A statistical appraisal of disproportional versus proportional microbial source tracking libraries

Brian J. Robinson[1]*, Kerry J. Ritter and*
Rudolph D. Ellender[2]

Abstract

Library-based microbial source tracking (MST) can assist in efforts to reduce or eliminate 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 unknown sources. 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 unknown sources into the library's dominant source category. We examined prediction bias for three of the most commonly used statistical matching algorithms in library-based MST when applied to disproportionately represented known source libraries. These include maximum similarity (MS), average similarity (AS), and discriminant analyses (DA). We found that MS was particularly sensitive to disproportional source representation, while AS and DA were more robust. We discovered that nearest neighbor (NN) analyses provides a compromise between correct prediction and sensitivity to disproportional libraries among the three statistical procedures. This includes increased matching success and stability that should be considered when matching to disproportionally represented libraries.

Introduction

Predicting sources of fecal contamination is important for managing a healthy water body and protecting against disease. By predicting the source(s) of fecal-associated bacteria, microbial sourcetracking (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). Commonly used library-based MST methods rely on the assembly of genetic or phenotypic "fingerprints" from pollution-indicative bacteria cultivated from known sources of fecal contamination (Harwood *et al*. 2003,Scott *et al*. 2002, Simpson *et al*. 2002). Scientists predict unknown sources of pollution using computer-based statistical analysis to match unknown source fingerprints to those from the known-source library (Bower 2001, Dombek *et al*. 2000, Hagedorn *et al*. 1999, Harwood *et al*. 2000, Whitlock *et al.* 2002, Wiggins 1996). The success of the library-based approach depends on: 1) the distribution of fingerprint patterns among source candidates, 2) how well each source candidate is represented in the library, and 3) which statistical algorithm or matching criterion is used to match unknown sources.

Construction of known-source libraries is often limited by the availability of known-source samples and our ability to collect and process those samples (Robinson 2004). As a result, libraries may contain disproportional representation of isolates among the source candidates. For example, collecting large numbers of 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 the statistical matching algorithm used to match unknown source isolates. Library-based methods employ a variety of statistical methods to match

[1] *National Oceanic and Atmospheric Administration, Charleston, NC*
[2] *The University of Southern Mississippi, Department of Biological Sciences, Hattiesburg, MS*

fingerprints of unknown origin to the known-source library (Bower 2001, Carson *et al.* 2003, Dombek *et al.* 2000, Hagedorn *et al.* 1999, Harwood *et al.* 2000, Hassan *et al.* 2005, Ritter *et al.* 2003, Whitlock *et al.* 2002, Wiggins 1996, Wiggins *et al.* 2003). Because 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 that classifies an unknown sources into the source group to which its most similar known member belongs (Applied Maths 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 2004). DA classifies unknown sources into source groups based on a "rule" developed from a calibration data set, e.g., library. 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 (SAS Institute 2004, Johnson 1998). With both AS and DA, disproportional libraries may create unstable estimates for the center of each group by allowing for a greater number of outliers that 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 bacteria in the water, forcing closure of the area by the state to recreational uses (Robinson 2004). Three potential sources of fecal contamination source (dog, gull, and sewer) were identified in this urban, mostly residential, watershed. These 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 led to disproportional representation among source candidates. The resulting library contained approximately five times as many human isolates as dog and gull isolates. Analysis of the data raised concerns that having a greater number of sewerage representatives in the library may have biased identification toward the sewer source.

This paper examines the use of library-based rep-PCR data and three common statistical methods (MS, AS, and DA) in the presence of disproportional source representation. The results are based on simulation studies using the enterococcal fingerprints from the study described above, where we estimate probabilities of correct and incorrect prediction for identifying three sources (sewer, gull, and dog) using disproportional libraries. In addition, we suggest an alternative statistical method, k-nearest neighbor (k-NN) as a valuable compromise among the other three matching strategies.

