Appendix A

Houston Engineering Censored Data Study

Statistical Methods For Analyzing
Censored Water Quality Data Sets

Red Lake Watershed District
Thief River Falls, MN 56701
November 2002

I hereby certify that this plan, specification, or report was prepared by me or under my direct supervision, and that I am a duly Registered Professional Engineer under the laws of the State of Minnesota.

Mark R. Deutschman
MN. Reg. No. 41259

Date: 11-12-02

Brent H. Johnson
MN. Reg. No. 20378

Date: 11-12-02

Houston Engineering, Inc.
10900 73rd Avenue N
Maple Grove, MN 55369
HE Project #1030-100

Introduction

The Red Lake Watershed District (RLWD) has received a Challenge Grant from the Board of Water and Soil Resources. As part of the Challenge Grant, RLWD requested assistance from Houston Engineering, Inc. to provide recommended methods for computing statistics with water quality monitoring data sets including data reported as less than the minimum detection limit (MDL). "Censored" refers to data sets where "*...values are only reported for those observations above some predetermined value.*" (Liu and others 1997).

Water quality data at select locations within the RLWD has been collected and recorded since 1980. Some of the water quality samples collected had certain parameters, such as nitrates, that were often below the minimum detection limit (BDL). **Table 1** lists the water quality parameters and the percentage of samples taken that had BDL readings. For statistical reporting, the BDL readings (censored data) pose a unique challenge. Simply skipping the BDL readings discards valuable data and produces estimates of the summary statistics that are biased high. (Liu and others 1997, Spooner 1991). Therefore, a method to "uncensor" the BDL readings is needed to compute summary statistics without excessive bias.

Methods for Estimating Summary Statistics

There are numerous methods that can be used to estimate summary statistics for data sets including BDL readings. Some of these methods are the simple substitution; distributional methods such as the probability plot, maximum-likelihood estimation (MLE), and fill-in with expected values MLE techniques; and the Helsel's Robust Method. As the censored portion of any data set increases, the reliability of the uncensoring techniques will decrease. The USEPA notes that: "*...power will generally decline as censoring increases; when the data are more than 60 to 80 percent nondetects, it is unlikely that any method will perform acceptably.*" (USEPA, 1998).

Table 1

Summary of Censored Data

Site	Censored Data Percentage	Turbidity	Nitrates	Ortho P	Organic P	Total P	Fecal Coliform	TKN	Ammonia	TSS
52	Number of samples	40	49	39	37	48	26	26	16	26
	Number of samples BDL or zero	3	15	2	1	0	21	8	8	2
	Percent listed as BDL or zero	7.5%	30.6%	5.1%	2.7%	0.0%	80.8%	30.8%	50.0%	7.7%
75	Number of samples	43	52	40	38	48	31	29	17	25
	Number of samples BDL or zero	2	16	5	1	0	8	8	8	1
	Percent listed as BDL or zero	4.7%	30.8%	12.5%	2.6%	0.0%	25.8%	27.6%	47.1%	4.0%
757	Number of samples	39	43	39	37	43	31	28	16	24
	Number of samples BDL or zero	1	7	1	0	0	7	5	4	0
	Percent listed as BDL or zero	2.6%	16.3%	2.6%	0.0%	0.0%	22.6%	17.9%	25.0%	0.0%
109	Number of samples	45	53	41	39	53	31	28	14	26
	Number of samples BDL or zero	1	8	0	0	0	4	7	3	1
	Percent listed as BDL or zero	2.2%	15.1%	0.0%	0.0%	0.0%	12.9%	25.0%	21.4%	3.8%
790	Number of samples	48	55	43	39	57	35	28	14	26
	Number of samples BDL or zero	1	6	3	2	3	7	6	4	0
	Percent listed as BDL or zero	2.1%	10.9%	7.0%	5.1%	5.3%	20.0%	21.4%	28.6%	0.0%

Based on the review of available literature, three methods were evaluated for estimating summary statistics on data sets with BDL readings. These included the simple substitution method, probability plot method and the Helsel's Robust Method.

Simple Substitution Method

The simple substitution method takes a numerical value, such as one-half the MDL, and substitutes this for each of the BDL readings. After all of the BDL's have been replaced with the substituted value (zero, one-half MDL, or MDL) the summary statistics are calculated.

As the name implies, the strength of this method is the simplicity of use (Oblinger-Childress and others, 1999). According to the U.S. Environmental Protection Agency (EPA), during a study of uncensoring methods they found that:

> "*General results from all simulations combined indicate that the simple substitution methods perform as well as or better than the more complicated censored data techniques in most situations. In particular, substitution of the detection limit when up to 40 percent of the data are nondetects, or one-half the detection limit when more than 40 percent of the data are nondetects, are methods that work reasonably well for small sample sizes in most cases ...*" (USEPA, 1998).

Capel, Gilliom and Larson note that distribution dependent methods may be used to determine the summary statistics for data sets with low rates of censoring, but recommend that a better approach is to "*...bound the analysis...*," "*...calculating the lowest possible value (setting all nondetections to zero) and the highest value (setting all nondetections to the lowest detected concentration or the MDL)...*" (Capel, Gilliom and Larson, 1996). The Australian and New Zealand Environment and Conservation Council recommend that statistics should be computed for complete data sets, with numeric values substituted for BDL observations, but caution that the impact of this substitution should be clearly understood and confidence limits, hypothesis testing or other inferential analyses should not be performed if 25% or more of the data set is BDL (Australian Guidelines, 2000).

The two weaknesses of the simple substitution method are the potential bias (high if substituting at the MDL and low if substituting at zero) and the fact that there is no theoretical basis for the arbitrary choice of the value used in the substitution (Helsel and Hirsch, 1992). **Table 2** provides the results of statistical summaries of Highlanding data using the simple substitution method in comparison to Helsel's Method and skipping all BDL observations.

Probability Plot Method

The probability plot method is one of several distributional methods that allow estimation of summary statistics based upon the assumption that the data above and below the detection limit follow a statistical distribution (Helsel and Hirsch, 1992). The probability plot method allows estimation of the summary statistics (mean and standard deviation) of a data set containing BDL's. The data are assumed to follow a normal or lognormal distribution. Data including BDL's are ranked from the lowest to highest value. The data above the MDL are plotted on a normal or lognormal distribution plot and a linear regression is performed on the data above the MDL. Porter and Ward explain that using this method: "*Censored observations are substituted with numbers taken from an extrapolation of the regression line into the region less than the censoring threshold.*" (Porter and Ward, 1991). The estimated mean and standard deviation are determined as the intercept and slope of the regression line, respectively.

Helsel and Hirsch found that distributional methods assuming log-normal distributions often perform well for estimating percentile statistics such as the 25th and 75th percentiles, the median and interquartile range—even when the true distribution is not log normal. However, distributional methods may perform poorly for computing moment statistics such as the mean and standard deviation unless the true distribution is log-normal, since moment statistics are sensitive to the largest observations and a distribution that does not fit the largest observations will yield poor moment statistics results. In addition, these methods are subject to transformation bias that occurs when logarithmic regression coefficients are transformed back to original units (Helsel and Hirsch, 1992).

