Harvard University
Harvard University Biostatistics Working Paper Series

Year 2005 *Paper* 17

A Nonstationary Negative Binomial Time Series with Time-Dependent Covariates: Enterococcus Counts in Boston Harbor

E. Andres Houseman* Brent Coull†

James P. Shine‡

*Harvard School of Public Health, ahousema@hsph.harvard.edu

†Harvard School of Public Health, bcoull@hsph.harvard.edu

‡Harvard School of Public Health, jshine@hsph.harvard.edu

http://biostats.bepress.com/harvardbiostat/paper17

A Nonstationary Negative Binomial Time Series with Time-Dependent Covariates: Enterococcus Counts in Boston Harbor

E. Andres Houseman, Brent Coull, and James P. Shine

Abstract

Boston Harbor has had a history of poor water quality, including contamination by enteric pathogens. We conduct a statistical analysis of data collected by the Massachusetts Water Resources Authority (MWRA) between 1996 and 2002 to evaluate the effects of court-mandated improvements in sewage treatment. Motivated by the ineffectiveness of standard Poisson mixture models and their zero-inflated counterparts, we propose a new negative binomial model for time series of Enterococcus counts in Boston Harbor, where nonstationarity and autocorrelation are modeled using a nonparametric smooth function of time in the predictor. Without further restrictions, this function is not identifiable in the presence of time-dependent covariates; consequently we use a basis orthogonal to the space spanned by the covariates and use penalized quasi-likelihood (PQL) for estimation. We conclude that Enterococcus counts were greatly reduced near the Nut Island Treatment Plant (NITP) outfalls following the transfer of wastewaters from NITP to the Deer Island Treatment Plant (DITP) and that the transfer of wastewaters from Boston Harbor to the offshore diffusers in Massachusetts Bay reduced the Enterococcus counts near the DITP outfalls.

A Nonstationary Negative Binomial Time Series with Time-Dependent Covariates: Enterococcus Counts in Boston Harbor

E. Andrés Houseman
Department of Biostatistics
Harvard School of Public Health
655 Huntington Ave., Boston, MA 02115, U.S.A.

Brent A. Coull
Department of Biostatistics
Harvard School of Public Health
655 Huntington Ave., Boston, MA 02115, U.S.A.

James P. Shine
Department of Environmental Health
Harvard School of Public Health
655 Huntington Ave., Boston, MA 02115, U.S.A.

September 3, 2005

Abstract

Boston Harbor has had a history of poor water quality, including contamination by enteric pathogens. We conduct a statistical analysis of data collected by the Massachusetts Water Resources Authority (MWRA) between 1996 and 2002 to evaluate the effects of court-mandated improvements in sewage treatment. Motivated by the ineffectiveness of standard Poisson mixture models and their zero-inflated counterparts, we propose a new negative binomial model for time series of Enterococcus counts in Boston Harbor, where nonstationarity and autocorrelation are modeled using a nonparametric smooth function of time in the predictor. Without further restrictions, this function is not identifiable in the presence of time-dependent covariates; consequently we use a basis orthogonal to the space spanned by the covariates and use penalized quasi-likelihood (PQL) for estimation. We conclude that Enterococcus counts were greatly reduced near the Nut Island Treatment Plant (NITP) outfalls following the transfer of wastewaters from NITP to the Deer Island Treatment Plant (DITP) and that the transfer of wastewaters from Boston Harbor to the offshore diffusers in Massachusetts Bay reduced the Enterococcus counts near the DITP outfalls.

Keywords

B-splines, Enterococcus, Fourier series, Penalized spline, Poisson-gamma, Orthogonal Basis, Overdispersion, Semiparametric regression, Smoothing, Water monitoring

1 Introduction

1.1 Background

Boston Harbor is a shallow marine estuary that has been a receptacle of both domestic and industrial wastes for hundreds of years. By the latter part of the twentieth century, decades of direct contaminant input from poorly treated sewage effluent and sewage sludge, combined with non-point inputs from storm water discharges and daily combined sewer overflows, resulted in such poor water quality that Boston Harbor was characterized as the most polluted harbor in the United States, nicknamed the "Harbor of Shame" (McGonagle and Otski, 1997; MacDonald, 1991). In 1985, in response to a class-action suit and a mandated federal court-order to clean the Harbor on a 13 year schedule, the state of Massachusetts created the Massachusetts Water Resources Authority (MWRA). The MWRA assumed responsibility for water and sewer systems in the Boston metropolitan area, serving 2.5 million people and 5500 industrial sources in 61 communities (Massachusetts Water Resources Authority, 2004). With respect to sewage, the MWRA operated two sewage treatment plants that discharged directly into Boston Harbor: the Nut Island Treatment Plant (NITP) and the Deer Island Treatment Plant (DITP). As late as 1999, the last year that sewage was discharged directly into the Harbor, the volume of sewage discharged through the MWRA's treatment plants compromised approximately 50% of the freshwater entering the Harbor, matching the combined contribution of freshwater from other sources such as the Charles, Mystic, and Neponsett Rivers (United States Geological Survey, 2004). In response to the federally mandated clean-up, the MWRA spent approximately 3.8 billion dollars to upgrade the quality of sewage treatment; a major portion of the project included diversion of the sewage outfall from direct discharge into the Harbor to a series of outfall diffusers located 15 km offshore in nearby Massachusetts Bay (McGonagle and Otski, 1997). A large frac-

tion of the clean-up costs were passed on to the local citizens, whose water rates became amongst the highest in the nation. Several events were significant in the subsequent history of Boston wastewater treatment. In 1991, the discharge of sewage sludge into the Harbor was completely halted. In 1998, flows from the Nut Island Treatment Plant ceased and all wastewater flows were transferred to the Deer Island Treatment Plant. A third major event occurred in September 2000 with the relocation of DITP outfalls from within Boston Harbor to the new offshore diffusers (Taylor, 2001, 2002, 2003). Figure 1 depicts DITP and NITP effluent flow, in millions of gallons per day (MGD), between 1996 and 2002.

It is well-known that the cessation of sludge input and an improved level of sewage treatment resulted in a decreasing trend in the loading of toxic chemicals to the Boston Harbor throughout the 1990's. This is reflected in decreasing concentrations of compounds such as heavy metals in the surface sediments (Zago et al., 2001). However, it is less clear how the Boston Harbor project affected the levels of pathogenic micro-organisms. Although poorly treated sewage can be a source of pathogens to the Harbor, previous studies have indicated that a large source of pathogens to the Harbor may have been through discharges via combined sewer overflows and storm drains (Ellis and Rosen, 2001). The presence of these enteric pathogens in the water can make the beaches unfit for recreation and the fish unsafe to consume. It is possible therefore, that although the $3.8 billion expenditure improved the level of sewage treatment and re-routed the effluent offshore, that the waters of Boston Harbor remained unfit for humans as both a recreational or fisheries resource.

Throughout the 1990's and early 2000's, the MWRA monitored a large number of water quality parameters throughout the Harbor. In particular, the MWRA has measured Enterococcus counts at 23 stations, illustrated in Figure 2, throughout Boston Harbor between 1995 and 2002. Enterococcus is an indicator bacterium correlated with the presence of other pathogenic bacteria and viruses, and is associated with an elevated risk of swimming-

related illnesses (Cabelli et al., 1979; Cabelli, 1982; Cabelli et al., 1982, 1983). Enterococcus counts, measured in colonies per 100 ml (in 5 colony increments), were typically collected every ten to fourteen days between 1995 and 2002. The time domain over which samples were collected generally depended on the station sampled, as did exact sampling frequency. Reported values consist of averages of replicate laboratory analyses, with acceptable precision calculated using Method 9020 B 4 of Standard Methods (APHA et al., 1995). Details are described in Tilton et al. (1998) and Margolin et al. (2002). We note that explicit laboratory recovery rates for Enterococcus were not measured. Inferences about time effects make the assumption that method differences in recovery rate had negligible effect on Enterococcus count. Table 1 presents a summary of the data collected during this period. Figure 3 depicts Enterococcus counts (in colonies per 20 ml) over time at Station 82, which is near NITP, and at Station 160, near DITP.

The primary goal of this article is the quantification of the effects of two of the important milestones described above, the cessation of flows from NITP in 1998 and the relocation of the DITP outfalls in 2000. These are of interest from a regulatory and policy perspective. Also of interest is characterizing the distribution to a degree sufficient for prediction, as such data are useful in the context of decision modeling for policy. For example, there is an interest in accurately predicting levels of Enterococcus at beaches to inform beach closures for reasons of public health (Morrison et al., 2003).

1.2 Statistical Methodology

The statistical analysis of the water pathogen time series presents several challenges. First, the counts do not apparently follow a standard distribution. The counts are clearly overdispersed relative to the Poisson distribution, as Figure 3 demonstrates. However, they also exhibit a high proportion of zero counts. Consequently, lognormal approximation to the dis-

tribution is inadequate, as can be shown by methods proposed by Houseman et al. (2004). This remains true even when models that account for zero-inflation are considered. Second, an important aspect of the data is the nonstationarity of both the outcomes and the effects of interest. Failing to address nonstationarity complicates the characterization of the underlying probability distribution. Moreover, as we demonstrate in Section 4, misleading inference may result if the time-dependence of the effects of interest in the context of the nonstationary outcomes is not taken into account.

Considerable work has been done in the area of time series for counts. Cox (1981) classified time-series models for serially-correlated data into two classes: observation-driven and parameter-driven. Observation-driven models specify the conditional distribution of a response at time t as a function of past responses; in contrast, parameter-driven models specify an underlying serially correlated latent process. Examples of parameter-driven (state-space) models include Zeger (1988), who developed a quasi-likelihood approach derived from a latent process assumption, and Kelsall et al. (1999), who proposed a frequency-domain approach. Other parameter-driven methods appear in a Bayesian context (e.g. Durbin and Koopman, 1997, 2000; Crainiceanu et al., 2003). Examples of observation-driven models include those proposed by Zeger and Qaqish (1998), who proposed a Markov approach, and Brumback et al. (2000), who condition a mean response on past observations. The *generalized autoregressive moving average* (GARMA) approach proposed by Benjamin et al. (2003) is similar but more general. Benjamin et al. (2003) provides an excellent overview of the history of time series for counts.

Although theory exists for irregularly spaced observations (Omori, 2003), which are common in environmental applications, many existing methods are difficult to adapt to such irregularity. This is particularly true of observation-driven methods, since some of the historical observations are essentially missing. Consequently, recent approaches to time series

data in environmental applications use a semiparametric approach, where a nonparametric time effect is included in the mean (Samoli et al., 2001; Coull et al., 2001). The advantage of such an approach is that it more flexibly models nonstationarity in the time effect. Recent semiparametric approaches to count data include Hunsberger et al. (2002), who applied locally-weighted scatterplot smoothing to Poisson time series, and Thurston et al. (2000), who proposed a semiparametric model for negative-binomial counts. The latter authors used locally-weighted scatterplot smoothing, and did not consider time series. In addition, they did not focus on selection of the bandwidth tuning parameter.

