Chapter III

KNOWN SOURCE LIBRARY CONSTRUCTION: WATERSHED CHARACTERISTICS, SAMPLE COLLECTION, SAMPLE ANALYSIS USING ANTIBIOTIC RESISTANCE ANALYSIS, AND STATISTICS

INTRODUCTION

The 1972 Clean Water Act, and subsequent legislation, requires that states must determine what waterbodies are impaired by pollutants and identify the source of those impairments. In accordance with this legislation, the Appomattox Total Maximum Daily Load (TMDL) development study dealt with determining the source of fecal contaminants for the Appomattox River watershed. This study is part of Phase I of the Virginia Department of Environmental Quality TMDL 10-year project. Within the Piedmont and South Central regions, forty sampling locations were selected. Water samples were collected monthly for twelve months, analyzed for bacterial counts and for source tracking by comparison with a known source library that was constructed from fecal samples collected within the watershed.

Over the past 8 years, several researchers (Wiggins 1996, *et al.,* 1997, *et al.,* 2003; Harwood *et al.,* 2000; Hagedorn *et al.,* 1999, *et al.,* 2003; and Whitlock *et al.,* 2001) have used Antibiotic Resistance Analysis (ARA) to successfully classify fecal contaminants in water. ARA classification is based upon a bacteria's resistance to antibiotics of varying concentrations (known as a resistance profile), compared to a database of known samples to relate profiles to sources. ARA is one of several library dependent methods; others include Ribotyping, Pulse Field Gel Electrophoresis, and Carbon Utilization Profiles. ARA is based on Multiple Antibiotic Resistance Profiles, which also relies on resistance patterns but without varying concentrations (Parveen *et al*., 1997). Both methods rely on the assumption that some animals have higher resistance to antibiotics (poultry, humans) because of frequent exposure, therefore there is "strong selection pressure on their fecal bacteria to become resistant" (Wiggins *et al.,* 1999). This assumption is used to distinguish between animals that have low resistance (wildlife, cattle) because of little to no exposure. Wiggins (1996) and Hagedorn *et al* (1999) tested hundreds of antibiotics at varying concentrations to develop a combination that, when used to classify a known source library, gave

adequate separation of each category. Wiggins (1996) ultimately settled on the five antibiotics reported in his study because of their generalized use in humans and animals.

The first step when starting an ARA project is to select a target organism. The target organism is the bacterium that will be isolated from feces and water and classified based on its resistance patterns. *Enterococcus* and *E. coli* have been the most widely used target organisms, primarily because they are the indicators approved for use by the Environmental Protection Agency (EPA) (Simpson *et al*.,2003). *E.coli,* a member of the fecal coliform group, is a short gram-negative rod which ferments lactose. *E.coli* is the most often used indicator because it correlates well with contamination in freshwater. *Enterococcus,* a gram-positive*,* is used for saltwater studies because it is able to grow at 6.5% NaCl (Holt 1994). The gram-positive Fecal Streptococcus group has also been used as ARA target organisms (Wiggins 1996, *et al.,* 1997, *et al.,* 2003). This group is especially useful because fecal coliforms are difficult to isolate from some sources such as composted animal and poultry litter (Hagedorn *et al.,* 1999). *Enterococcus spp.*, however, may provide the best all around indicator ability because they survive longer in marine environments than fecal coliforms, last longer during wastewater treatment processes, and "their numbers in recreational waters correlate with the risk of human pathogens and disease" (Harwood *et al.,* 2000).

The next step in the ARA process is to develop a source library using feces from known sources collected within the watershed being tested. This library will be the database against which all unknown water samples are compared and fecal contaminants classified. The library needs to be large enough to be representative of all possible fecal isolates within the watershed that might possibly contaminate the waterways (Hagedorn *et al.,* 1999). Library categories are gauged using rates of correct classification (RCC) values, or the rate at which a known source was correctly classified. Average rates of correct classification (ARCC), or the average of each category divided by the number of categories, is used to validate the library. Representativeness can be checked by periodically adding new source samples as unknowns to the library. If the library continues to classify them correctly, then the library is representative (Hagedorn *et al.,* 1999).

The known source library is divided into source types or categories. The number of categories is dependent on the types of sources present in the watershed. The most effective level of isolate classification occurs with two categories, human and non-human (Graves *et al.,* 2002, Carson *et al.,* 2003). The most commonly used number of categories is four, including a human, livestock, wildlife, and pet or bird category (Wiggins *et al*., 1999, Whitlock *et al.,* 2001, Hagedorn *et al.,* 1999). The greater number of categories used the lower the RCC and ARCC values tend to be (Wiggins *et al.,* 1996). It is more difficult to differentiate between types of animals, for example cattle and poultry than it is to distinguish between human and animals (Wiggins 1996).

Resistance patterns of bacteria change over time and can vary from one region or watershed to another (Wiggins *et al.,* 1999). However, larger libraries may be used to classify bacteria from different geographic regions (Hagedorn *et al.,* 1999). Before using libraries from different geographic regions (with different climates, rainfall, wildlife and or birds) to classify water samples, the library should first be tested, using source samples from the same region as the water samples (Hagedorn *et al*., 1999). For example, a library constructed from sources in the Tidewater region of Virginia is not likely to be sufficient for classification of isolates from far southwestern Virginia. Hagedorn *et al* (1999) obtained acceptable rates of classification by combing libraries from Montgomery and Clarke county Virginia. The library was continually verified by periodically adding known source samples to make sure they were classified correctly (Hagedorn *et al.,* 1999). Wiggins *et al* obtained similar results using multiple libraries from different watersheds (Wiggins *et al.,* 2003). He also discovered that bacterial profiles remained stable for at least a year (Wiggins *et al.,* 2003). Although researchers still do not know how big a library needs to be, Hagedorn *et al* (1999) recommended that each source should have at least a few hundred isolates in order to be representative. It is reasonable to say that a small watershed would require fewer sources, and a larger watershed would require more in order to attain representativeness.

