Microbial Source Tracking in a Mixed Use Watershed in Northern Virginia

Gregory D. Touchton

Thesis submitted to the faculty of the Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

Master of Science
In
Environmental Science and Engineering

Dr. Charles Hagedorn
Dr. Brooks Crozier
Dr. Carl Zipper

August 1, 2005
Blacksburg, Virginia

Keywords: (keywords) Microbial Source Tracking, *Enterococcus*, *E. coli*, PFGE, Prince William County, Water Quality

Microbial Source Tracking in a Mixed Use Watershed in Northern Virginia

Gregory Touchton

Abstract

Prince William County, located in the rapidly developing Northern Virginia region, contains watersheds of mixed rural and urban/suburban uses. As part of Virginia regulations, recreational waters must be tested and remain under a certain standard for levels of fecal indicator bacteria (FIB). The sources of fecal pollution in neighboring watersheds within the county were determined over the 12 months previous to this project by performing Antibiotic Resistance Analysis (ARA, a microbial source tracking protocol) on *Enterococcus* and *Escherichia coli* (*E. coli*). This study indicated that multiple sources of pollution were present at all sampling locations and that the dominant sources of contamination were related to the land-use patterns and human activities that were adjacent to each location.

The goal of the current project was to monitor and identify the sources of fecal pollution in eight streams in the Occoquan Basin (OQB) that have been classified as impaired waters due to high *E. coli* concentrations. Project objectives were i) employ microbial source tracking technology to identify the categories of sources that were responsible for the bacterial impairments; ii) develop and analyze appropriate Known Source Libraries (KSL's) to determine the best design for identifying the sources of water-sample isolates; and iii) evaluate the use of optical brighteners in freshwater by fluorometry as an indicator for human-origin pollution. One site on each of six streams and two sites on the remaining two (ten total) were selected for *E. coli* and *Enterococcus*

monitoring and microbial source tracking. Repeated sampling of the ten locations for thirteen months assessed the concentrations of the bacteria over time, while comparison of monthly bacterial concentrations to the U.S. standards was used to verify the impaired water designation.

Three thousand, four hundred and eighty-eight *Enterococcus* and 969 *E. coli* water-sample isolates were collected and evaluated to determine their sources. These isolates were compared to several known source libraries (KSL's) comprised of host-origin isolates collected from the Northern Virginia region. Linear discriminant analysis (LDA) using a KSL of unique isolates determined wildlife were the dominant source of fecal pollution. Results based on ARA were cross-validated through fluorometry of the water samples (to detect optical brighteners in detergents as human-derived pollution) and pulsed-field gel electrophoresis (PFGE, a DNA fingerprinting technique) of select *E. coli* isolates. In order to determine the best method to classify the water-sample isolates, variation in antibiotic resistance data representation, known source isolate inclusion, and LDA processing were compared. The KSL that used the most antibiotic resistance datapoints, contained no conflicting data, and performed most of the parameters associated with standard LDA, classified water-sample isolates the most successfully. This project involves the first thorough testing of fluorometry for the detection of human signatures in freshwaters.

Monitoring results showed consistent *Enterococcus* and *E. coli* contamination in all eight streams, demonstrating that each had been correctly placed on Virginia's impaired waters list by state regulatory agencies. Counts between *Enterococcus* and *E. coli* did not correlate well, although concentrations of both indicator organisms were

higher during dry months. Source tracking results determined a dominant wildlife signature at all sites. Few *Enterococcus* water-source isolates were classified as human and fluorescence at all sites was consistently low. KSL's with antibiotic resistance data represented as binary values classified isolates the best. Removal of conflicting isolates improved the KSL's rate of correct classification (RCC). Creation of an unknown category, clustering of the KSL, and only accepting results above a threshold did not appreciably improve the RCC.

The KSL with the binary representation was not used to classify isolates because it violates the normal distribution assumption of LDA. Differences in the results of *Enterococcus* and *E. coli* source classifications indicated that contributing sources vary in frequency. Human fecal matter was shown to be of little concern because both *Enterococcus* ARA and fluorometry indicated low presence. The positive predictive value (PPV) statistic was found to be preferable to the minimum detectable percentage (MDP) because it does not depend on KSL size. Establishing confidence intervals to determine completeness of KSL allows one to determine whether particular methods to refine the KSL will be helpful.

This project was successfully completed and the monitored streams were correctly identified by state authorities as impaired waters. Source tracking results often conflicted, although wildlife and pets were indicated as the major sources of impairment by ARA. More local source samples need to be taken to verify this result. The best ARA library design used only unique isolates, all pattern data points, and removed conflicting isolates. Continuing examination of the representation of library data as binary is necessary to determine whether the statistical assumptions in LDA prevent meaningful results.

Evaluation of fluorometry was partially successful as the absence of "hotspots" of high fluorescent brighteners agreed with ARA results that indicated little contamination form human sources. The fluorometer continues to have potential as a metric of waste in freshwater although more work needs to be done to fully prove its utility.

Table of Contents

List of Tables

List of Figures

Author's Acknowledgements

The author would like to thank Dr. Charles Hagedorn for all his help and guidance in this work. It would not be possible without him. He made Price 401 an excellent research environment. Further, the suggestions and guidance of committee members Dr. Brooks Crozier and Dr. Carl Zipper were quite beneficial.

Annie Hassall, Mike Saluta and Jay "Dr." Dickerson were a pleasure to work with. Their intellectual and elbow grease contributions to this project helped bring it to fruition. Justin Evanylo and Paul Youmans were a big help, particularly when the June crunch time began. I would like to thank *all* my co-workers for both their food and humor.

Of course my family and friends need a mention. However, I do believe that if you are reading this, you know well enough who you are. Therefore your names will remain anonymous.

Lastly, thank you, Uwe Kirste of Prince William County Public Works for funding this project, and to the employees of the Department for all their help, especially Mrs. Patty Dietz.

Chapter 1. Literature Review

I. Project Justification

Public health initiatives throughout the US have been steadily growing since the early 1900's. Starting with the 1848 Drug Importation Act, the federal government has been working to ensure a healthful environment for its people. In 1972 Congress passed legislation that would later become known as the Clean Water Act (CWA). The CWA required promulgation of rules and standards toward the maintenance of integrity of the nation's waters. The goal was to eliminate pollution discharge by 1985.

The standards of the CWA were set on both the federal and state levels. The federal standards were the minimum levels, below which no state could go. The states were then allowed to create standards more stringent than federal levels, but were not required to. Virginia has elected to maintain standards at the levels set by the federal government.

Pursuant to these standards, waters of the state had to be classified into one of three designations (CWA 303 (c)). Waters could be listed as protected, allowed to accept no new pollution; designated use, such as swimming and fishing; or of minimum historical standards. The waters of this study are of designated use and are therefore required to maintain such standards as can allow swimming and fishing.

In July 1992 the EPA promulgated rules to report waters not in compliance with water quality standards (40 CFR 130.7). In 1997, Virginia enacted the Water Quality Monitoring Information and Restoration Act (62.1-44.19:4- 19:8 Code of Virginia). Regular assessment of the waters is required under these rules. In Virginia, a schedule

has been set for assessment of waters every 5 years unless impaired. Impaired waters are those that fail standards 10% or more of the time. As of 2002 442 of 494 state identified watersheds had at least one impairment (VA DEQ 2004). Impairment can occur for numerous reasons (Simpson et. al. 2002), but pathogenic microbe contamination is most important for waters used in human recreation, drinking, and aquaculture. Specific pathogen monitoring can be used to assess impairment (Simpson et. al. 2002). Current research indicates that monitoring efforts fall below necessary levels to ensure public health (Whitman and Nevers to be Published).

In a 1999 legal case, the American Canoe Association vs. US EPA, the EPA was required through consent decree to enact the Total Maximum Daily Load (TMDL) program on impaired waters as required in CWA sec. 303 (d). The TMDL program is a pollution reduction program for impaired waters that do not meet the standards set by their designated use. The TMDL is the maximum amount of pollution the water body can receive and still meet acceptable standards (Simpson et. al. 2002). For 27 years the EPA took little or no action on TMDL's in Virginia but was "jump-started" by the consent decree to require all impaired waters of Virginia to have TMDL implementation within 11 years. Waters found to be impaired during this time were required to have TMDL implementation within 5-6 years. An estimated 648 TMDL's are due by 2010 (Virginia DEQ 2005)

In designated swimmable/fishable waters the water quality standards are designed for regular human contact. These standards are therefore dependent on public health levels of 8 gastrointestinal illnesses per 1000 swimmers (USEPA 1986). Any disease causing organisms can interact with swimmers and fishermen, and microbial levels are

expected to be kept to a minimum. Two EPA fecal indicator organisms are *Escherichia coli* (*E. coli*) and *Enterococcus*. The following discussion applies equally to both.

Unfortunately, this presents a further complication. Indicator levels require 24 hours or more to detect under standard methods. Therefore public health can not be safely protected by closing waters upon the detection of high fecal indicators, in such a case, officials would be 24 hours too late. Consequently, public health protection efforts have required the maintenance of the limit. The public health levels for these waters require freshwater *E. coli* to be kept below 235 colony forming units per 100 mL (CFU/100ml) and *Enterococcus* below 104.

Of the 11 year consent decree implementation time, 6 years have already passed. Counties within Virginia are being encouraged to create the TMDL plans for pollution reduction. However, reduction plans require knowledge of the pollution sources. In the case of the bacterial pollution commonly found throughout Virginia, its existence has one known source: fecal matter. Urban runoff, broken sewers and failed septic systems often contain microbe levels that far exceed TMDL limits (USEPA 2005). However, high levels of fecal indicators are not always attributable to these areas. Sources such as wildlife, including deer, raccoons, and waterfowl also play a significant role (USEPA 2005). Agricultural operations, both crop and animal husbandry, can have fecal runoff. Hence, the waters containing fecal indicators require knowledge of the source of fecal pollution influencing the water before pollution can be reduced.

The waters of this study are particular important for more than their designated use. The waters flow into two further environments: a reservoir for drinking water and a "multiple use, critical habitat", the Chesapeake Bay (Chesapeake Bay Program 2005).

For both these areas minimal pollution is desired. While ambient fecal indicator levels do not heavily influence the treatment of the drinking water, they may indicate more resistant pathogens able to survive the disinfection process. The Chesapeake Bay has a falling shellfish trade, less than 1% of historical levels. Shellfish harvests require that beds have less than moderate levels fecal bacteria (USDA 1995).

Source tracking of the Occoquan Basin watershed will have a strong impact on the Northern Virginia region because it will be used to design cleanup strategies, best management practices (BMP's). Establishment of BMP's is a required procedure to bring impaired water into compliance with TMDL limits. Should humans be the primary source of pollution, the BMP's could potentially influence millions of dollars of construction in the fast growing region (Connaughten 2005). Pet influence could require stronger laws requiring owners to clean after their animal. Should horses or livestock be the primary cause, changes may be required to farm and field management practices to prevent fecal encroachment. In the event that wildlife are the source, nutrient load reduction (Simmons et al. 1995, Simmons et al. 1998) or regulation exception might be in order. Further, public infrastructure and future planning could be influenced in similarly situated counties throughout the region.

II. Current Status of Microbial Source Tracking (MST)

As the result of recent papers described in this section, MST is undergoing a change in acceptable practices. Continuing efforts are being made to find unique characteristics of fecal organisms in water contaminated by particular waste. Established methods require multiple overlapping tests in order to build confidence by agreement of results. These tests require a set of training data, or library, in order to categorize results

with various statistics. Researchers continue to refine existing methods, however the goal is to switch over to rapid field tests independent of data libraries. It is not clear at this point just how feasible the desired methods will ultimately be.

The three main categories of source tracking research include host-source specific methods, species-specific methods, and chemical methods. Host-source methods are characterized by biological indicators that only exist in waste from a particular host. Species specific methods are dependent on the attributes of target enteric bacterial species to classify to different sources. Chemical methods use indicator molecules that are found in the waste of particular sources.

In the last two years two review papers (Scott et al. 2002, Simpson et al. 2002), two method comparison studies (SCCWRP, Stoeckel et al. 2004) and an EPA guide (USEPA 2005) have been published. These papers have covered the majority of species specific techniques, including ribotyping, carbon utilization procedures, coliphage typing, pulse field gel electrophoresis, antibiotic resistance analysis, and repetitive element PCR. As those papers covered the material in detail, only new developments and uncovered techniques will be addressed herein.

Several host-source specific methods require the detection of particular species or generic attributes in enteric bacteria. Sorbitol fermenting *Bifidobacteria* are one group unique to humans. Immunoglobulins can be linked to wide classes of hosts (Leonard 2001). *Rhodococcus coprophilus* has been evaluated and is an unlikely candidate (USEPA 2005, Long et al. 2003, Gilpin et al. 2003). Phages of *Bacillus fragilis* are also unique to humans. (Long et al. 2003). All of these microbes have proved lengthy or

difficult to cultivate (USEPA 2005) and will likely remain alternatives unless the main body of *Enterococcus* and *E. coli* methods prove to be failures.

A. MST Reviews

The review papers (Scott et al. 2002, Simpson et al. 2002) considered the main techniques available at the time. Several techniques were considered promising, but no method comparison studies had been attempted at that point. The most significant conclusion from the papers was that techniques would need to be evaluated together in a "toolbox" approach. The reviews were written based on publications at the time which mostly contained small data sets. The reviews were premature and have been superseded by both the EPA guide document and the method comparison studies.

B. EPA Guide Document

The EPA MST Guide (USEPA 2005) is the most complete description of MST to date. It includes significant treatment of the statistics involved. MST methods were categorized according to need for library and need to cultivate. The guide stressed quality assurance and described control samples for each method. Eight case studies were also included. Cost estimates were used as an additional consideration in evaluation of each method. Chemical source tracking methods were not included**.**

C. Method Comparison Studies

To date, two large methods comparison studies in MST were the USGS (Stoeckel et al. 2004) and SCCWRP (See SCCWRP) studies. Both studies sought a most reliable method of MST in order to channel research and development. Unfortunately, no one

method stood out as the best based on criteria of accuracy, cost and throughput. However, for accuracy, PFGE was the best method in the USGS comparison study (Stoeckel et al. 2004), and was one of the two best methods (with ribotyping) in the SCCWRP study. Both studies suffered from a low number of isolates tested (630 isolate library for USGS, 300 isolate library for SCCWRP).

Ritter made further recommendations for the data treatment of future studies (Ritter et al. 2003). The Ritter paper indicated that studies should attempt multiple methods and include a challenge set of known isolates to verify the goodness of training data. Only unique data should be used to prevent biasing unrepresentative sets to their own data. A further conclusion was to add an outside method of verification beyond the challenge set and internal training data metrics (Stewart et al. 2005). More recent statistical analysis indicate that Jacard methods of distance and band differentiation are appropriate (Stewart et al. 2005). The present study is one of the first designed to follow these recommendations.

1. Library Dependent Methods (Figure 1)

- Ribotyping is a system of categorization based on the 16S and 23S ribosomal RNA (rRNA) sequence.
- Carbon source utilization profiling (CUP) is based on nutrient metabolism.
- Repetitive element PCR (rep-PCR) uses particular primers to amplify DNA sequences of varying lengths. REP and BOX are typical primer sets.
- Real-time PCR quantifies the number of bacteria using an indicator of the amount of DNA amplified.
- Length heterogeneity and terminal restriction fragment length polymorphisms (LH-RFLP and T-RFLP) use cleaved 16S rDNA from PCR to categorize.
- Denaturing gel gradient electrophoresis (DGGE) uses the increased resistance of unraveling DNA top stop its movement through a gel.
- Amplified fragment length polymorphism (AFLP) uses specific primers and their partial complement to amplify certain genomic DNA.
- Randomly amplified polymorphic DNA (RAPD) uses nonspecific primers to create a strain specific fingerprint.

- Antibiotic Resistance Analysis (ARA) is an inexpensive phenotypic technique.

2. Library Independent Methods

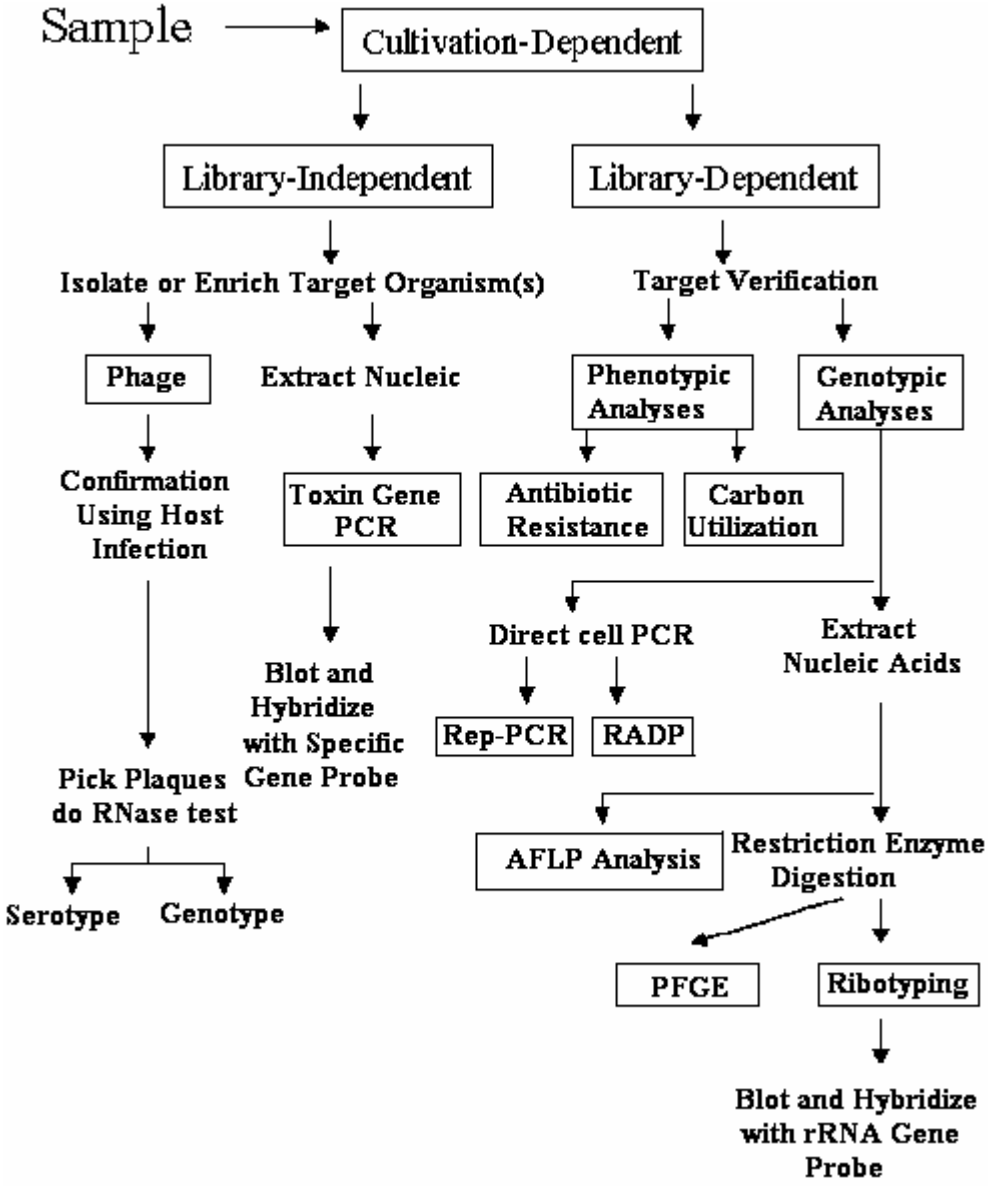
The majority of MST techniques require a library or training set of isolated organisms (isolates) on which to base determinations. While these training sets have been analyzed, the minimum number of isolates for a statistically sound study has yet to be determined (Simpson et. al. 2002). Several recommendations have been made (Ritter et al. 2003), but no definitive rule has been accepted.

Several MST methods rely on the presence or absence of markers. Most of these methods are known as library independent methods and may be cultivation independent methods as well (Figure 2). At this point, most of these methods have not been tested on a large enough data set to indicate their universal application (Field et al. 2003). This criticism mainly applies to biological markers, but many chemical approaches still need to check for potential interactions.

The two main classes of library independent methods are:

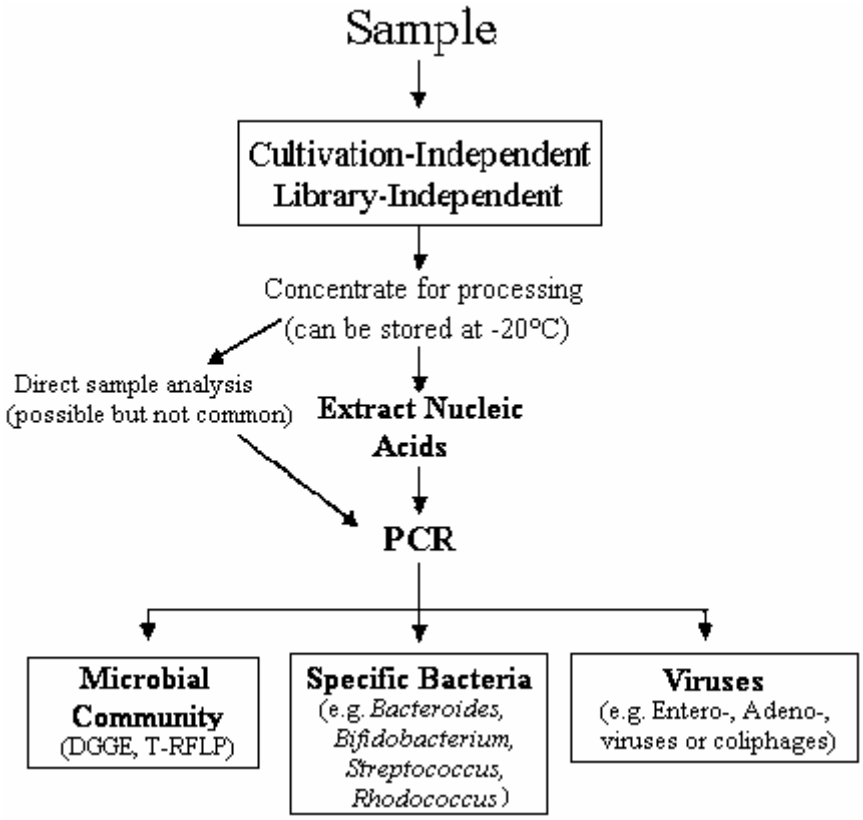
- F coliphage typing, differentiates between human and animal sources based on the presence of DNA or RNA from the virii of *E. coli*.
- Gene specific PCR based on known sequences occurring only in microbes of particular host animals.

Figure 1. Cultivation Dependent Source Tracking Methods[a]



[a] Partial Reproduction from US EPA 2005

Figure 2. Cultivation Independent Source Tracking Methods[a]



[a] Partial Reproduction from US EPA 2005

D. Statistical Analysis of MST

One large consideration for any library based method is whether the library is representative. The comparison studies (Stoeckel et al. 2004, SCCWRP) used challenge sets of unknown isolates to test. Wiggins has used rates of correct classification comparison between within library and out of library isolates (Wiggins et al. 2003). Another method he used was jackknife or leave-out analysis, in which isolates from a single fecal sample are removed and tested for correct classification. Reverse birthday analysis, based on repeated isolates, was suggested by computer scientists (Ritter 1994).

Because no statistical measurement of a library conclusively indicates ability to classify environmental samples, library methods will have multiple measurements.

A second consideration for training data is internal bias. In a random sample from a population, there is a chance of an imbalanced sample with a higher ratio of data points from one particular group than is in the population. This is an understood and accepted risk that is usually avoided by having a large sample set. There is also a possibility of a normal ratio, but an artifact of the statistical analysis may make an imbalanced ratio appear present. This artifact, artificial clustering, should be tested for by modifying the library data. The labeled source for the data is randomized and the statistical analysis of the data is redone. The artificial clustering would show rates of classification above one divided by the number of classes. So far, this technique has only been tested and included in a few source tracking programs.

In many of the current methods, fingerprints are created from gel band patterns. In most of these, the brightness of the band is not significant but rather culture and gel irregularity dependent (Stewart et al. 2005). Consequently band analysis matching requires use of matching methods that focus only on the presence or absence of bands. According to recent research, Jaccard gives the best results of Jaccard, Dice, Jeffrey's X, and Ochai (Ritter et al. 2003, Stewart et al. 2005, Hassan et al. 2004). Linear discriminant analysis, as done for PFGE by Simmons and refined by this lab, has yet to be analyzed in a comparison study.

During creation of a library and subsequent use for analysis, aberrant data points can be a considerable problem. One suggested method for library creation suggests creating an unknown category for unusual data points (Stewart et al. 2005). Another

method is to simply remove strong outliers due to possible process flaws. In the cases of conflicting data points, consideration has been given to assigning them to one category based on similar data, numerically prevalent data (Stewart et al. 2005) or accepting a lower rate of correct classification. Due to the nature of linear discriminant analysis, this study has included conflicting points to counterbalance bias except when local data points have conflicted with those out of watershed.

Analysis of retrieved data is problematic when classifications are made with low probability. Many methods rely on statistical processes that must assign environmental isolates to a particular source. In the USGS method comparison, a threshold was considered to remove problem isolates, but other problems with this study prevented it from giving a significant answer. Hassan recommends using a similarity or quality factor (an alternative threshold) for increasing confidence (Hassan et al. 2005). Wiggins et al. 2003, Harwood et al. 2003, and Hagedorn et al. 2003 have used a further threshold, the minimum detectable percentage (MDP), which is based on the variability in the misclassification rates, to ignore low numbers of classified isolates. The key issue of these analyses is their ability to remove that which one is unsure of. In studies that apply source tracking conclusions to real world problems, this is crucial.

Over the last four years, the within library rates of correct isolate classification, or average rate of correct classification (ARCC), have been reduced. Artificial clustering was the typical reason for the previous high ARCC's. The two chief changes have been an increase in library size and the elimination of within library clones. Mitigating this loss of specificity, method revisions not included in comparison studies have helped ARCC's from dropping as much. While ARA classification rates have been as high as

90%, typical average classification rates on uniques have been approximately 60% (Ram et al. 2004, Johnson, L. et al. 2004). This lower rate reduces the ability to predict secondary sources significantly, requiring a minimum detectable percentage (Wiggins et al. 2003) on a four way split, for example, to be at least 13% of the total signal.

E. Chemical (Non-Microbial) Source Tracking Methods

There are several source tracking methods based on the presence of specific chemicals. The main ones deal with the differentiation of human waste influenced waters from non-human influenced. Triclosan detection, a synthetic anti-microbial, is a promising method but expensive. Caffeine, arsenic and fluorescent brightener detection are three methods that depend on differentiation from background levels. Cotinine (digested nicotine) levels also show human water use. Arsenic levels in water can be source dependent. Fluorometry, a promising technique, will be discussed in study design.

Immunoglobulins can potentially differentiate between mammal and other fecal sources. Very little research has been done on environmental stability and survivability of secretory globulins. Current water research is being pioneered at the University of South Mississippi using Immunoglobulin A to differentiate between animal sources.

Caffeine is used to detect human wastewater. Caffeine is a naturally occurring compound that is included in many foods and beverages. Caffeine detection is based on the assumption that untreated human wastewater would have much higher levels of caffeine than found in the environment. Unfortunately, wastewater treatment does not reduce caffeine levels significantly enough to show low background levels (Johnson, A. et al. 2004). Further, environmental sources of caffeine are plentiful enough in plant life, litter, and runoff as to create highly variable background nonwaste signals.

Triclosan is used to detect human wastewater. Triclosan is a synthetic antimicrobial agent present in many antibacterial products. Detection of triclosan is dependent on higher levels in human wastewater than in natural systems. Unfortunately, triclosan levels vary greatly in human wastewater. Further, questions of triclosan safety may imply reduces in future usage (Rule et al. 2005). Continine, a nicotine derivative, also has reduced detection usability due to similar variability and future use considerations. Both chemicals require costly equipment for extraction and detection.

Arsenic levels have been used as potential wastewater indicators. While arsenic is a potential strong indicator for chicken feces (USEPA 2005), several environmental sources such as industrial discharge, treated wood, fertilizers and certain soils in the region can strongly interact. This indicator will probably not develop further in source tracking.

