Environmental Science & Technology

Article

pubs.acs.org/est

Implications of Stein's Paradox for Environmental Standard Compliance Assessment

Song S. Qian,*[,†] Craig A. Stow,[‡] and YoonKyung Cha[§]

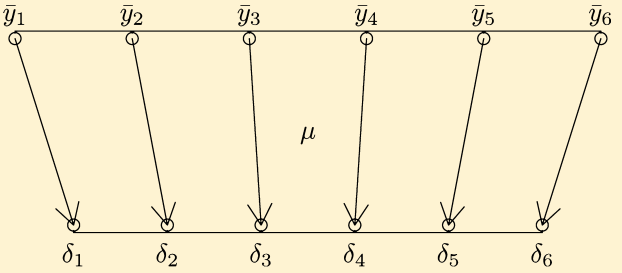
[†]Department of Environmental Sciences, The University of Toledo, Toledo, Ohio 43606, United States

[‡]Great Lakes Environmental Research Laboratory, National Oceanic and Atmospheric Administration, Ann Arbor, Michigan 48108, United States

[§]School of Natural Resources and Environment, University of Michigan, Ann Arbor, Michigan 48108, United States

Supporting Information

ABSTRACT: The implications of Stein's paradox stirred considerable debate in statistical circles when the concept was first introduced in the 1950s. The paradox arises when we are interested in estimating the means of several variables simultaneously. In this situation, the best estimator for an individual mean, the sample average, is no longer the best. Rather, a shrinkage estimator, which shrinks individual sample averages toward the overall average is shown to have improved overall accuracy. Although controversial at the time, the concept of shrinking toward overall average is now widely accepted as a good practice for improving statistical stability and reducing error, not only in simple estimation problems, but also in complicated modeling problems. However, the utility of Stein's insights are not widely recognized in the environmental management community, where mean pollutant concentrations of multiple waters are routinely estimated for management decision-making. In this essay, we introduce Stein's paradox and its modern generalization, the Bayesian hierarchical model, in the context of environmental standard compliance assessment. Using simulated data and nutrient monitoring data from wadeable streams around the Great Lakes, we show that a Bayesian hierarchical model can improve overall estimation accuracy, thereby improving our confidence in the assessment results, especially for standard compliance assessment of waters with small sample sizes.



INTRODUCTION

An important task of water quality management in the U.S. is compliance assessment as mandated by the Clean Water Act (CWA). Section 305(b) of the 2002 CWA amendment requires states to report the water quality status of all waters of the state (including rivers/stream, lakes, estuaries/oceans, and wetlands) every two years to support the U.S. Environmental Protection Agency's (EPA) strategy for achieving a broad-scale, national inventory of water quality conditions. States are further required to submit biennial reports with a list of impaired and threatened waters under Section 303(d) and an assessment of status and trends of significant publicly owned lakes under Section 314.

These assessments require states to quantify the levels of various pollutants for multiple waters and compare them to water quality criteria. Typically, these criteria are interpreted as mean values over some spatial and temporal domain.[1] As a result, various forms of sample averages are reported, typically one lake or river (or segment thereof) at a time. Given the large number of waterbodies in the U.S., this is a formidable task and the number of samples collected to evaluate compliance for any single waterbody is often relatively small. Results presented by Stein[2] and James and Stein,[3] commonly referred to as "Stein's Paradox" have useful implications for improving compliance assessment when many sites with small sample sizes are evaluated simultaneously.

Using simulated data and monitoring data from the U.S. Geological Survey's (USGS) National Water Quality Assessment program (NAWQA) study of wadeable streams,[4] we show that pooling data from multiple sites using a Bayesian hierarchical modeling approach will reduce estimation uncertainty, particularly for waters with small sample sizes. We start with an introduction to shrinkage estimators in the context of estimating mean nutrient concentrations for CWA compliance assessment, followed by detailed examples illustrating the implications of a shrinkage estimator. We end the paper with a general discussion on the use of the Bayesian hierarchical modeling approach beyond water quality compliance assessment.

SAMPLE AVERAGE AS AN ESTIMATE OF POPULATION MEAN

An important task of statistics is to estimate unobservable model parameters from observed data.[5] Because the parameter

Received: November 11, 2014
Accepted: April 13, 2015

DOI: 10.1021/acs.est.5b00656
Environ. Sci. Technol. XXXX, XXX, XXX−XXX

of interest cannot be observed directly, an estimator (a formula used to calculate an estimate) must be selected based on its performance, and assumptions about the distribution of the observed data. Assuming that data can be approximated by a normal distribution is prominent because of early works of Gauss and Laplace. Gauss derived the probability law, later known as the normal or Gaussian distribution, to justify the use of the least-squares method for estimating a mean. Laplace's central limit theorem,[6] which states that the distribution of sample averages of independent random variables can be approximated by the normal distribution regardless of the original distribution from which these random variables were drawn, cemented the importance of the normal distribution for statistical inference.

An important result of normal distribution theory is that the "best" estimator of the normal distribution mean is the sample average. It is the best because it is unbiased and, among all unbiased estimators, is least variable. Consequently, sample averages and standard errors are commonly reported in scientific studies. In environmental standard compliance assessment, these normal distribution properties helped justify the use of a hypothesis testing approach instead of the raw score assessment approach.[1,7]

For many environmental variables it is good practice to convert the observations to a log scale before making statistical inference so that properties of normal distributions can be used advantageously.[8,9] Concentration variables, in particular, are better approximated by a normal distribution after log-transformation, because they are bounded at zero and are usually right-skewed.

From the James−Stein Estimator to the Bayesian Hierarchical Model. Stein's paradox refers to the surprising features of a family of estimators originally introduced in the 1950s[2] and revised in 1961.[3] These estimators are paradoxical because they suggest that the best method for estimating the mean of one variable (calculating the sample average) is not the best approach when the means of several variables are to be estimated simultaneously. Specifically, James and Stein[3] showed that the overall accuracy (defined as the sum of squared differences between the estimated and true means) can be improved if we "shrink" the individually estimated averages toward the overall average—increasing those below and decreasing those above the overall average. In the context of nutrient criterion compliance assessment, Stein's paradox suggests that sample average is no longer the best estimator when estimating means of multiple waterbodies at the same time.

