A statistical appraisal of disproportional versus proportional microbial source tracking libraries

Brian J. Robinson, Kerry J. Ritter and R. D. Ellender

ABSTRACT

Library-based microbial source tracking (MST) can assist in reducing or eliminating fecal pollution in waters by predicting sources of fecal-associated bacteria. Library-based MST relies on an assembly of genetic or phenotypic "fingerprints" from pollution-indicative bacteria cultivated from known sources to compare with and identify fingerprints of unknown origin. The success of the library-based approach depends on how well each source candidate is represented in the library and which statistical algorithm or matching criterion is used to match unknowns. Because known source libraries are often built based on convenience or cost, some library sources may contain more representation than others. Depending on the statistical algorithm or matching criteria, predictions may become severely biased toward classifying unknowns into the library's dominant source category. We examined prediction bias for four of the most commonly used statistical matching algorithms in library-based MST when applied to disproportionately-represented known source libraries; maximum similarity (MS), average similarity (AS), discriminant analyses (DA), and k-means nearest neighbor (k-NN). MS was particularly sensitive to disproportionate source representation. AS and DA were more robust. k-NN provided a compromise between correct prediction and sensitivity to disproportional libraries including increased matching success and stability that should be considered when matching to disproportionally-represented libraries.

Key words | disproportional, library, microbial source tracking, proportional, statistics

Brian J. Robinson
National Oceanic and Atmospheric Administration,
219 Fort Johnson Road, Charleston SC 29412-9110,
USA
Tel: +1 843 762-8572
Fax: +1 843 762-8700
E-mail: *brianjrobinson1979@yahoo.com*

Kerry J. Ritter
Southern California Coastal Water Research Project,
7171 Fenwick Lane, Westminster CA 92683,
USA

R. D. Ellender (corresponding author)
The University of Southern Mississippi,
Department of Biological Sciences,
118 College Drive #5018,
Hattiesburg MS 39406-0001,
USA

INTRODUCTION

Predicting sources of fecal contamination is important for managing water bodies and protecting humans against waterborne disease. By predicting the source(s) of fecal-associated bacteria, Microbial Source Tracking (MST) allows scientists and regulators to prioritize and more effectively respond to health and environmental hazards associated with fecal-contaminated waters (Scott *et al.* 2002; Simpson *et al.* 2002; Stewart *et al.* 2003). Some of the commonly used MST methods are library-based and rely on the assembly of genetic or phenotypic "fingerprints" from pollution-indicative bacteria cultivated from known sources of fecal contamination (Scott *et al.* 2002; Simpson *et al.* 2002). Scientists predict unknown sources of fecal contamination using computer-based statistical analysis to match unknown source fingerprints to those from the known-source library (Wiggins 1996; Hagedorn *et al.* 1999; Dombek *et al.* 2000; Harwood *et al.* 2000; Bower 2001; Whitlock *et al.* 2002). The success of the library-based approach depends on the distribution of fingerprint patterns among source candidates, how well each source candidate is represented in the library, and which statistical algorithm or matching criterion is used to match unknowns (Ritter *et al.* 2003).

Construction of known-source libraries is often limited by the availability of known-source samples and our ability to collect and process those samples (Wiggins *et al.* 2003; Robinson 2004). As a result, libraries may contain disproportional representation of isolates among the source candidates. For example, collecting large numbers of

doi: 10.2166/wh.2007.044

samples from a wastewater treatment plant may be relatively easy while collecting an equivalent number of individual dog samples may require much more effort and may not be feasible. The concern is that libraries that are heavily "loaded" toward a particular source may bias predictions toward the dominant library source.

