Classification Tree Method for Bacterial Source Tracking with Antibiotic Resistance Analysis Data

Bertram Price,1* Elichia A. Venso,2,3 Mark F. Frana,2 Joshua Greenberg,1 Adam Ware,1 and Lee Currey4

Price Associates, One North Broadway, White Plains, New York 10601;1 Department of Biological Sciences2 and Environmental Health Science,3 Salisbury University, Salisbury, Maryland 21801; and Maryland Department of the Environment, 1800 Washington Blvd., Baltimore, Maryland 21230-17184

Received 11 October 2005/Accepted 27 February 2006

Various statistical classification methods, including discriminant analysis, logistic regression, and cluster analysis, have been used with antibiotic resistance analysis (ARA) data to construct models for bacterial source tracking (BST). We applied the statistical method known as classification trees to build a model for BST for the Anacostia Watershed in Maryland. Classification trees have more flexibility than other statistical classification approaches based on standard statistical methods to accommodate complex interactions among ARA variables. This article describes the use of classification trees for BST and includes discussion of its principal parameters and features. Anacostia Watershed ARA data are used to illustrate the application of classification trees, and we report the BST results for the watershed.

Bacterial source tracking (BST) with antibiotic resistance analysis (ARA) has been conducted using various statistical classification models to identify sources of bacterial contamination of surface waters (3–5, 9–15). Isolates are obtained from fecal samples from known sources, such as humans, pets, livestock, and wildlife) and 50 human source samples from sewage treatment facilities. A total of 1,155 known source enterococcal isolates were obtained from these fecal samples. Table 1 shows the distribution of samples and isolates among the four sources. Forty-two antibiotic-concentration combinations, subsequently referred to as the antibiotic variables, were used to form the ARA data. If the culture grew in the presence of a particular antibiotic, then it was considered resistant and recorded as “1.” If the culture did not grow in the presence of the antibiotic, then it was considered susceptible and recorded as “0.” Therefore, each observation (isolate) in the reference library is a sequence of 42 zeros and ones. Five of the 42 antibiotic variables were omitted from the statistical analysis: one because it was constant across all sources and four because they were not applied to any of the isolates from human sources. Of the 1,155 isolates, 1,074 had data for all of the remaining 37 antibiotic variables.

In addition to the reference library data that were used to build and validate a classification model, 1,565 water sample isolates were generated from 71 water samples collected from the Anacostia River; 1,512 of these isolates had complete ARA data. The procedures used to develop the reference library and water sample isolate data are described in a publication from the Maryland Department of the Environment (8).

Classification tree method. Classification tree methods build classification models by recursively splitting in two the reference library of isolates such that the resulting subsets, termed nodes, are increasingly homogenous with respect to isolate sources (1, 6). (We used the Classification and Regression Trees [CART] software developed by Salford Systems to build a classification tree for the ARA data.) The first step divides the reference library of isolates into two nodes by considering every binary split defined by the 0 or 1 outcome associated with every antibiotic variable. A variable is selected for the split that maximizes homogeneity of isolate sources within each node. The same procedure is applied to each resulting node. The collection of nodes at the ends of branches of the tree, called terminal nodes, is a classification model. Each terminal node is characterized by a unique antibiotic resistance profile and is associated with one source category. A bacterial isolate from a water sample that has the same antibiotic resistance profile as the profile for a particular terminal node would be assigned the source associated with that terminal node.

Criteria governing node splits. Various approaches for determining when to conclude the node-splitting process have been suggested and analyzed (1). Stopping criteria, or rules for determining when to stop splitting nodes, typically are based on a measure of source homogeneity for isolates within a node. Source homogeneity achieves its maximum when all isolates in a node have the same source. A node where an additional split would not yield a specified minimum increase in homogeneity would be a terminal node.