In the 1970s, Efron and Morris published a series of papers discussing the James−Stein estimator (JSE) (and its modifications) and its role in various estimation problems.[10−13] In their work, Efron and Morris used Bayes risk as a measure of estimation accuracy. Bayes risk is the average of the sum of squared differences:

$$R(\theta, \delta) = E_\theta \sum_{j=1}^{J} (\delta_j - \theta_j)^2 \tag{1}$$

where θ_j are unknown means (e.g., the true annual mean concentration of TP in a stream), δ_j is an estimator of θ_j (e.g., annual average of monthly monitoring data), and E_θ represents averaging over the distribution of θ_j. Bayes risk is often seen as the Bayesian version of the mean squared errors (MSE). A small Bayes risk is a good feature of an estimator. Efron and Morris showed that the Bayes risk of the JSE is always lower than the Bayes risk of the corresponding maximum likelihood estimator (MLE).

Efron and Morris further explored the relation between JSE and the estimator with the smallest Bayes risk−the Bayes estimator (BE). The connection can be easily represented in a normal data problem. Let y_{ij} be the i^{th} observed value (e.g., log TP concentration) from the j^{th} variable (e.g., one of the J rivers in a study) and $j = 1, \cdots, J$. We assume that the J variables are each normally distributed with a common variance, i.e.,

$$y_{ij} \sim N(\theta_j, \sigma_1^2) \tag{2}$$

where θ_j are the means we want to estimate and σ_1^2 is the variance of each of the J distributions. The assumption of a common variance is done here primarily for convenience of presentation. Distinct variances can be estimated if the common variance assumption seems inappropriate for any specific situation. The single variable MLE of θ_j is the sample average (average of log TP concentrations), i.e.,

$$\delta_j^{\text{mle}} = \bar{y}_j = \frac{1}{n_j} \sum_i y_{ij} \tag{3}$$

In Bayesian statistics, an unknown parameter is treated as random and a probability distribution is used to describe what we know about the parameter. Before we observe data, this distribution summarizes our knowledge about the parameter (the prior). A Bayesian analysis starts with a prior distribution of unknown parameters. Here, for the sake of illustration, we assume that σ_1 is known. Also, suppose that the rivers in our study are similar wadeable streams within the same ecoregion. From our prior experience, we may know the mean nutrient concentration of the region and the between stream variation. Such knowledge can be represented as the prior distribution of θ_j:

$$\theta_j \sim N(\mu, \sigma_2^2) \tag{4}$$

where μ is the mean of θ_j (the regional nutrient log mean) and σ_2^2 is the variance of θ_j (among stream variance). The prior distribution of eq 4 defines the location and spread of the J variable means. When the prior in eq 4 is known, the BE of θ_j is a weighted average of μ and δ_j^{mle}:

$$\delta_j^{\text{be}} = \mu + m_j^{\text{be}}(\delta_j^{\text{mle}} - \mu) \tag{5}$$

where $m_j^{\text{be}} = 1-((\sigma_1^2/n_j)/(\sigma_2^2+\sigma_1^2/n_j))$. The BE of eq 5, a shrinkage estimator, has the smallest Bayes risk among all estimators.[14]

While we must know μ and σ_2^2 to use the Bayes estimator, we can use the JSE of θ_j without such knowledge:

$$\delta_j^{\text{js}} = \mu + m_j^{\text{js}}(\delta_j^{\text{mle}} - \mu) \tag{6}$$

where μ, the mean of θ_j, is often estimated by the average of δ_j^{mle}, i.e., $\hat{\mu} = (1/J)\Sigma_j\delta_j^{\text{mle}}$, $m_j^{\text{js}} = 1 - (((J-2)\sigma_1^2/n_j)/S)$, and $S = \Sigma_j(\hat{\theta}_j - \hat{\mu})^2$. Judge and Bock[15] showed that m_j^{js} is an unbiased estimator of m_j^{be}. In other words, the JSE can be seen as deriving the prior from the data (an empirical Bayes approach). As the prior of eq 4 is often unavailable, the JSE is seen as the "next best thing".

Likewise, Bayesian hierarchical modeling (BHM) and multilevel modeling (MLM) are two alternative approaches for estimating m_j^{be} from data. BHM typically uses non-informative priors (priors with little information) on μ, σ_1^2,

DOI: 10.1021/acs.est.5b00656
Environ. Sci. Technol. XXXX, XXX, XXX−XXX

and σ_2^2 and often implemented using Markov chain Monte Carlo simulation.[16,17] MLM uses the maximum likelihood estimator of μ, σ_1^2, and σ_2^2 and typically based on approximate programs such as the ones implemented in the R package `lme4`.[18]

Intuitively, the improved estimation accuracy of a shrinkage estimator over the single variable MLE is achieved through the use of information on how, for example, nutrient mean concentrations of multiple streams are distributed—which are the overall average and the spread. Results from multivariate normal distribution suggest that a sample average tends to be farther away from the overall mean than the true mean does:

$$\Pr[\sum_j (\bar{y}_j - \mu)^2 > \sum_j (\theta_j - \mu)^2] > 0.5 \quad (7)$$

Consequently, shrinking a sample average toward the overall mean will improve the estimate. The question is then: shrinking by how much? Knowing the distribution of variable means represented in eq 4, we can answer this question. JSE, BHM, and MLM are three strategies of obtaining the same information from data.

Intuitively, eq 7 tells us that a sample average of the stream of interest is likely farther away from the overall mean than the true mean of the stream is. If the true nutrient concentration mean of the stream is larger than the overall mean, then the sample average is more likely to be larger than the true mean, and vice versa. As a result, shrinking the sample average toward the overall mean will improve our estimate. However, without the benefit of observing data from other streams in the region, we would not know which direction to shrink the sample average of the stream under study. Observing data from multiple streams provides us with information about the overall mean, thereby enables us to improve upon the sample average, which is otherwise the optimal estimator.