We found existing time series models for counts inadequate for the MWRA Enterococcus data. The frequency-domain methodology proposed by Kelsall et al. (1999) requires stationary outcomes, so is not easily adapted to Boston Harbor pathogen data. Originally, we applied the methods of Zeger (1988) to the Enterococcus data set, but found that the high proportion of zero counts, combined with spikes, resulted in estimating equations that were so ill-conditioned that reliable estimates could not be obtained. We also explored Bayesian fitting of the Poisson-lognormal model described in Crainiceanu et al. (2002) and Crainiceanu et al. (2003), but for our problem we found the mixing properties of the resulting MCMC chain to be poor even after extended computation time (about 12 hours on a 3.2 MHz PC for a single station) and using recommended methods to improve mixing. Finally, we considered several zero-inflation models (e.g. Lambert, 1992) and their time-series extensions (Wang, 2001) but ultimately decided against these approaches for reasons outlined in Min and Agresti (2005). First, zero-inflated models require that the data be zero inflated at every level of the covariates, which may not always be realistic. Violation of this assumption can yield unstable estimates, even in the presence of marginal zero-inflation. Second, the parameters are difficult to interpret, as one must distinguish whether an effect relates to a change in the probability of a zero count or to the mean count, given that the count

arises from the Poisson component of the model. Nevertheless, despite the drawbacks, we fit zero-inflated Poisson and zero-inflated negative binomial models to the Boston Harbor data. Our results mimicked the phenomena described by Min and Agresti (2005), with estimates proving unstable, as judged by the rank of the observed information matrix.

To address the limitations of existing approaches in the context of the MWRA data analysis, we propose a time series model for overdispersed count data in situations in which interest focuses on time-dependent covariates. Our model assumes negative binomial outcomes and employs a nonparametric, smooth function $f(t)$ to model simultaneously serial autocorrelation and zero-inflation, using the mixed-model approach proposed by Brumback et al. (1999) to select the tuning parameter. Our approach is related to that proposed in Thurston et al. (2000). However, the presence of smooth covariate processes complicates the analysis, leading to nonidentifiability of regression parameters unless basis functions are specified properly. As a result, we propose an approach that assumes a smooth residual process that is orthogonal to the known covariate processes. In addition, we provide methodology for standard errors that are robust to time-dependent dispersion.

Our paper is organized as follows. In Section 2, we describe our model in detail, discuss the identifiability of $f(t)$, motivate a basis for the space in which $f(t)$ should lie, and describe a penalized quasi-likelihood (PQL) approach to estimation. In addition, we propose procedures for robust standard error estimation, for goodness-of-fit and for spatial summarization. In Section 3 we use our model for hypothesis testing from data collected at four stations in Boston Harbor, and present a summary of an analysis of the nineteen remaining stations. We conclude with some closing remarks in Section 4.

2 Semiparametric Negative Binomial Model for Time Series

In this section we describe a statistical model for a time series of pathogen counts. Section 2.1 proposes the basic model, a negative-binomial outcome whose mean contains both a parametric specification, including time-dependent covariates, and a nonparametric time effect $f(t)$. In Section 2.2, we discuss identifiability of f in the presence of smooth covariate processes, and motivate a basis, orthogonal to the space spanned by the covariate processes, for the space in which f is assumed to lie. In Section 2.3 we briefly compare the negative-binomial model to the Poisson and zero-inflated Poisson models, demonstrating that in this context the Poisson model may not be sufficiently rich to model both autocorrelation and overdispersion, and that the negative-binomial model can accommodate zero-inflation. In Section 2.4, we discuss estimation and inference for the semiparametric negative binomial model. In Section 2.5 we describe how to obtain standard errors for the mean parameters that are robust to time-dependent overdispersion. In Section 2.6, we discuss goodness-of-fit. Finally, in Section 2.7 we describe the construction of a spatial summary of estimates obtained at different stations.

2.1 Negative Binomial Time Series with Time-Dependent Covariates

Fix a time interval $\mathcal{T} = [t_0, t_1]$ and let $L = t_1 - t_0$. To simplify notation, we assume, without loss of generality, that $t_0 = 0$ and $t_1 = L$. For $t \in \mathcal{T}$, assume $Y_t|Q_t \overset{iid}{\sim} Po(Q_t\mu_t)$, where $Q_t \overset{iid}{\sim} Ga(\sigma^{-1}, \sigma^{-1})$, $\mu_t = \exp[\mathbf{x}_t'\boldsymbol{\beta} + f(t)]$, $\mathbf{x}_t \in \mathbb{R}^p$, and the unknown function f is a member of $\mathcal{H}$, a Hilbert space of functions $\mathcal{T} \to \mathbb{R}$ with square integrable second derivative. Thus, Y_t is an overdispersed Poisson variable, with the overdispersion introduced through a Gamma variable Q_t having mean one and variance σ. This is equivalent to a negative

binomial model for Y_t:

$$\mathrm{P}\left(Y_t = y|\mu_t\right) = \frac{\Gamma(y+\sigma^{-1})}{\Gamma(\sigma^{-1})\Gamma(y+1)}\left(\frac{1}{1+\mu_t\sigma}\right)^{\sigma^{-1}}\left(\frac{\mu_t}{\mu_t+\sigma^{-1}}\right)^{y}. \tag{1}$$

We further assume that there is a subset $\mathcal{J} \subset \{1, ..., p\}$ corresponding to J time-dependent covariates, where for each $j \in \mathcal{J}$, $\mathbf{x}_{tj} = u_j(t)$, and $u_j \in \mathcal{H}$. Finally, assume that for a finite set of size n, $\mathcal{T}^* \subset \mathcal{T}$, we observe $(Y_t, \mathbf{x}_t')'$ for every $t \in \mathcal{T}^*$. We impose an assumption that the Y_t are missing completely at random, i.e. conditional on covariates, the probability that $t \in \mathcal{T}^*$ does not depend on $\{Y_t\}_{t\in\mathcal{T}}$. This assumption is reasonable for the Boston Harbor data, since sampling schedules were constructed independent of anticipated count levels.

The flows depicted in Figure 1, which are apparently smooth curves combined with noise, motivate our interest in time-dependent effects. We are not directly interested in estimating u_j, but the fact that u_j is time-dependent has implications for estimating f, particularly when the parametric effects of $\mathbf{x}_{tj}$ are of interest. In Section 2.2, we motivate further restrictions on the function f in order to identify f in the presence of the smooth covariate processes u_j.

2.2 Indentifiability

The model proposed in Section 2.1 is not identifiable. For a fixed $j \in \mathcal{J}$,

$$\beta_0 + \beta_j u_j(t) + f(t) = \beta_0 + (\beta_j + \gamma)u_j(t) + [f(t) - \gamma u_j(t)], \tag{2}$$

which produces the same distribution for every $\gamma \in \mathbb{R}$. Even if there are stationary zero-mean errors e_{tj} such that $\mathbf{x}_{tj} = u_j(t) + e_{tj}$, the model still has poor properties when e_{tj} is small relative to $u_j(t)$, since $\mathbf{x}_{tj}$ will still be highly collinear with the space of admissible functions f.

Identifiability can be obtained by restricting $f(t)$ to lie in a space $\mathcal{H}_1$ of smooth functions that are orthogonal to the subspace $\mathcal{H}_0$ spanned by all J of the functions $u_j(t)$. By standard

Hilbert space properties, $\mathcal{H}_0 \perp \mathcal{H}_1$ ensures a unique factorization of the smooth portion of $\log(\mu_t)$ into $\sum_{j\in\mathcal{J}} \beta_j u_j \in \mathcal{H}_0$ and $f \in \mathcal{H}_1$. A basis representation for $\mathcal{H}_1$ can easily be obtained from a standard basis for $\mathcal{H}$, such as the B-spline basis (Ramsay and Silverman, 1997, page 49), by Gram-Schmidt orthogonalization. Note that such a procedure depends critically on the definition of an appropriate inner product $\langle\cdot,\cdot\rangle$, which for square integrable functions is usually defined as $\langle g_1, g_2\rangle = \int_0^L g_1(t)g_2(t)dt$.

In particular, suppose $\mathcal{H} \supset \mathcal{H}_0$ is a Hilbert space for which f is assumed to lie in the linear complement $\mathcal{H}_1 = \mathcal{H} - \mathcal{H}_0$, and that all members of $\mathcal{H}$ are expressed through a finite-dimensional basis expansion $g(t) = \sum_{k=1}^K b_k \tilde{z}_k(t) = \mathbf{b}'\tilde{\mathbf{z}}_t$, where $\tilde{\mathbf{z}}_t' = [\tilde{z}_1(t), ..., \tilde{z}_K(t)]$ is a vector of spline functions evaluated at t. Then there is a $(K-J)\times K$ matrix $\mathbf{P}$ such that $\mathbf{P}\tilde{\mathbf{z}}_t$ is a finite-dimensional basis approximating $\mathcal{H}_1$. Thus, arbitrary members of $\mathcal{H}_1$ can be expanded as $\mathbf{a}'\mathbf{P}'\tilde{\mathbf{z}}_t$. We denote the orthogonal basis obtained from $\tilde{\mathbf{z}}_t$ as $\mathbf{z}_t = \mathbf{P}'\tilde{\mathbf{z}}_t$.

In functional data analysis problems such as this one, reduction of the infinite dimensional problem to finite dimensions typically involves regularization, where estimation involves a penalty such as the square-integral of the second derivative. In Section 2.4 we describe an estimation procedure that imposes regularization by penalizing solutions for which $\int_0^L |\partial^2 f/\partial t^2|^2 dt$ is large. The "roughness penalty" simplifies to a quadratic form $q(\mathbf{a};\lambda) = \lambda\mathbf{a}'\mathbf{D}\mathbf{a}$ (Ramsay and Silverman, 1997, Chapter 4), where λ is a tuning parameter and $\mathbf{D}$ is a "penalty matrix". The *fda* package written by J. Ramsay for the R statistical software environment (R Development Core Team, 2004) provides the B-spline basis representation of functions, as well as support for inner product computations and the regularization penalty $\mathbf{b}'\tilde{\mathbf{D}}\mathbf{b}$ obtained by integrating the square of the second derivatives of the B-spline basis functions. It is straightforward to show that, for members of $\mathcal{H}_1$ represented by the basis $\mathbf{z}$, the corresponding penalty matrix is $\mathbf{D} = \mathbf{P}\tilde{\mathbf{D}}\mathbf{P}'$. Consequently, once $\mathbf{P}$ has been computed, the penalty matrix for regularizing members of $\mathcal{H}_1$ can be obtained from

standard software such as *fda*. Note that the nullspace of $\tilde{\mathbf{D}}$ includes vectors that represent multiples of the identity function $\mathrm{id}(t) \equiv t$, so $\tilde{\mathbf{D}}$ is singular; however, if there is at least one smooth covariate u_j such that $\langle u_j, \mathrm{id} \rangle \neq 0$, then $\mathbf{D} = \mathbf{P}\tilde{\mathbf{D}}\mathbf{P}'$ will be nonsingular.