Table 2

Summary Statistics for Select Parameters

SITE 075 RED LAKE RIVER - HIGHLANDING

Parameter	Estimated Summary Statistic	Discard BDL Data	Helsel's Robust Method	Simple Substitution (BDL = 0)	Simple Substitution (BDL's = MDL)
Turbidity n=43 BDL=2	Mean	3.969	3.805	3.785	3.789
	Std Dev	3.792	3.776	3.796	3.791
	25th Percentile	1.700	1.600	1.600	1.600
	Median	2.700	2.610	2.610	2.610
	75th Percentile	4.900	4.650	4.650	4.650
Nitrates n=52 BDL=16	Mean	0.119	0.083	0.083	0.087
	Std Dev	0.241	0.207	0.208	0.206
	25th Percentile	0.010	0.002	0.000	0.010
	Median	0.023	0.010	0.010	0.020
	75th Percentile	0.135	0.045	0.045	0.045
Ammonia n=17 BDL=7	Mean	0.131	0.079	0.077	0.081
	Std Dev	0.127	0.115	0.116	0.113
	25th Percentile	0.026	0.007	0.000	0.010
	Median	0.110	0.019	0.019	0.019
	75th Percentile	0.200	0.124	0.124	0.124
TKN n=29 BDL=8	Mean	1.060	0.825	0.768	0.789
	Std Dev	0.873	0.834	0.882	0.863
	25th Percentile	0.730	0.253	0.000	0.100
	Median	0.900	0.760	0.760	0.760
	75th Percentile	1.000	0.910	0.910	0.910
Ortho P n=40 BDL=5	Mean	0.015	0.013	0.013	0.014
	Std Dev	0.018	0.018	0.018	0.017
	25th Percentile	0.005	0.004	0.004	0.005
	Median	0.007	0.007	0.007	0.007
	75th Percentile	0.016	0.014	0.014	0.014
Organic P n=38 BDL=1	Mean	0.042	0.041	0.041	0.041
	Std Dev	0.051	0.051	0.051	0.051
	25th Percentile	0.017	0.017	0.017	0.017
	Median	0.028	0.028	0.028	0.028
	75th Percentile	0.044	0.044	0.044	0.044
Total P n=49 ADL=1	Mean	0.073	0.078	0.092	0.133
	Std Dev	0.087	0.093	0.158	0.427
	25th Percentile	0.026	0.026	0.026	0.026
	Median	0.042	0.044	0.044	0.044
	75th Percentile	0.062	0.063	0.063	0.063
Fecal Coliform n=31 BDL=8	Mean	42.130	31.485	31.258	31.516
	Std Dev	57.253	52.351	52.488	52.331
	25th Percentile	4.500	1.947	1.000	1.500
	Median	13.000	6.000	6.000	6.000
	75th Percentile	61.000	32.500	32.500	32.500
TSS n=25 BDL=1	Mean	11.002	10.603	10.562	10.602
	Std Dev	8.300	8.366	8.418	8.368
	25th Percentile	3.910	3.640	3.640	3.640
	Median	10.000	9.000	9.000	9.000
	75th Percentile	17.000	16.000	16.000	16.000

Helsel's Robust Method

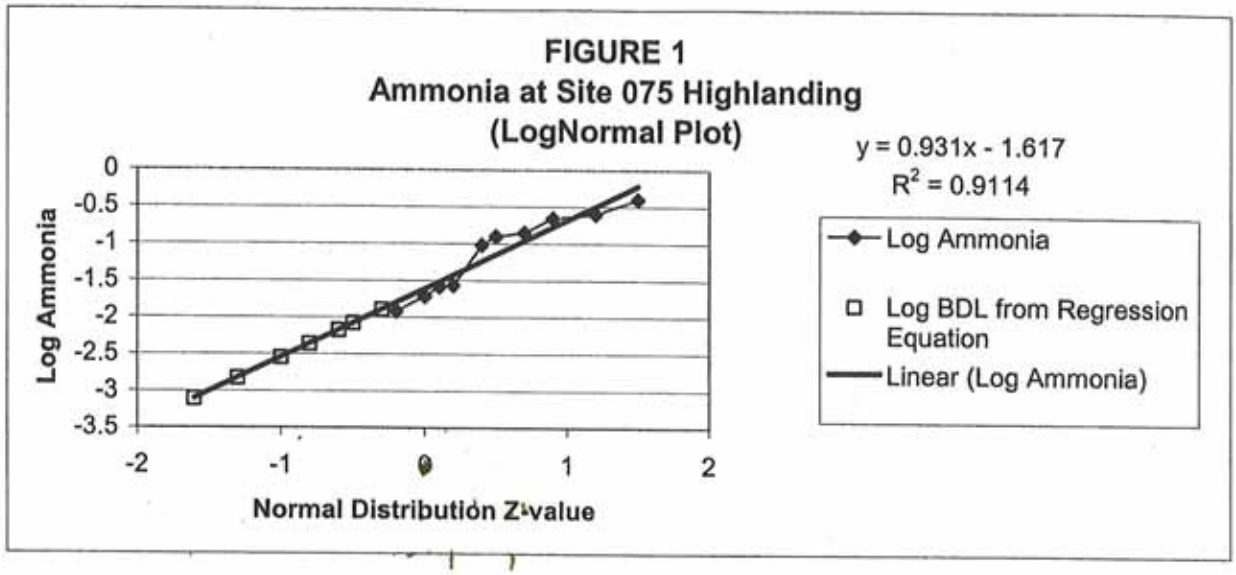
The Helsel's Robust Method combines data above the MDL with values assigned to the BDL readings by assuming a distributional shape (log-normal), to estimate summary statistics. This method calculates the log of the BDL reading using a regression line fitted to the log of the observed values above the MDL and their corresponding normal distribution z scores. The calculated value of the BDL reading is then back-transformed to the original units and the summary statistics are computed (Newman and others, UNCENSOR Users Manual). One note of caution; the calculated values of the BDL readings are not estimates of specific samples, but are used collectively to estimate summary statistics. Therefore, the values calculated for the BDL and zero readings by the Helsel's Robust Method are acceptable for estimating the mean, standard deviation, and other summary statistics but are not to be used as a value for a sample collected on a specific date (for trend analysis, etc.).

According to Helsel and Hirsch the Robust Method has two advantages over the probability plot, MLE and the fill-in with expected values MLE methods. Helsel's Robust Method is not as sensitive to the fit of a distribution for the largest observations because actual data is used rather than a fitted distribution. Also, the estimated summary statistics are computed in original units avoiding transformation bias (Helsel and Hirsch, 1992).

Figure 1 is a graph of the Ammonia data above detection limit at Highlanding as well as the values estimated for the below detection observations using Helsel's Robust Method. **Table 2** provides the results of statistical summaries using Helsel's Method and the simple substitution method in comparison to skipping all BDL observations. **Table 3** presents summary notes on the results of using Helsel's Robust Method. A list of the specific procedures of this method are presented in **Appendix A** as well as graphs of the results of application of the Robust Method to estimate values for the BDL observations for nine parameters at the Red Lake River at Highlanding (Site 075).

