Estimation of aquifer scale proportion using equal area grids: Assessment of regional scale groundwater quality

Kenneth Belitz,[1] Bryant Jurgens,[2] Matthew K. Landon,[1] Miranda S. Fram,[2] and Tyler Johnson[1]

Received 17 March 2010; revised 28 June 2010; accepted 11 August 2010; published 24 November 2010.

[1] The proportion of an aquifer with constituent concentrations above a specified threshold (high concentrations) is taken as a nondimensional measure of regional scale water quality. If computed on the basis of area, it can be referred to as the aquifer scale proportion. A spatially unbiased estimate of aquifer scale proportion and a confidence interval for that estimate are obtained through the use of equal area grids and the binomial distribution. Traditionally, the confidence interval for a binomial proportion is computed using either the standard interval or the exact interval. Research from the statistics literature has shown that the standard interval should not be used and that the exact interval is overly conservative. On the basis of coverage probability and interval width, the Jeffreys interval is preferred. If more than one sample per cell is available, cell declustering is used to estimate the aquifer scale proportion, and Kish's design effect may be useful for estimating an effective number of samples. The binomial distribution is also used to quantify the adequacy of a grid with a given number of cells for identifying a small target, defined as a constituent that is present at high concentrations in a small proportion of the aquifer. Case studies illustrate a consistency between approaches that use one well per grid cell and many wells per cell. The methods presented in this paper provide a quantitative basis for designing a sampling program and for utilizing existing data.

Citation: Belitz, K., B. Jurgens, M. K. Landon, M. S. Fram, and T. Johnson (2010), Estimation of aquifer scale proportion using equal area grids: Assessment of regional scale groundwater quality, *Water Resour. Res.*, *46*, W11550, doi:10.1029/2010WR009321.

1. Introduction

[2] Regional assessments of groundwater quality have been implemented in Europe [*Ward et al.*, 2004; *Grath et al.*, 2007; *Wendland et al.*, 2008], North America [*Lesage*, 2004; *Lapham et al.*, 2005], and elsewhere (as cited by *Rosen and Lapham* [2008] and *Mendizabal and Stuyfzand* [2009]). These assessments often include sampling for a large number of constituents (tens to >100) in a large number of wells (>100 to >1000). In addition, a comprehensive assessment program can be conducted in a large number of groundwater basins (sometimes >100) [*Belitz et al.*, 2003]. In recent years, robust measures and nonparametric statistical tests [*Helsel and Hirsch*, 2002] have been used to characterize the data from these assessments. For example, chemical concentrations are summarized using box plots that illustrate the median, quartiles, and range of the data. Although medians and quartiles are robust measures, they do include units of concentration, and therefore, it can be difficult to compare one chemical constituent to another. The issue of comparability can be addressed through the use of indices [*Backman et al.*, 1998; *Rentier et al.*, 2006; *Stigter et al.*, 2006] or if concentrations are normalized by a relevant value, such as a human health benchmark; *Toccalino and Norman* [2006] defined these normalized concentrations as benchmark quotients. Indices and benchmark quotients are dimensionless, thus allowing for a comparative analysis of different chemicals. *Worrall and Kolpin* [2003], in an evaluation of groundwater vulnerability, note that application of indices can involve subjective choices in the weighting of the component factors.

[3] This paper, recognizing the utility of dimensionless measures, addresses the issue of estimating the proportion of an aquifer where the concentration of a given constituent is above a specified threshold. The threshold of interest could be a human health benchmark, some fraction of a benchmark, or it could be an analytical detection level. For the purposes of discussion, concentrations above a threshold are referred to as high. In addition, the proportion of an aquifer with high concentrations is assessed on the basis of area rather than volume [*Reijnders et al.*, 1998]; the area-based proportion is defined here as the aquifer scale proportion.

[4] The use of aquifer scale proportion as a measure of water quality focuses attention on the aquifer and the constituent. From this perspective, one constituent may be more noteworthy than another, not because it has a larger median concentration or benchmark quotient, but because it is high in a larger proportion of the aquifer. Similarly, one aquifer might be considered more contaminated than another because it has a larger aquifer scale proportion for a particular constituent. Aquifer scale proportion can also be computed for a class of constituents, such as trace elements or organic compounds,

[1]USGS California Water Science Center, San Diego, California, USA.
[2]USGS California Water Science Center, Sacramento, California, USA.

thus allowing for an assessment of which constituent class has a greater impact on groundwater quality. Aquifer scale proportion can also be used to obtain a better understanding of important processes affecting noteworthy constituents [*Broers*, 2002, 2004]. The use of aquifer scale proportion as a measure of water quality does not necessarily equate to risk to human health. Other factors, including population served, toxicity, actual exposure levels, and potential synergistic effects of multiple constituents, also need to be considered.

[5] The use of proportion as a measure of groundwater quality is certainly not new. *Reijnders et al.* [1998] and *Broers* [2002, 2004] used detection frequency within a specified area as an estimator of proportion and evaluated uncertainty through the use of the cumulative binomial distribution and the *Blyth and Still* [1983] interval, respectively. *Broers* [2002] also used the binomial distribution to evaluate the probability of detecting contamination. These studies did not consider the potential influence of spatial bias due to clustering of data.

[6] The issue addressed in this paper is the extent to which an observed detection frequency is representative of the aquifer. This paper relies on equal area grids for providing spatially unbiased estimates of aquifer scale proportion and on the use of the binomial distribution for assessing the uncertainty associated with that estimate. An equal area grid could be developed for an aquifer system or for a part of an aquifer system; the term aquifer is used to refer to either situation.

[7] Equal area grids can be used to design a sampling network [*Gilbert*, 1987; *Alley*, 1993]. With this design, an aquifer is divided into cells of equal area. The cells can be defined using a rectilinear grid or they can have irregular shape [*Scott*, 1990]. One well is then randomly selected for sampling within each grid cell, thus avoiding the potential problem of clustered data. Given one well per cell, the aquifer scale proportion is equal to the detection frequency for the high concentrations. For the purposes of discussion, this is referred to as the grid-based approach.

[8] Equal area grids can also be used to provide spatially unbiased estimates of aquifer scale proportion when there is more than one well per grid cell (cell declustering) [*Journel*, 1983; *Isaaks and Srivastava*, 1989]. For the purposes of discussion, this is referred to as the spatially weighted approach, and it is discussed in more detail in the main body of the paper.

[9] The purpose of this paper is to obtain an estimate of aquifer scale proportion and to assess the uncertainty in that estimate. The main body of the paper is divided into eight sections (sections 2–9). In section 2, an idealized conceptualization of an aquifer is presented and used to illustrate the utility of equal area grids. In sections 3 and 4, the binomial distribution is introduced and then used to estimate a confidence interval for the grid-based aquifer scale proportion. Section 4 draws on important findings from the statistics research literature, particularly the shortcomings of using the standard interval as a basis for estimating confidence intervals. Sections 5 and 6 extend the analysis from a grid-based approach to a spatially weighted approach. In section 7, the binomial distribution is used to evaluate the adequacy of a grid of a given number of cells for identifying a small target (a constituent present at high concentrations in a small proportion of the aquifer). Section 7 addresses the issue of whether a small target is likely (or unlikely) to be detected when water quality samples are collected using a grid-based approach. In section 8, the binomial distribution is used to address the issue of prevalence when using a grid-based approach; prevalence can be used as a criterion for choosing which constituents, among a very large number, should be the subject of reporting and/or additional focus. In section 9, case studies are used to illustrate the approaches developed in the previous sections.