The potential bias resulting from disproportional libraries may be particularly problematic depending on which statistical matching algorithm is used to match unknown source isolates. Library-based methods employ a variety of statistical methods to match fingerprints of unknown origin to the known-source library (Wiggins 1996; Hagedorn *et al.* 1999; Dombek *et al.* 2000; Harwood *et al.* 2000; Bower 2001; Whitlock *et al.* 2002; Carson *et al.* 2003; Ritter *et al.* 2003; Wiggins *et al.* 2003; Hassan *et al.* 2005). Since each method relies on a different strategy for matching, some algorithms may be more sensitive than others to disproportional source representation in the library. Maximum similarity (MS), commonly used in MST data analysis, is a statistical matching algorithm which classifies an unknown into the source group to which its most similar known member belongs (Applied Maths Inc. 2004). Consequently, using MS may result in increased predictions to the dominant source category simply because there are more "opportunities" to match to the dominant source.

Average similarity (AS) and discriminant analysis (DA) provide alternative matching strategies to MS where isolates are matched to known sources based on proximities to the center of each source group, rather than on the proximity to a single library isolate. AS assigns unknown source fingerprints to the source group based on the average similarity of that fingerprint to all fingerprints within each known source group in the library (Applied Maths Inc. 2004). DA classifies unknowns into source groups based on a "rule" developed from a calibration data set (e.g. library) (SAS 2004). This "rule" is based on the distribution of distances between library fingerprints and the centroid of each source group in order to estimate the relative likelihood of belonging to each source group (Johnson 1998). In the case of both AS and DA, disproportionate libraries may create unstable estimates for the center of each group by allowing a greater number of outliers which may skew the estimated probabilities leading to incorrect prediction.

A study was performed in 2003 on a coastal watershed in Mississippi that consistently displayed elevated levels of fecal coliform bacteria in the water, forcing the area to be closed by the state for recreational use (Robinson 2004). Three potential sources of fecal contamination source (dog, gull, sewer) were identified in this urban, mostly residential, watershed. Source samples were collected and processed, based on availability, for enterococci by rep-PCR using BOX sequence (5′-CTA CGG CAA GGC GAC GCT GAC G-3′) primers (BOX-PCR). Although an attempt was made to build a library from equal numbers of isolates within each source, the variable rates of isolation, confirmation, and the selection of unique fingerprint patterns (i.e. clones) by removing identical fingerprints from the same sample, led to a disproportional representation among source candidates. The resulting library contained approximately three times as many sewer isolates as dog and gull isolates combined. Analysis of the data raised concerns that having a greater number of sewer representatives in the library may bias identification towards the sewer source.

This paper examines the use of library-based rep-PCR data and three common statistical methods (MS, AS, DA) and one alternative statistical method (k-NN) in the presence of disproportionate source representation. The results are based on simulation studies using the enterococcal fingerprints from the study described above, where we estimate the probabilities of correct and incorrect prediction for identifying three sources (sewer, gull, and dog) using disproportional libraries.

METHODS

To examine how disproportional source representation affects source identification, simulation studies estimated correct and incorrect prediction probabilities for MS, AS, DA, and k-NN across various libraries. These libraries differed in terms of the number and the relative proportion of sewer isolates that were randomly selected within each source group.

Data used in this study were obtained from the previously described 2003 study and consisted of samples ($N = 242$) collected from animals (dog and gull) and three sewer lift stations along the Mississippi gulf coast (Robinson 2004).

From 73 dog, 106 gull, and 63 sewer samples, 1,666 sewer, 343 dog, and 221 gull enterococci were isolated and confirmed biochemically (USEPA 2000). An average of 4 dog, 2 gull, and 26 sewer enterococci isolates were picked from each sample. These isolates were analyzed by BOX-PCR, visualized by gel electrophoresis to create individual isolate fingerprints, and assessed using BioNumerics v3.5 (Applied Maths, Sint-Martens-Latem, Belgium). Band-based binary data (presence/absence) were imported into SAS (v. 9.1, SAS Institute, Cary, NC). Clonal isolates or those isolates identified within individual samples as fingerprints having exact matching banding patterns (e.g. 100% similarity) were excluded from analyses. Dog and gull isolates isolated from the same sample had a high degree of clonality (data not shown) while sewer sample isolates were rarely clonal.