Table 3
Summary Notes On Helsel's Robust Method
SITE 075 RED LAKE RIVER - HIGHLANDING

Parameter	Summary Notes On Helsel's Robust Method
Turbidity n=43 BDL=2	Turbidity is predicted above detection in 2 of 2 observations. Summary statistic results are similar to the simple substitution method.
Nitrates n=52 BDL=16	Nitrates include multiple detection limits. Helsel's Method sets all BDL's to highest detection limit. Regression relation looks good and summary statistic results are similar to the simple substitution method.
Ammonia n=17 BDL=7	Ammonia is predicted above detection in 1 of 7 observations. Regression relation looks good and summary statistic results are similar to the simple substitution method.
TKN n=29 BDL=8	Total Kjeldahl Nitrogen is predicted above detection in 7 of 8 observations. Regression relation looks poor and mean estimate is higher than the simple substitution method result, and the 25th percentile estimate is poor.
Ortho P n=40 BDL=5	Reactive P includes multiple detection limits. Helsel's Method sets all BDL's to highest detection limit. Regression relation looks good and summary statistic results are similar to the simple substitution method.
Organic P n=38 BDL=1	Organic P is predicted above detection in 1 of 1 observations. Regression relation looks good and summary statistic results are similar to the either the simple substitution method or the skip BDL observations method.
Total P n=49 ADL=1	Total P includes 1 above detection observation, but the Robust Method predicts a value below the upper limit. The mean and standard deviation predictions are lower than those predicted by the simple substitution method--but not as low as predicted by the skip BDL observations method.
Fecal Coliform n=31 BDL=8	Fecal Coliform is predicted above detection in 3 of 8 observations. Regression relation looks good and summary statistic results are similar to the simple substitution method--except that the 25th percentile estimate is higher than given by the simple substitution method.
TSS n=25 BDL=1	TSS is predicted above detection in 1 of observations. Regression relation looks good and summary statistic results are similar to the simple substitution method.



Multiple Reporting Limits

Data sets may include multiple detection limits. This often results as lab procedures improve over time, or when data analyzed at several labs is combined into a common data set (Helsel and Hirsch, 1992). Data sets with multiple reporting limits can be set to the highest reporting limit, including censoring those values that were estimated or quantified below the highest reporting limit. This practice may result in a significant loss in information if detection limits have large changes over time, or if many quantified values must be censored (Oblinger-Childress and others, 1999). Spooner notes that data will indicate an artificial decreasing trend if detection limits decrease over time and multiple limits are used within the data set. Using the least sensitive reporting limit for all data will provide a more accurate trend estimate, but results in a loss of information from those samples analyzed with the more sensitive detection limit (Spooner 1991). The simple substitution method can be used with multiple reporting limits without losing information.

Summary

The Simple Substitution, Helsel's Robust Method, and the skipping all BDL observations method were applied to the parameters with data that contained BDL or zero readings from Site 075 Red Lake River Highlanding. The results of the summary statistics are listed in **Table 2.**

The computational spreadsheets and results are included in **Appendix B.** The Simple Substitution Method gave the most consistent and credible results. Helsel's Robust Method performed acceptably on some parameters but poorly on others. The outcome of Helsel's Method is dependent upon how well the data fit the assumed lognormal distribution.

Recommendations

Houston Engineering recommends calculating the estimated summary statistics using the simple substitution method with the BDL readings set first at zero and then recalculating the estimated summary statistics with the BDL and zero readings set at the MDL(s). This will clearly show the possible range of values, i.e. the "best" and "worst-case" scenarios, for parameters with censored data. This method is quickly completed in a spreadsheet, and more importantly is easy to comprehend. Those analyzing censored water quality data will have a clear understanding of the impact of substituting values for those listed below detection limits.

We also recommend that while using the Simple Substitution Method with censored data sets, the percentile statistics generated in Excel (or other statistical software package), should be checked manually for values of zero or the MDL. Computed percentile values (e.g. 25^{th} percentile) that are less than the detection limit should include the less than symbol (<) with the MDL (< MDL).

Bibliography

Australian Guidelines For Water Quality Monitoring And Reporting, National Water Quality Management Strategy No. 7, October 2000 http://www.ea.gov.au/water/quality/nwqms/pubs/mg-contents.pdf

Capel, P.D., Gilliom R.J., and Larson, S. J. Interpretation of Data On Low-Level Concentrations of Pesticides in Water. USGS National Water Quality Assessment, Pesticide National Synthesis Project, 1996

Helsel, D.R., and Hirsch, R.M. Statistical Methods in Water Resources. New York: Elsevier, 1992.

Liu S., Lu J.C., Kolpin, D.W., and Meeker, W.Q., Analysis of Environmental Data With Censored Observations, Environmental Science and Technology, v31, 1997

Newman M.C., Greene, K.D., and Dixon, P.M., UNCENSOR v4.0, Users Manual, Savannah River Ecology Laboratory, Aiken, South Carolina,

Oblinger-Childress C.J., Foreman, W.T., Connor, B.F. and Maloney, T.J., New Reporting Procedures Based on Long-Term Method Detection Levels and Some Considerations for Interpretations of Water-Quality Data Provided by the U.S. Geological Survey National Water Quality Laboratory, USGS Open File Report 99-193, 1999

Porter, S.P. and Ward, R.C. Estimating Central Tendency From Uncensored Trace Level Measurements, American Water Resources Association, Water Resources Bulletin, August 1991

Spooner, J. Censored Data Values: Description and Effect of Censoring on Statistical Trend Analyses (Part 2), North Carolina Cooperative Extension Service, NWQEP Notes No. 48, July 1991

United States Environmental Protection Agency. Evaluation of Dredged Material Proposed for Discharge in Waters of the U.S. – Inland Testing manual. Washington: GPO, EPA 823-B-98-004 February 1998 <http://www.epa.gov/ost/itm/ITM/appxd.htm>.

Appendix A

Simple Substitution and Helsel's Robust Method Procedures

Helsel's Robust Method

The following is a list of procedures used to calculate the estimated summary statistics using the Helsel's Robust Method. This procedure assumes that a standard spreadsheet program is used with available built-in functions. Appendix B shows the procedure using an Excel spreadsheet.

