Identifying Pathogen Sources and Dealing with Squirrely Data Sets

Robert Pitt[1] and Jane Clary[2]

[1] Cudworth Professor of Urban Water Systems, Department of Civil, Construction, and Environmental Engineering, The University of Alabama, P.O. Box 870205, Tuscaloosa, Al 35487; PH (205) 348-2684; FAX (205) 348-0783; email: rpitt@eng.ua.edu
[2] Senior Water Resources Scientist, Wright Water Engineers, Inc., 2490 W. 26th Ave., Suite 100A, Denver, CO 80211, email: clary@wrightwater.com

ABSTRACT

Stormwater bacteria data are challenging to analyze for several reasons:

- highly variable observation levels (more so than for any other commonly monitored stormwater constituent)
- frequent right-censored data (very high levels can exceed the upper limit of the method being used)
- non-conservative behavior can cause unusual observations and storage problems during sample collection
- sensitive to environmental conditions (especially temperature)
- most analytical methods have a limited range in bacteria levels that can be quantified

In addition, traditional bacteria data reporting is usually expressed as a geometric mean in order to moderate the effects of periodic very high observed bacteria levels (most bacteria standards or criteria are expressed using geometric means, for example), which hinder statistical analyses of existing information. These issues therefore hinder the types of statistical analyses that can be conducted using most stormwater bacteria data. However, many of these problems can be overcome by careful sampling and proper selection of the analytical method, as noted below.

Sampling plans must consider the sampling locations appropriate for the project objectives. Bacteria sources vary greatly in urban areas, but high levels can be observed in many locations. Roof runoff water can have very high levels during summer and spring months if heavily covered by trees, due to the increased numbers of birds and squirrels that can reside above the roof surfaces. During colder months, many of these animals may migrate, hibernate, or become inactive, resulting in significantly decreased bacteria levels in the roof runoff. Soil bacteria levels can remain high in areas where urban wildlife or pets defecate, with runoff bacteria levels dependent on the amount of erosion occurring. Bacteria levels in runoff from paved areas can also be high in areas where pets are "walked." Residential areas and park pathways usually have larger bacteria levels than industrial areas, for example. Outfall stormwater samples are affected by the relative contributions of flows from the different areas which vary mostly according to rain characteristics. Therefore, an experimental design for bacteria sampling must consider these varying sources and seasonal conditions, requiring many samples over an extended period.

INTRODUCTION

Analytical methods used for bacteria analyses usually have a limited range of quantification. Initial samples are likely to have many "right-censored" observations, with "too numerous to count" or other over-range indications instead of actual values. Most stormwater managers are familiar with "left-censored" data where the observations are below the detection limits. In cases with significant numbers of right-censored data, more appropriate and more sensitive methods can be used for future analyses, and several data substitution methods can be applied to the non-detected values. However, data substitutions are generally not available for the excessively high observations. The best approach is to expand the range of detectable levels by using a wider range of sample dilutions in the bacteria tests. This is done differently for different methods, but will result in additional analyses, and therefore higher analytical costs (assuming that the lower limit is to be preserved, and not shifted higher). As an example, using the IDEXX methods and Quanti-Tray/2000 chambers, a standard analytical range of <1 to 2420 MPN/100 mL is available. Most analysts using this method do not dilute the sample, with many stormwater bacteria observations exceeding this range. However, this can be supplemented with a second tray (at twice the analytical cost) with a sample diluted 10 to 1 to extend the range to 24,200 MPN/100 mL, a level that is only periodically exceeded. Further dilutions can even be used, but great care needs to be made, with the recommended use of replicate trays to reduce problems associated with sample dilution and non-discrete bacteria groups, further increasing the analytical costs. In most cases, having the complete data with minimum uncertainties is worth the extra costs associated with the expanded analytical method, especially if the data is to be used to calibrate a stormwater model, to identify bacteria sources, or to quantify the bacterial removal benefits of a stormwater control practice. For compliance purposes, it may only be necessary to know that the permit limit was exceeded; however, the actual value is needed to quantify the geometric mean value required by many regulatory agencies.