ARA has been shown to be effective in identifying non-point sources of fecal contamination in both rural and urban environments. With a library of 7,058 fecal streptococcus

isolates, Hagedorn *et al* (1999) was able to obtain an ARCC value of 88% using 892 known source samples from rural Virginia. Whitlock *et al* (2002) constructed a library with 2,398 isolates and obtained an ARCC value of 69% when using it to classify samples in an urban watershed. These results are similar to those obtained from similar studies conducted using molecular methods. Using 137 human and 346 non-human known fecal isolates, Carson *et al* (2003) compared rep-PCR and Ribotyping. Although both produced adequate classification rates, 88.14% for rep-PCR and 72.78% for Ribotyping, rep-PCR was the better of the two. This study provided insight into another phenomenon associated with both molecular and biochemical method libraries. When the libraries for both rep-PCR and Ribotyping were compared as simple two-way libraries, human verses non-human, classification rates were higher. For rep-PCR, the RCC values for the eight categories ranged between 83.33% for Goose and 97.06% for human with an ARCC of 88.14%. When the categories were reduced to two, human and non-human, the RCC values were 97.06% for human and 96.24 for non-human with and ARCC of 96.65%. For Ribotyping, the RCC values ranged from 50.98% for Turkey to 87.5% for human with an ARCC of 72.78%. When the eight categories were pooled, human verses non-human, the RCC values were 87.5% for human and 86.42% for non-human with an ARCC of 89.96. All molecular BST publications, thus far, have relied on small libraries to classify isolates (Carson *et al.,* 2003, DombecDombek *et al.,* 2000, Simmons and Herbein 2002). Small libraries give a false sense of reliability because of their high ARCC values. In actuality, small libraries are not representative of the bacterial strains in the environment and can not effectively classify isolates.

Once a library has been constructed and evaluated, water samples can be collected and analyzed against the library. Although there are issues to deal with before beginning a Bacterial Source Tracking (BST) project using any source tracking method, ARA is very useful because it is relatively inexpensive, quick, easy, and allows more samples to be analyzed than any other method available (Simpson *et al*., 2003).

MATERIAL AND METHODS

Watershed Characteristics

The James River begins in the Alleghany Mountains of Botetourt county, travels 366.9 km to the fall line in Richmond and 178.6 km to Hampton Roads where it enters the Chesapeake Bay (Appendix A Figure 53). The James River watershed is over 25,899.9 km² and is Virginia's largest tributary to the Chesapeake Bay (accounting for 15% of the bays' watershed). Land in the James River basin is 65% forested, 19% is used for agricultural purposes (pasture and cropland) and 12 % is urban (developed for residential, commercial, and industrial uses) (Appendix A Figure 54).

The Appomattox River is one of the major tributaries of the James River. The Appomattox watershed is 4009.3 km² and drains 414,338 hectares of agricultural, residential, and urban land spanning 12 counties in the Piedmont and South Central Regions of Virginia (Appendix A Figure 55). The Appomattox River meets the James River the town of Hopewell, where its waters then drain to the Chesapeake Bay. This River is officially part of the James River Basin or watershed. Studying a watershed or "all the land that drains into a given body of water" (DCR 2004) as opposed to just a segment of a stream or river is important so that all non-point pollution sources can be analyzed and accounted for.

Source Sample Collection

Fecal samples from known animal and human sources were collected throughout the watershed to build a known source library. As mentioned, the Appomattox Watershed Project is a small part of a much larger Virginia TMDL project. Since this was a statewide project, managed and funded by DEQ, four standard categories were established in order to best characterize the types of fecal contamination that may be encountered. For the sake of consistency, DEQ decided that the four source categories should be Human, Livestock, Pet and Wildlife. Birds are typically given their own category, however, DEQ decided to create a Pet category and place birds into either the Livestock or Wildlife category depending on the circumstances. The watershed was rural, with few migratory birds or waterfowl near the streams, so for the Appomattox TMDL Development Study, this was a reasonable decision. The Pet category was intended for more

urban areas where pet waste is often a problem, however, DEQ wanted the classification system to be consistent throughout the state. A library should only be constructed using sources that may contribute to pollution, because adding unnecessary sources to the library will cause misclassifications.

MapTech, Inc employees collected all source samples and delivered them to the laboratory. Composite human samples were collected from port-a-johns, septic trucks, and from volunteers within the watershed. Livestock samples were collected from horse, cattle (dairy and beef), swine, and poultry farms scattered throughout the watershed. Wildlife samples consisted of otter, muskrat, deer, fox, bear, geese, raccoon, and skunk. Wildlife samples were collected while hiking though the watershed and from dead animals when available. An animal tracking book was used to identify fecal samples collected from the forests (Peterson 1998). Cat and dog samples, for the pet category, were collected from animal shelters and from parks throughout the watershed. Fecal samples were collected and stored in coolers until brought back to the lab.

A total of 1,280 individual colonies (isolates) were obtained from the four source categories. From each category, 320 isolates were analyzed (8 isolates from each of 40 samples).

Isolates from Known Sources

Known source sample isolates were obtained by taking 10.0g of solid material, or 10.0 ml of septic water (for human samples), and adding it to 90 ml of prepackaged sterile phosphate buffer water (Hardy Diagnostics, CA) creating a 1:10 dilution. This dilution was spread plated onto prepackaged m-Tek agar (Teknova, CA) in 15x 100mm dishes and incubated for 2 hours at 35°C then placed in a 44.5°C water bath for 24 hours. Individual magenta-colored colonies were picked from the m-Tech plates and placed into 96-well micro-titer pates containing 200 micro-liters of colilert broth (Idexx Laboratories, Maine). After incubating at 37°C in a sealed plastic tub containing a damp paper towel for 24 hours, the presence of *E.coli* was confirmed by checking the microtiter wells for fluorescence under ultraviolet light (fluorescence means positive for *E.coli*).

Antibiotic Resistance Analysis (ARA)

After isolation of *E.coli* from fresh fecal samples, the samples were plated onto seven antibiotics at 28 different concentrations, including a Trypticase Soy Agar (TSA) plate as a positive control. Each antibiotic was prepared from commercial powders obtained from Simga Chemical Co (St Louis, MO)(Table 1).

