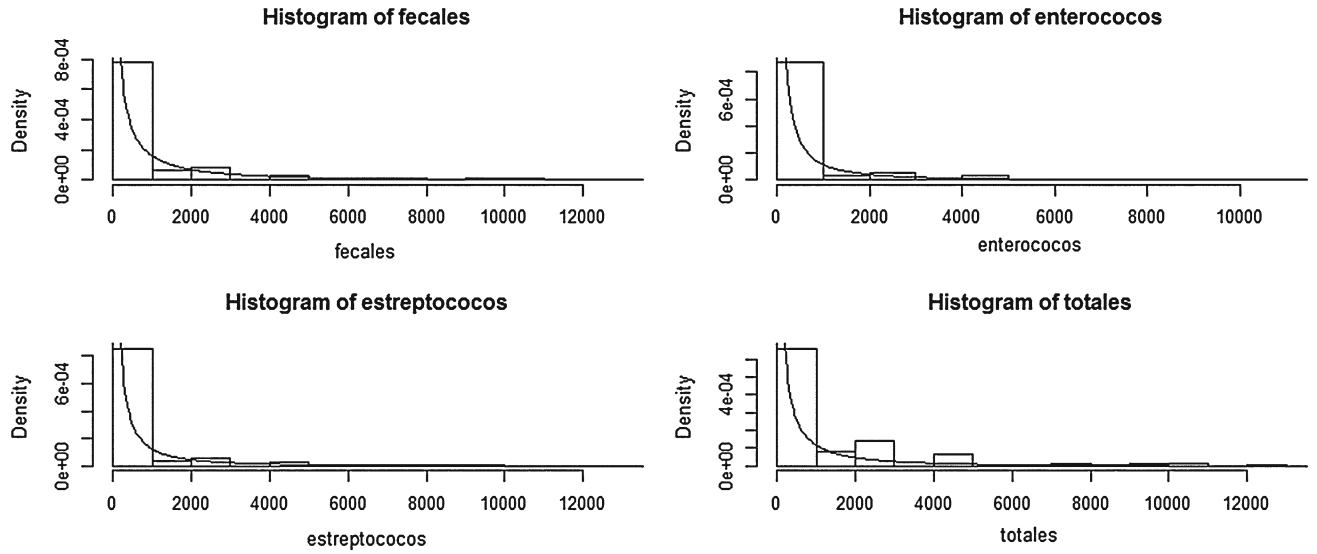

Fig. 1 *Boxplots* for each estimator for comparative purpose. The *horizontal line* indicates the theoretical percentile value ($P = 2.5$)

Table 2 Descriptive statistics of four groups of bacteria: fecal coliforms, streptococci, total coliforms and enterococci

Bacteria	Min	Median	Mean	Max	SD	Skew
Total coliform	2	460	1,467	13,000	2,315.38	2.50
Fecal coliform	2	220	1,014	13,000	2,077.52	3.30
Streptococci	2	80	652.4	13,000	1,600.24	3.96
Enterococci	2	43	552.5	11,000	1,374.94	3.96

The mean squared errors (*MSE*) were obtained to compare the performance of the corresponding estimators. Table 1 shows the results and Fig. 1 illustrates with a box plot for each percentile estimator, allowing the comparison of their properties. As can be seen, the *MSE* corresponding to the percentile obtained from Tweedie model is the smallest one.

Application to Real Data

The data were obtained from a study consisting of the monitoring and sampling of microbial water from the beaches of Mar del Plata, between 1999 and 2007, always in winter. There were four groups of bacteria: fecal coliforms, streptococci, total coliforms and enterococci; in Table 2 descriptive statistics are shown.

In a first step, we calculated for each bacteria the optimum value for p, to find the most suitable Tweedie distribution to fit the data. For *total coliforms* $p = 2.21$, for *fecal coliforms* $p = 2.071$ and for *streptococci* and *enterococci* $p = 2.5$. In Fig. 2 we show the histograms with the theoretical densities for the corresponding p superposed, it can be seen that the fit is more than acceptable.

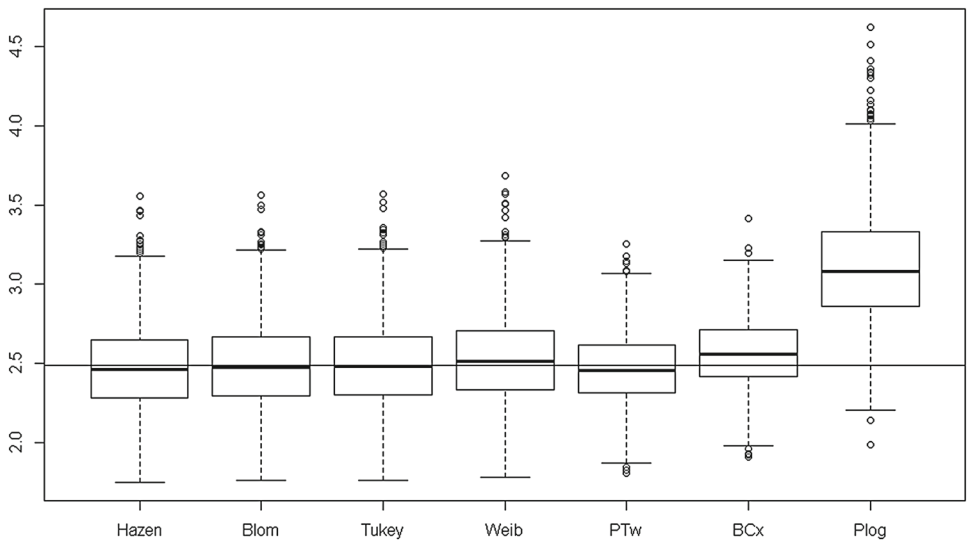

Fig. 2 *Histogram* for bacteria concentrations superposed with *density curve* of a Tweedie distribution with the corresponding value for p