Of the alternative source tracking techniques, a few show promise. Background levels and cost of assay are the key factors influencing their use. Immunoglobulin assays and fluorescence tracking are the only methods with significant potential.

III. The Future of MST Research

Four aspects of MST are open for considerably more work. Indicator genetic stability and population composition must be measured in relation to a variety of environmental factors. Markers and libraries must be expanded to more test data. Genomic and proteomic understanding of host organisms will provide many more areas of exploration. Follow up studies designed to verify the results and application of previous studies could return the focus of the science back to application.

The survivability and both the genetic and phenotypic stability of the indicators needs to be quantified. While research into the chemostatic composition of a ruminant's gut exists, it tells nothing of the expression and exchange of genetics going on among the bacteria within it. Marked differences in growth conditions between the environment and the gut could change the population composition and genetic expression of indicator organisms. Diurnal, solar, stream flow, and weather differences in sampling affect indicators (Whitman and Nevers, to be published). Continuous flow data for the sampled streams will be crucial to describe fresh or possible periodic waste influence (Jamieson et al. 2005). Continued basic research in environmental effects on indicators will lead to better MST conclusions.

As the fields of genomics and proteomics grow, MST should be poised to reap the benefit. Potential virulence factors unsuspected in previous work will become apt targets (Simpson et al. 2002). Proteomics will show possible receptor molecules specific to particular organisms that may interact with similarly shaped or interlocking host or indicator molecules. Using proteomics to find MST targets will be greatly enhanced by cheap computing power allowing researchers to calculate protein shapes in their own lab.

Successful application must always be the goal of MST. While basic science and method improvement are important, the goal of MST is pollution control. More studies need to occur that verify a previous studies results, particularly through the application of BMP's. Science is only as significant as its use, and MST should prove its worth.

IV. Study Design

A. Antibiotic Resistance Analysis (ARA)

ARA is a phenotoypic species specific method. ARA has been performed on the *Enterococci*, fecal coliforms, and *E. coli* (Booth et al., 2003; Graves et al., 2002; Hagedorn et al., 1999). This method relies on different antibiotic resistance patterns in fecal bacteria that can be related to specific sources of fecal pollution, and is predicated on the rationale that antibiotics exert selective pressure on the fecal flora of the animals that ingest or are treated with the antibiotic(s), and that different types of animals receive differential exposure to antibiotics. Benefits of ARA include use of simple laboratory techniques, requiring only basic equipment, and it can be performed at a relatively low cost compared to most other MST methods. To date, ARA has been used in more MST projects in the U.S. than any other method. In addition, high levels of separation between known source bacterial isolates have been found comparable to those reported for molecular methods.

Two methods of ARA are used in research. In the first method, fecal bacteria are isolated from fecal samples and challenged with antibiotics and scored for growth or no growth. This provides a library of resistance. A second less widely used method uses antibiotic zones of inhibition (Sayah et. al. 2004), rather than growth or no growth. Environmental isolates are then challenged with the same antibiotics and compared to the library set. The environmental isolates are then categorized through some form of discriminant analysis. Unlike molecular techniques, this high throughput, low cost method allows 10-fold or more increase in the isolates tested.

The antibiotic resistance variation between isolates from different sources tends to follow certain trends. Humans are typically found to have the most antibiotic resistant microbes, followed by livestock, pets, and least antibiotic resistant wildlife microbes. There are exceptions to this (Sayah et. al. 2004): pets may share microbes with humans and common low resistant microbes may be spread throughout animals. Other animals, such as gulls, feed in human sewage and may share human fecal bacteria. Areas, such as farms and hospitals, where antibiotic use is more common, tend to spread antibiotic resistance into the neighboring fauna in a manner similar to a chemostat (Levy 2005, Sayah et al. 2004). Lateral transfer of resistance genes also creates a source of variation in the data (Levy 2005).

All these sources of variation require thorough testing to assure that the library is representative of the research locale. The training data set must be significantly large, 1000 or more non-unique isolates depending on watershed size, to represent the local variation (Wiggins et al. 2003). Due to the evolving variation, it is expected that the library is time limited. Current data suggests that libraries are good for at least one year (Wiggins et al. 2003, O'Brian et al. 2005). Wildlife tests will likely prove more significant do to greater capacity for antibiotic level increases.

Reasons for ARA as Main MST Method

- Ease of Method
- Low Cost Per Isolate
- Comparison to Virginia DEQ results (MapTech) achieved through ARA
- Cross-Discipline Applicability (Bryan et al. 2004, Whittam 2005)
- Established Regional Library (Booth et al. 2003, Graves et al. 2003, Porter et al. 2003, Chapman et al. 2004)

B. Fluorometry

Fluorometry is used to detect human wastewater (Gilpin et al. 2003). Fluorometry detects fluorescent brighteners (brighteners, also known as optical brighteners and fluorescent whitener), detergent surfactants and fecal sterols that react to ultraviolet (300-400nm) light to fluoresce white. Other chemicals can be detected, but fluoresce at varying wavelengths (Holbrooke 2005) or at reduced strength (Hagedorn unpublished data 2004). A separate project (sponsored by NOAA) evaluated the use of a fluorometer in estuarine and coastal zone environments to determine if the equipment could detect a human waste signature. Brighteners are found in detergent, toilet paper, and tissues and are ubiquitous in human wastewater. Wastewater treatment facilities reduce brighteners through sediment and bacterial sorption (Poiger et al. 1998). There are at least two major potential human sources of contamination that could contain optical brighteners, and these include leachates from improperly functioning on-site wastewater systems (OWS) and leaking pipes from community wastewater treatment systems (Sargent and Castonguay 1998). In rural areas where the majority of homes are served by on-site systems, optical brighteners in water samples indicate failing conditions within OWS in close proximity to the sampled bodies of water. Fluorometry is currently under expanding use in several regions(USEPA 2000 , Waye 2000)

Fluorometry was used as a secondary indicator due to its widespread applicability. The ubiquitous presence of brighteners in wastewater lends to good detection of leaks throughout the system. The low background level relative to wastewater furthers its use as a hotspot indicator. Brightener's excitation/emission spectra is distinct from other fluorescing compounds in water. The cost of equipment is relatively low compared to

other chemical methods, which typically require a mass spectrometer. As this technique was pioneered in the Hagedorn lab, the body of knowledge, and its further extension, has been heavily dependent on continued research from this lab.

C. Pulse Field Gel Electrophoresis (PFGE)

In PFGE microbial DNA is isolated and then digested using restriction enzymes (Simmons and Herbein, 1998; Simmons et al., 1995; Simmons et al., 2000). PFGE differentiates closely related isolates through the position of restriction enzyme cleavage. The fragments of DNA are separated into distinct bands using a pulsed electrical field over 30 hours. Successful analysis requires a large culture collection, although identical matches can be easily classified. PFGE has been used as a cross-validation tool in numerous MST projects around the U.S.

PFGE was chosen as a further confirmation test due to its established accuracy. The presence of a large library at this location and an experienced technician also contributed to the choice of this technique as a verification technique. A low number of isolates were chosen for verification due to the time/cost limitations of PFGE.

D. Data Analysis

Four data representations, four LDA techniques and six statistics were selected for data analysis. The four data representations chosen were binary (Binary), maximum concentration of antibiotic resistance (High), highest concentration of antibiotic resistance before a concentration of no antibiotic resistance (Last), and a combination of High and Last (Combination). Data representations were based on the suggestions of Ritter et al 2003 and US EPA 2005. KSL's were designed using LDA techniques of

conflict removal, clustering, unknown class creation and thresholding. Techniques attempted were based on the method comparison studies (** and SCCWRP) and Stewart et al 2005. The six statistics selected were rate of correct classification (RCC), positive predictive value (PPV), negative predictive value (NPV), sensitivity, specificity and minimum detectable percentage (MDP). Statistic selection were based on Ritter et al 2003 and Stewart et al 2005.

E. Site Selection and Sampling Frequency

A total of ten sites were selected on eight streams in conjunction with Prince William County Department of Public Works in order to correspond closely with Virginia Department of Environmental Quality (DEQ) sampling locations. Sites were monitored with monthly frequency for 13 months in order to provide a more thorough analysis than DEQ sampling. Sites varied in drainage area and predominant land use, however all were within the Occoquan Basin (OQB).

V. References

Booth, A. M., A. K. Graves, C. Hagedorn, S. C. Hagedorn, and K. H. Mentz. 2003. Sources of Fecal Pollution in Virginia's Blackwater River. J. Environ. Engineering. 129:547-552.

Bryan, A., Shapir, N., Sadowsky, M.J. 2004. Frequency and Distribution of Tetracycline Resistance Genes in Genetically Diverse, Nonselected, and Nonclinical Escherichia coli Strains Isolated from Diverse Human and Animal Sources. Appl. Environ. Microbiol. 70:2503-2507.

Chapman,A., Hagedorn, C., Saluta, M. 2004. Identifying Sources of Fecal Pollution in Impaired Waters in Prince William County, Virginia. American Society for Microbiology, General Meeting. New Orleans, Lousiana.

Chesapeake Bay Program. 2005. http://chesapeakebay.net [online]. Last accessed June 26, 2005. Chesapeake Bay Program: Annapolis, Maryland.

Connaughten, S.T. 2005. State of the County Address. http://www.co.prince-william.va.us/default.aspx?topic=010010000810002865. Prince William County.

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

Field, K. G., Chern, E. C., Dick, L. K., Fuhrman, J., Griffith, J., Holden, P. A., LaMontagne, M. G., Le, J., Olson, B., Simonich, M. T. 2003. A comparative study of culture-independent, library-independent genotypic methods of fecal source tracking. J. Wat. Health 01.4:181-193.

Gilpin, B., James, T., Nourozi, F., Saunders, D., Scholes, P., and Savill, M. 2003. The use of chemical and molecular microbial indicators for faecal source identification. Wat. Sci. Technol. 47:39-43.

Graves, A. K., Hagedorn, C., Teetor, 11A., Mahal, M., Booth, A. M., and Reneau, Jr., R. B. 2002. Antibiotic resistance profiles to determine sources of fecal contamination in a rural Virginia watershed. J. Environ. Qual. 31:1300-1308.

Griffith, J. F., Weisberg, S. B., and McGee, C. D. 2003. Evaluation of microbial source tracking methods using mixed fecal sources in aqueous test samples. J. Water Health 1:141-151.

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

Hagedorn, C., Crozier, J. B., Mentz, K. A., Booth, A. M., Graves, A. K., Nelson, N. J., and Reneau, R. B. 2003. Carbon source utilization profiles as a method to identify sources of fecal pollution in water. J. Appl. Microbiol. 94:792-799.

Hassan, W.M., Wang, S.Y., Ellender, R.D. 2004. Methods to Increase Fidelity of Rpetitive Extragenic Palindromic PCR Fingreprint-Based Bacterial Source Tracking Efforts. Appl. Environ. Microbiol. 71:512-518.

Harwood, V. J., Wiggins, B., Hagedorn, C., Ellender, R. D., Gooch, J., Kern, J., Samadpour, M., Chapman, A. C. H., Robinson, B. J., Thompson, B. C. 2003. Phenotypic library-based microbial source tracking methods: Efficacy in the California collaborative study. J. Wat. Health 1:153-166.

Holbrooke, D. 2005. Evaluating the Characteristics and Behavior of Organic Materials in the Occoquan Watershed using Fluorescence Spectroscopy. April 15, 2005, Virginia Tech.

Jamieson, R.C., Joy, D.M., Lee, H., Kostaschuk, R., Gordon, R.J. 2005. Resuspension of Sediment-Associated Escherichia coli in a Natural Stream. J. Environ. Qual. 34:581-589.
Johnson, A., Carey, B., Golding,S. 2004. Results of a Screening Analysis for Pharmaceuticals in Wastewater Treatment Plant Effluents, Wells, and Creeks in the Sequim-Dungeness Area. Washington Dept. of Ecology.

Johnson, L. K., Brown, M. B., Carruthers, E. A., Ferguson, J. A., Dombek, P. E., and Sadowsky, M. J. 2004. Sample size, library composition, and genotypic diversity among natural populations of Escherichia coli from different animals influence accuracy of determining sources of fecal pollution. Appl. Environ. Microbiol. 70:4478-4485.

Leonard D.L. 2001. National Indicator Study: Is an international approach feasible? J. Shellfish Research. 20:1293-1298.

Levy, S.B. 2005. Spread of Antibiotic Resistance Among Animals and People.American Society for Microbiology General Meeting, Atlanta, GA.

Long, S.C., Shafer, E., Arango, F.C., Siraco, D. 2003. Evaluation of Three Source Tracking Indicator Organisms for Watershed Management. J. Water Supply Research Tech. - Aqua. 52:565-575.

Long,, S.C. and Plummer, J.D. 2004. Assessing Land Use Impacts on Water Quality Using Microbial Source Tracking. J. Amer. Water Res. Assoc. 40:1433-1448.

Myoda, S. P., Carson, C. A., Fuhrmann, J. J., Hahm, B.–K., Hartel, P. G., Yampara-Iquise, H., Johnson, L., Kuntz, R. L., Nakatsu, C. H., Sadowsky, M. J., and Samadpour, M. 2003. Comparison of genotypic-based microbial source tracking methods requiring a host origin database. J. Wat. Health 1:167-180.

Noble, R. T., Allen, S. M., Blackwood, A. D., Chu, W., Jiang, S. C., Lovelace, G. L., Sobsey, M. D., Stewart, J. R., and Wait, D. A. 2003. Use of viral pathogens and indicators to differentiate between human and non-human fecal contamination in a microbial source tracking comparison study. J. Wat. Health 1:195-207.

O'Brien, T.L., Bailey, D., Gill, A., Staton, P. 2005. Considerations for Using Sewage Versus Direct Human Samples to Generate a Bacterial Source Tracking Database. American Society for Microbiology Annual Meeting Abstracts. Atlanta, GA.

Porter, K. R. 2003. Identifying Sources of Fecal Pollution in Washington D.C. Waterways. M.S. Thesis. Virginia Tech.

Poiger, T., Field, J.A., Field, T.M., Siegrist, H., Giger, W. 1998. Behavior of Fluorescent Whitening Agents During Sewage Treatment. Water Research. 32:1939-1947

Ram, J. L., Ritchie, R. P., Fang, J., Gonzales, F. S., and Selegean, J. P. 2004. Sequence-based source tracking of Escherichia coli based on genetic diversity of □-glucuronidase. J. Environ. Qual. 33:1024-1032.s

Ritter, K. J., Carruthers, E., Carson, C. A., Ellendere, R. D., Harwood, V. J., Kingsley, K., Nakatsu, C., Sadowsky, M., Shear, B., West, B, Whitlock, J. E., Wiggins, B. A., and Wilbur, J. D. 2003. Assessment of statistical methods used in library-based approaches to microbial source tracking. J. Wat. Health 01.4:209-223.

Ritter, T. 1994. Estimating Populations from Repetitions in Accumulated Random Samples. Cryptologia. 18:155-190.[errata: http://www.ciphersbyritter.com/ARTS/BIRTHDAY.HTM [Online] Accessed June 26, 2005]

Krista L. Rule, Virginia R. Ebbett, and Peter J. Vikesland. 2005. Formation of Chloroform and Chlorinated Organics by Free-Chlorine-Mediated Oxidation of Triclosan. Environ. Sci. Technol. 39:3176 - 3185.

Sargent, D., and Castonguay, W. 1998. An Optical Brightener Handbook. http://www.naturecompass.org/8tb/sampling/index.html. Eight Towns and Bay Committee.

Sayah, R.S., Kaneene, J.B., Johnson, Y., Miller, R. 2005. Patterns of Antimicrobial Resistance Observed in Escherichia coli Isolates Obtained from Domestic- and Wild-Animal Fecal Samples, Human Septage, and Surface Water. Appl. Environ. Microbiol. 71:1394-1404.

Scott, T. M., Rose, J. B., Jenkins, T. M., Farrah, S. R., Lukasik, J. 2002. Microbial source tracking: Current methodology and future directions. Appl. Environ. Microbiol. 68:5796-5803.

SCCWRP see citations at SCCWRP heading below

Simpson, J. M., Santo Domingo, J. W., and Reasoner, D. J. 2002. Microbial source tracking: State of the science. Environ. Sci. Technol. 36:5279-5288.

Simmons, G. M., Jr., S. A. Herbein, and C. M. James. 1995. Managing Nonpoint Fecal Coliform Sources to Tidal Inlets. Water Resource Update 100:64-74.

Simmons, G. M., Jr., and S. A. Herbein. 1998. Shellfish and Water Column Comparison of Fecal Coliform Diversity Using NotI DNA Fingerprints of Escherichia coli Generated by Pulse Field Gel Electrophoresis. Final Report for the Virginia Coastal Resource Management Program Department of Environmental Quality, Richmond, VA.

Stewart, J. R., Ellender, R. D., Gooch, J. A., Jiang, S., Myoda, S. P., and Weisberg, S. B. 2003. Recommendations for microbial source tracking: Lessons learned from a methods comparison study. J. Water Health 1.4:225-231.

Stewart, J.R., Robinson, B., Hyer, K., Hagedorn, C., Whittam, T.S., Wilbur, J. 2005. Microbial Source Tracking Using Indicator Organisms. American Society for Microbiology General Meeting, Atlanta, GA.

Stoeckel, D. M., Mathes, M. V., Hyer, K. E., Hagedorn, C., Kator, H., Lukasik, J., O'Brien, T. L., Fenger, T. W., Samadpour, M., Strickler, K. M., and Wiggins, B. A.

2004. Comparison of seven protocols to identify fecal contamination sources using Escherichia coli. Environ. Sci. Technol. 38:6109-6117.

USDA. 1995. National Shellfish Sanitation Program Manual of Ops. U.S. Dept. of Agriculture: Washington, D. C.

US EPA. 1986. Quality Criteria for Water 1986. U. S. Environmental Protection Agency: Washington, D.C.

USEPA. .2000. Optical Brightener to Shed Light on Sewer and Septic Tank Leaks. Nonpoint Source News-Notes 63:11-12.

US EPA. 2005. Microbial Source Tracking Guide Document. U. S. Environmental Protection Agency, Office of Research and Development: Washington, D.C.

Virginia DEQ. 2004. 2004 305(b)/303(d) Water Quality Assessment Integrated Report. Virginia Dept. of Eviron. Quality: Richmond, Virginia,

Virginia DEQ. 2005. Virginia DEQ: Background Information on TMDLs. http://www.deq.virginia.gov/tmdl/backgr.html [online]. Last accessed June 26, 2005.

Waye, D. 2000. A New Tool for Tracing Human Sewage in Waterbodies: Optical Brightener Monitoring. Virginia Water Resources Research Symposium, November 2000.

Whitman, R.L., and Nevers, M.B. To Be Published. Escherichia coli Sampling Reliability at a Frequently Closed Chicago Beach: Monitoring and Management Implications.

Whittam, T.S. 2005. Genetic Diversity and Antibiotic Resistance in *E. coli* Populations in Nature. American Society for Microbiology General Meeting, Atlanta, GA.

Wiggins, B.A., Cash, P.W., Creamer, W.S., Dart, S.E., Garcia, P.P., Gerecke, T.M., Han, J., Henry, B.L., Hoover, K.B., Johnson, E.L., Jones, K.C., McCarthy, J.G., McDonough, J.A., Mercer, S.A., Noto, M.J., Park, H., Phillips, M.S., Purner, S.M., Smith, B.M., Stevens, E.N., Varner, A.K. 2003. Use of Antibiotic Resistance Analysis for

Representativeness Testing of Multiwatershed Libraries. Appl. Environ. Microbiol. 69:3399-3405.

SCCWRP

Field, K. G., Chern, E. C., Dick, L. K., Fuhrman, J., Griffith, J., Holden, P. A., LaMontagne, M. G., Le, J., Olson, B., Simonich, M. T. 2003. A comparative study of culture-independent, library-independent genotypic methods of fecal source tracking. J. Wat. Health 01.4:181-193

Griffith, J. F., Weisberg, S. B., and McGee, C. D. 2003. Evaluation of microbial source tracking methods using mixed fecal sources in aqueous test samples. J. Water Health 1:141-151.

Harwood, V. J., Wiggins, B., Hagedorn, C., Ellender, R. D., Gooch, J., Kern, J., Samadpour, M., Chapman, A. C. H., Robinson, B. J., Thompson, B. C. 2003. Phenotypic library-based microbial source tracking methods: Efficacy in the California collaborative study. J. Wat. Health 1:153-166

Myoda, S. P., Carson, C. A., Fuhrmann, J. J., Hahm, B.–K., Hartel, P. G., Yampara-Iquise, H., Johnson, L., Kuntz, R. L., Nakatsu, C. H., Sadowsky, M. J., and Samadpour, M. 2003. Comparison of genotypic-based microbial source tracking methods requiring a host origin database. J. Wat. Health 1:167-180

Noble, R. T., Allen, S. M., Blackwood, A. D., Chu, W., Jiang, S. C., Lovelace, G. L., Sobsey, M. D., Stewart, J. R., and Wait, D. A. 2003. Use of viral pathogens and indicators to differentiate between human and non-human fecal contamination in a microbial source tracking comparison study. J. Wat. Health 1:195-207.

Ritter, K. J., Carruthers, E., Carson, C. A., Ellendere, R. D., Harwood, V. J., Kingsley, K., Nakatsu, C., Sadowsky, M., Shear, B., West, B, Whitlock, J. E., Wiggins, B. A., and Wilbur, J. D. 2003. Assessment of statistical methods used in library-based approaches to microbial source tracking. J. Wat. Health 01.4:209-223.

Stewart, J. R., Ellender, R. D., Gooch, J. A., Jiang, S., Myoda, S. P., and Weisberg, S. B. 2003. Recommendations for microbial source tracking: Lessons learned from a methods comparison study. J. Water Health 1.4:225-231.

Chapter 2. Project Goals and Objectives

The project goal was to monitor and evaluate eight streams in the Occoquan Basin (OQB) that have been identified as impaired waters due to high *E. coli* concentrations. One site on each of six streams and two sites on the remaining two streams were identified for *E. coli* and *Enterococcus* monitoring and microbial source tracking. Repeated sampling of the ten locations for thirteen months assessed the concentrations of the bacteria over time, while comparison of monthly bacteria concentrations to the U.S. standards verified of the impaired water designation.

The project objectives are:

1. The categories of sources that cause bacterial impairment of the water will be determined through Antibiotic Resistance Analysis (ARA). The TMDL improvement plan will specify measures targeting the polluting source categories.
2. Known Source Libraries (KSL's) for use in ARA will be manipulated to determine the KSL design which best identifies the sources of environmental isolates. Comparison will be made of novel and recommended methods of KSL design. Configuration of the KSL will be attempted with both local and regional isolates in order to determine the success of regional library integration. The design with the greatest potential to correctly identify environmental isolates will be used for ARA.
3. Measurement of optical brighteners in freshwater by fluorometry as an indicator for human wastewater will be evaluated. Potential of fluorometry to indicate wastewater will be assessed through comparison to ARA results. Each of the ten sites will have monthly measurements of fluorometry data. Comparison will be

made to ARA results to determine correlation between identification of human bacterial isolates and fluorometric data.

Chapter 3. Materials and Methods

I. Study Area

The Occoquan Basin (Occoquan) of the Middle Potomac-Anacostia-Occoquan watershed (USGS Cataloguing Unit 02070010) within Prince William County was the site of this study. The Occoquan drains an area of 1528 km^2 and discharges into the Potomac River and thereafter the Chesapeake Bay (Appendix I).

Within the Occoquan a variety of land uses occur. This was traditionally a rural region with farming being the primary land use. Since the 1950's the area has been developing into a suburban center. Recently, development pressure(PWCOHCD 2005) has lead to limitations on further urbanification and farms continue to disappear(Connaughton 2005). As development continues, wildlife and humans are being forced to share more of the same greenways and undeveloped spaces.

As space is shared, so too is waste commingled. Visits to the Occoquan commonly indicated deer (*Odocoileus virginianus*), rabbits (*Sylvilagus floridanus*), fox (*Vulpes vulpes*), and raccoon (*Procyon lotor*) presence in the greenways. Geese (*Branta canadensis* and *Anser domesticus*) and various gulls (*Larus sp*.) are present bird sources. Pet waste, particularly dogs (*Canis familiaris*), are also distinct environmental concerns. Livestock are of reduced concern except one farm in the area and horse trails throughout the region. The livestock included were chicken(*Gallus gallus*), pig(*Sus scrofa*), horse(*Equus caballus*), and goat(*Capra aegagrus*). Emu (*Dromaius novaehollandiae*) were known to be at the farm, however no fecal samples were available. Human (*Homo sapiens*) waste, as sewer or septic leakage, was also of considerable concern.

Host-Origin isolates from previous studies and those recently collected were used in the current study. Fifty percent of the isolates included were collected from the watershed and 25% were collected from the Occoquan Basin (OQB). The rest of the isolates were collected from the Northern Virginia region. Isolates were collected until the known source library (KSL) showed representativeness of the OQB.

Isolates collected in the basin were collected in specific areas. Wildlife isolates were collected in the greenways with the majority coming from the Mannassas National Battlefield Park (under permit). Dog isolates were collected from the Battlefield Park, grooming services, a rest area along I-66, and owner donation. Human isolates were taken from the H.L. Mooney Water Reclamation Facility, Woodbridge, VA. Horse samples were taken from the Battlefield as well as from owner donation. Bird samples were taken from ponds in the Mannassas area, a local golf course and the Battlefield.

II. Water Sample Isolates

Ten sites in the watershed were sampled monthly throughout the 12 month sampling period. These sites were a continuation of the source tracking study from a separate part of the watershed (Chapman et al 2004). At each location a 250ml water sample was collected. Each month one site was selected for a duplicate sample and one other site was selected for a sediment sample. Over the 12 months every site had at least one sediment sample and one duplicate sample.

Site 1: Upper Bull Run –Blackburn's Ford - Longitude -77°26'58" Latitude 38°48'10" The sampling was conducted just downstream from a thin wooded zone separating older houses from the stream. Samples were collected at Centreville Rd (Rt 28). Although

nearby feeder streams drain the old housing, their flow was relatively minor and did not show signs of wastewater. ATV use was noted just upstream on the Prince William side. Summer months showed regular use of the waters by children, teenagers and families for swimming.

Site 2: Lower Bull Run – Marina - Longitude -77°23'15" Latitude 38°44'30"

This location was at a boat landing and tended to show very low flow rates. Samples were collected at the Yates Ford Rd crossing into Fairfax County. The waters here drain a long wooded stretch although feeder streams drain housing, none are near the sample site. This site was used for boating and fishing.

Site 3: Youngs Branch - Longitude -77°31'35" Latitude 38°49'04"

Here the water drains much of the battlefield. Upstream of this site were a couple of private ponds. Samples were taken at Sudley Rd (Rt 234). Numerous springs drain both the battlefield and further distant neighborhoods. Horse trails and a large deer heard were found throughout the park.

Site 4: Catharpin Run - Longitude -77°32'52" Latitude 38°50'39"

This site drains a wooded stretch that contains a few houses. Most of the year, flow rates were relatively rapid. Samples were collected at Robin Dr.

Site 5: Buckhall Branch - Longitude -77°26'12" Latitude 38°44'39"

The stream at this site collects waters from a neighborhood of single family homes and a few small farms. The samples were collected from a bridge on Signal Hill Road. An above ground pool was constructed during sampling on the land immediately upstream of the site. Just above that property was a small farming operation with horses, pigs, chickens, geese, and other animals.

Site 6: Flat Branch - Longitude -77°29'13" Latitude 38°46'54"

This site is located in a wooded buffer between housing developments. Sampling was done at the Lomond Dr bridge. Wildlife have been sighted repeatedly in and around the water of this site. Litter from runoff and local use have also been found at this site.

Site 7: South Run - Longitude -77°40'00" Latitude 38°46'10"

The creek drains a wooded zone that includes farms and houses. This site is on the creek as it enters Lake Manassas. Sampling was done at the Buckland Mill Rd bridge. Most of the year, the water here was backed up and showed little flow.