Hastie et al. (6) described three criteria, or indexes, that address homogeneity. As formulated, the indexes measure the complement of homogeneity, which Hastie et al. refer to as impurity. Homogeneity is maximized when impurity is minimized. The three impurity indexes are misclassification error, Gini index,
and cross-entropy deviance. The optimal split of a node at any stage of model development is the one that minimizes impurity as measured by the chosen index. We relied on the Gini index for developing the Anacostia classification tree model.

The Gini index for a node labeled i is calculated as \( \sum p_{jk} (1 - p_{jk}) \), where \( p \) is the proportion of isolates from source \( k \) in node \( m \). The minimum value taken by the Gini index is 0, which occurs when all isolates in a node are from the same source (i.e., minimum impurity corresponds to maximum homogeneity). The maximum of the Gini index occurs when the isolates in a node are equally distributed among the sources.

At this time, we know of no general guidance concerning a preference for one of the impurity indexes for BST modeling. Following the general suggestion that model building is an exploratory data analysis process, experimentation with the impurity indexes for BST modeling. Following the general suggestion that model building is an exploratory data analysis process, experimentation with each of the indexes may be beneficial.

### Source identification for terminal nodes.

The classification model, equivalently the set of rules for classifying isolates by source, is embodied in the collection of terminal nodes. Each terminal node in a tree represents isolates with a particular pattern of antibiotic resistance and assigns the source with highest probability in the node to all such isolates.

The source probabilities in a terminal node are posterior probabilities of the next source in the watershed, \( \pi_i \), for each source \( i \) in the reference library, assuming that the collection of fecal samples provide estimates of the true proportion of potential sources of bacteria. In the Anacostia reference library, we assume that the collection of fecal samples provide adequate coverage of the sources in the watershed, but the numbers of isolates for each source are not necessarily proportional to actual source contributions. The second obvious choice for prior probabilities, equal probabilities of 0.25 for each of the four sources in the Anacostia analysis, is a more acceptable choice. In the absence of specific reliable data on the source contributions of bacteria from the watershed, the equal prior probabilities properly reflect ignorance concerning the source contributions and, in that respect, lead to an unbiased classification model. Under choice of the equal prior probabilities, the posterior probabilities would be calculated using equation 1 with each \( P(S_j) \) replaced by 0.25. The numerical example in Table 2 demonstrates how source predictions are different depending on the choice of values for the prior probabilities.

#### Parameters governing model stability.

A classification model is stable if it consistently classifies new isolates from known sources accurately. Stability depends to a large extent on the representativeness and size of the reference library, two characteristics that are addressed in Discussion. For a given reference library, stability may be advanced by imposing size limitations on nodes that are candidates for splitting and size limitations on the terminal nodes. These limitations are intended to improve the statistical reliability of the classification model by increasing the statistical stability of terminal nodes. For example, requiring the number of isolates in a terminal node to be no less than a specified number is equivalent to imposing a statistical precision requirement on the estimates of posterior probabilities for that node.

### Evaluating alternative classification tree models.

A classification model for BST is judged by the likelihood of correctly predicting the source of isolates grown from water samples containing bacteria from unknown sources. The most basic assessment of this likelihood is a resubstitution estimate of the rate of correct classification (i.e., using the model to reclassify the reference library isolates). A relatively high value of this resubstitution estimate of the correct classification rate would be expected because the model is being used to classify the same isolates that were used to build it. Although a minimum acceptable limit for the correct classification rate has not been established, the model should at least have a correct classification rate that exceeds the rate for classification by chance (i.e., random assignment of isolates to sources). Therefore, the rate of correct classification for each source should be larger than the relative frequency of isolates in the reference library from that source.

