Tracing Pollution Sources in an Urban Watershed:

A GIS-based Predictive Model of Bacteria Contamination in Stormwater

Andrew J. Hrycyna

A Thesis in the Field of Sustainability and Environmental Management

for the Degree of Master of Liberal Arts in Extension Studies

Harvard University

May 2016

Abstract

This project explores whether it is possible to predict which urban stormwater pipe networks are most likely to be contaminated by wastewater inputs based on geographic information about the areas they drain. Wastewater pollution introduced into urban stormwater systems is a major source of impairment of water bodies in the United States, introducing pathogens and other pollutants, into rivers, streams, and lakes. Recent stormwater permits require extensive testing for bacteria. Efficiencies could be gained in remediating problems if an evidence-based prioritization scheme could target stormwater pipe networks based on publicly available information. I use a large data set of bacteria data in stormwater from the Mystic River watershed in Massachusetts, along with a GIS methodology, to explore a hypothesis that some features of the stormwater networks and the land they drain can usefully predict which networks will exhibit high bacteria values. Multiple regression analysis shows that pipe length, population density, and age of buildings in an area are significant predictors of high bacteria concentrations in the Mystic River dataset. In addition, I use the final regression model to estimate bacteria loads from stormwater outfalls. I conclude that the evidence supports a pollution-tracking prioritization scheme that tests large pipe networks first, at a minimum. I discuss the possible reasons for this somewhat surprising result, and suggest further ways to extend and refine this modeling approach.

Acknowledgements

I thank Mark Leighton enormously for his generous guidance, patience, keen insights, and contagious curiosity. Mark gave me exactly the help I needed to navigate the project, intellectually and otherwise.

Teams of interns and staff at the Mystic River Watershed Association helped create the catchment data used in this project. I especially thank Stephanie Clark, of Tufts University for helping edit and verify this data. Patrick Herron of the Mystic River Watershed Association suggested this topic to me and provided essential vision, detailed advice, and encouragement.

I thank my fellow classmates in the thesis proseminar for inspiration and camaraderie: Hawk Arachy, Elizabeth Youngblood, Claire Vaterlaus-Staby, Dylan Frazier, Allison Harwick, and Anna Cimini.

Finally, I thank Alice for encouraging me in the first place, and all along.

Table of Contents

List of Tables

List of Figures

Chapter I

Introduction

Polluted stormwater—rainwater flowing over paved and unpaved surfaces and through storm drain networks into rivers and streams—presents the most serious water pollution problem in many parts of the United States today, especially in urban areas (National Research Council, 2009). In Massachusetts, pollution introduced by stormwater pipes is the most important single source of violations of water quality standards in the state's rivers, streams and lakes, according to the Massachusetts Department of Environmental Protection (Massachusetts Department of Environmental Protection [MassDEP], 2012).

One major source of contaminants in stormwater systems is the wastewater sewer system—separate pipes that normally convey wastewater from homes and businesses to wastewater treatment plants, not directly to water bodies. Unwanted and untreated sewage in stormwater originating from leaking wastewater sewer pipes or illicit connections to the stormwater pipe network introduces harmful pathogens and other pollutants directly into water bodies, degrading ecosystems and posing a threat to public health. Tracing the sources of untreated sewage contamination in stormwater is therefore a high priority for water resource managers (Brown, Caraco, & Pitt, 2004). Bacteria in stormwater are principal targets of regulation both because bacteria themselves pose a public health threat and because certain bacteria serve as a robust marker of sanitary sewer contamination in general. Bacteria contamination is the leading cause of violations

of water quality standards in river and streams in the United States (72,305 miles officially designated impaired by pathogens) (U.S. Environmental Protection Agency [EPA], 2009).

Yet the task of locating the sources of bacteria in waterways is difficult in urban areas, where there can be hundreds of stormwater outfalls (ends of storm pipe networks emptying into a water body, see Figure 1) in a municipality. Some stormwater regulations issued by EPA require regular bacteria testing of every outfall (EPA, 2015a); but this is labor intensive and slow.

Figure 1. Stormwater infrastructure. Stormwater outfalls on a river (dots) and gravity storm sewer mains (lines) draining to them. Data source: ESRI, City of Medford, MA.

Research Significance and Objectives

Some methods of prioritizing outfalls with certain characteristics for pollution testing have been proposed (Neponset Stormwater Partnership, 2015; EPA, 2015a).

Some of these prioritization schemes emerge from common-sense considerations, but few if any are based on empirically tested models that would predict which outfalls to target first.

Significant efficiency could be gained if managers were able to predict in advance which outfalls were more likely to be introducing storm water with high concentrations of bacteria. Furthermore, a system that translated a prediction of high concentrations into a prediction of which outfalls are most likely to be contributing the highest total loads of bacteria would be of even higher interest to managers. These outfalls would be natural targets of prioritization.

This thesis explores a large dataset of water quality information from stormwater pipes in the Mystic River watershed in Massachusetts together with publicly available geographic information about population, pipe networks, impervious surface and other factors. My goal is explore to determine whether it is possible to create a predictive model that suggests—before bacteria testing—which stormwater networks in a geographic area are likely to be the greatest contributors of bacteria to water bodies, and thus which are the highest priorities for testing and remediation, based on the geographic characteristics of the land and infrastructure that drain to them. Such a model might help increase the pace of environmental improvements.

Background on Infrastructure and Management

The environmental problems posed by stormwater in urban areas emerge from the complex effects of the built environment on the rivers, streams, and lakes in urban landscapes.

Negative Impacts of Stormwater on Urban Waterways

Urbanized areas are characterized by a high percentage of impervious surface—sidewalks, roads, parking lots, and roofs. Stormwater systems comprise the infrastructure that carries rainwater from these impervious areas down storm drains and into catch basins and stormwater pipes, ultimately conveying this water to nearby water bodies.

A high density of impervious surface has negative effects on the hydrology and ecology of urban streams (Booth, Hartley, & Jackson, 2002; Klein, 1979; Konrad & Booth, 2005). Stormwater networks draining pavement vastly increase the speed with which water is conveyed to water bodies, compared to unpaved landscapes of forest or grassland. Channel erosion and sediment delivery are increased, temperature patterns are changed, peak flows are increased, and base flow diminished (Bernhardt & Palmer, 2007). Stormwater also sweeps material on impervious surfaces into the stream, introducing a variety of pollutants—from petroleum to salt to pesticides to toxic metals to excess nutrients such as phosphorus—into surface water bodies (Paul & Meyer, 2001). Stormwater systems are subject to other pollutions inputs as well, as described below.

The hydrologic effects of paving cities and the pollution introduced by stormwater systems have strikingly negative ecological impacts: even areas with impervious surface cover of 10-20% show significant negative biological effects on riverine ecosystems, with negative effects increasing with percent impervious surface (Schueler, Fraley-McNeal, & Cappiella, 2009). Heavily urbanized areas can easily have impervious surface cover of 40% or more. Stormwater pipe networks in such areas are thus highly significant factors in the health and safety of aquatic environments.

How Wastewater Enters Stormwater Systems

One important component of stormwater pollution in most urban areas is contamination of stormwater by wastewater. The mechanisms of this contamination are not well known by most citizens.

Municipal sewer systems are designed—for the most part—to keep waste and stormwater streams separate. Wastewater from household toilets and sinks and commercial facilities is typically directed through wastewater sewer pipes to a wastewater treatment plant. Only after treatment is wastewater reintroduced into the environment, often far from the originating source.

On a large scale, the United States has been successful in treating and rerouting wastewater, mitigating the negative effects of wastewater on water bodies. The clean up of Boston Harbor in Massachusetts is a well-known success story, founded on the building of a centralized state-of-the-art wastewater treatment facility. Reductions of 80-90% in nitrogen, phosphorus, total suspended solids in Boston Harbor have been documented, for example (Taylor, 2010). Material that used to be conveyed directly to a river or the shore is now removed or treated, reducing the impact of cities on the aquatic environment.

But not all sewage is conveyed successfully to water treatment facilities in urban areas. There are at least three major mechanisms by which untreated wastewater can mix directly with stormwater and thus be introduced into surface water bodies:

1. Combined sewer overflow systems (CSOs): In many older municipalities, including several in the greater Boston area, waste and stormwater pipe networks are combined

into a single stream, sent together to wastewater treatment facilities. In high volume rain events, when wastewater treatment plants reach their maximum input levels, these combined sewer systems are designed to overflow into pipes that lead directly to nearby rivers or streams. These are engineered, episodic releases of raw sewage, subject to a suite of regulations and controls.

2. Decaying infrastructure and infiltration: More typically, storm and wastewater systems are separated. But both storm and waste sewer systems in older cities can be many decades old. The two systems tend to follow streets and tend to be laid in close proximity to each other, even in the same trench. As old pipes decay, leak, and collapse, the two streams of material can intermix, and raw sewage finds its way into pipes or channels conveying storm water in modern American cities. Sewage inputs can be highest when it rains, as the ground becomes saturated, and cross-talk between the two systems is most likely. Studies have shown broken sanitary sewer lines to be a significant source of storm pipe contamination (Brown et al., 2004).
3. Illicit connections: Although against code and against the law, old cities typically have many sites where residential or commercial waste pipes are connected directly to the storm drain system, either by accident or intentionally. Illicit connections result in significant introduction of raw sewage into storm systems and thus into surface water bodies (Brown et al., 2004).

Combined sewer overflows are well-tracked sources of pollution. Their elimination through the creation of separate sewer networks is expensive, but their inputs can be relatively easily monitored.

The other two mechanisms of wastewater introduction into stormwater are more difficult to detect, trace, and quantify. It is these two mechanisms that are main subject of this study.

Why Wastewater Pollution Is a Problem

Wastewater inputs into surface water bodies have documented negative effects on both recreational safety and ecosystem health. Levels of pathogens like *E. coli* routinely exceed recommended safety standards for recreational boating and swimming in rivers and lakes when urban areas experience heavy rainfall events. To show this, I plotted in-stream data from three sampling locations on the Mystic and Malden Rivers, showing bacteria concentrations against the amount of rainfall in the previous 48 hours (Figure 2).