We note that the level of shrinkage (m_j) is largely determined by (1) the ratio of σ_1/σ_2 and (2) sample size n_j. A large standard deviation ratio (the standard deviation of individual variables, or within stream standard deviation, is large in comparison to the standard deviation among variable means, or among stream standard deviation) suggests a low confidence on the hypothesis that θ_j's are different. It leads to a small m_j^{be}, thereby a large level of shrinkage toward the overall mean. A small n_j (indicating a low confidence on δ_j^{mle}) leads to a small m_j^{be}, thereby a high level of shrinkage. In other words, using a shrinkage estimator is an effective way of addressing high uncertainty associated with a sample average estimated with a small sample size.

In comparing various shrinkage estimators, we measure the improvement of an estimator over the MLE using the ratio of Bayes risk of MLE over the Bayes risk of the estimator (the BR ratio). We note that the improvement is largely a function of how much we know about μ, σ_1^2, and σ_2^2. In practice, we use BHM and MLM because they are general purpose models and can be applied to many different types of problems.[17]

Bayesian Estimation of a Single Site Mean. The multivariate normal theory in eq 7 indicates the tendency of overdispersion of multiple sample averages. A shrinkage estimator corrects this tendency by shrinking sample averages toward the overall mean. The Bayes estimator (eq 5) has the lowest Bayes risk when the prior distribution of eq 4 is known. That is, when we know the distribution of multiple means we can benefit from the shrinkage effect without data from other sites—simply using the Bayes estimator. In our view, Stein's paradox gives a physical meaning of a prior distribution in the Bayes estimator—the distribution of multiple variable means. In the context of environmental standard compliance assessment, the prior distribution (eq 4) is the distribution of means from multiple sites. The prior mean is the average of all site means and the prior variance is the among-site variance of site means.

In many fields with extensive records (e.g., baseball), such priors are derived based on historical records (e.g., mean batting averages of all players in a league and the variance of these averages). In assessing water quality standard compliance, we can derive priors through the analysis of monitoring data from similar waters, as we will discuss in the next section. Once the prior is quantified (μ and σ_2^2 are estimated), we can directly use the Bayes estimator in eq 5 when estimation uncertainty on these two parameters is small. When estimation uncertainty of μ and σ_2^2 cannot be ignored, we can incorporate the uncertainty by using a proper prior distribution on μ and σ_2^2. For example, Qian and Reckhow[19] used the conjugate family of priors, where the joint distribution of μ and σ_2^2 is approximated by the normal-inverse gamma distribution, which summarizes the uncertainty in σ_2^2 using an inverse gamma distribution:

$$\sigma_2^2 \sim \text{Inv} - \text{gamma}(\alpha, \beta) \quad (8)$$

and the uncertainty in μ using a conditional normal distribution:

$$\mu|\sigma_2^2 \sim N(\mu_0, \sigma_2^2/n_0) \quad (9)$$

One advantage of using this prior is that the posterior distribution of the site mean θ_i is also a normal-inverse gamma distribution.[19] The estimated posterior site mean is $\delta_j^{be}=(n_j\, \bar{y}_j + n_0\mu_0)\ /(n + n_0)$. In other words, computation of the Bayes estimator is straightforward once the four parameters of the normal-inverse gamma distribution (α, β, μ_0, n_0) are quantified. In our example (next section), we use information derived from a BHM/MLM applied to monitoring data from multiple sites to quantify the prior distribution of μ and σ_2^2, following the steps outlined in Qian and Reckhow.[19]

EXAMPLES

We use a simulation study to illustrate the reduced overall estimation error and a real data example to demonstrate the practical implications in the compliance assessment process both for an individual waterbody and for a collection of multiple waterbodies.

Simulated Data—quantifying the Improvement of Shrinkage Estimators. *Bayes Risk.* A shrinkage estimator is known to have lower Bayes risk than the MLE. In the simulation example, we use the BR ratios to demonstrate the improvement. We illustrate the relative improvement of a shrinkage estimator over the MLE as a function of the within and among variable standard deviation ratio. A small standard deviation ratio suggests a large separation among variables, indicating a small value of information from other variables, and vice versa.

We draw data from a number of normal random variables to represent the situation of sampling from multiple streams. We assume that data for each stream are from a normal distribution with a known mean (θ_j) and known variance (σ_1^2). The means of different streams have a known prior distribution (i.e., μ and σ_2^2 of eq 4 are known). For each set of random variables, we use both the MLE and shrinkage estimators (JSE, BHM, and BE

DOI: 10.1021/acs.est.5b00656
Environ. Sci. Technol. XXXX, XXX, XXX−XXX

with a known prior) to estimate these group means. The estimated means (δ_j^{mle}, δ_j^{jse}, δ_j^{bhm}, δ_j^{be}) are then used to calculate the sum of squares statistics. For example, $SS^{mle} = \Sigma_j (\delta_j^{mle} - \theta_j)^2$. We repeat the process many times to represent repeated assessments over time and the average of the resulting SS's is an estimate of Bayes risk. We can interpret the Bayes risk as a mean squared error of estimating multiple means over many years. The smaller the Bayes risk, the smaller the chance of making a mistake in the assessment process.

In drawing the simulated data, we used five values of the standard deviation ratio ($r = \sigma_1/\sigma_2 = 0.25$, 0.5, 1, 2, and 4). For each value of the ratio, we draw $J = 50$ means from the standard normal distribution [$N(0,1)$] (i.e., $\mu = 0$, $\sigma_2 = 1$, and $\sigma_1 = r\sigma_2$). Each variable is then assigned a sample size randomly selected between 15 and 30. Random samples are then drawn for each variable and the 50 means are estimated by various estimators. With each set of random sample, we calculate one sum of squared errors for each estimator. Repeating this process 10 000 times, we have 10 000 sum of squares for each estimator and their mean is the estimated Bayes risk.