Cross-validation provides a more realistic estimate of the correct classification rate than resubstitution because the predicted source of an isolate is based on a model that did not employ the isolate for building the model. Cross-validation is accomplished for large reference libraries by developing the classification tree from a subset of reference library isolates by excluding a random subset of the isolates from the reference library. Using the resulting model to classify the isolates that were excluded leads to an independent estimate of the correct classification rate, which would be expected to be smaller than the resubstitution estimate. For smaller reference libraries, cross-validation may be structured to provide a bootstrap estimate of the correct classification rate (see references 1, 2, and 6 for general discussions of bootstrapping). In the software we employed, the bootstrap method is referred to as \( K \)-fold cross-validation. This method randomly establishes \( K \) nonoverlapping subsets of the reference library isolates, each consisting of \( 1/K \) of the total number of isolates in the reference library.
Each subset is “held out,” and the remainder of the reference library is used to build a tree classification model that, in turn, is used to classify the “held-out” isolates. Accumulating the results for all $K$ subsets leads to an alternative estimate of the correct classification rate. This estimate provides a more realistic assessment of model performance than the estimate based on resubstitution because the predicted classification of an isolate is based on a model that did not employ the isolate to build the model. The bootstrap estimate of the correct classification rate usually would be smaller than the resubstitution estimate. In our analysis, $K = 10$. (In addition to using $K$-fold cross-validation to estimate the rate of correct classification, CART, the software we used to build a classification model, uses information from the $K$-fold analysis to determine the optimal classification model. For details, see references 1 and 6.)

Classification probability thresholds. As described above, each terminal node is associated with the source that has the maximum posterior probability for the node. Also, an isolate that has the same antibiotic resistance profile as the node would be assigned the source associated with the node. In terminal nodes where the maximum probability for the source associated with the node is not much larger than probabilities for other sources represented in the node, classifications based on this node are likely to be uncertain. This uncertainty can be reduced by establishing a threshold for the maximum probability in the node before accepting the node as a basis for assigning isolates to one of the sources. If the maximum probability does not exceed the threshold, isolates in the terminal node would be classified as “source unknown.”

For example, assume that a node contains one pet isolate, one human isolate, two livestock isolates, and six wildlife isolates and, for purposes of this example, that the posterior probability is the relative frequency of isolates in the node. (For this example, we are assuming that the prior probability distribution for sources is the distribution of sources in the reference library. See “Source identification for terminal nodes” above for a discussion of prior probabilities.) The posterior probabilities would be 10% pet, 10% human, 20% livestock, and 60% wildlife. Without any threshold, all isolates falling into this node would be classified as wildlife. With a threshold of 70%, isolates falling into this node would be classified as unknown. If the threshold was 55%, the node and all isolates in it would be classified as wildlife.

A threshold probability reflects our confidence in the classification scheme. While increasing the threshold increases the correct classification rate, it also increases the number of isolates classified as unknown. The trade-off between correct classification rate and number of unknowns needs to be investigated by varying the value assigned to the probability threshold. Application of a probability threshold for the Anacostia Watershed is described in Discussion.

RESULTS

Here we present statistical results for the application of the tree classification method to data from the Anacostia Watershed.

Initial analysis. We used CART to build a tree classification model for four sources based on the 1,074 isolates from 135 samples with complete data in the Anacostia Watershed reference library (Table 1) The ARA data vary statistically both between samples for a particular source and within samples (i.e., across isolates grown from a particular sample). Both sources of variability must be accounted for in the modeling process, and they are important parameters for determining the number of samples and isolates needed for developing a satisfactory model. We used the 1,074 isolates for model building. The effects of both between- and within-sample variability are discussed with the results of the final model.
We chose the Gini index as the impurity criterion and implemented K-fold cross-validation with $K = 10$. Additionally, the minimum number of isolates in a parent node was set to 10 and the prior probabilities were set equal to 0.25, a choice that reflects the absence of information concerning the true distribution of sources in the watershed. The final tree consisted of 171 nodes, 86 of which were terminal nodes.

Figure 1 shows the first few nodes of the tree for illustration purposes. At the top is the root node, which is comprised of all isolates in the reference library. The first split was based on chlorotetracycline at a concentration of 100 μg/ml, minimizing the Gini index. None of the nodes shown in Fig. 1 are terminal nodes. Each node was subsequently split, leading to a classification model consisting of 86 terminal nodes.