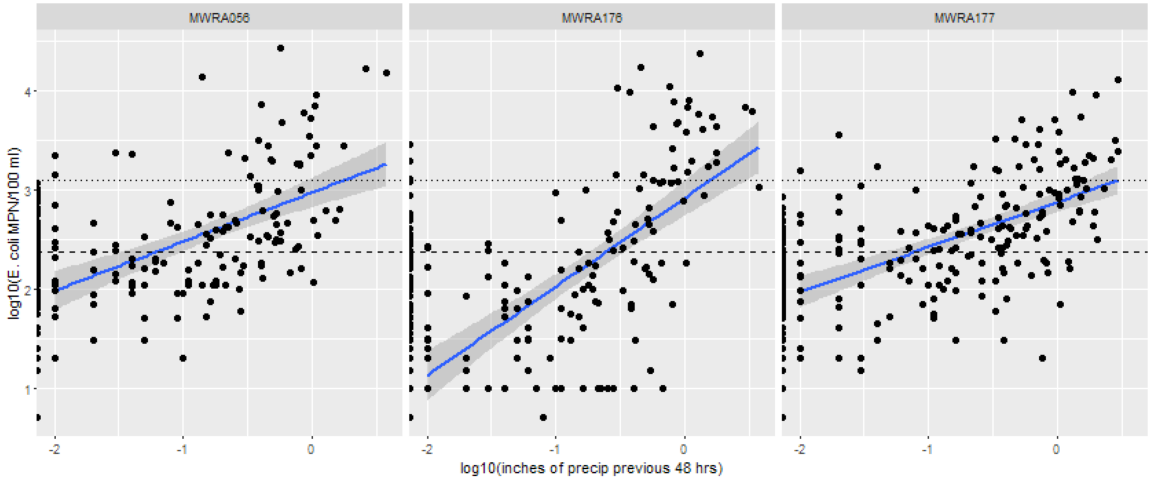

Figure 2. Evidence of stormwater pollution in streams. Plots show log of rainfall inches in previous 48 hours vs. log of *E. coli* concentrations in stream samples. Plots show increase in average *E. coli* levels in wet weather at three locations on the Malden and Mystic Rivers in Massachusetts, often exceeding water quality standards after rain. Dashed and dotted lines show MassDEP swimming and boating safety standards respectively. Data sources: MWRA; MyRWA; MassDEP.

Higher concentrations of bacteria in these streams tend to be associated with higher antecedent rainfall. The source of such high bacteria levels is presumed to be largely the effect of sewage-tainted stormwater, introduced by the three mechanisms described above. Polluted stormwater thus becomes a vector for disease transmission, reducing the value and safety of waterways as sites of recreation.

Of course, human pathogens are not the only pollutants conveyed in urban wastewater. Every substance residents or commercial establishments flush or rinse down drains ends up in wastewater as well. Many of these products and by-products of urban life are toxic or otherwise detrimental to freshwater ecosystems. Contaminants include: nutrients, in particular phosphorus, a limiting nutrient in freshwater systems, which can cause eutrophication of streams and blooms of toxic cyanobacteria; paints and solvents; cosmetic products; cleaning agents; and pharmaceuticals. As one example, endocrine disruptors still remaining in treated wastewater plant effluent have been shown to have negative effects on fish in receiving water bodies (Barber et al., 2011). It is reasonable to expect that the presence of such organic pollutants in raw sewage introduced to stormwater by illicit connections would have a similar effect.

Managing Wastewater Contamination in Stormwater

Given these negative human health and ecosystem effects, regulations emerging from the Clean Water Act in the United States impose on municipalities obligations to reduce and eliminate sources of wastewater pollution in stormwater. Under the jurisdiction of the Clean Water Act's National Pollutant Discharge Elimination System (NPDES), municipal stormwater systems are regulated by the Environmental Protection

Agency (EPA). In particular, state agencies delegated to enforce the Clean Water Act (and in some cases EPA itself) issue so-called "MS4" permits—"MS4" referring to "municipal separate sanitary sewer systems"— that require municipalities to develop programs to investigate, detect, and eliminate illicit connections and infiltration problems (Brown et al., 2004).

Protocols and procedures for tracking contamination of stormwater are known as Illicit Discharge Detection and Elimination (IDDE) programs. The methods used in IDDE are various and include dye testing, testing of receiving water bodies, citizen complaint hotlines, and other mechanisms. Commonly, IDDE projects start with testing of the water coming out at the ends of pipe networks—at outfalls—as they empty into receiving water bodies. This is a technique designed to find evidence of deteriorating pipes and illicit connections.

Outfall testing focuses on the search for pollutants that are reliable markers of wastewater from residential and commercial sources. A variety of indicators are used to detect the presence of wastewater in stormwater including chemical (e.g. caffeine, ammonia, detergents), sensory (e.g. odor), and microbiological indicators (Panasiuk, Hedström, Marsalek, Ashley, & Viklander, 2015).

Bacteria, in particular, *E. coli*, is commonly used as a cost-effective, adequately precise indicator of the presence of wastewater (Irvine, Rossi, Vermette, Bakert, & Kleinfelder, 2011).

Bacteria as Indicator Pollutants

Bacteria tested for in IDDE programs are so-called indicator pollutants. The presence of *E. coli*, common in fecal waste, signals wastewater intrusion and the likely presence of other infectious agents and other contaminants. *E. coli* concentrations are used widely as an indicator of possible or probable contamination by wastewater despite a number of factors that potentially complicate the interpretation of results. These include:

- Temporal variability in the patterns of wastewater production and introduction into the storm system (daily cycles)
- Variability in concentrations of bacteria collected in wet weather depending on when in the course of a storm the sample was collected
- Mortality of bacteria once introduced into the environment
- The documented ability of *E. coli* to reproduce outside host animals in ambient environments (Berthe, Ratajczak, Clermont, Denamur, & Petit, 2013)
- The fact that animal waste washed into storm drains by rainwater can introduce *E. coli* into the water, potentially making the result non-specific to wastewater contamination.

Despite these complications, regulatory schemes still call for measuring *E. coli* at outfalls to detect problems and to measure progress. In the spring of 2016, a new small MS4 permit for the state of Massachusetts was finalized by EPA (EPA, 2015a). One specific requirement of the permit is that municipalities screen all their outfalls for *E. coli* in dry weather and follow up with wet weather sampling at any outfall suspected of any of various so-called "system vulnerability factors." Subsequent stages in remediation

involve tracking the source up pipe networks by examining sequences of manholes; isolating the offending pipes; and ultimately coming up with a plan to replace or repair the breach in infrastructure. The goal is stormwater coming out of all municipal pipes with bacteria concentrations not exceeding the swimming and boating standards for the water bodies they drain to.

The Prioritization Problem

This legal requirement on towns and cities to sample all their stormwater outfalls is a significant burden. In a given municipality there can be dozens, even hundreds, of outfalls. In the relatively small Mystic River watershed, for instance, covering 76 square miles, and including parts of 22 municipalities, there are nearly 2000 stormwater outfalls. In small municipal public works departments, it can be an expensive and time-consuming challenge to screen all outfalls.

The scale of the testing requirement leads to a need to prioritize testing, to direct resources to the most likely sources of pollution. The MS4 permit recognizes this, and dictates grouping outfalls into high and low priority groups. Factors in assigning priorities include past evidence of illicit discharge, proximity to swimming beaches, the age of surrounding developments; documented impairment in receiving water bodies; and other factors. Municipalities are left to assign priorities based on these factors in a loose way; no quantitative method is recommended for assigning priorities.

Other organizations have given thought to the problem of outfall prioritization. The Neponset Stormwater Partnership recommends a GIS (geographic information system) methodology to assign rankings of low to high risk based on a similar set of

considerations listed in the permit, plus additional factors such as the density of pipe networks, number of old buildings in the area, and others (Neponset Stormwater Partnership, 2015). Following the permit language, the Neponset scheme identifies the fact that an outfall empties into an impaired water body as a factor contributing to a high-risk rating. In many highly urbanized areas, where outfalls are clustered on a severely impaired river, for instance, this might lead to most outfalls being classified as high risk.

The Potential of a Data-driven Model

The MS4 permit of these require labor intensive and relatively expensive and slow testing of outfalls. The prioritization schemes offered by the EPA and the Neponset Stormwater Partnership promise to increase the speed with which problems are resolved over the long term, if the factors used to prioritize outfalls truly reflect the underlying risk of contamination. But neither the permit nor the Neponset scheme shows empirical evidence for the weighting of their risk factors. The factors are presented as common-sense rules of thumb.

What if we could test a method for prioritizing by analyzing features of outfalls and the land that drains to them with respect to actual results of bacterial testing?

If we could establish that certain classes of outfalls are associated with higher bacteria values in a real world environment, and we could identify these classes of outfalls by using publicly available geographic information about landscape and infrastructure, then efficiencies in tracking down the worst-offending pipes could be achieved, and prioritization schemes could be put to empirical test and improved.

The Mystic River Watershed Dataset

Over the past 10-15 years, the Mystic River Watershed Association has collected a dataset uniquely positioned to inform such a project.

The Mystic River watershed is a highly urbanized 76 square-mile watershed in eastern Massachusetts. It includes portions of 22 municipalities, including parts of Boston, Cambridge, Somerville, Medford, Everett and other highly densely populated and industrialized communities. The Mystic River itself, like many urban rivers, shows bacterial impairment in wet weather conditions (see Figure 2, above). Many tributary streams show chronic bacterial impairments in both wet and dry weather, implying the presence of wastewater intrusions in the stormwater system throughout the watershed (EPA, 2015b).

Over several years, samplers, working according to EPA and Mass-DEP protocols, have acquired samples from stormwater outfalls in a variety of weather conditions, from a wide variety of outfalls in the Mystic watershed. The complete set of data includes 1251 bacteria measurements at 378 different outfall locations. An outfall dataset of this size and scope is relatively rare. It therefore has the potential of being uniquely valuable.

