Lognormal behaviour of untreated and treated wastewater constituents
S. C. Oliveira, I. Souki and M. von Sperling

ABSTRACT

This article presents an extensive study to select the theoretical probability distribution that represents the concentrations of major constituents from 35 wastewater treatment plants located in Brazil, as well as to verify the influence of the adopted treatment technology on the data distributions. Due to the lack of a solid basis for selecting the distributions that best fit the data, various distributions had to be tried and the choice was based on statistical tests and visual techniques, comprising a total of 3,444 tests undertaken. While comparing suitability of five distributions (normal, lognormal, gamma, exponential and rectangular) for analysis of data (influent and effluent biochemical oxygen demand, chemical oxygen demand, suspended solids, nitrogen, phosphorus, thermotolerant coliforms and flow), it was observed that the lognormal distribution was the most suitable, which is in agreement with the findings from other research studies. This conclusion requires a different position from the one currently adopted when analysing plant performance, in which symmetry of the data is generally implied, which has been shown not to be the case with the large array of data sets investigated.

Key words | lognormal behaviour, probability distribution, wastewater constituents

INTRODUCTION

Probability distribution of wastewater constituents

Numerous problems must be faced when applying statistical methods to environmental pollution studies. The data analyst must be aware that many statistical procedures were originally developed for data sets assumed to have been drawn from a population having normal distribution. However, environmental data sets are frequently asymmetrical and skewed to the right, so the validity of classical procedures may be questioned. Other problems associated with environmental data sets are large measurement errors, data near or below measurement detection limits, missing and/or suspect data values, complex trends and patterns in mean concentration levels over time and/or space, complicated cause-and-effect relationships, and the frequent need to measure more than one variable at a time. All these problems are also present in the context of wastewater treatment and have been discussed by several researchers (Dean & Forsythe 1976a, b; Berthouex et al. 1978; Niku et al. 1979, 1981, 1982; Berthouex & Hunter 1981, 1983; Yoo et al. 2002).

If the assumption of normality does not hold, significance levels may be distorted, the methods may suffer loss of power, and the estimates obtained based on this assumption may be severely inaccurate. Users frequently disregard assumptions required for parametric tests, and their results are likely to be correspondingly incorrect or unreliable (Potvin & Roff 1993; Modarres et al. 2005).

Thus, it is very important to know the probability distribution of the constituents present in wastewater treatment plants (WWTP), in order to be able to generate synthetic data, feed process models, select adequate statistical methods for evaluating performance and interpret measures of central tendency, all with a high degree of confidence, considering its variability and reliability.

Theoretically, positive skewness (distribution toward higher values) is to be expected with monitoring data as there are usually lower bounds on concentration data (lower limit equal to zero or detection limit, with no negative values), but there are no upper limits. In fact, many wastewater treatment research studies relate that, in most cases, the lognormal distribution provided a reasonable...
description of the effluent BOD (biochemical oxygen demand) and SS (suspended solids) concentration data (Dean & Forsythe 1976a, b, Niku et al. 1979, 1981, 1982; Berthouex & Hunter 1981, 1983; Metcalf & Eddy 2003; Charles et al. 2005).

Implications of non-normality or lack of symmetry in wastewater treatment performance evaluation

Implications for measurements of central tendency

In the context of data asymmetry, it is opportune to discuss the use of the arithmetic mean as a measure of central tendency for asymmetrical data. The notion of central tendency is an attempt to describe ‘typical’ or ‘average’ behaviour of a random quantity. For skewed or for heavily tailed distributions, the sample mean has generally been found inappropriate because of the drastic impact of observations far from the bulk of the data (outliers). Such observations occur with reasonably high probability when sampling from lognormal distributions, which often fits data for processes having an underlying multiplicative character, that is, when the observed variables are themselves products of more basic variables. In problems with a multiplicative character, such as WWTP data, the geometric mean is likely to be a more appropriate measure of central tendency than the arithmetic mean and should be used in setting standards in water quality management (Dean & Forsythe 1976a; Niku et al. 1979, 1981; Metcalf & Eddy 2003). For coliforms, the current practice is to calculate geometric means, but this practice has not been the case for the other constituents.