In practice, we select the tuning parameter λ using the mixed-model formulation proposed by Brumback et al. (1999). The quadratic form of the roughness penalty $q(\mathbf{a}; \lambda) = \lambda \mathbf{a}'\mathbf{D}\mathbf{a}$ suggests that $f(t)$ can be described as a stochastic process generated by considering the spline coefficients as random effects $\mathbf{a} \sim N(0, \tau^2 \mathbf{D}^{-1})$, where $2\tau^2 = \lambda^{-1}$. In this formulation, "rough" elements of $\mathcal{H}_1$ have small probability of arising. Thus, our final model is a negative binomial mixed model, closely related to the mixed model proposed by Booth et al. (2003). As described by Brumback et al. (1999), our proposed formulation facilitates a natural method for selecting the tuning parameter λ. We discuss additional details of estimation and inference in Sections 2.4, 2.5 and 2.7. For the remainder of this article, we assume that $f(t) = \mathbf{z}_t'\mathbf{a}$ and that

$$\log(\mu_t) = \mathbf{x}_t'\boldsymbol{\beta} + \mathbf{z}_t'\mathbf{a}. \tag{3}$$

Finally, we note that, even when $f \in \mathcal{H}$ is unconstrained and is thus unidentifiable, the penalized likelihood still produces a solution. However, the solution is not the same as that obtained under the orthogonality constraint. To see this, we concentrate on a single time-dependent covariate $\boldsymbol{\beta}_j$ and express equation (2) in terms of basis expansions:

$$\beta_0 + \beta_j \mathbf{z}_t'\mathbf{c}_j + \mathbf{z}_t'\mathbf{a} = \beta_0 + \mathbf{z}_t'(\beta_j + \gamma \mathbf{c}_j) + \mathbf{z}_t'[\mathbf{a} - \gamma \mathbf{c}_j],$$

where $f(t) = \mathbf{z}_t'\mathbf{a}$ and $u_j(t) = \mathbf{z}_t'\mathbf{c}_j$. If we assume that β_j and $\mathbf{a}$ are identified by the orthogonality constraint, then γ must be zero. However, in the unconstrained setting, the likelihood is determined only up to values of γ, and the likelihood is therefore singular. The penalty, whose general form is $\lambda(\mathbf{a} - \gamma \mathbf{c}_j)'\mathbf{D}(\mathbf{a} - \gamma \mathbf{c}_j)$, renders the penalized likelihood nonsingular and we are able to obtain a solution. Since the likelihood contains no information about γ,

the solution will minimize the penalty at $\gamma = \mathbf{a}'\mathbf{D}\mathbf{c}_j/\mathbf{c}_j'\mathbf{D}\mathbf{c}_j \neq 0$. Consequently, the solution obtained by the unconstrained problem differs from that obtained by the constrained problem. We illustrate this phenomenon in Section 4.

2.3 Comparison with Poisson Regression

Poisson models are often preferred over negative binomial models, since the negative binomial distribution is an exponential family only when σ is fixed. However, when counts are extremely overdispersed, the Poisson-gamma model is more realistic. Note that marginalizing over the random effects $\mathbf{a}$ in (3) induces overdispersion even in the Poisson case, $\sigma = 0$:

$$\text{Var}\,(Y_t) \;=\; \exp\left(\eta_t + \frac{\tau^2}{2}\mathbf{z}_t'\mathbf{D}^{-1}\mathbf{z}_t\right) + \exp\left(2\eta_t + \tau^2\mathbf{z}_t'\mathbf{D}^{-1}\mathbf{z}_t\right)\left[(\sigma+1)\exp\left(\tau^2\mathbf{z}_t'\mathbf{D}^{-1}\mathbf{z}_t\right) - 1\right],$$

where $\eta_t = \mathbf{x}_t'\boldsymbol{\beta}$. The term $\exp\left(2\eta_t + \tau^2\mathbf{z}_t'\mathbf{D}^{-1}\mathbf{z}_t\right)\left[\exp\left(\tau^2\mathbf{z}_t'\mathbf{D}^{-1}\mathbf{z}_t\right) - 1\right]$ is analogous to the overdispersion of a Poisson-lognormal distribution. However, for both Poisson and negative binomial models, the covariance has the same form:

$$\text{Cov}\,(Y_s, Y_t) \;=\; \exp\left(\eta_s + \eta_t + \frac{\tau^2}{2}\mathbf{z}_s'\mathbf{D}^{-1}\mathbf{z}_s + \frac{\tau^2}{2}\mathbf{z}_t'\mathbf{D}^{-1}\mathbf{z}_t\right)\left[\exp(\tau^2\mathbf{z}_s'\mathbf{D}^{-1}\mathbf{z}_t) - 1\right]. \quad (4)$$

Thus, for Poisson regression, overdispersion and autocorrelation are both determined by τ^2, but the negative binomial model includes an additional degree-of-freedom, $\sigma > 0$, allowing for a richer combination of autocorrelation and overdispersion. Although a Poisson-lognormal model would also provide a similarly rich space of models, the probability mass function of the Poisson-lognormal distribution does not have a closed form. Thus, the computational difficulties involved in implementing a Poisson-normal model motivate a preference for the negative binomial distribution.

Note that the semiparametric negative binomial model described in Section 2.1 accommodates zero inflation. From (1),

$$\text{P}\,(Y_t = 0|\mathbf{x}_t) = (1+\sigma\mu_t)^{-1/\sigma} = \{1 + \sigma\exp[\mathbf{x}_t'\boldsymbol{\beta} + f(t)]\}^{-1/\sigma}.$$

Thus, the probability of observing zero increases with σ and decreases with the Poisson mean μ_t. A run of pure zeros corresponds to small values of $f(t)$, while a time interval in which zeros alternate with spikes corresponds to a large value of σ combined with a moderate value of μ_t. Consequently, zero inflation is modeled in an integrated fashion that does not suffer from the stability and interpretation problems experienced with explicit zero-inflation models, as described by Min and Agresti (2005) and reviewed in Section 1. Note that a zero-inflated Poisson model is essentially a Poisson-binomial mixture, and is therefore, in a sense, less general than a Poisson-gamma mixture.

2.4 Estimation and Inference

The full mixed-model formulation described in Section 2.2 involves a likelihood containing an intractible integral:

$$\begin{aligned} \mathrm{P}\left(Y_t = y\right) &= \int_{\mathbb{R}^{2K}} \mathrm{P}\left(Y_t = y | \mu_t(\mathbf{a})\right) d\mathbf{a} \\ &= \frac{(2\pi\tau^2)^{-K}\Gamma(y+\sigma^{-1})}{\Gamma(\sigma^{-1})\Gamma(y+1)} \times \\ &\quad \int_{\mathsf{R}^{2K}} \left(\frac{1}{1+\sigma\mu_t(\mathbf{a})}\right)^{\sigma^{-1}} \left(\frac{\mu_t(\mathbf{a})}{\mu_t(\mathbf{a})+\sigma^{-1}}\right)^{y} \exp\left(-\frac{1}{2\tau^2}\mathbf{a}'\mathbf{D}\mathbf{a}\right) d\mathbf{a}, \end{aligned}$$

where $\mu_t(\mathbf{a})$ expresses the dependence of μ_t on the realized random effects vector $\mathbf{a}$. Following Breslow and Clayton (1993), a Laplace approximation can be used to obtain an approximate solution. Such an approach has been demonstrated to work well in smoothing applications (Hobert and Wand, 2000), and in particular overdispersed Poisson smoothing (Wager et al., 2004). Note that the approximation is only necessary for estimating the tuning parameter τ^2; assuming τ^2 is known, the methods of Thurston et al. (2000) can be used to estimate $\boldsymbol{\beta}$, σ, and $\mathbf{a}$ subject to a penalty constraint $\mathbf{a}'\mathbf{D}\mathbf{a} \leq C(\tau^2)$. This constraint is equivalent to maximizing a likelihood with an additional penalty term, $-2^{-1}\tau^{-2}\mathbf{a}'\mathbf{D}\mathbf{a}$. In addition, for proper choice of $\mathbf{D}$, the constraint can be interpreted as a roughness penalty

as measured by the square-integral of the second derivative of the function $f(t)$ (Ramsay and Silverman, 1997, Chapter 4). For fixed σ, iteratively reweighted least squares (IRLS) can be used to solve the corresponding score equation for $\boldsymbol{\beta}$ and $\mathbf{a}$:

$$\mathbf{U}_\mu(\boldsymbol{\beta}, \mathbf{a}; \sigma, \tau^2) = \tau^{-2}\mathbf{D}\mathbf{a} + \sum_{t \in \mathcal{T}^*} \left(\frac{Y_t - \mu_t(\mathbf{a})}{1 + \mu_t(\mathbf{a})\sigma} \right) \mathbf{w}_t = 0. \tag{5}$$

where $\mathbf{w}_t' = (\mathbf{x}_t', \mathbf{z}_t')$. Assuming $\boldsymbol{\beta}$ and $\mathbf{a}$ are known, maximum likelihood (ML) can be used to obtain an estimate for σ. The corresponding score equation is

$$U_\sigma(\sigma; \boldsymbol{\beta}, \mathbf{a}, \tau^2) = 0, \tag{6}$$

where $U_\sigma(\sigma; \boldsymbol{\beta}, \mathbf{a}, \tau^2)$ is equal to

$$\sigma^{-1} \sum_{t \in \mathcal{T}^*} \left[\psi_1(\sigma^{-1} + Y_t) - \psi_1(\sigma^{-1}) - \log(1 + \mu_t(\mathbf{a})\sigma) \right] + \sum_{t \in \mathcal{T}^*} \left[\frac{\sigma \mu_t(\mathbf{a}) Y_t + \mu_t(\mathbf{a})}{1 + \mu_t(\mathbf{a})\sigma} - Y_t \right],$$

and the digamma function $\psi_1(u)$ is the first derivative of $\log[\Gamma(u)]$. For a fixed value τ^2, the Laplace approximation L_τ to the log-likelihood involves solving (5) and (6) for $\widehat{\boldsymbol{\beta}}$, $\widehat{\mathbf{a}}$ and $\widehat{\sigma}$, and computing

$$L_\tau(\tau^2; \widehat{\boldsymbol{\beta}}, \widehat{\mathbf{a}}, \widehat{\sigma}) = \tilde{C}(\widehat{\boldsymbol{\beta}}, \widehat{\mathbf{a}}, \widehat{\sigma}) - \frac{1}{2\tau^2} \widehat{\mathbf{a}}' \mathbf{D} \widehat{\mathbf{a}} - K \log(\tau^2) - \frac{1}{2} \log \left| \mathbf{W}' \boldsymbol{\Omega} \mathbf{W} + \tau^{-2} \mathbf{D} \right|, \tag{7}$$

where $\mathbf{W} = (\mathbf{w}_{t_1}, ..., \mathbf{w}_{t_n})'$ is the matrix of covariates and basis vectors, $\tilde{C}(\widehat{\boldsymbol{\beta}}, \widehat{\mathbf{a}}, \widehat{\sigma})$, a normalization constant plus the sum of the natural logarithms of (1) computed for each $t \in \mathcal{T}^*$, is independent of τ^2, and $\boldsymbol{\Omega}$ is a diagonal $n \times n$ matrix of IRLS weights. The weights are computed as $\Omega_{tt} = \widehat{\mu}_t(\widehat{\mathbf{a}})(1 + \widehat{\mu}_t(\widehat{\mathbf{a}})\widehat{\sigma})^{-1}$, where the i^{th} diagonal element of $\boldsymbol{\Omega}$ is represented as $\Omega_{t_i t_i}$. Thus $\widehat{\tau}^2$ is obtained by maximizing (7). The procedure obtained by solving (5), solving (6), maximizing (7), and iterating until convergence, produces the penalized quasi-likelihood (PQL) estimate proposed by Breslow and Clayton (1993) for the negative binomial distribution.

Estimation of the variance of parameter estimates follows from the standard quasi-likelihood theoretical development of Breslow and Clayton (1993). In particular, conditional

on τ^2, the variance of $\widehat{\theta} = (\widehat{\boldsymbol{\beta}}', \widehat{\mathbf{a}}')'$ is approximated as

$$\text{Cov}\left(\widehat{\theta}\right) \approx (\mathbf{W}'\boldsymbol{\Omega}\mathbf{W} + \tau^{-2}\mathbf{D})^{-1}\mathbf{W}'\boldsymbol{\Omega}\mathbf{W}(\mathbf{W}'\boldsymbol{\Omega}\mathbf{W} + \tau^{-2}\mathbf{D})^{-1}. \tag{8}$$

Wald-type inference for linear combinations of $\widehat{\boldsymbol{\beta}}$ are easily obtained from (8). In addition, the estimator $\widehat{\sigma}$ is asymptotically uncorrelated with $\widehat{\theta}$ and its variance can be estimated as the inverse of the derivative of U_σ in (6). We emphasize, however, that all such Wald-type inference based on the approximation (8) is conditional on $\widehat{\tau}^2$. We view this as an acceptable and useful approximation, given that the asymptotic theory for penalized regression splines has not yet been fully developed (Ruppert et al., 2003; Crainiceanu and Ruppert, 2004).