For each simulation, isolates from each source category were randomly selected, without replacement, from the isolate archive using the SAS procedure 'PROC SURVEY-SELECT' and placed into a library. The first simulation library construction consisted of sampling an equal number of isolates from each source group in the archive (100 dog, 100 gull, and 100 sewer). One hundred isolates were chosen from each group because of limiting pools of dog ($n = 343$) and gull ($n = 221$) isolates. In the second set of simulations, libraries were constructed by sampling increasing numbers of sewer isolates (e.g. 200, 300, 400, 800), while keeping the remaining number of dog and gull isolates the same (e.g. 100).

The jaccard similarity coefficient was used as the similarity measure for both MS and AS, Mahalanobis distance was used for DA, and Euclidean distance was used for k-NN. Ties were excluded from analyses if the isolate tied to more than one source during assignment. If an isolate tied to two different isolates within the same source then ties were kept and matched to that source. No thresholds of fingerprint similarity were applied.

Simulations were repeated using k-nearest neighbor as a statistical alternative to MS, AS, and DA. In k-Nearest Neighbor (k-NN), source prediction is based on the unknown fingerprint's proximity to k of the most similar known individuals, rather than proximity to a single known individual or to the source group as a whole. We applied $k = 1, 2, 3, 30$, and 100 nearest neighbor strategies using the SAS procedure 'PROC DISCRIM' and Euclidean distance (SAS v 9.1).

Jackknife estimates of correct and incorrect prediction probabilities were calculated for each of the three sources in the library and for each of the four statistical matching procedures. The standard jackknife analysis, also known as "cross-validation" or "leave-one-out" analysis, calculates the bias of an estimator by deleting one isolate each time from the original data set and examining the similarity of that isolate to the remainder of the isolates in the library (Shao & Tu 1995; Wiggins *et al.* 2003). Jackknife estimates of correct and incorrect prediction probabilities for each source group are based on calculating the percentage of correct and incorrect source assignment across all (deleted) isolates within each source group. This emulates assignment of an unknown isolate to a library unit and provides an estimate of source group bias (correct versus incorrect assignment). Under simple random sampling, these jackknife estimates provide nearly unbiased estimates of library accuracy (and inaccuracy) for classifying unknown isolates for each source.

Final estimates of percent correct prediction and incorrect prediction probabilities (%CP and %IP) for each library construction were based on averaging jackknife estimates across 1,000 simulations. Overall rates of %CP and %IP were based on averaging prediction probabilities across the sources.

RESULTS AND DISCUSSION

The first simulation involved randomly selecting 100 isolates from each of the three source groups and classifying those isolates using jackknife analysis of MS, AS, DA, and 3-NN matching algorithms. Each source group was categorized best by different algorithms (Table 1). 3-NN best matched dog (67%) and sewer (58%) isolates. Gull isolates were best matched by AS (93%). For all source groups, AS showed the lowest (49%) average %CP while 3-NN showed the highest (62%). MS and DA exhibited similar average %CP at 58% and 54%, respectively.

The second set of simulations involved randomly selecting 100 isolates from each dog and gull source group (as in first simulation) and increasing the number of sewer isolates in the library (200 up to 800), and then performing a jackknife analyses on each library to determine the disproportional effect on %CP and %IP. MS exhibited an

Table 1 | The percent correct prediction (%CP) for dog, gull, and sewer sources against maximum similarity (MS), average similarity (AS), discriminant analysis (DA), and 3-nearest neighbor (3-NN) using a proportional group size library