In addition, the Mystic River Watershed Association (MyRWA) has assembled a body of so-called "catchment" data for municipal stormwater systems. A catchment is the area of land that drains to a pipe network that in turn drains to a particular storm water outfall. There is in general a unique association between an outfall and a catchment. Various characteristics of catchments can be quantified using GIS methodology and then analyzed together with water quality information. Using the Mystic River watershed as a

test case, we can ask whether data indicate properties of catchments that are useful in prioritizing the testing of stormwater outfalls in IDDE programs.

Research Questions, Hypotheses and Specific Aims

The project thus brings together the two distinct data types—water quality and geographic—to ask the following research questions:

Do the historical data from the Mystic River watershed suggest that high bacterial concentrations and loads from stormwater outfalls are associated with quantifiable characteristics of the geographic areas that drain to them?

In other words, are stormwater catchments with certain characteristics (population density, age of building stock, etc.) more likely to be significant sources of bacterial contamination?

At the level of practice, can municipalities use geographic information they probably already have to prioritize which storm water outfalls to investigate? Can this kind of geographic analysis be a useful addition to a municipality's tool kit?

At the level of regulatory policy, would it make sense for EPA to suggest a prioritization of outfall testing based on a set of factors that can be identified in this analysis? Do the results generalize to other urban watersheds? What are the limitations of such an approach?

Hypotheses

My main hypotheses with respect to these research questions include these:

1) Relationships exist between the likelihood and magnitude of bacterial contamination and some geographic features of stormwater catchments but not other features.

a) Outfalls at the end of longer pipe networks (or draining larger areas) will not be more likely to show high bacteria levels, because any increased volume of contaminants introduced will be balanced by increase stormwater flow in those catchments, diluting the impact. I would expect concentrations to be constant across catchments of different size, all things equal.

b) Outfalls in catchments with older housing stocks will be more likely to show high bacteria levels, because more time will have passed on average for neighborhood pipes to degrade and illicit connections to have been made.

c) Catchments with greater population density will be more likely to show high bacteria levels at outfalls.

2) The predicted bacteria load introduced by a stormwater outfall can be usefully modeled as a function of characteristics of the catchment and its infrastructure. With estimates of bacterial concentrations in hand, flow can be roughly estimated and loads computed using geographical variables.

3) As a matter of policy and practice, it makes sense to target first untested stormwater catchments that return the highest load estimates in the model described in hypothesis #2.

Specific Aims

The hypotheses articulated above generate specific research aims and associated techniques of analysis.

Specific aim 1: To explore the relationships between the likelihood of high bacteria values and geographic features of the catchments (see hypothesis 1). In order to explore relationships between bacteria levels and the specified characteristics of catchments (ex. length of pipe, population density, average age of building stock, etc.), regression analyses are called for on these predictor variables in turn. The response variable is the concentration of bacteria at outfalls.

Specific aim 2: To develop a multivariate predictive model for bacteria concentrations at outfalls in the Mystic watershed. Multiple regression analysis can test whether multiple variables improve predictive power, and identify which variables are significant.

Specific aim 3: To develop an appropriate model for bacterial loads in the Mystic watershed, not just concentrations. The contaminant data in the water quality data set represents information on concentrations only, at point locations that can be distant from the source of contamination. The data does not by itself provide an understanding of the total volume of contaminants being introduced to the water bodies, which is arguably the most important question from a management point of view. Using models of concentrations generated in Specific Aim 2 and other estimates, a rough estimate of relative bacteria loads of various catchments can be assembled.

Specific aim 4: To discuss the practical implications of the preceding analysis for policy and practice.

Chapter II

Methods

In order to explore the relationship between the likelihood of high bacteria values at outfalls and the geographic features of the catchments, I used two separate workflows to assemble: 1) water quality data from outfalls and 2) geographic data associated with stormwater outfalls in the Mystic River watershed. Water quality and geographic tables were then joined on a common field (outfall-catchment name) to make possible the regression analyses.

Geographic data were organized and analyzed and exported using ESRI's ArcGIS software (ESRI, 2011). The remaining organization and analysis of data were done principally in R, using the dplyr and ggplot2 packages for data organization and visualization (R Core Team, 2015; Wickham, 2011; Wickham & Francois, 2015).

Water quality data

Water quality data was extracted from the Mystic River Watershed Association (MyRWA) Water Quality Database. The database stores water quality data collected for multiple sampling programs conducted by MyRWA, as well as data shared from other agencies and organizations. The database contains important metadata such as field and lab methods as well as quality assurance information to ensure all data are properly characterized and used appropriately.

To extract the data, I used the myrwaR R package, currently under development by MyRWA. This package contains R functions for loading data from the MyRWA database, merging precipitation data with water quality records, and computing wet/dry conditions. Exported tables were then manipulated and analyzed in R.

Precipitation information was obtained by merging the sampling data with an hourly precipitation dataset obtained for Logan Airport from the Northeast Regional Climate Center and the NOAA Climate Data Online warehouse. Precipitation data is used to characterize each water quality sample as either dry or wet weather. This classification is based on a threshold of 48-hour antecedent precipitation > 0.25 in.

The core set of water quality data used in this analysis comes from the 15-year old "hotspot" sampling program. This program tests for bacteria and other parameters at stormwater outfalls and in streams.

I first assembled a table of results from the hotspot program. Each observation (row) represents a measured parameter (or Characteristic) at a location at a time. Each row includes fields characterizing the sampling event (Datetime, LocationID, VisitID, ProjectID, SampleTypeID); the measurement (CharacteristicID/Name, ResultValue, Units, Qualifier, FlagID); the location (MunicipalityID, WaterBodyID, Latitude, Longitude, LocationTypeID/Name); and the weather (precipation in last 48 hours, Wet/Dry code).

The hotspot program includes data from both streams (LocationTypeID=22) and stormwater sewer outfalls (LocationTypeID= 27). For this research project, I used only outfall data. Of the 1757 recorded outfall locations in the Mystic River watershed, the hotspot program has taken 1251 bacteria samples at 378 locations.

Bacteria Data

Bacteria data used in this analysis reported in units of most probable number (MPN) per unit volume of sample. This is a somewhat indirect measure of concentration, emerging from the statistical mechanisms behind the testing procedures. Bacteria test results are derived from tables of estimated counts, and therefore the possible results from any test are discrete values. It has been argued that for some modeling purposes these numbers should be translated through statistical manipulation into *in situ* concentration estimates (Gronewold, Borsuk, Wolpert, & Reckhow, 2008). But for the purposes of this study MPN values in a sample were taken to be the best estimates of bacteria densities in water, and the MPN value was treated as a continuous variable, despite any limitations of such an approach. Massachusetts public health standards for swimming and boating are defined in terms of results expressed in units of MPN/100 ml.

Three different bacteria parameters are represented in the MyRWA data set: *Escherichia coli (E. coli), Enterococcus,* and fecal coliform bacteria. Each variety is associated with a different set of water quality standards, and result values are not directly comparable. I chose to work with the subset of *E. coli* data only, because it is the largest data set (see Figure 3) and is the most common measure in freshwater samples. *E. coli* values at outfalls, understood as representing estimated concentrations of *E. coli* in the flow emerging from stormwater pipes, comprise the data that will inform the water quality response variable of the regression model.

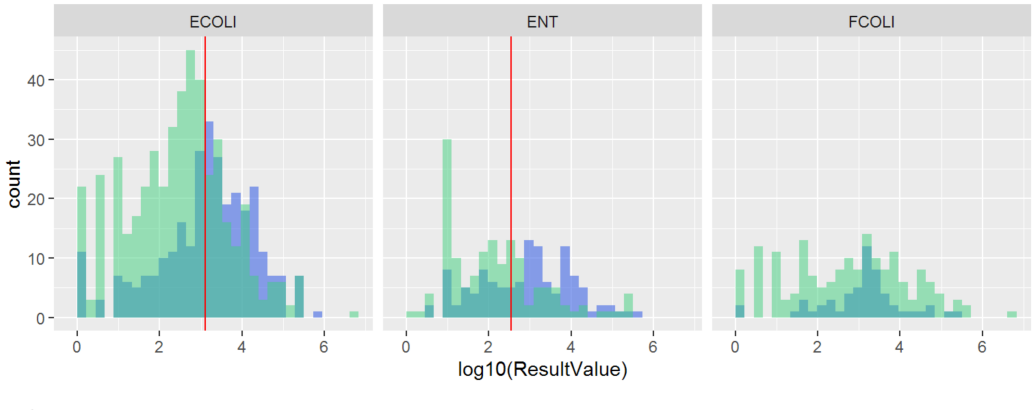

Figure 3. Distribution of bacteria values from Mystic River dataset. Three parameters measured are *E. coli, Enterococcus,* and fecal coliform bacteria. Data approximate log normal distributions for each. Green bars show dry weather data, and blue bars show wet weather data (>0.25 in rain in previous 24 hours). Red line is the regulatory standard for safe boating.

The distributions of values approximate a normal distribution when log-transformed, as is common in environmental data, including microbiological data (Limpert, Stahel, & Abbt, 2001). Taken as a whole, samples in wet weather (for *E. coli* and *Enterococcus*, at least) seem to show higher mean values than dry weather values. I chose to aggregate wet and dry weather *E. coli* data in this study despite this difference for three reasons. First, aggregating increases sample size for bacteria results. Second, aggregating wet and dry data is motivated by features of collection methods and the nature of the research questions. Dry weather data is recognized as especially important by regulators, because evidence of sewage flows in dry weather is especially strong evidence of illicit connections and large breaches of infrastructure (EPA, 2015a). In addition, much "dry" weather data in this dataset was collected in the rain: MyRWA samplers target rain events. Samples collected early in storms will still count as "dry" weather data, despite active stormwater flow in pipes. In the absence of further ability to determine exact conditions at the time of sample collection, I have no way to distinguish

events categorized as dry that were taken in rainy conditions. Finally, chronic dry weather flows are potentially large contributors to load estimates from a given outfall.