The following discussion summarizes some of the issues and solutions that can be applied when statistically analyzing stormwater bacteria data. Much of this discussion is summarized from the stormwater sampling book by Burton and Pitt (2002), supplemented with various examples from past and on-going research on stormwater bacterial sources, transport, and fate being conducted by researchers at the University of Alabama.

ANALYTICAL METHODS FOR BACTERIA DATA ANALYSIS

The analysis of data requires at least three elements: 1) quality control/quality assurance of the reported data, 2) an evaluation of the sampling effort and methods (and associated expected errors), and finally, 3) the statistical analysis of the information. Quality control and quality assurance basically involves the identification and proper handling of questionable data. When reviewing previously collected data, it is common to find obvious errors that are associated with improper units or sampling locations. Other potential errors are more difficult to identify and correct. In some cases, the identification and rejection of "outliers" may result in the dismissal of rare data observations.

Selection of Statistical Procedures

Most of the objectives of receiving water studies can be examined through the use of relatively few statistical evaluation tools. The following briefly outlines some simple experimental

objectives and a selected number of statistical tests (and their data requirements) that can be used for data evaluation (Burton and Pitt 2001).

Basic Characterizations

One of the first tasks usually conducted with monitoring data is to prepare basic characterization statistics. For most of the examples in this memo, the follow data will be used (from Sumandeep Shergill's MS thesis at the University of Alabama, 2004). These were collected during a six month period in 2002 from the campus of the University of Alabama and from surrounding areas in Tuscaloosa, AL. The data are presented with three different options: the first set of columns reflect the native limited range of the analytical method (<1 to 2,419.2), the second set of columns includes the results of the additional ten-fold dilutions that were also evaluated, and the third set of columns has data substitutions or 0.5 in place of the <1 low detection limit. Also shown on the table are the statistical summaries for each set of data.

Roof Runoff E. coli Observations (MPN/100 mL)

	Without dilution		**With 10X dilution**		**With 10X dilution and substitution for <1**	
	birds	**no birds**	**birds**	**no birds**	**birds**	**no birds**
29-Aug-02	145.5	<1	145.5	<1	145.5	0.5
21-Sep-02	461.1	30.5	461.1	30.5	461.1	30.5
25-Sep-02	18.7	2	18.7	2	18.7	2
25-Sep-02	1,413.6	5.2	1,413.6	5.2	1,413.6	5.2
10-Oct-02	410.6	344.8	410.6	344.8	410.6	344.8
27-Oct-02	>2,419.2	161.6	17,329	161.6	17,329	161.6
5-Nov-02	>2,419.2	29.2	12,033	29.2	12,033	29.2
29-Jan-03	2	<1	2	<1	2	0.5
6-Feb-03	<1	>2,419.2	<1	5,298	0.5	5,298
Minimum	<1	<1	<1	<1	0.5	0.5
Maximum	>2,419	>2,419	17,329	5,298	17,329	5,298
Median	411	29	411	29	411	29
Analyses based on quantifiable data:						
Number of useful observations	6	6	8	7	9	9
Average (mean)	409	95.6	3,977	839	3,535	652
Geometric mean	106	28.3	363	59.8	175	20.7
Standard deviation	529	136	6,772	1,970	2,157	582
COV	1.3	1.4	1.7	2.3	1.8	2.7
Significantly different from normal distribution?	Yes (<0.001)	Yes (<0.001)	Yes (<0.001)	Yes (<0.001)	Yes (<0.001)	Yes (<0.001)

These are paired observations obtained from two different residential roofs; one roof had an extensive canopy of trees covering the building, while the other did not. The building with the canopy had a significant amount of observed bird and squirrel activity during the spring and summer months, while few were observed at the uncovered roof. As noted, three of the samples had no response during the test and therefore had <1 MPN/100 mL, while three were over-range. These samples were also analyzed using a ten-fold dilution to extend the upper range of the test

to 24,192 MPN/100 mL, resulting in numeric results for the over-range values if no additional dilution was used.