Table 1: Antibiotic Formulations and Preparation

Antibiotic	**Commercial Formulation**	**Solvent used for Preparation**	**Stock Concentration (mg/mL)**
Rifampacin	Rifampacin	Methanol	10
Oxytetracycline	Oxytetracycline HCl	1:1 water-methanol	10
Streptomycin	Streptomycin Sulfate	Distilled water	10
Cephalothin	Cephalothin	Distilled water	10
Erythromycin	Erythromycin	1:1 water-ethanol	10
Tetracyclin	Tetracyclin HCl	Methanol	10
Neomycin	Neomycin Sulfate	Distilled water	10

Trypticase Soy Agar (BBL Cockeysville MD) was prepared, according to label instructions, in 100ml aliquots, autoclaved at 121°C for 15 minutes, and cooled to 50°C so that the agar would not melt the plastic petri dishes. Antibiotics were added to each beaker to obtain the desired concentration (Table 2), swirled, and poured into five 15 X 100ml sterile plastic petri dishes, each 20ml aliquot having the same amount of antibiotic. Control plates, to be used as positive controls, containing only TSA were also poured. Samples were transferred from the 96 well microtiter plates to each concentration of antibiotic (28 concentrations) via a stainless steel 48 pronged replica plater, ethanol sterilized between each type of antibiotic. After the innoculant had dried on the plates, the plates were incubated in a plastic tub containing a damp paper towel at 37°C for 24 hours. Each isolate on each plate was scored for growth or no growth, as compared to the control. If no growth occurred on the control plate for a particular isolate, then all antibiotics were scored zero for that particular isolate.

Table 2: Antibiotic Concentrations

Antibiotic	Plate Concentration (ug/L)
Rifampacin	60, 75, 90
Oxytetracycline	2.5, 5.0, 7.5, 10.0, 15.0
Streptomycin	2.5, 5.0, 7.5, 10.0, 15.0
Cephalothorin	150, 250, 350
Erythromycin	60, 70, 90, 100
Tetracyclin	2.5, 5.0, 7.5, 10.0, 15.0
Neomycin	2.5, 5.0, 10.0

Statistics

Isolates were analyzed using JMP-In software (Version 5.0 for Windows, SAS Institute, Inc.). A total of 1,280 isolates were assigned to the source from which it came (Human, Livestock, Wildlife, and Pet), 320 isolates per category. Discriminate Analysis (DA) and Logistic Regression (LR) were both run on the database to place each isolate in a category based on its source and its profile as compared to all the other isolates in the library.

RESULTS AND DISCUSSION

Analysis of the known source libraries

The Appomattox River Watershed Project is one of the first BST projects to use Logistic Regression for the classification of isolates. Discriminate Analysis has been the model of choice for both molecular and biochemical methods (Wiggins 1996, Parveen *et al*., 1999, Harwood *et al*., 2000). LR was used because the DA classifications were questionable, little to no Pet and Human signatures were expected for such a rural watershed. Although many statistical methods could have been used for this comparison, LR was chosen because it is adapted to data where the dependent variable is binary, or has only two outcomes (Hicks and Turner 1999). In this case, the outcome was either growth, or no-growth, so LR was a good match. Unlike DA, LR also assumes that the data is randomly distributed as bacterial populations should be (Hicks and Turner 1999).

The RCC values, for the known source library, obtained from the Discriminate Analysis model were similar to those from the Logistic Regression model. Table 3 is a summary of the Discriminate Analysis Regular Library. For this library, the original 1,280 known source isolates were analyzed using the Discriminate Analysis model, using four source categories, without deleting any isolates. For each source category, 320 isolates were analyzed. Discriminate

Analysis produced RCC values of 99.69% for humans, 89.69% for Livestock, 90.31% for Pets, 73.13% for Wildlife and an ARCC value of 88.20%.

Table 4 is a summary of the Logistic Regression Regular Library. The original 1,280 known source isolates were analyzed using the Logistic Regression statistical model, using four source categories, without deleting any isolates. For each source category, 320 isolates were analyzed. Logistic Regression analysis produced RCC values of 100% for human, 91.25% for livestock, 91.25% for pets, and 75.94% for wildlife and a ARCC value of 89.61%.

The classification rates from the DA and LR Regular libraries are similar to those obtained by other ARA studies by Hagedorn *et al* 1999, Wiggins 1996, and to molecular method results such as those from Carson *et al* (2003). The RCC values for each model are very high for Human. This is most likely due to sampling too many isolates from a limited number of source samples.

Table 3: Discriminate Analysis Regular Library

Predicted Source	Percent and number of known source isolates assigned to each source				
	Human	Livestock	Pet	Wildlife	Total Isolates Assigned to Source
Human	**99.69** (319)	.63 (2)	.63 (2)	0	323
Livestock	0	**86.69** (287)	5 (16)	19.38 (62)	365
Pet	0	2.5 (98)	**90.31** (289)	7.5 (24)	321
Wildlife	.31 (1)	7.19 (23)	4.06 (13)	**73.13** (234)	271
Total Isolates	320	320	320	320	1280

*****ARCC 88.20%

**Numbers in Bold indicate the RCC values for that particular column, numbers in parenthesis indicate the number of isolates

Table 4: Logistic Regression Regular Library

Predicted Source	Percent and number of known source isolates assigned to each source				
	Human	Livestock	Pet	Wildlife	Total Isolates Assigned to Source
Human	**100** (320)	0	.31 (1)	0	321
Livestock	0	**91.25** (292)	4.69 (15)	18.75 (60)	367
Pet	0	.94 (3)	**91.25** (292)	5.31 (17)	312
Wildlife	0	7.81 (25)	3.75 (12)	**75.94** (243)	280
Total Isolates	320	320	320	320	1280

*ARCC 89.61%

** Numbers in Bold indicate the RCC values for that particular column, numbers in parenthesis indicate the number of isolates

Test for Library Representativeness

The Discriminate Analysis and Logistic Regression Regular libraries were each tested for representativness using an artificial clustering test. Before running each model, Type and Source were randomly assigned to each known fecal sample. A representative library with four categories should randomly classify each known source approximately 25% of the time. The Discriminate Analysis Regular library randomly classified the known fecal samples as 30.06% Human, 30.94 % Livestock, 24.69% Pet and 21.25% Wildlife (Table 5). The average between these four categories was 29%.