We close this section with a comment on model degrees-of-freedom, which in semiparametric regression is an important tool for comparing the complexity of different models. In analogy with ordinary least squares, it is defined as the rank of the *smoother matrix*, which is a linear transformation that maps an outcome vector to its corresponding predictor. In the generalized linear model setting, this is the rank of the corresponding matrix for the working residuals. In our context, the methods of Ruppert et al. (2003, Section 3.13) produce

$$df = \text{trace}\left[\mathbf{W}(\mathbf{W}'\boldsymbol{\Omega}\mathbf{W} + \tau^{-2}\mathbf{D})^{-1}\mathbf{W}'\boldsymbol{\Omega}\right] = \text{trace}\left[(\mathbf{W}'\boldsymbol{\Omega}\mathbf{W} + \tau^{-2}\mathbf{D})^{-1}\mathbf{W}'\boldsymbol{\Omega}\mathbf{W}\right].$$

2.5 Time-Dependent Overdispersion

Equation (1) assumes a constant overdispersion parameter σ. However, time-dependent overdispersion is possible, in which case σ_t would be substituted for σ in (1). Incorporating an additional smooth term $\sigma_t = s(t)$ in (1) would vastly increase the complexity of the computation; in addition, it could lead to stability problems, as Brumback et al. (2000) have suggested in a similar context.

Estimation of mean parameters $\boldsymbol{\beta}$ and $\mathbf{a}$ is robust to time-dependent overdispersion,

since estimating equation (5) is unbiased for $\boldsymbol{\beta}$ and $\mathbf{a}$, aside from the $\tau^{-2}\mathbf{Da}$ term necessary for regularized estimation of $f(t)$. However, a remaining problem is that standard error estimates arising from (8) will be incorrect.

We employ methods analogous to those outlined in Brumback et al. (2000) for computing standard errors robust to time-varying overdispersion. Let $R_t = (Y_t - \widehat{\mu}_t)^2/(\widehat{\mu}_t + \widehat{\mu}_t^2\widehat{\sigma})$, where $\widehat{\mu}_t$ and $\widehat{\sigma}$ are obtained as in Section 2.4. If the model is correctly specified, $\mathrm{E}[R_t] = 1$; otherwise,

$$\mathrm{E}[R_t] = \frac{\mu_t + \mu_t^2\sigma_t}{\widehat{\mu}_t + \widehat{\mu}_t^2\widehat{\sigma}} = h(t). \tag{9}$$

Assuming $h(t)$ is a smooth function, a standard generalized additive model (GAM, Hastie and Tibshirani, 1990) can be employed to obtain an estimate $\widehat{h}(t)$, which then serves as an estimate of the left hand side of (9). The appropriately adjusted IRLS weight Ω_{tt}^* is then computed as

$$\Omega_{tt}^* = \frac{\widehat{\mu}_t^2}{\widehat{\mu}_t + \widehat{\mu}_t^2\widehat{\sigma}_t} = \frac{\widehat{\mu}_t^2}{(\widehat{\mu}_t + \widehat{\mu}_t^2\widehat{\sigma})\widehat{h}(t)} = [\widehat{h}(t)]^{-1}\Omega_{tt}.$$

Robust standard errors are obtained by substituting $\boldsymbol{\Omega}^*$, the matrix of IRLS weights Ω_{tt}^*, in place of $\boldsymbol{\Omega}$ in (8).

The standard chi-square test used in the GAM setting to test a smooth effect can be employed to test a null hypothesis of time-independent overdispersion. For if the overdispersion has no time dependence, then $h(t) \equiv 1$, in which case the chi-square test should fail to reject the null hypothesis of a constant smooth effect. Since $\mathrm{E}[R_t] > 0$, a logarithm link function for fitting the GAM seems reasonable. That is, $h(t)$ is assumed to be the exponent of a smooth real process. Since R_t can be quite skewed, a reasonable variance function for fitting $h(t)$ is $\mathrm{E}[R_t]^2$, corresponding to a Gamma distribution; thus, residuals having large expectation are down-weighted considerably.

2.6 Goodness-of-Fit

A common method of assessing fit is the deviance statistic, which is based on a likelihood ratio test that compares the target model with the "saturated" model in which each outcome is its own prediction (Ruppert et al., 2003, Chapter 4). For the negative binomial model (1), the deviance, which we denote G, is

$$G = 2\tilde{L} - 2L(\widehat{\tau}^2; \widehat{\boldsymbol{\beta}}, \widehat{\mathbf{a}}, \widehat{\sigma}) = 2\sum_{t \in \mathcal{T}^*} G_t - 2K\log(\tau^2) - \log\left|\mathbf{W}'\boldsymbol{\Omega}\mathbf{W} + \tau^{-2}\mathbf{D}\right|, \qquad (10)$$

where $\tilde{L}$ is the log-likelihood for the saturated model, computed with $\mu_t(\mathbf{a}) = Y_t$, $\mathbf{a} = \mathbf{0}$, $\sigma = \widehat{\sigma}$, and $\tau^2 = 0$, and

$$G_t = Y_t \log Y_t - Y_t - Y_t \log \widehat{\mu}_t(\widehat{\mathbf{a}}) + \sigma^{-1}\log\left|1 + \widehat{\mu}_t(\widehat{\mathbf{a}})\sigma\right|.$$

Details of the derivation are provided in the Appendix.

An assumption of asymptotic normality might be used to motivate the approximation $G \sim \chi^2_d$, where d is equal to the residual smoothing degrees-of-freedom, defined by Ruppert et al. (2003, Section 3.14). However, Crainiceanu and Ruppert (2004) have shown that normal approximations are not necessarily valid for penalized regression splines. Consequently, Ruppert et al. (2003, Section 4.8.2) recommend simulating the null distribution of likelihood ratio statistics, with parameters set to their true values. Thus, a large number M of data sets $\{(Y_t^{(m)}, \mathbf{x}_t')'\}$, $m \in \{1, ..., M\}$, are simulated using the model described in Sections 2.1 and 2.2, with $\boldsymbol{\beta}$, σ, and $\mathbf{a}$ set to their estimated values. The corresponding statistics $G^{(m)}$ are tabulated for each such data set, using $Y_t^{(m)}$, $\widehat{\mathbf{a}}$, $\widehat{\boldsymbol{\beta}}$, $\widehat{\sigma}$ and $\widehat{\tau}^2$. By refitting $\mathbf{a}$, $\boldsymbol{\beta}$, and σ for every m, it is possible to account for the effect of parameter estimation on the distribution of G. Note that fitting τ^2 by (7) is the most time-consuming step of the algorithm described in Section 2.4. However, τ^2 directly relates to the smoothness of $\widehat{f}$, which is fixed in the simulation by fixing $\mathbf{a} = \widehat{\mathbf{a}}$. Therefore, a reasonable approximation to the distribution of

G is still obtained by holding τ^2 constant at $\widehat{\tau}^2$ for each m, with considerable savings in computation.

The null distribution of G is taken to be the empirical distribution of $G^{(m)}$. For example, the P-value is approximately $M^{-1}\sum_{m=1}^{M} 1\left(G^{(m)} > G\right)$. An expression similar to (10) exists for the assumption $\sigma = 0$ in the computation of $\tilde{L}$; the corresponding null distribution may also be obtained by simulation.

2.7 Spatial Summary

Until this point we have ignored the spatial aspect of the Boston Harbor data. As described in the Introduction, Enterococcus was sampled at 23 different locations in the harbor. Since effects may differ considerably from site to site, we have focused the statistical methodology on analysis of a single site at a time. Nevertheless, spatial summarization of the effects of NITP and DITP flow are of interest. In particular, we wish to construct a map illustrating spatial heterogeneity in the harbor.

To this end, we make use of a two-stage approach (Fitzmaurice et al., 2004, Chapter 8.4,). At the first stage we apply the methodology described in Sections 2.4 and 2.5. For the second stage, we fit a geoadditive model (Kammann and Wand, 2003) to effect estimates obtained in the first stage. The geoadditive model uses thin-plate splines in the GAM setting to model geographic effects, and corresponds to a form of generalized Kriging (Ruppert et al., 2003, Chapter 13). Other Kriging solutions (Christensen, 1991, Chapter VI) could also be employed. However, we wish to weight the effects $\widehat{\beta}_j$ by the inverse of their variances, an operation that is much more easily accomplished using GAMs.

In this context, a spatial model can be described as follows:

$$\beta_j(\mathbf{s}) = \beta_{j0} + v_j(\mathbf{s}), \tag{11}$$

where $\beta_j(\mathbf{s})$ is the effect corresponding to element j of $\mathbf{x}_t$ at location $\mathbf{s}$, β_{j0} is the mean effect, and $v_j(\mathbf{s})$ is a smooth function over two-dimensional location vectors $\mathbf{s}$. Assuming models (1) and (11) are correct, fitting (11) to the parameter estimates $\widehat{\beta}_{j\mathbf{s}^*}$ from each sampling station $\mathbf{s}^*$, weighted by the inverse of their variances, produces valid estimates. Maps showing the spatial heterogeneity of $\beta_j(\mathbf{s})$ are obtained from the estimate $\widehat{\beta}_{j0}$ and spatial predictions $\widehat{v}_j(\mathbf{s})$ over a grid of locations $\mathbf{s}$ in the region of interest.

3 Enterococcus Counts in Boston Harbor

We now turn to the analysis of Enterococcus counts in Boston Harbor. For brevity, we concentrate on four representative stations: Stations 82 and 139, which are near the NITP outfalls, and Stations 159 and 160, which are near the pre-2000 DITP outfalls. All four of these stations were sampled every ten to fourteen days between 1995 through 2002. Subsequently, we present summary results for all 23 stations sampled and the results of a pooled spatial analysis. For effects involving NITP, we did not include Station 130 because the station was sampled only after NITP flows ceased in mid-1998. All analyses were conducted using the R software package (R Development Core Team, 2004, version 1.9.1), including the libraries *fda* as described in Section 2.4 and *mgcv* for the GAMs described in Section 2.5.

The Boston sewer system combines both sanitary flow and storm flow; during extreme storm events when flow is unusually high, raw sewage can be released from combined sewer overflows (CSOs) present on the shoreline. Thus, we expect the effects of base flow, which reflect "normal" plant operations, to differ from the transient effects of storm overflow. This motivates the decomposition of DITP and NITP flows into base flow (a lower envelope of the total flow) and positive residuals whose spikes reflect storm overflow. Figure 1 depicts

effluent flows at DITP and NITP, along with smooth curves representing base flow. The smooth curves were obtained by fitting an unpenalized B-splines in a quantile regression model for the 5^{th} percentile. Seventy-four knots were used for DITP and forty-one for NITP: eight regularly spaced knots per year were used for DITP throughout the time period studied and for NITP up through 1997; ten knots were used for NITP in 1998, and one knot per year thereafter. The cessation of NITP flows in mid-1998 motivated the irregular choice of knots. Of primary interest were the base flow effects, which we consider to be smooth processes whose indices belong to $\mathcal{J}$. The transient overflow effects, of secondary interest, were obtained as the residuals, observed flow minus the smoothed base flow. Since these residuals varied from day to day, we did not treat them as smooth. Thus, we had four flow covariates: the smooth base flows *DITP base* and *NITP base* and the residual flows *DITP overflow* and *NITP overflow*.