2. Idealization of an Aquifer for the Purpose of Estimating Aquifer Scale Proportion

[10] A spatially unbiased estimate of aquifer scale proportion can be obtained by dividing an aquifer into grid cells of equal area. For example, consider the idealization of a two-dimensional aquifer divided into 100 grid cells (Figure 1).

[11] In Figures 1a–1d, the dark cells are characterized by uniformly high concentrations, and the white cells are characterized by uniformly low concentrations. Since there are nine dark cells, the aquifer scale proportion is 0.09. The proportion obviously does not depend on location (corner or center of domain), shape (square or rectangular), or distribution (compact or distributed). If one were to obtain one water quality sample per cell, then one would obtain exactly nine samples with high concentration, and the detection frequency (number of samples with high concentrations/total number of samples) would be 0.09. For this highly idealized (and highly structured) representation, there is no uncertainty associated with the estimate of aquifer scale proportion.

[12] In Figures 1e–1h, there are 100 cells total, of which 36 are gray. Within each of the gray cells, one quarter of the cell is characterized by high concentrations, and the remainder is characterized by low concentrations. The gray cells can be conceptualized as a 2 × 2 grid with one dark "subcell" (Figure 1i) or as a 4 × 4 grid with four dark subcells (Figure 1j). Independent of conceptualization, the proportion of the aquifer with high concentrations is 0.09 (36/100 × 1/4). The aquifer scale proportion does not depend on the location, shape, or distribution of the gray cells nor does it depend on the location, shape, or distribution of dark subcells within the gray cells. If one were to randomly obtain one water quality sample from each cell, then one would obtain 64 low samples from the 64 white cells and one would expect to obtain 9 high samples and 27 low samples from the 36 gray cells; by chance, one could obtain fewer or more than nine high samples. Unlike Figures 1a–1d, there is uncertainty associated with the estimate of the aquifer scale proportion in Figures 1e–1h.

[13] In Figure 1k, there are 100 light gray cells. Within each and every light gray cell, 9% of the cell is characterized by high concentrations, and the remainder is characterized by low concentrations. Each of the light gray cells can be conceptualized as consisting of 100 subcells, as illustrated in Figures 1a–1d or as in Figures 1e–1h (with corresponding distributions as shown in Figures 1i and 1j). If one were to randomly obtain one water quality sample from each cell, then one would expect to obtain 9 high samples and 91 low samples. The probability of obtaining exactly nine high samples (and exactly 91 low samples) is lower in Figure 1k than in Figures 1e–1h; additional uncertainty is associated with the more widespread distribution of the high concentrations over the entire aquifer (100 cells rather than 36 cells). Additional subdivision of the 100 light gray cells,

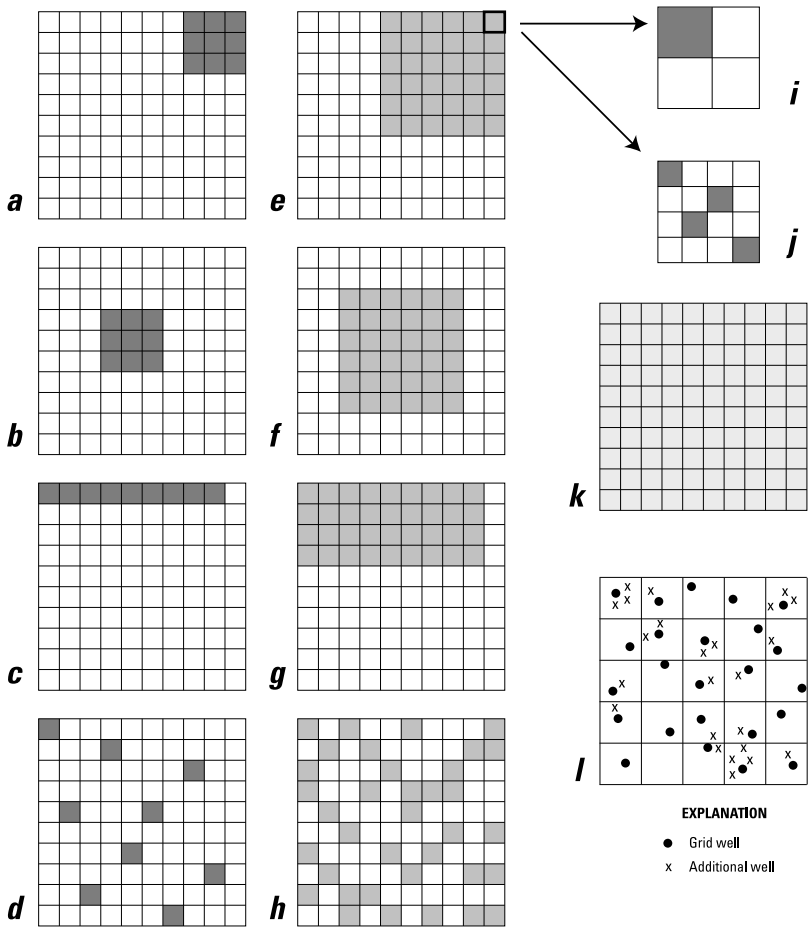

Figure 1. (a–h) The proportion of the aquifer with high concentrations is 9%. The nine dark cells in Figures 1a–1d indicate concentrations that are uniformly high throughout the cell. The 36 gray cells in Figures 1e–1h indicate an area where the target occupies 25% of the cell. For example, one gray cell can be conceptualized as a 2 × 2 grid, with one dark cell (Figure 1i). Alternatively, one gray cell can be conceptualized as a 4 × 4 grid, with four dark cells (Figure 1j). (l) The proportion of the aquifer with high concentrations is 9%, the proportion within each cell is also 9%, and the configuration within each of the cells could be as illustrated in the preceding representations. The aquifer is divided into 25 cells; the proportion and distribution of high concentrations is unknown.

for example, into a 20 × 20 or 30 × 30 subgrid rather than a 10 × 10 subgrid, would not introduce additional uncertainty: within each light gray cell, the probability of sampling a dark cell is 0.09. As described, the high concentrations in Figure 1k are areas of finite size but of unknown location within any given cell. The locations could be structured or widely distributed within the cell. At the limit, the high concentrations could be fully dispersed.

[14] Figures 1a–1k illustrate an important point: if one divides an aquifer into cells of equal area and obtains one water quality sample per cell, then the expected value of the aquifer scale proportion does not depend on the spatial distribution of high concentrations; an assumption of homogeneity is not required. However, the uncertainty associated with the grid-based estimate does depend on the spatial distribution of high concentrations.

[15] In Figure 1l, a 5 × 5 sampling grid is overlain on an aquifer in which the distribution and proportion of high concentrations are unknown. For the purposes of estimating aquifer scale proportion, the high concentrations might be restricted to a few areas (Figures 1a–1h), distributed across the aquifer system (Figure 1k), or have other characteristics such as second-order stationarity [*Journel and Huijbregts*, 1978]. For the purposes of estimating a confidence interval, it is assumed that the high concentrations are uniformly distributed (Figure 1k), but not necessarily fully dispersed. The 5 × 5 equal area grid could be used to identify wells for sampling (grid-based approach) or it could be used for cell declustering (spatially weighted approach). In this paper, the focus is at the scale of the entire aquifer, with no attempt made to evaluate the internal distribution of high concentrations.