Because of the typically wide range of bacteria values typically observed during a monitoring period, managers are uncomfortable with the extra effects that the very large values have on resulting calculated average values. The median values are therefore commonly used when describing this type of data as the extremes and uncertain values have little effects on its value (unless the uncertain events number more than 50% of the data set!). Unfortunately, medians are not very useful when comparing to standards written using geometric mean values, or when calculating loads. If a data set was symmetrical (not necessarily normally distributed), then the medians and the means would have the same value, but as the distribution skewness increases, the means and medians can vary greatly, as in this example. Methods to moderate the effects of these very large values are typically used for reporting purposes. The median and geometric mean values are shown in this example to be significantly smaller than the averaged values, with the geometric means being 20+ times smaller than the averages. If geometric means are used in "mass" calculations, it is obvious that the results would cause large errors. Similarly, if geometric mean summaries of past observations are all that are available, the statistical tests that can be applied are limited.

This table also shows the results of the Shapiro-Wilk test (SigmaPlot version 11) used to test normality of the data. In this example, the test failed for all of the data sets. According to SigmaPlot, "a test that fails indicates that the data varies significantly from the pattern expected if the data was drawn from a population with a normal distribution." Therefore, statistical tests that require normally distributed data should not be used with these data. This result is common for most stormwater observations (Maestre, *et al*. 2005, and many others), especially for bacteria data. The relatively large coefficient of variation (the standard deviation divided by the mean) values (1.8 and 2.7 for the final data set with substitutions) also indicate likely non-normal behavior.

Comparison Tests

Probably the most common experimental objective is to compare data collected from different locations, or seasons. Comparison of data with reference sites, of influent with effluent, of upstream to downstream locations, for different seasons of sample collection, of different methods of sample collection, can all be made with comparison tests. If only two groups are to be compared (above/below; in/out; test/reference), then the two group tests can be effectively used, such as the simple Student's *t*-test or nonparametric equivalent. If the data are collected in "pairs," such as for concurrent influent and effluent samples, or for concurrent above and below samples, then the more powerful and preferred paired tests can be used. If the samples cannot be collected to represent similar conditions (such as large physical separations exist in sampling location, or different time frames), then the independent tests must be used.

If multiple groupings are used, such as from numerous locations along a stream, but with several observations from each location; or at one location; or from one location, but for each season, then a one-way ANOVA is needed. If one has seasonal data from each of the several stream locations for multiple seasons, then a two-way ANOVA test can be used to investigate the effects of location, season, and the interaction of location and season together. Three-way ANOVA tests

can be used to investigate another dimension of the data (such as contrasting sampling methods or weather for the different seasons at each of the sampling locations), but that would obviously require substantially more data to represent each condition.

There are various data characteristics that influence which specific statistical test can be used for comparison evaluations. The parametric tests require the data to be normally distributed and that the different data groupings have the same variance, or standard deviation (checked with probability plots and appropriate test statistics for normality, such as the Shapiro-Wilk, the Kolmogorov-Smirnov one-sample test, the chi-square goodness of fit test, or the Lilliefors test). If the data do not meet the requirements for the parametric tests, the data may be transformed to better meet the test conditions (such as taking the $\log_{10}$ of each observation and conducting the normality test on the transformed values). The non-parametric tests are less restrictive, but are not free of certain requirements. Even though the parametric tests have more statistical power than the associated non-parametric tests, they lose any advantage if inappropriately applied. If uncertain, then non-parametric tests should be used.

Many statistical analysis tools may not be applicable to stormwater bacteria data. The large data variations hinder sufficient data to verify many of the required data characteristics (generally restricting available methods to some of the non-parametric procedures), and there are typically many missing data in the observed data sets (especially problematic are the over-range observations). In addition, historical bacteria data is usually reported as geometric means that do not reflect the flow-weighted values that are needed for load analyses. Therefore, the most obvious methods that can be used to evaluate stormwater bacteria data may be restricted to the following:

Basic Data Summaries:
- central tendency measures appropriate for the project objectives (geometric means for compliance with water quality standards; means for calculating flow-weighted discharges and TMDL compliance and for model calibrations)
- measures of variation (tests for data normality, standard deviations, COVs, and limitations due to sample numbers)