Compliance Assessment. Under the current EPA water quality assessment guidelines, states often use a binomial test, which is equivalent to testing whether the 90th percentile of the concentration distribution exceeds the water quality criterion.[7] The comparable process of using a shrinkage estimator is to estimate the predictive distributions of multiple sites. A site is out of compliance if the 90th percentile of its distribution exceeds the criterion. The multivariate statistics theorem (eq 7) suggests that a shrinkage estimator will outperform its MLE counterpart on average. As a result, the improvement in the compliance assessment cannot be easily shown using data from one-time sampling. We borrow analytic results of total phosphorus (TP) concentrations in Saginaw Bay[20] and designed a simulation to compare the compliance assessment using a shrinkage estimator and using the hypothesis testing method. The goal is to calculate the likelihood of wrongly concluding a site to be out of compliance when the site mean is below the criterion.

We draw random samples from a number of "sites" with known means ranging from 10 to 20. A simulated data set includes a number of data points (randomly drawn between 6 and 24) for each site from log-normal distributions with fixed log means ranging from log(10) to log(20), and a within site log standard deviation of 0.35 based on the Saginaw Bay TP data.[20] The binomial test is applied to the simulated data from each site to test whether the site is in compliance of the criterion of 15. A MLM is also used to estimate site means and its standard deviation, which in turn is used to calculate the 90th percentile. For each set of data, we record whether a site is declared to be out of compliance. Repeating this process 10 000 time, we record the fractions of times the two approaches declared noncompliance. These fractions are plotted against the known site means. When the site mean is below the criterion of 15, these fractions are the probability of wrongly declaring a site to be noncompliance.

Nutrient Monitoring Data in Streams near the Great Lakes. In the simulation study, we are able to calculate and compare Bayes risk because we know the true underlying variable means (θ_j). In analyzing real data, true means are always unknown. In the data example, we illustrate that the improvement of a shrinkage estimator (BHM) is manifested in a more stable estimate.

The data were from a USGS study,[4] including nutrient (total phosphorus and total nitrogen) concentrations from 64 sampling sites in wadeable streams surrounding the Great Lakes. The data were compiled to facilitate the development of regional nutrient criteria. We use the data to discuss the nutrient criterion compliance assessment process.

The sampling sites represent three broad nutrient ecoregions:[21] VI, cornbelt and northern Great Plains; VII, mostly glaciated dairy region, and VIII, nutrient poor largely glaciated upper Midwest and Northeast. These nutrient ecoregions are an aggregation of the Level III ecoregion of Omernik.[22] The data were sampled between 1993 and 2006. We set aside data from the only site in western New York for illustrating the Bayesian estimation process in the next section. We also removed sites without TP measurements. A total of 59 sampling sites are used for the analysis.

Using this example, we show that the reduced Bayes risk is manifested in the improved stability in the estimated means over time. We use both the MLE and BHM to estimate annual seasonal mean TP concentrations for each site (up to four seasonal means each year). We then compare the sum of squares (SS) statistics using the site average (over the entire period) as an estimate of the true site mean (sum of the squared differences between annual seasonal means and the estimated site mean). The sum of squares statistics is part of the Bayes risk calculation, measuring the variation of the annual seasonal averages around the long-term mean. We understand that TP concentrations in streams vary by season, and seasonal averages are not good estimates of a long-term site mean. The large differences among seasonal means are used to illustrate the shrinkage effect.

We use an additive model similar to an ANOVA analysis to represent sources of variation:

$$y_{ijkl} = \beta_0 + \beta_{1i} + \beta_{2j} + \beta_{3k} + \epsilon_{ijkl} \tag{10}$$

where y_{ijkl} is the l^{th} observed log TP concentration value (collected in the k^{th} season of the j^{th} year, from the i^{th} site), β_0 is the overall mean for all sites, years, and months, β_{1i} is the site effect (e.g., $\beta_{13} = -2$ indicates the mean of site 3 is 2 units below overall average), β_{2j} is the year effect (e.g., $\beta_{21} = 1.2$ indicates that the year 1 mean is 1.2 units above the overall average), β_{3k} is the seasonal effect ($\beta_{32} = 2$ represents that season 2 mean is 2 units above the overall average), and ϵ_{ijkl} is a normal random variable with mean 0 and a constant variance (σ_1^2) representing the measurement error and other random noise (within site variance).

We use the linear model (ANOVA) to obtain (unshrunk) estimates of all model coefficients. The BHM approach imposes common prior distributions for β_{1i}, β_{2j}, and β_{3k}:

$$\beta_{1i} \sim N(0, \sigma_{\beta_1}^2) \tag{11a}$$

$$\beta_{2j} \sim N(0, \sigma_{\beta_2}^2) \tag{11b}$$

$$\beta_{3k} \sim N(0, \sigma_{\beta_3}^2) \tag{11c}$$

where $\sigma_{\beta_1}^2$ is the between site variance, $\sigma_{\beta_2}^2$ is the between year variance, and $\sigma_{\beta_3}^2$ is the between season variance. We use the R function `lmer` for the shrinkage (BHM/MLM) estimator. In fitting the MLM, we nested sites in Level III ecoregions.

Bayesian Estimation of the Mean from a Single Site. We use the data from the set-aside site from New York to

DOI: 10.1021/acs.est.5b00656
Environ. Sci. Technol. XXXX, XXX, XXX−XXX

illustrate the Bayesian estimation process, which results in predictive distributions of TP concentrations and site mean concentration. The focus is the use of the posterior predictive distribution in the compliance assessment process. Specifically, we derive an informative prior as in eq 4 using the MLM/BHM model in eq (11) fit to data from all sites except the New York one. The MLM output provides point estimates of the two standard deviations (σ_1 and σ_2 in eqs 2 and 4) and the overall mean (μ). The point estimate of β_0 is used as the prior mean μ_0 (eq 9). The prior sample size n_0 is set as the sample size in ecoregion 83, where the New York site is located. The point estimate for $\sigma^2_{\beta_1}$ represents the prior mean of the between site variance. Based on eq 8, the prior mean of the between site variance is $\beta/(\alpha - 1)$. To quantify the prior of σ_2^2 (estimating α and β) we need one more piece of information. We chose $\alpha = n_0/2$ such that, information in the prior for σ_2^2 is more or less equally represented in the posterior.[19]

Data from the set-aside site are then used to derive predictive distribution of TP concentrations one year at a time to compare with the assessment results using the current compliance assessment method (using hypothesis testing).