3. Binomial Distribution and the Grid-Based Approach

[16] The binomial distribution assigns a probability b to achieving a given number of successes k in a given number of trials n, where the probability of success is p [*Ott and Longnecker*, 2001]:

$$b(k;n,p) = \binom{n}{k} p^k (1-p)^{n-k} = \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k}. \quad (1)$$

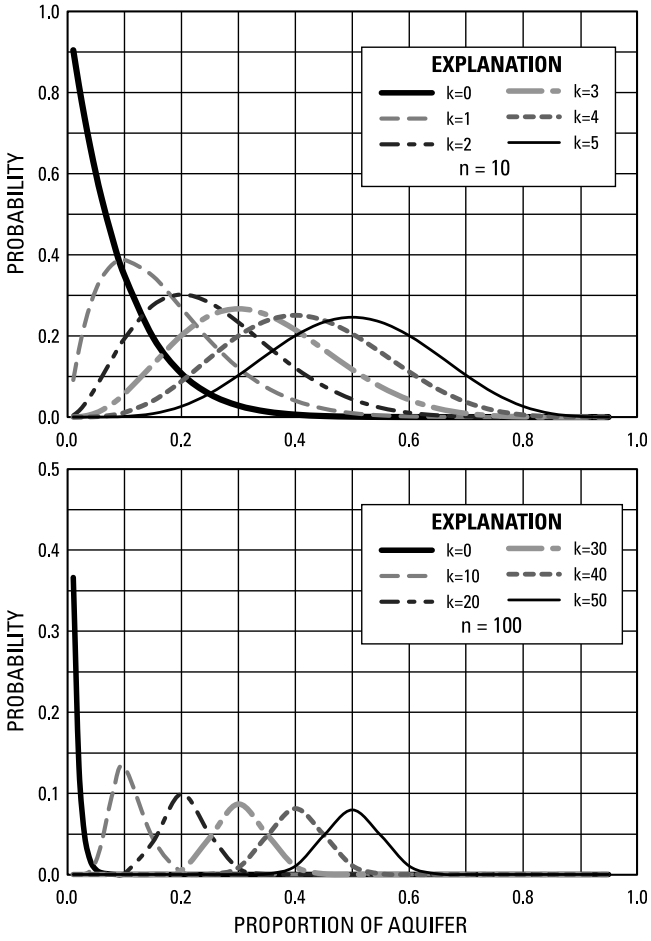

Figure 2. Binomial distribution illustrating the probability of obtaining k detections using a grid-based approach (one sample per cell, n = number of cells), as a function of aquifer proportion. (a and b) The number of cells is different, but the detection frequency (f) is the same (f = 0, 0.1, 0.2, 0.3, 0.4, 0.5).

In terms of water quality samples obtained using a grid-based approach (one well sampled per cell), the parameters of the binomial distribution are defined as the number of samples with high concentrations (k), the number of cells sampled (n), and the proportion of the aquifer with high concentrations (p). Three notable characteristics of the binomial distribution, as applied to a grid-based approach, are illustrated in Figure 2. First, for any given observed detection frequency (variously labeled curves in Figure 2), the highest probability is associated with an aquifer where the proportion is equal to the detection frequency; this characteristic is independent of the number of cells in the grid. Second, the distribution (for a given detection frequency) is narrower for a grid with more cells than for a grid with less cells. Third, the distribution is asymmetric at low (and high) detection frequency.

[17] Generally, the proportion of an aquifer with high concentrations is unknown, and we seek to estimate its value and to provide a confidence interval for that estimate. If a grid-based approach is used to obtain samples, then the most likely estimate of the unknown proportion ($\hat{p}$) is the observed detection frequency (f),

$$\hat{p} = f = k/n. \tag{2}$$

The confidence interval for this estimate is discussed in section 4.

4. Estimation of Confidence Intervals for the Binomial Proportion

[18] The appropriate method for estimating a two-sided confidence interval for the binomial proportion has been the subject of considerable research [*Vollset*, 1993; *Agresti and Coull*, 1998; *Brown et al.*, 2001, 2002; *Cai*, 2005]. In particular, these researchers have used coverage probability as a criterion for evaluation. The coverage probability of a confidence interval, for a fixed value of a parameter, is the probability that the interval contains that value. These researchers evaluated more than 15 methods for estimating a confidence interval; most perform poorly, while a few have acceptable coverage properties.

[19] The standard confidence interval (CI_s) for the estimated proportion (also known as the Wald interval) is

$$CI_s = \hat{p} \pm z_{\alpha/2}\sqrt{\sigma^2/n}, \tag{3a}$$

where $z_{\alpha/2}$ is the $(1 - \alpha/2)$ quantile of the standard normal distribution [*Ott and Longnecker*, 2001], α is the significance level associated with the confidence interval, and σ^2 is the variance. For a 90% confidence interval, $\alpha = 0.10$.

[20] Given that $\sigma^2 = \hat{p}(1 - \hat{p})$for a binomial distribution, equation (3a) becomes

$$CI_s = \hat{p} \pm z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n}. \tag{3b}$$

On the basis of the criterion of coverage probability, numerous researchers recommend against using the Wald interval. *Brown et al.* [2001, 2002] are unequivocal in their rejection of the Wald interval, stating that it should not be used under any circumstance (interested readers might want to see the extensive discussions that accompany the article by *Brown et al.* [2001]). For example, given a nominal confidence interval of 0.95, the average coverage probability for the Wald interval is less than 0.92 for $n < 100$ and less than 0.8 for $n < 20$ [*Brown et al.*, 2001]. In addition to poor coverage probability, the Wald interval is symmetric and can provide negative values and values larger than one, which are problematic for proportions. *Vollset* [1993] demonstrated that continuity-corrected modifications of the Wald interval, such as the *Blyth and Still* [1983] interval, have the same shortcomings as the Wald interval. *Agresti and Coull* [1998] propose a modified Wald interval: add two successes and two failures. *Brown et al.* [2001] have shown that the Agresti-Coull interval provides better coverage probability than the Wald interval and suggest that the Agresti-Coull interval can be used for $n > 40$.

[21] The exact method [*Clopper and Pearson*, 1934] is a commonly recommended alternative to the Wald interval. The lower and upper bounds of the Clopper-Pearson interval are obtained by inverting equal-tailed binomial tests of the null hypothesis. Several researchers [*Vollset*, 1993; *Agresti and Coull*, 1998; *Brown et al.*, 2001; *Brown et al.*, 2002] have shown that the Clopper-Pearson interval is inherently conservative. For example, given a nominal confidence interval of 0.95, the average coverage probability exceeds 0.98 for $n < 50$ and can approach 1.0 for $n < 10$. The

conservatism of the Clopper-Pearson interval leads to confidence intervals that are unnecessarily wide.

[22] *Agresti and Coull* [1998] note that the Wilson score confidence interval (CI_W) represents a compromise between the Wald and Clopper-Pearson intervals; it is based on inverting the approximately normal test at the null value of the hypotheses,

$$CI_W = \left(\hat{p} + \frac{z^2_{\alpha/2}}{2n} \pm z_{\alpha/2}\sqrt{\left[\hat{p}(1-\hat{p}) + z^2_{\alpha/2}/4n\right]/n}\right)/\left(1 + z^2_{\alpha/2}/n\right). \quad (4)$$

The two-sided Wilson score interval has good coverage probability [*Vollset*, 1993; *Agresti and Coull*, 1998; *Brown et al.*, 2001, 2002] and is relatively easy to compute. *Brown et al.* [2001] have shown that the average coverage probability for the Wilson interval is closer to the nominal value than the Agresti-Coull interval. *Brown et al.* [2002] also showed that the Jeffreys interval and the likelihood ratio interval provide two-sided coverage properties comparable to the Wilson interval and suggest that any of the three can be used.