Site 8: Broad Run - Longitude -77°32'05" Latitude 38°44'13"

This part of the creek had considerable flow. It drains a wooded area and was buffered from new houses by a grassy plain. Sampling was done from the below the Nokesville Rd (Rt 28) bridge.

Site 9: Lower Kettle Run – Bristow Manor - Longitude -77°32'00" Latitude 38°42'10"

This site is just past a neighborhood of single family homes and in the middle of a golf course. Sampling was done at Valley View Dr bridge. The stream at this site was obstructed by debris on several occasions, but usually showed strong flow.

Site 10: Upper Kettle Run – West - Longitude -77°35'38" Latitude 38°42'22"
This site is in a wooded area that contains several horse trails. The waters also drain from older housing. Samples were taken from the bridge on Reid La (Rt 657).

Sample collection and preparation is the same as detailed in previous studies. (Graves 2003, Porter 2003, Booth 2003).

Samples were collected through immersion of a "Nalgene" container into the water. Samples were collected in the center of the flow when collected from a bridge, and within the main flow when collected from the banks. For sediment samples, bed sediment was agitated first before waters were collected. Duplicate samples were collected immediately following the original sample.

After collection, samples were immediately placed in ice. The samples remained in an ice filled cooler throughout transportation to the lab. A blank sample was included in every batch of samples to measure temperature conditions upon reaching the lab. Temperature measurement of the blank was used to indicate the sample temperatures. Blank samples were below 1°C every month. Samples were transported to the lab according to EPA methods 1603 and 1600. All samples reached the lab in less than 6 hours in accordance with Bordner et al. 1978.

III. Processing

Samples were processed in accordance with EPA Method 1603: Escherichia coli (*E. coli*) in water by membrane filtration using modified membrane-thermotolerant Escherichia coli agar (Modified mTEC). Upon arrival at the laboratory, each sample was given a presumptive test for *E. coli* using Colilert (Idexx). The following day, positive samples were filtered through four 40mm 10 micron filters in two aliquots between 1 and 20ml. One filter of each concentration was then placed on either modified mTEC (BD Diagnostic Systems) or modified *Enterococcus* agar (mENT, Fischer). Modified mTEC plates were placed in the 35°C incubator for one to two hours before 22 hours in a 44.5°C water bath incubation. mENT plates were incubated in the 35°C incubator for 48 hours.

After incubation, dark purple colonies, positive for *E. coli*, were counted on modified mTEC plates. On the mENT plates the deep red colonies, positive for *Enterococcus*, were counted. The colony forming units (CFUs) of *Enterococcus* and *E. coli* were then calculated per 100 ml sample.

Every month 24 *Enterococcus* colonies were randomly picked from each sample's mENT plates. These colonies were then aseptically transferred to Enterococcoseal Broth (Ent Broth) wells of a 96 well microtiter plate in order to grow as isolated colonies. This plate was then incubated at 35°C for 48 hours. The growth of *Enterococcus* was confirmed by a change of broth color from yellow to black, caused when the *Enterococcus* isolates hydrolyze esculin. Antibiotic Resistance Analysis (ARA) was then performed on the isolates.

Every quarter, 24 *E. coli* colonies were randomly picked from the mTEC plates and transferred in the same manner to Colilert (Idexx) broth in a 96 well plate. This plate

was incubated at 37°C for 24 hours. The growth of *E. coli* was confirmed through fluorescence under long wave ultraviolet radiation. ARA was then performed on the *E. coli* isolates.

A. ARA

Antibiotic Resistance Analysis (ARA) was performed on both *Enterococcus* and *E. coli* isolates. *Enterococcus* ARA was performed monthly on the samples successfully grown in the Ent Broth. *E. coli* ARA was performed on the quarterly samples grown in the Colilert broth. Each analysis used different concentrations of antibiotics mixed with 1% Tryptic Soy Agar (TSA). Colonies were transferred onto TSA plates using a 48 prong multiplater (Sigma, Inc.) to deposit 5µl of colony broth onto the plates. *Enterococcus* ARA plates were incubated at 37°C for 48 hours while *E. coli* ARA plates incubated 24 hours. *Enterococcus* ARA used the nine antibiotics at the levels in Table 1 made from the stock solutions of Table 2 as well as a control plate with no antibiotics. *E. coli* ARA used seven antibiotics at levels in Table 1 made from the stock solutions of Table 2. Stock solutions were remade at least quarterly. Levels were determined by previous studies (Wiggins et al 1999, Hagedorn et al 1999).

Following incubation, plates were read for growth or no growth of colonies. Colony growth consisted of at least a complete ring of cell growth at the edge of the 5µl inoculation. No growth consisted of any area of inoculation lacking a complete circle of growth at the edge of the growth area. All growth data was entered into SAS-JMP (v. 5.0.1, SAS Inst., Cary, NC) and Excel 2000 (Microsoft Corp, Redmond, WA).

Table 1. Antibiotic Concentrations after addition to TSA

Antibiotic	Plate Concentrations (μg/L)	
	Enterococcus	*E. coli*
Amoxicillin	2.5	-
Cephalothin	10, 15, 30, 50	15, 25, 35
Chlorotetracyclin	60, 80, 100	-
Erythromycin	10, 15, 30, 50	60, 70, 90, 100
Neomycin	40, 60, 80	2.5, 5.0, 10
Oxytetracycline	20, 40, 60, 80, 100	2.5, 5.0, 7.5, 10, 15
Rifampicin	-	60, 75, 90
Streptomycin	40, 60, 80, 100	2.5, 5.0, 7.5, 10, 15
Tetracycline	10, 15, 30, 50, 100	2.5, 5.0, 7.5, 10, 15
Vancomycin	2.5	-

Table 2. Antibiotic Stock Solution Preparations

Antibiotic	Commerical Formulation	Solvent	Stock Concentration (mg/mL)
Amoxicillin	Amoxicillin	1:1 water:methanol	2.5
Cephalothin	Cephalothin	Distilled water	10
Chlorotetracycline	Chlorotetracycline HCl	1 N NaOH	10
Erythromycin	Erythromycin	1:1 water:ethanol	10
Neomycin	Neomycin Sulfate	Distilled water	10
Oxytetracycline	Oxytetracycline HCl	1:1 water:methanol	10
Rifampicin	Rifampicin	Methanol	2.5
Streptomycin	Streptomycin Sulfate	Distilled water	10
Tetracycline	Tetracycline HCl	Methanol	10
Vancomycin	Vancomycin Sulfate	1:1 water:ethanol	10

B. Fluorometry

All water samples were measured for fluorescence as compared to µg/L of Fluorescent Brightener 28 (FB28) standard. This was done on the Turner Systems AU-10 Fluorometer. Before each month's measurement the fluorometer was calibrated with FB28 at concentrations of 15, 30, 60, 125, 250 and 500 µg/L. Blanking, at 3 – 4%, was conducted with deionized water. The span was set to between 38 and 44%. The 250 µg/L FB28 solution was run as the standard solution. Standard were used for no more than three months. As the calibration curve did not vary significantly and a time of suspension was unknown for the FB28, no adjustments were made to the data based on the calibration curve. Excitation wavelengths were filtered to between 300-400nm and observation wavelengths 410-600nm.

C. PFGE

Pulse Field Gel Electrophoresis (PFGE) was performed as detailed in Simmons and Herbein, 1998; Simmons et al., 1995; Simmons et al., 2000. Twenty four *E. coli* samples split between January and December were processed through PFGE. All isolates characterized by PFGE were first verified as *E. coli* using the API 20E test strips (bioMerieux sa). Preparation of cell suspension plugs, restriction endonuclease digestion with NotI, and specifications for pulsed field gel electrophoresis with the BioRad GenePath System have been described and were identical to that used for the USGS method comparison study (Stoeckel et al. 2004). Gels were photographed with a Polaroid DS34 camera, the prints were scanned (Microtek ScanMaker III) into a computer and

bands were recognized and recorded with SigmaGel (v. 1.0) software. Results were compared to the library of known PFGE band patterns.

IV. Statistical Analysis

Isolates from ARA and PFGE underwent a variety of processes in Excel and JMP. Isolates were compared in Excel for uniqueness and repeats were calculated for each month, site and for the entire set of samples. Clones were converted into augmented doubles according to the combinatoric: number of clones choose 2 (mathematical formula: NC2). Population size was estimated using Reverse Birthday Analysis (Ritter 1995) and recalculated throwing out common isolates that repeated more than 3 standard deviations above the average. All results of ARA and PFGE were categorized in JMP according to the discriminant analysis of the known source library (KSL).

A. Known Source Isolate Protocol

Known Source Isolates (KSI) were obtained from fecal samples through a three step process. A portion of the fecal sample was suspended in sterile water. Between 0.05 and 0.5 ml of suspension was spread plated onto mENT and mTEC plates. Isolates were obtained through incubation, transfer and re-incubation in 96 well plates as described above. *E. coli* and *Enterococcus* isolates were then subjected to ARA as described above.

B. KSL Creation

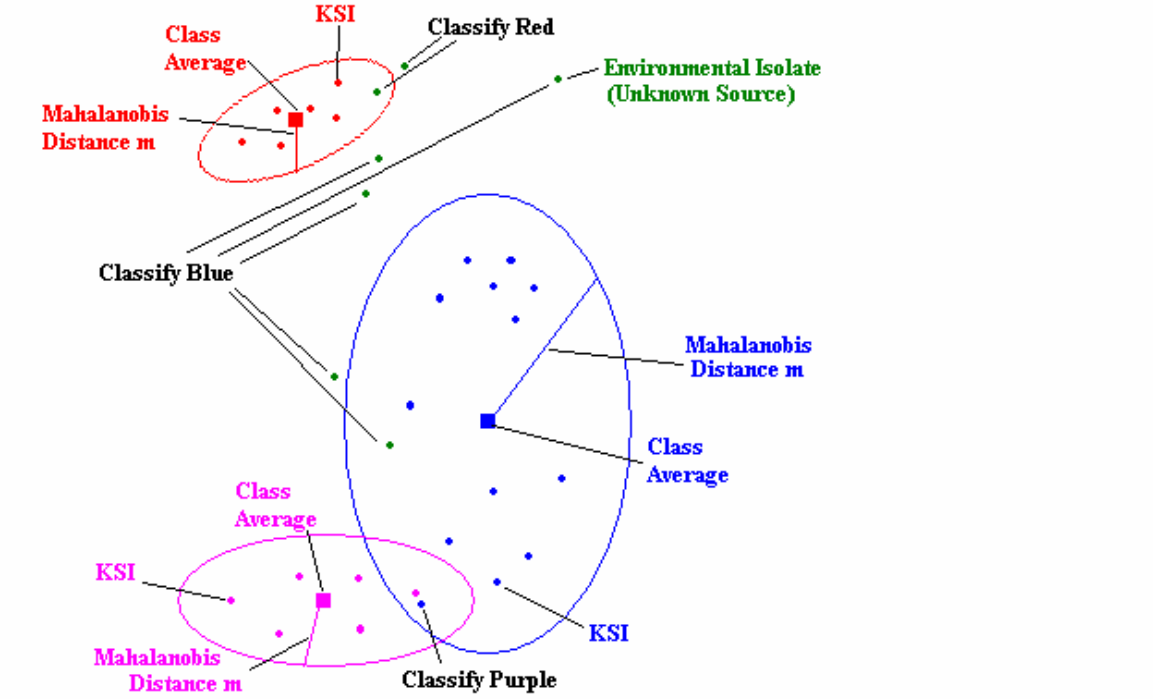
The Known Source Library (KSL or library) was created from the compiled Virginia and DC region known source isolates of previous studies (Porter 2003, Graves 2003, Booth 2003) as well as one half the known source isolates collected concurrent to

sampling. Only unique isolates were included in the library. Unique isolates with identical ARA patterns found in multiple source categories were included as one isolate of each category. Using JMP Linear Discriminant Analysis (LDA) was run on the data both in numeric format of highest resistance, highest isolate resistance below any antibiotic concentration where no growth occurred, a combination of the two and in binary format of all 31 variables for *Enterococcus*. *E. coli* was only run in the binary format with all 28 variables. Prior to LDA, Ward based Clustering was done to determine loci within the each category. Categories were subdivided into up to six separate loci based on the dissimilarity between clusters.

C. Linear Analysis Discriminant (LDA)

LDA classifies isolates by creating a generalization for each source category and comparing the probability that the isolate is within each separate source category (Figure 3). For each source class an average, standard deviation and covariance matrix is calculated from the KSL. The standard deviation and the covariance matrix are used to create a class specific unit of distance (Mahalanobis distance). Probability that the isolate is in a particular class is inversely proportional to the class specific distance of the isolate from the class average. Therefore the more similar the ARA pattern of the isolate is to he average of the class, the higher the probability the isolate is of the class. The probability calculation for each class is made independently. Each isolate is classified into only its most probable class.

Figure 3. Linear Discriminant Analysis



Unlabeled isolates classify according to their color.

D. KSL Analysis

Five measurements were obtained for the KSL in order to assure its quality. Rates and ratios of self-classification (Self) of KSI's included in the KSL were calculated, artificial clustering was inspected, a challenge set of KSI's not included in the KSL (CSC) was analyzed and two field verification sites were compared to expected data classification. The first two measurements are library internal measurements (Labeled KSI in Figure 3), independent of outside data, while the second two are dependent on library external data.

The rates and ratios of self-classification were sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), average rate of correct classification (ARCC) and minimum detectable percentage (MDP) as outlined by Stewart et al 2004 and an overall rate of correct classification (RCC). A true positive is a correctly classified isolate. A false positive is an isolate misclassified into a particular category. A true negative is an isolate correctly classified out of a category. A false negative is an isolate incorrectly classified out of a particular category. Sensitivity is the proportion of true positives that are correctly identified by the test. Specificity is the proportion of true negatives that are correctly identified by the test. PPV was calculated as the proportion of true positives over true and false positives. NPV was calculated as the number of true negatives over the total number of true and false negatives. ARCC was the sensitivity of each category averaged between the categories. MDP was four times the standard deviation of the average number of the false positives from each

category. RCC is the number of true positives for all categories over the total number of isolates in the library.

Artificial clustering of the library was checked for by randomly reassigning isolates into four equal categories. DA was run on these categories as well as these categories further split into up to six loci each. Artificial clustering was found to exist when categories substantially differed from a rate of classification of 25% (100% divided by the number of categories).

The challenge set was created from a randomly selected half of the KSI obtained during the course of this study. The challenge set was classified according to the DA of the KSL. The rate of correct classification was determined.

The field verification was determined by comparison of classification results of the library on the samples from sites with the known conditions. These conditions were withheld from the researcher until preliminary results were given. Classification was compared to known parameters. This is a qualitative measurement of the library.

V. References

Booth, A. M., C. Hagedorn, A. K. Graves, S. C. Hagedorn, and K. H. Mentz. 2003. Sources of Fecal Pollution in Virginia's Blackwater River. J. Environ. Engineering 129:547-552.

Bordner R., J.A. Winter, P.V. Scarpino (eds.). 1978. Microbiological Methods for Monitoring the Environment: Water and Wastes, EPA-600/8-78-017. Office of Research and Development, USEPA. Washington, D.C.

Chapman,A., Hagedorn, C., Saluta, M. 2004. Identifying Sources of Fecal Pollution in Impaired Waters in Prince William County, Virginia. American Society for Microbiology, General Meeting. New Orleans, Lousiana.

Connaughten, S.T. 2005. State of the County Address. http://www.co.prince-william.va.us/default.aspx?topic=010010000810002865. Prince William County.

Graves, A. K. 2003. Identifying Sources of Fecal Pollution in Water as a Function of Sampling Frequency Under Low and High Stream Flow Conditions. Ph.D. diss. Virginia Tech.

Hagedorn, C., Robinson, S. L., Filtz, J. R., Grubs, S. M., Angier, T. A., and Reneau, R. B. 1999. Determining Sources of Fecal Pollution in a Rural Virginia Watershed with Antibiotic Resistance Patterns in Fecal Streptococci. Appl. Environ. Microbiol. 65:5522-5531.

PWCOHCD. 2005. DRAFT Consolidated Housing and Community Development Plan Fiscal Years 2006-2010 and FY06 Annual Action Plan. Prince William County Office of Housing and Community Development: Prince William, VA.

Simmons, G. M., Jr., S. A. Herbein, and C. M. James. 1995. Managing Nonpoint Fecal Coliform Sources to Tidal Inlets. Water Resource Update 100:64-74.

Simmons, G. M., Jr., and S. A. Herbein. 1998. Shellfish and Water Column Comparison of Fecal Coliform Diversity Using NotI DNA Fingerprints of Escherichia coli Generated by Pulse Field Gel Electrophoresis. Final Report for the Virginia Coastal Resource Management Program Department of Environmental Quality, Richmond, VA.

Simmons, G. D. Waye, S. Herbein, S. Myers, E. Walker. 2000. Estimating Nonpoint Fecal Coliform Sources in Northern Virginia's Four Mile Run Watershed. Virginia Water Research Symposium 2000, Blacksburg, VA.

Stewart, J.R., Robinson, B., Hyer, K., Hagedorn, C., Whittam, T.S., Wilbur, J. 2005. Microbial Source Tracking Using Indicator Organisms. American Society for Microbiology General Meeting, Atlanta, GA.

Stoeckel, D. M., Mathes, M. V., Hyer, K. E., Hagedorn, C., Kator, H., Lukasik, J., O'Brien, T. L., Fenger, T. W., Samadpour, M., Strickler, K. M., and Wiggins, B. A. 2004. Comparison of seven protocols to identify fecal contamination sources using Escherichia coli. Environ. Sci. Technol. 38:6109-6117.

Szeles, C. L. 2003. Determining Sources of Fecal Contamination in Two Rivers of Northumberland County, Virginia. M.S. Thesis. Virginia Tech.

Porter, K. R. 2003. Identifying Sources of Fecal Pollution in Washington D.C. Waterways. M.S. Thesis. Virginia Tech.

Wiggins, B. A., Andrews, R. W., Conway, R. A., Corr, C. L., Dobratz, E. J., Dougherty, D. P., Eppard, J. R., Knupp, S. R., Limjoco, M. C., Metttenburg, J. M., Rinehardt, J. M., Sonsino, J., Torrijos, R. L., and Zimmerman, M. E. 1999. Use of Antibiotic Resistance Analysis to Identify Nonpoint Sources of Fecal Pollution. Appl. Environ. Microbiol. 65:3483-3486.

Chapter 4. Results

I. Monitoring Results

Environmental monitoring at the sample sites included quantification of *Enterococcus* and *E. coli* in the water. The sample sites are judged on a criterium of a single sample not exceeding 235 colony forming units (CFU) per 100mL of water *E. coli* 10% of the time. The geometric mean for these sites had to remain less than 126CFU/100mL for *E. coli* according to EPA 1986 ambient water standards (Table 3). *Enterococcus* standards require a geometric mean less than 33CFU/100mL. These standards are significantly relaxed depending on the amount and type of use (Table 3), however the sites in this study were judged according to the most stringent standards.

E. coli monitoring showed that 9 out of 10 sites were above 235CFU/100mL in 10% or more of the samples (Table 4). Only Upper Bull Run was below this standard of impairment. The geometric mean of Upper Bull Run, Lower Bull Run, South Run and Broad Run was below the 126CFU/100mL standard (Table 4), however no site met the 33CFU/100mL *Enterococcus* standard (Table 5). According to research from US EPA 1986 such sites pose a health hazard due to their likelihood to cause gastrointestinal disease in greater than 8 of 1000 swimmers. The South Run and Bull Run sites are the only sites with water deep enough for swimming. Both Bull Run sites had swimmers present during sampling in June 2005.

Table 3. 1986 Criteria for Indicators for Bacteriological Densities [a]

Freshwater	Acceptable swimming associated gastroenteritis rate per 1000 swimmers	Steady state geometric mean indicator density	Single sample maximum allowable density			
			Designated beach area (upper 75% C.L.)[bc]	Moderate full body contact recreation (upper 82% C.L.)	Lightly used full body contact recreation (upper 90% C.L.)	Infrequently used full body contact recreation (upper 95% C.L.)
Enterococci	8	33CFU [d] /100 ml	61	78	107	151
E. coli	8	126CFU /100 ml	235	298	409	575

[a] Table is a Partial Reproduction from the US EPA 1986 Ambient Water Quality Criteria for Bacteria

[b] Where C.L. is confidence limit

[c] Standard applied to sites tested in study

[d] Colony Forming Units per 100ml of water

Table 4. *E. coli* Monthly Sampling Counts (CFU/100mL)[a]

Site	June 04	July	Aug	Sept	Oct	Nov	Dec	Jan	Feb	Mar	April	May	June 05	Geometric Mean[b]
Upper Bull Run	172.5	50	218.75	47.5	60	120	145	35	57.5	32.5	197.5	57.5	210	***86***
Lower Bull Run	50	15	280	40	545	925	35	50	1	95	402.5	120	45	***73***
Youngs Branch	292.5	200	205	375	385	480	730	105	45	20	160	147.5	285	189
Catharpin	132.5	87.5	220	990	295	265	850	90	132.5	15	167.5	42.5	105	152
Buckhall Branch	990	105	952.5	**5760**[c]	680	715	655	277.5	57.5	7.5	80	212.5	**2150**	339
Flat Branch	720	1067.5	607.5	**3000**	200	105	955	37.5	110	10	110	175	55	206
South Run	32.5	5	12.5	115	320	125	177.5	177.5	42.5	40	240	35	695	***76***
Broad Run	157.5	32.5	410	202.5	70	70	287.5	27.5	10	27.5	202.5	40	220	***82***
Lower Kettle Run	167.5	130	830	300	355	590	6480	62.5	95	70	360	287.5	130	270
Upper Kettle Run	**1360**	**1135**	**4080**	395	**1595**	**915**	**1520**	**1005**	105	**245**	**920**	**825**	695	819
x-bar[d]	407.5	282.8	781.6	1122.5	450.5	431.0	1183.5	186.8	65.6	56.3	284.0	194.3	459.0	322
s[e]	456.7	435.6	1196.6	1856.0	445.5	340.1	1915.6	297.9	43.7	71.9	245.2	237.1	640.2	

[a] Colony Forming Units per 100ml of water

[b] Mean of the monthly counts for each site as calculated by (June 04*July*Aug…*June 05)^1/13

Bold Italic means meet water quality standards

[c] Counts 2 or more Standard Deviations above average are in **Bold**

[d] Average across all sites for the month

[e] Standard Deviation across all sites for the month

Table 5. *Enterococcus* Monthly Sampling Counts (CFU/100mL)[a]

Site	June 04	July	Aug	Sept	Oct	Nov	Dec	Jan	Feb	Mar	April	May	June 05	Geometric Mean [b]
Upper Bull Run	502.5	135	667.5	122.5	175	55	465	10	100	5	155	30	205	101
Lower Bull Run	35	55	780	92.5	215	**1205** [c]	97.5	35	72	**60**	**347.5**	10	100	105
Youngs Branch	**2507.5**	650	980	860	555	340	1200	30	**1100**	5	107.5	37.5	260	276
Catharpin	195	290	622.5	3600	360	245	580	37.5	105	7.5	135	10	130	156
Buckhall Branch	1620	425	1440	**3920**	385	535	670	85	55	27.5	40	70	**1285**	310
Flat Branch	502.5	**1960**	1315	**4800**	350	80	1720	112.5	307.5	25	45	92.5	530	331
South Run	82.5	85	85	780	705	140	197.5	40	17.5	2.5	142.5	12.5	100	76
Broad Run	55	160	947.5	485	235	100	440	10	105	12.5	120	0	192.5	124
Lower Kettle Run	340	397.5	**1940**	652.5	**1010**	555	**4800**	100	400	**77.5**	90	20	640	359
Upper Kettle Run	1165	490	1502.5	**4400**	**1300**	**970**	1820	**195**	225	**72.5**	**262.5**	**222.5**	315	558
x-bar [d]	700.5	464.8	1028	1971.3	529	422.5	1199	65.5	248.7	29.5	144.5	50.5	375.8	309
s [e]	818.7	559.9	533.6	1940.8	372.5	396.8	1399.1	58.1	322.7	29.5	94.9	67.1	367.1	

[a] Colony Forming Units per 100ml of water

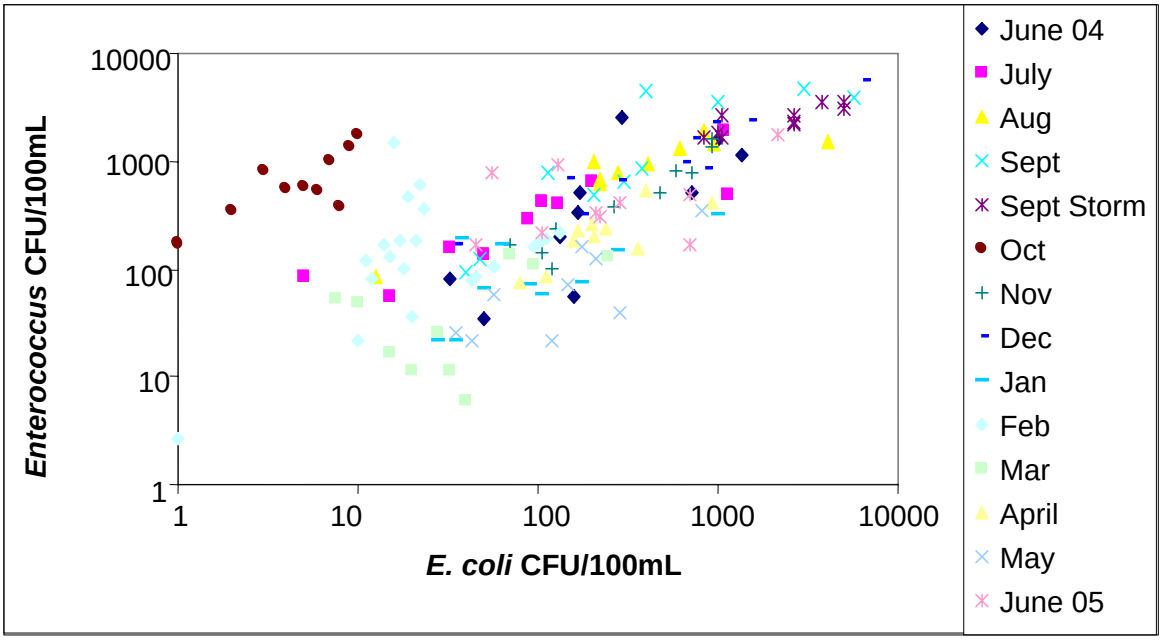
[b] Mean of the monthly counts for each site as calculated by (June 04*July*Aug…*June 05)^1/13

[c] Counts 1 or more Standard Deviations above average are in **Bold**

[d] Average across all sites for the month

[e] Standard Deviation across all sites for the month

Figure 4. Scattergram of Counts *Enterococcus* vs. *E. coli* Counts



Trendlines (Not shown due to lack of fit)
Linear: *Enterococcus* = 0.7351**E. coli* + 222.21 $R^2 = 0.5373$
Logarithmic: *Enterococcus* = 372.09Ln(*E. coli*) - 1349.5 $R^2 = 0.3631$

Particularly high bacteriological counts were found at two sites. The Buckhall Branch site showed considerable contamination (Tables 4 and 5). A small farm was located 20m upstream from the sample site. The Upper Kettle Run site was also particularly contaminated. At Upper Kettle Run 8 of the 13 samplings were more than two standard deviations above the month average for all ten sites (Table 4). The bacterial counts at Upper Kettle Run have a geometric mean more than six times higher than the *E. coli* standard. Conversely, Upper and Lower Bull Run , South Run and Broad Run all had geometric means of less than the 126CFU/100ml threshold of impairment. Counts at Upper Bull Run were below the 235CFU/100ml single sample threshold for all samplings. Upper Bull Run has no indicators of *E. coli* impairment.

The variability between the sites was great enough that 8 of the 13 months showed a standard deviation greater than the average. Most of the variability was accounted for by the counts of Upper Kettle Run. The high counts of Buckhall Branch, in Spetember and June 05, the highest counts for that site, accounted for most of the variability of September and June 05 samplings.