RESULTS

Simulation Study—Bayes Risk. We used a Monte Carlo simulation algorithm to calculate the Bayes risks of the MLE and the three shrinkage estimators (JSE, BHM, and BE). Using the MLE as a basis for comparison, we present the Bayes risk ratios—the Bayes risk of MLE (BR_{mle}) over the Bayes risk of the shrinkage estimators: BR_{mle}/BR_{bhm}, BR_{mle}/BR_{jse}, and BR_{mle}/BR_{be}—as measures of the improvement. If a risk ratio is larger than 1, then the MLE has a higher Bayes risk than the shrinkage estimator. In our simulation, the Bayes risk ratio is always larger than 1 (Figure 1), indicating that shrinkage estimators are always better than MLE.

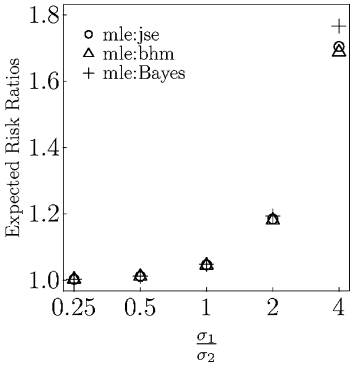


Figure 1. Simulated risk ratios compare the estimated Bayes risk of MLE to three shrinkage estimators at five different σ_1/σ_2 ratios. All risk ratios are larger than 1. A small σ_1/σ_2 ratio represents a large difference in group means, and vice versa.

The level of improvement is a function of the standard deviation ratio. When the standard deviation ratio is small (0.25 and 0.5), the BR ratios are very close to 1, indicating little shrinkage benefit. When the standard deviation ratio is large (1 or larger), the BR ratios are substantially higher than 1. In calculating the JSE, we set $\mu = 0$ to contrast with BHM where μ is estimated from data. The Bayes risk ratio BR_{mle}/BR_{bhm} is always slightly lower than the ratio BR_{mle}/BR_{jse} (Figure 1), suggesting that using information about μ is advantageous.

Compliance Assessment. Because we are comparing the 90th percentile of the TP distribution to the criterion, we expect that both processes will overstate the likelihood of noncompliance. However, the hypothesis testing method has a consistently higher probability of wrongly declaring a site to be noncompliance (Figure 2).

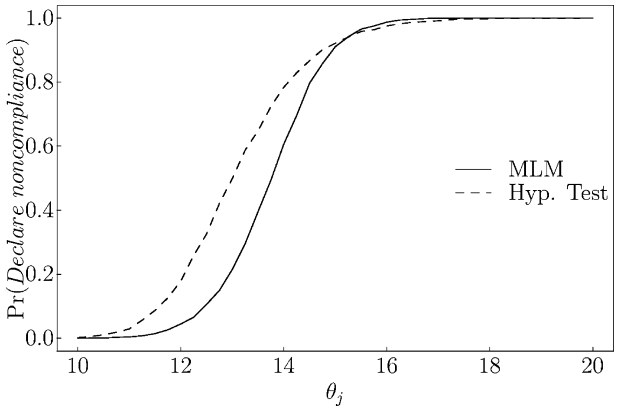


Figure 2. Simulation estimated probabilities of declaring a site as noncompliant are plotted against the known site mean concentrations. The hypothetical water quality criterion is 15.

Nutrient Monitoring Data in Streams near the Great Lakes. The MLE of annual seasonal means is based on the linear (ANOVA) model fitted using the R function `lm`. The shrinkage estimates are based on BHM implemented in the R function `lmer`. These models are then used to predict annual seasonal means for each site. Site means are approximated using site-average log TP concentrations (with site-specific sample sizes ranging from 6 to 288).

We compare the estimated annual seasonal means between the two models for the site with the largest site-specific sample size (site 50−66, $n_j = 288$). The two models resulted in slightly different estimates of the site mean (Figure 3). For the linear model (MLE), seasons with small sample sizes tend to result in estimates with large estimation uncertainty. When using the BHM, all the MLE estimated seasonal means were pulled toward the estimated site mean. As the shrinkage factor (m_j) is a function of season-specific sample size and the standard deviation ratio (within divided among season), we observe large shrinkage for seasons with small sample sizes and seasons with averages far away from the site mean. We note that the shrinkage estimator improves our confidence on estimates (smaller standard errors) with small sample sizes.

The shrinkage effect of BHM resulted in smaller SS statistics in 54 (91.5%) of the 59 sampling sites (Figure 4). The lowest SS from MLE is higher than the 75th percentile of the SS from BHM. In other words, using BHM we are far less likely have large year to year swings in estimated seasonal averages.

Bayesian Estimation of a Single Site Mean. Of the 27 observations from the single site in New York observed in 1996 ($n = 9$), 1997 (15), and 1998 (3), 15 of them (6 in 1996 and 9 in 1997) are below the method reporting limit (MRL). The numbers of observed TP concentrations exceeding the EPA recommended TP criterion of 0.02413 mg/L[23] are 3, 5, and 2, respectively, for years 1996, 1997, and 1998. If these observations are used for TP criterion compliance assessment using the binomial test with a null hypothesis of the river being in compliance (probability of exceeding the criterion being 0.1

DOI: 10.1021/acs.est.5b00656
Environ. Sci. Technol. XXXX, XXX, XXX−XXX

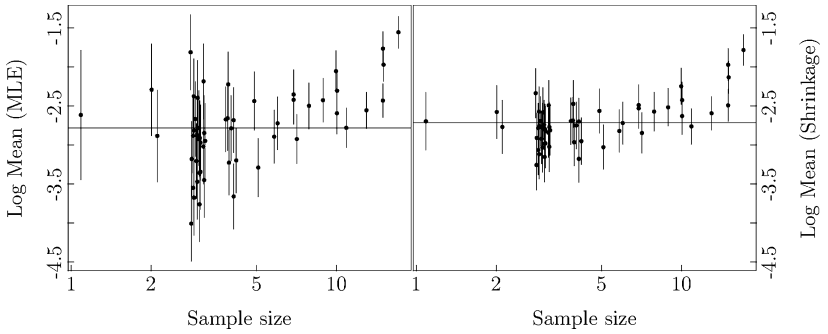


Figure 3. Estimated annual seasonal means using the linear model (left panel) are compared to the same using the multilevel model (right panel). The comparison illustrates the shrinkage effect as a function of sample size.