[23] *Brown et al.* [2001], in their response to comments, noted that a particular method can provide a two-sided confidence interval with good coverage probability, even while failing to provide satisfactory one-sided intervals. This apparent anomaly occurs because of compensating one-sided errors. *Cai* [2005] evaluated coverage probability for one-sided confidence intervals. He found that the upper bound for the Wilson interval provides overcoverage for small proportions ($p < 0.3$) and undercoverage for large proportions ($p > 0.7$). The bias in the coverage, for a nominal confidence level of 0.98, ranged from 0.01 to 0.02. In contrast, the coverage probability of the Jeffreys interval is close to the nominal confidence level for all values of p. *Cai* [2005] did not evaluate the coverage probability for the upper bound of the likelihood ratio interval.

[24] The Wald, Clopper-Pearson, and Wilson intervals are derived from a frequentist perspective [*Brown et al.*, 2001; *Berger*, 1985]. In contrast, the Jeffreys interval is derived from application of Bayes' theorem,

$$g(p;k,n) = \frac{b(k;n,p)g(p)}{\int_0^1 b(k;n,p)g(p)\mathrm{d}p}, \quad (5)$$

where $g(p; k, n)$ is the posterior distribution of the aquifer scale proportion (p), given k and n, and $g(p)$ is the prior distribution of p. The lower and upper bounds of the $100(1-\alpha)$% confidence interval are the $\alpha/2$ and $1-\alpha/2$ quantiles of the posterior distribution. Alternatively stated, the confidence interval is obtained by trimming the tails on the posterior distribution.

[25] If $g(p) = 1$ (uniform prior distribution of p), then the posterior distribution would simply be the binomial distribution evaluated for the specified value of p (and k), normalized by the cumulative distribution for all values of p (and the fixed value of k). For example, graphs of $g(p; k, n)$ would have the same shapes as the graphs of Figure 2, but the y values would be normalized by the area under the curves. For a uniform prior, the cumulative distribution in the denominator of equation (5) is equal to $1/(n+1)$. *Reijnders et al.* [1998], by trimming the tails on the cumulative binomial distribution, used a uniform prior for evaluating a confidence interval.

[26] A prior distribution can be identified using information relevant to the problem at hand or, in the absence of sufficient information, a noninformative prior can be used [*Berger*, 1985]. For analytical and computational purposes, it is useful to select a prior distribution that provides a posterior distribution with the same mathematical form as the prior distribution (a conjugate prior). The beta distribution is the standard conjugate prior for the binomial distribution [*Berger*, 1985].

[27] If a beta distribution [$Beta(m_1, m_2)$] is used in equation (5), then

$$g(p;k,n) = Beta(k+m_1, n-k+m_2), \quad (6)$$

where m_1 and m_2 are shape parameters. For a uniform prior, $m_1 = m_2 = 1$. Although a uniform prior could be used, it is not as noninformative as the Jeffreys prior [*Berger*, 1985]. The use of Jeffreys prior leads to a confidence interval referred to as the Jeffreys interval ($m_1 = m_2 = {}^1/_2$ in equation (6)). *Brown et al.* [2001] noted that the Jeffreys interval can be regarded as a continuity-corrected version of the Clopper-Pearson interval, thus providing a frequentist rationale for a Bayesian method.

[28] The lower bound ($L_{1-\alpha}$) and upper bound ($U_{1-\alpha}$) of the Jeffreys interval are found by inverting equation (6) at the appropriate points of the distribution,

$$L_{1-\alpha}(0,n) = 0 \qquad \text{for } k = 0, \quad (7a)$$

$$L_{1-\alpha}(k,n) = B^{-1}(\alpha/2; k+{}^1/_2, n-k+{}^1/_2) \qquad \text{for } k > 0, \quad (7b)$$

$$U_{1-\alpha}(k,n) = B^{-1}(1-\alpha/2; k+{}^1/_2, n-k+{}^1/_2) \quad \text{for } k < n, \quad (7c)$$

$$U_{1-\alpha}(n,n) = 1 \qquad \text{for } k = n, \quad (7d)$$

where B^{-1} (α; m_1, m_2) is the inverse of the cumulative beta distribution. The inverse beta function can be evaluated using commonly available software packages including MS-Excel®. *Brown et al.* [2001] also provide a formula for approximating the bounds of the interval.

[29] For the purposes of illustration, the aquifer scale proportion and confidence intervals associated with a single detection are computed using the Jeffreys, Wilson, and Agresti-Coull methods for grids ranging in size from 10 to 100 cells (Figure 3). The value of the aquifer scale proportion is closer to the lower bound than to the upper bound for all three methods, reflecting the asymmetry of the binomial distribution. For the range of values shown ($10 \le n \le 100$), the width of the Jeffreys interval is about 90% of the width of the Wilson interval and about 70% of the Agresti-Coull interval. As the number of detections increase (not shown in Figure 3), the widths of the intervals become closer; for example, for three detections the width of the Jeffreys interval is about 97% of the width of the Wilson interval and about 88% of the Agresti-Coull interval. The Jeffreys interval also provides a narrower confidence interval than an interval based on a uniform prior (not shown in Figure 3). For a single

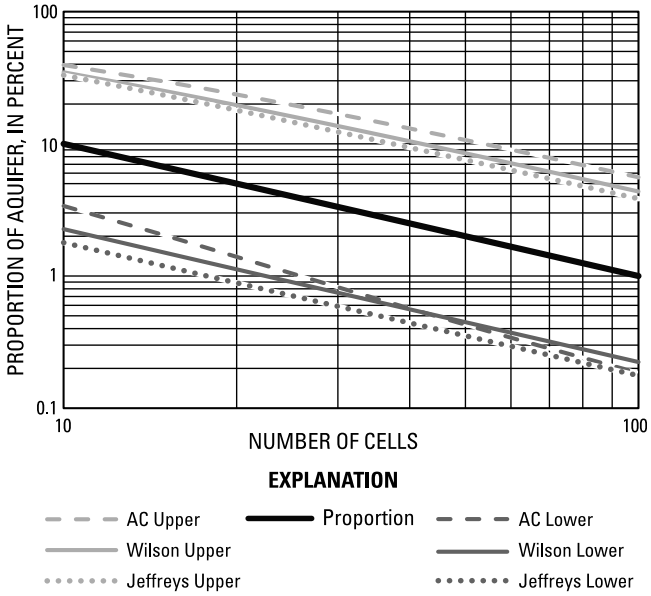

Figure 3. Estimated aquifer proportion and corresponding 90% confidence interval for a single detection, as a function of the number of cells in the sampling grid.

detection ($10 \le n \le 100$), the Jeffreys interval is about 88% of the width of the interval obtained using a uniform prior, and for three detections, the Jeffreys interval is about 96% of the interval obtained using a uniform prior.

[30] In this paper, the Jeffreys method is used for computing confidence intervals for aquifer scale proportions. The Jeffreys interval is used because it provides a narrower confidence interval and better one-sided coverage probability [*Cai*, 2005] than the Wilson and Agresti-Coull intervals. The Jeffreys interval also provides a narrower confidence interval than an interval obtained using a uniform prior.

5. Incorporation of Additional Data: Spatially Weighted Estimation of Aquifer Scale Proportion

[31] Incorporation of additional water quality data, beyond one sample per cell, into the analysis of aquifer scale proportion requires consideration of the potential effects of clustering. For example, consider Figures 1a–1h. If one were to obtain one water quality sample from the black cells and more than one from the white cells, then the observed detection frequency would be lower than the actual aquifer scale proportion. *Journel* [1983] proposed the use of "cell declustering" to estimate a global mean based on data that are not uniformly distributed across the domain.