	Source			
Matching algorithm	**Dog**	**Gull**	**Sewer**	**Ave. %CP**
MS	55%	67%	53%	58%
AS	29%	93%	27%	49%
DA	53%	61%	48%	54%
3-NN	67%	62%	58%	62%

increase in correct prediction for sewer isolates (+38%) and a decrease in correct prediction for dog (−25%) and gull (−17%) isolates when additional sewer isolates ($n = 800$) were added to the library (Figure 1). These increases in %CP for sewer were followed by an increase in %IP for dog (+42) and gull (+30%) (Figure 2). AS exhibited a stable (~0% change) %CP across the three sources as sewer isolates were added to the library (Figure 1). DA also exhibited a moderately stable %CP for sewer (+16%), dog (−7%) and gull (−2%) sources upon addition of sewer isolates to the library. These %CP rates were higher than AS (except gull sources) and more stable than MS. AS and DA algorithms exhibited slight changes in %IP as the library became increasingly disproportional (Figure 2). Although changes in AS %IP were negligible across the three sources, DA resulted in modest increases in %IC for dog (+10%) and gull (+6%) as sewer source with the addition of sewer isolates to the library ($n = 800$).

Additional simulations of nearest neighbors were performed (k = 1, 2, 3, 10, 30, 100) (k = 1, 2, 10, 30 and 100 data not shown). 1-NN exhibited results similar to MS. In fact, in terms of matching strategies, 1-NN is equal to MS. The only difference here is that different similarity or distance measures were used; Euclidean was used for 1-NN and Jaccard was used for MS. As more nearest neighbors were added (k = 2,3), bias was reduced and %IP stabilized. As k increases (i.e. k = 100), results generally mirrored AS. There were, however, some differences. For example, dog was classified best using 100-NN. Observed differences between the two strategies are likely to be the effect of different similarity measures (as with 1-NN and MS) and slightly different matching algorithms; DA uses probability-based methods and AS uses average similarity. Our results indicated that when MS appears to be a better matching algorithm with a given data set in proportional library conditions, a lower number of nearest neighbors (i.e. 1–3) may be most appropriate in disproportional conditions. Conversely, when AS seems to be a better choice in proportional conditions, a larger number of nearest neighbors (i.e. 100–200), depending on library size and proportion, may yield more reliable results. In this study, k = 3 nearest neighbor provided an optimum balance of %CP and %IP (Figures 1 and 2) among the sources.

Researchers have suggested that removing clonal isolates from the library improves prediction and library representativeness within a source tracking library (Wiggins *et al.* 2003; Hassan *et al.* 2005). The library sources in this study exhibited clonality within individual sample by type. Dog and gull samples were frequently clonal while sewer samples were rarely clonal. Improvement in prediction rates by removal of

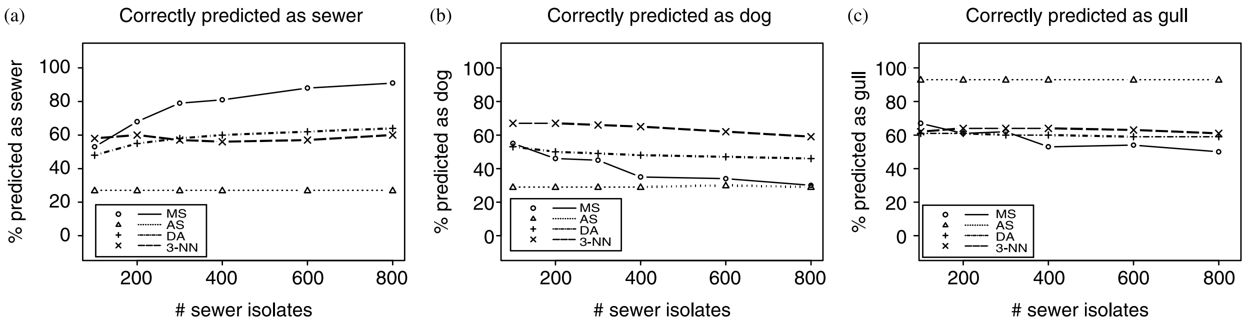

Figure 1 | Estimated probability of correctly predicted isolates into each source, sewer (a), dog (b), and gull (c), as a function of increasing numbers of library sewer isolates for maximum similarity (MS), average similarity (AS), discriminant analyses (DA), and 3-nearest neighbor (3-NN) analysis.