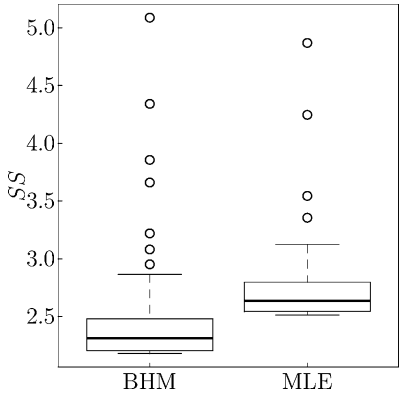


Figure 4. Boxplots of the SS statistics calculated from individual sampling sites using BHM and MLE are compared.

or less),[7] then we would conclude that the river is in compliance in 1996, but out of compliance in the subsequent two year (with p-values of 0.053, 0.013, and 0.028, respectively). We note that the three p-values are fairly close to each other and all close to the customary cut of point of 0.05. A small change in sample size can result in a change in the assessment outcome.

Alternatively, we can estimate the distribution of TP concentration using data from this site, assuming that TP can be adequately modeled by the log-normal distribution, with log-mean θ_{ny} and log-variance σ_1^2. However, because of the high level of data censorship, the commonly used substitution method (e.g., replacing the censored values with half of the MRL) is inadvisible. Instead, we can use a Bayesian estimation method with noninformative prior ($\pi(\theta_{ny}, \sigma_1^2) \propto 1/\sigma_1^2$). However, we were unable to estimate these parameters by year because of the small sample sizes and relatively large number of observations below MRL. Combining data from three years, the estimated log mean and log standard deviation are −5.24 and 3.21, respectively. On the basis of this distribution, the probability of individual TP concentrations exceeding the TP criterion is 0.32. In other words, the site is out of compliance using the current assessment method (probability of exceeding the criterion is larger than 0.1).

In order to use BE, we first derive prior distributions of parameters in eq 4 (μ, σ_1^2, and σ_2^2)—using the BHM model reported mean for Ecoregion 83 to derive a prior for μ, the BHM estimated between site variance to derive a prior for σ_2^2, and the BHM estimated model residual variance to derive a prior for σ_1^2. On the basis of data from all three years, the estimated mean of the posterior distribution of θ_{ny} (the log mean) is −4.00, and the mean of posterior distribution of σ_1 (within site log standard deviation) is 1.41. Using the predictive distribution of TP concentration, we estimated that the probability of TP concentrations exceeding the criterion is 0.41.

With an informative prior, we are able to estimate these parameters for each of the three years (posterior means of θ_{ny} are −3.97, −3.81, and −3.02, and posterior mean of σ_1 are 1.07, 1.33, and 0.79, for 1996, 97, and 98, respectively). The estimated probabilities of exceeding the criterion for these three years are 0.41, 0.47, and 0.77, respectively.

The posterior means (estimated using data from each of the three years and from all three years) lie in between the prior mean and the data average (Figure 5), as expected from a shrinkage estimator. In Figure 5, we contrasted the predictive distribution of TP concentration based on data from all three years (the dark shaded histogram) to the predictive distribution

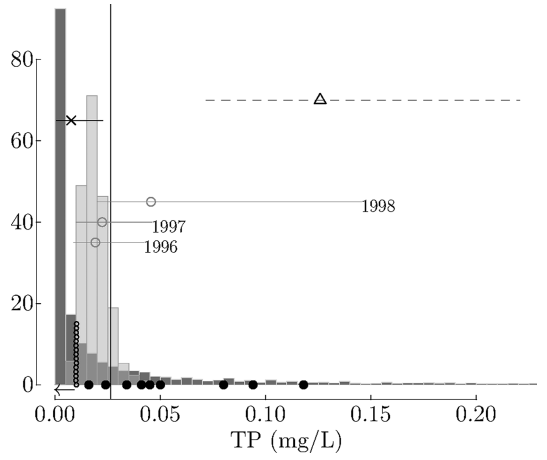


Figure 5. Posterior (based on all three years data) predictive distributions of site mean (light-shaded histogram) and individual TP values (dark-shaded histogram) are compared to the prior mean distribution (the horizontal dashed-line shows the 95% credible interval and the triangle is the mean) derived from similar streams in the same Ecoregion. The short horizontal solid line (near top-left corner) shows the 95% credible interval of the estimated site mean distribution using data from the site only, and the estimated mean is shown by the cross (×). The three gray horizontal lines show the estimated 95% credible intervals of the posterior predictive distributions of site mean using data from the three years separately (with estimated means shown in open circles). The vertical line is the EPA recommended TP criterion. The solid black dots at the bottom are the observed TP values (three values above 0.2 mg/L not shown) and the black open dots are TP values below MRL (shown at the MRL).

DOI: 10.1021/acs.est.5b00656
Environ. Sci. Technol. XXXX, XXX, XXX−XXX

of site mean (the light shaded histogram based on data from all three years, and the 95% credible intervals of the predictive site mean distributions based on data from each of the three years shown in shaded horizontal lines). In this case, the difference between the data distribution (the black horizontal line) and the prior distribution (dashed line) is large. Given a large between site variance (shown by the wide 95% credible interval of the prior), the Bayesian estimates are fairly stable with a moderate sample size. When the sample size is small (e.g., 1998 had three observations: 0.016, 0.034, and 0.045 mg/L), we see a large shrinkage even though the within site variability for 1998 is low (log standard deviation of 0.53, compared to the previous two years: log standard deviations 1.32 in 1996 and 1.52 in 1997).