[32] In the cell declustering approach, an equal area grid is overlain on the domain. A local value of aquifer scale proportion ($\hat{p}_i$) is then computed for the ith grid cell, and a global value for the domain is obtained by averaging the local values,

$$\hat{p}_i = f_i = k_i / nw_i, \tag{8a}$$

$$\hat{p} = \frac{1}{n}\sum_{i=1}^{n} \hat{p}_i, \tag{8b}$$

where k_i is the number of water quality samples with high concentrations in the ith cell and nw_i is the number of wells sampled in the ith cell. From a global perspective, the weight assigned to any given well is inversely proportional to the number of cells in the grid and the number of wells in the same cell as that well. The value of $\hat{p}$ computed using equations (8a) and (8b) is defined as the spatially weighted aquifer scale proportion. If there is only one well per cell, the spatially weighted value is identical to the grid-based detection frequency (equation (2)).

[33] *Isaaks and Srivastava* [1989] note that the global value obtained by cell declustering can be a function of the number of cells and recommend finding the number of cells that minimize (or maximize) the estimate. Finding an optimal number of cells in a grid requires systematically varying the number of rows and columns, and computing the global value for each possible combination. In this paper, the global value (spatially weighted aquifer scale proportion) is computed using previously defined equal area grids. As will be shown in the case studies, there are often a large number of constituents present at high concentrations. Evaluation of an optimal grid for each constituent could require considerable effort.

[34] For the purposes of discussion, the uncorrected detection frequency computed using all the data is defined as the raw detection frequency. In this paper, the grid-based aquifer scale proportion ($\hat{p}_{\text{grid}}$) is compared to both the raw detection frequency and the spatially weighted value. If one is making a comparison for several constituents, the overall difference can be evaluated using the mean absolute deviation (MAD),

$$\text{MAD} = \frac{\sum_{i=1}^{\text{nc}} \left|\hat{p}_{\text{grid}} - \hat{p}_a\right|}{\text{nc}}, \tag{9}$$

where nc is the number of constituents being compared and $\hat{p}_a$ is the value computed using all of the data (either the raw detection frequency or the spatially weighted value).

6. Confidence Intervals for Spatially Weighted Aquifer Scale Proportion

[35] Estimation of a confidence interval for the spatially weighted aquifer scale proportion requires consideration of the potential effects of clustering. Although the error associated with the spatially weighted proportion cannot be directly assessed [*Journel*, 1983], some understanding can be obtained through consideration of the problem from three perspectives: clustered sampling design, stratified sampling design, and an aquifer with high concentrations that are fully dispersed. For the first two perspectives, the distribution of high concentrations is assumed to be unknown.

[36] *Kish* [1965] proposed the use of a design effect (DE) for complex sampling designs, such as clustered sampling and stratified sampling [also see *Cochran*, 1977; *Kish*, 1995; *Campbell et al.*, 2007]. The design effect (DE) is used to adjust the total number of samples for the purpose of computing a confidence interval,

$$n^* = N_{\text{wells}}/\text{DE}, \tag{10}$$

where n^* is the effective number of samples and N_{wells} is the actual number of wells sampled.

[37] A clustered sampling design is one in which clusters of samples are obtained from a population (e.g., voters interviewed at selected polling stations). If the equal area grid,

with multiple wells per cell, is viewed as a clustered sampling design, then

$$DE_c = 1 + (m_a - 1)\rho, \tag{11a}$$

$$m_a = N_{wells}/n \tag{11b}$$

$$\rho = \sigma_{bc}^2/\left(\sigma_{bc}^2 + \sigma_{wc}^2\right) = \sigma_{bc}^2/\sigma_T^2, \tag{11c}$$

where DE_c is the design effect for a clustered design, m_a is the average cluster size (average number of wells in a cell), ρ is the intraclass correlation, σ_{bc}^2 is the between-cluster (between-cell) variance, σ_{wc}^2 is the within-cluster (within-cell) variance, and σ_T^2 is the total variance.

[38] Given an effective sample size, one can compute an effective number of successes (k^*),

$$k^* = n^*\hat{p}. \tag{12}$$

The effective number of successes need not be an integer. The confidence interval is computed by substituting n^* and k^* into equations (7a)–(7d).

[39] The between-cell variance can be computed in terms of the previously defined local and aquifer scale proportions (equations (8a) and (8b)),

$$\sigma_{bc}^2 = \frac{1}{n}\sum_{i=1}^{n}(\hat{p}_i - \hat{p})^2. \tag{13}$$

The within-cell variance can be computed as an average of the variances computed for each cell,

$$\sigma_{wc}^2 = \frac{1}{n}\sum\nolimits_{i=1}^{n}\sigma_i^2, \tag{14a}$$

$$\sigma_i^2 = \hat{p}_i(1 - \hat{p}_i). \tag{14b}$$

The total variance can be computed as the sum of the between-cell and within-cell variances,

$$\sigma_T^2 = \sigma_{bc}^2 + \sigma_{wc}^2, \tag{15a}$$

or from the spatially weighted aquifer scale proportion,

$$\sigma_T^2 = \hat{p}(1 - \hat{p}). \tag{15b}$$

Computation using equations (15a) and (15b) provides a check on the accuracy of the computations. From a geostatistical perspective, equation (15a) is the dispersion variance [*Journel and Huijbregts*, 1978].

[40] The characteristics of the design effect can be understood by substituting equations (11a) through (11c) into (10) and rearranging

$$\frac{1}{n^*} = \frac{1 - \rho}{N_{wells}} + \frac{\rho}{n} \tag{16a}$$

or

$$\frac{\sigma_T^2}{n^*} = \frac{\sigma_{wc}^2}{N_{wells}} + \frac{\sigma_{bc}^2}{n}. \tag{16b}$$

Equation (16a) illustrates that the effective number of samples (n^*) is a weighted function of the number of wells and the number of cells. At a minimum, n^* is equal to the number of cells, and at a maximum, it is equal to the number of wells. A confidence interval computed using the number of cells would be wider than a confidence interval computed using the number of wells.

[41] Equation (16b) provides additional insight into the design coefficient. The left-hand side of equation (16b) can be viewed as the standard error associated with the spatially weighted estimate. The first term on the right-hand side can be viewed as the standard error associated with a stratified design [*Kish*, 1965; *Cochran*, 1977], and the second term on the right-hand side can be viewed as an additional source of error associated with dividing the aquifer into n cells for the purpose of declustering the data. Acceptance of the equal area grid as a clustered design requires that the variance between cells (σ_{bc}^2) arises from uncertainty, rather than from deterministic differences between cells. This assumption could be met if areas with high concentrations are uniformly distributed across the aquifer (Figure 1l); note that a uniform distribution is not identical to a fully dispersed distribution.

[42] In a stratified design, a finite population is divided into subsets based on relevant criteria. If the equal area grid is viewed as a stratified sampling design (with each cell treated as an individual stratum) and there are m_a wells in each cell, the design effect (DE_s) becomes [*Kish*, 1965; *Cochran*, 1977],

$$DE_s = \frac{N_{wells}}{n^*} = \frac{\sigma_{wc}^2}{\sigma_T^2}. \tag{17}$$

Application of equation (17) results in an effective number of samples (n^*) equal to or larger than the actual number of wells because $\sigma_{wc}^2 \le \sigma_T^2$. Acceptance of the equal area grid as a stratified design requires that the variance between cells (σ_{bc}^2) arises from deterministic differences between cells, rather than contributing to uncertainty. For example, if high concentrations are structured (Figures 1a–1h) and the orientation of the grid with that structure were known, equation (17) could be used for computing n^*.