For the lognormal distribution, the mode is lower than the median, which is lower than the arithmetic mean. The geometric mean is also lower than the arithmetic mean, and may be somewhat close to the median. Both the geometric mean and the median become increasingly smaller than the arithmetic mean as the coefficient of variation increases.

It is entirely understood that researchers are very used to report central tendency by arithmetic means, and also readers are more accustomed to this form of expression. This publication does not intend to change this traditional behaviour, but makes the point that the other central tendency measures such as median and geometric mean could be also included in the descriptive statistical tables, and that care must be exercised when applying parametric tests (which rely on the symmetry assumption).

Implications for measurement of dispersion of data

In many publications, when reporting monitoring data and expressing central tendency and dispersion, researchers frequently use the format \( x \pm S \) (arithmetic mean \( \pm \) standard deviation). When including the plus and minus signs, it is implied that a symmetry around the mean value exists, and that the distribution of values to the left and right of the mean are equal, which is not in accordance with the findings from this research. If one wishes to report descriptive statistics of WWTP performance including arithmetic mean and standard deviation, it would be appropriate not to report them with the traditional ‘ \( \pm \) ’ sign.

Implications for the assessment of the compliance with wastewater discharge standards

Several papers have dealt with the application of statistical concepts to the setting of standards for wastewater discharges (Dean & Forsythe 1976a; Niku et al. 1979, 1981, 1982; McBride 2001; Metcalf & Eddy 2003) but, in general, legal requirements have not included statistical considerations. If the performance of a plant obeys lognormal statistics, there is always a concrete probability of exceeding any finite upper limit, even though this probability may be very small. Use of probabilistic methods in setting discharge standards is a realistic and practical approach from an operational point of view. By knowing or assuming the variability of plant performance, it should be possible to determine what average effluent quality must be maintained to meet a standard with some predetermined probability. If the legislation is not based on maximum permissible values, but rather allows the constituent to exceed the limit value up to a certain percentage of cases, the required performance of the plant can be estimated. As discussed in Metcalf & Eddy (2003) and Oliveira & von Sperling (2008), depending on the desired reliability level, the plant should be designed for higher or lower effluent concentrations.

Implications for reliability assessment

The reliability of a system can be defined as the probability of achieving adequate performance for a specified period of time under specified conditions. In terms of performance of a wastewater treatment plant (WWTP), the reliability can be understood as the percentage of time at which the expected effluent concentrations comply with specified discharge standards or treatment targets (Dean & Forsythe 1976a, b;

Because of variations in performance, a WWTP should be designed to produce an average effluent concentration below the discharge standards. A mean value should be used to guarantee an effluent concentration consistently less than the discharge standard with a certain reliability level. The coefficient of reliability developed by Niku et al. (1979), based on the assumed lognormality of the data, can be used to estimate the reliability of the treatment plants, and relates mean constituent values (i.e. design or operational value) to the standards that must be achieved on a probability basis.

Oliveira & von Sperling (2008) applied the methodology developed by Niku et al. (1979) and calculated the values of the coefficient of reliability (COR). If one multiplies the value of the legal discharge standard by the COR, the resulting value corresponds to the arithmetic mean that needs to be obtained in the WWTP in order to comply with the standards for a certain reliability level. For a reliability level of 95% (95% of the samples complying with the discharge standards), Oliveira & von Sperling (2008) obtained values of COR ranging between 0.4 and 0.6 for BOD (biochemical oxygen demand), COD (chemical oxygen demand), TSS (total suspended solids), TN (total nitrogen) and TP (total phosphorus), and between 0.3 and 0.4 for FC (thermotolerant or faecal coliforms). For instance, for COR equal to 0.5 and a discharge standard of 20 mg/L for BOD, the WWTP needs to be designed and operated to achieve a mean arithmetic effluent concentration of $0.5 \times 20 = 10 \text{ mg/L}$, in order to comply with the standards for 95% of the time.