Another covariate of interest was the effect of transferring DITP discharges from outfalls in the harbor to the offshore diffusers in Massachusetts Bay. We represented this effect with an indicator *Mass Bay* having value one for samples collected after September 6, 2000, and zero otherwise. Strictly speaking, *Mass Bay* is not smooth in the sense described in Section 2.1. However, it is constant except for the singularity at September 6, 2000. Therefore, at each station, we projected it onto the B-spline basis $\tilde{\mathbf{z}}$ used to represent the function space $\mathcal{H}$, and henceforth treated it as smooth. The smooth covariates *DITP base* and *NITP base* were also projected onto the B-spline basis $\tilde{\mathbf{z}}$ after having been estimated using a different B-spline basis.

Other nuisance covariates included temperature (in degrees Celcius), salinity (in PSU), tide height (in meters), and a smooth seasonal effect represented by sinusoidal terms $\sin(2\pi t/T)$ and $\cos(2\pi t/T)$, where $T = 365.25$ days. We centered temperatures by subtracting a sinusoidal fit at each station, resulting in the covariate *Residual Temperature*. Similarly, we

centered salinity values to obtain *Residual Salinity*. We centered *Tide height* simply by subtracting the mean computed at each station.

The two base flows, the projected *Mass Bay* covariate, and the sinusoidal terms were all smooth. Thus, we constrained the residual process f to lie in the space orthogonal to these five smooth, time-varying covariates, in addition to an intercept. For each station we used a separate B-spline basis expansion with 50 knots determined by quantiles of the days on which the station was sampled.

For four Stations 82, 139, 159, and 160, complete results of the regression models described in Sections 2.1 through 2.5 appear in Table 2 and Figure 3. As mentioned in Section 2.4, inference is conditional on the value of τ^2. None of the base DITP flow effects were significant at the 5% level. For Stations 82 and 139, near NITP, the NITP flow effects were large and highly significant: 42 and 23 log colonies per 20 ml per 100 MGD at Station 82 and 139 respectively. Thus, the NITP flows appeared to have a strong negative impact on the water quality at nearby stations. The *Mass Bay* coefficients were significantly negative for the stations near NITP (159 and 160) but not for the other two, suggesting that the tranfer of discharges from the harbor to the offshore diffusers in Massachusetts Bay had a positive effect on water quality at stations near DITP.

Interestingly, there were highly significant and moderately strong DITP overflow effects at all four stations. These may represent the negative impact of storm overflow events in which untreated wastewater was discharged from CSOs. There were no significant NITP overflow effects, but this may be due to collinearity between *DITP overflow* and *NITP overflow* and the "loss of information" incurred by the cessation of NITP flows in mid-1998. In addition to the flow effects, there was evidence of tidal variation at Stations 82 and 159, residual temperature variation at Stations 82 and 160, seasonal variation at all four stations. There was no significant salinity effect at any of the four stations. Note that only Station

82 demonstrated significant time-dependent overdispersion.

The time series plots in Figure 3 show the nonparametric time effects at the four stations, together with 95% pointwise standard error bands obtained from appropriate quadratic forms involving (8). Only Station 160 seemed to have residual autocorrelation. Note that τ^2 was negligible for the other three stations, from which, after applying equation (4), we conclude that $f(t) \equiv 0$ and there was no temporal autocorrelation.

Tables 3 summarizes the effect of transferring flow from NITP to DITP at twenty-two stations. We captured this effect by subtracting the coefficient for *base DITP* from that for *base NITP*, with the result representing the *decrease* in log mean Enterococcus concentration associated with a transfer of 100 MGD from NITP to DITP. Many of the positive values were significant, further supporting the conclusion that the NITP wastewater discharges led to high Enterococcus counts throughout the harbor. The only significant negative value was at Station 160, which is quite close to the pre-2000 DITP outfalls. Table 4 summarizes the *Mass Bay* effects at the 14 stations for which sufficient post-2000 data were collected to estimate the effect. The effects were significantly negative at Stations 130, 159, and 160 and significantly positive at Stations 44 and 48.

Figure 4 shows the geographic impact of the MWRA policies, computed using the GAMs described in Section 2.7 to fit two-dimensional smooth functions to the estimated regression coefficients $\widehat{\beta}_j$ weighted by Var $\left(\widehat{\beta}_j\right)^{-1}$. In the maps presented in Figure 4, each contour line represents an integer level of the effect depicted, dark solid lines represent effects above the estimated GAM intercept (which we interpret as the geographic mean), and light dashed lines represent effects below the estimated intercept. The caption provides the intercepts, their 95% confidence intervals, and P values for tests of the hypothesis that the spatial effect is constant, using the chi-square test available through the GAM software. Figures 4(A) and 4(B) suggest that the base DITP flow effects were much flatter than the base

NITP flow effects. Figure 4(C) suggests that the transfer of NITP wastewater to DITP had the most significant impact near NITP, where the differences between the NITP and DITP transfer effects on Enterococcus counts were signficantly positive. Figure 4(D) suggests a somewhat flat *Mass Bay* effect that is nevertheless more strongly negative in the outer harbor and Massachusetts bay than in the inner harbor. Note that Figures 4(B) and 4(C) suggest anisotropic dispersion of poor-quality NITP effluent, with greater dispersion along a northwest-southeast axis; these figures are consistent with known tidal exchange and dispersion patterns in Boston Harbor (Signell and Butman, 1992).

Table 5 reports the results of goodness-of-fit tests for Poisson and negative binomial models at Stations 82, 139, 159 and 160, including P-values and simulation summaries for the test described in Section 2.6. This table also lists the P-values one would obtain naively using a chi-square distribution. It is evident from Table 5 that the negative binomial models fit much better than the Poisson models. This is consistent with the estimates and confidence intervals for the gamma mixing parameters σ presented in Table 2. Note that the distribution of the likelihood ratio statistic did not appear to be well-approximated by a chi-square distribution.

4 Closing Remarks

In this article we have proposed a negative binomial model for time series of Enterococcus counts in Boston Harbor, where nonstationarity and autocorrelation, are modeled using a nonparametric smooth function $f(t)$ of time in the predictor. Since f is not identifiable in the presence of smooth, time-dependent covariates, we restricted it to lie in a space orthogonal to that spanned by these smooth covariates. From the statistical analysis described in Section 3 we have concluded that Enterococcus counts were greatly reduced near the NITP

outfalls following the transfer of wastewaters from NITP to DITP and that the transfer of wastewaters from Boston Harbor to the offshore greatly reduced the Enterococcus counts near the DITP outfalls.

Figure 5 illustrates what would happen if we had not restricted f, but rather allowed it to be any twice-differentiable function. The figure shows the estimated residual process at Stations 139 and 160, obtained by using an unrestricted B-spline basis. Note that they differ from the estimates shown in Figure 3. In particular, at Station 139, Enterococcus levels appeared to drop steadily between 1995 and 2002, an effect apparently unexplained by parametric covariates. The *base NITP* flow effect in this model was 12.0, with 95% confidence interval $(-16.8, 5.5)$. In the corresponding model with restricted f, illustrated by Figure 3(B), the Enterococcus levels unexplained by covariates appeared to have negligible change over the eight-year monitoring period. The *base NITP* effect in the properly restricted model was 23.1, with 95% confidence interval $(11.4, 34.9)$. Thus, in the unrestricted model, much of the NITP effect was absorbed into $f(t)$ due to collinearity. In the restricted model, the effects of NITP flow were much more evident. A similar phenomenon occurred with the *Mass Bay* effect at Station 160: in the unrestricted model, the effect estimate was -1.4, with confidence interval $(-4.0, 1.2)$, while in the restricted model the estimate was -3.5, with confidence interval $(-4.7, -2.3)$.

In Section 1 we mention that a secondary goal of this analysis was to develop predictive models. As one reviewer pointed out, $f(t)$ cannot be predicted beyond the range of the data. However, marginalizing over $\mathbf{a}$ in (3) produces the prediction $\log(\mu_t) = \mathbf{x}_t'\boldsymbol{\beta}$ for values of t beyond the last time sampled. Since f was used primarily to account for temporal autocorrelation and seasonal effects were addressed by the sinusoidal terms in the model, such an approach seems adequate for the types of predictions required.

Figure 3(a) seems to suggest two outliers before 1998. However, both of these points

corresponded to high NITP flows and do not appear quite so extreme when compared with their predicted values, obtained either by including or excluding the two values. They also corresponded to high fecal coliform counts, so they are unlikely to reflect laboratory or data-entry error. Thus, removing them could bias the results. In general, negative binomial models are less sensitive to outliers than are Poisson models, where values would be weighted by μ_t^{-1} rather than the potentially much smaller negative binomial weight $(\mu_t + \sigma\mu_t^2)^{-1}$.

We used PQL to estimate the tuning parameter for smoothing f. This produced considerable computational savings for our application. However, while PQL has been shown to work well as an approximation when random effects are used for smoothing (Hobert and Wand, 2000; Wager et al., 2004), it can potentially lead to bias when random effects are used for more traditional purposes such as accounting for additional levels of clustering (Breslow and Lin, 1995; Lin and Breslow, 1996). Consequently, it is unclear whether this approach would perform well in applications involving hierarchical random effects.

The maps in Figure 4 hint at a more integrated analysis where the spatial components are estimated simultaneously with the other covariates. However, such an analysis would be complicated not only by the spatial heterogeneity of effects, but also by the differing levels of smoothness at each station. For example, Table 2 shows that the smoothing parameter τ^2 was quite different at Station 160 than at Stations 82, 139, and 159. This would require multiple smoothing parameters or a hierarchical treatment. In addition, the model would need to consider spatially-dependent overdispersion; presumably, this could be addressed in a manner similar to that described in Section 2.5, although $h(t)$ would become a function $h(t, \mathbf{s})$ of three dimensions with anisotropy. The philosophy employed in this paper was to minimize model complexity, a view that motivated not only the two-stage approach used to address the spatial dimension of the problem, but also the decision to avoid the problematic issues of stability and interpretation associated with the zero-inflated Poisson model.

Data and software are available from the authors upon request.

References

APHA, WPCF, and AWWA (1995), *Standard Methods for the Examination of Waters and Wastewaters*, Washington, D.C.: American Public Health Association, 19th ed.

Benjamin, M. A., Rigby, R. A., and Stasinopoulos, D. M. (2003), "Generalized autoregressive moving average models," *Journal of the American Statistical Association*, 98, 214–223.

Booth, J. G., Casella, G., Friedl, H., and Hobert, J. P. (2003), "Negative binomial loglinear mixed models," *Statistical Modelling*, 3, 179–191.

Breslow, N. E. and Clayton, D. G. (1993), "Approximate inference in generalized linear mixed models," *Journal of the American Statistical Association*, 88, 9–25.

Breslow, N. E. and Lin, X. H. (1995), "Bias correction in generalized linear mixed models with a single component of dispersion," *Biometrika*, 82, 81–91.

Brumback, B. A., Ruppert, D., and Wand, M. P. (1999), "Comment to 'Variable selection and function estimation in additive nonparametric regression using a data-based prior'," *Journal of the American Statistical Association*, 94, 794–797.

Brumback, B. A., Ryan, L. M., Schwartz, J. D., s, L. M. N., Stark, P. C., and Burge, H. A. (2000), "Transitional Regression Models, with Application to Environmental Time Se ries," *Journal of the American Statistical Association*, 95, 16–27.

Cabelli, V. J. (1982), "Microbial indicator systems for assessing water quality," *Antoine Van Leeuwenhoek Journal of Microbiology*, 48, 613–618.

Cabelli, V. J., Dufour, A. P., Levin, M. A., McCabe, L. J., and Haberman, P. W. (1979), "Relationship of microbial indicators to health effects at marine bathing beaches," *American Journal of Public Health*, 69, 690–696.