Table 3 displays the model predictions for library isolates by using resubstitution. The overall correct classification rate for isolates by resubstitution is 81.0%. Table 3 also displays the distribution of predicted sources by actual source. The minimum correct classification rate by resubstitution was 75.9%, for livestock isolates.

Table 4 displays correct classification rates for samples by using resubstitution. These results address the effect of within-sample variability (i.e., variation across isolates). We consider a sample to be correctly classified if a minimum cutoff of correct classification is achieved for the isolates grown from that sample. For example (Table 4), a sample may be considered correctly classified if at least 60% of the isolates grown from that sample are correctly classified. As the cutoff percentage increases, the percentage of samples classified declines. This pattern is a manifestation of within-sample variation across isolates. These results are discussed further in “Reference library representativeness and size” below, which discusses determination of the number of samples (and isolates per sample) for a BST field study design.

Using 10-fold cross-validation, which provides a more accurate assessment of the model’s predictive ability, the overall correct classification rate is estimated to be 72.2% (Table 5), which is 8.8% lower than the estimate obtained by resubstitution. The estimates of correct classification rates by source range from 64.1% for livestock isolates to 82.0% for human isolates.

Applying the tree classification model to the 1,565 Anacostia River water isolates yielded the following distribution of sources: 468 (29.9%) pet, 222 (14.2%) human, 437 (27.9%) livestock, and 438 (28.0%) wildlife. These results were determined from analysis of all the water isolates, which represent six monitoring stations with samples collected monthly for 1 year. Therefore, the source distribution presented here does not account for the distribution of high-flow and low-flow periods, which may contribute different sources to the streams. Also, note that bacterial sources can be site specific in a watershed, given the nonconservative nature of bacterial transport. For the purpose of this analysis, all the water isolates from the six monitoring stations were used to estimate the overall watershed relative source contributions. The results based on this averaging method indicate that humans contribute the least bacterial contamination to the Anacostia River. The other sources of bacterial contamination are evenly distributed among pet animals, livestock, and wildlife.

**Model refinement with probability thresholds.** We conducted an analysis to determine whether a probability threshold would improve the classification model. We evaluated the trade-off between increasing the rate of correct classification and the percentage of isolates that would be classified as unknown by plotting these two quantities for a range of potential probability threshold values.

Figure 2 shows the results of applying different probability thresholds to isolates in the reference library. If the threshold was 50%, the percentage of isolates classified as unknown would be less than 5%, but the improvement in the rate of correct classification would be minimal. Thresholds of 70% or greater are associated with at least a 10% improvement in the correct classification rate, but the percentage of unknowns also increases. We chose 80% as the threshold. For this threshold, the percent correctly classified increased from 81% to over 90%; the percentage of unknowns for this threshold is 44%.

Table 6 shows how the model classified the library isolates when an 80% probability threshold was used. We observe a significant increase in the correct classification rates, both for

| TABLE 3. Resubstitution results for classification of Anacostia Watershed reference library isolates by using the tree classification model |
|---|---|---|---|---|---|---|
| Actual source | Pet | Human | Livestock | Wildlife | Total (%) |
| Pet | 187 | 10 | 13 | 26 | 236 | 22.0 | 79.2 |
| Human | 15 | 346 | 17 | 21 | 399 | 37.2 | 86.7 |
| Livestock | 12 | 11 | 186 | 36 | 245 | 22.8 | 75.9 |
| Wildlife | 14 | 15 | 14 | 151 | 194 | 18.1 | 77.8 |
| Total (%) | 228 (21.2) | 382 (35.6) | 230 (21.4) | 234 (21.8) | 1,074 |

* Overall correct classification, 81.0%.