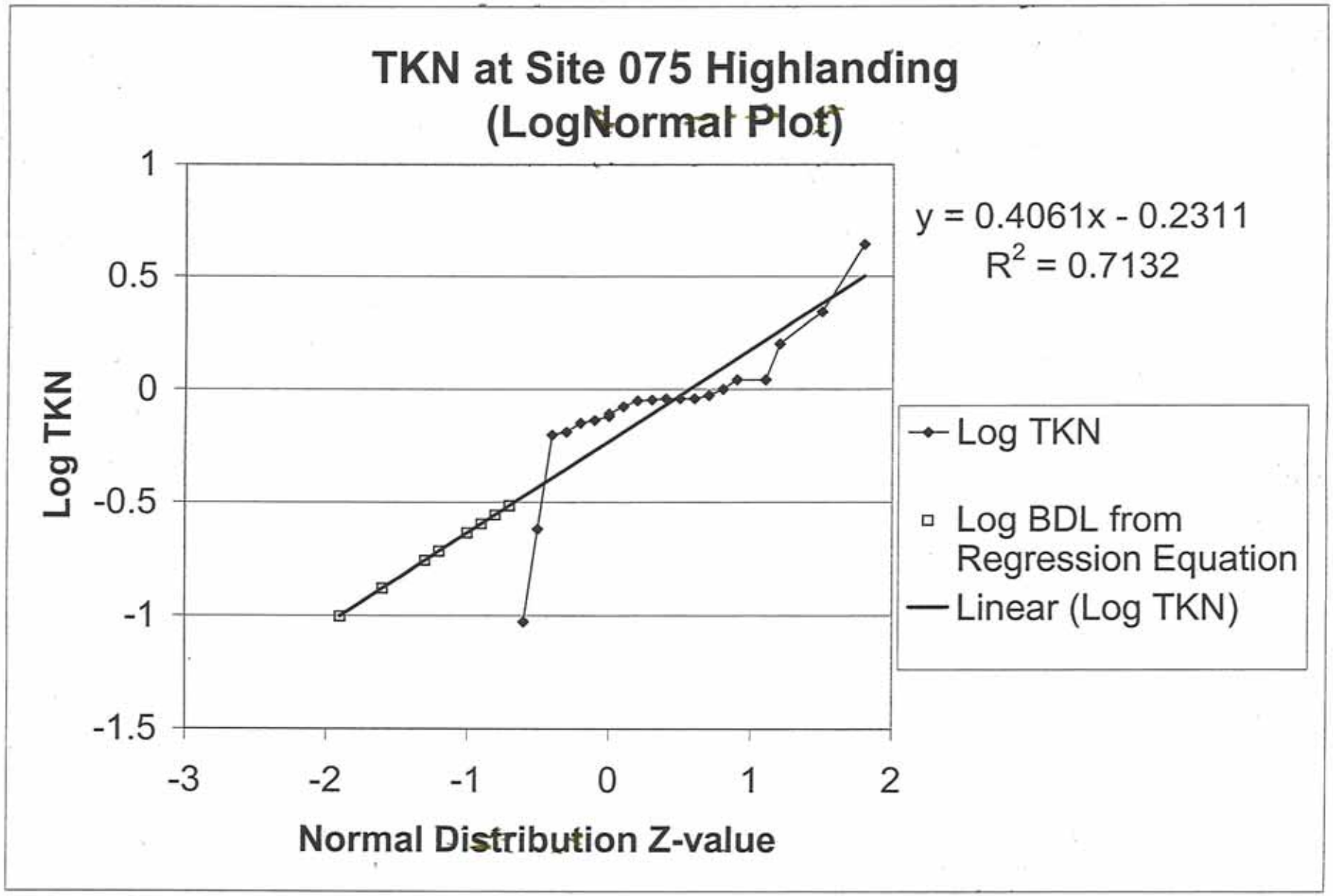
1. Sort data for the selected parameter from smallest to largest. The BDL and zero readings should be listed ahead of the smallest recorded value.
2. Assign a rank to all of the sorted data, including the BDL and zero readings, from 1 to n.
3. Convert the rank to the probability plotting position using the formula

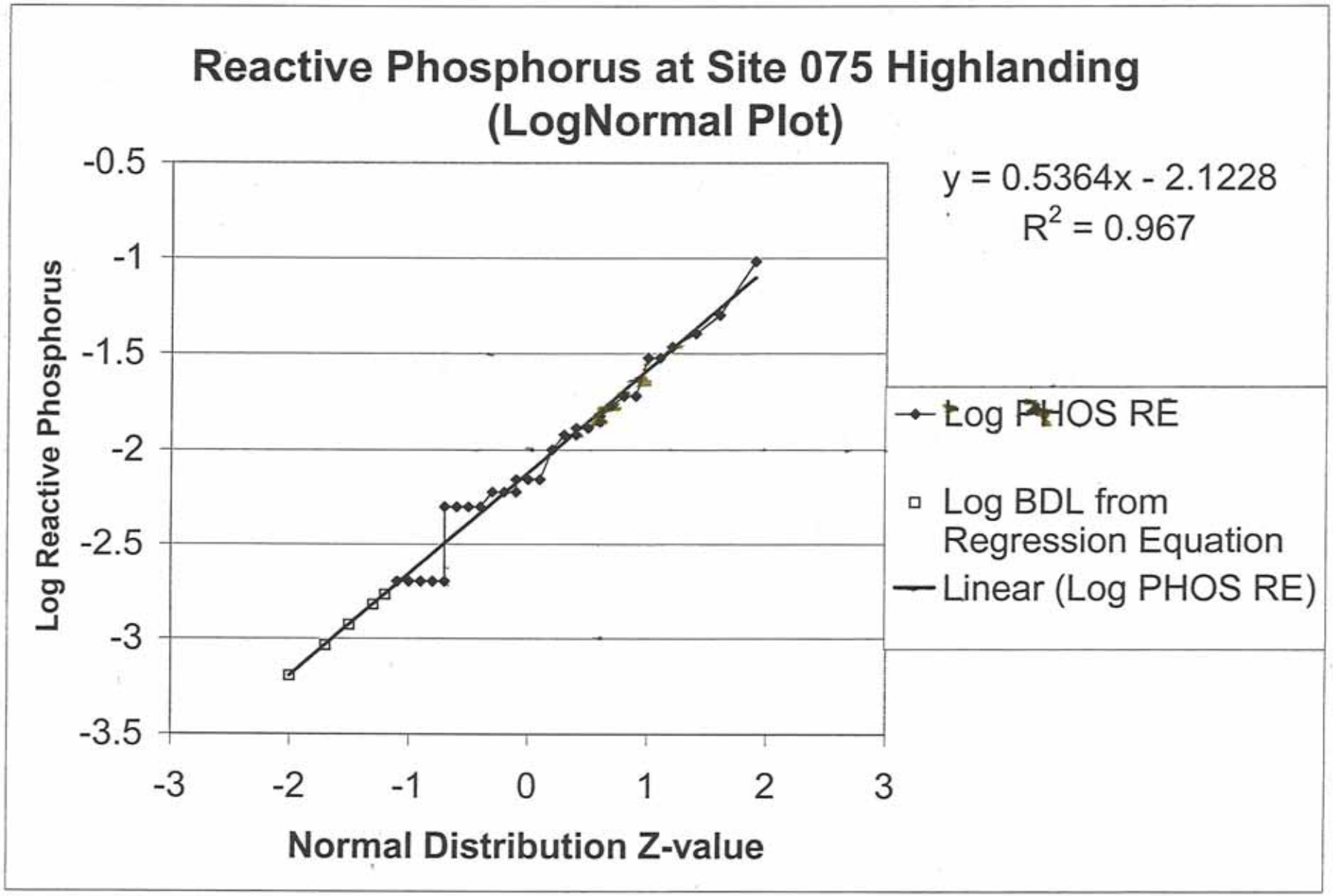
   $P = r/(n+1)$
   where;
   P = probability plotting position
   r= rank of value
   n = total number of values, including the BDL and zero readings.