Geographic data

The geographic data used in this study consists of three main datasets: 1) data from the GIS systems of local municipalities in the Mystic River watershed, which was shared with the Mystic River Watershed Association (MyRWA); 2) catchment data which was created by MyRWA; 3) public geographic data for the state of Massachusetts, which was downloaded as GIS layers from the Massachusetts Office of Geographic Information (MassGIS).

Municipal Data

The Mystic River Watershed Association acquired, with the cooperation of all municipalities in the Mystic River watershed, data from municipal stormwater systems showing the location of all stormwater outfalls and stormwater mains leading to those outfalls. Names of outfalls were changed to agree with labels in the MyRWA database, where necessary, after careful verification that they referred to the same locations.

Catchment Data

A catchment is defined for the purposes of this study as the area of land generating the stormwater runoff that drains through a pipe network to a particular stormwater outfall. Catchment polygons for the Mystic River watershed have been created in ArcGIS in the past few years by MyRWA staff and interns.

Catchments polygons were drawn in ArcGIS in this way:

- Pipe networks leading to named outfalls in municipal GIS systems were identified as defining the core extent of catchments. Each outfall and each catchment is associated with exactly one network of stormwater pipes.
- Exact boundaries of catchment polygons were drawn taking into account topographical information from imported layers in ArcGIS. Estimates were made about direction of flow from paved surfaces leading to stormwater mains, and each municipality was divided into a mainly continuous patchwork of catchments (see Figure 4).

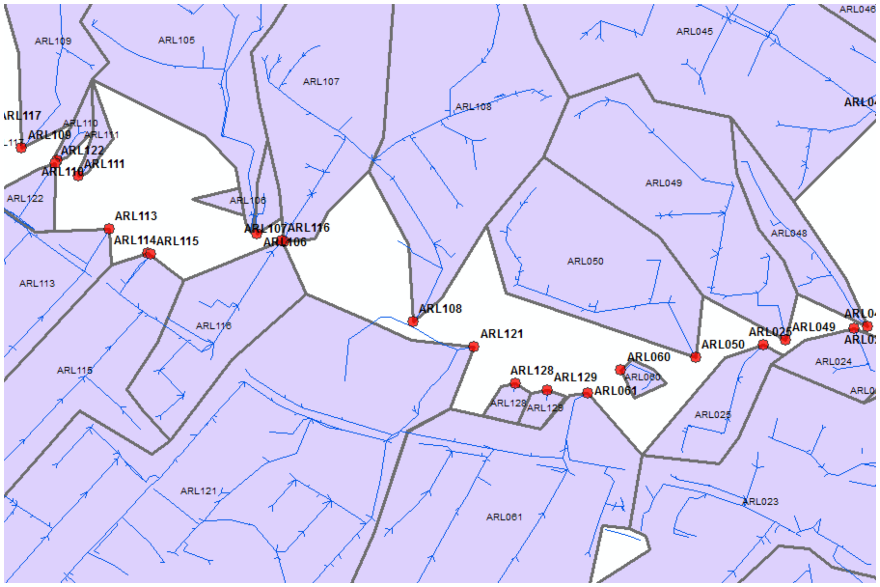

Figure 4. Catchments, pipe networks, and outfalls. Catchments in Arlington, MA, emptying into Mill Brook, not depicted. Each catchment, outlining a single pipe network, is given the same name as the outfall associated with it. Note variation in size.

I supervised and participated in a process of vetting and editing catchment data for five municipalities. Stephanie Clark and I reviewed provisional drawn catchments, corrected mistakes, and met with municipal officials to confirm ambiguities and

questions about pipe networks and labels. A small number of catchments whose pipe network information could not be verified were removed from the data set. A small number of other networks were redrawn, based on further information from municipalities. The result was a set of catchments from five municipalities that empty their stormwater system completely or partially into the Mystic and Malden Rivers for which we have catchment information in which we are relatively confident (see Figure 5). (Some of these municipalities lie only partially in the Mystic River watershed, and thus are only partially covered.)

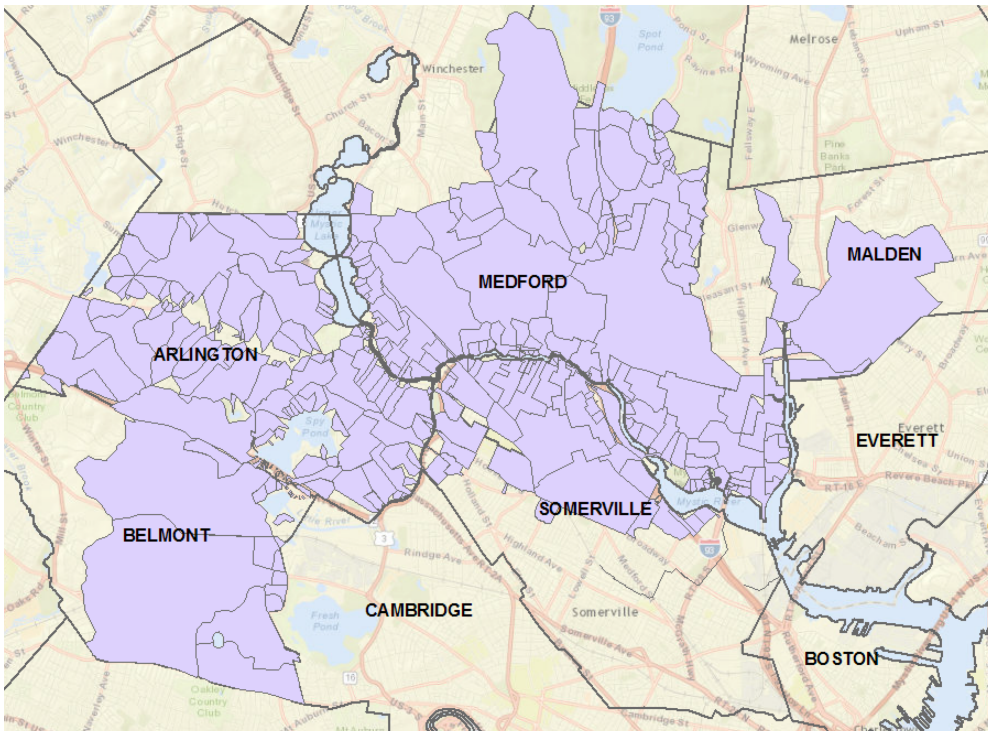

Figure 5. Map of study area. Catchments (purple) in Medford, Arlington, Somerville, Malden, and Belmont, MA.

These five municipalities—Medford, Arlington, Somerville, Malden, and Belmont—are the source of a large proportion of outfall samples in the complete dataset.

47% of the bacteria samples from outfalls in the MyRWA database come from these five towns and cities. 37% of tested outfalls are in these five municipalities.

Pipe network length and area per catchment were calculated in ArcGIS after catchment shapes were finalized.

MassGIS data

In order to create additional geographical variables associated with catchments, I downloaded several data layers from the Massachusetts Office of Geographic Information (MassGIS), the state repository of public GIS information (Table 1). These layers pertain to the population and features of the built environment.

Table 1. GIS layers from MassGIS.

Layer	**Source**	**Resolution**	**Projection**	**Type**
Impervious surface	MassGIS	1 m	Massachusetts State Plane.	Raster
Census, 2010, blocks	MassGIS	--	As above.	Polygon
Building structures, 2D (roofprints)	MassGIS	--	As above.	Polygon
Level 3 Assessors' parcels (property data)	MassGIS	--	As above.	Polygon

Overlay Analysis

Next, I used overlay analysis in ArcGIS on catchment data and other geographic data to calculate various quantities for each catchment that could serve as predictor variables in regression analysis. Here I summarize the process that generated geographic variables.

Impervious surface raster data was used to calculate the percent impervious surface (roofs and pavement) in each catchment. The impervious raster has values of 1 or 0, representing impervious and pervious surface, respectively. The Zonal Statistics as Table tool calculates mean value per catchment, which corresponds to proportion impervious. Multiplying proportion impervious by area also gave impervious area as a variable for each catchment.

Census block information was used to calculate population and population density per catchment. I intersected the two polygon layers, assigning a percentage of the census block's population to each new subdivided polygon based on the percentage area of the original census block it represented. Aggregating the subdivided polygons then by catchment yielded (estimated) populations for each catchment. Dividing by catchment area generated a population density per catchment.

Intersecting assessors' parcel data with the building structures layer yields a layer in which each building is associated with parcel data, including building construction date. I intersected this layer in turn with the catchment layer, and exported the resulting attribute table. In R, I was able to aggregate rows by catchment and, with some simple calculations, create columns for number of buildings per catchment, number of old buildings (50+ years old) per catchment, and average age of building per catchment.

Combined datasets

The next step in the analysis was to join two data tables:

- Water quality table: I began with the raw data table in which each of the *E. coli* bacteria samples is a row with many values, including outfall name, municipality,

location type (outfall vs. stream), latitude/longitude, weather (wet/dry, based on quantity of rain in past 48 hours), bacteria concentration, ammonia concentration, and detergent concentration. The variety of fields allowed flexible filtering of subsets of data in the course of the analysis.

In R, I created a table creating for each catchment a column for geometric mean of bacteria concentrations of samples taken, and a column for n, number of *E. coli* measurements in the dataset from that catchment. Geometric mean is understood to better represent conditions than the mean in data with log normal distributions such as this; rare very high values artificially draw the mean high. The result is a water quality table in which each tested catchment now has columns for geometric mean and sample size. I then created separate columns (labeled N01, N02, etc.) which contained geometric mean values only for those catchments for which n>=1, n>=2, etc. In this way, I could define response variables corresponding to subsets of data of different minimum sample sizes. Seventy-six catchments had at least one sample; 53 catchments had at least two samples; and 36 catchments had at least 3 samples.

- Catchment table: Exported from ArcGIS and imported into R, this table included catchment name, area, pipe length contained, number of old buildings, number of buildings, average age of building, population, population density, percent impervious surface, and area impervious.