**Implications for quality control charts**

Statistical process control (SPC) involves statistical techniques to measure and analyze the variation in processes. The main feature of SPC is the control chart, in which the measures of quality or quantity are plotted as a time series graph, on which three horizontal lines are drawn to aid decision-making. The central line is drawn at an appropriate target value. This target value may be set at the mean of past data or it may be related to the specification limits. The upper and lower action lines, or control lines, are placed three standard deviations above and below the centre line. If the plotted points lie between the two action lines and are scattered at random around the centre line, then the process is stable or ‘in control’ (the process is influenced only by random causes or common causes). However, if a point falls outside the action line, or a non-random pattern of points is observed, then an assignable cause, or special cause, has influenced the process. Unfortunately, if the analyzed population is markedly non-normal, then using simple ideas based on testing means may give rise to inappropriate charts that will either fail to detect real changes in the process or which will generate spurious warnings when the process has not changed (Caulcutt 1995; Cheng & Xie 2000; Albers & Kallenberg 2004).

**Objectives of the work**

Although there have been studies on the distributions of influent BOD and SS (cited earlier in the paper), very little has been researched or published about the behaviour of other constituents in wastewater (untreated and treated) and the relationship with the treatment technology. Also, few works have concentrated on the specific conditions of wastewaters in developing countries. Therefore, this paper aims at addressing this lack of information through a series of tests that focused on determining the normality or non-normality of a large data set – influent and effluent biochemical oxygen demand, chemical oxygen demand, total suspended solids, total nitrogen, total phosphorus, thermotolerant or faecal coliforms – as well as the influent flow rate. It also attempted to verify the best fitting of a specific theoretical distribution to the various constituents. The data used in this research came from operational records from wastewater treatment plants, covering different treatment technologies, in a developing country (Brazil).

**METHODS**

**Data sources**

The data used in this research came from 35 wastewater treatment plants, considering 11 different treatment technologies, located in Brazil (latitudes 20 to 22° South, tropical climate, average liquid temperatures between 20 and 25 °C). All treatment plants receive wastewater from urban origin, mostly from domestic sources, but also incorporating occasional minor components of industrial effluents. Separate sewerage systems are adopted in the 35 systems investigated, but cross-connections occur at different degrees, meaning that a fraction of storm water may be present in the influent to the treatment plants during storm events.

The data used were obtained directly from the operational records of the Water and Sanitation companies...
responsible for the operation of the treatment plants. The data obtained span 27 years, with variations for each specific plant. The selected treatment plants represent a wide range of geographic locations with different environmental, social and economic conditions, as well as process types and plant sizes. Monitoring frequency was very variable from plant to plant, with weekly, fortnightly, monthly and quarterly frequencies, together with undefined monitoring routines. Also sampling collection varied: grab and composite samples have been obtained (it is not possible to discriminate them from the plant’s records). All data were analysed in this work in order to check for consistency.

The 55 evaluated wastewater treatment technologies were: overland flow (one plant), trickling filter (one plant), septic tank + anaerobic filter (four plants), facultative aerated lagoon (two plants), completely mixed aerated lagoon (one plant), anaerobic pond + facultative pond (six plants), facultative pond (seven plants), facultative pond + maturation pond (one plant), activated sludge (nine plants), upflow anaerobic sludge blanket reactor (one plant), UASB reactor + facultative pond (two plants).

Evaluation of normality

In order to arrive at the distribution that best fits the data, sequential steps were taken in order to make a systematic analysis: (i) preliminary evaluation based on skewness and kurtosis; (ii) assessment of normality or non-normality; and (iii) goodness-of-fit tests for candidate distributions (normal, lognormal, rectangular, exponential or gamma distributions).