Cabelli, V. J., Dufour, A. P., McCabe, L. J., and Levin, M. A. (1982), "Swimming-associated gastroenteritis and water quality," *American Journal of Epidemiology*, 115, 606–616.

— (1983), "A marine recreational water quality criterion consistent with indicator concepts and risk analysis," *Journal of Water Pollution Control Federation*, 55, 1306–1314.

Christensen, R. (1991), *Linear Models for Multivariate, Time Series, and Spatial Data*, New York: Springer Verlag.

Coull, B. A., Schwartz, J., and Wand, M. P. (2001), "Respiratory health and air pollution: additive mixed model analyses," *Biostatistics*, 2, 337–349.

Cox, D. R. (1981), "Statistical analysis of time-series – some recent developments," *Scandanavian Journal of Statistics*, 8, 93–115.

Crainiceanu, C. M. and Ruppert, D. (2004), "Likelihood ratio tests in linear mixed models with one variance component," *Journal of the Royal Statistical Society, Series B*, 66, 165–185.

Crainiceanu, C. M., Ruppert, D., Stedinger, J. R., and Behr, C. T. (2002), "Improving MCMC Mixing for a GLMM Describing Pathogen Concentrations in Water Supplies," in *Case Studies in Bayesian Statistics*, Springer Verlag, pp. 207–221.

Crainiceanu, C. M., Stedinger, J., Ruppert, D., and topher Behr, C. (2003), "Modeling the National Distribution of Waterborne Pathogen Concentrations with Application to Cryptosporidium parvum," *Water Resources Research*, 39, Article No. 1235.

Durbin, J. and Koopman, S. J. (1997), "Monte Carlo maximum likelihood estimation for non-guassian state-space models," *Biometrika*, 84, 669–684.

— (2000), "Time series analysis of non-gaussian observations based on state-space models from both classical and [B]ayesian perspectives," *Journal of the Royal Statistical Society, Series B*, 62, 3–56.

Ellis, B. and Rosen, J. (2001), "Statistical Analysis of Combined Sewer Overflow Receiving Water Data, 1989 – 1999 (ENQUAD Report 2002-06)," Tech. rep., Massachusetts Water Resources Authority (http://www.mwra.state.ma.us/harbor/enquad/).

Fitzmaurice, G. M., Laird, N. M., and Ware, J. H. (2004), *Applied Longitudinal Analysis*, Hoboken, New Jersey: John Wiley & Sons.

Hastie, T. and Tibshirani, R. (1990), *Generalized Additive Models*, Boca Raton, Florida: Chapman & Hall/CRC.

Hobert, J. P. and Wand, M. P. (2000), "Automatic generalized nonparametric regression via maximum likelihood," Tech. rep., Department of Biostatistics, Harvard School of Public Health.

Houseman, E. A., Ryan, L. M., and Coull, B. A. (2004), "Cholesky Residuals for Assessing Normal Errors in a Linear Model with Correlated Outcomes," *Journal of American Statistical Association*, 99, 383–394.

Hunsberger, S., Albert, P. S., Follmann, D. A., and Suh, E. (2002), "Parametric and semiparametric approaches to testing for seasonal trend in serial count data," *Biostatistics*, 3, 289–298.

Kammann, E. E. and Wand, M. P. (2003), "Geoadditive models," *Journal of the Royal Statistical Society, Series C*, 52, 1–18.

Kelsall, J. E., Zeger, S. L., and Samet, J. M. (1999), "Frequency domain log-linear models; air pollution and mortality," *Journal of the Royal Statistical Society, Series C*, 48, 331–344.

Lambert, D. (1992), "Zero-inflated Poisson regression, with an application to defects in manufacturing," *Technometrics*, 34, 1–14.

Lin, X. H. and Breslow, N. E. (1996), "Bias correction in generalized linear mixed models with multiple components of dispersion," *Journal of the American Statistical Association*, 91, 1007–1016.

MacDonald, D. A. (1991), "Status and Trends in Concentrations of Selected Contaminants in Boston Harbor Sediments and Biota (NOS OMA 56)," Tech. rep., National Oceanic and Atmospheric Administration, National Ocean Service.

Margolin, A. B., Abramson, S. L., Gagnon, C., and Baptiste-Carpenter, E. M. (2002), "Combined Work/Quality Assurance Project Plan (CW/QAPP) for Anthropogenic Virus Survey: 2002-2005 (ENQUAD Report MS-073," Tech. rep., Massachusetts Water Resources Authority (http://www.mwra.state.ma.u s/harbor/enquad/).

Massachusetts Water Resources Authority (2004), "About MWRA," http://www.mwra.state.ma.us/02org/html/whatis.htm.

McGonagle, T. C. and Otski, R. M. (1997), "Toward a Healthy Harbor," *Civil Engineering*, 67, 46–49.

Min, Y. Y. and Agresti, A. (2005), "Random effect models for repeated measures of zero-inflated count data," *Statistical Modelling*, 5, 1–19.

Morrison, A. M., Coughlin, K., Shine, J. P., Coull, B. A., and Rex, A. C. (2003), "Receiver

operating characteristic curve analysis of beach water quality indicator variables," *Applied and Environmental Microbiology*, 69, 6405–6411.

Omori, Y. (2003), "Estimation for unequally spaced time series of counts with serially correlated random effects," *Statistics and Probability Letters*, 63, 1–12.

R Development Core Team (2004), *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-00-3.

Ramsay, J. O. and Silverman, B. W. (1997), *Functional Data Analysis*, New York: Springer Verlag.

Ruppert, D., Wand, M. P., and Carroll, R. J. (2003), *Semiparametric Regression*, Cambridge, UK: Cambridge University Press.

Samoli, E., Schwartz, J., Wojtyniak, B., Touloumi, G., Spix, C., Balducci, F., Medina, S., Rossi, G., Sunyer, J., Bacharova, L., Anderson, H. R., and Katsouyanni, K. (2001), "Investigating regional differences in short-term effects of air pollution on daily mortality in the APHEA project: A sensitivity analysis for controlling long-term trends and seasonality," *Environmental Health Perspectives*, 109, 349–353.

Signell, R. P. and Butman, B. (1992), "Modeling tidal exchange and dispersion in Boston Harbor," *Journal of Geophysical Research – Oceans*, 97, 15591–15606.

Taylor, D. I. (2001), "Comparison of water quality in Boston Harbor before and after inter-island transfer (ENQUAD Report 2001-09)," Tech. rep., Massachusetts Water Resources Authority (http://www.mwra.state.ma.u s/harbor/enquad/).

— (2002), "Water quality improvements in Boston Harbor during the first year aft er offshore transfer of Deer Island flows (ENQUAD Report 2002-04)," Tech. rep., Massachusetts Water Resources Authority (http://www.mwra.state.ma.u s/harbor/enquad/).

— (2003), "24 months after "offshore transfer": an update of water quality improve ments in Boston Harbor (ENQUAD Report 2003-04)," Tech. rep., Massachusetts Water Resources Authority (http://www.mwra.state.ma.u s/harbor/enquad/).

Thurston, S. W., Wand, M. P., and Wiencke, J. K. (2000), "Negative binomial additive models," *Biometrics*, 56, 139–144.

Tilton, K. S., Margolin, A. B., Albro, C. S., and Baptiste-Carpenter, E. M. (1998), "Combined Work/Quality Assurance Project Plan (CW/QAPP) for Anthropogenic Virus Survey: 1998-2000 (ENQUAD Report MS-047," Tech. rep., Massachusetts Water Resources Authority (http://www.mwra.state.ma.u s/harbor/enquad/).

United States Geological Survey (2004), "Real-Time Data for Massachusetts: Streamflow," Http://waterdata.usgs.gov/ma/nwis/current/.

Wager, C., Coull, B. A., and Lange, N. (2004), "Modelling spatial intensity for replicated inhomogeneous point patterns in brain imaging," *Journal of the Royal Statistical Society Series B*, 66, 429–446.

Wang, P. (2001), "Markov zero-inflated Poisson regression models for a time series of counts with excess zeros," *Journal of Applied Statistics*, 28, 623–632.

Zago, C., Giblin, A. E., and Bergamasco, A. (2001), "Changes in the metal content of surficial sediments of Boston Harbor since the cessation of sludge discharge," *Marine Environmental Research*, 51, 389–415.

Zeger, S. L. (1988), "A regression model for time series of counts," *Biometrika*, 75, 621–629.

Zeger, S. L. and Qaqish, B. (1998), "Markov regression models for time series – a quasi-likelihood approach," *Biometrics*, 44, 1019–1031.

Acknowledgements

This paper was supported by a grant from the Howard Hughes Medical Institute Predoctoral Fellowship Program and by National Institutes for Environmental Health Sciences grant ES05947 (Houseman), and by National Institutes for Environmental Health Sciences grant NIH/NIEHS ES05940 (Coull). The authors thank the editor, the associate editor, and three referees for insightful comments that improved our approach to the analysis of the data, and Kelly Coughlin (Massachusetts Water Resources Authority) for data and useful comments.

A Technical Details for Goodness-of-Fit Statistic

In this section we motivate the form of the goodness-of-fit statistic described by (10). Note that

$$\begin{aligned} K\log(\tau^2) - \frac{1}{2}\log\left|\mathbf{W}'\mathbf{\Omega}\mathbf{W} + \tau^{-2}\mathbf{I}_{2K}\right| &= K\log(\tau^2) - \frac{1}{2}\log\left(\tau^{-4K}\left|\tau^2\mathbf{W}'\mathbf{\Omega}\mathbf{W} + \mathbf{I}_{2K}\right|\right) \\ &= -\frac{1}{2}\log\left|\tau^2\mathbf{W}'\mathbf{\Omega}\mathbf{W} + \mathbf{I}_{2K}\right| \\ &\to 0 \text{ as } \tau^2 \to 0. \end{aligned}$$

Therefore, $\tilde{L}$ is the sum of the logarithms of (1) computed with $\mu_t(\mathbf{a}) = Y_t$ and $\mathbf{a} = \mathbf{0}$. When $\tau^2 > 0$ the corresponding log-likelihood is equal to the sum of the logarithms of (1) computed at the estimated values, plus the term $K\log(\tau^2) - \frac{1}{2}\log|\mathbf{W}'\mathbf{\Omega}\mathbf{W} + \tau^{-2}\mathbf{I}_{2K}|$. A similar statistic can be obtained for the Poisson distribution by substituting the probability mass function for the Poisson distribution for (1).

Note that, a priori, one could set the value of $\sigma = 0$ in the computation of $\tilde{L}$. This leads to a different statistic whose null distribution can also be obtained via simulation. The conclusions of Table 5 do not change with this modification.