Methods

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

The isolates used in this study contained 242 samples collected from animals and lift stations along the Mississippi gulf coast during the 2003 calendar year (Robinson 2004). From these samples, 1,666 sewer, 343 dog, and 221 gull enterococci were isolated and confirmed biochemically (USEPA 2000). 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 (SAS Institute 2004) for statistical evaluation. Clones were removed prior to analysis.

For each simulation, isolates from each source category were randomly selected, without replacement, from the isolate archive using the SAS procedure PROC SURVEYSELEC (SAS Institute 2004), 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, and 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, while Mahalanobis distance was used for DA, and Euclidean distance was used for NN. Ties were excluded from analysis 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-NN as a statistical alternative to MS, AS, and DA. Using 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 NN strategy using the SAS procedure PROC DISCRIM and Euclidean distance (SAS Institute 2004).

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. 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 (Wiggins 1996, Shao and Tu 1995). 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 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 for each statistical method.

Results

The first set of 1,000 simulations 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. The percent of correctly identified isolates for each source group varied depending on the statistical algorithm used to match isolates (Table 1). The 3-NN method resulted in the highest %CP for 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.

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 (e.g., n = 100 isolates from each source).

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

The second set of simulations involved randomly selecting 100 isolates from dog and gull (as in the first simulation) and increasing the number of sewer isolates in the library (200 up to 800). Jackknife analyses were performed on each library to determine the disproportional effect on %CP and %IP. MS exhibited a maximum increase in correct prediction for sewer isolates (+38%) and a maximum decrease in correct prediction for dog (-25%) and gull (-17%) isolates as the number of sewer isolates represented in the library increased to n = 800 (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. Although changes in AS %IP were negligible across the three sources, DA resulted in modest increases in %IP for dog (+10%) and gull

(+6%) as sewer source with the addition of sewer isolates to the library (n = 800). As sewer isolates were added to the library, 3-NN exhibited relatively higher and more stable %CP than MS, AS, and DA (Figure 1). The 3-NN method misclassified an additional 7% of dog and 4% of gull isolates as sewer source when comparing proportional and disproportional libraries (Figure 2).

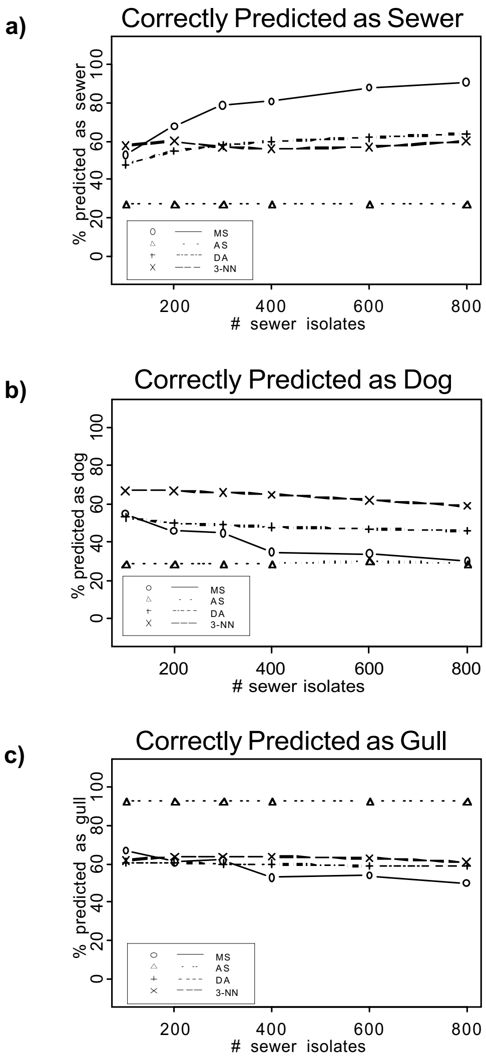

Figure 1. Estimated probability of correctly predicted isolates into each source, sewer (a), dog (b), and gull (c), as a function of increasing numbers of library sewer isolates for maximum similarity (MS), average similarity (AS), discriminant analyses (DA), and 3-nearest neighbor (3-NN) analysis.