Exploratory Data Analyses:
- probability plots (with truncated distributions reflecting missing data)
- box and whisker plots (possibly only using reported values)
- trend plots showing bacteria level changes with time
- line plots contrasting paired data sets

Comparison Tests:
- Sign test for paired observations
- Wilcoxon signed-rank test for paired observations with few missing data
- Mann-Whitney rank sum test for independent observations in two sample sets
- Kurskal-Wallis ANOVA on ranks to detect significant subsets of the data

Correlation Tests:
- Spearman Rank order test for simple correlations of non-normal data

- Cluster and principal component advanced analyses to identify complex relationships of data; requires substantial information and few missing data

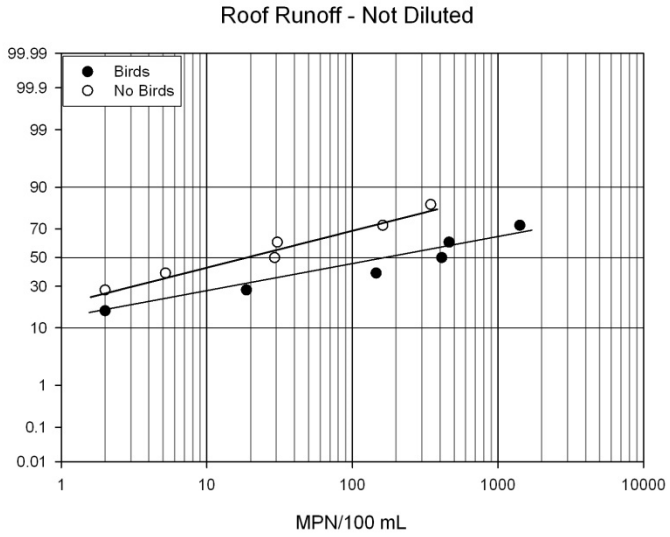
Trend Analyses and Model Building:
- Graphical analyses, usually based on time series of observations over long periods of time
- Nonparametric trend tests, depending on available data and their characteristics
- Factorial analyses to identify significant factors affecting observations, if sufficient data are available

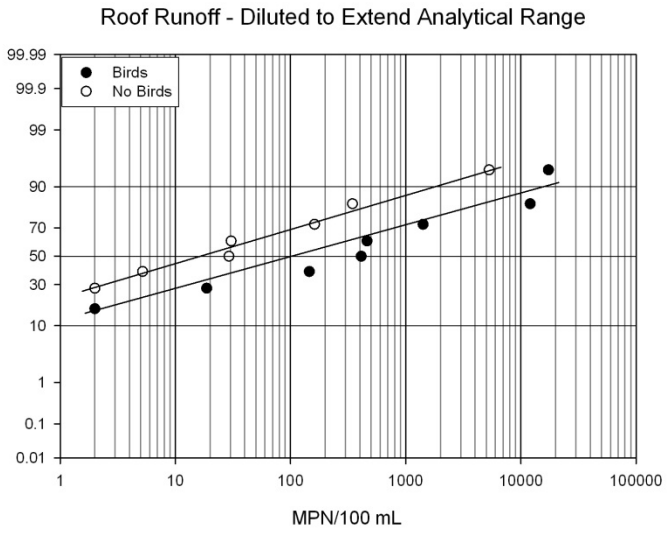
A few of the more basic options are described in the following paragraphs:

Exploratory Data Analyses
Exploratory data analyses (EDA) is an important tool to quickly review available data before a specific data collection effort is initiated. It is also an important first step in summarizing collected data to supplement the specific data analyses associated with the selected experimental designs. A summary of the data's variation is most important and can be presented using several simple graphical tools. Another important reference for basic analyses is *Exploratory Data Analysis* (Tukey 1977) which is the classic book on this subject and presents many simple ways to examine data to find patterns and relationships. Besides plotting of the data, exploratory data analyses should always include corresponding statistical test results, if available.

