Water Quality Analysis of Drinking Water by Identification of Their Distribution and Process Capability

Smita Jain

Department of Mathematics, Jaipur Engineering College and Research Centre, Jaipur, India

Email address: smitajain.maths@jecrc.ac.in

To cite this article:
Smita Jain. Water Quality Analysis of Drinking Water by Identification of Their Distribution and Process Capability. American Journal of Biological and Environmental Statistics. Vol. 4, No. 1, 2018, pp. 20-30. doi: 10.11648/j.ajbes.20180401.14

Received: November 27, 2017; Accepted: December 12, 2017; Published: February 8, 2018

Abstract: The present study deals with the water quality management using Water Quality Index (WQI) and Distribution Identification model. The drinking water quality in Jaipur city has been used for the various physio-chemical parameters such as temperature, pH, turbidity, conductance, total hardness, alkalinity, magnesium, calcium, nitrate, chloride, fluoride, sodium, potassium. Drinking water suitability for domestic purposes was examined by WHO and BIS standards, which indicate the drinking water in almost all the areas, were not much suitable for drinking purpose. It is observed that Water Quality Index (WQI) is very high in Sanganer and VKI areas and Gamma distribution model fits on all the parameters for different sites of water data. Thus it was observed that potable water quality has degraded with respect to all the parameters and almost in all the Sites parameters.

Keywords: Water Physio-Chemical Parameters, WQI, Distribution Identification Model, Process Capability and Health Effects

1. Introduction

Potable water or drinking water has always been an important and life sustaining drinks to humans and is essential to the survival of all living organisms. The potable water should be free from pathogenic agents and chemical constituents, pleasant to taste and usable for domestic purposes. Parameters for drinking water quality typically fall under two categories - Chemical and Physical. Chemical parameters causes’ chronic health risks through built up of heavy metals whereas physical parameters affect the aesthetics and taste of drinking water [10]. Water supply in Jaipur comes either from surface water or ground water sources. The ground water is characterized by multiple quality problems. Many hazardous pollutants viz., colored dyes, heavy metals, nitrates and fluoride, agriculture runoffs, waste water treatment plants, combined sewage overflows, storm water discharges, failed septic tanks, prescription drugs etc. Pollute it.

Water Quality Index may be defined as a rating, reflecting the composite influence on overall quality characteristics of water of individual quality parameters, which is being regarded as one of the most effective way to communicate water quality.

This will give us only certain numerical value but for estimating exact quality of water an indexing system has been developed known as Water Quality Index, which gives us the idea of water of whole system [2].

Quality rating equal to zero means a complete absence of pollutants. While 0 to 100 implies that the pollutants are under consideration means within the prescribed standards. When quality rating is more than 100 it implies that the pollutants are above the standards. Similarly when WQI is 0 to 100 indicate that the water is under consideration and is safe for use.

Distribution Identification Model is used to evaluate the optimal distribution based on the probability plots and goodness-of-fit tests prior to conducting a capability analysis study [2].

Capability analysis study is used to produce a process capability report when data do not follow a normal distribution. The report includes a capability histogram overlaid with a distribution curve, and a table of overall capability statistics [7]. The report also includes statistics of the process data, such as the mean, distribution parameter estimates and process specifications; the actual overall
capability (Pp, Ppk, PPU, and PPL), and the observed and expected overall performance. The report can be used to visually assess the distribution of the process relative to the target, whether the data follow a specified distribution, and whether the process is capable of meeting the specifications consistently.

Study Area: The present study was carried out in Jaipur city which is situated in eastern part of Rajasthan surrounded on three sides by the rugged Aravali hills. The city has total geographical area of 14,068 sq. km. (4.11 percent of the state). It is located at 26.92°N 75.82°E. It has an average elevation of 431 meters.