Table 1: Summary of Enterococcus Data in Boston Harbor

	Jan. 1995 – July 1998			July 1998 – Sept. 2000			Sept. 2000 – Dec. 2002		
Station	n	% ND	med	n	% ND	med	n	% ND	med
44	72	68%	< 5	53	77%	< 5	67	70%	< 5
47	90	63%	< 5	47	79%	< 5	0		
48	65	77%	< 5	51	82%	< 5	52	73%	< 5
65	88	66%	< 5	64	66%	< 5	40	80%	< 5
77	97	71%	< 5	83	92%	< 5	83	83%	< 5
79	94	24%	32.5	47	68%	< 5	0		
80	88	34%	10	44	80%	< 5	0		
81	88	39%	10	46	85%	< 5	0		
82	93	43%	5	70	84%	< 5	82	84%	< 5
106	93	71%	< 5	82	82%	< 5	80	92%	< 5
116	73	45%	5	43	53%	< 5	0		
117	78	56%	< 5	44	80%	< 5	0		
118	86	77%	< 5	81	88%	< 5	17	88%	< 5
124	97	61%	< 5	80	86%	< 5	80	88%	< 5
129	87	72%	< 5	64	73%	< 5	0		
130	12	33%	12.5	113	69%	< 5	84	89%	< 5
135	81	73%	< 5	80	86%	< 5	17	88%	< 5
136	76	80%	< 5	60	92%	< 5	0		
139	104	54%	< 5	83	80%	< 5	84	76%	< 5
141	96	59%	< 5	79	95%	< 5	83	90%	< 5
142	88	74%	< 5	74	91%	< 5	78	94%	< 5
159	76	64%	< 5	84	64%	< 5	18	83%	< 5
160	80	68%	< 5	83	66%	< 5	80	94%	< 5

Enterococcus counts at various MWRA monitoring stations in Boston Harbor, measured in colonies per 100 mL. Sample size, percent of observations falling below the detection limit (5 colonies per 100 mL), and median count value are given for three time periods: January 5, 1995 through July 15, 1998; July 16, 1998 through September 6, 2000, and September 7, 2000 through December 19, 2002.

Table 2: Negative Binomial Regression Model Estimates for Enterococcus Counts at Selected Boston Harbor Stations

	Station 82					Station 139				
		Naive		Robust			Naive		Robust	
Covariate	Est.	SE	P value	SE	P value	Est.	SE	P value	SE	P value
(Intercept)	-1.60	1.47	0.2749	1.55	0.3024	0.93	1.58	0.5533	1.59	0.5564
Tide Height	-0.92	0.16	< 0.0001	0.17	< 0.0001	0.24	0.14	0.0895	0.14	0.0903
Res. Temperature	-0.39	0.12	0.0007	0.12	0.0010	-0.10	0.10	0.3274	0.10	0.3314
Res. Salinity	0.00	0.12	0.9748	0.10	0.9671	0.29	0.21	0.1654	0.21	0.1635
DITP overflow	7.11	2.19	0.0012	1.91	0.0002	5.87	2.30	0.0108	2.20	0.0076
NITP overflow	-6.95	6.29	0.2692	7.84	0.3758	-7.49	7.67	0.3290	8.53	0.3799
DITP base	-2.70	5.20	0.6039	5.41	0.6185	-9.47	5.61	0.0915	5.62	0.0920
NITP base	41.26	5.54	< 0.0001	5.98	< 0.0001	23.13	6.00	0.0001	6.27	0.0002
$\sin(2\pi t/T^*)$	0.08	0.33	0.8220	0.33	0.8189	0.16	0.32	0.6067	0.31	0.6014
$\cos(2\pi t/T^*)$	0.49	0.25	0.0475	0.25	0.0475	1.64	0.22	< 0.0001	0.22	< 0.0001
Mass Bay	0.25	0.50	0.6198	0.41	0.5466	0.68	0.41	0.0993	0.38	0.0721
Dispersion (σ)	3.4, 95% CI = $(2.5, 4.7)$					2.8, 95% CI = $(2.0, 3.8)$				
Smoothing (τ^2)	7.3×10^{-4}, df = 0.0					3.8×10^{-4}, df = 0.0				
Time indep σ	$P = 0.035$					$P = 0.161$				

	Station 159					Station 160				
		Naive		Robust			Naive		Robust	
Covariate	Est.	SE	P value	SE	P value	Est.	SE	P value	SE	P value
(Intercept)	-1.75	1.82	0.3359	1.61	0.2758	-2.29	1.77	0.1962	1.74	0.1881
Tide Height	0.46	0.19	0.0138	0.13	0.0004	0.18	0.19	0.3472	0.19	0.3454
Res. Temperature	0.18	0.14	0.2065	0.11	0.0876	0.43	0.16	0.0058	0.16	0.0071
Res. Salinity	-0.10	0.18	0.5766	0.13	0.4164	0.13	0.14	0.3470	0.13	0.3075
DITP overflow	6.49	2.37	0.0062	1.93	0.0008	8.07	2.50	0.0012	2.52	0.0014
NITP overflow	4.52	8.52	0.5958	7.93	0.5683	9.47	8.96	0.2906	8.08	0.2412
DITP base	6.60	6.35	0.2986	5.47	0.2281	10.10	6.21	0.1037	6.11	0.0987
NITP base	0.01	6.71	0.9992	6.16	0.9992	-5.18	6.90	0.4527	6.54	0.4277
$\sin(2\pi t/T)$	-0.35	0.37	0.3485	0.32	0.2724	-0.56	0.37	0.1279	0.37	0.1291
$\cos(2\pi t/T)$	-0.50	0.32	0.1169	0.22	0.0197	0.54	0.31	0.0810	0.31	0.0831
Mass Bay	-2.06	0.88	0.0198	0.65	0.0016	-3.52	0.61	< 0.0001	0.69	< 0.0001
Dispersion (σ)	5.4, 95% CI = $(3.8, 7.6)$					5.9, 95% CI = $(4.2, 8.3)$				
Smoothing (τ^2)	7.6×10^{-6}, df = 0.0					1.7×10^{1}, df = 5.0				
Time indep σ	$P = 0.059$					$P = 0.575$				

Negative binomial regression estimates for enterococcus counts in colonies per 20 ml. Stations 82 and 139 are near Nut Island Treatment Plant (NITP) and Stations 159 and 160 are near Deer Island Treatment Plant (DITP). Residual temperature is in ^{o}C units and residual salinity is in PSU units, each subtracted from a sinusoidal trend. Tide Height reflects tide height in meters, centered at the station mean. The overflow and base DITP and NITP effects, in 100 MGD units, are described in detail in Section 3. Time t represents days since December 31, 1994, and $T = 365.25$. The dispersion parameter σ is the variance of a Gamma mixing variable, shown with a 95% confidence interval. The smoothing parameter τ^2 is the variance of the normal random effect used to regularize fitting of a residual error process, corresponding to the nonlinear degrees-of-freedom labelled "df". The P value labelled "Time indep σ" summarizes the test for time-independent dispersion described in Section 2.5.

Table 3: Effect of Flow Transfers on Enterococcus Counts in Boston Harbor

	NITP - DITP Flow				
		Naive		Robust	
Station	Est.	SE	P value	SE	P value
44	38.1	8.0	< 0.0001	7.4	< 0.0001
47	2.6	4.9	0.6044	5.6	0.6481
48	28.8	7.7	0.0002	5.4	< 0.0001
65	7.2	4.6	0.1174	6.5	0.2679
77	27.2	5.2	< 0.0001	4.9	< 0.0001
79	29.5	4.2	< 0.0001	4.9	< 0.0001
80	24.1	3.9	< 0.0001	4.3	< 0.0001
81	41.3	5.7	< 0.0001	11.9	0.0005
82	44.0	4.7	< 0.0001	4.8	< 0.0001
106	4.5	5.1	0.3811	6.0	0.4549
116	3.5	3.7	0.3442	3.5	0.3178
117	18.9	4.8	< 0.0001	3.6	< 0.0001
118	-8.9	5.5	0.1068	5.6	0.1083
124	16.8	4.7	0.0003	5.4	0.0018
129	10.0	5.2	0.0546	6.8	0.1391
135	23.3	6.0	< 0.0001	4.3	< 0.0001
136	31.8	0.0	< 0.0001	12.8	0.0130
139	32.6	4.3	< 0.0001	4.4	< 0.0001
141	37.3	5.9	< 0.0001	9.0	< 0.0001
142	-18.6	9.3	0.0445	12.7	0.1423
159	-6.6	5.2	0.2073	3.6	0.0638
160	-15.3	5.5	0.0055	5.2	0.0031

Negative binomial regression estimates for the base flow effects of transferring flow from Nut Island Treatment Plant (NITP) to Deer Island Treatment Plant (DITP), calculated as *base NITP Flow - base DITP Flow*. Effects for 22 stations in Boston Harbor are given; Sufficient samples were not collected at Station 130 to estimate an NITP effect. See Figure 2 for station locations. Flow effects are presented in natural logarithms of colonies per 20 ml per 100 MGD flow. Estimates are adjusted for tide height, temperature, salinity, overflow effects, and a nonparametric residual time effect. When possible, estimates were also adjusted for the effect of offshore diffusers.

Table 4: Effect of Offshore Diffusers on Enterococcus Counts in Boston Harbor

	Offshore Mass Bay Diffuser Effect				
		Naive		Robust	
Station	Est.	SE	P value	SE	P value
44	1.3	0.5	0.0086	0.5	0.0050
48	1.1	0.5	0.0400	0.4	0.0080
65	0.6	0.5	0.2502	0.5	0.2196
77	0.7	0.5	0.1456	0.4	0.0970
82	0.2	0.5	0.6198	0.4	0.5466
106	-0.7	0.5	0.1762	0.5	0.1610
118	-0.6	0.9	0.5006	1.1	0.5634
124	0.2	0.5	0.7462	0.6	0.7886
130	-1.8	0.5	0.0005	0.7	0.0059
139	0.7	0.4	0.0993	0.4	0.0721
141	0.1	0.6	0.8567	1.0	0.9068
142	-1.0	0.8	0.2163	0.8	0.1994
159	-2.1	0.9	0.0198	0.7	0.0016
160	-3.5	0.6	< 0.0001	0.7	< 0.0001

Estimated effect of transferring wastewater flows from Boston Harbor to the offshore diffusers in Massachusetts Bay on September 6, 2000, at 14 stations in Boston Harbor. See Figure 2 for station locations. Only 14 of the 23 stations had sufficient data to estimate this effect. Estimates are adjusted for tide height, temperature, salinity, base and overflow treatment plant effects, and a nonparametric residual time effect.

Table 5: Goodness-of-Fit for Poisson and Poisson-Gamma Models for Enterococcus in Boston Harbor Stations 82 and 160

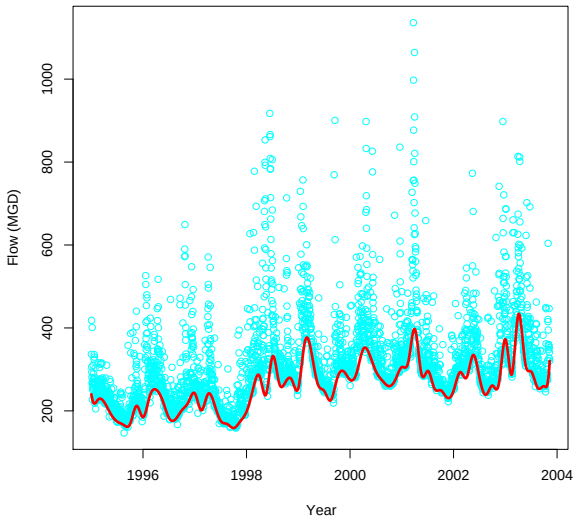
	Negative Binomial Station				Poisson Station			
	82	139	159	160	82	139	159*	160*
L.R. Statistic	170.3	191.1	129.2	151.5	3841.8	806.4	698.3	873.1
Degrees-of-Freedom	234.0	260.0	167.0	225.2	216.9	224.5	122.6	192.2
χ^2 P-value	> 0.99	> 0.99	0.99	> 0.99	< 0.01	< 0.01	< 0.01	< 0.01
Simulation P-value	0.24	0.67	0.51	0.54	< 0.01	< 0.01	< 0.01	< 0.01
Simulated 5% Crit. Val.	178.9	216.9	146.8	173.4	503.8	363.1	362.4	361.7
Simulated Mean Under H_0	161.6	197.6	128.4	151.4	457.9	332.3	337.0	333.9
Simulated SD Under H_0	11.1	11.6	11.8	13.5	26.4	18.0	17.1	20.1

Goodness-of-fit statistics for Poisson and Poisson-Gamma models described in Section 3. The likelihood ratio statistic is described in Section 2.6. Residual degrees-of-freedom are based on the formula described by Ruppert et al. (2003, Section 3.14); the corresponding P-value from the χ^2 test based on a normal approximation is shown, as well as the simulation-based P-value suggested by Ruppert et al. (2003, Section 4.8.2) and described in Section 2.6. For each case, the null distribution was estimated from 500 simulations; the corresponding mean and critical value for $\alpha = 0.05$ are given.