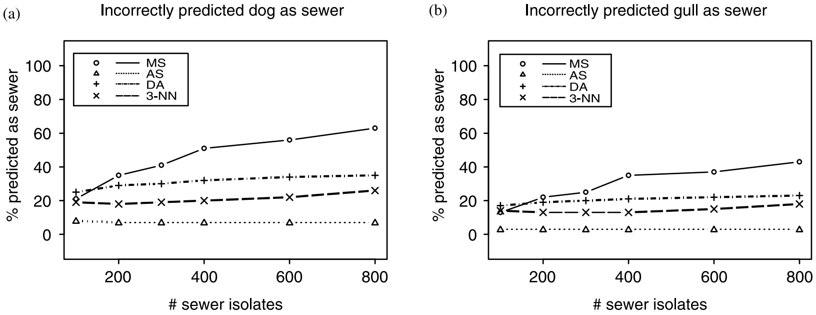

Figure 2 | Estimated probability of incorrectly predicted dog (a) and gull (b) isolates as sewer as a function of increasing numbers of library sewer isolates for maximum similarity (MS), average similarity (AS), discriminant analyses (DA), and 3-nearest neighbor (3-NN) analysis.

clonal isolates by sample depends on which statistical matching algorithm is used. For example, removing clones has no effect on prediction rates with MS since matching relies only on a single isolate. With DA and AS, removing clones may alter predictions substantially depending on the number of clones since means or centroids depend on all the data. We suggest that additional clones should be removed if they belong to the same individual within the same source as additional clones may bias the distribution of fingerprints and artificially reduce variability. Furthermore, removing clones provides a more conservative estimate of correct and incorrect prediction probabilities. When building a source tracking library of this type and addressing sample, library, or analysis bias, one must consider the goals of the study. Is it a goal to obtain the highest rate of correct prediction possible or to correctly identify a contaminating source in a water body, even if the library does not have close to 100% rates of correct prediction? These things must be considered when building a source tracking library.

Bias within a MST library may be caused by additional measures such as overlap in fingerprint distribution among sources. Overlap in fingerprint distribution occurs when fingerprint patterns from one source are more similar to a different source than to its own source group. Unlike sample isolate clonality, fingerprint overlap among library sources affects all matching algorithms tested. The data set exhibited some overlap of rep-PCR fingerprints between sources (Robinson 2004, data not shown). This introduces bias and difficulty of correctly predicting fecal sources completely unrelated to disproportional library size.

In order to deal with overlap, some researchers have suggested the use of similarity thresholds to make it more difficult for isolates to match to different sources (Whitlock *et al.* 2002; Hassan *et al.* 2005). With similarity thresholds, an unknown isolate is eliminated from consideration if the similarity or average similarity coefficient is below some threshold value. The reasoning is that if the similarity coefficient is low, then there is not enough evidence for identifying the source. While this may improve the rates of correct prediction for both proportional and disproportional libraries, the omission of unknown isolates is problematic when attempting to determine the ratio among sources or the greatest source among a collection of samples (Ritter *et al.* 2003). If the decision is made to use thresholds, then we recommend reporting the number of isolates that were omitted as well as the individual predictions for the remaining isolates.

The use of other types of library-based MST methods may or may not result in similar bias as those shown here with rep-PCR when disproportional libraries are used. More specific methods, such as PFGE, that create a larger number of bands and increase resolution among different fingerprints may result in more fingerprint overlap, especially in the presence of multiple subtypes within a source. Less discriminative methods, such as ribotyping, are less source-specific but allow easier matches of library isolates to unknown source isolates which can increase bias as disproportionality increases. The effect disproportional libraries have on library-based methods depends on the distribution of fingerprints. If increasing the resolution separates out the sources from each other then %CP will be greater and bias will not change with increased disproportionality. However, if increased resolution only creates more overlap then bias may increase. We recommend that the user estimates the jackknife probabilities of %CP and %IP associated with a particular method. Choosing a method that discriminates among library sources yet can accurately identify unknown environmental (extralibrary) isolates is ideal.