Probability Plots
The most basic and important exploratory data analysis method is to prepare a probability plot of the available data. The plots indicate the possible range of the values expected, their likely probability distribution type, and the data variation. The values and corresponding probability positions are plotted using normal-probability scales. These have a y-axis whose values are spread out for the extreme small and large probability values. When plotted using these scales, the values form a straight line if they are normally distributed (Gaussian). If the points do not form an acceptably straight line, they can then be plotted using a log scale for the observed values to indicate if they are log-normally distributed.

The following two figures are probability plots of the above presented bacteria data for the roof runoff from the two sampling locations, with the bird and on bird data shown on the same plots for comparison. These are truncated plots not showing information for the non-detected left-censored or right-censored observations. The second plot is for the data set that includes the diluted samples with an extended range and no right-censored observations. These are accurate plots in that they do not include any assumed or substituted data, and reflect the actual observations. They are both log-normal plots and indicate reasonably straight line relationships, indicating that data transformations would possibly be advantageous and allow extended parametric statistical analyses. The lower limits of these plots are truncated and do not show the <1 MPN/100 mL non-detected values that were at about 11 and 22% of the datasets. The upper limits are truncated at the right-censored values. For the first plot not having the extended range associated with the extra sample dilutions, the data are truncated at about 70 and 80%. For the extended range plot, the upper limits are only truncated at the maximum values obtained. The plotted median values are seen to shift between the two sets of data, especially for the site having birds.





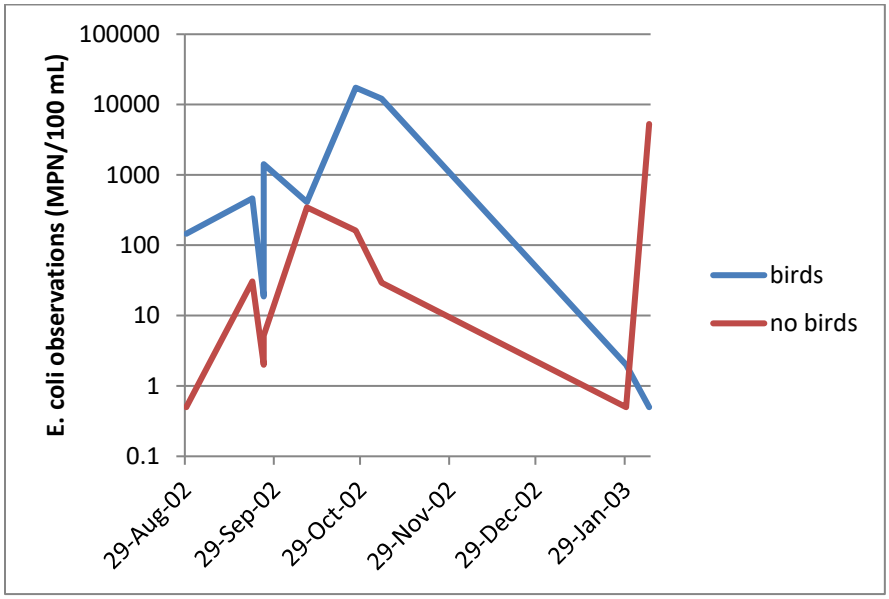
Generally, water quality observations do not form a straight line on normal probability plots, but do (at least from about the 10 to 90 percentile points) on log-normal probability plots, as shown above. This indicates that the samples generally have a log-normal distribution and many parametric statistical tests can probably be used, but only after the data is log-transformed. These plots indicate the central tendency (median) of the data, along with their possible distribution type and variance (the steeper the plot, the smaller the COV and the flatter the slope of the plot, the larger the COV for the data).

Probability plots should be supplemented with standard statistical tests that determine if the data are normally distributed. These tests, include the Kolmogorov-Smirnov one-sample test, the chi-square goodness of fit test, and the Lilliefors variation of the Kolmogorov-Smironov test. They

basically are paired tests comparing data points from the best-fitted normal curve to the observed data. The statistical tests may be visualized by imagining the best-fitted normal curve data and the observed data plotted on normal probability graphs. If the observed data crosses the fitted curve data numerous times, it is much likely to be normally distributed than if it only crossed the fitted curve a few times. As indicated previously, these roof runoff bacteria data are not normally distributed for statistical test purposes, but may be log-normally distributed.