Table 3 The 95th percentile estimator obtained using the proposed methods, from four data sets of bacteriological counts from beaches of Mar del Plata

Bacteria	P_{obs}	P_H	P_B	P_T	P_W	P_{log}	P_{BC}^{λ}	P_{Tw}
Total Coliform	7,000	7,100	7,400	6,900	7,300	11,130	7,117	6,901
Fecal Coliform	5,000	5,400	5,500	5,450	5,500	6,634	5,706	5,113
Streptococci	3,500	3,725	3,781	3,800	3,940	3,199	4,201	3,531
Enterococci	3,000	2,900	3,100	3,220	3,950	2,436	4,100	2,973

Later, we calculated percentiles using all the above methods and compared them to the actual percentile value of the data (Table 3). The percentile obtained from Tweedie distribution is the one that most closely fits the observed percentile for all groups.

Discussion and Conclusions

It has been found that bacteria count is not normally nor log normally distributed.

Among others, Chawla and Hunter (2005), found that their datasets "were not log normally distributed on at least 85 % of occasions and these finding fatally undermine the validity of using a parametric method for calculating 95th percentiles to classify bathing water quality".

Other percentile estimators frequently used have been proposed in the literature (Hunter 2002), they are 'non-parametric and use a limited amount of information because they only consider the order of each observation, not the exact value. Crabtree et al. (1987) affirm that "the arbitrary use of non-parametric techniques may fail to make the most effective use of the information contained in the data". On the other hand, Beamonte et al. (2007) state that parametric methods gave better results than non-parametric ones.

Another alternative is to antitransform percentiles obtained from data that has been transformed to approach normality. (see Taylor 1985)

In this paper, we suggest estimating the percentile of bacteriological counts in water, from a probability density function that takes into account the asymmetric distribution of this kind of data. We used the Tweedie family proposed by Tweedie (1984) and characterized as an exponential dispersion model by Jørgensen (1992, 1997). This model is appropriate for fitting asymmetrical data sets and eliminates the need to alter the original scale of the data by applying transformations.

In comparing the *MSE* of different percentile estimates, we found that the lowest mean square error was obtained using the Tweedie family. So we can conclude that this is a better estimator, in the sense that it is more precise.

It has also a more direct calculation. The numerical method implemented in the R package, allows choosing optimal values for the p parameter, as the one that maximizes the profile likelihood curve. Then, the 95th percentile estimator can easily been obtained from the optimal distribution function.

References

Bartram J, Rees G (2000) Monitoring bathing waters. E and FN Spon, London

Beamonte E, Bermúdez JD, Casino A, Veres E (2007) A statistical study of the quality of surface water intended for human consumption near Valencia (Spain). J Environ Manage 83(3):307–314 (ISSN 0301–4797). http://dx.doi.org/10.1016/j.jenvman.2006.03.010

Box GE, Cox DR (1964) An analysis of transformed data. J R Stat Soc B 39:211–252

Chawla R, Hunter PR (2005) Classification of bathing water quality based on the parametric calculation of percentiles is unsound. Water Res 39(18): 4552–4558 (ISSN 0043–1354). http://dx.doi.org/10.1016/j.watres.2005.08.022

Crabtree RW, Cluckie ID, Forster CF (1987) Percentile estimation for water quality data. Water Res 21(5):583–590 (ISSN 0043–1354). http://dx.doi.org/10.1016/0043-1354(87)90067-4

Dunn PK, Smyth GK (2005) Series evaluation of Tweedie exponential dispersion model densities. Stat Comput 15(4):267–280

Dunn P (2004) Tweedie exponential family models. R package version1.02. http://www.r-project.org/

Ellis JC (1989) Handbook on the design and interpretation of monitoring programmes. Report NS 29: Medmenham, England: WRc Environment, Water. Research Centre

Feller W (1978) Introducción a la Teoría de Probabilidades y sus Aplicaciones, vol II. Limusa

Freeman J, Modarres R (2006) Inverse Box–Cox: the power-normal distribution. Stat Probab Lett 76:764–772

Hunter PR (2002) Does calculation of the 95th percentile of microbiological results offer any advantage over percentage exceedence in determining compliance with bathing water quality standards? Lett Appl Microbiol 34(4):283–286

Jørgensen B (1992) The theory of exponential dispersion models and analysis of deviance. Mathematical Monographs no 51. IMPA, Rio de Janeiro, Brasil

Jørgensen B (1997) The theory of dispersion models. Chapman and Hall, Boca Raton

Modarres R, Nayak TK, Gastwirth JL (2002) Estimation of upper quantiles under model and parameter uncertainty. Comput Stat Data Anal 39:529–554

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/

Taylor JM (1985) Measures of location of skew distributins obtained through Box–Cox transformations. Am Stat Assoc 80(390):427–432

Tweedie MCK (1984) An index which distinguishes between some important exponential families. In Ghosh JK, Roy J (eds) Statistics: applications and new directions. Proceedings of the Indian Statistical Institute Golden Jubilee International Conference, pp 579–604. Indian Statistical Institute, Calcutta