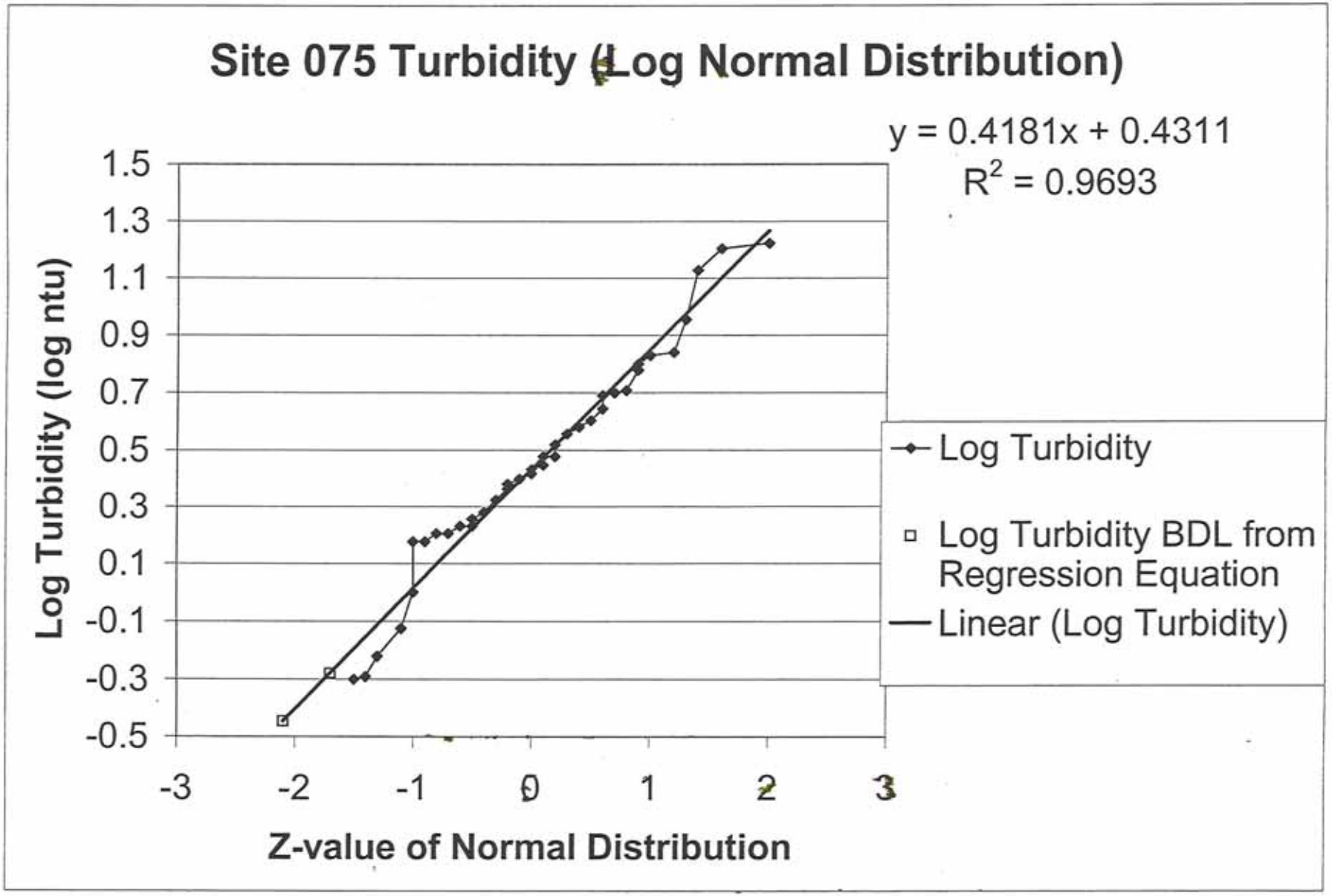
4. Look up the normal score (z value) from a Standard Normal Distribution Table and record next to each value, including BDL and zero readings.
5. Calculate the Log of non-BDL and zero readings.
6. Plot the Log of non-BDL and zero readings against their corresponding normal score (z-value).
7. Fit a regression line to the plotted data and record the regression line equation and R^2 value.
8. Calculate the Log value of the BDL and zero readings using the regression line equation and the corresponding normal score (z value) of each BDL and zero reading.

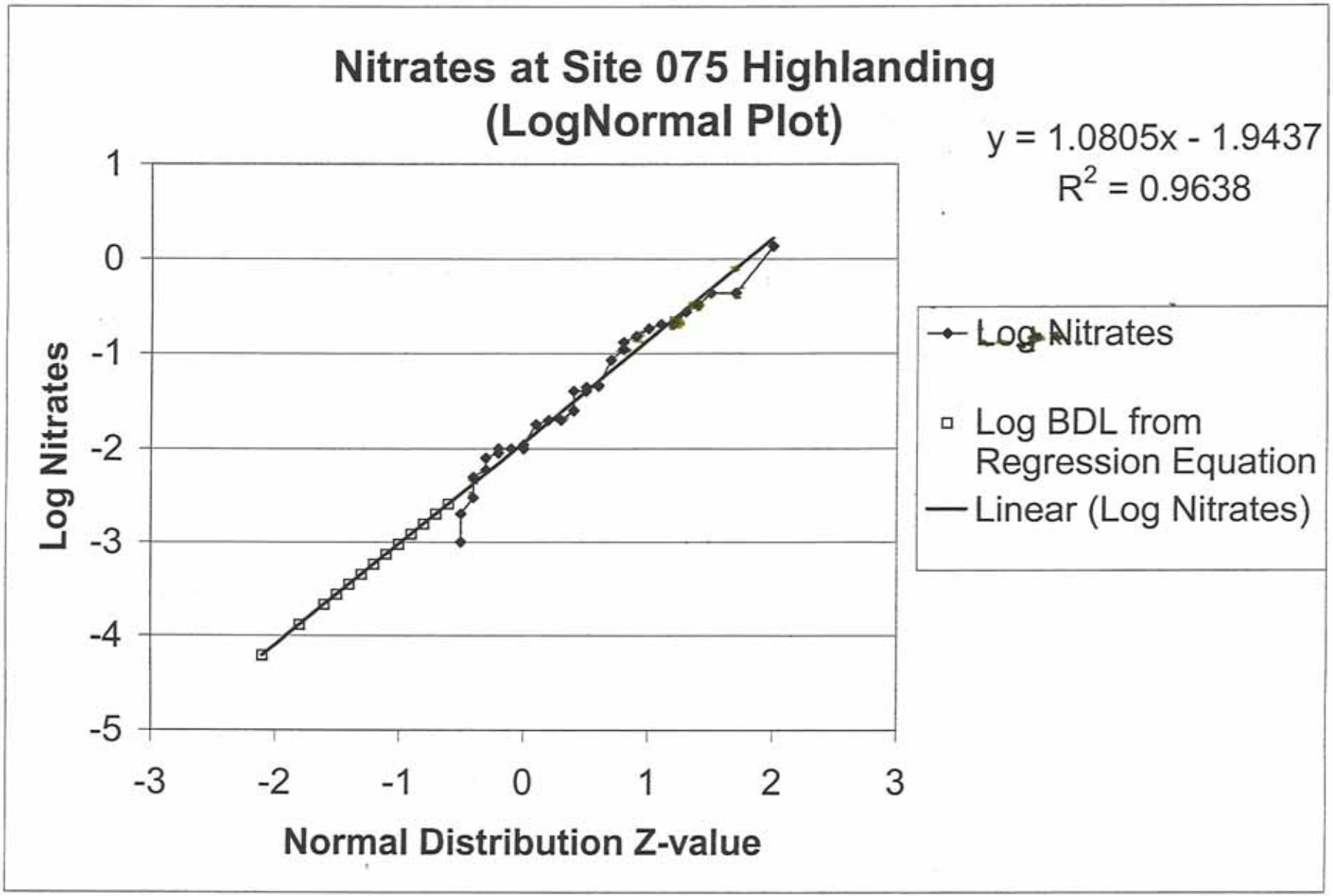
   $\text{Log BDL} = m*z + b$
   where;
   z = corresponding normal score (z value) of each BDL or zero reading
   m = slope from regression line of non-BDL and zero readings
   b = y-intercept from regression line of non-BDL and zero readings.