The climate of the city is hot semiarid with extremes of temperature (15-45°C) and rainfall 650 mm (26 inch.). In this city water samples of selected for 10 different areas such as VKI, Jhothwara, Chitrkoot, Vaishali Nagar, Transport Nagar, Kanota, Bari Chopar, Subhash Chowk, Gopalpura and Sanganer were used with the view to understand the quality of potable water.

2. Statistical Analysis

2.1. Analysis for Water Quality Index

The basic equation for Water Quality Index is W. Q. I. = Σ qi wi
Where qi = quality ratings of parameters
wi = unit weight of different parameters
qi wi = parameters sub-index

2.1.1. Quality Ratings of the Parameters

Quality ratings qi = \[\frac{VO - VI}{VS - VI}\] * 100
Where VO = observed value of the parameter
VS = standard value of the parameter
VI = ideal value of the parameter
For pH = 7 and for dissolved oxygen = 14.6 mg/l
While VI is zero for other parameters.

2.1.2. Unit Weight for Various Parameters

Wi = K/ Si
Where K = 1+1/si1 + 1/\sqrt{s2} +1/\sqrt{s3}+ \ldots \ldots +1/\sqrt{sn}
Si = recommended standard value for the corresponding Parameter
si = standard value for the corresponding parameter

2.2. Distribution Identification Model

By the help of Mini Tab (Software) Distribution Identification Model is used to evaluate the optimal distribution based on the probability plots and goodness-of-fit tests prior to conducting a capability analysis study.

2.2.1. MLE

Maximum likelihood estimates of the parameters are calculated by maximizing the likelihood function with respect to the parameters. The likelihood function describes, for each set of distribution parameters, the chance that the true distribution has the parameters based on the sample [5].

The Newton-Raphson algorithm is used to calculate maximum likelihood estimates of the parameters which define the distribution. The Newton-Raphson algorithm is a recursive method for computing the maximum of a function. The percentiles are calculated from that distribution.

2.2.2. Pp

Pp is calculated using the parameter values estimated using maximum likelihood method for a distribution used in the analysis.

\[Pp = \frac{Z_{ucl} - Z_{lcl}}{6}\]

Where: Z_{lsl} = F^{-1}(p_1)
Z_{usl} = F^{-1}(p_2)
F^{-1} (p) = p * 100th percentile of a standard normal distribution
p_1 = Prob (X \leq LSL)
p_2 = Prob (X \leq USL)
X follows the distribution used in the analysis

2.2.3. PPL

PPL is calculated using the parameter values estimated using maximum likelihood method for a distribution used in the analysis.

\[PPL = \frac{-F^{-1}(p_1)}{3}\]

where: F^{-1} (p) = p * 100th percentile of a standard normal distribution
p_1 = Prob (X \leq LSL)
X follows the distribution used in the analysis

2.2.4. PPU

PPU is calculated using the parameter values estimated using maximum likelihood method for a distribution used in the analysis.

\[PPU = \frac{F^{-1}(p_2)}{3}\]

Where: F^{-1} (p) = p * 100th percentile of a standard normal distribution
p_2 = Prob (X \leq USL)
X follows the distribution used in the analysis

2.2.5. Ppk

The minimum of PPU and PPL

3. Result and Discussion

Table 1 and Table 2 give the detail of Quality Ratings, unit Weight and Water Quality Index of different areas for all quality parameters. It is observed that WQI is not in the prescribed limit for all the areas and for Sanganer and VKI, WQI is so high.
Identification of distribution and process capability for different parameters:

**Ph**

Descriptive statistics The table of descriptive statistics provides the summary information for the whole column of data. All the statistics are based on the non-missing (N = 10) values. For these data, \( \mu = 8.16 \) and \( \sigma = 0.483506 \).

Goodness-of-fit test The table includes Anderson-Darling (AD) statistics and the corresponding p-value for a distribution. For a critical value \( \alpha \) a p-value greater than \( \alpha \) suggests that the data follow that distribution. The p-value 0.259, 0.003, 0.250, and 0.238 indicate that the gamma distributions fit the data well because this has the largest p value which is greater than \( \alpha \).