[43] If high concentrations were fully dispersed across an aquifer, then it is not necessary to decluster the data. Without structure or autocorrelation, all data have equal value [*Journel and Huijbregts*, 1978]. In that case, the aquifer scale proportion is computed as an unweighted mean and the confidence interval is computed using the actual number of wells.

[44] In the case studies presented in this paper, the effective number of samples is computed from the perspective of a clustered sampling design; DE_c (equations (11a)–(11c)) is substituted for DE in equation (10). Consequently, the confidence intervals could be too wide. The clustered sampling design effect is chosen because no additional work is done to evaluate the distribution of high concentrations within the aquifer. Analysis of the distribution for a large number of constituents in a large number of aquifers could require substantial effort.

[45] Conceptualization of the equal area grid as a stratified design (with each cell as a separate stratum) differs from the strategy of dividing an aquifer into selected subareas based on hydrogeologic characteristics [*Broers*, 2002]. In the latter

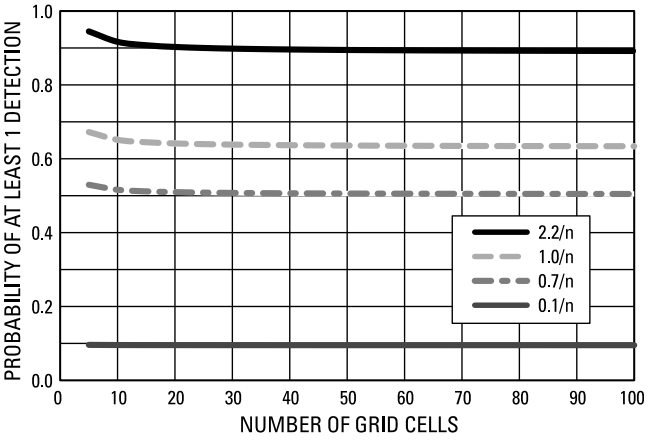

Figure 4. A small target is defined in terms of the size of the grid. Each of the curves is for a target of specified size (proportion of aquifer with high concentrations). The curve labeled $1.0/n$ is a target present at a proportion equal to 1 out of n grid cells; the probability of detecting a target of this size is about 0.63. The curve labeled $2.2/n$ can be defined as a target that is unlikely to be missed. The curve labeled $0.1/n$ can be defined as a target that is unlikely to be detected. The curve labeled $0.7/n$ defines a target with an equal probability of being detected or not detected.

case, one is using prior information to identify subareas such that points within the subarea are relatively similar to one another and that points in different subareas are relatively different from each other. In the latter case, one could use equal area grids for each subarea, compute proportions and confidence intervals for each subarea using methods described in this paper, and then combine the results for the larger aquifer system using methods for stratified sampling [*Cochran*, 1977; *Broers*, 2002].

7. Detecting a Small Target Using a Grid-Based Approach

[46] A target is defined as that part of the aquifer with a constituent present at high concentrations. The target might be contiguous or it might be distributed (Figure 1). The size of the target, when expressed as a function of the areal extent of an aquifer, is equal to the aquifer scale proportion. From this perspective, target size is a nondimensional parameter. Given an equal area grid of n cells with one sample obtained per cell, a target is defined as too small if it is unlikely to be detected. Similarly, a target is defined as sufficiently large if it is unlikely to be missed. Given these two definitions, one can obtain a lower and upper bound for the size of a small target. In turn, one can then assess the utility of the grid-based approach for identifying a small target.

[47] The probability of detecting a small target $[D(n, p_s)]$ can be computed in terms of the probability of not detecting the target (substituting $k = 0$ in equation (1)),

$$D(n, p_s) = 1 - b(0; n, p_s). \tag{18a}$$

For the purposes of analysis, the size of a small target (p_s) can be expressed in terms of the number of cells in the grid,

$$D(n, c/n) = 1 - b(0; n, c/n). \tag{18b}$$

For example, if a target is present at a proportion equal to one out of n grid cells, then $c = 1$. In Figure 4, $D(n, c/n)$ is plotted as a function of n for several values of c. For n ranging from 20 to 100, the probability of detecting a small target is relatively constant; for example, the probability of detecting a target present at a proportion of $1.0/n$ is about 0.63. For $n <$ 20, the probability of detection is somewhat greater than 0.63. This indicates that if a target has a size equal to a single cell, then there is at least a 63% chance of detecting that target.

[48] A target that is too small is defined as one that will be detected at a probability π', and a target that is sufficiently large is defined as one that will be detected at a probability π, where $\pi > \pi'$. If we require that $\pi' = (1 - \pi)$, then the lower and upper bounds are expressed with respect to a single probability level π. Establishment of this requirement reflects a "balancing of errors" [*Smith et al.*, 2001; *McBride and Ellis*, 2001; *McBride*, 2003]; the probability of not detecting a target at the lower bound is equal to the probability of detecting the target at the upper bound.

[49] Given a specified probability level (π), a small target is bounded in size by a lower limit ($p_{l|\pi}$) and an upper limit ($p_{u|\pi}$),

$$p_{l|\pi} \le p_{s|\pi} \le p_{u|\pi}. \tag{19a}$$

The lower and upper limits are found by inverting equation (18b),

$$p_{l|\pi} = D^{-1}(\pi'; n, c_1/n) = D^{-1}(1 - \pi; n, c_1/n), \tag{19b}$$

$$p_{u|\pi} = D^{-1}(\pi; n, c_2/n). \tag{19c}$$

The subscripts for the constant c are different in equations (19b) and (19c) and reflect the different probabilities associated with detecting a target that is too small as compared to a target that is sufficiently large.

[50] For $\pi = 0.9$, the lower and upper bounds on the size of a small target are approximately (Figure 4),

$$\frac{0.10}{n} \le p_{s|\pi=0.9} \le \frac{2.2}{n}. \tag{20a}$$

Given a grid of 50 cells, a small target is unlikely to be detected if it is present in less than 0.2% of the aquifer system, and it is unlikely to be missed if it is present in more than 4.4% of the aquifer system.

[51] For $\pi = 0.95$, the size of a small target is approximately (not shown in Figure 3),

$$\frac{0.05}{n} \le p_{s|\pi=0.95} \le \frac{3.0}{n}. \tag{20b}$$

If $\pi = 0.5$, then the small target is one where the probability of being detected is equal to the probability of being missed. For n ranging from 20 to 100,

$$p_{s|\pi=0.5} = D^{-1}(0.5; n, c/n) \approx 0.7/n. \tag{20c}$$

If a target is present at a proportion less than $0.7/n$, then the target is more likely than not to be missed. Conversely, if

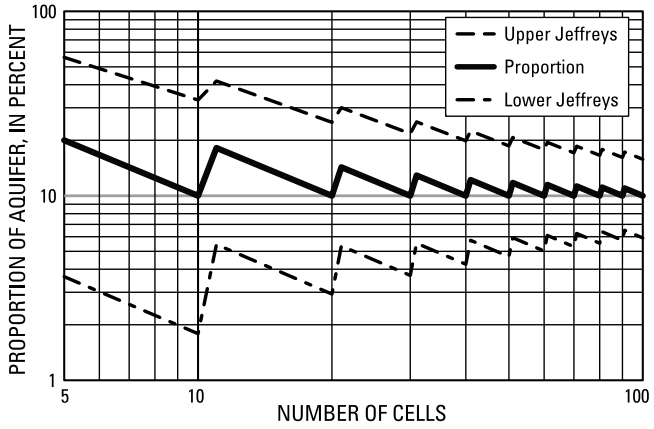

Figure 5. 90% confidence intervals for a prevalent compound (≥10% detection frequency).

the target is present at a proportion more than $0.7/n$, then the target is more likely than not to be detected.