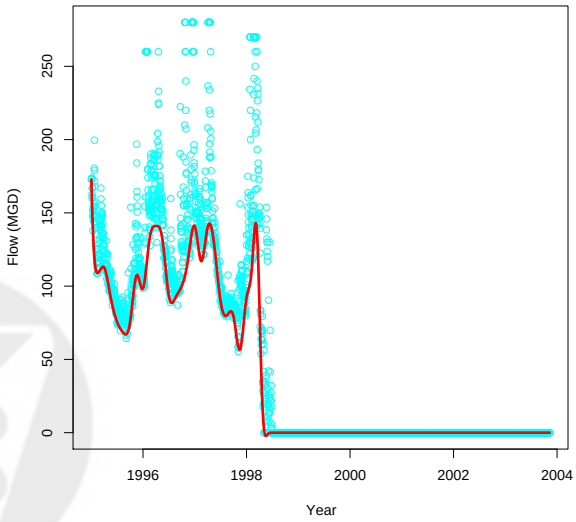
* For Poisson models, one extremely large LR statistic from Station 159 and two extremes from Station 160 were removed before calculating the simulation mean and standard deviation; however, they were included in the P value and critical value computation.

Figure 1: Time Series Plots for Plant Effluent Flows

A. Flow from Deer Island Treatment Plant

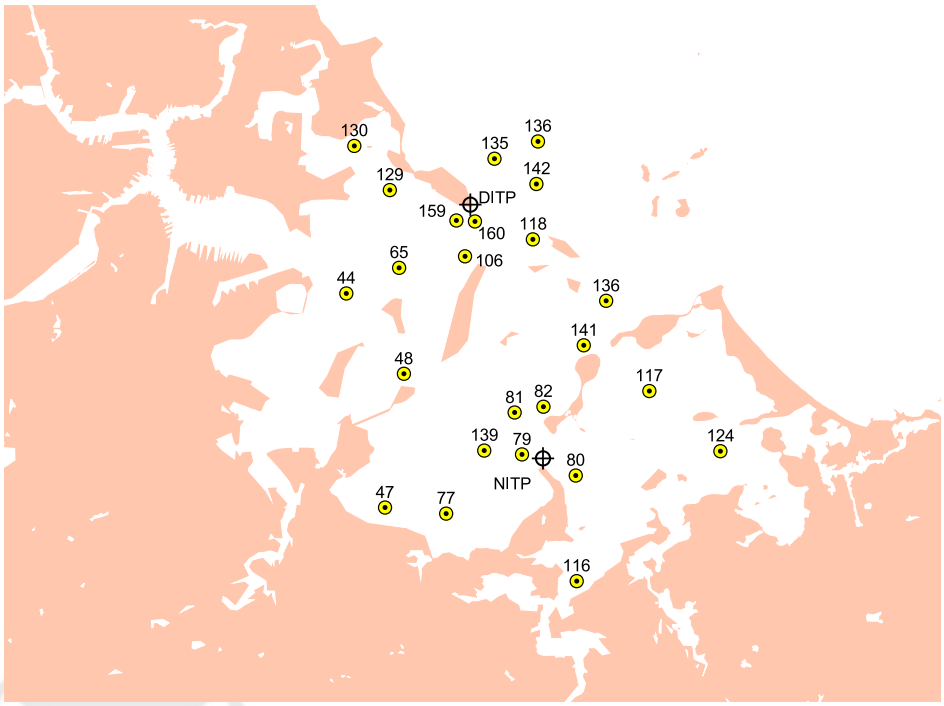


B. Flow from Nut Island Treatment Plant



Effluent flows from Nut Island Treatment Plant (A) and Deer Island Treatment Plant (B). Flows are in MGD units. The smooth solid line represents a smoothed "lower envelope", obtained by 5th percentile regression upon B-splines. 41 knots were used for NITP flow (at approximately 0.125 year intervals until 1999 and 1 year intervals thereafter), and 74 knots were used for DITP flow (at 0.125 year intevals).

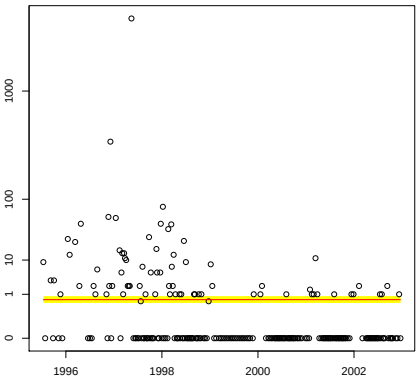
Figure 2: MWRA Sampling Stations



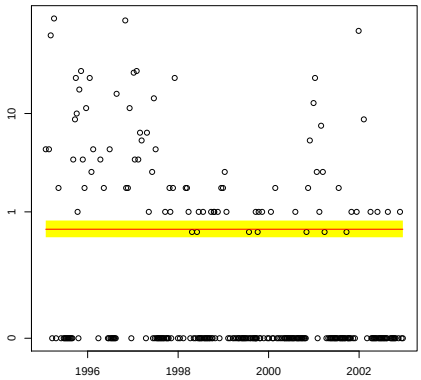
Massachusetts Water Resources Authority sampling stations in Boston Harbor.

Figure 3: Time Series Plots for Enterococcus Counts at Selected Boston Harbor Stations

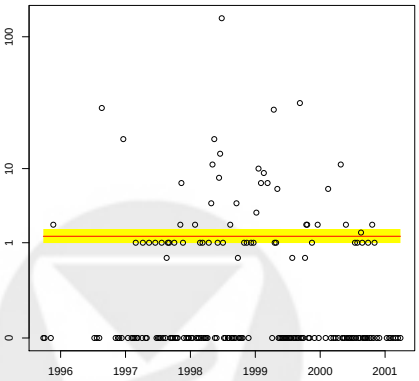
A. Station 82 (Near Nut Island Treatment Plant)



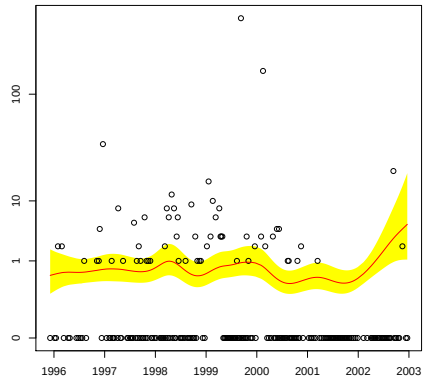
B. Station 139 (Near Nut Island Treatment Plant)



C. Station 159 (Near Deer Island Treatment Plant)



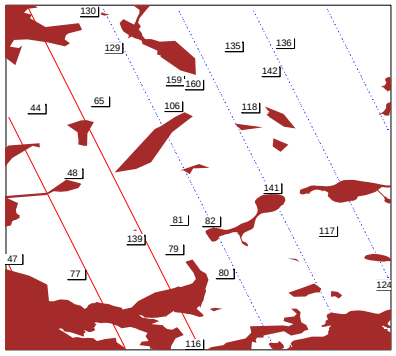
D. Station 160 (Near Deer Island Treatment Plant)



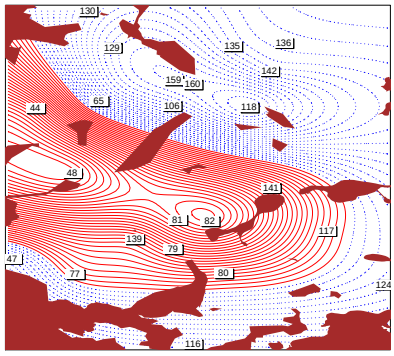
Estimated residual time effect $f(t)$ at Stations 82 and 139 (near NITP) and Stations 159 and 160 (near DITP). Shaded region represents an approximate 95% pointwise confidence band. For presentation, the physical scale of the vertical axis is 0.25 power.

Figure 4: Effects by Location

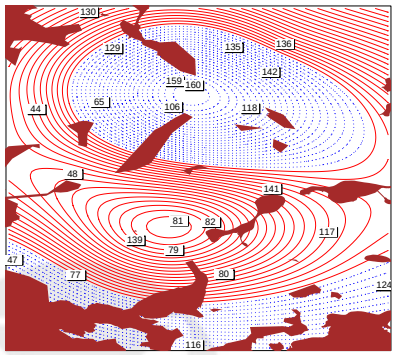
A. DITP Effect



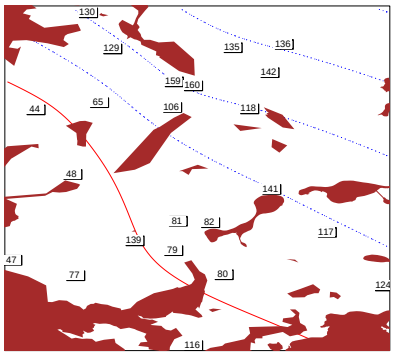
B. NITP Effect



C. Transfer Effect (NITP - DITP)



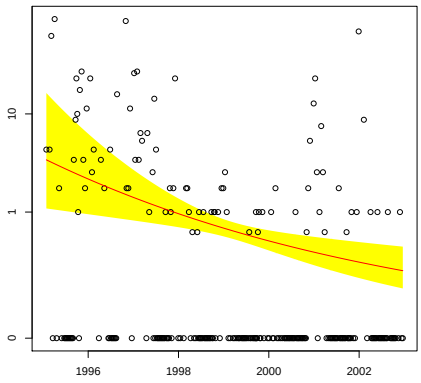
D. Massachusetts Bay Diffuser Effect



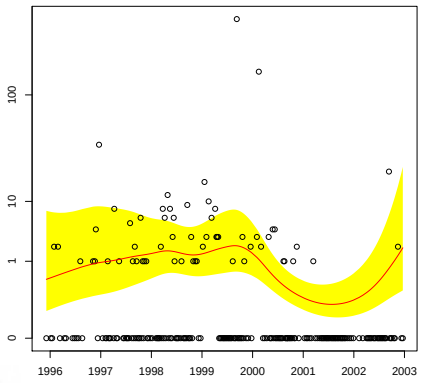
Geographic depiction of effects of interest, as described in Section 3. Each contour line represents a one unit difference in regression coefficient estimate, with units as in Table 2. Dark solid lines represent effects greater than the mean, light dashed lines represent effects less than the mean. The regression intercept, interpreted as the geographic mean effect, and corresponding 95% confidence interval were, respectively, (A) -1.3 $(-5.3, 2.7)$, (B) 15.9 $(13.6, 18.3)$, (C) 16.7 $(12.4, 21.0)$, and (D) -0.34 $(-0.76, 0.09)$. P values from chi-square tests for spatial homogeneity were, respectively, (A) 0.80, (B) 0.002, (C) 0.003, and (D) 0.018.

Figure 5: Examples of Penalized Regression Splines without Orthogonalization

Station 139



Station 160



Estimated residual time effect $f(t)$ at Station 139 (near NITP) and Station 160 (near DITP), where $f(t)$ was not constrained to lie in a space orthogonal to smooth covariates. Figure 3 shows the corresponding estimates for $f(t)$ orthogonal to the smooth covariates. For Station 139, much of the effect of NITP is absorbed into $f(t)$ due to collinearity of the function space with the time-dependent effects.