Table 5: Discriminate Analysis Test for Artificial Clustering

Predicted Source	Percent and number of known source isolates assigned to each source				
	Human	Livestock	Pet	Wildlife	Total Isolates Assigned to Source
Human	**39.06**(125)	34.69 (111)	33.13 (106)	34.69 (111)	453
Livestock	25.31 (81)	**30.94** (99)	24.38 (78)	24.06 (77)	335
Pet	21.25 (68)	18.13 (58)	**24.69** (79)	20.00 (64)	269
Wildlife	14.38 (46)	16.25 (52)	17.81 (57)	**21.25** (68)	223
Total Isolates	320	320	320	320	1280

* Numbers in Bold indicate the RCC values for that particular column, numbers in parenthesis indicate the number of isolates

The Logistic Regression Regular library classified the known fecal samples as 30.06% Human, 30.63% Livestock, 24.69% Pet and 21.56% Wildlife, also resulting in an average of 29% (Table 6). Both libraries resulted in classification rate averages (29%) that were close to random, so the 1,280 isolate library constructed for the Appomattox River Watershed is representative of the fecal material that may end up in the streams, at least according to this test.

Table 6: Logistic Regression Test for Artificial Clustering

Predicted Source	Percent and number of known source isolates assigned to each source				
	Human	Livestock	Pet	Wildlife	Total Isolates Assigned to Source
Human	**39.06**(125)	34.06 (109)	33.13 (106)	34.06 (109)	449
Livestock	25.31 (81)	**30.63** (98)	24.38 (78)	24.09 (77)	334
Pet	20.94 (67)	18.75 (60)	**24.69** (79)	20.31 (65)	271
Wildlife	14.69 (47)	16.56 (53)	17.81 (57)	**21.56** (69)	226
Total Isolates	320	320	320	320	1280

*Numbers in Bold indicate the RCC values for that particular column, numbers in parenthesis indicate the number of isolates

Increasing Representativeness

Although the artificial clustering test indicated that the DA and LR Regular libraries were fairly representative, extra known source samples were added to the database to try and bring the random rates of correct classification closer to 25%. Table 7 is a summary of the RCC values obtained from adding adding Human and Livestock isolates to the Logistic Regression Regular library, from a neighboring region, in an attempt to make the library more representative. Two hundred and forty isolates were added to the library and the test for artificial clustering was run on the new data set, more would have been added had they been available. As compared to Table 6 (RCC values of 36.06% for Human and 30.63% for Livestock) both the Livestock and Human categories became less representative (RCC values of 44.21% for Human and 18.95% for Livestock (Table 7), their artificial clustering RCC values with the addition of extra isolates were further away from 25%. The Wildlife and Pet were not significantly affected.

Table 7: Logistic Regression Test for Artificial Clustering with Addition of Extra Isolates

Predicted Source	Percent and number of known source isolates assigned to each source				
	Human	Livestock	Pet	Wildlife	Total Isolates Assigned to Source
Human	**44.21** (168)	36.84 (140)	38.95 (148)	36.32 (138)	594
Livestock	13.68 (52)	**18.95** (72)	13.42 (51)	16.32 (62)	237
Pet	23.95 (91)	23.68 (90)	**27.37** (104)	22.63 (86)	371
Wildlife	18.16 (69)	20.53 (78)	20.26 (77)	**24.74** (94)	318
Total Isolates	380	380	380	380	1520

*Numbers in Bold indicate the RCC values for that particular column, numbers in parenthesis indicate the number of isolates

Comparision of Discriminate Analysis and Logistic Regression for Sample Classification

Researchers have typically used the Discriminate Analysis statistical model to classify unknown isolates (Hagedorn *et al.,* 1999, *et al.,* 2003; Wiggins 1996, *et al.,* 1999, *et al.,* 2003; Whitlock *et al.,* 2002). In order to determine which model was the most appropriate model to use for the Appomattox TMDL Development Study, several variations of the Discriminate Analysis

and Logistic Regression libraries were compared: DA and LR Regular libraries, DA and LR 80% Delete Libraries, and DA and LR with Unknown category libraries. All water samples, whose source of fecal contamination was unknown, were classified using each of the six libraries. Ten water samples were randomly selected to test the six library variations. **Group A** (Table 8) consisted of five water samples that produced significantly different source classifications when analyzed by the DA and LR Regular library models (human signature for one model /no human signature for the other) (Table 10 and 11). **Group B** (Table 9) consisted of five water samples that produced source classifications that were the same for the DA and LR Regular library models (Table 12).

Table 8: **Group A Water Samples**

Lab ID	Station ID	River/Stream
D191	2APP110.93	Appomattox River
D411	2APP012.79	Appomattox River
D824	2APP050.23	Appomattox River
D1017	2APP110.93	Appomattox River
D2093	2ANG003.35	Angola Creek

* Group A: Source Classifications were different for DA and LR

Table 9:Group B Water Samples

Lab ID	Station ID	River/Stream
D398	2ANG001.27	Angola Creek
D1005	2ANG003.35	Angola Creek
D1550	2APP085.85	Appomattox River
D2091	2ANG001.27	Angola Creek
D2293	2APP118.04	Appomattox River

*Group B: Source Classifications were the same for DA and LR

When Group A was analyzed using the DA Regular library, four out of five samples had no human signature (D824 being the only on that did) (Table 10). The exact opposite is true with the LR Regular library, four out of the five samples had a human signature with an average between the samples of 21.4% Human (Table 11). The DA Regular library averaged the remaining isolates as 31.3% Livestock, 36.3% Pet and 23.3 % Wildlife. The LR Regular library averaged the remaining isolates as 40.6% Livestock, 20.00% Pet, and 18.2% Wildlife. Group B were classified the same by the DA and LR Regular library model (Table 12).