8. Confidence Intervals for a Constituent Detected at 10% Frequency Using a Grid-Based Approach (Prevalent Constituents)

[52] Regional assessments of groundwater quality can include analyses for dozens or even hundreds of constituents. Given the potential for detecting a large number of constituents, one might choose to focus more attention on those constituents that are prevalent (frequently detected) and less attention on those that are not. The detection frequency for defining prevalence is subjective; one possibility is a detection frequency equal to or exceeding 10%.

[53] The aquifer scale proportion for a prevalent constituent and the associated bounds for the 90% confidence interval are plotted in Figure 5. Each of the upward steps in the graphs represents a step change from k high values to $k + 1$ high values; the transition from k to $k + 1$ is required so that the detection frequency is equal to or exceeds 10%. For example, when n increases from 10 to 11, k increases from 1 to 2. For a small number of cells ($n \leq 20$, with the exception of $n = 10$), the lower bound of the 90% confidence interval is above 2%. For a larger number of cells ($20 < n \leq 30$), the lower bound is above 3%. For $n \geq 60$, the lower bound is above 5%. Noting that the lower bound of a 90% confidence interval is also a one-sided 95% confidence level, inferences can be drawn about prevalent constituents: for a small grid ($n \leq 20$, with the exception of $n = 10$), there is at least a 95% confidence that a prevalent constituent is present in more than 2% of the aquifer; for a larger grid ($n \geq 60$), there is at least a 95% confidence that a prevalent constituent is present in more than 5% of the aquifer.

9. Case Studies From California's GAMA Program

[54] The U.S. Geological Survey (USGS), in collaboration with the California Water Board's Ground Water Ambient Monitoring and Assessment program (GAMA), is implementing an evaluation of groundwater quality in about 120 groundwater basins in California [*Belitz et al.*, 2003]; these evaluations are called priority basin assessments. The priority groundwater basins, along with selected areas outside of basins, have been aggregated into study units; it is anticipated that about 35 study units will be evaluated. The priority basin assessments are based on groundwater quality data from existing wells, primarily wells used for public supply. In this paper, data from two study units are presented as examples. *Mendizabal and Stuyfzand* [2009] discuss the utility of using public supply wells for the purposes of assessing regional groundwater quality.

[55] The GAMA program uses equal area grids to design a sampling network. Within each grid cell, one public supply well is randomly selected for sampling [*Scott*, 1990]. If there are no public supply wells available in a cell, then an attempt is made to sample a well tapping the same depth zone as public supply wells located in nearby cells. For the purposes of discussion, wells sampled as part of the equal area grid network are referred to as grid wells. Additional wells are also sampled for the purposes of evaluating the human and natural factors that may affect water quality; these wells are referred to as understanding wells.

[56] All grid and understanding wells are sampled for an extensive suite of organic constituents and selected field parameters, but not all wells are sampled for inorganic constituents. The California Department of Public Health (CDPH) maintains a database containing chemical analyses conducted for the purposes of regulatory compliance; these data are used to provide additional coverage for those cells where inorganic data were not collected by the USGS. The USGS anticipates sampling about 2500 wells for the GAMA program, and there are about 15,000 public supply wells with chemical data in the CDPH database. The USGS uses analytical methods that evaluate water quality samples for a larger suite of organic constituents and at lower detection levels (1–2 orders of magnitude lower) than the analytical methods used for regulatory compliance. The analytical methods used for regulatory compliance are suitable for evaluating concentrations relative to health-based and aesthetic thresholds. Detections of anthropogenic organic compounds at very low concentrations can provide additional information about the potential impact of human activities on groundwater quality [*Shelton et al.*, 2001; *Worrall and Besien*, 2005].

9.1. Case Study 1: Concentrations Above Health-Based Thresholds in the Central Eastside Study Unit

[57] The Central Eastside study unit (Figure 6) is located in California's San Joaquin Valley. The Central Eastside study unit was sampled with four equal area grids. Three of the grids correspond to Quaternary alluvial deposits in each of the Modesto, Merced, and Turlock groundwater basins, and the fourth grid corresponds to Quaternary-Pleistocene consolidated deposits that form elevated terraces along the eastern portions of the three basins (Figure 6). A total of 60 grid cells were identified, with each cell having an area of 100 km^2. Fifty-eight grid wells were sampled by the USGS from March through June 2006. Twenty additional wells were sampled for the purposes of understanding; of these, three are screened within the depth zone used for public supply. For the purposes of illustration, the study unit is evaluated as a single aquifer, and data from the four grids are combined. A more accurate estimate for aquifer scale proportion could be obtained if each of the grids were evaluated separately. *Landon and*

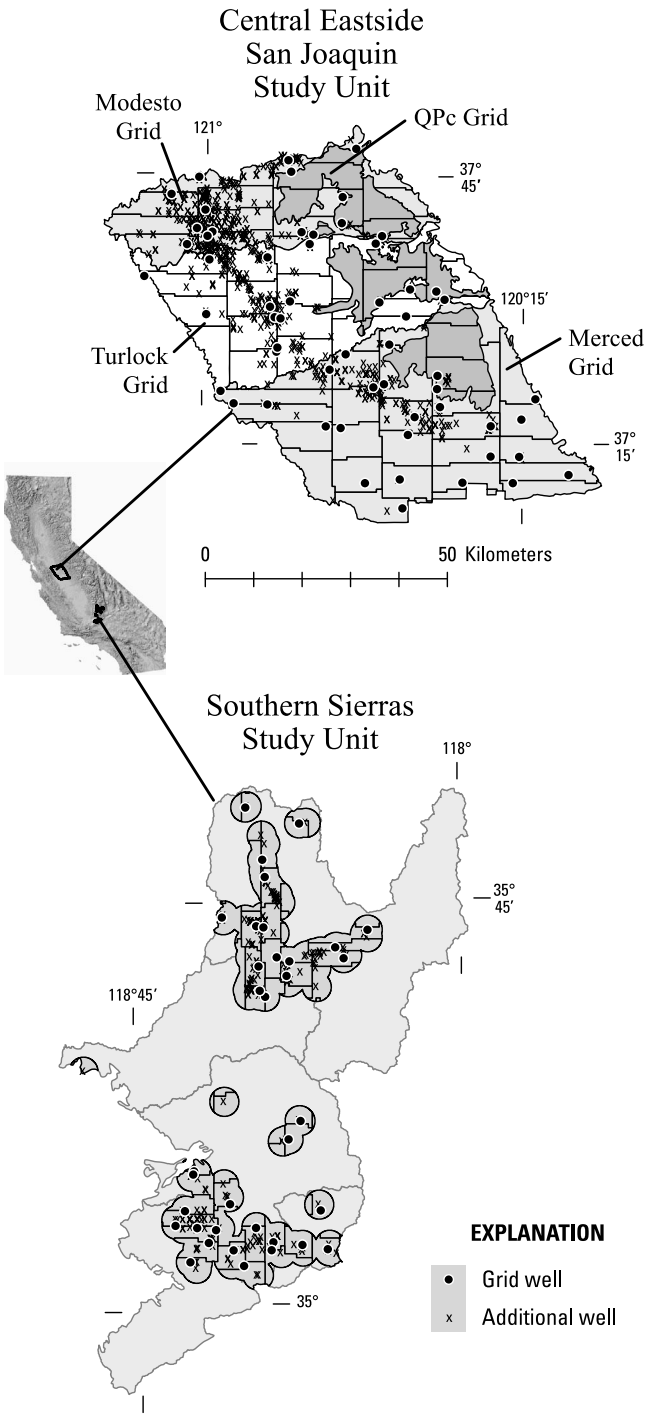

Figure 6. Maps of equal area grids, grid wells (solid circles), and nongrid wells (open circles) in the Central Eastside San Joaquin Study Unit and Southern Sierras Study unit.