Joining these two tables on the common catchment/outfall name field, allowed analysis of patterns in bacteria data at outfalls in the context of geographic properties of the areas of land that drain to those outfalls, as described in the next chapter.

Chapter III

Results

Relationships among the assembled bacteria data and geographic predictor variables were analyzed using the tools of linear regression. To explore the relationships between response and candidate predictor variables, I first calculated correlations among all variables, including the subsets of geometric mean of bacteria data that represent outfalls for which I have at least one, at least two, or at least three samples (variables N01-N03) (Table 2).

Table 2. Correlation matrix for untransformed variables.

	N01	**N02**	**N03**	**Length_Pipes**	**Shape_Area**	**Pop_total**	**imp_area**	**count_bldg**	**avg_age**	**count_old**
N01										
N02	1.00***									
N03	1.00***	1.00***								
Length_Pipes	0.06*	0.06	0.07							
Shape_Area	0.03	0.03	0.03	0.90***						
Pop_total	0.06*	0.08*	0.09	0.61***	0.76***					
imp_area	0.05	0.06	0.07	0.81***	0.92***	0.93***				
count_bldg	0.06*	0.08	0.09	0.79***	0.88***	0.94***	0.97***			
avg_age	0.00	0.00	0.04	0.01	0.01	0.04	0.02	0.04		
count_old	0.06*	0.09*	0.10	0.75***	0.84***	0.93***	0.95***	0.99***	0.05*	
Pop_density	0.01	0.01	0.04	0.01	0.01	0.01	0.00	0.00	0.28***	0.00

Non-adjusted R-squared values for each pairing of variables. Stars correspond to the following significance levels: * = $p<0.05$; ** = $p<.01$; *** = $p<.001$.

Predictors of Bacterial Concentrations

A few predictor variables show significant weak but positive correlations with bacteria level (N01) in this table: pipe length, total population count of buildings, and count of old buildings (all have R-squared=0.06, $p<0.05$), suggesting that a linear model might be appropriate.

The case of count of old buildings is typical. Linear regression suggests a significant if weak relationship ($p = 0.04$, R-squared = 0.06) (see Table 3).

Table 3. Linear regression results for count of buildings as predictor of geometric mean of bacteria concentration.

	Estimate	Std. Error	t value	Pr(>|t|)
count_bldg	0.7017	0.341	2.058	0.04327
(Intercept)	1421	396.3	3.585	0.000614

Observations	Residual Std. Error	R^2	Adjusted R^2
73	2991	0.05629	0.043

Diagnostic tests of residuals, however, indicate that assumptions for regression are not met, and a linear model is not appropriate (Figure 6). The residuals vs. fitted graph shows that the spread of residuals increases as predicted values increase-that is, that the data are heteroscedastic (Mendenhall & Sincich, 2012). Linear regression assumes a uniform distribution of residuals (or errors) across the range of predicted values. Moreover, the strongly curved normal Q-Q graph further indicates that the residuals are not normally distributed, violating the assumption that the errors around the regression line follow a normal model,

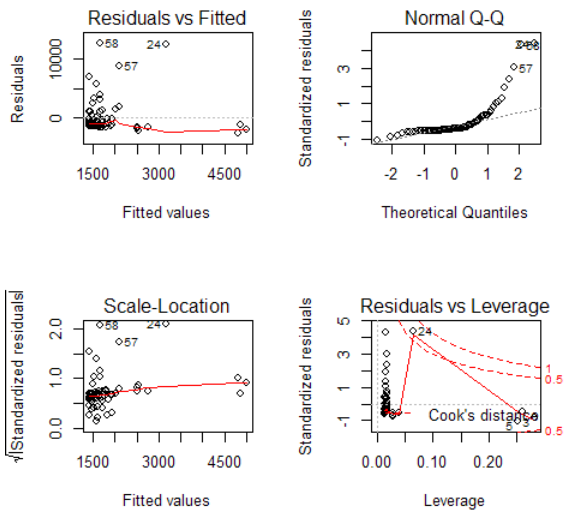

Figure 6. Diagnostic residual plots for linear regression of count of buildings as predictor of geometric mean of bacterial concentration.

Log-transformed Variables

In order to make the data more nearly satisfy the assumptions of linear regression, I transformed both x-values and y-values (Mendenhall & Sincich, 2012).

When log-transformed, both geometric mean and count of buildings show near normal distributions, indicating the kind of skewed distributions for which log-transforms are often appropriate (Figures 7 and 8).

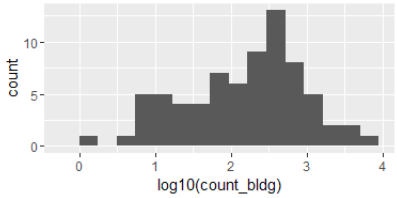

Figure 7. Distribution of count of buildings values.

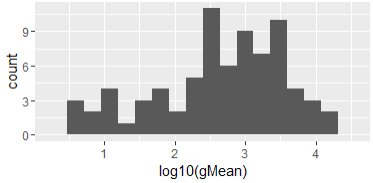

Figure 8. Distribution of *E. coli* geometric mean values.

I calculated correlations among log-transformed response variables and transformed predictor variables as appropriate. The transformed response variables (N01-N03) now all show significant correlations with many more variables (Table 4).

Table 4. Correlation matrix for transformed variables.

	logN01	logN02	logN03	logArea	logPipes	Log CountBldg	Log Imp Area	arcsinPerc Imp	avg_age	count_old	arcsinPerc Old	Log Pop
logN01												
logN02	1.00***											
logN03	1.00***	1.00***										
Log Area	0.23***	0.16**	0.19**									
Log Pipes	0.25***	0.14**	0.20**	0.90***								
Log CountBldg	0.18***	0.06	0.21**	0.79***	0.76***							
Log Imp Area	0.25***	0.19**	0.24**	0.97***	0.88***	0.76***						
arcsinPerc Imp	0.01	0.02	0.06	0.02	0.01	0.05	0.00					
avg_age	0.00	0.00	0.02	0.03	0.03	0.25***	0.03	0.01				
count_old	0.15***	0.11*	0.19*	0.71***	0.65***	0.75***	0.69***	0.02	0.15***			
arcsinPerc Old	0.01	0.01	0.07	0.05*	0.06*	0.26***	0.05	0.01	0.75***	0.14***		
Log Pop	0.19***	0.07	0.11*	0.70***	0.63***	0.87***	0.69***	0.01	0.24***	0.59***	0.24***	
Pop_density	0.03	0.01	0.02	0.01	0.01	0.10**	0.02	0.12**	0.28***	0.03	0.27***	0.25***

Non-adjusted R-squared values for each pairing of variables. Stars correspond to the following significance levels: * = $p<0.05$; ** = $p<.01$; *** = $p<.001$.

Among the predictor variables, log-transformed measures of pipe length, area, impervious area, population, and count of buildings all showed higher R-squared values and lower p-values in their correlations with log bacteria values than when

untransformed. (Non-transformed predictor values are not shown in Table 3.2.) I used the arcsine of the square root transformation for the variables expressed as percentages.

Scatterplots of the predictor variables that show the strongest correlations display relatively linear patterns, with substantial scatter around the line of best fit (see Figure 9, for one example).

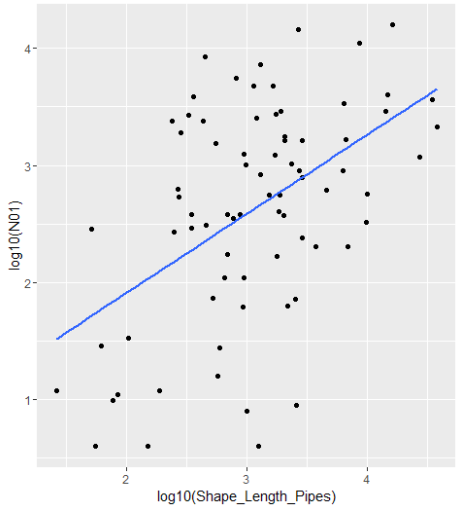

Figure 9. Scatterplot of log of pipe length vs. log of geometric mean of bacteria values. For outfalls with samples n>=1, with line of best fit.

Linear regressions of the transformed response variable against predictor variables now show diagnostic residual plots that suggest that the assumptions for regression are met. This is true for variables that showed relatively strong linear relationships and those that did not (see Figures 10 and 11).

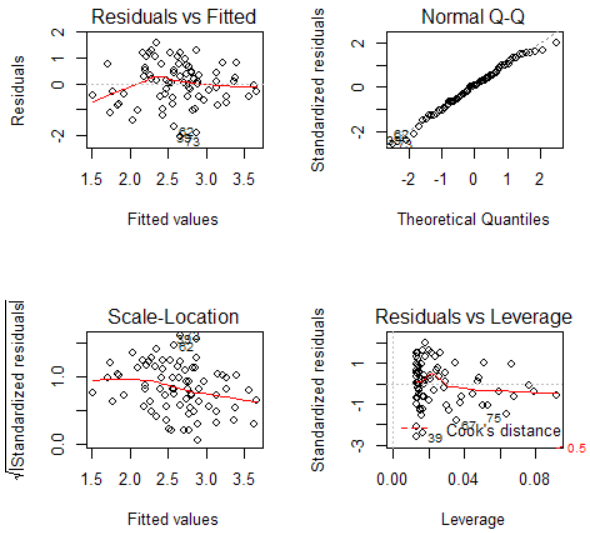

Figure 10. Diagnostic residual plots for linear regression of log of pipe length vs. log of geometric mean of bacteria level. Outfalls with samples n>=1. ($p<0.001$, R-sq = .25).