This variability between sites was also reflected to a lesser extent in the *Enterococcus* data (Table 5). The between site *Enterococcus* variation had a standard deviation greater than the average 5 of the 13 months, of which June 04, July, December and May also had a standard deviation greater than average for the *E. coli* data. Upper Kettle Run was not the main contributor to *Enterococcus* variability (as it was for *E. coli*) and high counts were distributed throughout the sites.

Enterococcus and *E. coli* count data did not correlate well (Figure 4). Individual sites, such as Buckhall Branch, showed highest *Enterococcus* counts the same months

that *E. coli* counts were highest (Table 4 and 5). The *E. coli* counts did not predict the *Enterococcus* counts and vice versa. Correlation attempts on linear and logarithmic scales had R^2 values of 0.54 and 0.36 respectively. The linear fit had 8% more variation explained by regression than by error. The logarithmic regression explained more than two-thirds of the variation by regression error.

Differentiation between sites and months was performed using the Student's T test and Tukey-Kramer HSD tests of multicomparisons (Tables 6-9). The Student's T test distinguished pairwise similarity as membership of a group of indistinguishable values. Groups were created inclusive of all values between two endpoints that were indistinguishable from one another. The Tukey-Kramer HSD test designated groups of values when all compared values could potentially be samples from a single population set. In both of these tests values that are indistinguishable are indicated by the presence of an identical letter in the same column. Both tests were always used with alpha=0.05.

Monthly *E. coli* data in Table 6 indicated that only two months, December and September, could be separated from most of the rest by Student's T. December, which was different from groups B and C, had higher counts than January through May, June 04, and July. September, which was different than group C, had higher counts than January through May and July. The counts for all other months could not be distinguished from each other at an alpha=0.05. Despite the differences between individual months, Tukey-Kramer HSD indicated that all *E. coli* count data could be considered from a single sample set.

Table 6. *E. coli* Difference Between Months (alpha =0.05) [a]

Month	Student's T [b]			Tukey-Kramer HSD [c]	Average [d] (CFU/100mL)
Dec [e]	A [f]			A	1183.5
Sept	A	B		A	1122.5
Aug	A	B	C	A	781.6
June 05	A	B	C	A	459.0
Oct	A	B	C	A	450.5
Nov	A	B	C	A	431.0
June 04		B	C	A	407.5
April			C	A	284.0
July			C	A	282.8
May			C	A	194.3
Jan			C	A	186.8
Feb			C	A	65.6
Mar			C	A	56.3

[a] Student's T and Tukey-Kramer HSD comparison tests were calculated at alpha= 0.05
[b] Student's T tests difference between individual pairs of months
[c] Tukey-Kramer HSD tests whether all months could be a members of a single set
[d] Average monthly count across all sites of colony forming units per 100 ml of water
[e] Each month includes data from every sample site
[f] Presence of the same letter in a column indicates similarity

Table 7. *E. coli* Difference Between Sites (alpha =0.05)[a]

Site	Student's T [b]		Tukey-Kramer HSD [c]	Average [d] (CFU/100mL)
Upper Kettle Run [e]	A [f]		A	1138.1
Buckhall Branch	A		A	972.5
Lower Kettle Run	A	B	A	758.3
Flat Branch	A	B	A	550.2
Youngs Branch		B	A	263.8
Catharpin		B	A	261.0
Lower Bull Run		B	A	200.3
South Run		B	A	155.2
Broad Run		B	A	135.2
Upper Bull Run		B	A	108.0

[a] Student's T and Tukey-Kramer HSD comparison tests were calculated at alpha= 0.05
[b] Student's T tests difference between individual pairs of sites
[c] Tukey-Kramer HSD tests whether all months could be a members of a single set
[d] Average site count across all months of colony forming units per 100 ml of water
[e] Each sample site includes data from every month
[f] Presence of the same letter in a column indicates similarity

Site comparison by Student's T of *E. coli* (Table 7) showed that Upper Kettle Run and Buckhall Branch had higher counts than the remaining sites. The remaining eight sites were indistinguishable in *E. coli* concentration. Tukey-Kramer HSD(alpha=0.05) again indicated all sites of *E. coli* count data could again be considered from a single set.

Monthly comparison of *Enterococcus* data (Table 8) showed much more differentiation than the *E. coli* data. This was a byproduct of the relatively lower standard deviation in the *Enterococcus* data (Table 5). September was distinguished from all other months by Student's T pairwise comparison. Tukey-Kramer HSD determined that September counts were only considered from the same set as December and August. December was different from July, November, January through May, and June 05 by pairwise comparison. December could not be considered part of the same group of counts as March by Tukey-Kramer HSD. August was similar by pairwise comparison only to June 04, July, October through December and June 05. All remaining months could be considered as being part of the same set.

Site comparison of *Enterococcus* data (Table 9) showed lesser differences between sites than between months. All sample sites could be considered part of the same set of *Enterococcus* counts. In pairwise comparison Upper Kettle Run had higher counts than Bull Run, Broad Run and South Run. Upper Kettle Run had higher counts than Upper Bull Run, Broad Run and South Run. All other sites had indistinguishable counts at alpha=0.05.

Table 8. *Enterococcus* Difference Between Months (alpha =0.05) [a]

Month	Student's T [b]				Tukey-Kramer HSD [c]			Average [d] (CFU/100mL)
September [e]	A [f]				A			1971.3
December		B			A	B		1199.0
August		B	C		A	B	C	1028.0
June 04		B	C	D		B	C	700.5
October		B	C	D		B	C	529.0
July			C	D		B	C	464.8
November			C	D		B	C	422.5
June 05			C	D		B	C	375.8
February				D		B	C	248.7
April				D		B	C	144.5
January				D		B	C	65.5
May				D		B	C	50.5
March				D			C	29.5

[a] Student's T and Tukey-Kramer HSD comparison tests were calculated at alpha= 0.05
[b] Student's T tests difference between individual pairs of months
[c] Tukey-Kramer HSD tests whether all months could be a members of a single set
[d] Average monthly count across all sites of colony forming units per 100 ml of water
[e] Each month includes data from every sample site
[f] Presence of the same letter in a column indicates similarity

Table 9. *Enterococcus* Difference Between Sites (alpha =0.05) [a]

Site	Student's T [b]			Tukey-Kramer HSD [c]	Average [d] (CFU/100mL)
Upper Kettle Run [e]	A [f]			A	995.8
Flat Branch	A	B		A	910.8
Lower Kettle Run	A	B	C	A	847.9
Buckhall Branch	A	B	C	A	812.1
Youngs Branch	A	B	C	A	664.0
Catharpin	A	B	C	A	486.0
Lower Bull Run		B	C	A	238.8
Broad Run			C	A	220.2
Upper Bull Run			C	A	202.1
South Run			C	A	183.8

[a] Student's T and Tukey-Kramer HSD comparison tests were calculated at alpha= 0.05
[b] Student's T tests difference between individual pairs of sites
[c] Tukey-Kramer HSD tests whether all months could be a members of a single set
[d] Average site count across all months of colony forming units per 100 ml of water
[e] Each sample site includes data from every month
[f] Presence of the same letter in a column indicates similarity

II. Library (Training Data)

A. Summary

Four libraries were developed and six statistical and five methodological processes were performed on each. As detailed in the Materials in Methods section, the libraries were created from both local and regional isolates. The isolate patterns in each library were all unique, and a clone of each antibiotic resistance pattern was only included if that pattern was found in multiple classes. The libraries were compared by statistical measure (Table 11), challenge set classification (Table 22) and known environmental factor detection (Table 23).

Libraries created with isolates included from both multiple classes and classes that conflicted with isolates from the OQB (Table 11). These were found to have lower average rates of correct classification (ARCC) and lower rates of correct challenge set classification (CSC). The non-clustered binary library without interclass or spatial conflicts correctly classified isolates at the highest rate, however statistical assumptions detailed in the discussion section required the use of the combination library.

Table 10. Library Known Source Isolates

	Human	Livestock	Pet	Wildlife	Total
Total Unique ARA Patterns	339	192	82	542	918 [a]
Single [b] Class ARA Patterns	191	107	57	381	736
Two [c] Class ARA Patterns	100	42	9	113	132
Three [d] Class ARA Patterns	42	38	10	42	44
Four [e] Class ARA Patterns	6	6	6	6	6

[a] Total isolates of all sources is not the sum of each source do to isolates of multiple classes
[b] ARA Patterns found only in one source category
[c] ARA Patterns found only in two source categories
[d] ARA Patterns found only in three source categories
[e] ARA Patterns found in all four source categories

B. Design

Libraries were created from a pool of 918 unique ARA patterns (Table 9). Each pattern corresponded to either an isolate of a particular class (single class isolate) or isolates from two or more classes (multiclass isolates). Of the 918 unique ARA patterns, 736 corresponded to single class isolates. Only 6 patterns corresponded to all four classes and could not be used for differentiation between classes. The largest class of known source isolates (KSI's) was wildlife with 381 single class isolates and 161 multiple class isolates. Pets had the least number of isolates, and 30% of the Pet isolates were multiclass isolates.

Four methods of library creation were considered (Table 11): High, Last, Binary and Combination. Each method treated the isolate growth data differently. These methods were also used to treat the environmental isolate data for subsequent analysis. The combination method was chosen for isolate processing.

One method of library creation used only the highest concentration of antibiotic to which the isolate resisted (High dataset). This ignored lower concentrations at which the isolate failed to grow.

A second method of library creation used the highest concentration of antibiotic before isolates failed to grow (Last dataset). This method ignores all higher concentrations of antibiotic above the concentration at which the isolates fails to grow.

Table 11. Data Interpretation of Example Data According Library Design

		Design Interpretation			
Tetracycline Concentration	Growth [a]	High [b]	Last [c]	Binary [d]	Combination [e]
10 μg/ml	Yes	Ignores	10 μg/ml	Growth	10 μg/ml
15 μg/ml	No	Ignores	No growth	No growth	No growth
30 μg/ml	Yes	Ignores	Ignores	Growth	Ignores
50 μg/ml	No	Ignores	Ignores	No growth	Ignores
100 μg/ml	Yes	100 μg/ml	Ignores	Growth	100 μg/ml
Recorded Value		100 μg/ml	10 μg/ml	1,0,1,0,1	100 μg/ml, 10 μg/ml

[a] Indicated by at least a complete ring of cell growth at the edge of the 5μl inoculation
[b] Dataset which uses only the highest antibiotic concentration where growth occurred
[c] Dataset which uses the antibiotic concentration where growth occurred below the lowest concentration where no growth occurred
[d] Dataset which uses growth data from all antibiotic concentrations as binary values
[e] Dataset which uses the highest antibiotic concentration where growth occurred and the antibiotic concentration where growth occurred below the lowest concentration where no growth occurred

The third method, the Binary dataset made the absence of growth of an isolate at a particular concentration a unique data point. The design was similar to banding pattern classification, where the presence or absence of any data point is significant when considered independently. In order to include all test data a binary sample set was run.

In order to create a library containing as much data as possible but following the statistical assumptions of linear discriminant analysis (USEPA 2005) a fourth category of library was created. This library combined both Last and High datasets into a combined resistance (Combination dataset).

Each data design was compared with a complete dataset and one in which conflicts were removed in Table 12. Two sets of statistics were calculated for each library. The designs were separated into two libraries by completeness of dataset (with or without conflicts). The measurements of each library was calculated for library classification of itself (Self) and challenge set classification (CSC), classification of the challenge set by the library. The rate of correct classification (RCC), the overall rate at which isolates were placed in the correct category, was highest for the Binary dataset without any conflicts. The average positive predictive value, the PPV of each class averaged, showed that the Binary or High datasets could classify isolates with the most accuracy. The class averaged negative predictive value (ANPV) showed that the Binary dataset was again most accurate in classification. The negative predictive value (NPV) is the rate of correctly classifying an isolate out of a class over all isolates classified out of a particular class.

The class averaged specificity of Table 12 showed that the Binary dataset, with the highest value. The specificity is the rate of isolates correctly classified out of a

particular class divided by all isolates not of that particular class. High values of specificity give a low chance of Type I error. Type I error is misclassifying an isolate out of a class. The specificity differs from the NPV in that the denominator of the specificity statistic has not been classified by the library.

The average rate of correct classification (ARCC), which also the class averaged sensitivity, had the highest value in the Binary dataset without conflicts (Table 12). Sensitivity is a measurement of the correctly classified isolates over all isolates of that class. This indicated that the Binary dataset had the lowest risk of Type II error. Type II error is misclassifying an isolate into a class. Sensitivity differs from PPV in that the denominator of the sensitivity statistic has not been classified by the library.

Average minimum detectable percentage (A-MDP) was only calculated for library self-classification in Table 12 because values depended on the number of isolates in each class. The calculations of this test are based on the relations between values within the library. Results of this test on challenge sets and data outside the library don't refer to these values and therefore invalidate the statistic. The A-MDP showed that the Binary dataset was most able to detect classified isolates.

Positive Prediction Value (PPV) is the rate of correct classifications given the classification in a particular area. This rate is the fraction of correctly classified isolates to all isolates classified into that class. The PPV projects a rate of correct classification of environmental isolates given the rate at which a known set of isolates are correctly classified.

The Binary dataset had the highest rates of correct classification (RCC) and the highest positive prediction value (PPV) of any library (Table 12). Positive predictive

values from the Combination dataset were better on average than those of the High or Last data libraries (Tables 12, 13).

Table 12. Library and Challenge Set Classification (CSC) According to Library Conflict Removal and Library Design

Dataset Representation	Binary [a]		Combination [b]		High [c]		Last [d]		Binary		Combination		High		Last	
Interclass conflict [e]	Yes		Yes		Yes		Yes		No		No		No		No	
Interspatial conflict [f]	Yes		Yes		Yes		Yes		No		No		No		No	
Classification Type [g]	Self	CSC	Self	CSC	Self	CSC	Self	CSC	Self	CSC	Self	CSC	Self	CSC	Self	CSC
RCC [h]	0.51	0.49	0.45	0.42	0.44	0.44	0.40	0.34	0.69	0.56	0.59	0.48	0.56	0.55	0.52	0.45
APPV [i]	0.47	0.40	0.46	0.31	0.44	0.36	0.41	0.27	0.63	0.42	0.57	0.32	0.52	0.43	0.51	0.32
ANPV [j]	0.83	0.76	0.81	0.74	0.81	0.75	0.80	0.72	0.88	0.77	0.85	0.75	0.84	0.77	0.84	0.75
A-Specificity [k]	0.83	0.77	0.82	0.74	0.81	0.75	0.80	0.72	0.89	0.76	0.87	0.75	0.85	0.78	0.85	0.75
ARCC [l]	0.53	0.40	0.51	0.30	0.49	0.38	0.45	0.25	0.67	0.39	0.63	0.33	0.58	0.42	0.57	0.31
A-MDP [m]	0.29		0.37		0.35		0.35		0.25		0.26		0.3		0.31	

[a] Dataset which uses growth data from all antibiotic concentrations as binary values
[b] Dataset which uses the highest antibiotic concentration where growth occurred and the antibiotic concentration where growth occurred below the lowest concentration where no growth occurred
[c] Dataset which uses only the highest antibiotic concentration where growth occurred
[d] Dataset which uses the antibiotic concentration where growth occurred below the lowest concentration where no growth occurred
[e] Isolates from two or more categories which have the same ARA pattern. A No value indicates removal from the library
[f] Isolates from two or more categories which have the same ARA pattern but different categories within OQB than without. A No value indicates removal of out of basin isolates
[g] Indicates whether statistics below were created from classification of the library(Self) or classification of the challenge set (CSC)
[h] Rate of Correct Classification. Values are the fraction of correctly classified isolates over all isolates
[i] Average Positive Predictive Value. Values are the fractional average of the correctly classified isolates of each source classes over the all isolates classified into the same source class

[j] Average Negative Predictive Value. Values are the fractional average of the isolates correctly classified out of each source classes over all isolates incorrectly classified out of the same source class

[k] Average Specificity. Values are the fractional average of the isolates correctly classified out of each source classes over all isolates not of the same source class

[l] Average Rate of Correct Classification. Values are the fractional average of the correctly classified isolates of each source classes over the isolates of that source class

[m] Average Minimum Detectable Percentage. Values are the average of the minimum fraction for a source to be considered not noise

Tables 12 and 13 indicate that the Binary dataset is the most accurate design. The Last dataset did not correctly classify any human isolates but half the pet isolates that it classified were correct. The combination dataset, which follows statistical assumptions of linear discriminant analysis (LDA), had highest PPV of human isolates but could only classify one quarter of isolates classified as pets were actually pet isolates.

The High dataset showed PPV's almost as large as the Combination dataset in Table 13. The human PPV showed the largest difference when 8% less of the isolates that the High dataset classified as human were human. The High data gave PPV's with for human and wildlife with little difference from the Binary dataset. The assumptions for LDA were valid for the High dataset.

Table 13. Positive Predictive Value (PPV) of Libraries without Interspatial or Interclass Conflicts as Affected by Design

	Human	Livestock	Pet	Wildlife	APPV [a]
Binary [b]	0.70	**0.68** [c]	0.33	**0.81**	0.63
Combination [d]	**0.74**	0.48	0.26	0.79	0.57
High [e]	0.66	0.45	0.21	0.77	0.52
Last [f]	0.00	0.17	**0.48**	0.63	0.32

[a] Average Positive Predictive Value. Values are the fractional average of the correctly classified isolates of each source classes over the all isolates classified into the same source class
[b] Dataset which uses growth data from all antibiotic concentrations as binary values
[c] Values are the fractional average of the correctly classified isolates of each source classes over the all isolates classified into the same source class. The largest value for each class is in **Bold**
[d] Dataset which uses the highest antibiotic concentration where growth occurred and the antibiotic concentration where growth occurred below the lowest concentration where no growth occurred
[e] Dataset which uses only the highest antibiotic concentration where growth occurred
[f] Dataset which uses the antibiotic concentration where growth occurred below the lowest concentration where no growth occurred

C. Library Processing

Four kinds of processes were attempted to increase classification rates of libraries: i) removal of conflicts in the data set, ii) relabeling poorly classified isolates into an "unknown" category, iii) clustering of classes to identify similar subclasses and iv) thresholding of poorly classifying isolates. Conflict removal was the most successful, although unknown and subclass creation had limited success (Tables 14-19).

i) For each dataset two types of conflicts were considered: interclass clones and spatial clones. For spatial clones, isolates from outside the basin that conflicted with those from the basin, removal showed a significant increase in challenge set classification (Table 12). This resulted in fractions of a percentage change increases in other measures of classification (Table 14). The largest of these changes was a 4% difference in A-MDP.

Interclass clones, isolates that were clonal across two, three or four classes were also removed. At the removal of isolates that were in three or more classes, the average rate of correct classification increased by more than the amount mandated by such a removal (Table 15). Removal of isolates in two categories showed an increase in all self-classification statistics except A-MDP and better rates of challenge set classification (Tables 12 & 15). Both of these isolate removal techniques significantly reduced the size of the libraries (Table 10).

Table 14. Library Classification as Affected by Interspatial Conflicts in a Combination Dataset

	Full Dataset	Interspatial Conflicts [a] Removed
RCC [b]	0.449	0.453
APPV [c]	0.456	0.461
ANPV [d]	0.815	0.816
A-Specificity [e]	0.820	0.821
ARCC [f]	0.512	0.506
A-MDP [g]	0.367	0.325

[a] Isolates from two or more categories which have the same ARA pattern but different categories within OQB than without. Removal leaves only the isolates from within the OQB
[b] Rate of Correct Classification. Values are the fraction of correctly classified isolates over all isolates
[c] Average Positive Predictive Value. Values are the fractional average of the correctly classified isolates of each source classes over the all isolates classified into the same source class
[d] Average Negative Predictive Value. Values are the fractional average of the isolates correctly classified out of each source classes over all isolates incorrectly classified out of the same source class
[e] Average Specificity. Values are the fractional average of the isolates correctly classified out of each source classes over all isolates not of the same source class
[f] Average Rate of Correct Classification. Values are the fractional average of the correctly classified isolates of each source classes over the isolates of that source class
g Average Minimum Detectable Percentage. Values are the average of the minimum fraction for a source to be considered not noise

Table 15. Library Classification as Affected by Interclass Conflicts [a] in a High Dataset

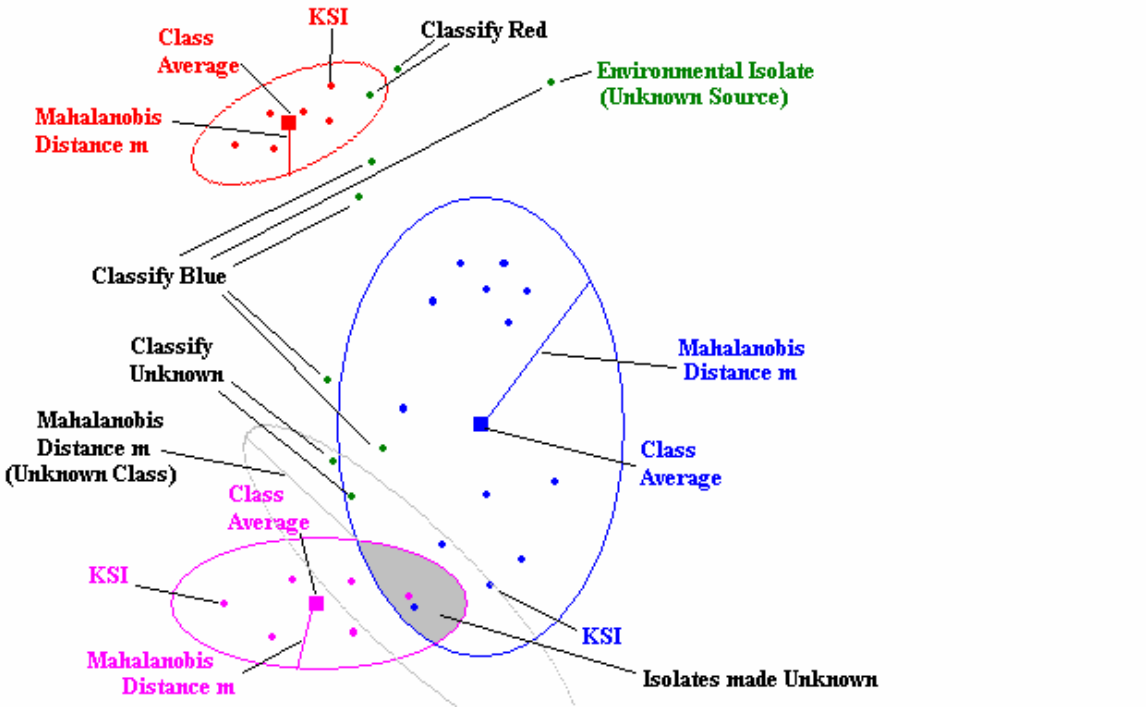
	With Conflicts	3 & 4 Class Conflicts Removed	All Interclass Conflicts Removed
RCC [b]	0.51	0.46	0.55
APPV [c]	0.47	0.45	0.52
ANPV [d]	0.83	0.81	0.84
A-Specificity [e]	0.83	0.82	0.85
ARCC [f]	0.53	0.51	0.58
A-MDP [g]	0.29	0.33	0.32

[a] Isolates from two or more categories which have the same ARA pattern. Removal leaves no isolate with that pattern in the library.
[b] Rate of Correct Classification. Values are the fraction of correctly classified isolates over all isolates
[c] Average Positive Predictive Value. Values are the fractional average of the correctly classified isolates of each source classes over the all isolates classified into the same source class
[d] Average Negative Predictive Value. Values are the fractional average of the isolates correctly classified out of each source classes over all isolates incorrectly classified out of the same source class
[e] Average Specificity. Values are the fractional average of the isolates correctly classified out of each source classes over all isolates not of the same source class
[f] Average Rate of Correct Classification. Values are the fractional average of the correctly classified isolates of each source classes over the isolates of that source class
g Average Minimum Detectable Percentage. Values are the average of the minimum fraction for a source to be considered not noise

ii) The creation of an unknown category reduced the effective size of the library. While these methods increased rates of artificial clustering, some benefit was observed (Table 16). The basic concept of unknown removal is shown in figure 5. In each method isolates were removed from their category into an unknown class. Three methods of unknown isolate removal were attempted. In the first method isolates classified under a certain threshold were labeled unknown (confidence removal). A second method labeled unknown only those isolates classified into the wrong category (misclassification removal). A third method labeled any misclassified isolates that were above a threshold of probability for a separate class unknown (outlier removal). The fourth method classified as unknown any misclassified isolate that was more than a threshold of probability greater than the correct probability (difference removal).

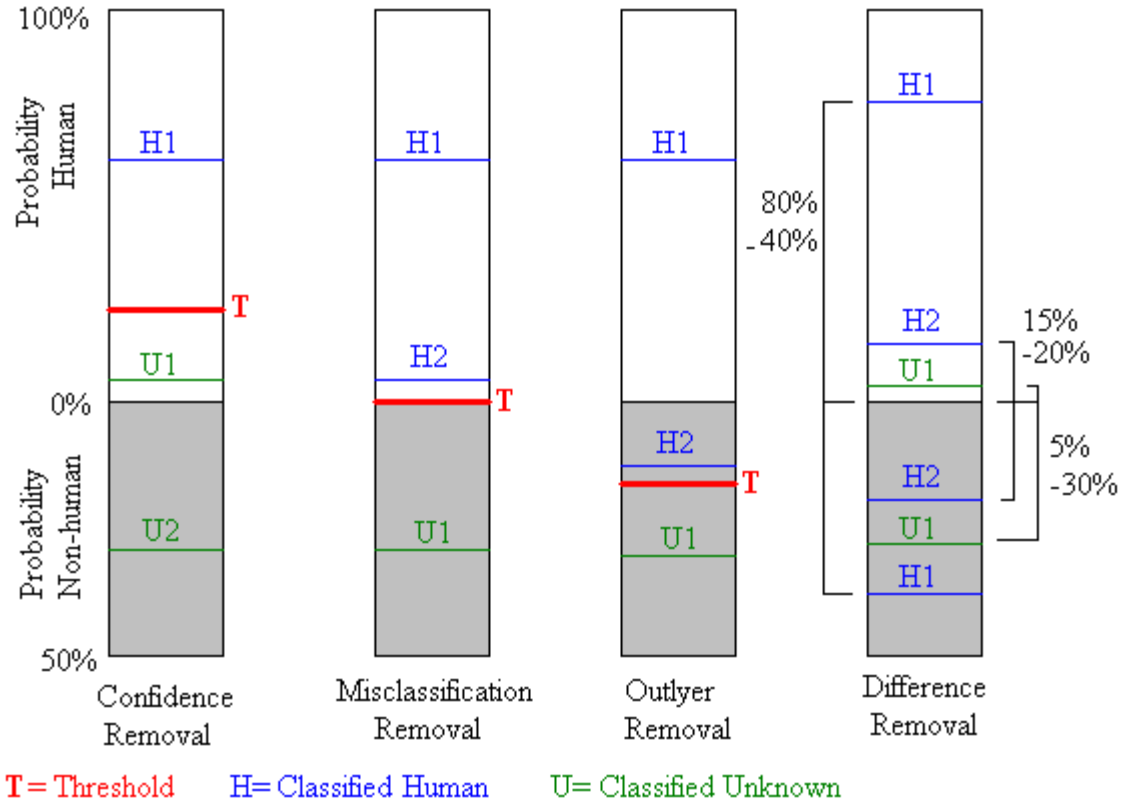
Figure 6 illustrates the different methods of removal to an unknown class. The figure depicts the classification probability of each isolate as a bar ranging from 0% upward into a human classification and downward into a nonhuman classification. The threshold of each removal method is depicted by a line with a "T" next to it. Those isolates which classify below the threshold of each removal method are relabeled unknown. In difference removal the connected lines describe the multiple probabilities given by LDA for a particular isolate. The unknown isolate is the isolate, which has a greater probability of being non-human by an amount above the threshold (in this case an arbitrary 20%), as suggested by Ritter et al., 2003.

Figure 5. Unknown Class Creation in Linear Discriminant Analysis



Unlabeled isolates classify according to their color

Figure 6. Classification of Hypothetical Known Human Isolates



Difference removal shown is based on a 20% threshold.

