A Bayesian approach to the assessment of contaminant spread Part II. Continuous case[1]

Tommy Norberg
Department of Mathematical Statistics
Chalmers University of Technology
and Göteborg University

21 February, 2002

Abstract

The paper provides a Bayesian method for estimating totals such as the polluted area or cost of remediation of a possibly contaminated hazardous waste site. After specifying a prior distribution on the total polluted area (here expert opinion may be taken into account), its focus is on how to chose the number of measurements to make in order to achieve a specified accuracy goal. This paper treats the continuous case. An accompanying paper [1] treats the discrete case in which the site is partitioned into a finite number of remediation units.

Key words: *beta-binomial distribution, binomial measurement model, pre-posterior analysis, cost distribution, value at risk, expected shortfall.*

[1]Preprint no 2002:10, `http://www.math.chalmers.se/Stat/Research/Preprints/`

1 Introduction

Let R denote a possibly contaminated area and for a point $p = (p_x, p_y) \in R$, let $f_{\mathrm{COC}}(p)$ be the concentration of some contaminant of concern (COC). Given is an action level a_{COC} such that any point $p \in R$ is considered polluted if $f_{\mathrm{COC}}(p) > a_{\mathrm{COC}}$ and not polluted otherwise. The purpose of this paper is to provide a simple method that accurately estimates the polluted proportion

$$\theta_{\mathrm{COC}} = \frac{|\{p \in R : f_{\mathrm{COC}}(p) > a_{\mathrm{COC}}\}|}{|R|}$$

where $|\cdot|$ denotes Lebesgue measure (i e, the area of). Since this paper only treats the univariate case with one COC hereafter all references to it will be dropped in the notation. Thus, instead of θ_{COC} we will henceforth write θ and so on.

Denote by $\mathrm{Beta}(\alpha, \beta)$ the beta distribution, the density of which is

$$p(\theta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\, \theta^{\alpha-1}\,(1-\theta)^{\beta-1}$$

for $0 < \theta < 1$ and $\alpha, \beta > 0$ (Γ denotes the Euler gamma function, cf Appendix A of [1]).

The approach of this paper is Bayesian and it is based on the well known fact that if θ prior to doing any measurement is $\mathrm{Beta}(\alpha, \beta)$-distributed, and a random sample of n points are investigated in order to determine the number X of polluted points, making X binomial with parameters n and θ, then the posterior distribution of θ, given $X = x$, is $\mathrm{Beta}(\alpha + x, \beta + n - x)$. Its focus is on providing means for appropriately selecting the number n of randomly sampled points to analyse.

The paper is a continuous case companion to Norberg [1], which treats the discrete case. Most derivations and results are identical or similar to the corresponding ones in [1]. Therefore this paper is densely written and the readers who wants to see details of derivations are referred to [1].

It may be argued that assessment of the polluted area is best handled by Kriging methods. However, it may also be argued that the intrinsic hypothesis (see e g Wackernagel [6, p 36]) is questionable for many hazardous man made waste sites. We do regard the Bayesian method of this paper as a robust and simple technique for assessing totals such as the polluted area or the total remediation cost, which is preferable to use at least in the early stages of an investigation of a possibly contaminated area.

The mathematical essence of the method is written down in Section 2. It consists mainly of some formulae for how to calculate pre-posterior (or predictive) means for various posterior quantities of interest. Section 4 of [1] contains some faked case studies, their purpose of which are to illustrate our results for the discrete case and show how they may be used in the process of determining a suitable number n of cells to investigate. These examples are with few obvious changes easily transferred to the continuous case that this paper handles. We thus refer our readers to [1] for examples.

2 Bayesian analysis

Recall that θ denotes the proportion of the region that is polluted. Notice that if a point p is chosen uniformly in R, then $P(f(p) > a) = \theta$. Hence, if n points $p_1, \ldots, p_n$ are chosen independently and uniformly in R, then the number X of points p_i, such that $f(p_i) > a$, is binomial with n trials and success probability θ.

It is well known that the beta distribution is conjugate to the binomial. For this reason and this reason only, the prior distribution of the polluted proportion θ is taken to be $\mathrm{Beta}(\alpha, \beta)$. Refer back to the introduction of this paper for the density of $\mathrm{Beta}(\alpha, \beta)$. The prior mean and variance are

$$E[\theta] = \frac{\alpha}{\alpha + \beta}$$

and

$$\mathrm{V}[\theta] = \frac{\alpha}{\alpha + \beta} \frac{\beta}{\alpha + \beta} \frac{1}{\alpha + \beta + 1}$$

respectively. Moreover, let θ_p be the pth quantile of θ. Thus θ_p is the unique number satisfying

$$p = \int_0^{\theta_p} p(\theta)\, d\theta$$

where $p(\theta)$ is the density of $\mathrm{Beta}(\alpha, \beta)$.

The probability mass function of X, given θ, is

$$p(x|\theta) = \binom{n}{x} \theta^x (1-\theta)^{n-x}$$

for $x = 0, 1, \ldots, n$. A straightforward application of Bayes' rule now shows that the posterior distribution of θ, given $X = x$, is $\mathrm{Beta}(\alpha + x, \beta + n - x)$.