The coefficient of skewness and kurtosis were used as descriptive and inferential measures for evaluating normality, as suggested by D’Agostino & Pearson (1973), Pearson et al. (1977) and D’Agostino et al. (1990). The coefficient of skewness is a measure of deviation from symmetry. The coefficient is close to zero when the distribution is a symmetric curve. The coefficient tends to be positive for distributions skewed to the right (i.e., the data are clustered more to the left of the mean with most of the extreme values to the right) and negative for those skewed to the left. The coefficient of kurtosis is a measure of whether the data are peaked or flat relative to a normal distribution, which has a coefficient of kurtosis equal to 3.

Goodness-of-fit tests for candidate distributions

The characterization of the distributions of flow and influent and effluent concentrations of six constituents (BOD, COD, TSS, TN, TP, FC) was performed in two steps. The first one attempted to check the data normality or non-normality, by using the Kolmogorov–Smirnov test for normality, the Shapiro–Wilks W test, the Lilliefors test and the graphical test normal probability plot. This plot is a graphical representation of the data that will be approximately a straight line if the underlying distribution is normal. Deviations from linearity correspond to various type of non-normality. Some of these deviations reflect skewness and/or kurtosis. Others reflect features such as the presence of outliers, mixtures in the data or truncation (censoring) in the data (D’Agostino et al. 1990).

In the second step, the following goodness-of-fit tests were performed: chi-squared test, Kolmogorov–Smirnov test and Lilliefors test, to choose the probability distribution that fits the WWTP concentration data. These tests were used to examine whether wastewater constituents concentration followed the normal, lognormal, rectangular, exponential or gamma distributions. Theoretical distributions have been selected for this analysis, as opposed to simple fittings of empirical functions, because it is assumed that the former are more likely to be able to represent data behaviour outside the domains in which they have been gathered. To compare the suitability of the several candidate distributions, the probability–probability plot was used. Plotting the data is particularly recommended because goodness-of-fit tests are incapable of discriminating between distributions. Goodness-of-fit tests were made at the 5% significance level, which may be defined as the probability of rejecting a hypothesis that is, in fact, true, the so-called type I error. The software Statistica 6.1 was used for statistical analysis of data throughout the study.

RESULTS

Data characteristics

The number of data of the constituents analyzed varied substantially, from only 10 data in some treatment plants to 909 in others. In some cases, the WWTP did not monitor all constituents, and the number of inflow rate data was largely reduced due to the lack of monitoring of this specific variable.

The monitoring frequency, the number of WWTP and percentage of treatment plants within each category showed a great variability. In terms of sampling frequency and monitoring period variations from one to 2,879 days were observed considering the first and the last sample.
The majority of treatment plants had no clearly identifiable monitoring frequency. This situation, in a way, reflects usual conditions found in operational control of wastewater treatment plants in developing countries. A further discussion of these points can be found in Oliveira & von Sperling (2011).

Evaluation based on skewness and kurtosis

The coefficients of skewness and kurtosis of the 35 WWTP were calculated for BOD, COD, TSS, TN, TP, FC concentration of the raw and treated wastewater, and the inflow rate (results not shown here). In most cases, it was found that the skewness coefficient was positive, a fact that indicates that the data, generally, were not symmetrical and were skewed to the right of the most frequent values. These results confirm theoretical observations that positive skewness is to be expected for wastewater treatment data as there are, usually, no negative values and the upper physical limits lie far away from the usual effluent concentration.

Examination of the results of the coefficient of kurtosis indicated that most values were different from 3, which is the reference value for a normal distribution.

Goodness-of-fit tests

As discussed in the Methods section, the characterization of the distributions of the influent and effluent BOD, COD, TSS, TN, TP, FC concentrations and flow was performed in two steps. The first one checked the data normality and the second one tested the goodness-of-fit for candidate distributions (normal, lognormal, rectangular, exponential or gamma distributions). Considering the two steps, 3,444 tests were applied, one by one, and the results are presented in Table 1.