Probability plot: The probability plot includes percentile points for corresponding probabilities of an ordered data set. The middle line is the expected percentile from the distribution based on maximum likelihood parameter estimates. The left and right line represents the lower and upper bounds for the confidence intervals of each percentile. The probability plots shows that the data points fall approximately on a straight line and within the confidence intervals for Normal, Weibul and gamma distribution.

### Table 3. Descriptive Statistics for the parameter Ph.

| N  | N* | Mean | StdDev | Median | Minimum | Maximum | Skewness | Kurtosis |
|----|----|------|--------|--------|---------|---------|----------|----------|
| 10 | 0  | 8.16 | 0.483506 | 8.3    | 7.5     | 8.9     | -0.187851 | -1.25583 |

### Table 4. Goodness of Fit Test for the Parameter Ph.

| Distribution | AD  | P    |
|--------------|-----|------|
| Normal       | 0.420 | 0.259 |
| Exponential  | 4.104 | 0.003 |
| Weibull      | 0.397 | 0.250 |
| Gamma        | 0.483 | 0.238 |
Table 5. ML Estimates of Distribution for Parameters Ph

| Distribution  | Location  | Shape  | Scale  | Threshold |
|---------------|-----------|--------|--------|-----------|
| Normal*       | 8.16000   | 0.48351|        |           |
| Exponential   | 8.16000   |        |        |           |
| Weibull       | 20.63355  | 8.37381|        |           |
| Gamma         | 313.90323 | 0.02600|        |           |

Figure 1. Probability Plot for the Parameter Ph.

The histogram shows the evidence of serious discrepancies between the assumed model and the data.

We can see that the right tail of the distribution falls outside the upper specification limit. This means at all the areas the parameter Ph shows the values higher than the upper specification.

The Ppk and PPU indices tell us whether the process is within the tolerance limits. Both indices are -0.79, below the guideline. Thus this process of Ph cannot be considerable.

Likewise, the observed PPM > UCL is 1000000. This value means that 1000,000 out of 1,000,000 points more than the upper specification limit that is in all areas Ph has more than the upper specification limit of 7.

Figure 2. Process Capability for the Parameter Ph.
Turbidity

Descriptive statistics: The table of descriptive statistics provides the summary information for the whole column of data. All the statistics are based on the non-missing (N = 10) values. For these data, μ = 9.36 and σ = 0.839577.

Goodness-of-fit test: The table includes Anderson-Darling (AD) statistics and the corresponding p-value for a distribution. For a critical value α a p-value greater than α suggests that the data follow that distribution. The p-value 0.68, 0.003, 0.250, and 0.250 indicate that the Normal, Weibull and gamma distributions fit the data well because this has the largest p-value which is greater than α.

Probability plot: The probability plot includes percentile points for corresponding probabilities of an ordered data set. The middle line is the expected percentile from the distribution based on maximum likelihood parameter estimates. The left and right line represents the lower and upper bounds for the confidence intervals of each percentile.

The probability plots shows that the data points fall approximately on a straight line and within the confidence intervals for Normal, Weibull and gamma distribution.

The histogram show the evidence of serious discrepancies between the assumed model and the data.

We can see that the right tail of the distribution falls outside the upper specification limit. This means at all the areas the parameter Turbidity shows the values higher than the upper specification.