Time-Series Plots

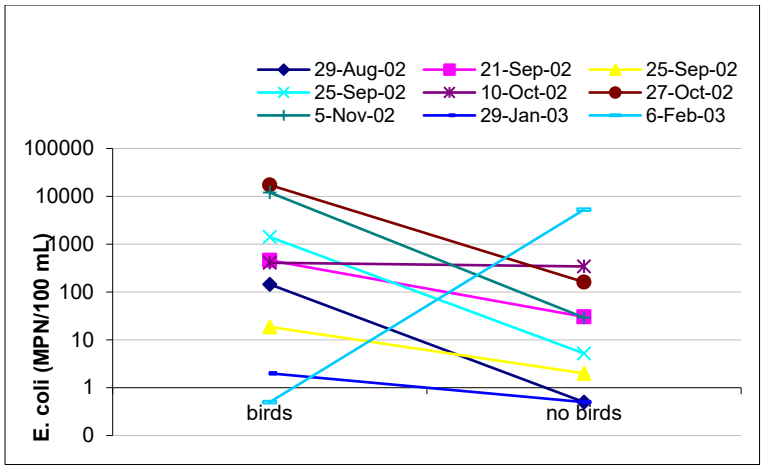
Berthouex and Brown (1994) point out that since the best way to display data is with a plot, it makes little sense to present the data only in a table. A basic time series plot indicates any obvious data trends with time. The following figure shows the E. coli observations for the roof runoff from the building having birds vs. no birds above the roof. It is obvious that the presence of the birds (in the absence of any other factor) affected the observed values during much of the study period. However, these effects notably decreased in the late fall.



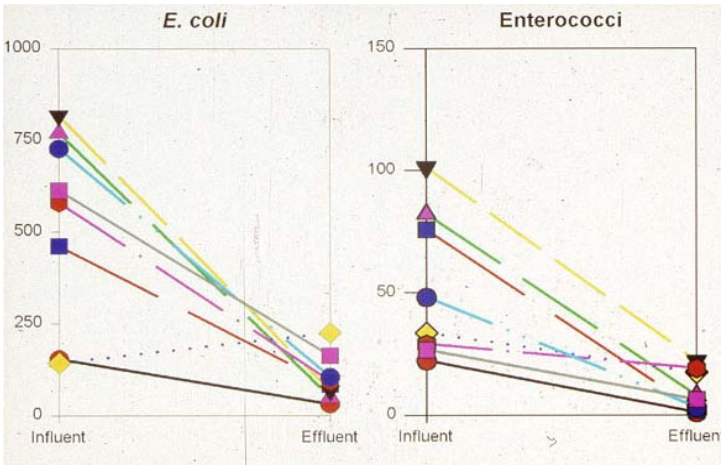
***E. coli* observations in roof runoff.**

Line Plots

Line plots are a type of scatterplot that contrasts paired observations. The following figure is a line plot for the roof runoff data contrasting the roofs affected by birds and those not affected by birds. This plot indicates that almost all of the data pairs were higher for the roof having birds than for the roof without birds, with one exception.



The following line plots illustrate the removal of stormwater bacteria during filter tests (Clark 1996). The common downward trending lines indicate consistent and significant removals of the bacteria, along with a much reduced range of discharged bacterial levels compared to influent bacteria levels.