Table 16. Library and Challenge Set Classification as Affected by Method of Unknown Class Creation in a High Dataset [a]

	Confidence Removal [b] (30%)		Misclassification Removal [c]		Outlier Removal [d] (30%)		Difference Removal [e] (30%)		No Unknown Class	
Set [f]	Self	CSC	Self	CSC	Self	CSC	Self	CSC	Self	CSC
RCC [g]	0.42	0.48	0.42	0.32	0.42	0.32	0.46	0.47	0.47	0.48
A-PPV [h]	0.45	0.37	0.51	0.34	0.51	0.34	0.48	0.36	0.46	0.38
A-NPV [i]	0.81	0.77	0.81	0.74	0.81	0.74	0.82	0.75	0.82	0.76
A-Specificity [j]	0.83	0.79	0.88	0.87	0.88	0.87	0.84	0.77	0.82	0.76
ARCC [k]	0.48	0.36	0.44	0.22	0.44	0.22	0.48	0.35	0.51	0.39
A-MDP [l]	0.35		0.25		0.24		0.29		0.33	

[a] Data is based on the High dataset with interspatial, 3 and 4 conflicts removed

[b] Reclassification of isolates unknown when the greatest probability of the isolate being in any class is below 30%

[c] Reclassification of isolates unknown when discriminant analysis determines that the isolate is of a different source category

[d] Reclassification of isolates unknown when discriminant analysis determines that the isolate is of a different source category at a greatest probability of at least 30%

[e] Reclassification of isolates unknown when discriminant analysis determines that the isolate has a probability of being in a different source category than the known category by 30% or more.

[f] Indicates whether statistics were created from classification of the library(Self) or classification of the challenge set (CSC)

[g] Rate of Correct Classification. Values are the fraction of correctly classified isolates over all isolates

[h] Average Positive Predictive Value. Values are the fractional average of the correctly classified isolates of each source classes over the all isolates classified into the same source class

[i] Average Negative Predictive Value. Values are the fractional average of the isolates correctly classified out of each source classes over all isolates incorrectly classified out of the same source class

[j] Average Specificity. Values are the fractional average of the isolates correctly classified out of each source classes over all isolates not of the same source class

[k] Average Rate of Correct Classification. Values are the fractional average of the correctly classified isolates of each source classes over the isolates of that source class

[l] Average Minimum Detectable Percentage. Values are the average of the minimum fraction for a source to be considered not noise

No method of unknown removal increased the RCC above that of LDA without unknown removal (Table 16). Confidence Removal and Difference Removal made no more than a few percentage point changes to the data. Misclassification and Outlier Removal each increased the self-classification APPV and A-Spec and the CSC A-Spec. The better rates of self-classification reduced the A-MDP by ten and eleven percent respectively (Table 16).

Each method of unknown removal showed particular benefits (Disparate data not shown). Confidence Removal tended to increase the eccentricities of the library: categories that classified well increased in sensitivity, while those with low sensitivity became even lower. Sensitivity is the fraction of correctly classified isolates of a source class over the isolates of that source class. Even stronger sensitivity increases were seen with the removal of an entire category, such as livestock. Misclassification Removal only classified isolates that were of high confidence in the class. Outlier Removal increased statistical measures on limited occasions, particularly when very few isolates were reclassified unknown. Difference removal rarely changed any of the library metrics by more than 1%.

iii) Clustering was performed on each of the libraries in order to determine any nodal groupings within each class. Nodal groupings are groups of isolates within a class that are closer by Mahanalobis distance to a significantly different average than the class average. Significant groups were found that were distinct from other isolates within the class. Each class was clustered into at least two groups using a Complete or Average similarity method. These methods determine group similarity by total distance from the entire group (Complete) or average distance from members of the group (Average).

Single similarity, suggested during Stewart et al. 2005, detected few three to five isolate groups but did not otherwise detect potential subclasses.

The challenge set results of the clustered libraries were generally equal to or worse than un-clustered libraries (Table 17). For each library the Complete similarity clustering produced poorer, although not more than a few percentage points, results than the Average clustering. The largest difference in CSC due to clustering was in the High data set Complete clustering. The PPV was reduced nine percent and the RCC was reduced ten percent. This corresponded with a twelve percent drop in the sensitivity (Table 17).

In most instances, clusters within the wildlife subtype tended to increase classification rates (Table 18). While complete clustering in the High dataset reduced wildlife sensitivity by few percentage points, all other measures of wildlife classification were raised by a few points. Complete clustering conversely increased the wildlife specificity and, with sensitivity values, indicates an overall lower rate of wildlife classification

Clustering of the High dataset by Average clustering increased library measures a few percentage points, while Complete clustering reduced them a few percentage points (Table 18). The CSC values were all reduced through clustering. The challenge set did not reflect the same wildlife ratio as was in the library (Tables 10 & 21).

Table 17. Challenge Set Classification (CSC) in Three Datasets as Affected by Clustering Method

Dataset	High without interspatial, 3 or 4 class conflicts [a]			Binary with all conflicts			Combination without interspatial, 2, 3 or 4 class conflicts		
Cluster [b]	None	Complete	Average	None	Complete	Average	None	Complete	Average
RCC [c]	0.48	0.38	0.43	0.49	0.41	0.45	0.48	0.46	0.44
APPV [d]	0.38	0.29	0.36	0.4	0.43	0.42	0.32	0.26	0.32
ANPV [e]	0.76	0.70	0.73	0.76	0.74	0.77	0.75	0.69	0.73
A-Specificity [f]	0.76	0.71	0.73	0.77	0.75	0.77	0.75	0.71	0.73
ARCC [g]	0.39	0.27	0.27	0.4	0.37	0.39	0.33	0.25	0.26

[a] Interclass conflicts are isolates from two or more categories and have the same ARA pattern. Interspatial conflicts are isolates are from two or more categories, have the same ARA pattern, and are in different categories within OQB than without. Removal leaves isolates from within the OQB

[b] Similarity detection method used for clustering. None indicates no clustering was done. Complete uses the complete similarity algorithm. Average uses the average similarity algorithm.

[c] Rate of Correct Classification. Values are the fraction of correctly classified isolates over all isolates

[d] Average Positive Predictive Value. Values are the fractional average of the correctly classified isolates of each source classes over the all isolates classified into the same source class

[e] Average Negative Predictive Value. Values are the fractional average of the isolates correctly classified out of each source classes over all isolates incorrectly classified out of the same source class

[f] Average Specificity. Values are the fractional average of the isolates correctly classified out of each source classes over all isolates not of the same source class

[g] Average Rate of Correct Classification. Values are the fractional average of the correctly classified isolates of each source classes over the isolates of that source class

Table 18. Classification within the Complete High Dataset as Affected by Clustering Method

	Wildlife [a]			Library Average [b]			Challenge Set Average [c]		
Cluster [d]	None	Complete	Average	None	Complete	Average	None	Complete	Average
PPV [e]	0.63	0.69	0.70	0.44	0.43	0.46	0.36	0.34	0.32
NPV [f]	0.59	0.60	0.63	0.81	0.82	0.83	0.75	0.71	0.71
Specificity [g]	0.80	0.86	0.83	0.81	0.82	0.83	0.75	0.73	0.72
Sensitivity [h]	0.38	0.35	0.45	0.49	0.45	0.49	0.38	0.19	0.22
MDP [i]	0.38	0.25	0.24	0.35	0.44	0.31			

[a] Values below are given for only the wildlife source category. This category was the most influenced by clustering.

[b] Values below are the averages across the four source categories for the classification of the library

[c] Values below are the averages across the four source categories for the classification of the challenge set

[d] Similarity detection method used for clustering. None indicates no clustering was done. Complete uses the complete similarity algorithm. Average uses the average similarity algorithm.

[e] Positive Predictive Value. Values are the fraction of the correctly classified isolates of each source classes over the all isolates classified into the same source class

[f] Negative Predictive Value. Values are the fraction of the isolates correctly classified out of each source classes over all isolates incorrectly classified out of the same source class

[g] Specificity. Values are the fraction of the isolates correctly classified out of each source classes over all isolates not of the same source class

[h] Sensitivity. Values are the fraction of the correctly classified isolates of each source classes over the isolates of that source class. Average sensitivity is ARCC

[i] Detectable Percentage. Values are the minimum fraction for a source to be considered not noise

iv) Library self-classification results were thresholded as in Confidence Removal at a variety of values (Table 19). Thresholding relabels isolates unclassified if the isolate classified at a probability below a certain percentage (Figure 7). The library is not then reprocessed, as in Confidence Removal, but instead the threshold results are considered the final classification. Thresholding was performed at 90, 80, 60, 40, 30 and 20 percent however only 80, 60 and 40 percent were reported in Table 19. The 90 percent results rarely classified any isolates, and the 30 and 20 percent results rarely removed any isolates.

Isolate removal by thresholding increased PPV and specificity of library self-classification (Table 19). At the same time RCC, ANPV and ARCC decreased. A-MDP had no trend at low threshold percentages but greatly reduced at 60% and 80%. Thresholding removed poorly classifying isolates and did not distinguish between correctly and incorrectly classifying isolates at the thresholds used.

Table 19. Library Classification of Three Datasets as Affected by Threshold Value

Dataset	Combination without interspatial or 2, 3 and 4 interclass conflicts [a]				Binary Full Dataset				High without interspatial or 3 or 4 interclass conflicts			
Threshold [b]	80%	60%	40%	None	80%	60%	40%	None	80%	60%	40%	None
RCC [c]	0.14	0.34	0.57	0.59	0.02	0.13	0.38	0.51	0.02	0.13	0.38	0.47
APPV [d]	0.85	0.64	0.58	0.57	0.81	0.59	0.49	0.47	0.81	0.59	0.49	0.46
ANPV [e]	0.78	0.81	0.85	0.85	0.75	0.77	0.81	0.83	0.75	0.77	0.81	0.82
A-Specificity [f]	1.00	0.96	0.88	0.87	1.00	0.98	0.88	0.83	1.00	0.98	0.88	0.82
ARCC [g]	0.17	0.37	0.62	0.63	0.02	0.15	0.43	0.53	0.02	0.15	0.43	0.51
A-MDP [h]	0.03	0.12	0.24	0.26	0.01	0.06	0.31	0.29	0.01	0.06	0.31	0.33

[a] Interclass conflicts are isolates from two or more categories and have the same ARA pattern. Interspatial conflicts are isolates are from two or more categories, have the same ARA pattern, and are in different categories within OQB than without. Removal leaves isolates from within the OQB

[b] The minimum probability percentage of an isolate before it is considered classified. None indicates no threshold used.

[c] Rate of Correct Classification. Values are the fraction of correctly classified isolates over all isolates

[d] Average Positive Predictive Value. Values are the fractional average of the correctly classified isolates of each source classes over the all isolates classified into the same source class

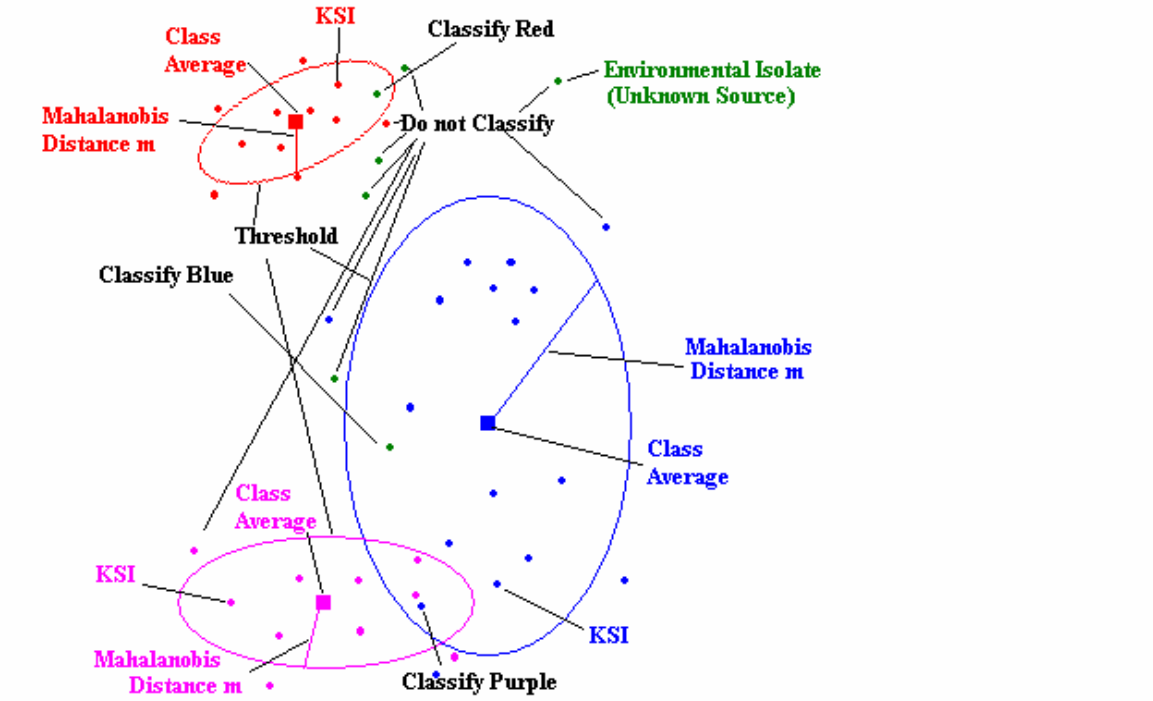
[e] Average Negative Predictive Value. Values are the fractional average of the isolates correctly classified out of each source classes over all isolates incorrectly classified out of the same source class

[f] Average Specificity. Values are the fractional average of the isolates correctly classified out of each source classes over all isolates not of the same source class

[g] Average Rate of Correct Classification. Values are the fractional average of the correctly classified isolates of each source classes over the isolates of that source class

[h] Average Minimum Detectable Percentage. Values are the average of the minimum fraction for a source to be considered not noise

Figure 7. Thresholding in Linear Discriminant Analysis



Isolates outside of threshold circles are considered unclassified. Unlabeled isolates within threshold circles are classified by color.

D. Evaluation

The average rate of correct classification and other statistical metrics for each library are shown in Table 12. The classification by artificial clustering due to isolate removal and data interpretation is shown in Table 20. The artificial clustering represents potentials from randomization and is not attributable only to the analysis. Artificial clustering is a measure of the comparison between variability of the data used to create the library and the artifacts of analysis. Artificial clustering ranged from 2% to 7% above the desired 25%. The desired value is 25% because random classification of isolates into four classes should approach 25%.

The Binary dataset showed the highest rate of artificial clustering. The Last dataset showed the lowest rate of artificial clustering. The combination dataset had more artificial clustering than the High dataset. In each case the inclusion of conflicts reduced the rate of artificial clustering. The reduction in artificial clustering for the Last dataset with conflicts was two thousandth's of a percent.

Table 20. Artificial Clustering Above 25% of Selected Datasets

Artificial Clustering [a]	Full Dataset	Interspatial 2, 3 and 4 class conflicts removed [b]
Binary [c]	6.83 [d]	9.58
Combination [e]	3.21	6.18
High [f]	2.91	4.77
Last [g]	2.77	2.77

[a] Discrimination between classes attributable to artifacts of analysis
[b] Interclass conflicts are isolates from two or more categories and have the same ARA pattern. Interspatial conflicts are isolates are from two or more categories, have the same ARA pattern, and are in different categories within OQB than without. Removal leaves isolates from within the OQB
[c] Dataset which uses growth data from all antibiotic concentrations as binary values
[d] Values given are percentages above 25% that indicate artifacts of analysis
[e] Dataset which uses the highest antibiotic concentration where growth occurred and the antibiotic concentration where growth occurred below the lowest concentration where no growth occurred
[f] Dataset which uses only the highest antibiotic concentration where growth occurred
[g] Dataset which uses the antibiotic concentration where growth occurred below the lowest concentration where no growth occurred

Table 21. Challenge Set Composition

	Human	Livestock	Pet	Wildlife	Total [a]
Total Unique ARA Patterns	8	17	37	48	97
Single Class ARA Patterns [b]	8	10	30	40	88
Two Class ARA Patterns [c]	0	3	3	4	5
Three Class ARA Patterns [d]	0	4	4	4	4
Maximum Rate of Correct Classification [e]	1.00	0.63	0.79	0.81	0.89

[a] Total is not the sum of isolates in each class because some isolates are in multiple classes
[b] ARA Patterns found only in one source category
[c] ARA Patterns found only in two source categories
[d] ARA Patterns found only in three source categories
[e] Correct classification of multiple class isolates implies incorrect classification in the remaining classes. The incorrect classifications are factored into the maximum rates for each category.

Half the unique PWC isolates went into the challenge set while the other half went into the library. This allowed for no exact matching of isolates in the challenge set by those in the library. Exact matching is a potential alternative method of classification to that of LDA. Table 21 described the contents of the challenge set. The 9 of 97 challenge set isolates were multiclass isolates. The correct classification of a multiple class isolate in one category also incorrectly classified the isolate out of the remaining categories. Therefore the maximum rate of correct classification was not 100% but 89%. While each category could individually classify 100% of the isolates correctly, the resulting incorrect isolates of the remaining categories of the multiclass isolates were factored into the maximum rate of challenge set classification for each category.

Table 22 shows the results of each library on the challenge set and its relation to the sensitivity. The highest rate of correct classification (RCC) was the Binary dataset with conflicts removed at 56%. The sensitivity of each dataset and the average sensitivity or ARCC do not correlate to the CSC rate of correct classification. The Combination dataset has higher self-classification than the High dataset but has a CSC RCC that is seven percent lower. Higher sensitivity and better rates of classification were found in the datasets lacking conflicts.

Table 22. ARCC, Class Sensitivity [a] and RCC CSC of Selected Libraries

Conflicts	Dataset	ARCC [b]	Human	Livestock	Pet	Wildlife	CSC RCC [c]
Interspatial, 2, 3 and 4 conflicts removed [g]	Binary [d]	0.67 [e]	0.63	0.69	0.65	0.72	0.56
	Combination [f]	0.63	0.56	0.68	0.71	0.56	0.48
	High [h]	0.58	0.55	0.69	0.54	0.52	0.55
Full Dataset	Binary	0.53	0.47	0.55	0.59	0.51	0.49
	Combination	0.51	0.38	0.54	0.71	0.42	0.42
	High	0.49	0.40	0.56	0.64	0.38	0.44

[a] Sensitivity. Values are the fraction of the correctly classified isolates of each source classes over the isolates of that source class

[b] Average Rate of Correct Classification. Values are the fractional average of the correctly classified isolates of each source classes over the isolates of that source class

[c] Rate of Correct Classification of the Challenge Set. Values are the fraction of correctly classified isolates over all isolates

[d] Dataset which uses growth data from all antibiotic concentrations as binary values

[e] ARCC values are the average of sensitivities of the four source categories

[f] Dataset which uses the highest antibiotic concentration where growth occurred and the antibiotic concentration where growth occurred below the lowest concentration where no growth occurred

[g] Interclass conflicts are isolates from two or more categories and have the same ARA pattern. Interspatial conflicts are isolates are from two or more categories, have the same ARA pattern, and are in different categories within OQB than without. Removal leaves isolates from within the OQB

[h] Dataset which uses only the highest antibiotic concentration where growth occurred

Table 23. Rates of Correct Challenge Set Classification (CSC)

Class Conflicts	Design	Full Dataset	Without Interspatial Conflicts [a]
Full Dataset	Binary [b]	0.45 [c]	0.49
	Combination [d]	0.42	0.40
	High [e]	0.44	0.46
3 and 4 class conflicts removed [f]	Binary	0.47	0.49
	Combination	0.42	0.45
	High	0.47	0.48
2, 3 and 4 class conflicts removed	Binary	0.57	0.56
	Combination	0.47	0.48
	High	0.50	0.55

[a] Interspatial conflicts are isolates are from two or more categories, have the same ARA pattern, and are in different categories within OQB than without. Removal leaves isolates from within the OQB

[b] Dataset which uses growth data from all antibiotic concentrations as binary values

[c] Rate of Correct Classification. Values are the fraction of correctly classified isolates over all isolates

[d] Dataset which uses the highest antibiotic concentration where growth occurred and the antibiotic concentration where growth occurred below the lowest concentration where no growth occurred

[e] Dataset which uses only the highest antibiotic concentration where growth occurred

[f] Interclass conflicts are isolates from two or more categories and have the same ARA pattern.

Table 23 shows the rate of correct classification of each library on the challenge set. The highest rate was the Binary dataset with 2, 3, and 4 class conflicts removed at 57%. Removal of interspatial conflicts increased the CSC of 7 out of the 9 libraries. The Combination dataset classified the challenge set the poorest out of every design. The rate of challenge set correct classification increased with the removal of interclass conflicts for each of the 6 libraries. The RCC of CSC had a lowest value of 0.40. The median value was 0.47 (Table 23). The variations in the RCC of CSC corresponded to a difference of 8 classified isolates.

At the sampling site on Buckhall Branch, slightly upstream was a farm (detailed in Materials and Methods). One metric for each library was their detection of a livestock signature at this site. None of the libraries showed a dominant livestock signature. Dogs were found off leash in the farm area. All of the libraries continued to detect a strong wildlife signature at this location.

All libraries were used to classify and estimate the number of correct classifications of Livestock at Buckhall Branch (Table 24). The number of Enterococcus, which classified livestock for the entire sampling period were multiplied by the livestock PPV in order to estimate the actual number of livestock isolates from the site. The highest value was given by the Combination library with 3 and 4 class conflicts removed. The Binary and High datasets detected less livestock than the Combination dataset. Removal of conflicts did not produce a trend in the data.

The sample site at Flat Branch was above a suspected leaking sewer line (detailed in Materials and Methods). A human signature was found at this site by all libraries

(Table 24). Pets were again detected as the secondary signature. The same method of livestock isolate estimation was done in Table 24 for human isolates at Flat Branch.

Table 24. Counts, PPV, and Estimated Correct Counts of Humans at Flat Branch and Livestock at Buckhall Branch

	Site	Flat Branch			Buckhall Branch		
Dataset	Method	Human Counts [a]	Human PPV [b]	Estimated Human [c]	Livestock Counts	Livestock PPV	Estimated Livestock
Full	Binary [d]	38	0.57	22	49	0.41	20
	Combination [e]	25	0.62	16	59	0.36	21
	High [f]	27	0.59	16	51	0.36	18
3 and 4 class conflicts removed [g]	Binary	39	0.60	23	28	0.48	13
	Combination	16	0.63	10	81	0.38	30
	High	27	0.60	16	51	0.37	19
Spatial, 2, 3 and 4 class conflicts removed [h]	Binary	19	0.70	13	13	0.68	9
	Combination	9	0.74	7	40	0.48	19
	High	13	0.33	4	43	0.24	10

[a] Counts are the sum of classified environmental data at the sample site for all 13 months

[b] Positive Predictive Value. Values are the fraction of the correctly classified isolates of each source classes over the all isolates classified into the same source class

[c] Estimations are made by multiplying counts by PPV and rounding to the nearest whole number

[d] Dataset which uses growth data from all antibiotic concentrations as binary values

[e] Dataset which uses the highest antibiotic concentration where growth occurred and the antibiotic concentration where growth occurred below the lowest concentration where no growth occurred

[f] Dataset which uses only the highest antibiotic concentration where growth occurred

[g] Interclass conflicts are isolates from two or more categories and have the same ARA pattern

[h] Interspatial conflicts are isolates are from two or more categories, have the same ARA pattern, and are in different categories within OQB than without. Removal leaves isolates from within the OQB

The Binary dataset consistently estimated more human isolates than the either the Combination or High datasets. The Binary dataset with 3 and 4 class conflicts removed estimated 24 human isolates, while the entire Binary dataset estimated 23. Removal of spatial conflicts along with all interclass conflicts resulted in estimates 1/3 or less of full dataset estimates.

At both of these sites neither livestock nor human was the dominant signature (Table 25). Buckhall Branch has a dominant signature of pets or wildlife. Flat Branch has a dominant signature of wildlife.

III. Environmental Isolates

The environmental isolates from the monthly *Enterococcus* sampling showed a majority wildlife signature at 5 sites and a majority pet signature at 5 sites according to MDP adjustment (Table 25). Secondary signatures for MDP were of pets or wildlife were found at all sites. After adjustment according to library PPV, wildlife was the dominant source at every location.

Of the 3488 environmental *Enterococcus* colonies collected, 957 unique ARA patterns were found. The median isolate repetition was 1 isolate. Fifty ARA patterns accounted for 50% of the total environmental isolates. Of these 50 isolates, all were resistant to chlortetracycline and none were resistant to oxytetracycline or tetracycline. Twenty-five of the 50 most numerous isolates were pet isolates and they accounted for 20% of all environmental isolates.

The relative fraction of each classification after adjusting for the remaining number of classified isolates is shown in Table 25. The MDP adjustment considers any data above the MDP for the class significant. The PPV adjustment applies the fraction

correct during library self-classification to environmental isolates by multiplying counts by the PPV.

Enterococcus data for the ten sites showed majority pet or wildlife signature (Table 26). At no site was a human signature present above the MDP of 24% or the second most numerous class by PPV adjustment. At three sites, based on MDP, the livestock signature was present. South Run isolates classified as wildlife were below the MDP. On a monthly basis livestock was the dominant number of isolates for 8 of the months and wildlife the remaining 5 months. Pets and humans were both below their respective MDPs for study wide monthly results and never were above 15% by PPV (Appendix B). All samples combined showed a majority livestock signature and a secondary wildlife signature, however the wildlife signature was slightly more dominant after PPV adjustment.

The MDP findings were partially rejected by quarterly environmental sampling (Appendix C). The library used for *E. coli* classification was not subject to any of the methods of *Enterococcus* analysis but was instead confirmed through pulse field gel electrophoresis (Appendix E).

Table 25. Major and Minor Signatures at Each Sample Site Using the Combination Library with all Conflicts Removed [a]

	MDP Adjusted Sources [b]				PPV Adjusted Sources [c]			
Site	Major [d]	Fraction	Minor [e]	Fraction	Major	Fraction	Minor	Fraction
Upper Bull Run	Wildlife	0.47 [f]	Pet	0.35	Wildlife	0.37	Pets	0.09
Lower Bull Run	Pet	0.47	Wildlife	0.39	Wildlife	0.31	Pets	0.12
Youngs Branch	Pet	0.38	Wildlife [f] Livestock	0.37 0.23	Wildlife	0.30	Livestock	0.11
Catharpin	Wildlife	0.37	Pet Livestock	0.33 0.27	Wildlife	0.29	Livestock	0.13
Buckhall Branch	Pet	0.47	Wildlife	0.29	Wildlife	0.23	Pets	0.12
Flat Branch	Wildlife	0.39	Pet	0.39	Wildlife	0.31	Pets	0.10
South Run	Pet	0.48	None	-	Wildlife	0.23	Pets	0.13
Broad Run	Wildlife	0.46	Pet	0.31	Wildlife	0.37	Livestock	0.09
Lower Kettle Run	Wildlife	0.40	Pet	0.37	Wildlife	0.32	Livestock Pets	0.10 0.10
Upper Kettle Run	Pet	0.41	Wildlife Livestock	0.32 0.23	Wildlife	0.25	Livestock	0.11

[a] Dataset which uses the highest antibiotic concentration where growth occurred and the antibiotic concentration where growth occurred below the lowest concentration where no growth occurred. Interclass conflicts are isolates from two or more categories and have the same ARA pattern. Interspatial conflicts are isolates are from two or more categories, have the same ARA pattern, and are in different categories within OQB than without. Removal leaves isolates from within the OQB

[b] Calculated sources at a sample site than the minimum detectable percentage from the compiled data of 13 months were removed

[c] The PPV was multiplied by the fraction of sources calculated for each site from the compiled data of 13 months.