The Ppk and PPU indices tell us whether the process within the tolerance limits. Both indices are -1.66, below the

---

**Table 6.** Descriptive Statistics for Turbidity.

| N      | N* | Mean | StDev  | Median | Minimum | Maximum | Skewness | Kurtosis |
|--------|----|------|--------|--------|---------|---------|----------|----------|
| 10     | 0  | 9.36 | 0.839577 | 9.4    | 7.8     | 10.9    | -0.0277117 | 0.913103 |

**Table 7.** Goodness of Fit Test for Turbidity.

| Distribution | AD    | P       |
|--------------|-------|---------|
| Normal       | 0.245 | 0.680   |
| Exponential  | 3.881 | <0.003  |
| Weibull      | 0.318 | >0.250  |
| Gamma        | 0.255 | >0.250  |

**Table 8.** ML Estimates of Distribution Parameters for Turbidity.

| Distribution | Location | Shape | Scale | Threshold |
|--------------|----------|-------|-------|-----------|
| Normal*      | 9.36000  | 0.83958 |       |           |
| Exponential  |          | 9.36000 |       |           |
| Weibull      | 12.50797 | 9.72588 |       |           |
| Gamma        | 136.58407| 0.06853 |       |           |

**Figure 3.** Probability Plot for the Parameter Turbidity.
Likewise, the observed PPM > UCL is 1000000. This value means that 1000,000 out of 1,000,000 points more than the upper specification limit that is in all areas Turbidity has more than the upper specification limit of 5.00.

TDS

Descriptive statistics. The table of descriptive statistics provides the summary information for the whole column of data. All the statistics are based on the non-missing (N = 10) values. For these data, \( \mu = 1388.5 \) and \( \sigma = 244.427 \).

Goodness-of-fit test. The table includes Anderson-Darling (AD) statistics and the corresponding p-value for a distribution. For a critical value \( \alpha \) a p-value greater than \( \alpha \) suggests that the data follow that distribution. The p-value 0.054, 0.003, 0.045, and 0.071 indicate that the gamma distributions fit the data well because this has the largest p value which is greater than \( \alpha \).

Probability plot The probability plot includes percentile points for corresponding probabilities of an ordered data set. The middle line is the expected percentile from the distribution based on maximum likelihood parameter estimates. The left and right line represents the lower and upper bounds for the confidence intervals of each percentile. The probability plots shows that the data points fall approximately on a straight line and within the confidence intervals for Normal, Weibull and gamma distribution.

Table 9. Descriptive Statistics for TDS.

| N   | N* | Mean | StDev | Median | Minimum | Maximum | Skewness | Kurtosis |
|-----|----|------|-------|--------|---------|---------|----------|----------|
| 10  | 0  | 1388.5 | 244.427 | 1255   | 1100    | 1760    | 0.584265 | -1.44218 |

Table 10. Goodness of Fit Test for TDS.

| Distribution | AD   | P     |
|--------------|------|-------|
| Normal       | 0.674| 0.054 |
| Exponential  | 3.286| <0.003|
| Weibull      | 0.733| 0.045 |
| Gamma        | 0.700| 0.071 |

Table 11. ML Estimates of Distribution Parameters for TDS.

| Distribution | Location | Shape | Scale   | Threshold |
|--------------|----------|-------|---------|-----------|
| Normal\*     | 1388.50000 |       | 244.42734 |           |
| Exponential  | 1388.50000 |       |         |           |
| Weibull      | 6.39721  | 1489.81653 |         |           |
| Gamma        | 37.28603 | 37.23915 |         |           |
The histogram shows the evidence of serious discrepancies between the assumed model and the data. We can see that the right tail of the distribution falls outside the upper specification limit. This means at all the areas the parameter TDS shows the values higher than the upper specification.

The Ppk and PPU indices tell us whether the process is within the tolerance limits. Both indices are -1.10, below the guideline. Thus this process cannot be considerable.

Likewise, the observed PPM > UCL is 1000000. This value means that 1000,000 out of 1,000,000 points more than the upper specification limit that is in all areas Turbidity has more than the upper specification limit of 500.

Descriptive statistics The table of descriptive statistics provides the summary information for the whole column of data. All the statistics are based on the non-missing (N = 10) values. For these data, $\mu = 5.75$ and $\sigma = 0.479004$.

Goodness-of-fit test The table includes Anderson-Darling test results for normal, exponential, Weibull, and gamma distributions. The P-values for the normal, exponential, and gamma distributions are 0.054, 0.045, and 0.071, respectively, indicating a reasonable fit for these distributions.