Table 10: Discriminate Analysis Regular Library with Group A Samples

Count Col %	D1017	D191	D2093	D411	D824	Total Isolates	% Average
Human	0 **0**	0 **0**	0 **0**	0 **0**	11 **45.83**	11	9.1
Livestock	8 **33.33**	4 **16.67**	8 **33.33**	9 **37.50**	9 **37.50**	38	31.3
Pets	11 **45.83**	7 **29.17**	12 **50.00**	14 **58.33**	0 **0.00**	44	36.3
Wildlife	5 **20.83**	13 **54.17**	4 **16.67**	1 **4.17**	4 **16.67**	28	23.3
	24	24	24	24	24	120	100

* First number in each box in the number of isolates, number in bold is the percent of isolates classified as the source indicated in the corresponding row

Table 11: Logistic Regression Regular Library with Group A Samples

Count Col %	D1017	D191	D2093	D411	D824	Total Isolates	% Average
Human	7 **29.17**	5 **20.83**	7 **29.17**	6 **25.00**	0 **0.00**	26	21.4
Livestock	8 **33.33**	2 **8.33**	9 **37.50**	10 **41.67**	20 **83.33**	49	40.6
Pets	7 **29.17**	7 **29.17**	5 **20.83**	5 **20.83**	0 **0.00**	24	20
Wildlife	2 **8.33**	10 **41.67**	3 **12.50**	3 **12.50**	4 **16.67**	22	18.2
	24	24	24	24	24	120	100

* First number in each box in the number of isolates, number in bold is the percent of isolates classified as the source indicated in the corresponding row

Table 12: Discriminate Analysis and Logistic Regression Regular Library with Group B Samples

Count Col %	D1005	D1550	D2091	D2293	D398	Total Isolates	% Average
Human	0 **0.00**	0 **0.00**	0 **0.00**	0 **0.00**	0 **0.00**	0	0
Livestock	24 **100.00**	24 **100.00**	24 **100.00**	4 **50.00**	24 **100.00**	100	96
Pets	0 **0.00**	0 **0.00**	0 **0.00**	4 **50.00**	0 **0.00**	4	4
Wildlife	0 **0.00**	0 **0.00**	0 **0.00**	0 **0.00**	0 **0.00**	0	0
	24	24	24	8	24	104	

* First number in each box in the number of isolates, number in bold is the percent of isolates classified as the source indicated in the corresponding row

<u>80% Delete DA Library Model</u>

The first variation of the DA and LR Regular libraries was to delete any known source samples whose probability of correct classification was below 80%, as suggested by Simmons and Herbein (2002), and analyze the remaining isolates using the DA and LR models. This was intended to disregard any isolates whose classification or association to a pattern in the library was weak. For Group A, the 80% Delete DA library produced average values for Pet (40.1%) (Table 13) that were higher than the Pet average (36.3%) from the Regular DA library (Table 10). The Wildlife average (19.7%) was slightly lower than that obtained from the DA Regular Library

(23.3%). The Human signature was between 8-9% for both the DA Regular and the 80% Delete DA library. Only 13 Group A isolates were deleted from the library. For the Group B samples, there was essentially no change when using the 80% Delete DA library, greater than 97% of the isolates were classified as Livestock (Table 14). Twenty-two Group B isolates were deleted from the library.

Table 13: 80% Delete Library Discriminate Analysis with Group A Samples

Count Col %	D1017	D191	D2093	D411	D824	Total Isolates	% Average
Human	0 **0.00**	0 **0.00**	0 **0.00**	0 **0.00**	9 **45.00**	9	8.4
Livestock	8 **36.36**	2 **9.52**	7 **31.82**	8 **36.36**	9 **45.00**	34	31.8
Pets	11 **50.00**	7 **33.33**	12 **54.55**	13 **59.09**	**0** **0.00**	43	40.1
Wildlife	3 **13.64**	12 **57.14**	3 **13.64**	1 **4.55**	2 **10.00**	21	19.7
	22	21	22	22	20	107	100

* First number in each box in the number of isolates, number in bold is the percent of isolates classified as the source indicated in the corresponding row

Table14: 80% Delete Library Discriminate Analysis with Group B Samples

Count Col %	D1005	D1550	D2091	D2293	D398	Total Isolates	% Average
Human	0 **0.00**	0 **0.00**	0 **0.00**	0 **0.00**	0 **0.00**	0	0
Livestock	24 **100.00**	24 **100.00**	24 **100.00**	3 **50.00**	20 **100.00**	95	97
Pets	0 **0.00**	0 **0.00**	0 **0.00**	3 **50.00**	0 **0.00**	3	3
Wildlife	0 **0.00**	0 **0.00**	0 **0.00**	0 **0.00**	0 **0.00**	0	0
	24	24	24	6	20	98	100

* First number in each box in the number of isolates, number in bold is the percent of isolates classified as the source indicated in the corresponding row

<u>80% Delete LR Library Model</u>

For the Group A water samples, The 80% Delete LR library produced averages of 25.6% for Human (lower than LR Regular Library of 21.4%) 40.5% for Livestock (approximately the same as than Regular LR Library-40.6%), 18% for Pet (lower than the LR Regular Library-20%) and 15.9% for Wildlife (slightly higher than the LR Regular Library-18.2%) (Table 15). Twenty-Six Group A isolates were deleted from the library. For Group B, four out of five remained the same (Table 16), only D2293 had isolates with classifications other than Livestock. Forty-Four Group B isolates were deleted from the library. Logistic Regression is a more stringent statistical model, requiring that the association between the known source patterns and the unknown pattern be

more similar in order for a classification to be made. This results in a higher probability of correct classification.