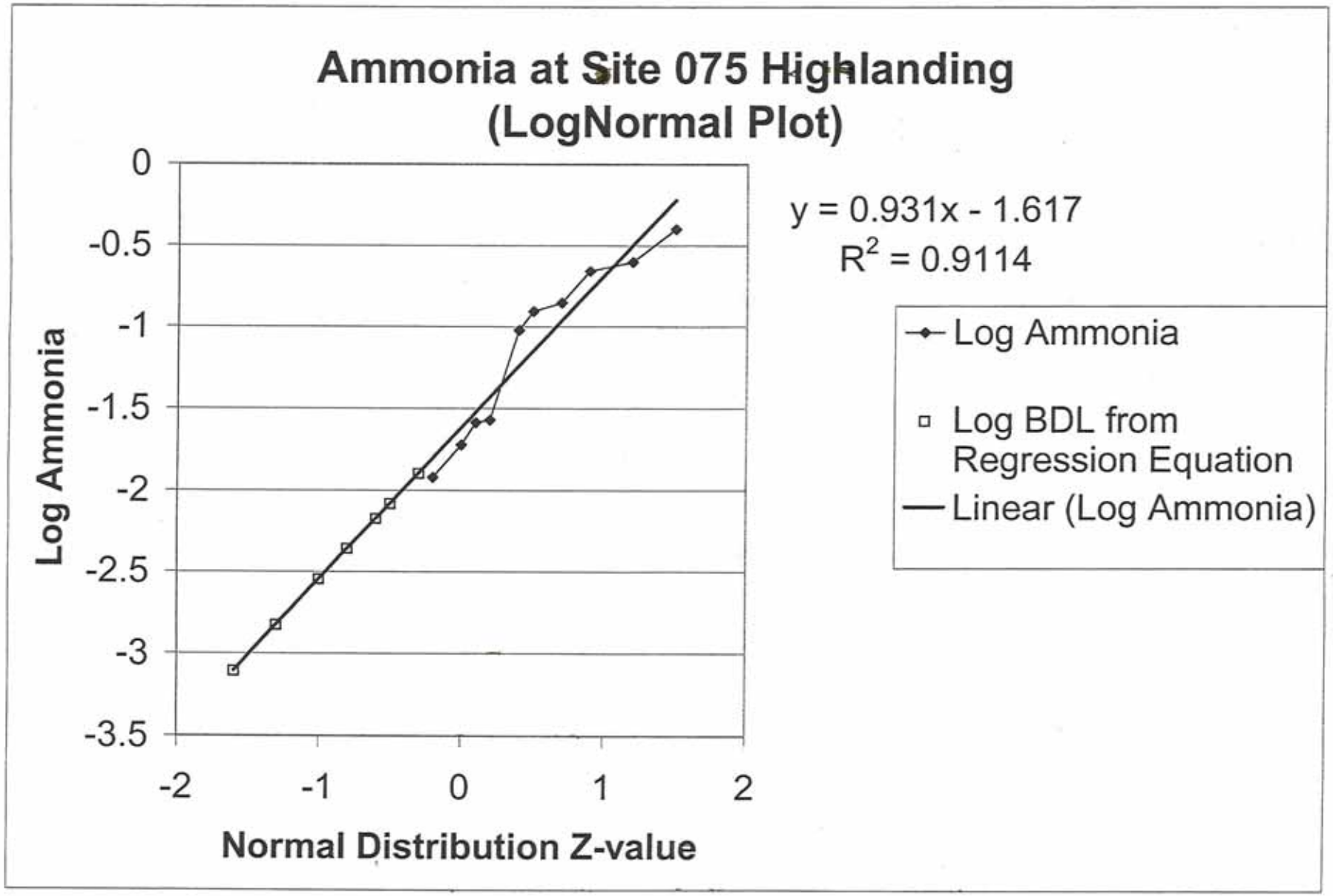
9. Transform the Log BDL value to original units by calculating the antilog.
10. Calculate the estimated summary statistics using the calculated values for the BDL or zero readings and all observed values.

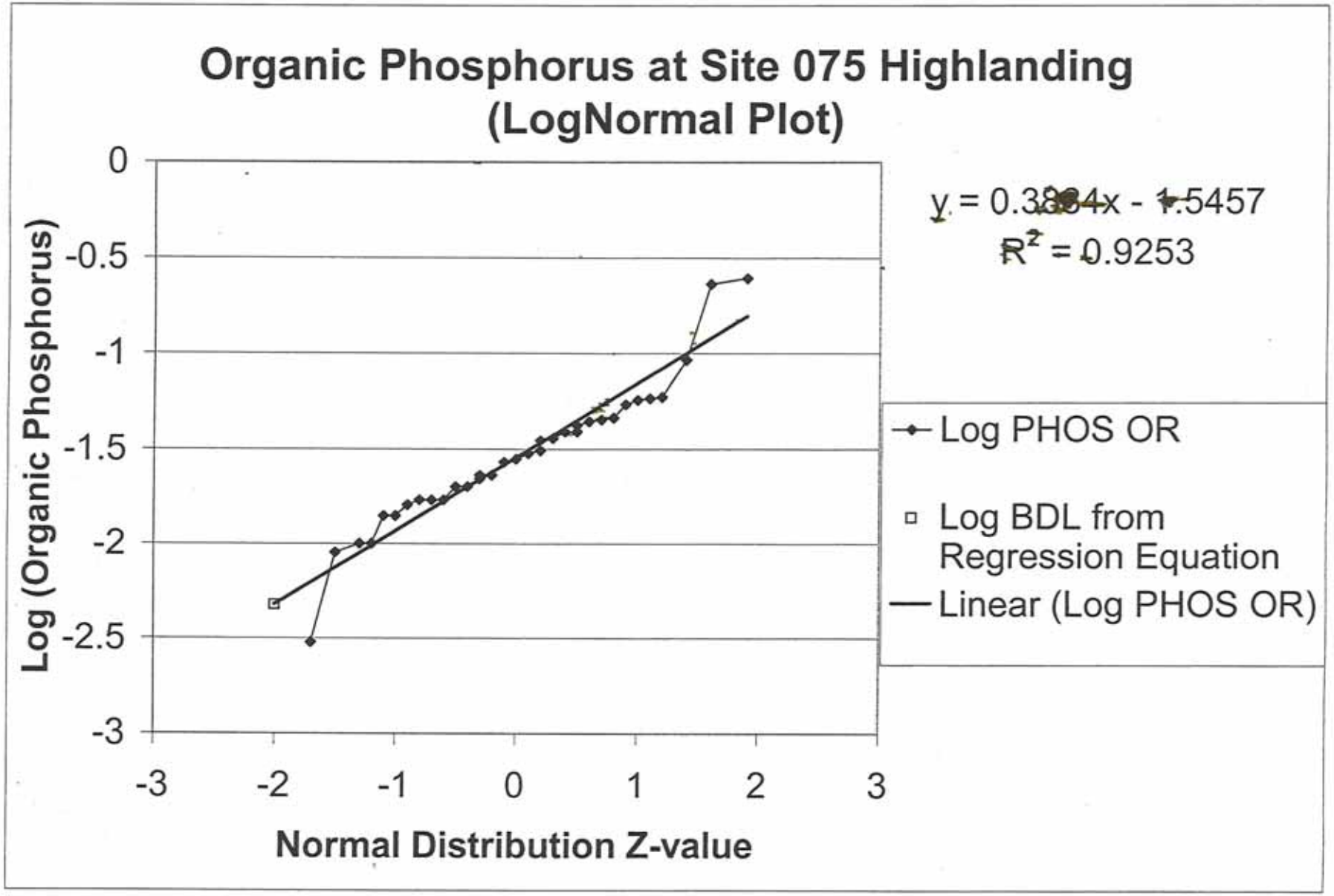

TKN at Site 075 Highlanding
(LogNormal Plot)
y = 0.4061x - 0.2311
R2 = 0.7132
Log TKN
Log BDL from Regression Equation
Linear (Log TKN)
Log TKN
Normal Distribution Z-value
1
0.5
0
-0.5
-1
-1.5
-3
-2
-1
0
1
2