The conclusion on the ‘normality’ of the data was based on a joint evaluation of the four methods described: Kolmogorov–Smirnov test for normality, Shapiro–Wilk’s W test, Lilliefors test and graphical test normal probability plot. It should be mentioned that there was a great variability in the results of the tests for normality, depending on the test applied. The differences between Kolmogorov–Smirnov and Shapiro–Wilk tests were substantial. It is also important to report that some authors (Shapiro & Wilk 1965; D’Agostino et al. 1990) mention that the Kolmogorov–Smirnov test has shown poor power properties and D’Agostino et al. (1990) strongly recommend that it should not be used when testing for normality. In general, the normality tests indicated the prevalence of the non-normality of the concentration data, especially when considering the effluent concentration of all constituents.

The results of the goodness-of-fit tests, including the chi-squared, Kolmogorov–Smirnov and Lilliefors tests, were used to make inferences about the underlying population. It should be emphasized that the goodness-of-fit tests do not discriminate between distributions, but provide an analysis of whether or not a hypothesized distribution gives a good fit. Thus, it was necessary to make a united analysis of the goodness-of-fit tests and to use charts (probability–probability plot) for choosing between the candidate distributions. The term ‘other’ was used when any

Table 1 | Best theoretical distribution that fits the influent and effluent data (% of WWTPs)

<table>
<thead>
<tr>
<th>Concentration data</th>
<th>Results of the tests for normality</th>
<th>Results of the goodness-of-fit tests</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Normal (%)</td>
<td>Non-normal (%)</td>
</tr>
<tr>
<td>Raw wastewater</td>
<td>BOD</td>
<td>47</td>
</tr>
<tr>
<td></td>
<td>COD</td>
<td>35</td>
</tr>
<tr>
<td></td>
<td>TSS</td>
<td>48</td>
</tr>
<tr>
<td></td>
<td>TN</td>
<td>28</td>
</tr>
<tr>
<td></td>
<td>TP</td>
<td>50</td>
</tr>
<tr>
<td></td>
<td>FC</td>
<td>38</td>
</tr>
<tr>
<td></td>
<td>Flow rate</td>
<td>38</td>
</tr>
<tr>
<td>Treated wastewater</td>
<td>BOD</td>
<td>44</td>
</tr>
<tr>
<td></td>
<td>COD</td>
<td>55</td>
</tr>
<tr>
<td></td>
<td>TSS</td>
<td>32</td>
</tr>
<tr>
<td></td>
<td>TN</td>
<td>22</td>
</tr>
<tr>
<td></td>
<td>TP</td>
<td>33</td>
</tr>
<tr>
<td></td>
<td>FC</td>
<td>31</td>
</tr>
</tbody>
</table>

Note: Values in bold highlight the distribution with the highest percentage of occurrence (best fitting).
other distribution, among the proposed models, provided an adequate fit to the data or when it was not possible to decide between the candidate distributions, considering the results of the tests.

The lognormal distribution should not be rejected for most plants at a significance level of 5%. Generally, the results reported in Table 1 show that a lognormal distribution provides a reasonable description of the BOD, COD, TSS, TN and FC concentration, considering the raw and treated wastewater (influent and effluent) data and flow, as well. The only exception was observed for the influent TP (raw wastewater), whose results showed that most plants presented data better described by the normal distribution. It is important to emphasize that two studies were performed, one with and one without the exclusion of the outliers, and the results of the second one were even more evident about the prevalence of the lognormal distribution.

These results are in agreement with the findings from several authors who have studied the distribution of the concentration data from WWTP (most of them considering BOD and TSS), and have reported that the lognormal distribution gives a good overall fit to effluent values (Dean & Forsythe 1976a; Niku et al. 1979, 1981, 1982; Berthouex & Hunter 1981, 1983; Metcalf & Eddy 2003; Charles et al. 2005).

Some other important inferences can be drawn from the data analysis (results not shown here). The type of wastewater treatment technology does not seem to have influence in the selection of the best theoretical probability distribution that represents the WWTP data. It was observed that the quality and size of the sample seem to have a greater impact than the process used to treat the wastewater. The plants that had no clearly identifiable monitoring frequency, with sparse sampling throughout the time, represented in this study by overland flow (OF) and facultative ponds (FP and AP + FP), provided less conclusive information about the distribution that fits the data. In general, ‘other’ distribution was found to fit most data of these plants.