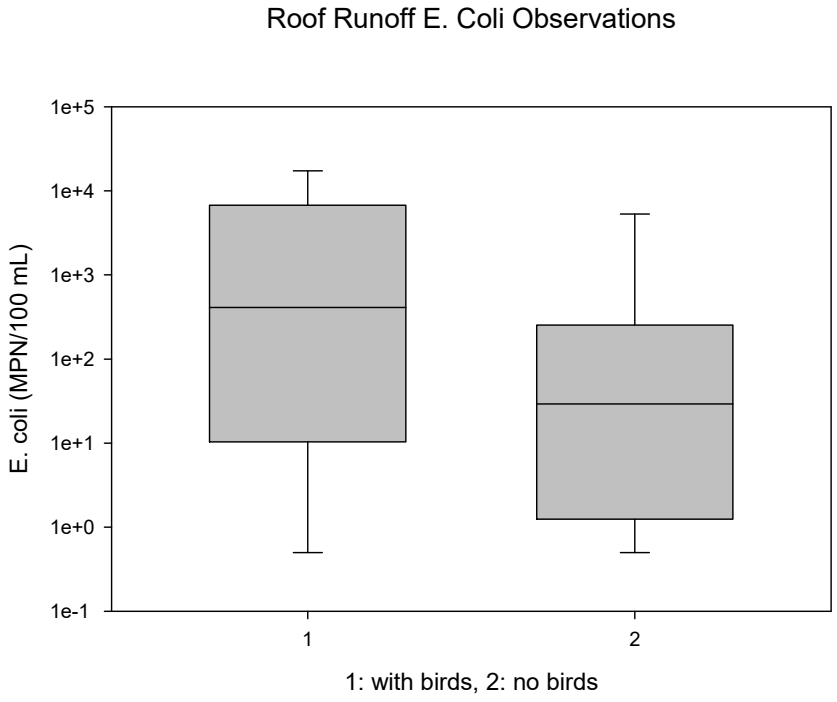


Grouped Box and Whisker Plots

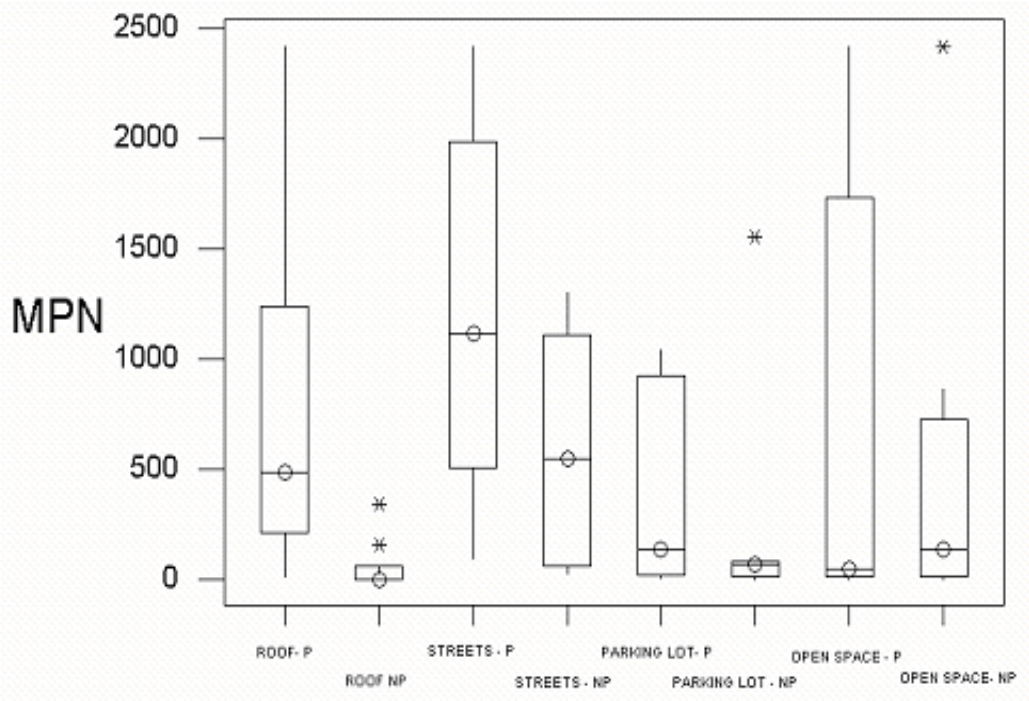
Another primary exploratory data analysis tool, especially when differences between sample groups are of interest, is the use of grouped box and whisker plots. Examples of their use include examining different sampling locations (such as above and below a discharge), influent and effluent of a treatment process, different seasons, etc.