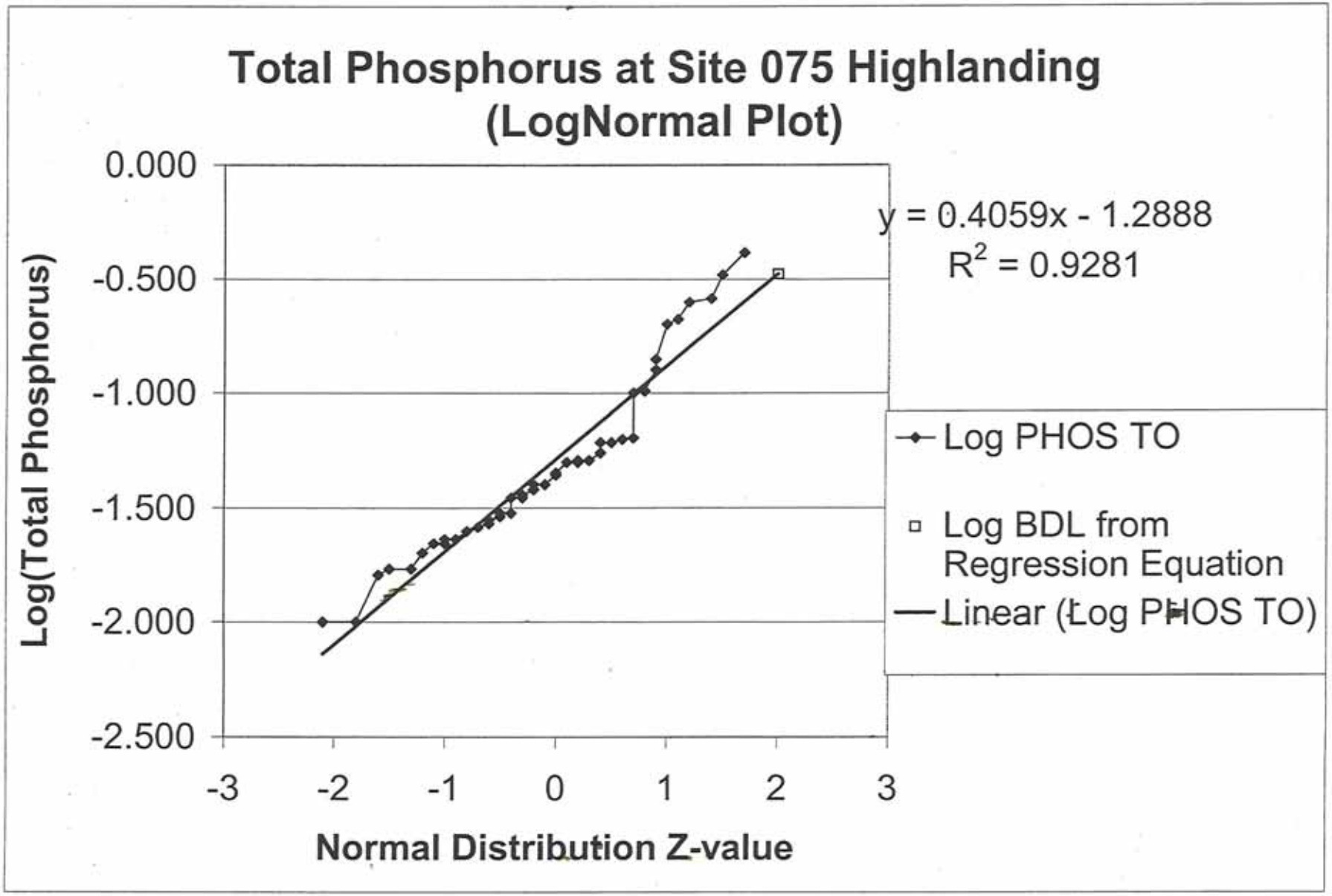

Reactive Phosphorus at Site 075 Highlanding
(LogNormal Plot)
y = 0.5364x - 2.1228
R2 = 0.967
Log Reactive Phosphorus
-0.5
-1
-1.5
-2
-2.5
-3
-3.5
-3
-2
-1
0
1
2
3
Normal Distribution Z-value
Log PHOS RE
Log BDL from Regression Equation
Linear (Log PHOS RE)