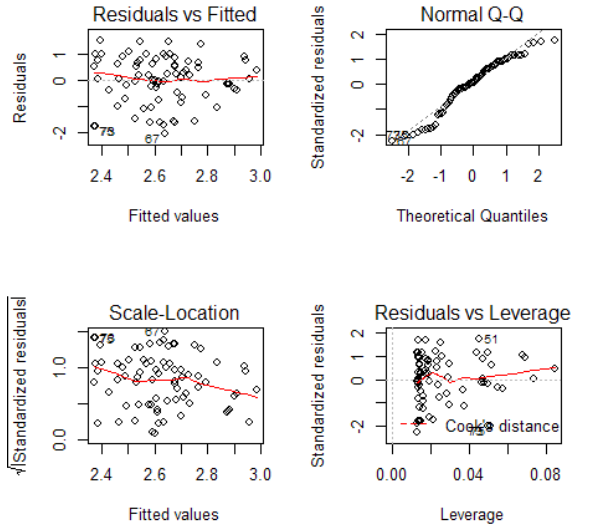

Figure 11. Diagnostic residual plots for linear regression of population density against log of geometric mean of bacteria. Outfalls with samples n>=1 ($p=0.15$, R-squared = .03).

In summary, simple linear regression analysis suggests significant relationships between bacteria geometric means and several variables: log of total area, log of impervious area, log of total population, log of number of buildings, log of pipe length and log of number of old buildings.

A Multiple Regression Model for Bacterial Concentrations at Outfalls

Having determined significant linear relationships between the response variable (limited sample size n>=1 in the discussion below) and several individual predictor variables, I explored whether incorporating more predictor variables into a multiple linear regression could account for more of the variability in bacteria data.

Evidence of multicollinearity

First, it is necessary to note that many of the predictor variables a show high degree of correlation with each other (>0.75 R-squared), or multicollinearity (Table 4). Best practice dictates that predictor variables in multiple regression models should not be correlated with one another because of the difficulty of teasing out the unique contribution of each variable and other difficulties in interpretation (Gotelli & Ellison, 2013).

Faced with a dataset exhibiting multicollinearity, step-wise regression can be used to screen out variables contributing redundant information (Mendenhall and Sincich, 2012).

Several factors that are related to the physical shape of the infrastructure (pipe length, area, impervious area, number of buildings) showed evidence of multicollinearity. I screened these factors alone by running a stepwise regression in R. The AIC stepwise model selected only pipe length alone as a final product, suggesting that the other variables do not make significant contributions to a linear model once pipe length is included (Gotelli & Ellison, 2013). So I proceeded in using log of pipe length only from among this set of variables.

Final Multiple Regression Model

Next, I ran a similar stepwise regression on the model:

$$\text{Geometric\_mean} = b1 \times \log10(\text{pipe\_length}) + b2 \times \text{Pop\_density} + b3 \times \log10(\text{tot\_pop}) + b4 \times \text{avg\_age} + b5 \times \text{count\_old}$$

Stepwise regression yields a final model ($p<0.001$, adj. R-squared =0.29) of:

$$\text{Geometric mean bacteria} = 0.7131(\log10(\text{pipe\_length})) + 0.0011(\text{Pop\_density}) - 0.013(\text{avg\_age})$$

The coefficient of each term contributes significantly to the model (Table 5).

This improves on regression results from a regression on pipe length vs. geometric mean alone ($p<0.001$, adj R-squared = 0.24):

$$\text{Geom\_mean bacteria} = 0.674(\text{pipe\_length})$$

The multiple regression model suggests that 29% of the variation in bacteria values at outfalls can be explained by the relationship between bacteria levels and the three variables--pipe length, population density, and average age of building in a catchment (Table 5).

Table 5. Multiple regression output in R for final model.

```
##
## Call:
## lm(formula = log10(N01) ~ log10(Shape_Length_Pipes) + Pop_density +
##     avg_age, data = (how_fit))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.63488 -0.40974 -0.04086  0.53515  1.51439
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                1.0834869  0.5943284   1.823   0.0726 .
## log10(Shape_Length_Pipes)  0.7131906  0.1366487   5.219  1.8e-06 ***
## Pop_density                0.0011202  0.0004404   2.544   0.0132 *
## avg_age                   -0.0131466  0.0065017  -2.022   0.0471 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7702 on 69 degrees of freedom
##   (3 observations deleted due to missingness)
## Multiple R-squared:  0.3187, Adjusted R-squared:  0.289
## F-statistic: 10.76 on 3 and 69 DF,  p-value: 6.942e-06
```

Variation inflation factor (VIF) results for the model give low values for each coefficient, suggesting no issues of multicollinearity (Mendenhall and Sincich, 2012) (Table 6).

Table 6. Variation inflation factors (VIF) for each coefficient.

```
## log10(Shape_Length_Pipes)          Pop_density
##                  1.033293             1.384535
##                   avg_age
##                  1.423440
```

Note that all are close to 1.0, and much less than 10, indicating an absence of evidence of multicollinearity, output in R.

Diagnostic residual plots for the final model suggest that assumptions for regression are met (Figure 12).

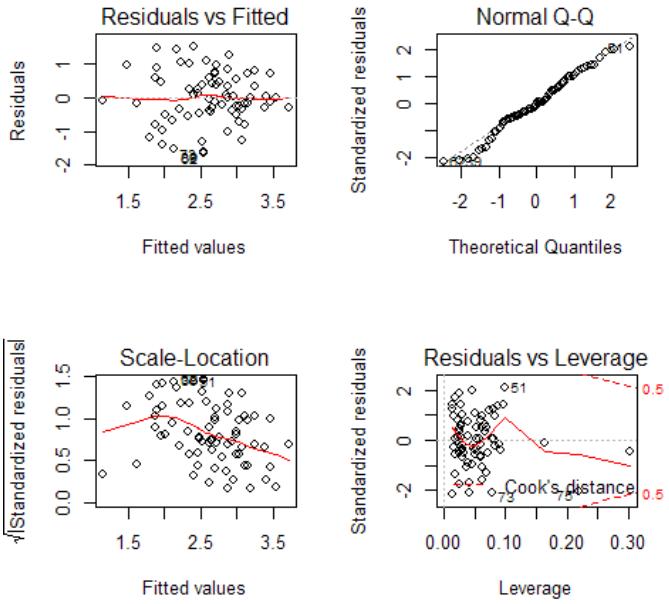

Figure 12. Diagnostic residual plots for final multiple regression model.

Chapter IV

Discussion

The results of the multiple regression analysis allow a review of my initial hypotheses, some of which were confirmed and others not confirmed by the data. With a predictive model of bacteria concentrations in place, I next estimate not only bacterial concentrations at outfalls, but also bacterial loads by catchment. Finally, I discuss limitations and caveats and suggest ways that this model and approach could be improved.

Interpretation of Multiple Regression Model

The multiple regression model included terms representing pipe-length, population, and age of building, representing factors that correspond to my three initial quantitative hypotheses.

Pipe-length Hypothesis

Hypothesis 1(a) was the hypothesis that there is no relationship between pipe length and bacteria concentrations. The regression analysis suggests that we can reject this hypothesis. Large pipe networks are significantly associated with higher concentrations of bacteria contamination. In fact, the relationship between pipe length and bacteria level ($p = <0.001$, R-squared = .25) was the strongest of any predictor variable. This surprising result demands analysis.

The original hypothesis was based on a key assumption: that the probability of contamination per unit length of pipe is the same in large pipe networks as it is in smaller pipe networks. If that assumption were true, one would expect the concentrations of bacteria in large networks to be similar to those in smaller networks. Increased bacteria inputs in larger networks would be balanced by increase dilution, because large pipe networks will carry larger volumes of incoming rainwater in a rain event.

One explanation for the failure of the hypothesis may be that this assumption was wrong. It may be the case that the probability of contamination per unit length of pipe in long networks is not the same as in shorter pipe networks.

There are several mechanisms that might account for this:

1. Larger pipe networks may be older. Core areas of cities had their pipes laid in a short period of in single networks. Larger networks may contain higher proportions of old and decrepit pipe. In other words, there may be a lurking variable--age or condition of pipes—to which we do not have direct access that explains the association with pipe length.

2. Larger pipe networks may be built in the most important drainages in a city, low-lying areas that receive the greatest inundations in wet weather. Evidence for this is that some of the largest catchments are buried streams, converted into culverts. This in turn would result in these networks being subject to the greatest risk of cross-talk contamination between leaking sanitary sewer pipes and the stormwater network.

3. Large pipe networks might be more likely to harbor reservoirs of viable bacteria during dry weather and between storms, resulting in higher measured levels

when it rains. There is some evidence that *E. coli* populations can subsist in pipes for some time.

4. Finally, the pipe networks mapped are primarily gravity mains. Left out of the calculation is the number of lateral line inputs. Larger pipe networks probably represent wider diameter pipes with more inputs per unit length, with a corresponding increase in the probability of sewage inputs.

While there are no very large pipe networks with very low geometric means, there are some small catchments with relatively high geometric means; there is great variability in contaminant values for small catchment sizes. Perhaps the variability we see among smaller catchments can be explained by similar factors: older smaller pipe networks may be those with systematically higher bacteria levels, for instance.

Interpretation of Multicollinearity

The presence of mulitcollinearity among variables reflects real world dynamics. Pipe length is highly correlated with other variables: area, impervious area, and number of buildings. These associations all make sense in the context of this urbanized watershed. Catchment shapes are drawn around pipe networks, so larger networks will mainly define larger catchment areas; some variation will emerge from differential density of pipes in a given catchment. Storm water pipe networks follow streets, and homes and buildings are built close to streets, so it makes sense that pipe length and percent impervious surface are strongly correlated. Number of buildings in this highly urbanized set of catchments will also track closely the amount impervious area.

In choosing variables for the multiple regression, I eliminated these correlated variables in favor of pipe length on the basis of step-wise regression screening. But these other correlated variables urge caution in interpreting the results. Substituting log of impervious area for log of pipe length, for instance, results in a model with nearly the same fit.

I argue that the choice of pipe length is justified by the nature of the question under investigation: it most directly captures a key physical feature of the stormwater system whose contamination is under investigation. But it is important to keep in mind that other variables show much the same pattern in evaluating proposed mechanisms that might explain patterns in the data.