The characterization of the lognormal distribution as the more generally applicable to the data sets investigated also opens the way for its utilization in making predictions for future situations. This statement is supported by other studies, in which the lognormal distribution has been used successfully to make predictions about future performances of treatment plants, considering the compliance with quality standards and reliability analysis of the systems, as can be seen in Niku et al. (1979, 1982), Crites & Tchobanoglous (2000), Metcalf & Eddy (2003), Oliveira & von Sperling (2007) and Oliveira & von Sperling (2008).

**Degree of departure from normality**

Even though it has been shown that the investigated flow and concentration data are better represented by the lognormal distribution, it is appropriate to analyse whether the departure from the typical normal distribution was substantial. This concern is due to the fact that most researchers still imply symmetry when doing the interpretation of monitoring data or undertaking statistical tests.

The probability density function (pdf) for a lognormal distribution shows higher departures from normality with increasing values of the coefficient of variation CV (standard deviation/arithmetic mean). Ott (1995) states that for CV values less than 1/6 (0.16667), the probability density function of the lognormal distribution shows a very similar behaviour to the normal distribution. Figure 1 shows an example of 1,000 synthetically generated data of BOD concentration, for a fixed arithmetic mean (55.2 mg L\(^{-1}\)) and different CV values (CV = 0.1, CV = 0.5, CV = 1.0 and CV = 2.0). It is clear that for CV values greater than 0.5 the departure from normality is already substantial.

Data obtained by Oliveira & von Sperling (2008) from 166 WWTP plants in Brazil (in which the 35 plants from the present study are included) indicated that the mean values of the coefficient of variation CV varied between 0.4 and 1.0 for BOD, COD, TSS, TN and TP, and between 1.0 and 5.0 for FC. This stresses the fact that, for all cases, considerable data asymmetries were observed. This result is a strong message for researchers that great caution should be exercised when doing statistical analysis of monitoring data and applying tests that are based on normality or symmetry of the data.

![Figure 1](https://iwaponline.com/wst/article-pdf/65/4/596/442402/596.pdf)
SUMMARY AND CONCLUSIONS

The data from the wastewater treatment plants investigated were generally not symmetrical. The distributions were skewed to the right and the coefficient of kurtosis indicated that most values were different from 3, which is the usual value for a normal distribution.

Due to the lack of a solid basis for selecting the distributions that best fit the data, various distributions had to be tried and the choice was based on statistical tests and visual techniques. Practically all constituents present in raw and treated wastewater (BOD, COD, suspended solids, total nitrogen, total phosphorus and thermotolerant coliforms), together with flowrate, reasonably followed the lognormal distribution. These results are in agreement with the findings from several authors who studied the distribution of the concentration data from WWTPs. However, the present paper advanced on this aspect, because most of the publications so far dealt only with BOD and suspended solids.

The type of wastewater treatment technology did not seem to have influence on the selection of the best theoretical probability distribution that represents the WWTP data. It was observed that the quality and size of the sample seem to have greater impact than the process used to treat the wastewater. The plants that had no clearly identifiable monitoring frequency, with sparse sampling throughout the time, provided less conclusive information about the distribution that fits the data.

The general conclusion of the paper is that, regardless of the factors in the urban area that dictated the characteristics of the influent wastewater and of the treatment process and operational strategies that influenced the behaviour of the effluent, the lognormal distribution has been shown to give the overall best fit to the data. This conclusion requires a different position from the one currently adopted when analysing plant performance, in which symmetry of the data is generally implied, which has been shown not to be the case with the large array of data sets investigated.

ACKNOWLEDGEMENTS

The authors would like to thank FINEP, FAPEMIG, CNPq, SABESP, COPASA, FEAM, USP and municipalities for their support and cooperation during the course of the research.

REFERENCES


D’Agostino, R. & Pearson, E. S. 1973 Tests for departure from normality, empirical results for the distributions of b2 and \(|z|b1). Biometrika 60 (3), 613–622.


First received 1 May 2011; accepted in revised form 10 October 2011