<p>| TABLE 4. Correct classification percentages for samples based on resubstitution |</p>
<table>
<thead>
<tr>
<th>Source</th>
<th>No. of samples</th>
<th>% of samples correctly classified with indicated minimum cutoff (%)&lt;sup&gt;a&lt;/sup&gt;</th>
<th>60</th>
<th>70</th>
<th>75</th>
<th>80</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pet</td>
<td>28</td>
<td>89.3</td>
<td>78.6</td>
<td>67.9</td>
<td>60.7</td>
<td></td>
</tr>
<tr>
<td>Livestock</td>
<td>32</td>
<td>78.1</td>
<td>68.8</td>
<td>65.6</td>
<td>65.6</td>
<td></td>
</tr>
<tr>
<td>Human</td>
<td>50</td>
<td>98.0</td>
<td>90.0</td>
<td>90.0</td>
<td>70.0</td>
<td></td>
</tr>
<tr>
<td>Wildlife</td>
<td>24</td>
<td>79.2</td>
<td>66.7</td>
<td>58.3</td>
<td>54.2</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>134</td>
<td>88.1</td>
<td>78.4</td>
<td>73.9</td>
<td>64.2</td>
<td></td>
</tr>
</tbody>
</table>

<sup>a</sup> Minimum cutoffs refer to percentages of isolates per sample used to define a correctly classified sample.
sources individually (minimum equal to 91.9%) and overall (93.0%). Also, the number of classification errors is small compared to the resubstitution and cross-validation results (Tables 3 and 5), and there is no systematic pattern of misclassification by actual source. However, we must accept an overall rate of unknowns equal to 44.3%. The percentage of unknowns for the wildlife source is over 70%, suggesting that this source was not well differentiated by these ARA data.

Table 7 displays the results of applying the model, with and without the probability threshold of 80%, to water sample isolates. The threshold model classifies 500 of the 1,565 water isolates (31.9%) as unknown. The most notable change in the source prediction results occurs for livestock, which represented 27.9% of the water isolates based on the initial model but only 15.1% based on the threshold model. Also, the overall distribution of isolates among sources based on the threshold model is different from the distribution determined by the initial model. The initial model indicated that the human source, at approximately 14%, was about 50% of the other sources, which were approximately 28%. Based on the threshold model, the human and livestock sources (17.5% and 15.1%, respectively) are each the source of approximately half the percentages of the water isolates attributed to the pet and wildlife sources (34.5% and 33.0%, respectively). The threshold model, which we prefer even with its moderate rate of unknowns, suggests potentially different risk management strategies than the initial model for reducing the bacterial content of Anacostia River water.

## DISCUSSION

Here we discuss a variety of issues that have been raised in BST classification modeling, including missing data, cross-validation, and size and representativeness of the reference library.

### Missing data

The Anacostia Watershed reference library consisted of 1,155 isolates, each with outcomes for 37 antibiotic-concentration combinations. Of the 1,155 isolates, 1,074 had data for all antibiotic-concentration combinations; the remaining 81 isolates were missing data for 12 antibiotic-concentration combinations. Because most statistical classification methods require complete data on all variables, either isolates with one or more missing values are omitted from statistical analysis or missing values are imputed (i.e., replaced with estimates from similar isolates with complete data). CART, the tree classification software we employed, includes a procedure similar to imputation that is based on “surrogate splitters” (1, 6). In addition, the classification tree method, due to its flex-

---

**TABLE 5. Classification of Anacostia Watershed library isolates with cross-validation**

<table>
<thead>
<tr>
<th>Actual source</th>
<th>No. of isolates with predicted source:</th>
<th>% of isolates in reference library</th>
<th>% Correctly classifieda</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Pet</td>
<td>Human</td>
<td>Livestock</td>
</tr>
<tr>
<td>Pet</td>
<td>155</td>
<td>22</td>
<td>19</td>
</tr>
<tr>
<td>Human</td>
<td>31</td>
<td>327</td>
<td>20</td>
</tr>
<tr>
<td>Livestock</td>
<td>22</td>
<td>19</td>
<td>157</td>
</tr>
<tr>
<td>Wildlife</td>
<td>22</td>
<td>19</td>
<td>17</td>
</tr>
<tr>
<td>Total (%)</td>
<td>230 (21.4)</td>
<td>387 (36.0)</td>
<td>213 (19.8)</td>
</tr>
</tbody>
</table>