Hypotheses about Other Predictor Variables

The multiple regression model suggests that two other factors might, considered together, explain some of the variability in bacteria not explained by pipe length alone: Population density and average age of building in the catchment.

Among catchments of a given pipe length and average building age, it is reasonable to expect that more densely populated catchments would have higher bacteria concentrations at outfalls. The total amount of raw sewage contamination from residential units, for instance, might be expected to increase with population density. More people generate more waste. Areas traversed by long pipe networks that contain fewer people might simply have fewer sewage inputs (they might be in commercial areas, etc.). Population density in this way serves as a measure of intensity of residential land use in a way that pipe-length itself may not completely capture. It is therefore reasonable to find

that this factor makes a contribution to bacteria levels in stormwater in the context of this model.

Note that this result is consistent with population density not being a significant predictor alone of bacteria concentrations. Multiple mechanisms might account for this divergence. In any case, the model suggests among catchments with a given length of pipe network and age of building stock, those with higher population densities will tend to have higher bacteria concentrations.

Among catchments with a given pipe length and population density, the model predicts that catchments with younger buildings on average would tend to have higher bacteria levels. (The coefficient for log average age of building is negative.)

This result is unexpected. If the age of buildings were an effective proxy for the condition of pipes, one would expect catchments with older buildings to have higher mean bacteria levels. But there was no significant relationship between average age of buildings and geometric mean of *E. coli*, when considered alone in a simple regression model (see Table 4, correlation matrix). So this measure may not be capturing the intended information about the conditions of pipes.

In the context of a multiple regression with pipe length and population density, average age of building in a catchment becomes a significant variable, but with a negative coefficient. There are multiple possible explanations for this unexpected result:

1. Lurking variables. Average age of building may be associated with some other unidentified variable that would explain the negative coefficient.

2. Hidden multicollinearity. If pipe length is indeed associated with pipe age, as discussed above, the average age of building may contribute redundant information, making interpretation of coefficients difficult.

It is worth noting that a simplified model that drops the variable average age and includes only pipe length and population density outperforms a model (higher adjusted R-squared) with pipe length alone, although the population density term is itself not significant ($p=0.2$) in the context of the multiple regression model.

It must also be noted that for the subsets of bacteria data for which there are two or more samples (variables N02 and N03), terms in addition to pipe length were not significant in the context of the multiple regression models I tested. It may be the case that if there were more data points representing two or more samples, additional terms would become significant. But caution is warranted in putting emphasis on the additional terms. On the other hand, the correlations between bacteria levels and pipe length or impervious area are significant and nearly as strong for the subsets of data where $n>=2$ and $n>=3$ as for the full dataset where $n>=1$ (see Table 4 above). The result that longer pipe networks are associated with higher bacteria concentrations is robust across these different response variables.

Utility of the Model in Prioritizing Pipe Sampling

The original problem this project hoped to address was whether one could usefully narrow the search for contaminated pipes in an urbanized watershed by identifying characteristics of catchments that could be calculated before going to the field to collect water quality samples.

The results here suggest some recommendations for those tasked with sampling outfalls in urban areas:

- Map out catchment shapes.
- Prioritize the catchments with the largest pipe networks first.
- Do additional GIS work to calculate population density and building age, if possible.
- Apply the multivariate model to prioritize the catchments that the model predicts will have the highest concentration values.

The finding that larger pipe networks are associated with higher bacteria values in the Mystic River watershed dataset is a lesson for other stormwater managers and regulators in other urban areas. As noted, the current draft of the Massachusetts MS4 Permit requires testing at all outfalls. This study provides empirical evidence that a prioritization based on the factors in the model (even just pipe length) might help speed the search for sewage infiltrations and illicit connections.

Of course, the longest pipe networks will also be those in which the detective work will be most difficult. The largest catchments in this dataset contain many miles of pipe terminating at one outfall. Finding the precise locations of infrastructure failure will be difficult in long networks. But it is an important result of this study that there is some empirical evidence that this should be the first place to start looking.

Estimating Loads from Catchments

Ultimately the goal of bacteria tracing in pipe networks is not simply to detect which pipes have the highest concentrations of bacteria, but to reduce bacteria loads to

the water bodies the pipes pollute. Very small volumes of high concentration runoff may not be a major concern when diluted by a river's flow. Large volumes of more modest concentrations can severely impact water quality in a lake or stream.

We can extend the analysis here to create order of magnitude estimates of relative bacteria load contributions of various catchments.

Bacteria loads can be roughly modeled as:

Total bacteria load from a catchment to water body =

(average bacteria concentration) x (volume of flow from outfall)

We cannot know the true average concentration of bacteria in a given pipe. But the final multiple regression model in this study provides an estimate of average bacteria concentration based on other factors.

Similarly, we do not have data on the true volume of flow in a catchment pipe. But this volume can be roughly estimated by remembering that the vast majority of water in stormwater pipes in storms runs off of impervious surfaces (roofs and pavement) in a catchment. The volume of flow from an outfall in a rain event, therefore, will be roughly proportional to the impervious area in a catchment, at least as a first approximation. So we can use impervious area as a proxy for volume of flow.

Therefore load from a catchment in a rain event can be roughly modeled as:

Load Index = (model equation) x impervious area

Or:

Load Index = [0.7131(log10(pipe_length)) + 0.0011(Pop_density) - 0.013(avg_age)] x impervious_area

This equation yields a unitless load index that can assign to catchments estimated relative bacteria loads.

By mapping this index value onto a set of catchments, we can now visualize which catchments on this model are introducing the greatest bacteria loads in an area. Applied to the original set of catchment data in this study, the results can be displayed in a map (Figure 13).

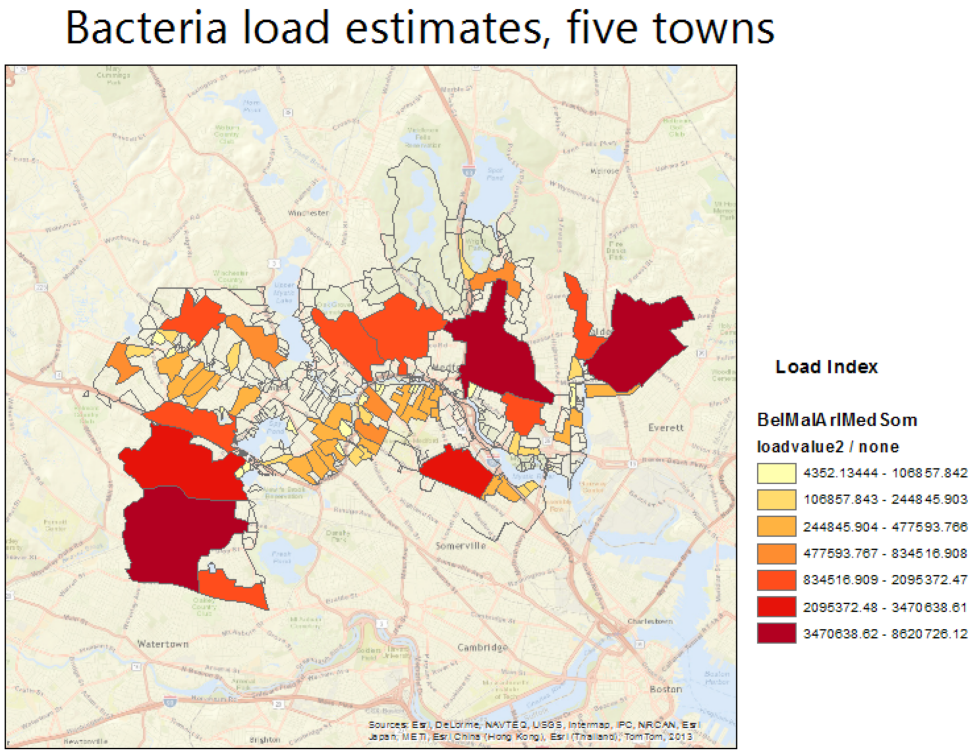

Figure 13.Map of bacteria load estimates. Map showing relative bacteria loads by catchment as predicted by the final model and the load index model above. Darker colors indicate higher bacteria loads.

Visualizing the load index value for a set of catchments in the form of a tree map allows us to see the relative contributions of each catchment to the total estimated load in an area. When applied to the study area in our data set, a striking pattern is that the majority of estimated bacterial load is contributed by a relatively few catchments (Figure 14).

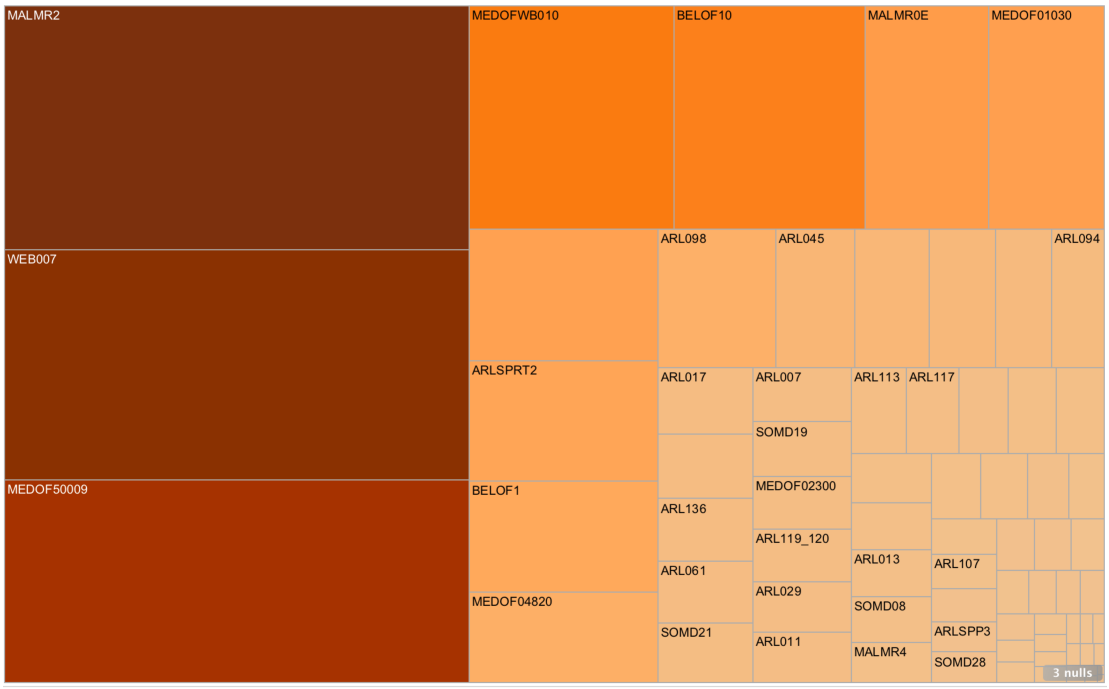

Figure 14. Tree map of relative bacteria load by catchment. The majority of load is contributed by a relatively small number of catchments.