Table15: 80% Delete Library Logistic Regression Group A Samples

Count Col %	D1017	D191	D2093	D411	D824	Total Isolates	% Average
Human	7 **50.00**	5 **23.81**	7 **43.75**	5 **26.32**	0 **0.00**	24	25.6
Livestock	2 **14.29**	2 **9.52**	5 **31.25**	9 **47.37**	20 **83.33**	38	40.5
Pets	5 **35.71**	6 **28.57**	3 **18.75**	3 **15.79**	0 **0.00**	17	18
Wildlife	0 **0.00**	8 **38.10**	1 **6.25**	2 **10.53**	4 **16.67**	15	15.9
	14	21	16	19	24	94	100

* First number in each box in the number of isolates, number in bold is the percent of isolates classified as the source indicated in the corresponding row

Table 16: 80% Delete Library Logistic Regression Group B Samples

Count Col %	D1005	D1550	D2091	D2293	D398	Total Isolates	% Average
Human	0 **0.00**	0 **0.00**	0 **0.00**	1 **16.67**	0 **0.00**	1	1.5
Livestock	15 **100.00**	7 **100.00**	22 **100.00**	4 **66.67**	10 **100.00**	58	97
Pets	0 **0.00**	0 **0.00**	0 **0.00**	1 **16.67**	0 **0.00**	1	1.5
Wildlife	0 **0.00**	0 **0.00**	0 **0.00**	0 **0.00**	0 **0.00**	0	0
	15	7	22	6	10	60	100

* First number in each box in the number of isolates, number in bold is the percent of isolates classified as the source indicated in the corresponding row

Comparison of 80%Delete DA and LR library Results

The results for Group A obtained from each 80% library differed considerably for the Human and Pet categories. The 80% Delete DA library classified 8.4% of the isolates as Human and 40.1% of the isolates as Pet. The 80% Delete LR library classified 25.6% of the isolates as Human and only 18% of the isolates as Pet. The Livestock values varied slightly between models (DA-31.8%) (LR-40.5%), as did the Wildlife (DA-19.7%) (LR-15.9%). For Group A, more isolates were deleted by the 80% Delete Logistic Regression library (22) than with the 80% Delete Discriminate Analysis library (13). For Group B more isolates were deleted by the 80% Delete Logistic Regression library (44) than with the 80% Delete Discriminate Analysis library (22). The source classification were almost identical for Group B isolates when classified by each library. The results obtained from the 80% Delete LR library are more likely because isolates with weak associations were thrown out in greater numbers than with the 80% Delete DA library. The 80% Delete DA library was not as selective as the 80% Delete LR library.

Libraries with Unknown Category

The final variation of the DA and LR Regular libraries was to place known sources whose probabilities fell below 80% into an unknown category, creating a fifth category, instead of deleting them. This allowed the isolates to be classified only if there was a strong correlation between their pattern and an existing pattern in the database. Creating an unknown category prevents weak classifications from being made and allows greater confidence in the classifications that are made.

The DA Library with Unknown Category

DA library (Table 17) with an unknown category resulted in no human signature for 4 out of 5 of the Group A samples, the same as the DA Regular library. As compared to the DA Regular library, the Group A sample averages were 28.4% Livestock (lower), 35.8% Pet (lower) and 17.6% Wildlife was (lower) (Table 17). The unknown category accounted for 9.1% of the isolates. As compared to the 80% Delete DA library, the Unknown Category DA library Group A samples resulted in a higher percentage for Human (the same number of isolates for each), and lower for all other categories. The Unknown Category DA results for Group B samples were virtually identical to those obtained from the DA Regular library and the 80% Delete DA library, two samples deviated by an isolate or two (Table 11, 14, 18). Only 11 Group A and 10 of Group B isolates were classified as unknown.

Table 17: Unknown Category Discriminate Analysis Library with Group A Samples

Count Col %	D1017	D191	D2093	D411	D824	Total Isolates	% Average
Human	0 **0.00**	0 **0.00**	0 **0.00**	0 **0.00**	11 **45.83**	11	9.1
Livestock	6 **25.00**	3 **12.50**	7 **29.17**	9 **37.50**	9 **37.50**	34	28.4
Pets	11 **45.83**	7 **29.17**	12 **50.00**	13 **54.17**	0 **0.00**	43	35.8
Unknown	7 **29.17**	1 **4.17**	2 **8.33**	0 **0.00**	1 **4.17**	11	9.1
Wildlife	0 **0.00**	13 **54.17**	3 **12.50**	2 **8.33**	3 **12.50**	21	17.6
	24	24	24	24	24	120	100

* First number in each box in the number of isolates, number in bold is the percent of isolates classified as the source indicated in the corresponding row

Table 18: Unknown Category Discriminate Analysis Library with Group B Samples

Count Col %	D1005	D1550	D2091	D2293	D398	Total Isolates	% Average
Human	0 **0.00**	0 **0.00**	0 **0.00**	0 **0.00**	0 **0.00**	0	0
Livestock	23 **95.83**	24 **100.00**	24 **100.00**	4 **50.00**	15 **62.50**	90	86.5
Pets	0 **0.00**	0 **0.00**	0 **0.00**	4 **50.00**	0 **0.00**	4	4
Unknown	1 **4.17**	0 **0.00**	0 **0.00**	0 **0.00**	9 **37.50**	10	9.5
Wildlife	0 **0.00**	0 **0.00**	0 **0.00**	0 **0.00**	0 **0.00**	0	
	24	24	24	8	24	104	

* First number in each box in the number of isolates, number in bold is the percent of isolates classified as the source indicated in the corresponding row

LR Library with Unknown Category

The LR library (Table 19) with an unknown category resulted in a slight human signature in two Group A samples (average of 1.6%), as opposed to 20.83% from the Regular LR library and 25.6% in the 80% Delete LR library. As compared to the LR Regular and the 80% Delete libraries, the Livestock average was 23.33% (lower), the Pet average 10% (lower), and the Wildlife average was 36.67% (higher). The unknown category accounted for 28.33% of the isolates.

The Group B samples were averaged as 53.8% Livestock and 44.23% Unknown (Table 20), significantly different from the LR Regular and the 80% Delete libraries where the isolates were classified as at least 96% Livestock. Thirty-Four Group A isolates were classified as unknown, and forty-six from Group B.