Overall, the process capability analysis suggests that the TDS parameter does not meet the desired performance levels, with significant deviations from the target values and specification limits.
(AD) statistics and the corresponding p-value for a distribution. For a critical value \( \alpha \) a p-value greater than \( \alpha \) suggests that the data follow that distribution. The p-value 0.016, 0.003, 0.138, and 0.017 indicate that the Weibull distributions fit the data well because this has the largest p value which is greater than \( \alpha \).

Probability plot The probability plot includes percentile points for corresponding probabilities of an ordered data set.

The middle line is the expected percentile from the distribution based on maximum likelihood parameter estimates. The left and right line represents the lower and upper bounds for the confidence intervals of each percentile.

The probability plots shows that the data points fall approximately on a straight line and within the confidence intervals for Normal, Weibull and gamma distribution.

### Table 12. Descriptive Statistics for DO.

| N  | N* | Mean  | StDev  | Median | Minimum | Maximum | Skewness | Kurtosis |
|----|----|-------|--------|--------|---------|---------|----------|----------|
| 10 | 0  | 5.75  | 0.479004 | 5.9    | 4.6     | 6.3     | -1.75910 | 3.51392  |

### Table 13. Goodness of Fit Test for DO.

| Distribution | AD   | P        |
|--------------|------|----------|
| Normal       | 0.871| 0.016    |
| Exponential  | 3.988| <0.003   |
| Weibull      | 0.561| 0.138    |
| Gamma        | 0.960| 0.017    |

### Table 14. ML Estimates of Distribution Parameters for DO.

| Distribution | Location | Shape    | Scale      | Threshold |
|--------------|----------|----------|------------|-----------|
| Normal*      | 5.75000  |          |            |           |
| Exponential  | 5.75000  |          |            |           |
| Weibull      | 19.15438 | 5.93017  |            |           |
| Gamma        | 146.45689| 0.03926  |            |           |

The histogram shows the evidence of serious discrepancies between the assumed model and the data.

WE can see that the right tail of the distribution falls within the upper specification limit. This means at all the areas the parameter DO shows the values lower than the upper specification.

Likewise, the observed PPM > UCL is 0 and PPM < LCL is 600000. This value means that 6000,00 out of 1,000,000 points less than the lower specification limit that is in almost all areas DO has less than the lower specification limit of 6.
Nitrate

Descriptive statistics The table of descriptive statistics provides the summary information for the whole column of data. All the statistics are based on the non-missing (N = 10) values. For these data, \( \mu = 102.8 \) and \( \sigma = 18.4860 \).

Goodness-of-fit test The table includes Anderson-Darling (AD) statistics and the corresponding p-value for a distribution. For a critical value \( \alpha \), a p-value greater than \( \alpha \) suggests that the data follow that distribution. The p-values 0.823, 0.003, 0.250, and 0.250 indicate that the Normal, Weibull and gamma distributions fit the data well because this has the largest p value which is greater than \( \alpha \).

Probability plot The probability plot includes percentile points for corresponding probabilities of an ordered data set. The middle line is the expected percentile from the distribution based on maximum likelihood parameter estimates. The left and right line represents the lower and upper bounds for the confidence intervals of each percentile. The probability plots shows that the data points fall approximately on a straight line and within the confidence intervals for Normal, Weibull and gamma distribution.