[d] The calculated source with the largest percentage of isolates over the course of the 13 months

[e] The calculated remaining source(s) by MDP and the second most numerous source(s) by PPV

[f] Fraction of environmental isolates classifying as that source

Table 26. Adjusted Relative Fraction of Classified Isolates

Site	MDP Fraction [a]				PPV Fraction [b]			
	Human	Livestock	Pet	Wildlife	Human	Livestock	Pet	Wildlife
Upper Bull Run	0.00 [c]	0.00	0.43	0.57	0.04	0.13	0.16	0.67
Lower Bull Run	0.00	0.00	0.54	0.46	0.04	0.11	0.24	0.61
Youngs Branch	0.00	0.24	0.38	0.38	0.02	0.22	0.19	0.57
Catharpin	0.00	0.28	0.34	0.38	0.04	0.24	0.16	0.55
Buckhall Branch	0.00	0.00	1.00	0.00	0.16	0.13	0.25	0.47
Flat Branch	0.00	0.00	0.50	0.50	0.04	0.17	0.19	0.60
South Run	0.00	0.00	1.00	0.00	0.12	0.15	0.26	0.47
Broad Run	0.00	0.00	0.40	0.60	0.05	0.16	0.14	0.65
Lower Kettle Run	0.00	0.00	0.48	0.52	0.04	0.18	0.18	0.60
Upper Kettle Run	0.00	0.24	0.42	0.33	0.06	0.22	0.21	0.51

[a] Fractions below are the amount of environmental samples classified into the category over all those classified after categories below the MDP were reduced to zero

[b] Fractions below are are the amount of environmental samples classified into the category multiplied by the category PPV over the sum of all categories multiplied the respective PPV's

[c] Data is based on 13 months of sampling for each site. Values of zero were below category MDP at the particular site.

Enterococcus but not *E. coli* showed general agreement with the fluorometry data (Appendix D). No hotspots of dominant human isolates or fluorescence were found. Levels of fluorescence remained below those indicating contamination in other studies (Hagedorn et al. 2003).

IV. References

Hagedorn, C., R. B. Reneau Jr., M. Saluta, and A. Hassall. 2003. Impact of Onsite Wastewater Systems on Water Quality in Coastal Regions. Final project report to the National Oceanic and Atmospheric Administration, Charleston, SC.

US EPA. 2005. Microbial Source Tracking Guide Document. U. S. Environmental Protection Agency, Office of Research and Development: Washington, D.C.

Chapter 5. Discussion

I. Monitoring Results (Tables 3-9, 24-26, Figure 3)

At each sample site the *E. coli* data conflicted with the *Enterococcus* data. Not only was there little or no correlation between the magnitudes of the bacterial concentration, the source data also conflicted. This suggests that the animals contributing *E. coli* and *Enterococcus* occur with different frequency. This is further supported by different concentrations of bacteria in fecal samples used for known source isolates. A similar lack of bacterial magnitude correlation was found in another study of the Occoquan Basin (OQB) by Hagedorn et al. 2004. The correlation between the two indicators may depend differentially on more factors, such as temperature, flow, time of day and local animal populations. Time of day may be of particular import due to both solar disinfection and animal defecation habits.

Monthly differences of *E. coli* and *Enterococcus* counts were greater than sample site differences. This indicates that the locations are affected by most of the same factors. *E. coli* data showed higher statistically higher counts for December and September, however August through December, dry months, did have higher counts. *Enterococcus* data showed September, December, and August as having statistically higher counts. June 04 through December, dry months, did have higher counts. Dry months do influence sites.

The *E. coli* data showed a majority wildlife signature and a minority human signature at all locations. For the month of March human was the dominant signature and wildlife the minority. This conflicted with the *Enterococcus* data, which showed majority pet or wildlife. The human presence indicates human *E. coli* waste materials entering the watershed in multiple locations.

E. coli contamination, the main standard of impairment for these sites, has a primary contributor of wildlife. Cleanup of the sample sites to fall under the regulation levels of contamination will be difficult, as there are few established control mechanisms for wildlife. While increasing vegetation in the riparian buffer reduces waterfowl, this can result in an increase in beaver, opossum, deer, songbirds and nuisance birds.

The *Enterococcus* data varied depending on the adjustment of the raw data according to minimum detectable percentage (MDP) or positive predictive value (PPV). The values given by the adjustment are heavily weighted by the pet PPV being about 1/3 the wildlife PPV. The PPV adjusted data confirmed the results of the *E. coli* data and therefore should be considered the more probable analysis. However, the MDP adjusted data is still useful for determining areas of potential remediation consideration.

The *Enterococcus* data indicates three things: wildlife is a significant source, pets contribute at all locations, and livestock is significant at particular sites. This suggests that countywide cleanup should focus on pet influence, while particular sites should focus on livestock influence. The pet influence at Buckhall Branch and South Run is the only class, or source category, above MDP and therefore must be reduced before other targets for site cleanup become apparent.

The only classification that is not a concern at any site is human fecal contamination. At the two highest levels detected, human fecal signature was the third and fourth strongest signature. This is inline with the fluorometry data that only showed slight variation between sites. The monthly fluorometry data show Young's Branch, Kettle Run, and Buckhall Branch as higher fluorescence sites. Of these sites *Enterococcus* counts confirmed Buckhall Branch as having a higher, although

insignificant human signature. The lack of strong differentiation between sites confirms the similarity between sites with the MST results. The normality of each month's fluorometry data as well as the minor site trends indicate only background levels of fluorescence and therefore no substantial human source of fecal matter.

The wildlife counts at each site could be explained by suburban and rural deer, field mice, rats, rabbits, raccoons, geese, ducks, songbirds, and nuisance birds, such as starlings, throughout the region. The lower signature at Buckhall Branch and South Run is likely due to pets or livestock diluting the presence of wildlife. Continued sampling of this area will likely show wildlife as the main contributor.

II. Library (Training Data)

A. Linear Discriminant Analysis (LDA) (Table 10)

LDA determines isolate class by comparing the probability that the isolate is within the confidence interval of each class. The closer the isolate is to the mean of the class, the higher the probability the isolate is of the class. Distance from the mean is traditionally determined through relation to the covarying standard deviation (Mahalanobis distance). The probability calculation for each class is made independently. Each isolate is classified into only its most probable class.

Thresholding of results can effectively classify isolates by suggesting they are either unknown or of multiple categories. Should an isolate not fall within the confidence interval of a class or subclass, it should be categorized unknown. The threshold establishes a particular probability that the isolate is of the class, given that the training data is representative of the population. Pairwise comparison should be made of each isolate's class probability to assure that an isolate is not of multiple classes. Isolates

which have a probability above a threshold should be considered potentially of either class. Such a threshold would establish the singularity of isolates and describe overlap between classes. Singularity or uniqueness would be established by indication that the isolate is only of the classes listed.

In order for the library to be representative of the population all KSI's should be included in the creation of an LDA classification algorithm. The KSI's which classify above the threshold used for environmental isolates should be used to calculate library statistics. Statistics should also be generated for overlapping classes so as to determine whether there is significant difference between classes within the library. A library that is representative by confidence interval may need more isolates to be useful for differentiation.

Binary data, similar to an unknown category, is disrupted in LDA classification. While a highly variable class has a shorter Mahalanobis distance from the average, a binary class has less meaningful distance. The covariance calculation of the Binary dataset is hampered by the lack of intermediate values to become a simple ratio of covariance. The established ratio of binary values biases the dataset if it is assumed to be a normal distribution. The JMP software calculation of the covariance and the distance may have assumed normality, however the similarity of results between the Binary dataset and the High and Combination datasets suggests that this is not a major source of bias.

Clustering has greater potential to work in larger data sets. While increases in wildlife classification without change in PPV indicated multiple subclasses, no other classes benefited from clustering. There were 542 wildlife KSI's, which is slightly more

than half the KSI's in the dataset. In every library process the subclasses within the wildlife dataset were apparent. Clustering for the other classes was hampered by a lack of apparent clusters. Clustering is most likely beneficial to datasets larger than 918 isolates when used to create subclasses with tighter confidence intervals than the original class.

B. Library Design (Tables 11-13)

The multiple methods of library design were attempted due to the recommendation that library data not be binary (Stewart et al. 2005). Of the four methods attempted, High, Last, Binary, and Combination, only the Last method was found to be poor and unusable. The differences between the three successful methods were due to challenge set composition and which laboratory errors each could account for. Laboratory error, in which no isolate is plated at a particular concentration, is ignored by the High dataset. Laboratory error, in which contamination occurs during plating a particular concentration, is ignored by the Last dataset. Neither Binary nor Combination datasets can ignore this kind of error.

The High library was probably more successful in challenge set classification because of growth conditions that match the High dataset. Thirteen of the 97 isolate challenge set had different growth data due to differences in interpretation for the Last and High datasets. The Binary library was similarly successful because it could include intermediate data in its classification and therefore put less weight on low concentration datapoints. The Combination library had both the advantages of the High library and disadvantages of the Last library but was chosen as the library for environmental classification because of its self-classification ability. The self-classification allowed for

higher PPVs. The Binary library would have been the natural choice for environmental classification except that it violates more of the normality and independence assumptions of linear discriminant analysis than the other libraries.

Most analyses assume data points to be independent and normally distributed (USEPA 2005). Ongoing research continually shows correlation because microbes have a higher incidence of resistance to antibiotic B if they are resistant to antibiotic A. Therefore a binary representation typically shows dependent data that cannot be normally distributed. The representation is independent if and only if a microbe can be resistant to an antibiotic at a high concentration while susceptible at a low concentration. This susceptibility follows from an assumption that resistance genes require a threshold of antibiotic before they activate. Multiple datapoints describing a single attribute, level of resistance to an antibiotic, does not harm the results because the overspecification of datapoints is being used to predict the results rather than regress the variables. Overspecification obscures the influence of each factor by causing multicollinearity. Multicollinearity is when two or more factors cannot be distinguished because the two are, or nearly are, linear combinations of one another. As we are seeking only to predict the result and not calculate precise values of each individual factor, the independence of variables is less of a problem.

The Combination library does have an advantage over all other datasets. While this dataset does not make the absence of isolate growth a distinct data point, the combined library includes more available data on the isolate than other methods. The Combination library has no dependence between its concentrations of antibiotic columns. The Combination dataset is similar to both the High and Last datasets in that it has

potential for a normal distribution of data. It is possible that the Combination library had higher PPV's because of library overclassification, or discrimination on trivial differences. Overclassification would be exhibited by a lower rate of correct classification (RCC) in the challenge set classification (CSC). This lower rate could also be due to unrepresentative data from joining the High and Last datasets.

C. Library Processing (Tables 14-19)

LDA ignores any minor differences in probability that suggest an isolate could be of multiple classes. LDA also forces categorization of every isolate into a class, despite low probability for the isolate to be any class. The creation of an unknown category and the creation of a threshold are both proposed solutions to this problem. Because the threshold does not obscure classification data it is preferable to the unknown category.

Four kinds of processes were attempted to increase correct LDA classification by libraries: i) conflicts in the data set were removed, ii) isolates that classified poorly were relabeled unknowns, iii) clustering was performed on classes to identify similar subclasses, and iv) thresholds were made for isolate classification consideration. Conflict removal was the most successful, although unknown and subclass creation had limited success.

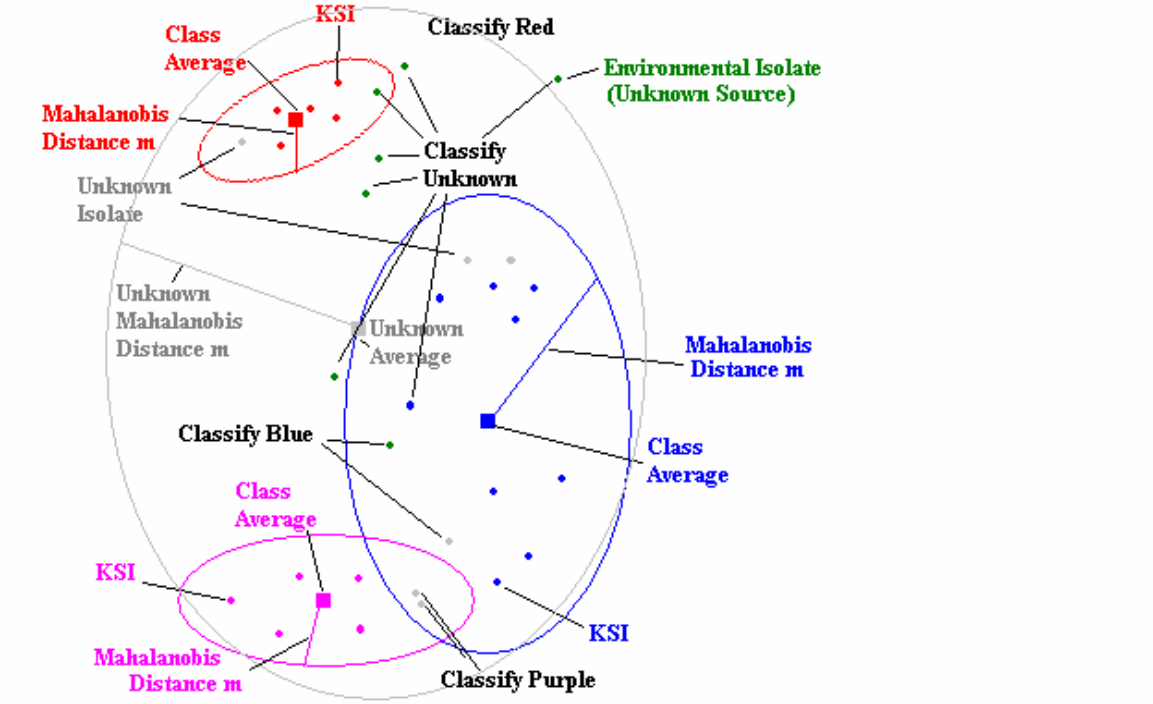
i) Conflict removal was suggested in (Ritter et al. 2003) and it significantly improves training data, libraries. The removal of conflicting data allows the discriminant analysis algorithm to train on data of a reduced number of classes. Removal of all conflicts creates a more specific classification algorithm. While this dataset size reduction does increase the artificial clustering of the data, the more specific algorithm

classifies with a higher probability. This is a significant way to reduce the amount of training data.

ii) The creation of an unknown category allows reasonable classification of isolates unlike those in any other category. This helps to prevent forcing an isolate that is of low probability in all categories into a class. When excluding isolates, the alternative to complete removal is removal into an unknown category. This reduces the influence of the isolate on the discriminant analysis of its original classification. By the isolate exclusion, the algorithm for classification into a particular class becomes more specific. Algorithm specificity is useful to properly exclude members of other classes.

Reclassification of a KSI to an unknown category, rather than isolate removal, allows the isolate to obscure further categorization. In both cases the KSI no longer affects the LDA calculation for its original class. However, the unknown class becomes a possible category for environmental isolates. A highly variable class, such as one with any poorly classifying or overlapping KSI's, has a wider confidence interval than a low variability class. This larger interval is reflected by a shorter Mahalanobis distance and therefore greater category probability (Figure 8). Environmental isolates which are near the mean value of an unknown class, despite a high probability of categorizing into a known class, have a greater probability of being unknown. Removal of KSI's prevents the additional category from obscuring the comparison between the known classes. KSI's not meaningful to classification are better removed than categorized unknown under LDA.

Figure 8. High Variable Unknown Class in Linear Discriminant Analysis



Unlabeled isolates classify according to color.

In the creation of the unknown category the drawbacks outweighed the advantages. For most of the unknown categories, the removal methods created an unknown category that was too large. This resulted in the unknown classification becoming the dominant class. Difference removal was the only unknown category creation algorithm that did not result in worse results. Difference removal should continue examination in a larger dataset as a way of removing overlapping data that does not directly conflict. Unknown categories might have better success in refining libraries with higher rates of correct classification, as they had less detrimental influence in the better classifying libraries.

iii) The advantage of performing a cluster analysis, relabeling similar isolates of a class into a subclass, before linear discriminant analysis (LDA) is because clustering removes subclasses that expand the confidence interval of a class. This improves the classification algorithm by making tightly clustered uniform groups.

Subclasses are groups within a library of similar isolates that together have a lower variability than that of the library. Multiple subclasses may be distinct within a class. The definition of a subclass is particularly useful for linear discriminant analysis in that it allows the training algorithm to find a more specific requirement for data to be of a particular class. However, only when the separation of a subclass reduces the confidence interval of the remainder of the library does the separation of a subclass aid correct classification.

While clustering tended to increase rates of classification when definite subclasses existed, the lack of numerical method may have resulted in over-clustering.

Unlike other methods of processing the clustering mechanism had no percentage similarity or similar such function. This made designation of groups entirely subjective. Clustering without a numerical indicator had poor results, however a numerically thresholded approach may attain the benefits of separating subclasses.

iv) Improvement of results data by thresholding is based on misclassified isolates receiving low probabilities of classification. The poor thresholding results imply that misclassified isolates have probabilities approaching those of correctly classified isolates. The LDA classification of the data set did not provide tight enough classification intervals, or confidence limits, to exclude other isolates. Even when these classification limits were further tightened by thresholding, the resulting loss in correct or all isolate classification outweighed the reduction of misclassified isolates. Were classifications made with tighter confidence limits, this technique might have more success in removing marginal isolates.

D. Library Evaluation (Tables 20-23, 26)

Library classification rates determine the success of environmental isolate classification. The rate at which the libraries corrrectly classify themselves indicates the rate at which they classify environmental isolates, provided there is no artificial clustering and issues with library representativeness.

Libraries require little or no artificial clustering to indicate their lack of bias. Artificial clustering suggests there are too few isolates in the library to determine whether the library is representative of all possible isolates. This suggestion is based on the conclusion that isolates randomly assigned to a group would have no discernable similarity other than the similarity between all groups. That conclusion is dependent on

each class, or source category, having normal variability in its isolates. The Binary and Combination datasets tend to increase artificial clustering because the increase in dependent variables gives more opportunities for similarity. Unequivalent variability would create biases in the randomly assigned groups when outliers or subclasses collect in a particular group. Artificial clustering is a good measure of a library when subclasses maintain the normal variation of datapoints within a class.

Bias is the tendency of a sample to deviate from the population norm. It exists when samples are pre-selected for a particular feature that is not equally distributed throughout the population. A severely biased dataset can appear to successfully classify itself, but have no relationship to the population. Bias exists at higher levels in sample sets that increase until the average value does not change. This is due to the greater likelihood of a similar sample set average close to the prior sample average than both a sample set average that brings the set closer to the mean and then a similar sample set average. For known source libraries the addition of isolates until they classify as well as isolates already within the library would show this kind of bias. It can be avoided by the addition of samples until confidence limits tighten. In the case of known source libraries the confidence interval of each class must be calculated and should decrease to a plateau with added samples. For a low bias library, addition of isolates should be stopped when the confidence interval reaches a particular threshold. No threshold was established for isolate addition for this study. Confidence intervals were not calculated in this study.

In this study known source isolates were not enumerated. Choosing only uniques for training data inclusion increases variability and biases the dataset toward the more infrequently occurring isolates. This false variability widens confidence limits and

thereby the range of samples identified as part of a class. Enumerating allows one to give proportional weight to the isolates given. However applying enumeration techniques requires data on animal population, feces size and production rate.

The statistical analysis of the source tracking data allows better comparison of this work with that of other classifications. Particularly significant are the positive predictive value (PPV) and the sensitivity of the analysis. Both of these measurements indicate the likelihood of correct classification of the library. The high levels for human and wildlife predict the usefulness of the library in classifying these two significant classes.

Correct classification of humans is of particular concern due to the possibility of wastewater contamination. Human wastewater contains both chemicals and pathogenic organisms that contribute to public morbidity. For waters designated swimmable, it is of upmost importance to maintain low levels of human based contamination. Untreated wastewater release into the watershed is also illegal. Reduction of human impact is the goal of environmentalism as well as clean water legislation.

Wildlife isolate classification is important due to the lack of effective cleanup mechanism. Once one knows the percentage of wildlife isolates, one obtains the number of isolates that have little or no potential for elimination. In the case of impaired waters, wildlife bacteria concentrations above regulation levels suggest permanent impairment. Such impaired waters can have no effective TMDL so long as there is no effective cleanup mechanism.

There are three mechanisms for reduction of wildlife fecal influence: reduce wildlife populations, reduce wildlife access, reduce wildlife habitat. None of these methods are practical in OQB. Wildlife population can be reduced through hunting or

poisoning. While these methods effect a temporary drop in the population, migration and reproduction often cause a rebound. Hunting is further unlikely in this region due to the proximity of wildlife areas to high-density housing. Reduction of wildlife access to water is only effective for waterfowl. Increased riparian vegetation reduces waterfowl access, but provides habitat for other wildlife. Reduction of habitat will naturally happen through the urbanification of the basin, however removal of woodland tracts is generally unacceptable to suburban and rural citizens.

Known Source Isolates (KSIs) from the Occoquan Basin were limited in this study. The low number of pet *Enterococcus* isolates probably gave this class wider confidence limits. This was exhibited as more isolates classifying as pets and a higher PPV. While results data was adjusted by the PPV, the low total number of samples reduces the likelihood that the library directly relates to the population. Consequently the actual PPV and number of pet isolates could vary significantly from what was reported.

It is unknown whether enough samples were included in the library to make it representative. While artificial clustering indicated a low level of statistical artifact, confidence interval thresholds were not employed. Jack-knife analysis was also not employed to determine whether more isolates were necessary. Either or both of these measures could increase confidence in the library.

The process of localizing the library and removing interspatial conflicts also removed some interclass conflicts. The removal of interspatial conflicts only kept interclass conflicts that were from the Occoquan basin. This may have resulted in some KSI's influencing the analysis toward a single class despite being occurring in multiple classes. The lack of KSI's from the Basin did not allow analysis to compare the

influence of out of basin isolates on the training data. It is possible that the out of basin KSI's bias the data because within basin conflicts were not detected for lack of isolates.

While *E. coli* data analysis applied the PPV and MDP, the lack of local *E. coli* isolates may have resulted in a significantly biased verification. None of the data processing techniques were applied to the *E. coli* library. It also consisted of repetitive data. This likely focused the library more closely to the region from which the repetitive data came.

The method of PFGE data processing applied is not the same as that of the published journal articles. While it has been used in studies (Hagedorn et al. 2003), the method used has yet to undergo rigorous statistical review. The method may not be the most efficient as the influence of band size is dependent on the number of bands and the influence of the number of bands is dependent on the band size. The order in which the band size and number of bands are combined makes the sets points significantly different. However, the number of bands as dependent on the band size significantly reduces the resolution of the band size. The lack of library PFGE isolates does not allow high resolution linear discriminant analysis (LDA) through the SAS-JMP (v. 5.0.1, SAS Inst., Cary, NC). This lack of resolution may bias the data of the environmental isolates to greater similarity to library isolates.

III. Environmental Isolates

A. PPV and MDP (Tables 12, 13, 25 & 26)

Positive predictive value (PPV) and minimum detectable percentage (MDP) were used to recompile the results according to confidence in the data. Both statistics are measures of the rate at which results are considered significant. These statistics allow

one to remove data from consideration as insignificant. Both methods require the assumption that the statistical measures of the classification of training data applies to the environmental data.

Positive predictive value is the ratio in a particular category of correctly classified KSI's to all both KSI's classified into that same category. The PPV is the rate at which a classification is correct for that category. The average of these values for all categories is not the same as the rate of correct classification (RCC) of the library because the PPV's are not weighted by the number of isolates classified into each category. This shows that the PPV is independent of the number of KSI's in each class.

The minimum detectable percentage is calculated using the ratio of false positives in one category from a second category to the total number of KSI's of the second category. Another way to describe the false positives from their original category is as the false negatives into the calculating category. The MDP statistic is equal to the average plus four times the standard deviation of the ratios of false positives from each category to the number of KSI's in the library of their category. Should any category misclassify more or less than other categories into the calculating category, the standard deviation increases the value of the MDP. If only one category does not misclassify into the calculating category, this tends to make the MDP value higher and therefore worse.

While both statistics depend on false positives classifications MDP is heavily influenced by the size of library classes, while the PPV depends only on the number of isolates classified. Because the MDP is affected by category size it has poorer statistical fidelity than PPV unless library classes are proportional to the classes of the

environmental population. As the environmental population is unknown, results from PPV should be considered more accurate.

PPV was used to adjust results rather than sensitivity because sensitivity is calculated using the known category of analyzed isolates. The PPV was more robust for this data analysis than sensitivity. Using sensitivity as the modifier would invalidate the results modification because the actual source of the environmental data is not known.

Negative predictive value (NPV) was not applied to results because it is chiefly designed for a two way test. NPV is the ratio of KSI's correctly identified as not of a particular category. These percentages do not depend on where the isolates were actually classified, so long as they aren't in the original category. While a matrix of NPV's can be used to adjust data, the higher complication of accurate application encouraged use of the reciprocal statistic of NPV, which is designated as PPV.

B. Caveats of Analysis (No Tables)

While linear discriminant analysis assumes environmental data and KSI's are normally distributed, they are in fact skewed. This skew is disregarded in the KSI analysis due to the removal of conflicting isolates and repeats. Differential survival of isolates is the source of the asymmetry. Metabolic requirements select for highly repetitive low resistance organisms. Antibiotic exposure selects for high resistance organisms. As antibiotic exposure is not constant, during the times in which antibiotics are not in use low resistance organisms will flourish. These low resistance patterns have a metabolic advantage and skew the environmental isolates towards the classification of these multi-category repetitive isolates. Given environmental and gut isolate survival and growth rates, this deviation could possibly be accounted for.

Fecal samples are assumed representative of the area. Tests were not performed to determine if enough samples from the OQB were included for representativeness of the OQB. Confidence limits, as described above, would indicate when additional samples from the Basin do not improve library representativeness. Thresholds of these limits would determine significance of the data analysis.

Fecal indicator survival time, ratio of excretion, recovery from water and attribute expression were assumed to be equal or normally distributed by lack of isolate significance weighting in the analysis. These assumptions are due to lack of data and continued research may change their influence. These assumptions are acceptable for classification as a best guess, but not as an exact probability. If these factors are defined, they may make use of enumeration techniques based on population estimate. In this way population estimates can give an expected range environmental isolate ratios. Should results fall outside of estimated boundaries, other factors would have to be considered or population estimate would have to be revised.

The current analysis assumes that each isolate is of a particular class. Many isolates and groups of isolates occur in multiple classes. Through enumeration data ratios of isolates can be determined to either give each multiclass isolate a most probable class or to reflect the distribution in identified isolates. A third option is to classify isolates as though they were from multiple classes. Interclass conflicts of KSI's would continue to be used for classification. Thresholding the results to determine if there is a distinct difference in class probabilities would allow LDA methods to describe multiple class isolates. Isolate class would change from one categorical variable to four separate binary values in order to best support multiclass KSI designation.

Multiple methods were used to test the validity of results. *E. coli* and *Enterococcus* data conflicted in results and therefore couldn't confirm either analysis. The only agreement between the techniques was that wildlife was the primary source of pollution. While humans and pets tend to share gut bacteria, the libraries pointed to different classes as the secondary source. In both cases the PPV of the secondary class was below that of wildlife. Therefore less confidence can be given to results of that class. The wildlife class, with its probable subclasses, may have been the only representative class. The close proximity of wildlife and human areas may have increased the availability of rubbish to wildlife, and thereby provided wildlife with unrepresentative fecal bacteria.

The fluorescence data was standardized across each month due to fluctuating methods of calibration. The changes in detection method and standard concentrations reduced the significance of between month data comparison. Consequently, when all the sample sites gave higher counts for a particular month, fluorometry data could not confirm this. Until there is a standardized protocol , the ongoing development of fluorometric technique reduces its confirmation potential.

IV. Real Value of Analysis

The main utility of this study is that it provides locations and targets for reduction *E. coli* and *Enterococcus*. As these are indicator organisms, this in turn should lead to general reduction of waterborne disease. The confirmation of the *Enterococcus* result by the *E. coli* data provides a target of wildlife as the primary contributor to bacterial pollution. This target will allow necessary reduction measures to be taken to reduce wildlife fecal pollution. Directed action by Prince William County (PWC) will allow

consideration to be given by the state to enforcement of water quality standards in wildlife impacted areas.

The impact of the research on the TMDL and TMDL reduction plan are determined both by levels of *E. coli* and *Enterococcus* reported and by the sources of fecal pollution. As wildlife was the largest identified source, the most resources and planning will be given to economically feasible methods of wildlife fecal reduction. Secondary targets of pets and livestock will provide PWC with minor impetus to reduce fecal influence in the affected areas.