The following grouped box and whisker plot (Shergill 2004) contrasts the roof runoff samples collected from buildings affected by birds vs. those without birds. These are all of the data, without regard to season and indicate large variations and overlapping data sets. For small data sets, the median line in one box (the "central" line in the box) needs to be above or below the corresponding 25th or 75th percentile box ends in the adjacent box for a statistically significant difference. The whiskers indicate the 5th and 95th percentile observations of the data sets. In this

example, the “with bird” median value is barely larger than the “no bird” 75th percentile, but he “no bird” median is not lower than the “with birds” 25th percentile value. Because of the large variations in these data, the level of confidence in the differences may be marginal at best. With larger data sets, the amount of allowable overlap for significantly different data sets can be larger (large variations require larger sample numbers).



In contrast, the following plot only examines the warm weather source area *E. coli* values from areas likely affected by urban pets and wildlife compared to other areas. The differences for the roof data are most obvious, but the other areas (streets, parking lots, and open space) also show large decreases (although the large variations and greater overlapping of the boxes indicate that they may not have differences that are statistically significant.



SUMMARY OF STORMWATER BACTERIA STATISTICAL ANALYSES

Stormwater bacteria data are characterized by large variations and missing data. This can be overcome by carefully designing the monitoring program to focus on the most critical elements to monitor so sufficient data can be obtained. In addition, appropriate laboratory methods need to be used to enable the wide range of bacteria levels to be quantified, such as expanding the dilution series.

Data summaries and statistical analyses, as always, must be chosen to correspond to the objectives of the research effort. Geometric mean values are commonly used for bacteria standards, but they are misleading when applied to statistical analyses and model building. Flow-weighted average values are most suitable for these analyses. In most cases, nonparametric statistical analyses are needed for analyzing stormwater bacteria data. There are many tools that can be used, but data requirements must be verified before their use, especially related to right-censored values. Also, because of the large variability in the data, it may be most suitable to accept somewhat less demanding data quality objectives, especially for initial exploratory investigations.

REFERENCES

Berthouex, P.M. and L.C. Brown. Statistics for Environmental Engineers. Lewis Publishers, Boca Raton, FL, 1994.

Burton, G.A. Jr., and R. Pitt. Stormwater Effects Handbook: A Tool Box for Watershed Managers, Scientists, and Engineers. CRC Press, Inc., Boca Raton, FL. August 2001. 911 pgs.

Clark, Shirley. *Evaluation of Filtration Media for Stormwater Runoff Treatment*. MSCE thesis to the Department of Civil and Environmental Engineering, The University of Alabama at Birmingham, 1996.

Gilbert, R. 0., Statistical Methods for Environmental Pollution Monitoring. New York, NY: Van Nostrand Reinhold, 1987.

Lehman, E.L. and H.J.M. D'Abrera. Nonparametrics: Statistical Methods Based on Ranks. Holden-Day and McGraw-Hill. 1975.

Maestre, A., R. Pitt, S.R. Durrans, and S. Chakraborti. "Stormwater quality descriptions using the three parameter lognormal distribution." *Effective Modeling of Urban Water Systems,* Monograph 13. (edited by W. James, K.N. Irvine, E.A. McBean, and R.E. Pitt). CHI. Guelph, Ontario, pp. 247 – 274. 2005.

Pitt, R., and J. McLean. Toronto Area Watershed Management Strategy Study: Humber River Pilot Watershed Project. Ontario Ministry of the Environment, Toronto, Ontario. 486 pgs. 1986.

Shergill, S. *Quantification of Escherichia Coli and Enterococci Levels Wet Weather and Dry Weather Flows*. MSCE thesis prepared for the Department of Civil and Environmental Engineering, The University of Alabama. 2004.

Sokal, R.R and F.J. Rohlf. Biometry: The Principles and Practice of Statistics in Biological Research. W.H. Freeman and Co. New York. 1969.

Tukey, John W. Exploratory Data Analysis. Addison-Wesley Publishing Co. 1977.