Table 15. Descriptive Statistics for Nitrate.

| N  | N* | Mean   | StDev | Median | Minimum | Maximum | Skewness | Kurtosis |
|----|----|--------|-------|--------|---------|---------|----------|----------|
| 10 | 0  | 102.8  | 18.4860 | 102.5  | 76      | 135     | 0.262407 | -0.788338|

Table 16. Goodness of Fit Test for Nitrate.

| Distribution  | AD    | P      |
|---------------|-------|--------|
| Normal        | 0.205 | 0.823  |
| Exponential   | 3.185 | <0.003 |
| Weibull       | 0.239 | >0.250 |
| Gamma         | 0.227 | >0.250 |

Table 17. ML Estimates of Distribution Parameters for Nitrate.

| Distribution | Location | Shape | Scale    | Threshold |
|--------------|----------|-------|----------|-----------|
| Normal*      | 102.80000| 18.48603|          |           |
| Exponential  | 102.80000|        |          |           |
| Weibull      | 6.40291  | 110.30535|         |           |
| Gamma        | 34.42168 | 2.98649 |          |           |
The histogram shows the evidence of serious discrepancies between the assumed model and the data. We can see that the right tail of the distribution falls outside the upper specification limit. This means at all the areas the parameter Nitrate shows the values higher than the upper specification. The Ppk and PPU indices tell us whether the process within the tolerance limits. Both indices are -0.84, below the guideline. Thus this process cannot be considerable. Likewise, the observed PPM > UCL is 1000000. This value means that 1000,000 out of 1,000,000 points more than the upper specification limit that is in all areas Turbidity has more than the upper specification limit of 50.

4. Conclusion

Statistical analysis of water quality parameters has been studied by using MINI TAB. It is observed that WQI is so high in all the areas of Jaipur under study and for Sanganer it is highest.

The textile industry is one of the oldest and largest organized sectors in Sanganer area with geographical coordination 26° 49' 0" North, 75° 47' 0" East. Which houses almost 400 small-scale textile printing units, a large volume
of textile wastewater (>10,000 KL/day) is discharged. Further, untreated and sometimes even treated effluents from these industries are released into surface waters of Amanishah drainage or through the drainage systems, seep into the ground water and adjoining water bodies. Since many textile dyes are known carcinogens and mutagens12. These disposal practices have contaminated the environment and caused adverse effects on the flora, crops and fauna in Sanganer town13 and had higher fluoride (2.8 mg/L) compared to other areas.

We study that all the parameters are responsible for this increased WQI in the city. Since all the parameters show the higher values then UCL in Process Capability.

DO is essential for living organism and we has been observed that in all the areas DO has near to its minimum value.

Now we have to think that what we are getting by this urbanization, industries, textile mills etc.

References

[1] Imo TS, Oomori T, Toshihiko M, Tamaki F (2007) the comparative study of trihalomethanes in drinking waters. Int J Environ Sci Tech 4 (4): 421–426.

[2] Kumar A, Dua A (2009) Water quality index for assessment of water quality of river Ravi at Modhopur (India). G J Env Sci 8 (1): 49–57.

[3] Cheng, Bing, and D. M. Titterington. “Neural Networks: A Review from a Statistical Perspective.” Statistical Science 9, no. 1 (1994): 2-30.

[4] Frew, James E., and Jeff Dozier. “Environmental Informatics.” Annual Review of Environment and Resources 37 (2012): pp. 449-472.

[5] Lebanon, Guy. Bias, Variance, and MSE of Estimators. Technical Notes, Georgia: Georgia Institute of Technology, 2010.

[6] Shkurin, Aleksei, and Alfredo Vellido. “Random Forests for quality control in G-Protein Coupled Receptor databases.” 2016.

[7] Starkweather, Jon. Cross Validation techniques in R: A brief overview of some methods, packages, and functions for assessing prediction models. Review, University of North Texas, 2011.

[8] Alam Md JB, Muyen Z, Islam MR, Islam S, Mamun M (2007) Water quality parameters along rivers. Int J Environ Sci Tech 4 (1): 159–167.

[9] Bhardwaj R, Parmar KS (2013b) Wavelet and statistical analysis of river water quality parameters. App Math Comput 219 (20): 10172–10182.

[10] Damodhar U, Reddy MV (2013) Impact of pharmaceutical industry treated effluents on the water quality of river Uppanar, South east coast of India: a case study. Appl Water Sci 3: 501–514.