DISCUSSION

The accuracy of a statistical estimator is measured by its bias and precision (variation).[24] Instead of balancing the trade-off between bias and precision, we traditionally emphasize bias over precision, and often require estimators to be unbiased. The preference of unbiasedness is, however, often misguided. Given the multivariate result in eq 7, we know that "deliberately induced biases" (shrinking the unbiased sample averages toward the overall mean) "can drastically improve estimation properties".[10] These improved properties can lead to improved environmental decision-making. For example, we reduced estimation uncertainty by using the BHM (Figure 3), especially when sample sizes are small. The reduced uncertainty will translate to fewer compliance assessment errors (Figure 2). Given the large number of waters in the U.S. assessed each year, using a shrinkage estimator for water quality standard compliance assessment can have a significant impact.

Stein's paradox and the James-Stein estimator have experienced rigorous debates in statistical articles.[12] In the end, statisticians agree on the merits of the shrinkage class of estimator in many situations. Intuitively, the difference between shrinkage estimators and nonshrinkage estimators lies in how we treat the information that the multiple variables in question are related. In MLE, we ignore this information by treating these variables as independent of each other. In a shrinkage estimator, we make use of this information. For example, eq 4 summarizes this information—θ_j's are both different (they are random variables) and related (share the same prior distribution), but we do not know the nature of the difference. Although this information is vague, we nevertheless can use it to improve the estimation. We note that because shrinkage estimators are biased, commonly used residual-based model diagnostic methods may be misleading. Rather, we should treat a BHM as an improved version of the corresponding MLE and assessing the adequacy of a proposed model form before pooling data.[25]

An early point of contention about the James−Stein estimator was on the relevancy of the mean of one variable on the estimation of another variable mean. An often used example of the James−Stein estimator is the estimation of batting averages of baseball players.[26] The question is why the batting average of the best player is relevant when we are estimating the batting average of a rookie. Efron[27] showed that sample averages ($\bar{y}_j$) tend to be farther away from the origin than the underlying means (θ_j) do (eq 7), or in other words, as surrogates for the unknown means sample averages tend to be too far away from the center. As a result, we can derive an improved estimator by shrinking sample averages toward the overall mean. However, this result is based on the assumption that these multiple variables are from a multivariate normal distribution. Consequently, the most consistent improvement occurs when these variables are related and distributed similarly. We used ecoregion as a grouping factor because nutrient criteria were recommended by ecoregion.

Ideally, the grouping should be based on the concept of exchangeable means. Equation 4 is also known as the exchangeability assumption. Intuitively, when we say the means θ_j are exchangeable, we mean that they are different, but we are ignorant about the nature of the difference. Deciding whether waters in a region are exchangeable requires a careful examination of all sources of information. Taking full advantage of available information is always a good practice in environmental management.

The exchangeability assumption can also be applied in statistical modeling. For example, Qian et al.[25] applied this assumption to a series of linear regression model coefficients. These linear regression models describe stream ecosystem's response to watershed urbanization in nine regions in the U.S. The regional difference in regression coefficients was shown to be the result of regional differences in climate and land use history. Qian and Cuffney[28] used a zero-inflated Poisson model to study macroinvertebrate taxon response to urban stressor. They imposed the exchangeable assumption on model coefficients of a group of taxa representing the pollution sensitive mayfly, stonefly, and caddisfly community. We can apply the same concept to the multiple linear regression models used by Dodds and Oakes[29] for deriving reference nutrient concentrations.

Our discussion of the Stein's paradox suggests that BHM be applied in a wide range of applications to reap the benefit of improved accuracy and numerical stability. BHM was used for computational necessity[30,31] (necessary for handling large amount of censored data) and for assessing multiple sites simultaneously.[20] BHM is naturally suited for assessing water quality criteria compliance of multiple waters, where pooling data from similar waters is always advantageous.

In the U.S., efforts have been made in developing state level nutrient criteria for rivers/streams, lakes, wetlands, and estuaries, and EPA is advocating the development of nutrient ecoregion based nutrient criteria. Such effort should improve the relevancy of the resulting criteria. Furthermore, pooling data from similar sites will allow us to improve assessment confidence for sites with small sample sizes (Figure 3).

Yuan et al.[32] pooled data nationwide to develop a nitrogen threshold for predicting harmful algal bloom in U.S. lakes. The connection between a Bayes estimator and a shrinkage estimator suggests that such effort can be used to improve local or regional scale risk assessment, because the advantage of a Bayesian modeling approach lies in the use of proper prior information to reduce Bayes risk. As a result, developing national or regional prior distributions for different types of waters will help water quality compliance assessment at a local level. The resulting regional priors can be updated to incorporate newly acquired data in an adaptive management framework.[33−35]

ASSOCIATED CONTENT

Supporting Information

R setup; simulation data; Great Lakes data; and Bayesian Estimator for single site. This material is available free of charge via the Internet at http://pubs.acs.org.

DOI: 10.1021/acs.est.5b00656
Environ. Sci. Technol. XXXX, XXX, XXX−XXX

AUTHOR INFORMATION

Corresponding Author

*E-mail: song.qian@utoledo.edu.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

S.S.Q.'s work was partly supported by U.S. EPA and University of Michigan Water Center. We thank George Arhonditsis and Yong Cao, Thomas F. Cuffney, Wesley Daniel, Yin-Phan Tsang, and Lester Yuan for their helpful comments and discussion. Jeffrey W. Frey and Amanda H. Bell kindly provided the wadeable stream data. Comments from three reviewers and the associate editor are greatly appreciated. GLERL contribution number 1758.

REFERENCES

(1) Gibbons, R. A statistical approach for performing water quality impairment assessment. *J. Am. Water Resour. Assoc.* **2003**, *39*, 841−849.

(2) Stein, C. Inadmissibility of the usual estimator for the mean of a multivariate normal distribution Proceedings of the third Berkeley Symposium **1955**; pp 197−206.

(3) James, W.; Stein, C. *Proceedings of the 4th Berkeley Symposium Mathematics Statistics and Probability*; University of California Press: Berkeley, CA, Vol. *1*, 1961; pp 361−379.