Additional simulations of nearest neighbors were performed (k = 1, 2, 10, 30, 100) (data not shown). For 1-NN exhibited results similar to MS. For k greater than or equal to 2, a substantial increase in %CP was observed, followed by a decrease in %IP for gull and dog isolates in disproportional library conditions. Nearest neighbor k = 30 and k = 100 exhibited similar stability to that of AS. However, with this data set, k = 3 seemed the most stable and best choice for classification in disproportional conditions. In general, analysis using the nearest neighbor algorithm was comparable with MS's higher %CP without the instability associated with increased disproportional sewer source representation in the library.

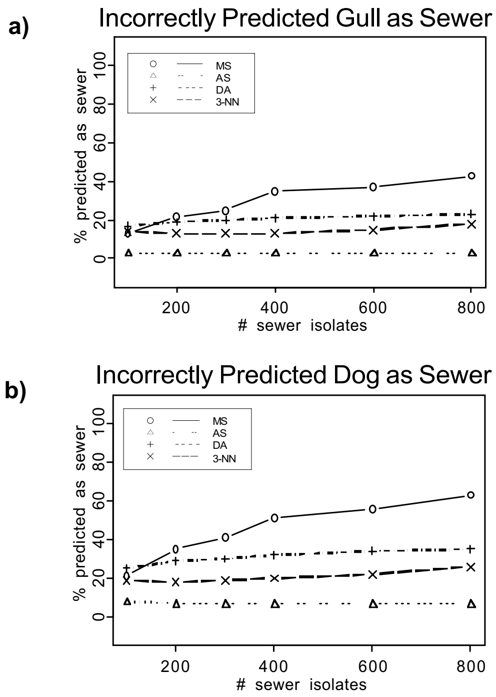

Figure 2. Estimated probability of incorrectly predicted dog (a) and gull (b) isolates as sewer as a function of increasing numbers of library sewer isolates for maximum similarity (MS), average similarity (AS), discriminant analyses (DA), and 3-nearest neighbor (3-NN) analysis.

Discussion

This study showed that 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 unknown sources. 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 to disproportional libraries, they were not always the best for correctly matching unknown sources. This may be due, in part, to the high degree of overlap among the distribution of fingerprint patterns, multimodality, and the presence of subtypes within each sources (Ritter *et al*. 2003. Such patterns have been noted by others and were confirmed by cluster analyses of the original library in the 2003 study.

The k-NN method allows for the identification of subtypes in the library (a strength of MS), and is robust to increases in prediction bias associated with using disproportional libraries (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. In this study, discrepancies among the k-NN and MS and AS resulted from the choice of distance measure; Euclidean distance was used for NN instead of Jaccard that was used with MS and AS. the k-NN (k = 3) method tended to perform as well as MS when proportional libraries were used, and correct and incorrect prediction rates were nearly as stable as with DA and AS when disproportional libraries were tested. The success and stability of k-NN matching strategy is a compromise between the matching to a single isolate and matching to the group as a whole.

Bias within an MST library may be caused by additional measures such as fingerprint overlap between the sources. The data set exhibited some overlap of rep-PCR fingerprints between sources, data not shown (Robinson 2004). This introduces bias and difficulty of correctly predicting fecal sources completely unrelated to disproportional library size. The biological characterization method specificity, rep-PCR in this case, can be used to access the fidelity of a source tracking library. Other options may include omitting overlapping/homologous fingerprints from the library or applying similarity thresholds, which may increase the accuracy of library matches.

MST researchers frequently evaluate the effectiveness of their source library and the reliability of their statistical matching algorithm by estimating probabilities of correct prediction. However, little, if any, attention is given to prediction bias among the various source categories. Bias during analysis can lead to incorrect prediction of the true pollution sources and funds may be spent to remediate the wrong source(s) of pollution. These types of situations are particularly problematic for water resource managers.