a Overall correct classification, 72.2%.
To accommodate missing ARA data directly, a fixed numerical value other than 0 or 1 would be assigned to each missing data entry in the reference library. Then, the analysis treats the missing value simply as another outcome value in the tree-building process. If missing data occur at random, the replacement of missing data with a value other than 0 or 1 should not alter the result that would have been achieved if there were no missing data. However, missing data may not occur at random but may be correlated with the sources. For example, if a particular antibiotic applied to isolates from a particular source often produced a laboratory result recorded as missing, the classification analysis would indicate a correlation with that source. The correlation could be an indication that a biological characteristic of the source is responsible for an unsuccessful laboratory outcome with the antibiotic in question. Therefore, the correlation could be useful for building the classification tree. If a water sample isolate from an unknown source registered “missing data” for the same antibiotic, the likelihood that the isolate came from the source that had a high percentage of missing data in the reference library for this antibiotic would be increased.

The missing data in the Anacostia Watershed reference library were limited to 81 isolates and the same 12 antibiotic-concentration combinations. All isolates were tested with an initial panel of antibiotics in various concentrations. Later in the project, additional antibiotics and concentrations were added to the panel, requiring the growth and testing of frozen isolates. Missing data resulted when 81 random isolates would not regrow for reasons unrelated to source. Therefore, the pattern of missing data contained no information that could help identify isolate sources. We omitted the 81 isolates with missing data from all subsequent analyses.

**Reference library representativeness and size.** Two important recurring questions concerning the validity of all BST classification modeling methods are (i) whether the reference library is representative of the watershed and (ii) whether the reference library consists of a sufficient number of samples and isolates to develop a reliable model. Neither question can be simply and completely answered. However, we provide guidance based on general statistical principles and the specifics of BST using ARA data.

Concerning representativeness, statistical principles dictate that samples should be collected in a manner that would provide coverage for the total watershed. Stratification and proportional sampling by strata could be used if there was information indicating that certain areas within the watershed were more likely as sources of fecal bacteria than other areas. Also, if information on the relative population sizes of the contributing sources (i.e., pet, human, livestock, and wildlife for the Anacostia Watershed) was available, the source categories could be used to define strata and a proportional sampling scheme for these strata could be designed. However, in any initial BST investigation such as the investigation conducted for the Anacostia Watershed, it is unlikely that the information needed to define strata would be well developed. Therefore, a sampling design based on a combination of expert scientific judgment and convenience is a practical choice.

For the Anacostia Watershed, sampling was conducted after an initial field survey designed to identify likely “major” contributors to fecal bacteria contamination. Field personnel were instructed to concentrate their scat collection efforts around these major sources.

The size of the reference library is determined by specifying the number of field samples to be collected, the number of isolates to be grown from each sample, and the number of isolates analyzed by ARA. Determining the size (i.e., number of samples and isolates) required for a reference library may follow the usual approach for setting requirements for the number of observations in a statistical sample: (i) state a well-defined quantitative objective for the statistical analysis, (ii) determine the size of the statistical error considered acceptable for an estimate of the objective, and (iii) select a probability or confidence level for ensuring that the estimation error is not exceeded. This approach, although conceptually straight-
forward, is complicated in application and has not been applied in the BST literature.

As an introduction to the procedure for determining the size of the reference library for classification tree modeling, we provide an example of this procedure and its complications with a more familiar statistical classification method, i.e., linear discriminant analysis. As a first step, it would be useful to know whether the means of the antibiotic-concentration combinations used to build the classification model (the variables in the discriminant function) are statistically different across the sources. For two sources and data that have an approximately normal distribution, the test is straightforward and depends explicitly on the number of isolates from each source (7). It follows that 95% confidence intervals could be developed for the mean vectors for each source and inspected to determine whether these contours overlapped. If the overlap was pronounced, it is likely that the classification model would perform poorly. The widths of the confidence contours depend on the inherent variation in the data and the number of isolates from each source. As such, the number of isolates from each source could be determined to achieve a specified separation of the means for the two sources. The complications for BST using linear discriminant analysis to model ARA data are that (i) ARA data do not follow an approximately normal distribution and (ii) extensions of the two-source confidence contour analysis to multiple sources have not been fully developed. These complications affect all statistical classification methods that rely on the assumption that the data have a normal distribution.