CONCLUSIONS

Unequal source representation in the library may substantially bias source prediction toward the more dominant library source. The magnitude of bias is affected both by the amount of disproportionality among source candidates and by the choice of statistical algorithm used to match unknowns. Of the three commonly used statistical algorithms (MS, AS, and DA) investigated in this paper, MS was the most sensitive to disproportional source representation. While AS and DA were more robust in the case of disproportional libraries, they were not always the best for correctly matching unknowns. K-Nearest neighbor tended to perform as well as MS when proportional libraries were used, and was as stable as DA and AS when disproportional libraries were tested. The success and stability of k-NN matching strategy is a compromise between matching to a single isolate and matching to the group as a whole.

When analyzing library-based MST data, it is important not only to consider the %CP of sources groups, but also the %IP and the proportional library size. Disproportional library conditions arise frequently due to sampling and processing limitations. However, it is not necessary to eliminate samples from a data set simply to create a proportional library. When disproportional libraries arise, it is necessary to survey the data statistically and compare results using different statistical algorithms as well as considering the possible bias associated with disproportional libraries and some matching algorithms.

Our results suggest that k-NN offers a valuable compromise when working with disproportional libraries, incorporating the strengths of both MS and AS/DA. K-NN allows for the identification of subtypes in the library, a strength of MS, and is robust where increases in prediction bias associated with using disproportionate libraries occur, a strength of AS/DA. The choice of k allows the researcher the flexibility to address each issue: subtype identification and library-based bias as the situation demands. At $k = 1$, k-NN is equivalent to MS. As k increases, k-NN takes on more characteristics of AS/DA. K-NN is suggested for those cases where disproportionate libraries are used and where MS typically performs better than AS or DA using proportional libraries. In choosing k, we suggest calculating jackknife estimates of both correct and incorrect prediction rates for various levels of k. In this way, the researcher can weigh the trade-offs associated with increased correct prediction probabilities and prediction bias. We found that for the 2003 study, $k = 3$ provided an optimum balance. We believe that k-nearest neighbor offers a promising statistical matching algorithm that should be considered when using disproportional libraries. Future research should investigate how disproportional libraries and statistical matching algorithms affect bias in prediction rates for other library-based methods including PFGE, multiple antibiotic resistance (MAR), antibiotic resistance analysis (ARA), and ribotyping data sets.

ACKNOWLEDGEMENTS

The original data set used in this research was generated in 2003 through funding by The National Oceanic and Atmospheric Administration through the Coastal Impact Assistance Program grants MS-17 and MS-24. Thanks also to Kyle Kingsley and Brian West of Applied Maths, Inc. for explanation of the BioNumerics software and to J. Christopher Jolly for aid with organization and writing of this manuscript.

DISCLAIMER

The National Ocean Service (NOS) does not approve, recommend, or endorse any proprietary product or material mentioned in this publication. No reference shall be made to NOS, or to this publication furnished by NOS, in any advertising or sales promotion which would indicate or imply that NOS approves, recommends, or endorses any proprietary product or proprietary material mentioned herein or which has as its purpose any intent to cause directly or indirectly the advertised product to be used or purchased because of NOS publication.

REFERENCES

Applied Maths Inc. 2004 *BioNumerics User Manual Version 4.0*. Applied Maths BVBA, Austin, TX.

Bower, R. J. 2001 Fecal source identification using antibiotic resistance analysis. *Puget Sound Notes* **45**, 3–8.

Carson, C. A., Shear, B. L., Ellersieck, M. R. & Schnell, J. D. 2003 Comparison of ribotyping and repetitive extragenic palindromic-PCR for identification of fecal *Escherichia coli* from humans and animals. *Appl. Environ. Microbiol.* **69**, 1836–1839.

Dombek, P. E., Johnson, L. K., Zimmerley, S. T. & Sadowsky, M. J. 2000 Use of repetitive DNA sequences and the PCR to differentiate *Escherichia coli* isolates from human and animal sources. *Appl. Environ. Microbiol.* **66**, 2572–2577.