At least for the Mystic River watershed, this analysis suggests that if managers could reduce bacteria contamination in a few catchments to zero, the majority of bacterial inputs to water bodies would be eliminated. As noted above, this is a significant but daunting result. These few catchments tend to be precisely the largest. Tracking sources of bacterial contamination in large catchments is difficult. But Figure 14 suggests in a striking way why this should be a high priority of managers.

A similar analysis could be done in any watershed or municipality by managers armed with GIS data about catchments, pipe length, population, impervious surface and building age only. Most of this data is publicly available in most areas. Maps and tree maps like those in Figures 13 and 14 can communicate priorities to communities and identify the areas in which work should likely be prioritized.

Research Limitations and Caveats

The results and conclusions presented here are subject to a number of caveats and limitations:

- The coefficient of determination of the final model is relatively low. The prediction interval will be wide. The result is at best a loose predictive ability. The small sample sizes for many of the catchments in the analysis certainly contribute to the low coefficients of determination. Presumably more samples would reduce estimation errors for the geometric mean for the tested catchments in this study that have relatively few samples. More samples would give greater confidence that the geometric mean represents conditions well.
- Attempts were made to capture accurate geographic and water quality data. But the conclusions can be only as accurate as the data. If pipe network data or catchment delineation are incomplete or inaccurate, results of the analysis would be affected.
- Similar caveats apply to application of this approach in other watersheds. Creating catchment data demands much careful vetting of geographic information and

careful delineation of catchment shapes. This sometimes requires intensive collaboration with municipal officials and on-site ground truthing.

- The study only covered five cities with particular characters. Results should be generalized only to similarly urbanized environments.

Questions for Further Research

The final regression model included terms that represent the geometry of the infrastructure (area, length, impervious area); intensity of land use and human impact (population density); and a variable meant to reflect the condition of the infrastructure (average age of building).

One feature of this analysis in this thesis is that the variables used to capture the physical condition of pipes is necessarily a proxy variable—average age of building, number of old buildings in a catchment, etc. What we really want to know is something about the physical condition of pipes—number of cracks per kilometer, amount of cross talk with sanitary sewers, whether the pipe is sound, etc. Because I did not have access to this kind of direct physical data, I invoke an assumption that older neighborhoods have older pipes that are more likely to be in bad physical condition, and more likely to leak. Age of buildings is a loose proxy for physical condition of pipes.

But old developments can have new, replaced stormwater mains. Nothing guarantees that age of building will track the condition of stormwater infrastructure.

A more direct proxy for physical condition of pipes would be the age of pipe in a catchment. Old pipe could be reasonably expected to be a more likely source of contamination. A pipe-age variable if available would be a stronger proxy for physical

condition than building age. As discussed, it may be that the first variable (pipe length in the model) is actually capturing the influence of a lurking variable, namely age of pipe network.

The municipal data I had access to did not include age of pipe or other direct measure of pipe condition, so I turned to relatively weak proxies of those variables. Because of the likelihood of bacteria contamination is directly causally linked to the physical integrity of pipes, I would predict that age of pipes or other variables (perhaps categorical) directly reflecting the condition or repair history of pipes would be positively associated with bacteria levels at outfalls.

I would expect that the explanatory power of the model would be improved if I were able to gain access to data more directly related to the condition of pipes. Municipalities with data related to the age or condition of pipes themselves would do well to prioritize those pipe networks for which there is active evidence of old or decaying infrastructure.

References

Barber, L. B., Brown, G. K., Nettesheim, T. G., Murphy, E. W., Bartell, S. E., & Schoenfuss, H. L. (2011). Effects of biologically-active chemical mixtures on fish in a wastewater-impacted urban stream. *Science of The Total Environment*, *409*(22), 4720–4728. http://doi.org/10.1016/j.scitotenv.2011.06.039

Bernhardt, E. S., & Palmer, M. A. (2007). Restoring streams in an urbanizing world. *Freshwater Biology*, *52*(4), 738–751. http://doi.org/10.1111/j.1365-2427.2006.01718.x

Berthe, T., Ratajczak, M., Clermont, O., Denamur, E., & Petit, F. (2013). Evidence for coexistence of distinct Escherichia coli populations in various aquatic environments and their survival in estuary water. *Applied and Environmental Microbiology*, *79*(15), 4684–4693. http://doi.org/10.1128/AEM.00698-13

Booth, D. B., Hartley, D., & Jackson, R. (2002). Forest cover, impervious-surface area, and the mitigation of stormwater impacts. *Journal of the American Water Resources Association*, *38*(3), 835–845. http://doi.org/10.1111/j.1752-1688.2002.tb01000.x

Brown, E., Caraco, D., & Pitt, R. (2004). *Illicit discharge detection and elimination: a guidance manual for program development and technical assessments*. Water Permits Division, Office of Water and Wastewater, US Environmental Protection Agency.

ESRI. (2011). *ArcGIS Desktop: Release 10*. Redlands, CA: Environmental Systems Research Institute.

Gotelli, N. J., & Ellison, A. M. (2013). *A primer of ecological statistics* (2nd ed.). Sunderland, Massachusetts: Sinauer Associates, Inc., Publishers.

Gronewold, A. D., Borsuk, M. E., Wolpert, R. L., & Reckhow, K. H. (2008). An assessment of fecal indicator bacteria-based water quality standards. *Environmental Science & Technology*, *42*(13), 4676–4682. http://doi.org/10.1021/es703144k

Irvine, K., Rossi, M. C., Vermette, S., Bakert, J., & Kleinfelder, K. (2011). Illicit discharge detection and elimination: Low cost options for source identification and trackdown in stormwater systems. *Urban Water Journal*, *8*(6), 379–395. http://doi.org/10.1080/1573062X.2011.630095

Klein, R. D. (1979). Urbanization and stream quality impairment. *Journal of the American Water Resources Association*, *15*(4), 948–963. http://doi.org/10.1111/j.1752-1688.1979.tb01074.x

Konrad, C. P., & Booth, D. B. (2005). Hydrologic changes in urban streams and their ecological significance. In *American Fisheries Society Symposium* (Vol. 47, pp. 157–177). Retrieved from http://water.usgs.gov/nawqa/urban/pdf/157-178_Konrad.pdf

Limpert, E., Stahel, W. A., & Abbt, M. (2001). Log-normal distributions across the sciences: Keys and clues. *BioScience*, *51*(5), 341–352. http://doi.org/10.1641/0006-3568(2001)051[0341:LNDATS]2.0.CO;2

Massachusetts Department of Environmental Protection. (2012). Massachusetts Stormwater Handbook. Retrieved December 6, 2015, from http://www.mass.gov/eea/agencies/massdep/water/regulations/massachusetts-stormwater-handbook.html

National Research Council. (2009). *Urban stormwater management in the United States*. National Academies Press.

Neponset Stormwater Partnership. (2015). Neponset River watershed illicit discharge detection and elimination plan, draft. Retrieved from http://neponsetstormwater.org/member-resources/

Panasiuk, O., Hedström, A., Marsalek, J., Ashley, R. M., & Viklander, M. (2015). Contamination of stormwater by wastewater: A review of detection methods. *Journal of Environmental Management*, *152*, 241–250. http://doi.org/10.1016/j.jenvman.2015.01.050

Paul, M. J., & Meyer, J. L. (2001). Streams in the urban landscape. *Annual Review of Ecology and Systematics*, *32*, 333–365.

R Core Team. (2015). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org/

Schueler, T., Fraley-McNeal, L., & Cappiella, K. (2009). Is Impervious Cover Still Important? Review of Recent Research. *Journal of Hydrologic Engineering*, *14*(4), 309–315. http://doi.org/10.1061/(ASCE)1084-0699(2009)14:4(309)

Taylor, D. I. (2010). The Boston Harbor Project, and large decreases in loadings of eutrophication-related materials to Boston Harbor. *Marine Pollution Bulletin*, *60*(4), 609–619. http://doi.org/10.1016/j.marpolbul.2009.10.006

U.S. Environmental Protection Agency. (2009). *2004 National Water Quality Inventory Report to Congress* (Data and Tools). Retrieved from

https://www.epa.gov/waterdata/2004-national-water-quality-inventory-report-congress

U.S. Environmental Protection Agency. (2015a). General Permits for Stormwater Discharges from Small Municipal Separate Storm Sewer Systems in Massachusetts [DRAFT]. Retrieved from http://www3.epa.gov/region1/npdes/stormwater/ma/2014DraftMASmallMS4GeneralPermit.pdf

U.S. Environmental Protection Agency. (2015b). New Grading System for Mystic River Watershed Gives Public Better Localized Information. Retrieved April 18, 2016, from https://yosemite.epa.gov/opa/admpress.nsf/0/F6B639348AF5E63685257E880049908B

Wickham, H. (2011). ggplot2. *Wiley Interdisciplinary Reviews: Computational Statistics*, *3*(2), 180–185. http://doi.org/10.1002/wics.147

Wickham, H., & Francois, R. (2015). *dplyr: A Grammar of Data Manipulation*. Retrieved from http://CRAN.R-project.org/package=dplyr