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 consider 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. We suggest applying k-NN strategy 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 classification rates for various levels of k. In this way, the researcher can weigh the trade-offs associated with increased correct classification probabilities and prediction bias. We found that for the 2003 study, k = 3 provided an optimum balance. We believe that k-NN offers a promising statistical matching algorithm that should be considered when using disproportional libraries. These findings could be applicable to any disproportional source tracking library multivariate data including rep-PCR, multiple antibiotic resistance, antibiotic resistance analysis, and ribotyping data sets.

Literature Cited

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., B.L. Shear, M.R. Ellersieck and J.D. Schnell. 2003. Comparison of ribotyping and repetitive extragenic palindromic-PCR for identification of fecal Escherichia coli from humans and animals. *Applied Environmental Microbiology* 69:1836-1839.

Dombek, P.E., L.K. Johnson, S.T. Zimmerley and M.J. Sadowsky. 2000. Use of repetitive DNA sequences and the PCR to differentiate Escherichia coli isolates from human and animal sources. *Applied Environmental Microbiology* 66:2572-2577.

Hagedorn, C., S. Robinson, J. Filtz, S. Grubbs, T. Angier and R. Reneau Jr. 1999. Determining sources of fecal pollution in a rural Virginia watershed with antibiotic resistance patterns in fecal Streptococci. *Applied Environmental Microbiology* 65:5522-5531.

Harwood, V.J., C. Hagedorn, R.D. Ellender, J. Gooch, J. Kerns, M. Samadpour, A.C. Chapman and B.J Robinson. 2003. Phenotypic library-based microbial source tracking methods: Efficacy in the California Collaborative Study. *Journal of Water and Health* 1:153-166.

Harwood, V.J., J. Whitlock and V. Withington. 2000. Classification of antibiotic resistance patterns of indicator bacterial by discriminant analysis: use in predicting the source of fecal contamination in subtropical waters. *Applied Environmental Microbiology* 66:3698-3704.

Hassan, W.M., S.Y. Wang, and R.D. Ellender. 2005. Methods to increase fidelity of repetitive extragenic palindromic PCR fingerprint-based bacterial source tracking efforts. *Applied Environmental Microbiology* 71:512-518.

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

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

Robinson, B.J. 2004. Source analysis using Enterococci and its application to a coastal watershed. Thesis, Department of Biological Sciences, University of Southern Mississippi. Mississippi, MO.

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

Scott, T., J. Rose, T. Jenkins, S. Farrah and J. Lukasik. 2002. Microbial source tracking: current methodology and future directions. *Applied Environmental Microbiology* 68:5796-5803.

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

Simpson, J., J. Santo Domingo and D. Reasoner. 2002. Microbial source tracking: state of the science. *Environmental Science Technology* 36:5280-5288.

Stewart, J., R.D. Ellender, J.A. Gooch, S. Jiang, S. Myoda and S. Weisberg. 2003. Recommendations for microbial source tracking: lessons from a methods comparisons study. *Journal of Water and Health* 1:225-231.

US 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, DC.

Whitlock, J.E., D.T. Jones and V.J. Harwood. 2002. Identification of the sources of fecal coliforms in an urban watershed using antibiotic resistance analysis. *Water Research* 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. *Applied Environmental Microbiology* 62:3997-4002.

Wiggins, B.A., Philip W. Cash, Wes S. Creamer, Scott E. Dart, Preston P. Garcia, Todd M. Gerecke, Jennifer Han, Brian L. Henry, Kylie B. Hoover, Erika L. Johnson, K.C. Jones, Jacquie G. McCarthy, Justin A. McDonough, Sarah A. Mercer, Michael J. Noto, Haewon Park, Matthew S. Phillips, Stephanie M. Purner, Brian M. Smith, Erin N. Stevens and Amy K. Varner. 2003. Use of antibiotic resistance analysis for representativeness testing of multiwatershed libraries. *Applied Environmental Microbiology* 69:3399-3405.

Acknowledgements

The original data set used in this research was generated in 2003 with 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 Bio-Numerics software, and to J. Christopher Jolly for aid with the organization and writing of this manuscript.