Hagedorn, C., Robinson, S., Filtz, J., Grubbs, S., Angier, T. & Reneau, R. Jr 1999 Determining sources of fecal pollution in a rural Virginia watershed with antibiotic resistance patterns in fecal streptococci. *Appl. Envion. Microbiol.* **65**, 5522–5531.

Harwood, V. J., Whitlock, J. & Withington, V. 2000 Classification of antibiotic resistance patterns of indicator bacterial by discriminant analysis: use in predicting the source of fecal contamination in subtropical waters. *Appl. Environ. Microbiol.* **66**, 3698–3704.

Hassan, W. M., Wang, S. Y. & Ellender, R. D. 2005 Methods to increase fidelity of repetitive extragenic palindromic PCR fingerprint-based bacterial source tracking efforts. *Appl. Environ. Microbiol.* **71**, 512–518.

Johnson, D. E. 1998 *Applied Multivariate Methods for Data Analysts*. Brooks/Cole Publishing Co., Pacific Grove, CA, USA.

Ritter, K. J., Carruthers, E., Carson, C. A., Ellender, R. D., Harwood, V. J., Kingley, K., Nakatsu, C., Sadowsky, M., Shear, B., West, B., Whitlock, J. E., Wiggins, B. A. & Wilbur, J. D. 2003 Assessment of statistical methods used in library-based approaches to microbial source tracking. *J. Wat. Health.* **1**, 209–223.

Robinson, B. J. 2004 *Source Analysis using Enterococci and its Application to a Coastal Watershed*. A thesis. Department of Biological Sciences. The University of Southern Mississippi, USA.

SAS Institute Inc. 2004 *SAS OnlineDoc® 9.1.3*. SAS Institute Inc., Cary, NC.

Scott, T., Rose, J., Jenkins, T., Farrah, S. & Lukasik, J. 2002 Microbial source tracking: current methodology and future directions. *Appl. Environ. Microbiol.* **68**, 5796–5803.

Shao, J. & Tu, D. 1995 *The Jackknife and Bootstrap*. Springer-Verlag, New York, USA.

Simpson, J., Santo Domingo, J. & Reasoner, D. 2002 Microbial source tracking: state of the science. *Env. Sci. Technol.* **36**, 5280–5288.

Stewart, J., Ellender, R. D., Gooch, J. A., Jiang, S., Myoda, S. & Weisberg, S. 2003 Recommendations for microbial source tracking: lessons from a methods comparisons study. *J. Wat. Health.* **1**, 225–231.

U.S. Environmental Protection Agency (USEPA) 2000 *Improved Enumeration Methods for the Recreational Water Quality Indicators: Enterococci and Escherichia coli*. EPA Office of Water, Office of Science and Technology, Washington, D.C., USA.

Whitlock, J. E., Jones, D. T. & Harwood, V. J. 2002 Identification of the sources of fecal coliforms in an urban watershed using antibiotic resistance analysis. *Wat. Res.* **36**, 4273–4282.

Wiggins, B. A. 1996 Discriminant analysis of antibiotic resistance patterns in fecal streptococci, a method to differentiate human and animal sources of fecal pollution in natural waters. *Appl. Environ. Microbiol.* **62**, 3997–4002.

Wiggins, B. A., Cash, P. W., Creamer, W. S., Dart, S. E., Garcia, P. P., Gerecke, T. M., Han, J., Henry, B. L., Hoover, K. B., Johnson, E. L., Jones, K. C., McCarthy, J. G., McDonough, J. A., Mercer, S. A., Noto, M. J., Park, H., Phillips, M. S., Purner, S. M., Smith, B. M., Stevens, E. N. & Varner, A. K. 2003 Use of antibiotic resistance analysis for representativeness testing of multiwatershed libraries. *Appl. Environ. Microbiol.* **69**, 3399–3405.

Available online May 2007