Livestock farm targets near the sampling regions will be required to implement best management practices (BMPs) for runoff reduction. Possible methods include fencing of creeks and creation of separate watering holes. Horse trails in the region could either be diverted away from waterways, or additional storm water management ponds could be added to the area. In areas where such changes are not possible, diapering may be appropriate for horses.

Pet sources were found in a wider part of the region. Although these are not the primary sources of fecal pollution, the research provides basis for increased/enforced bagging of dog feces. As housing increases in the Occoquan Basin, the reduction of potential pet fecal influence will become more important for meeting bacterial water quality standards.

V. Conclusions

The goal of this project was to monitor and evaluate the identification of eight streams as *E. coli* impaired. The goal was successfully completed. The monitored streams were correctly identified by state authorities as impaired waters.

One objective of this project was to determine the source of bacterial impairment. This objective was completed but more work is necessary due to conflicting source tracking results. Wildlife and pets were indicated as the source of impairment by ARA. The bacterial impairment of the eight streams could be significantly reduced if the identification of pollution is correct. More local source samples need to be taken to verify this result.

A second objective was to determine the best design for an Antibiotic Resistance Analysis library. This objective was completed in part, because the extent which older isolates and those from outside the region can be used was not determined due to lack of unique KSI's. The best ARA library design used only unique isolates, all pattern data points and removed conflicting isolates. Continuing examination of the representation of library data as binary is necessary to determine whether the statistical assumptions in LDA prevent meaningful results.

Local libraries must remain dominant but regional information is useful in filling in gaps. The multi-year library created for this study may have contained regional or dated data, which inflated the variation and importance of the pet class. This study cannot confirm the use of library data from both an entire region and multiple years.

The third objective for the study was to evaluate the fluorometer for measurement of optical brighteners in fresh water. This was partially successful due to the lack of hotspots of high fluorescent brighteners or high human isolate counts. The fluorometer continues to have potential as a metric of waste in freshwater. More work needs to be done to prove its utility. It currently shows minor intrusion of wastewater at two locations, but it has yet to show a high brightener hotspot.

This study indicates that ARA and fluorometry can continue to be a method to indicate sources of bacterial pollution. Due to low performance measurements of ARA, it should be used cautiously because results made with less than enough isolates can be highly influenced my artifacts of analysis. Fluorometry should be evaluated in its detection of freshwater human microbial hotspots but should continue to be used for future monitoring. Library design should be examined with larger KSL's before definitive statements can be made about the utility of the binary data representation, interspatial conflict removal, clustering, unknown class creation and thresholding.

VI. Study Revision Recommendations

Stronger libraries would contain more known source isolates and more data on each isolate. Data for the location and the fecal sample for each isolate should be taken. This would allow geographic weighting of known source isolates to the individual stream site. Such an approach has not been used in MST.

In order to compare the usefulness of known source isolates from a region or multiple years, a complete local library should be created. Resubstitution of isolates into a local library could show any degradation of results from out of region or several year old isolates. Nine hundred or more known source isolates from the OQB would have enabled a comparison for this study.

A larger variety of stream data would have been useful for this study. While precipitation was used to indicate flow rates, a flowmeter or USGS monitoring station interpolation would have allowed a stronger comparison of bacterial concentrations to studies indicating flow dependence.

One assumption of these studies is that fecal pollution will not be obfuscated through other pollution. Chemical testing of the streams during sampling could show influx of inhibitory or stimulatory compounds. Any chemicals affecting indicator growth or loss could greatly affect the results, particularly if isolates are differentially affected.

More statistically advanced clustering and discriminant analysis should be attempted on this dataset. While clustering was attempted on the library in order to identify subclasses, there was no numerical threshold to determine the breakout of a particular subclass. Based on previous studies, when faced with a highly varying pet class, wildlife data might be better represented in multiple subclasses. Quadratic discriminant analysis was suggested as a possible way to satisfy statisticial assumptions when using binary data (USEPA 2005).

While quadratic discriminant analysis and neural network algorithms were attempted, these methods need to be revised. Current results showed overclassification, however this suggests that some limit would allow acceptable classification. Given more time for statistical analysis, this could be well explored.

VII. Suggested Further Research

Verification of the results of this study could be completed in two alternative ways: this study could be repeated after cleanup and remediation measures had been enacted, or by comparison of data throughout the region. For the data comparison independent variables of population densities, housing age, weather, temperature, livestock populations, pet population estimates, wildlife estimates, park land, and land use would be compared against bacterial counts and classification. Other data could be considered for regression as available. This could be done with stream data throughout

the state or EPA region. Any area that closely correlates in either the stream data or the local statistics could indicate watershed cleanup targets or factors.

Two areas of statistical measure of known source libraries (KSLs) should be pursued: standard confidence intervals of classes should be explored, and the existence of subclasses should be determined. The establishment of confidence intervals for each class, even ones that overlap, would help demonstrate representativeness of data. Distinct confidence limits, which are currently incorporated by discriminant analysis, demonstrate whether classes can be separate in analysis. With the inclusion of confidence limits, subclasses could be determined as those groups, which through separation create two groups with a tighter confidence interval than the original class.

Weighting datapoints could also significantly affect the library. Recent work by Ritter et al. 2003 suggested that only unique isolates should be included in the library. The three reasons for a library built of uniques are the lack of standard enumeration data relating each isolate to a percentage of the fecal matter, the rate of feces production a given species, and a population estimate of every species included in the study. Further, environmental survival data for each isolate is too rudimentary for widespread use. Lacking these four factors, repetition of isolates would make the library less representative of the environmental population. Were sample weighting or repetition to be used, consideration should also be given to the geographic relationship between the sample and the sample site.

VIII. References

Hagedorn, C., R. B. Reneau Jr., M. Saluta, and A. Hassall. 2003. Impact of Onsite Wastewater Systems on Water Quality in Coastal Regions. Final project report to the National Oceanic and Atmospheric Administration, Charleston, SC.

Hagedorn, C., A. Hassall, and M. Saluta. 2004. Identifying Sources of Fecal Pollution in Impaired Waters in Prince William County, Virginia. Final project report to Prince William County Department of Public Works, Woodbridge, VA.

Ritter, K. J., Carruthers, E., Carson, C. A., Ellendere, R. D., Harwood, V. J., Kingsley, K., Nakatsu, C., Sadowsky, M., Shear, B., West, B, Whitlock, J. E., Wiggins, B. A., and Wilbur, J. D. 2003. Assessment of statistical methods used in library-based approaches to microbial source tracking. J. Wat. Health 01.4:209-223.

Stewart, J.R., Robinson, B., Hyer, K., Hagedorn, C., Whittam, T.S., Wilbur, J. 2005. Microbial Source Tracking Using Indicator Organisms. American Society for Microbiology General Meeting, Atlanta, GA.

US EPA. 2005. Microbial Source Tracking Guide Document. U. S. Environmental Protection Agency, Office of Research and Development: Washington, D.C.

Glossary

Antibiotic Resistance Analysis (ARA) – a method to determine the source of bacteria based on similarity of antibiotic resistance between bacterial isolates

Binary – a dataset representation used in antibiotic resistance analysis indicating bacterial growth at a particular antibiotic concentration with a 0 or 1.

Category – several sources of bacteria, in this study the four categories of bacteria source are humans, livestock animals, pets and wildlife. Synonym of Class

Challenge Set – several known source isolates used to test the classification ability of a library. These isolates are not used in library creation.

Class – several sources of bacteria, in this study the four categories of bacteria source are humans, livestock animals, pets and wildlife. Synonym of Category

Cluster – a group of antibiotic resistance patterns that are similar to one another. May be separated out from a class to improve classification or alternatively used instead of linear discriminant analysis to determine source categories of environmental isolates

Combination – a dataset representation used in antibiotic resistance analysis indicating the highest concentration of each antibiotic at which bacterial growth occurred as well as the highest concentration below the lowest concentration at which growth failed to occur

Confidence Interval – the range in which a measured of calculated value is likely to occur. May be used interchangeably with Confidence Limit

Confidence Limit – the upper and lower boundaries in between which a measured or calculated value is likely to occur. May be used interchangeably with Confidence Interval

Environmental Isolate – bacterial colony separated into pure culture from a water sample

Fluorometry – measurement of particular chemicals in water by light to indicate contamination with human wastewater

High – a dataset representation used in antibiotic resistance analysis indicating the highest concentration of each antibiotic at which bacterial growth occurred

Isolate – a bacterial colony separated into pure culture. Used to perform antibiotic resistance analysis

Known Source Isolate (KSI) – a bacterial colony separated into pure culture from a the fecal sample of a particular species or source category of animal

Known Source Library (KSL) – a collection of known source isolates used to create a classification rule for

Last – a dataset representation used in antibiotic resistance analysis indicating the greatest antibiotic concentration with bacterial growth below the lowest concentration at which growth failed to occur.

Library – a set of training data used to create a classification rule for linear discriminant analysis. This study uses known source isolates for the training data.

Linear Discriminant Analysis – a statistical method of classification based on the difference between an unclassified value and the average value for each class. The unclassified value is compared to each class average independently.

Sample – unprocessed feces or volume of water containing multiple bacterial colonies

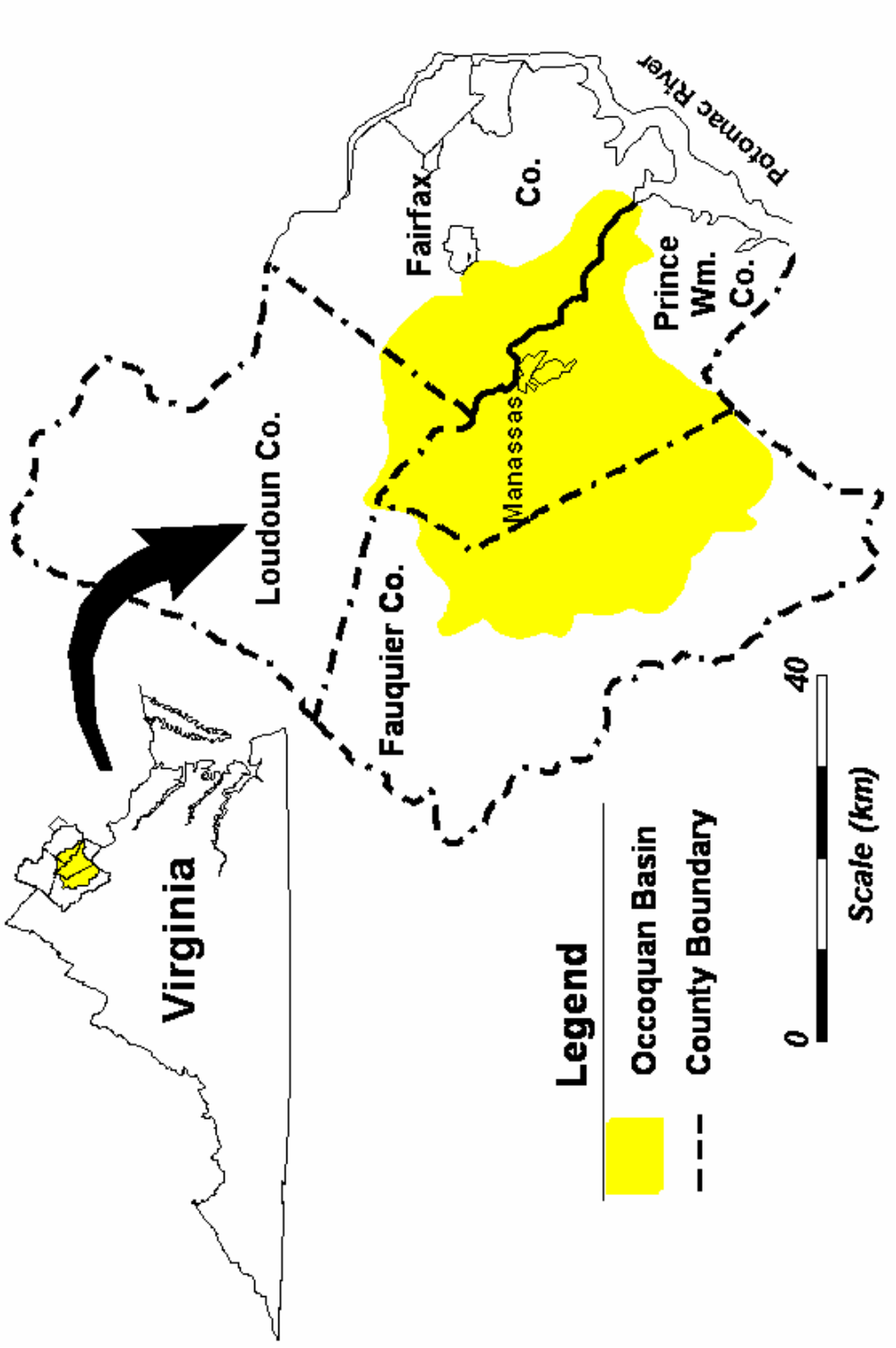
Source – the species or category of animals from which a particular bacterial colony comes from

Subclass – a group of isolates within a class or category. Used to refer to the similar isolate groups identified through clustering

Appendix

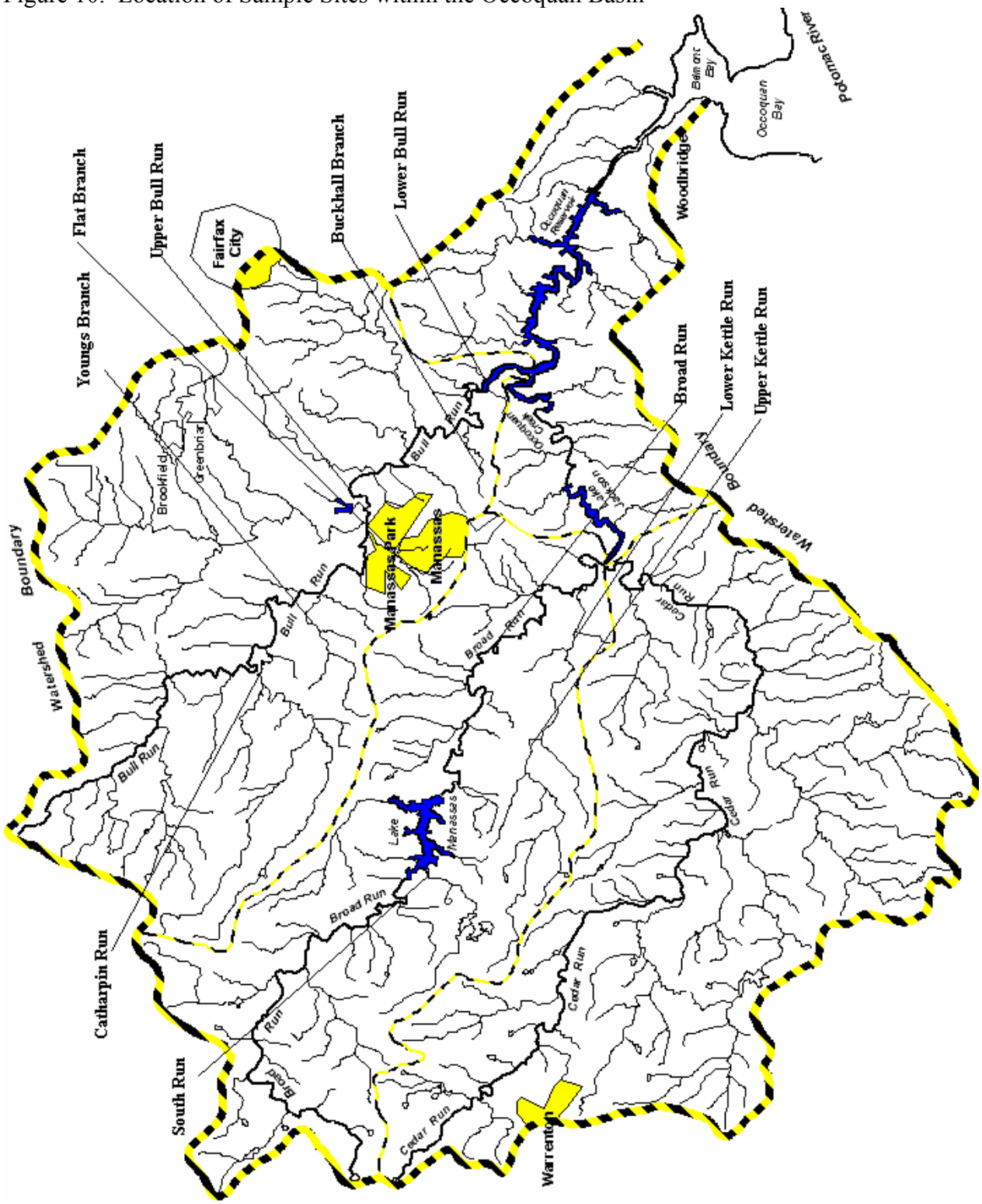
A. Occoquan Basin Maps

Figure 9. Location of Occoquan Basin in Virginia



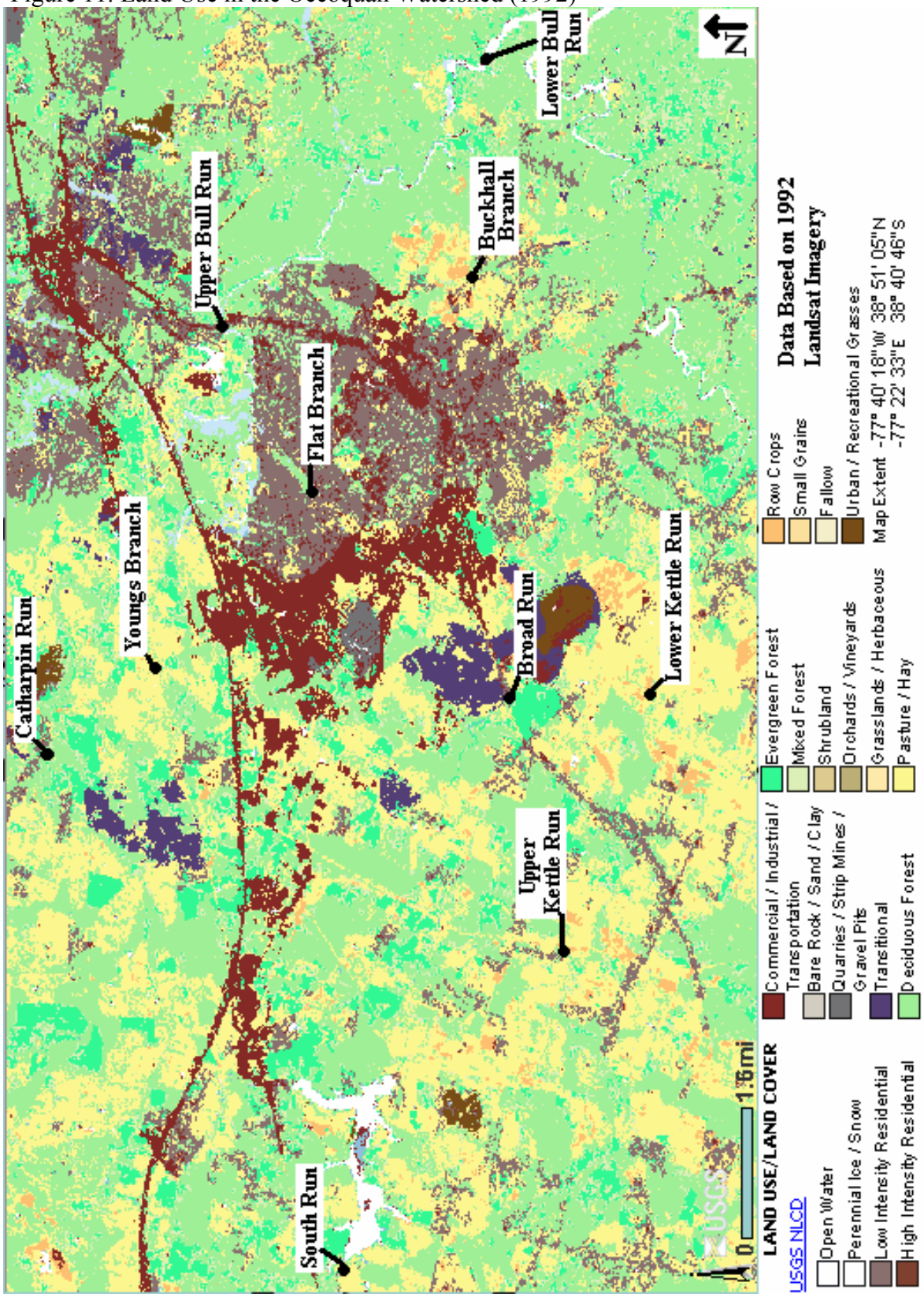
Courtesy of the Northern Virginia Regional Commission - www.novaregion.org/

Figure 10. Location of Sample Sites within the Occoquan Basin



Courtesy of the Northern Virginia Regional Commission - www.novaregion.org/

Figure 11. Land Use in the Occoquan Watershed (1992)



USGS National Land Cover Data National Map

B. Enterococcus Site Data

Table 27. Monthly Upper Bull Run *Enterococcus* Classification

Site	Upper Bull Run		Number of Enterococcus colonies classified			
Sampling	CFU/100[a]	Isolates	Human	Livestock	Pet	Wildlife
June 04	502.5	24	0	4	1	19
July	135	24	0	0	20	4
August	667.5	23	0	0	6	17
September	525	24	0	0	9	15
October	175	24	1	0	6	17
November	55	23	0	0	19	4
December	465	24	0	11	4	9
January	10	24	0	3	3	18
February	100	24	0	7	8	9
March	5	24	3	0	19	2
April	155	20	4	10	1	5
May	30	24	1	11	4	8
June 05	205	23	0	0	0	23
Total	N/A	305	9	46	100	150
Average	233	23.46	0.69	3.54	7.69	11.54
Std. Dev	226	1.13	1.32	4.59	7.15	6.91

[a] Colony Forming Units per 100ml

Table 28. Monthly Lower Bull Run *Enterococcus* Classification

Site	Lower Bull Run		Number of Enterococcus colonies classified			
Sampling	CFU/100[a]	Isolates	Human	Livestock	Pet	Wildlife
June 04	35	4	1	1	0	1
July	55	24	1	0	7	7
August	780	24	1	1	15	7
September	780	22	0	0	13	9
October	215	19	1	0	5	13
November	1205	16	1	0	19	4
December	97.5	18	0	0	16	2
January	35	24	0	5	10	8
February	72	24	0	3	4	17
March	60	24	1	0	19	4
April	347.5	23	1	8	1	13
May	10	24	0	13	9	2
June 05	100	23	0	0	6	17
Total	N/A	269	7	31	124	104
Average	292	20.69	0.54	2.38	9.54	8
Std. Dev	383	5.69	0.52	4.03	6.44	5.54

[a] Colony Forming Units per 100ml

Table 29. Monthly Youngs Branch *Enterococcus* Classification

Site	Youngs Branch		Number of Enterococcus colonies classified			
Sampling	CFU/100[a]	Isolates	Human	Livestock	Pet	Wildlife
June 04	2507.5	7	0	0	0	7
July	650	24	1	1	20	2
August	980	24	0	0	15	9
September	980	24	0	0	13	11
October	555	24	0	0	11	13
November	340	24	0	0	7	17
December	1200	23	0	19	1	3
January	30	24	0	8	4	12
February	1100	24	0	0	21	3
March	5	24	1	10	5	8
April	107.5	24	0	14	1	9
May	37.5	24	1	16	4	3
June 05	260	23	1	1	9	13
Total	N/A	293	4	69	111	110
Average	673	22.54	0.31	5.31	8.54	8.46
Std. Dev	701	4.68	0.48	7.15	7.05	4.72

[a] Colony Forming Units per 100ml

Table 30. Monthly Catharpin *Enterococcus* Classification

Site	Catharpin		Number of Enterococcus colonies classified			
Sampling	CFU/100[a]	Isolates	Human	Livestock	Pet	Wildlife
June 04	195	23	0	11	11	1
July	290	24	0	0	16	8
August	622.5	24	0	0	10	14
September	622.5	24	1	0	10	13
October	360	24	0	0	4	20
November	245	24	2	0	6	16
December	580	24	0	20	1	3
January	37.5	24	2	5	5	12
February	105	24	2	2	17	3
March	7.5	24	1	13	8	2
April	135	24	1	16	3	4
May	10	24	0	16	3	5
June 05	130	23	1	0	9	13
Total	N/A	310	10	83	103	114
Average	257	23.85	0.77	6.38	7.92	8.77
Std. Dev	226	0.38	0.83	7.64	4.92	6.21

[a] Colony Forming Units per 100ml

Table 31. Monthly Buckhall Branch *Enterococcus* Classification

Site	Buckhall Branch		Number of Enterococcus colonies classified			
Sampling	CFU/100[a]	Isolates	Human	Livestock	Pet	Wildlife
June 04	1620	24	10	3	11	0
July	425	24	1	0	18	5
August	1440	24	1	0	16	7
September	1440	24	2	0	22	0
October	385	24	0	0	4	20
November	535	32	0	5	0	19
December	670	24	0	0	24	0
January	85	24	0	11	7	3
February	55	24	9	0	9	6
March	27.5	24	1	0	17	6
April	40	24	2	13	5	4
May	70	22	7	7	6	4
June 05	1285	23	0	1	6	16
Total	N/A	317	33	40	145	90
Average	621	24.38	2.54	3.08	11.15	6.92
Std. Dev	611	2.36	3.62	4.57	7.5	6.96

[a] Colony Forming Units per 100ml

Table 32. Monthly Flat Branch *Enterococcus* Classification

Site	Flat Branch		Number of Enterococcus colonies classified			
Sampling	CFU/100[a]	Isolates	Human	Livestock	Pet	Wildlife
June 04	502.5	20	0	6	2	12
July	1960	24	0	0	14	10
August	1315	24	1	0	8	15
September	1315	24	0	0	19	5
October	350	24	1	0	2	21
November	80	24	2	11	0	11
December	1720	24	0	1	19	4
January	112.5	24	0	2	9	13
February	307.5	24	1	0	15	8
March	25	24	1	1	18	4
April	45	24	2	16	2	4
May	92.5	23	0	18	2	3
June 05	530	24	1	3	9	11
Total	N/A	307	9	58	119	121
Average	643	23.62	0.69	4.46	9.15	9.31
Std. Dev	687	1.12	0.75	6.41	7.19	5.33

[a] Colony Forming Units per 100ml

Table 33. Monthly South Run *Enterococcus* Classification

Site	South Run		Number of Enterococcus colonies classified			
Sampling	CFU/100[a]	Isolates	Human	Livestock	Pet	Wildlife
June 04	82.5	8	0	5	1	2
July	85	20	0	1	13	6
August	85	24	1	0	19	4
September	85	24	0	0	16	8
October	705	24	17	0	4	3
November	140	24	0	5	0	19
December	197.5	23	1	0	22	0
January	40	24	0	2	4	17
February	17.5	15	2	0	10	3
March	2.5	24	0	0	17	7
April	142.5	24	1	17	0	6
May	12.5	20	0	12	6	2
June 05	100	24	0	0	22	2
Total	N/A	278	22	42	134	79
Average	130	21.38	1.69	3.23	10.31	6.08
Std. Dev	182	4.84	4.64	5.42	8.32	5.78

[a] Colony Forming Units per 100ml

Table 34. Monthly Broad Run *Enterococcus* Classification

Site	Broad Run		Number of Enterococcus colonies classified			
Sampling	CFU/100[a]	Isolates	Human	Livestock	Pet	Wildlife
June 04	55	8	0	4	0	4
July	160	24	0	0	14	10
August	947.5	24	0	1	16	7
September	947.5	23	0	0	5	18
October	235	24	1	0	6	17
November	100	14	0	5	0	9
December	440	23	0	0	23	0
January	10	23	1	2	5	16
February	105	24	0	15	2	7
March	12.5	23	3	0	10	10
April	120	24	3	18	0	3
May	0	23	2	8	7	6
June 05	192.5	23	1	0	0	23
Total	N/A	280	11	53	88	130
Average	256	21.54	0.85	4.08	6.77	10
Std. Dev	329	4.86	1.14	6.08	7.21	6.7