We describe an approach for determining the reference library size for BST by using the tree classification method, but practical implementation of this approach must await further development. For example, set the well-defined quantitative objective to be the rate of correct classification as estimated by resubstitution, and require a reference library large enough to ensure no more than a 5% error (plus or minus) with 99% confidence for the estimate. Then, by considering models with different numbers of terminal nodes, the required number of isolates could be determined, which, based on a determination of the number of isolates per sample, leads to the number of samples. Although this objective can be stated simply, it has not been translated into a requirement for the size of a reference library for use with the classification tree method. Another approach for determining the required number of isolates would be to require a small error for the estimates of the maximum posterior probability in terminal nodes. Recall that the maximum posterior probability in a terminal node determines the source for that node.

A concept referred to as artificial clustering has been suggested by some researchers as a way to judge whether or not the reference library is large enough (5, 12). However, it is unlikely that the approach employing the concept of artificial clustering is informative for determining the necessary size of a reference library. As described by Harwood et al. (3), a test of the library size would be conducted by randomly assigning the isolates in the library to sources. For example, if there were four sources, each isolate would have a probability equal to 0.25 of being assigned to one of the four sources. The statistical classification procedure would then be applied to the library isolates, and the resulting model would be used to reclassify the isolates among sources. Because the isolates were assigned to sources at random (i.e., the assignment was independent of the ARA outcomes), if the statistical classification procedure is not biased, it would function as a random classification procedure. Each isolate would have a 25% chance of being assigned to each source. The expected resubstitution result would be “25% of the isolates classified into each source.” Various authors have relied on empirical results showing that as the library size increases, the source classifications deviate less from the expected 25%. Therefore, the realization of large deviations from the expected 25%, a circumstance referred to as artificial clustering, is interpreted as an indication that the library is too small.

Artificial clustering as defined, implemented, and interpreted does not provide useful information about the adequacy of library size. Let N be the number of isolates in a reference library. Assume that the isolates came from four sources. Any individual random assignment of the N isolates to the four sources, which subsequently is analyzed by a statistical classification method, would be expected to classify 25% of the N isolates into each source category. A deviation from the 25% target has two possible causes. First, the statistical classification method may be biased. It is generally assumed that this is not the case. Second, the deviation could be a consequence of statistical variation, and larger deviations would be more likely for smaller reference libraries. Deviations from the 25% target can be predicted on a probability basis using the multinomial distribution, but even then they provide minimal guidance for determining if the size of the reference library is acceptable.

At the current stage of BST model development, the only way to assess the size of a reference library is through sensitivity analysis by varying the size of the library or adding more isolates from known sources and evaluating the stability of the model.

**Interpretation of water sample isolate source predictions.** We used a statistical method, classification tree analysis, to build a model for classifying isolates among sources and applied the model to water sample isolates with unknown sources. The resulting distribution of isolates among sources has implications for risk management interventions designed to reduce bacterial contamination of river waters. Our analysis suggests that the human and livestock sources contribute fewer bacteria to the Anacostia River than the pet and wildlife sources (Table 7). Although we are confident that the results indicate the correct attribution for sources, we have not conducted a statistical test to determine if the percentages are statistically significantly different, and at present we have not developed the methodology for conducting such a test. We are currently investigating methods for efficiently evaluating statistical differences among predicted source percentages derived from classification tree models.

**REFERENCES**


---

There appears to be a typographical error in the source identifier "http://aem.asm.org/". It is not recognized or decrypted by the system for display or further action.