Table 19: Unknown Category Logistic Regression Library with Group A Samples

Count Col %	D1017	D191	D2093	D411	D824	Total Isolates	% Average
Human	0 **0.00**	0 **0.00**	1 **4.17**	1 **4.17**	0 **0.00**	2	1.6
Livestock	1 **4.17**	6 **25.00**	2 **8.33**	3 **12.50**	16 **66.67**	28	23.3
Pets	0 **0.00**	4 **16.67**	2 **8.33**	6 **25.00**	0 **0.00**	12	10
Unknown	22 **91.67**	1 **4.17**	8 **33.33**	2 **8.33**	1 **4.17**	34	28.3
Wildlife	**1** **4.17**	13 **54.17**	11 **45.83**	12 **50.00**	7 **29.17**	44	36.6
	24	24	24	24	24	120	100

* First number in each box in the number of isolates, number in bold is the percent of isolates classified as the source indicated in the corresponding row

Table 20: Unknown Category Logistic Regression Library with Group B Samples

Count Col %	D1005	D1550	D2091	D2293	D398	Total Isolates	% Average
Human	0 **0.00**	0 **0.00**	0 **0.00**	0 **0.00**	0 **0.00**	0	0
Livestock	13 **54.17**	7 **29.17**	22 **91.67**	4 **50.00**	10 **41.67**	56	53.8
Pets	0 **0.00**	0 **0.00**	0 **0.00**	1 **12.50**	0 **0.00**	1	1
Unknown	11 **45.83**	17 **70.83**	2 **8.33**	2 **25.00**	14 **58.33**	46	44.23
Wildlife	0 **0.00**	0 **0.00**	0 **0.00**	1 **12.50**	0 **0.00**	1	1
	24	24	24	8	24	104	

* First number in each box in the number of isolates, numbers in bold are the percent of isolates classified as the source indicated in the corresponding row

Comparison of DA and LR with Libraries with Unknown Category

The DA library with Unknown category classified 9.1% of the Group A isolates as Human, 28.4% as Livestock, 35.8% as Pets, 9.1% as Unknown, and 17.6% as Wildlife. The LR library with Unknown category classified 1.6% of the Group A isolates as Human, 23.3% as Livestock, 10% as Pet, 28.3% as Unknown, and 36.6% as Wildlife. More isolates in Group A were classified as Unknown by the LR library with Unknown category (34) than with the DA library with Unknown category (11), as with the Group B samples LR classified 46 isolates as Unknown and DA classified 10 as Unknown.

LOGISTIC REGRESSION LIBRARY WITH UNKNOWN CATEGORY FOR ARA CLASSIFICATION

By comparing the land use for the Appomattox watershed and the isolate classifications for the unknown samples, the LR library with unknown category appears to be the most appropriate library to classify the Appomattox water samples. As with the Regular DA and Regular LR libraries, the RCC values for each category are very high (Table 10, 11, 12). This is most likely due to sampling too many isolates from a given source sample, thereby having fewer original or unique patterns. The LR library with unknown category ARCC is 97.03%, and the RCC values are 100% for Human, 99.51% for Livestock, 96.61% for Pets, 91.32% for Unknown, and 97.72% for Wildlife (Table 21).

Table 21: Unknown Category Logistic Regression Library used for Sample Classification

Predicted Source	Percent and number of known source isolates assigned to each source					
	Human	Livestock	Pets	Unknown	Wildlife	Total Isolates Assigned to Source
Human	320 **100.00**	0	0.34 (1)	0	0	321
Livestock	0	**99.51 (203)**	1.69 (5)	3.72 (9)	0	217
Pet	0	0	**96.61** (285)	4.13 (10)	0.91 (2)	297
Unknown	0	0.49 (1)	0.68 (2)	**91.32** (221)	1.37 (3)	227
Wildlife	0	0	0.689 (2)	0.83 (2)	**97.72** (214)	218
Total Isolates	320	204	295	242	219	1280

*ARCC 97.03%
**Numbers in Bold indicate the RCC value for that particular column

Summary

Based on the artificial clustering test, the isolates used to construct the Appomattox watershed DA and LR Regular libraries were representative. ARCC values from both models were greater than 88%, similar to other ARA studies from rural watersheds. Wiggins *et al* (1999) constructed four fecal streptococcus libraries, each with four-way classification, from various rural watersheds. ARCC values were 64% for library 1 (5,990 isolates), 66% for library 2 (2,635 isolates), 65% for library 3 (2,844 isolates) and 78% for library 4 (3,032). Graves *et al* (2002) analyzed 1,174 enterococcus isolates in rural Millwood VA. With three-way classification, the ARCC value was 92.02%. Although *E.coli* was the target organism for the Appomattox watershed, the studies sited are comparable as shown by Harwood *et al* (2000) in a comparison of Fecal Streptococcus and fecal coliforms (*E.coli*) libraries for the identification of isolates. The ARCC value for Fecal Streptococcus was 62.3% and 63.9% using fecal coliforms (Harwood *et al* 2000). Although the larger libraries are most likely representative of the strains present in the environment, they have lower ARCC values than the smaller libraries. Wiggins *et al* (2003) explains this by saying that “the more isolates of each source type that are contained in the library (i.e. the more representative it is), the greater the chance they will vary in their resistance patterns”. Another excellent test of library representativeness is to calculate the number of unique known source patterns in the database. Using a SAS program developmed by Dr. Bruce Wiggins, the number of unique resistance patterns for the Appomattox River Watershed project was

calculated to be 371 (original library 1,280 patterns generated from 1,280 isolates). Although some repetition of patterns in a library in normal, the number for the Appomattox River Watershed is high and does suggest that the library was not representative. Fortunately, by using the LR library with Unknown category for classification of isolates that is taken care of.

After comparing each library and its variations (Regular DA/ LR, 80% Delete DA/LR, Unknown DA/LR), the library that produced the most logical results was the Logistic Regression library with unknown category. Because the Appomattox River watershed is 65% forested and 19% of the land is used for agricultural purposes, the high Pet signatures that resulted from each of the Group A DA libraries (Regular 36.6%, 80% Delete 49.39%, and Unknown 35.83%), as well as the two other Group A LR libraries (Regular 20.00%, 80% Delete 19.76%), are unlikely. Similarly, it also appears unlikely that these water samples would have the high human signatures produced by the Group A Regular LR (20.83%) and the Group A 80% Delete LR (32.11%) libraries. Due to these concerns, the LR library with unknown category is the most appropriate to analyze the Appomattox water samples because the source classifications most closely fit the land use.