[a] Colony Forming Units per 100ml

Table 35. Monthly Lower Kettle Run *Enterococcus* Classification

Site	Lower Kettle Run		Number of Enterococcus colonies classified			
Sampling	CFU/100[a]	Isolates	Human	Livestock	Pet	Wildlife
June 04	340	24	2	5	0	17
July	397.5	24	0	0	16	8
August	1940	24	0	0	19	5
September	1940	23	2	0	8	13
October	1010	24	1	0	23	0
November	555	X[b]	X	X	X	X
December	4800	24	0	0	21	3
January	100	24	0	2	4	18
February	400	24	0	18	4	2
March	77.5	24	2	10	2	10
April	90	24	0	13	4	7
May	20	23	0	10	5	8
June 05	640	24	0	0	1	23
Total	N/A	286	7	58	107	114
Average	947	23.83	0.58	4.83	8.92	9.5
Std. Dev	1326	0.39	0.9	6.34	8.39	7.03

[a] Colony Forming Units per 100ml
[b] Isolates were contaminated

Table 36. Monthly Upper Kettle Run *Enterococcus* Classification

Site	Upper Kettle Run		Number of Enterococcus colonies classified			
Sampling	CFU/100[a]	Isolates	Human	Livestock	Pet	Wildlife
June 04	1165	22	0	0	6	16
July	490	24	0	14	2	8
August	1502.5	19	0	0	16	3
September	1502.5	24	0	0	15	9
October	1300	24	0	0	23	1
November	970	X[b]	X	X	X	X
December	1820	16	0	0	13	3
January	195	24	2	0	16	1
February	225	24	0	4	8	12
March	72.5	24	0	11	8	5
April	262.5	24	1	14	1	8
May	222.5	24	8	9	2	5
June 05	315	24	0	11	5	8
Total	N/A	273	11	63	115	79
Average	773	22.75	0.92	5.25	9.58	6.58
Std. Dev	620	2.6	2.31	6.03	6.95	4.5

[a] Colony Forming Units per 100ml
[b] Isolates were contaminated

C. E. coli *Site Data*

Table 37. Quarterly Upper Bull Run *E. coli* Classification

Site	Upper Bull Run		Number of *E. coli* colonies classified			
Quarter	CFU/100[a]	Isolates	Human	Livestock	Pet	Wildlife
September	3830	19	2	13	0	4
December	145	24	8	2	1	13
March	32.5	23	8	2	6	7
June 05	210	24	11	1	1	11
Total	N/A	90	29	18	8	35
Average	1054	22.50	7.25	4.50	2.00	8.75
Std. Dev	1852	2.38	3.77	5.69	2.71	4.03

[a] Colony Forming Units per 100ml

Table 38. Quarterly Lower Bull Run *E. coli* Classification

Site	Lower Bull Run		Number of *E. coli* colonies classified			
Quarter	CFU/100[a]	Isolates	Human	Livestock	Pet	Wildlife
September	4960	17	3	8	0	6
December	35	8	4	0	0	4
March	95	24	9	8	4	3
June 05	45	24	12	1	1	10
Total	N/A	73	28	17	5	23
Average	1284	18.25	7.00	4.25	1.25	5.75
Std. Dev	2451	7.59	4.24	4.35	1.89	3.10

[a] Colony Forming Units per 100ml

Table 39. Quarterly Youngs Branch *E. coli* Classification

Site	Youngs Branch		Number of *E. coli* colonies classified			
Quarter	CFU/100[a]	Isolates	Human	Livestock	Pet	Wildlife
September	2660	24	7	1	1	15
December	730	24	11	0	0	13
March	20	24	14	1	2	7
June 05	285	22	8	0	0	14
Total	N/A	94	40	2	3	49
Average	924	23.50	10.00	0.50	0.75	12.25
Std. Dev	1194	1.00	3.16	0.58	0.96	3.59

[a] Colony Forming Units per 100ml

Table 40. Quarterly Catharpin *E. coli* Classification

Site	Catharpin		Number of *E. coli* colonies classified			
Quarter	CFU/100[a]	Isolates	Human	Livestock	Pet	Wildlife
September	2675	24	5	7	2	10
December	850	24	19	2	1	2
March	15	24	11	1	0	12
June 05	105	24	18	2	0	4
Total	N/A	96	53	12	3	28
Average	911	24.00	13.25	3.00	0.75	7.00
Std. Dev	1234	0.00	6.55	2.71	0.96	4.76

[a] Colony Forming Units per 100ml

Table 41. Quarterly Buckhall Branch *E. coli* Classification

Site	Buckhall Branch		Number of *E. coli* colonies classified			
Quarter	CFU/100[a]	Isolates	Human	Livestock	Pet	Wildlife
September	2615	24	1	6	2	15
December	655	24	8	4	6	6
March	7.5	24	13	6	2	3
June 05	2150	24	13	6	2	3
Total	N/A	96	35	22	12	27
Average	1357	24.00	8.75	5.50	3.00	6.75
Std. Dev	1228	0.00	5.68	1.00	2.00	5.68

[a] Colony Forming Units per 100ml

Table 42. Quarterly Flat Branch *E. coli* Classification

Site	Flat Branch		Number of *E. coli* colonies classified			
Quarter	CFU/100[a]	Isolates	Human	Livestock	Pet	Wildlife
September	1050	22	0	4	0	18
December	955	24	11	2	0	11
March	10	24	12	3	1	8
June 05	55	20	4	1	1	14
Total	N/A	90	27	10	2	51
Average	518	22.50	6.75	2.50	0.50	12.75
Std. Dev	562	1.91	5.74	1.29	0.58	4.27

[a] Colony Forming Units per 100ml

Table 43. Quarterly South Run *E. coli* Classification

Site	South Run		Number of *E. coli* colonies classified			
Quarter	CFU/100[a]	Isolates	Human	Livestock	Pet	Wildlife
September	840	24	5	0	1	18
December	177.5	24	11	5	0	8
March	40	24	9	2	2	11
June 05	695	24	15	0	1	8
Total	N/A	96	40	7	4	45
Average	438	24.00	10.00	1.75	1.00	11.25
Std. Dev	389	0.00	4.16	2.36	0.82	4.72

[a] Colony Forming Units per 100ml

Table 44. Quarterly Broad Run *E. coli* Classification

Site	Broad Run		Number of *E. coli* colonies classified			
Quarter	CFU/100[a]	Isolates	Human	Livestock	Pet	Wildlife
September	1040	24	3	3	1	17
December	287.5	24	16	0	0	8
March	27.5	24	7	6	2	9
June 05	220	23	2	1	1	19
Total	N/A	95	28	10	4	53
Average	394	23.75	7.00	2.50	1.00	13.25
Std. Dev	445	0.50	6.38	2.65	0.82	5.56

[a] Colony Forming Units per 100ml

Table 45. Quarterly Lower Kettle Run *E. coli* Classification

Site	Lower Kettle Run		Number of *E. coli* colonies classified			
Quarter	CFU/100[a]	Isolates	Human	Livestock	Pet	Wildlife
September	1000	23	9	3	1	10
December	6480	24	1	2	8	13
March	70	24	15	3	1	5
June 05	130	24	13	2	0	9
Total	N/A	95	38	10	10	37
Average	1920	23.75	9.50	2.50	2.50	9.25
Std. Dev	3070	0.50	6.19	0.58	3.70	3.30

[a] Colony Forming Units per 100ml

Table 46. Quarterly Upper Kettle Run *E. coli* Classification

Site	Upper Kettle Run		Number of *E. coli* colonies classified			
Quarter	CFU/100[a]	Isolates	Human	Livestock	Pet	Wildlife
September	5060	24	6	1	1	16
December	1520	24	13	1	0	10
March	245	24	14	0	9	1
June 05	695	24	10	1	0	13
Total	N/A	96	43	3	10	40
Average	1880	24.00	10.75	0.75	2.50	10.00
Std. Dev	2185	0.00	3.59	0.50	4.36	6.48

[a] Colony Forming Units per 100ml

D. Fluorometry Site Data

Table 47. Fluorometry Z values at Upper Bull Run Compared to Human Microbial Counts

Site	Upper Bull Run						
	Enterococcus			*E. coli*			Fluorescence
Month	CFU/100[a]	Isolates	Human	CFU/100	Colonies	Human	Z valueb[b]
June 04	502.5	24	0	-[c]	-	-	0.15
July	135	24	0	-	-	-	-0.19
August	667.5	23	0	-	-	-	-0.04
September	525	24	0	3830	19	2	-1.06
October	175	24	1	-	-	-	-0.96
November	55	23	0	-	-	-	-0.19
January[d]	10	24	0	-	-	-	-1.52
February	100	24	0	-	-	-	-0.19
March	5	24	3	32.5	23	8	0.00
May	30	24	1	-	-	-	-0.71
June 05	205	23	0	210	24	11	-1.05
Total	N/A	305	9	4217.5	90	29	-5.75
Average	233	23.46	0.69	1054	22.50	7.25	-0.52
Std. Dev	226	1.13	1.32	1852	2.38	3.77	0.56

[a] Colony Forming Units per 100 ml
[b] Z values are calculated from the average and standard deviation from each month
[c] ARA was not performed on *E. coli* for these months
[d] December and April data unavailable

Table 48. Fluorometry Z values at Lower Bull Run Compared to Human Microbial Counts

Site	Lower Bull Run						
	Enterococcus			*E. coli*			Fluorescence
Month	CFU/100[a]	Isolates	Human	CFU/100	Colonies	Human	Z value[b]
June 04	35	4	1	-[c]	-	-	-0.57
July	55	24	1	-	-	-	-0.77
August	780	24	1	-	-	-	0.15
September	780	22	0	4960	17	3	-0.76
October	215	19	1	-	-	-	0.40
November	1205	16	1	-	-	-	-0.77
January[d]	35	24	0	-	-	-	-1.64
February	72	24	0	-	-	-	-2.34
March	60	24	1	95	24	9	1.74
May	10	24	0	-	-	-	1.11
June 05	100	23	0	45	24	12	-0.03
Total	N/A	269	7	5135	73	28	-3.50
Average	292	20.69	0.54	1284	18.25	7.00	-0.32
Std. Dev	383	5.69	0.52	2451	7.59	4.24	1.17

[a] Colony Forming Units per 100 ml
[b] Z values are calculated from the average and standard deviation from each month
[c] ARA was not performed on *E. coli* for these months
[d] December and April data unavailable

Table 49. Fluorometry Z values at Youngs Branch Compared to Human Microbial Counts

Site	Youngs Branch						
	Enterococcus			*E. coli*			Fluorescence
Month	CFU/100[a]	Isolates	Human	CFU/100	Colonies	Human	Z value[b]
June 04	2507.5	7	0	-[c]	-	-	-0.89
July	650	24	1	-	-	-	1.99
August	980	24	0	-	-	-	1.46
September	980	24	0	2660	24	7	-0.09
October	555	24	0	-	-	-	1.47
November	340	24	0	-	-	-	1.99
January[d]	30	24	0	-	-	-	0.78
February	1100	24	0	-	-	-	0.11
March	5	24	1	20	24	14	0.78
May	37.5	24	1	-	-	-	0.81
June 05	260	23	1	285	22	8	1.27
Total	N/A	293	4	3695	94	40	9.69
Average	673	22.54	0.31	924	23.50	10.00	0.88
Std. Dev	701	4.68	0.48	1194	1.00	3.16	0.90

[a] Colony Forming Units per 100 ml
[b] Z values are calculated from the average and standard deviation from each month
[c] ARA was not performed on *E. coli* for these months
[d] December and April data unavailable

Table 50. Fluorometry Z values at Catharpin Compared to Human Microbial Counts

Site	Catharpin						
	Enterococcus			*E. coli*			Fluorescence
Month	CFU/100[a]	Isolates	Human	CFU/100	Colonies	Human	Z value[b]
June 04	195	23	10	-[c]	-	-	0.12
July	290	24	1	-	-	-	0.13
August	622.5	24	1	-	-	-	0.58
September	622.5	24	2	2675	24	5	0.78
October	360	24	0	-	-	-	-0.06
November	245	24	0	-	-	-	0.13
January[d]	37.5	24	0	-	-	-	-0.90
February	105	24	9	-	-	-	-0.14
March	7.5	24	1	15	24	11	-0.89
May	10	24	7	-	-	-	-1.08
June 05	130	23	0	105	24	18	0.53
Total	N/A	310	33	3645	96	53	-0.81
Average	257	23.85	2.54	911	24.00	13.25	-0.07
Std. Dev	226	0.38	3.62	1234	0.00	6.55	0.63

[a] Colony Forming Units per 100 ml
[b] Z values are calculated from the average and standard deviation from each month
[c] ARA was not performed on *E. coli* for these months
[d] December and April data unavailable

Table 51. Fluorometry Z values at Buckhall Branch Compared to Human Microbial Counts

Site	Buckhall Branch						
	Enterococcus			*E. coli*			Fluorescence
Month	CFU/100[a]	Isolates	Human	CFU/100	Colonies	Human	Z value[b]
June 04	1620	24	10	-[c]	-	-	1.13
July	425	24	1	-	-	-	0.98
August	1440	24	1	-	-	-	-0.31
September	1440	24	2	2615	24	1	1.05
October	385	24	0	-	-	-	0.14
November	535	32	0	-	-	-	0.98
January[d]	85	24	0	-	-	-	0.15
February	55	24	9	-	-	-	-0.57
March	27.5	24	1	7.5	24	13	0.29
May	70	22	7	-	-	-	0.36
June 05	1285	23	0	2150	24	13	0.25
Total	N/A	317	33	5427.5	96	35	4.45
Average	621	24.38	2.54	1357	24.00	8.75	0.40
Std. Dev	611	2.36	3.62	1228	0.00	5.68	0.57

[a] Colony Forming Units per 100 ml
[b] Z values are calculated from the average and standard deviation from each month
[c] ARA was not performed on *E. coli* for these months
[d] December and April data unavailable

Table 52. Fluorometry Z values at Flat Branch Compared to Human Microbial Counts

Site	Flat Branch						
	Enterococcus			*E. coli*			Fluorescence
Month	CFU/100[a]	Isolates	Human	CFU/100	Colonies	Human	Z value[b]
June 04	502.5	20	0	-[c]	-	-	-1.07
July	1960	24	0	-	-	-	-0.30
August	1315	24	1	-	-	-	-1.06
September	1315	24	0	1050	22	0	-1.34
October	350	24	1	-	-	-	-1.16
November	80	24	2	-	-	-	-0.30
January[d]	112.5	24	0	-	-	-	0.16
February	307.5	24	1	-	-	-	-0.06
March	25	24	1	10	24	12	-0.14
May	92.5	23	0	-	-	-	-0.36
June 05	530	24	1	55	20	4	-0.38
Total	N/A	307	9	2070	90	27	-6.00
Average	643	23.62	0.69	518	22.50	6.75	-0.55
Std. Dev	687	1.12	0.75	562	1.91	5.74	0.51

[a] Colony Forming Units per 100 ml
[b] Z values are calculated from the average and standard deviation from each month
[c] ARA was not performed on *E. coli* for these months
[d] December and April data unavailable

Table 53. Fluorometry Z values at South Run Compared to Human Microbial Counts

Site	South Run						
	Enterococcus			*E. coli*			Fluorescence
Month	CFU/100[a]	Isolates	Human	CFU/100	Colonies	Human	Z value[b]
June 04	82.5	8	0	-[c]	-	-	-1.39
July	85	20	0	-	-	-	-1.65
August	85	24	1	-	-	-	-1.97
September	85	24	0	840	24	5	-0.86
October	705	24	17	-	-	-	-1.68
November	140	24	0	-	-	-	-1.65
January[d]	40	24	0	-	-	-	0.73
February	17.5	15	2	-	-	-	0.80
March	2.5	24	0	40	24	9	-1.79
May	12.5	20	0	-	-	-	-1.92
June 05	100	24	0	695	24	15	-1.11
Total	N/A	278	22	1752.5	96	40	-12.49
Average	130	21.38	1.69	438	24.00	10.00	-1.14
Std. Dev	182	4.84	4.64	389	0.00	4.16	1.00

[a] Colony Forming Units per 100 ml
[b] Z values are calculated from the average and standard deviation from each month
[c] ARA was not performed on *E. coli* for these months
[d] December and April data unavailable

Table 54. Fluorometry Z values at Broad Run Compared to Human Microbial Counts

Site	Broad Run						
	Enterococcus			*E. coli*			Fluorescence
Month	CFU/100[a]	Isolates	Human	CFU/100	Colonies	Human	Z value[b]
June 04	55	8	0	-[c]	-	-	-0.11
July	160	24	0	-	-	-	-0.53
August	947.5	24	0	-	-	-	-0.17
September	947.5	23	0	1040	24	3	-0.05
October	235	24	1	-	-	-	0.09
November	100	14	0	-	-	-	-0.53
January[d]	10	23	1	-	-	-	0.30
February	105	24	0	-	-	-	1.31
March	12.5	23	3	27.5	24	7	-0.75
May	0	23	2	-	-	-	1.06
June 05	192.5	23	1	220	23	2	-1.40
Total	N/A	280	11	1575	95	28	-0.79
Average	256	21.54	0.85	394	23.75	7.00	-0.07
Std. Dev	329	4.86	1.14	445	0.50	6.38	0.78

[a] Colony Forming Units per 100 ml
[b] Z values are calculated from the average and standard deviation from each month
[c] ARA was not performed on *E. coli* for these months
[d] December and April data unavailable

Table 55. Fluorometry Z values at Lower Kettle Run Compared to Human Microbial Counts

Site	Lower Kettle Run						
	Enterococcus			*E. coli*			Fluorescence
Month	CFU/100[a]	Isolates	Human	CFU/100	Colonies	Human	Z value[b]
June 04	340	24	2	-[c]	-	-	1.35
July	397.5	24	0	-	-	-	-0.20
August	1940	24	0	-	-	-	1.17
September	1940	23	2	1000	23	9	0.79
October	1010	24	1	-	-	-	0.91
November	555	X[d]	X	-	-	-	-0.20
January[e]	100	24	0	-	-	-	1.06
February	400	24	0	-	-	-	0.84
March	77.5	24	2	70	24	15	0.83
May	20	23	0	-	-	-	0.58
June 05	640	24	0	130	24	13	0.34
Total	N/A	286	7	7680	95	38	7.47
Average	947	23.83	0.58	1920	23.75	9.50	0.68
Std. Dev	1326	0.39	0.90	3070	0.50	6.19	0.51

[a] Colony Forming Units per 100 ml
[b] Z values are calculated from the average and standard deviation from each month
[c] ARA was not performed on *E. coli* for these months
[d] November isolates lost due to contamination
[e] December and April data unavailable

Table 56. Fluorometry Z values at Upper Kettle Run Compared to Human Microbial Counts

Site	Upper Kettle Run						
	Enterococcus			*E. coli*			Fluorescence
Month	CFU/100[a]	Isolates	Human	CFU/100	Colonies	Human	Z value[b]
June 04	1165	22	0	-[c]	-	-	1.29
July	490	24	0	-	-	-	0.54
August	1502.5	19	0	-	-	-	0.18
September	1502.5	24	0	5060	24	6	1.55
October	1300	24	0	-	-	-	0.86
November	970	X[d]	X	-	-	-	0.54
January[e]	195	24	2	-	-	-	0.89
February	225	24	0	-	-	-	0.23
March	72.5	24	0	245	24	14	-0.06
May	222.5	24	8	-	-	-	0.14
June 05	315	24	0	695	24	10	1.58
Total	N/A	273	11	7520	96	43	7.72
Average	773	22.75	0.92	1880	24.00	10.75	0.70
Std. Dev	620	2.60	2.31	2185	0.00	3.59	0.58

[a] Colony Forming Units per 100 ml
[b] Z values are calculated from the average and standard deviation from each month
[c] ARA was not performed on *E. coli* for these months
[d] November isolates lost due to contamination
[e] December and April data unavailable

E. PFGE E. coli *Cross-Verification*

Table 57. PFGE *E. coli* Cross-Verification		
Isolate ID	*E. coli* ARA	PFGE
December1	Human	wildlife
December2	wildlife	Human
December3	Human	Pets
December4	Human	Human
December5	wildlife	Livestock
December6	Human	Pets
December7	Human	Human
December8	Human	Wildlife
January1	Human	Human
January2	Human	Human
January3	Human	Human
January4	wildlife	Pets
January5	Human	Human
January6	Human	Wildlife
January7	Human	Wildlife
January8	Human	Pets
January9	wildlife	Livestock
January10	wildlife	Human
January11	Human	Human
Rate of Correct Classification 7/19		

F. Population Estimation

Table 58. Population Estimation at Upper Bull Run

Site	Upper Bull Run			*Enterococcus*		
Sampling	CFU/100[a]	Isolates	Unique ARA Patterns	Total Isolates	New ARA Patterns	Estimated Population
June	502.5	24	16	24	16	18
July	135	24	20	48	20	51
August	667.5	23	17	71	15	69
September	525	24	12	95	4	48
October	175	24	6	119	3	21
November	55	23	15	142	6	26
December	465	24	21	166	20	35
January	10	24	12	190	6	33
February	100	24	13	214	8	31
March	5	24	21	238	20	38
April	155	20	17	258	17	45
May	30	24	20	282	12	45
June 2	205	23	13	305	5	40
Total	N/A	305	203	305	152	40

[a] Colony Forming Units per 100ml

Table 59. Population Estimation at Lower Bull Run

Site	Lower Bull Run			*Enterococcus*		
Sampling	CFU/100[a]	Isolates	Unique ARA Patterns	Total Isolates	New ARA Patterns	Estimated Population
June	35	4	4	4	4	4
July	55	24	8	28	8	17
August	780	24	21	52	19	40
September	780	22	13	74	9	37
October	215	19	10	93	8	23
November	1205	16	16	109	6	24
December	97.5	18	18	127	11	29
January	35	24	16	151	13	37
February	72	24	12	175	8	31
March	60	24	21	199	21	39
April	347.5	23	18	222	18	48
May	10	24	16	246	9	51
June 2	100	23	14	269	8	52
Total	N/A	269	187	269	142	52

[a] Colony Forming Units per 100ml

Table 60. Population Estimation at Youngs Branch

Site	Youngs Branch			*Enterococcus*		
Sampling	CFU/100[a]	Isolates	Unique ARA Patterns	Total Isolates	New ARA Patterns	Estimated Population
June	2507.5	7	7	7	7	N/A
July	650	24	23	31	23	465
August	980	24	17	55	17	149
September	980	24	10	79	8	62
October	555	24	13	103	7	50
November	340	24	11	127	7	25
December	1200	23	20	150	20	34
January	30	24	13	174	7	39
February	1100	24	10	198	10	47
March	5	24	16	222	11	50
April	107.5	24	17	246	13	53
May	37.5	24	13	270	1	53
June 2	260	23	17	293	8	52
Total	N/A	293	187	293	139	52

[a] Colony Forming Units per 100ml

Table 61. Population Estimation at Catharpin

Site	Catharpin			*Enterococcus*		
Sampling	CFU/100[a]	Isolates	Unique ARA Patterns	Total Isolates	New ARA Patterns	Estimated Population
June	195	23	16	23	16	16
July	290	24	14	47	14	33
August	622.5	24	17	71	15	41
September	622.5	24	11	95	10	42
October	360	24	9	119	6	26
November	245	24	11	143	6	17
December	580	24	20	167	19	23
January	37.5	24	15	191	10	27
February	105	24	13	215	10	31
March	7.5	24	18	239	13	36
April	135	24	17	263	11	40
May	10	24	12	287	4	40
June 2	130	23	15	310	12	46
Total	N/A	310	188	301	146	46

[a] Colony Forming Units per 100ml

Table 62. Population Estimation at Buckhall Branch

Site	Buckhall Branch			*Enterococcus*		
Sampling	CFU/100[a]	Isolates	Unique ARA Patterns	Total Isolates	New ARA Patterns	Estimated Population
June	1620	24	17	24	17	28
July	425	24	17	48	17	35
August	1440	24	23	72	21	71
September	1440	24	17	96	16	97
October	385	24	8	120	4	32
November	535	32	12	152	11	43
December	670	24	15	176	13	54
January	85	24	13	200	11	64
February	55	24	19	224	16	77
March	27.5	24	21	248	17	85
April	40	24	21	272	17	93
May	70	22	21	294	19	107
June 2	1285	23	20	317	20	123
Total	N/A	317	224	317	199	123

[a] Colony Forming Units per 100ml

Table 63. Population Estimation at Flat Branch

Site	Flat Branch			*Enterococcus*		
Sampling	CFU/100[a]	Isolates	Unique ARA Patterns	Total Isolates	New ARA Patterns	Estimated Population
June	502.5	20	13	20	13	19
July	1960	24	16	44	16	50
August	1315	24	18	68	14	63
September	1315	24	12	92	9	60
October	350	24	6	116	4	22
November	80	24	13	140	12	26
December	1720	24	18	164	13	34
January	112.5	24	16	188	10	33
February	307.5	24	19	212	17	42
March	25	24	21	236	17	47
April	45	24	17	260	13	50
May	92.5	23	11	283	5	51
June 2	530	24	16	307	10	55
Total	N/A	307	196	307	153	55

[a] Colony Forming Units per 100ml

Table 64. Population Estimation at South Run

Site	South Run			*Enterococcus*		
Sampling	CFU/100[a]	Isolates	Unique ARA Patterns	Total Isolates	New ARA Patterns	Estimated Population
June	82.5	8	5	8	5	N/A
July	85	20	11	28	10	14
August	85	24	18	52	16	33
September	85	24	8	76	7	38
October	705	24	14	100	12	48
November	140	24	7	124	6	26
December	197.5	23	9	147	6	32
January	40	24	9	171	5	22
February	17.5	15	9	186	9	26
March	2.5	24	23	210	12	31
April	142.5	24	17	234	14	32
May	12.5	20	7	254	1	31
June 2	100	24	14	278	12	36
Total	N/A	278	151	278	115	36

[a] Colony Forming Units per 100ml

Table 65. Population Estimation at Broad Run

Site	Broad Run			*Enterococcus*		
Sampling	CFU/100[a]	Isolates	Unique ARA Patterns	Total Isolates	New ARA Patterns	Estimated Population
June	55	8	8	8	8	N/A
July	160	24	11	32	11	45
August	947.5	24	22	56	17	67
September	947.5	23	13	79	8	44
October	235	24	10	103	6	20
November	100	14	8	117	7	24
December	440	23	13	140	10	30
January	10	23	15	163	11	30
February	105	24	15	187	13	37
March	12.5	23	17	210	13	42
April	120	24	20	234	15	48
May	0	23	20	257	13	53
June 2	192.5	23	16	280	12	60
Total	N/A	280	188	280	144	60

[a] Colony Forming Units per 100ml

Table 66. Population Estimation at Lower Kettle Run

Site	Lower Kettle Run			*Enterococcus*		
Sampling	CFU/100[a]	Isolates	Unique ARA Patterns	Total Isolates	New ARA Patterns	Estimated Population
June	340	24	15	24	15	15
July	397.5	24	17	48	17	42
August	1940	24	14	72	12	50
September	1940	23	11	95	8	51
October	1010	24	14	119	10	63
November	555	24	X[b]	119	X	63
December	4800	16	7	143	4	55
January	100	24	7	167	7	43
February	400	24	14	191	14	54
March	77.5	24	14	215	12	45
April	90	24	21	239	21	55
May	20	23	17	262	17	65
June 2	640	24	11	286	10	73
Total	N/A	302	162	286	147	73

[a] Colony Forming Units per 100ml
[b] Isolates were contaminated

Table 67. Population Estimation at Upper Kettle Run

Site	Upper Kettle Run			*Enterococcus*		
Sampling	CFU/100[a]	Isolates	Unique ARA Patterns	Total Isolates	New ARA Patterns	Estimated Population
June	1165	22	8	22	8	3
July	490	24	16	46	16	13
August	1502.5	19	11	65	11	22
September	1502.5	24	15	89	10	28
October	1300	24	13	113	9	36
November	970	24	X[b]	113	X	36
December	1820	16	7	129	3	39
January	195	24	12	153	8	45
February	225	24	16	177	9	50
March	72.5	24	13	201	9	49
April	262.5	24	20	225	20	60
May	222.5	24	17	249	17	72
June 2	315	24	16	273	16	85
Total	N/A	297	164	273	136	85

[a] Colony Forming Units per 100ml
[b] Isolates were contaminated