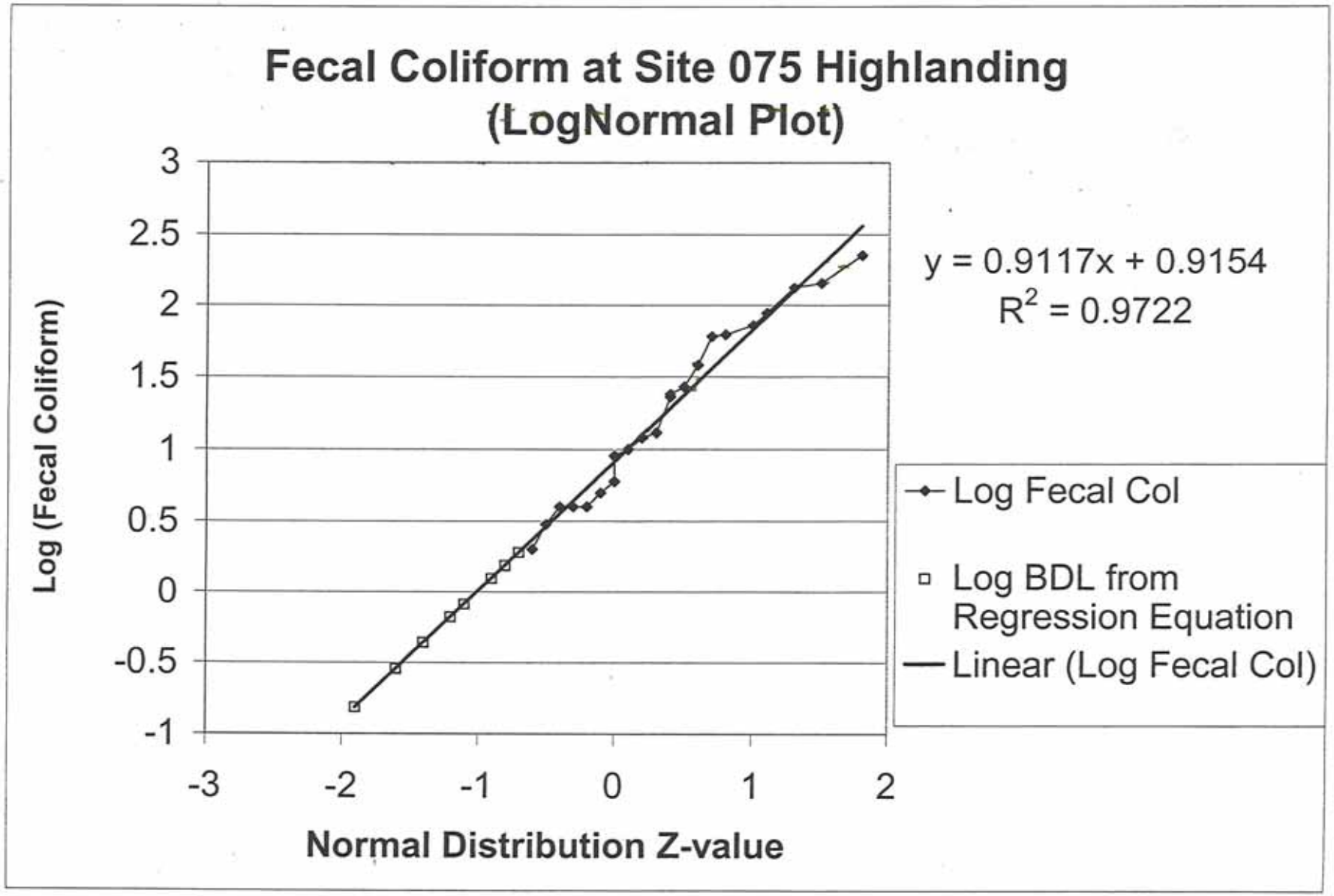

Site 075 Turbidity (Log Normal Distribution)
y = 0.4181x + 0.4311
R2 = 0.9693
Log Turbidity (log ntu)
1.5
1.3
1.1
0.9
0.7
0.5
0.3
0.1
-0.1
-0.3
-0.5
-3
-2
-1
0
1
2
3
Z-value of Normal Distribution
Log Turbidity
Log Turbidity BDL from Regression Equation
Linear (Log Turbidity)

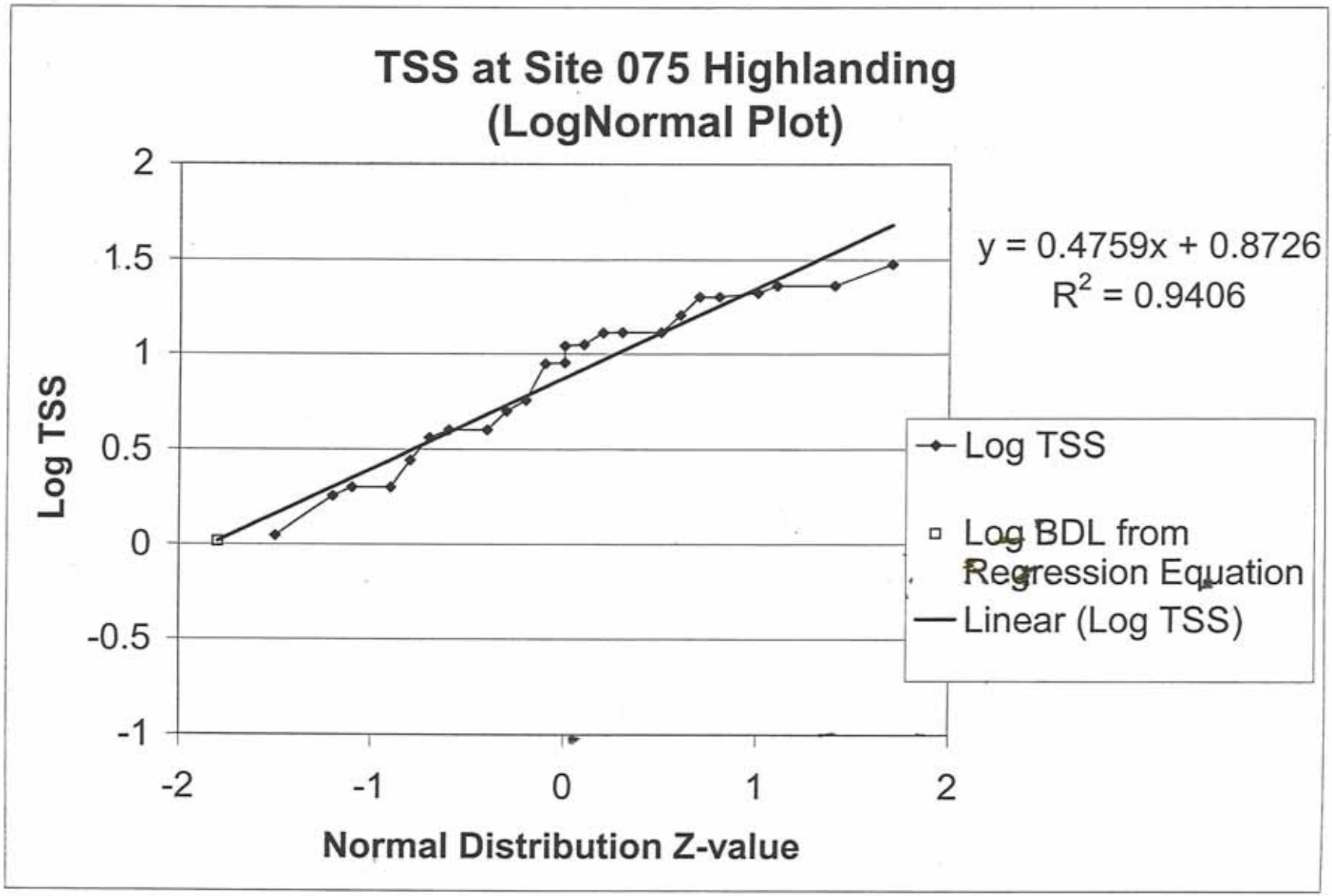

Nitrates at Site 075 Highlanding
(LogNormal Plot)
y = 1.0805x - 1.9437
R2 = 0.9638
Log Nitrates
Normal Distribution Z-value
1
0
-1
-2
-3
-4
-5
-3
-2
-1
0
1
2
3
Log Nitrates
Log BDL from Regression Equation
Linear (Log Nitrates)


Ammonia at Site 075 Highlanding
(LogNormal Plot)
0
-0.5
-1
-1.5
-2
-2.5
-3
-3.5
Log Ammonia
-2
-1
0
1
2
Normal Distribution Z-value
y = 0.931x - 1.617
R2 = 0.9114
Log Ammonia
Log BDL from Regression Equation
Linear (Log Ammonia)


Organic Phosphorus at Site 075 Highlanding
(LogNormal Plot)
Log (Organic Phosphorus)
0
-0.5
-1
-1.5
-2
-2.5
-3
-3
-2
-1
0
1
2
3
Normal Distribution Z-value
y = 0.3884x - 1.5457
R² = 0.9253
Log PHOS OR
Log BDL from Regression Equation
Linear (Log PHOS OR)


Total Phosphorus at Site 075 Highlanding
(LogNormal Plot)
Log(Total Phosphorus)
0.000
-0.500
-1.000
-1.500
-2.000
-2.500
-3
-2
-1
0
1
2
3
Normal Distribution Z-value
y = 0.4059x - 1.2888
R2 = 0.9281
Log PHOS TO
Log BDL from Regression Equation
Linear (Log PHOS TO)


Fecal Coliform at Site 075 Highlanding
(LogNormal Plot)
Log (Fecal Coliform)
3
2.5
2
1.5
1
0.5
0
-0.5
-1
-3
-2
-1
0
1
2
Normal Distribution Z-value
y = 0.9117x + 0.9154
R2 = 0.9722
Log Fecal Col
Log BDL from Regression Equation
Linear (Log Fecal Col)


TSS at Site 075 Highlanding
(LogNormal Plot)
Log TSS
Normal Distribution Z-value
2
1.5
1
0.5
0
-0.5
-1
-2
-1
0
1
2
y = 0.4759x + 0.8726
R2 = 0.9406
Log TSS
Log BDL from Regression Equation
Linear (Log TSS)