This project is an important advance in the science of Bacterial Source Tracking, because of the use of Logisitc Regression as a statistical model for the classification of isolates and for the use of an Unknown category. As mentioned, Discriminate Analysis has been the statistical model of choice for classifying isolates. In this study, it was shown that Logistic Regression is also an adequate model for classification. Not only is Logistic Regression a suitable statistical model, when coupled with an unknown category it provides a more stringent classification of isolates. Misclassification of isolates is a fundamental concern in Bacterial Source Tracking, which may lead to inaccurate contamination identification. Misclassification may occur if the known source library is not representative of the fecal pollution in the watershed. The Unknown category can also help deal with issues of library representativness. By using the Unknown category, unknown isolates whose resistance pattern is not represented in the library patterns will not be classified as an identifiable source. This provides a better alternative to allowing these isolates to be classified incorrectly. As seen using the LR library with Unknown

category, 30% of the Group A isolates were unable to be classified. This provides a more realistic answer than reporting the source of pollution as 30% Pet in a rural heavily forested watershed. Although this data was unavailable at the time that the Appomattox River TMDL was developed, these results should be helpful in the development of the TMDL implementation plan. Using the unknown category from the beginning would have prevented the misclassification of isolates and prohibited the development of improper Total Maximum Daily Loads for the Appomattox watershed.

REFERENCES

American Public Health Association. 1995. Standard methods for the examination of water and wastewater, 19 th ed. American Public Health Association, Washington, D.C.

Baron, Samuel. 1991. Medical Microbiology, 4th edition. University of Texas Medical Branch at Galvaston.

Carson, A. C., B. L. Shear, M. L. Ellersieck, and J. D. Schnell. 2003. Comparison of Ribotyping and Repetative Extragenic palindroic-PCR for identification of fecal Escherichia coli from human and non-humans. Applied Environmental Microbiology. 69: 1836-1839.

Department of Conservation and Recreation. 2004b. Environmental Education. http://www.dcr.state.va.us [Online] Last accessed 20 April 2004.

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 non-human and animal sources. Applied and Environmental Microbiology. 66:2572-2577.

Hicks, C., K. Turner. 1999. Fundamental Concepts in the Design of Experiments, 5th edition. Oxford University Press, New York, New York.

Holt, J. G. (ed). 1994. Bergey'e manual of determinative bacteriology, 9th ed. Williams & Wilkins, Baltimore, Md.

James River Association. 2002. State of the James. 52 pages. James River Association, Mechanicsville, Va.

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

Hagedorn, C., J. B. Crozier. K. A. Mentz, A. M. Booth, A. K. Graves, N. J. Nelson, and R. B. Reneau, Jr. 2003. Carbon cources untilization profiles as a method to identify sources of fecal pollution in water. Journal of Applied Microbiology. 94: 792-799.

Harwood, Valerie J., Whitlock, John, and Withington, Victoria Classification of Antibiotic Resistance Patterns of Indicator Bacteria by Discriminant Analysis: Use in Predicting the Source of Fecal Contamination in Subtropical Waters. Applied and Environmental Microbiology. 2000. 66: 3698-3704.

Parveen, S., R. L. Murphee, L. Edmiston, C. W Kasper, K. M. Portier, and M. L. Tamplin. 1997. Association of multiple-antibiotic-resistance profiles with point and nonpoint sources of Escherichia coli in Apalachicola Bay. Applied Environmental Microbiology. 63; 2607-2612.

Parveen, Salina, Portier, Kenneth M., Robinson, Kevin, Edmiston, Lee, Tamplin, Mark L. Discriminant Analysis of Ribotype Profiles of Escherichia coli for Differentiating Human and Nonhuman Sources of Fecal Pollution Appl. Environ. Microbiol. (1999) 65: 3142-3147

Peterson, Roger. A guide to Animal Tracks. 1998. Houghton Mifflin Company. New York, New York.

Simmons, G. M., Jr., D. F. Waye, S. Herbein, S.Myers, and E. Walker. 2002. Estimating nonpoint source fecal cliorm sources using DNA profile analysis, p. 143-168. In T Younos (ed). Advances in water monitoring research. Water Resources Publications, LLC, Denver, Colorado.

Simpson, J.M, J. W Santo Domingo, and D. J Reasoner. 2002. Microbial Source Tracking: Sate of the Science. Environmental Science and Technology. 36:5279-5288.

Whitlock, J. E., D. T. Jones, and V. J Harwood. 2002. Identification of the sources of fecal coliforms in and urban watershed using antibiotic reistance analysis. Water Research. 36:4273-4282

Wiggins, B. A. 1996. Discrimminant 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., R. W. Andrews, R. A. Conway, C. L. Corr, E. J. Dobratz, D. P. Dougherty, J. R. Eppard, S. Rknupp, M C. Limjoco, J. M. Mettenburg, J. M. Rinehardt, J. Sonsino, R. L. Torrijos, and M. E. Zimmerman. 1999. Use of antibiotic resistance analysis to indentify nonpoint sources of fcal pollution. Applied Environmental Microbiology. 65:3483-3486.

Wiggins, B. A., P. W. Cash, W. S Creamer, S. E. Dart, P. P. Garcia, T. M. Gerecke, J. Hahn, B. L. Henry, K. B. Hoover, E. L Johnson, K. C. Jones, G. G. McCarthy, J. A. McDonough, S. A Merces, M. J. Noto, H. Park, M. S. Phillips, S. M. Perner, B. M. Smith, E. N. Stevens, and A. K. Varner. 2003. Use of Antibiotic Resistance Analysis for representativeness testing of multiwatershed libraries. Applied Environmental Microbiology. 69: 3399-3405.