Hence, the pre-posterior (or predictive) mean and variance of θ are

$$E[\theta|X] = \frac{\alpha + X}{\alpha + \beta + n}$$

and

$$\mathrm{V}[\theta|X] = \frac{\alpha + X}{\alpha + \beta + n} \, \frac{\beta + n - X}{\alpha + \beta + n} \, \frac{1}{\alpha + \beta + n + 1}$$

respectively. Clearly, by the double expectation formula,

$$E[E[\theta|X]] = E[\theta]$$

Thus, pre-posterior to doing the n measurements we cannot expect that our posterior mean $E[\theta|x]$ will be different from the prior mean $E[\theta]$. (This statement, of course, is more or less true depending on how well the parameters α, β are chosen.)

The calculations of other means of pre-posterior quantities such as the variance $\mathrm{V}[\theta|X]$ are more rewarding. Notice, however, first the well known fact that the pre-posterior distribution of X is beta-binomial with parameters n and α, β, to be referred to as $\mathrm{BB}(n, \alpha, \beta)$. A proof is given in Appendix A of [1]. Thus the probability mass function of X is

$$p(x) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \, \frac{n!}{\Gamma(\alpha + \beta + n)} \, \frac{\Gamma(\alpha + x)}{x!} \, \frac{\Gamma(\beta + n - x)}{(n - x)!}$$

for $x = 0, 1, \ldots, n$, and

$$E[X] = n \frac{\alpha}{\alpha + \beta}$$

and

$$\mathrm{V}[X] n \frac{\alpha}{\alpha + \beta} \, \frac{\beta}{\alpha + \beta} \, \frac{\alpha + \beta + n}{\alpha + \beta + 1}$$

(cf [1]).

A straightforward application of Lemma 1 of [1] now shows that

$$E[\mathrm{V}[\theta|X]] = \frac{\alpha+\beta}{\alpha+\beta+n}\mathrm{V}[\theta]$$

Thus, $\mathrm{V}[\theta] > E[\mathrm{V}[\theta|X]] \searrow 0$ as $1 \leq n \to \infty$. The posterior pth quantile $\theta_p(x)$ is defined by

$$p = \int_0^{\theta_p(x)} p(\theta|x)\,d\theta$$

were $p(\theta|x)$ is the density of Beta$(\alpha+x, \beta+n-x)$. The mean of $\theta_p(X)$ can of course be calculated by means of the formula

$$E[\theta_p(X)] = \sum_x \theta_p(x)p(x)$$

If one furthermore is interested in the probability distribution of, say, the pre-posterior variance $\mathrm{V}[\theta|X]$ (e g, in order to calculate a probability such as $P(\mathrm{V}[\theta|X] > c)$ similar to what was done in Case study A of [1]), then one just tabulates $\mathrm{V}[\theta|x]$ vs $p(x)$ for $0 \leq x \leq n$ and the various n's of interest.

Next, denote by C the total cost of remediating the site. We will assume below that C, given θ, is normal with mean $\mu\theta$ and variance $\sigma^2\theta$. As motivation we refer the reader to Section 2 of [1]. Assume, however, first only that $E[C|\theta] = \mu\theta$ and $\mathrm{V}[C|\theta] = \sigma^2\theta$. Then

$$\mathrm{V}[C] = \sigma^2 E[\theta] + \mu^2\mathrm{V}[\theta]$$

from which

$$E[\mathrm{V}[C|X]] = \sigma^2 E[\theta] + \mu^2\frac{\alpha+\beta}{\alpha+\beta+n}\mathrm{V}[\theta]$$

follows as in [1].

Assume next that the total cost C, given θ, is normal with mean $\mu\theta$ and variance $\sigma^2\theta$. Denote by $F(c)$ and $F(c|x)$ the prior and posterior distribution function of C. Then, cf [1],

$$F(c) = E\left[\Phi\left(\frac{c-\mu\theta}{\sigma\sqrt{\theta}}\right)\right]$$

and

$$F(c|x) = E\left[\Phi\left(\frac{c-\mu\theta}{\sigma\sqrt{\theta}}\right)\middle| X = x\right]$$

where Φ is the standard normal distribution function, given by

$$\Phi(z) = \int_{-\infty}^{z} \varphi(t)\, dt = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}}\, e^{-t^2/2}\, dt$$

The densities $f(c)$ and $f(c|x)$ are obtained by reversing the order of integration. Thus,

$$f(c) = E\left[\varphi\left(\frac{c-\mu\theta}{\sigma\sqrt{\theta}}\right)\frac{1}{\sigma\sqrt{\theta}}\right]$$

and

$$f(c|x) = E\left[\varphi\left(\frac{t-\mu\theta}{\sigma\sqrt{\theta}}\right)\frac{1}{\sigma\sqrt{\theta}}\middle|\, X = x\right]$$

Notice that the prior and posterior pth quantile or *value at risk* (VaR) $1-p$, c_p and $c_p(x)$, respectively, solves $F(c) = p$ and $F(c|x) = p$, and that the associated prior and posterior *expected shortfall*, e_p and $e_p(x)$ are defined exactly as in [1]. Notice also that the means of pre-posterior quantities such as $c_p(X)$ and $e_p(X)$ are easily calculated by means of the formula

$$E[h(X)] = \sum_x h(x)p(x)$$

where h is either c_p or e_p.

Acknowledgement

I am grateful to Lars Rosén for many valuable conversations regarding remediation of contaminated sites.

References

[1] Norberg, Tommy: A Bayesian approach to the assessment of contaminant spread. Part I. Discrete case. Preprint no 2002:09. Department of Mathematical Statistics, Chalmers University of Technology, 2002.

[2] Wackernagel, Hans: *Multivariate Geostatistics.* Springer, 1995.