(4) Frey, J., Bell, A., Hambrook-Berkman, J., Lorenz, D. *Assessment of nutrient enrichment by ese of algal-, invertebrate-, and fish-community attributions in wadeable streams in ecoregions surrounding the Great Lakes*; Scientific Investigations Report 2011−5009; National Water-Quality Assessment Program, U.S. Geological Survey: Reston, VA, 2011.

(5) Fisher, R. On the mathematical foundations of theoretical statistics. *Philos. Trans. R. Soc. London, Ser. A* **1922**, *222*, 309−368.

(6) Stigler, S. Napoleonic statistics: the work of Laplace. *Biometrika* **1975**, *62*, 503−517.

(7) Smith, E.; Ye, K.; Hughes, C.; Shabman, L. Statistical assessment of violations of water quality standards under section 303(d) of the Clean Water Act. *Environ. Sci. Technol.* **2001**, *35*, 606−612.

(8) Ott, W. *Environmental Statistics and Data Analysis*; Lewis Publishers: Boca Raton, FL, 1995.

(9) van Belle, G. *Statistical Rules of Thumb*, 2nd ed.; Wiley: New York, 2002.

(10) Efron, B. Biased versus unbiased estimation. *Advances in Mathematics* **1975**, *16*, 259−277.

(11) Efron, B.; Morris, C. Stein's estimation rule and its competitors—an empirical Bayes approach. *J. Am. Stat. Assoc.* **1973**, *68*, 117−130.

(12) Efron, B.; Morris, C. Combining possibly related estimation problems. *J. R. Stat. Soc. Ser. B (Method.)* **1973**, *35*, 379−421.

(13) Efron, B.; Morris, C. Data analysis using Stein's estimator and its generalizations. *J. Am. Stat. Assoc.* **1975**, *70*, 311−319.

(14) Lehmann, E., Casella, G. *Theory of Point Estimation*, 2nd ed.; Springer: New York, 1998.

(15) Judge, G., Bock, M. *The Statistical Implications of Pre-Test and Stein-Rule Estimators in Econometrics*; North-Holland: Amsterdam, 1978.

(16) Gelman, A.; Carlin, J.; Stern, H.; Dunson, D. B.; Vehtari, A.; Rubin, D. *BayesianData Analysis*, 3rd ed.; CRC Press: Boca Raton, FL, 2014.

(17) Gelman, A., Hill, J. *Data Analysis Using Regression and Multilevel/Hierarchical Models*; Cambridge University Press: New York, 2007.

(18) Bates, D. *lme4: Mixed-Effects Modeling with R*; Springer: New York, 2010.

(19) Qian, S.; Reckhow, K. H. Combining model results and monitoring data for water quality assessment. *Environ. Sci. Technol.* **2007**, *41*, 5008−5013.

(20) Stow, C.; Cha, Y.; Qian, S. A Bayesian hierarchical model to guide development and evaluation of substance objectives under the 2012 Great Lakes Water Quality Agreement. *J. Great Lakes Res.* **2014**, *40* (Supplement 3), 49−55.

(21) Potapova, M.; Charles, D. Diatom metrics for monitoring eutrophication in rivers of the United States. *Ecol. Indic.* **2007**, *7*, 48−70.

(22) Omernik, J. Ecoregion of the conterminous United States, Map (scale 1:7 500 000). *Ann. Assoc. Am. Geogr.* **1987**, *77*, 118−125.

(23) U.S. EPA, *Ambient Water Quality Criteria Recommendations—Information Supporting the Developlment of State and Tribal Nutrient Criteria—Rivers and Streams in Nutrient Ecoregion VII*; EPA 822-B-00−018; U.S. EPA Office of Water: Washington, CD.C., 2000.

(24) Gilbert, R. O. *Statistical Methods for Environmental Pollution Monitoring*; Van Nostrand Reinhold: New York, 1987.

(25) Qian, S.; Cuffney, T.; Alameddine, I.; McMahon, G.; Reckhow, K. On the Application of Multilevel Modeling in Environmental and Ecological Studies. *Ecology* **2010**, *91*, 355−361.

(26) Efron, B.; Morris, C. Stein's Paradox in Statistics. *Sci. Am.* **1977**, *236*, 119−127.

(27) Efron, B. Controversies in the foundations of statistics. *Am. Math. Monthly* **1978**, *85*, 231−246.

(28) Qian, S.; Cuffney, T. A hierarchical zero-inflated model for species compositional data - from individual taxon responses to community response. *Limnol. Oceanogr.: Methods* **2014**, *12*, 498−506.

(29) Dodds, W.; Oakes, R. A technique for establishing reference nutrient concentrations across watersheds affected by humans. *Limnol. Oceanogr.: Methods* **2004**, *2*, 333−341.

(30) Qian, S.; Schulman, A.; Koplos, J.; Kotros, A.; Kellar, P. A hierarchical modeling approach for estimating national distributions of chemicals in public drinking water systems. *Environ. Sci. Technol.* **2004**, *38*, 1176−1182.

(31) Wu, R.; Qian, S.; Hao, F.; Cheng, H.; Zhu, D.; Zhang, J. Modeling contaminant concentration distributions in China's centralized source waters. *Environ. Sci. Technol.* **2011**, *45*, 6041−6048.

(32) Yuan, L.; Pollard, A.; Pather, S.; Oliver, J.; D'anglada, L. Managing microcyctin: identifying national-scale thresholds for total nitrogen and chlorophyll a. *Freshwater Biol.* **2014**, *59*, 1970−1981.

(33) Walters, C.; Holling, C. Large-scale management experiments and learning by doing. *Ecology* **1990**, *71*, 2060−2068.

(34) Williams, B. Adaptive management of natural resources—framework and issues. *J. Environ. Manage.* **2011**, *92*, 1346−1353.

(35) Rist, L.; Campbell, B.; Frost, P. Adaptive management: where are we now? *Environ. Conserv.* **2012**, *40*, 5−18.

DOI: 10.1021/acs.est.5b00656
Environ. Sci. Technol. XXXX, XXX, XXX−XXX