Belitz [2008] provide an overview of the study area, a description of the sampling program and summarize the data collected for the Central Eastside study unit.

[58] The CDPH database, along with the data collected by the USGS, was used to identify constituents present at concentrations above health-based benchmarks (Table 1) in the Central Eastside (Table 2). For the purpose of assessing current groundwater quality, the analysis was restricted to the most recent data for each constituent at each well in the CDPH database during the 3 year period from March 2003 through February 2006. For the constituents listed in Table 2, the number of grid wells ranged from 28 to 58. The total number of wells available for each constituent ranged from 133 to 372, and the effective number of wells (equations (10) and (11a)–(11c)) used for computing a confidence interval ranged from 72 to 287.

[59] Four constituents were detected above health-based benchmarks in the grid wells and in the CDPH database (Table 2a) and an additional seven constituents were detected above benchmarks, but only in the CDPH database (Table 2b). For 10 of the 11 constituents, the spatially weighted aquifer proportion is within the 90% confidence interval computed from the grid wells. In contrast, the raw detection frequency is within the 90% grid-based confidence interval for 8 constituents and outside the range for 3. For the 11 constituents as a group, the MAD (equation (9)) is 1.8% for the spatially weighted values and 3.0% for the raw detection frequencies. Overall, spatial weighting shifts the aquifer scale proportion toward the value determined from the equal area grid sampling.

[60] Given the grid-based and spatially weighted estimates of aquifer scale proportion (Table 2), it is illustrative to calculate the size of a small target. The median number of grid wells for the constituents listed in Table 2 is 43 (the average is 44). At a 90% confidence level (equation (20a)), a target is too small to be detected if it is present in less than 0.2% of the aquifer, and it is unlikely to be missed (sufficiently large) if it is present in more than 5.2%.

[61] Table 2a includes constituents detected above benchmarks in both the grid wells and in the CDPH database. All of these constituents have spatially weighted proportions larger than 0.2%. Targets that are too small to be detected (at a 90% confidence level) were not detected using the grid-based sampling design.

[62] Table 2b includes constituents above thresholds only in the CDPH database. Six of the seven constituents have spatially weighted proportions less than 5.2%. One constituent (gross-alpha) has a spatially weighted proportion of 5.9%, and for $n = 37$ a sufficiently large target would be 6%. None of the targets that were missed by the grid-based sampling design were sufficiently large.

Table 1. Health-Based Benchmarks for Constituents Listed in Tables 2 and 3[a]

Constituent	Threshold Type	Threshold Concentration
Antimony	MCL-US	6 mg/L
Arsenic	MCL-US	10 μg/L
Boron	NL-CA	1000 mg/L
Copper	MCL-US	1300 mg/L
DBCP	MCL-US	0.2 mg/L
Fluoride	MCL-CA	2 mg/L
Gross alpha	MCL-US	15 pCi/L
Lead	MCL-US	15 mg/L
Nitrate	MCL-US	10 mg/L
PCE	MCL-US	5 mg/L
Radium	MCL-US	5 pCi/L
Selenium	MCL-US	50 mg/L
Uranium	MCL-US	30 mg/L
Vanadium	NL-CA	50 mg/L

[a]MCL, maximum contaminant level; NL, notification level; US, threshold established by U.S. Environmental Protection Agency; CA, threshold established by California Department of Public Health.

Table 2. Case Study for Concentrations Above Health-Based Benchmarks Listed in Table 1 in the Central Eastside[a]

	Grid Data (One Sample per Cell)				Grid + Additional Data (Many Samples per Cell)					
	Num. Wells		90% CI		Num.	Raw	Spatially Weighted	Effective Num.	90% CI	
Constituent	(Cells)	Prop.	Lower	Upper	Wells	Det. Freq.	Prop.	Wells	Lower	Upper
Constituents With High Concentrations in Grid Wells										
Arsenic	45	15.6%	8.3%	25.9%	348	8.9%	12.7%	72	7.4%	20.3%
Vanadium	28	3.6%	0.6%	13.1%	133	5.3%	1.4%	81	0.3%	5.0%
Lead	42	2.4%	0.4%	8.9%	325	0.6%	0.4%	202	0.1%	1.7%
Nitrate	48	2.1%	0.4%	7.9%	480	5.0%	3.4%	140	1.5%	6.7%
Constituents With High Concentrations in Additional (CDPH) Data Set but not in Grid Wells										
Gross-alpha	37	0.0%	0.0%	3.6%	282	8.2%	5.9%	106	3.0%	10.6%
Uranium	33	0.0%	0.0%	4.0%	191	6.3%	3.6%	107	1.5%	7.5%
DBCP	58	0.0%	0.0%	2.3%	372	3.5%	1.0%	211	0.3%	2.7%
Copper	44	0.0%	0.0%	3.0%	332	0.3%	0.3%	186	0.0%	1.6%
Antimony	43	0.0%	0.0%	3.1%	336	0.3%	0.2%	199	0.0%	1.5%
PCE	58	0.0%	0.0%	2.3%	365	0.8%	0.2%	287	0.0%	1.0%
Selenium	43	0.0%	0.0%	3.1%	337	0.3%	0.2%	228	0.0%	1.2%

[a]For constituents with high concentrations in grid wells. For constituents with high concentrations in additional data set but not in grid wells. The lower and upper confidence limits are computed using the Jeffreys interval. For nonzero proportions, the confidence interval is computed as a two-sided interval; for zero, a one-sided interval is computed. CDPH, California Department of Public Health; CI, confidence interval; Num., number; Prop., proportion; Det. Freq., detection frequency.

9.2. Case Study 2: Concentrations Above Health-Based Thresholds in the Southern Sierra Study Unit

[63] The Southern Sierra study unit is located in the southern end of the Sierra Nevada (Figure 6). The study unit included several small basins of Quaternary fluvial and alluvial deposits and selected areas of Mesozoic granitic and Mesozoic-Paleozoic rocks. The Southern Sierra study unit included all areas within 3 km of a public supply well and was sampled with a single equal area grid, consisting of 40 cells, each with an area of 30 km^2 (Figure 6). Thirty-five grid wells and 15 understanding wells were sampled by the USGS. *Fram and Belitz* [2007] provide an overview of the study area, a description of the sampling program and summarize the data collected for the Southern Sierra study unit.

[64] The USGS sampled wells in the Southern Sierra in June 2006 and the most recent CDPH data were from February 2006. For the purpose of assessing current groundwater quality, the analysis was restricted to the most recent data for each constituent at each well in the CDPH database during the 3 year period from January 2003 through February 2006. For the constituents listed in Table 3, the number of grid wells ranged from 13 to 33, the total number of wells ranged from 43 to 204, and the effective number ranged from 21 to 115. The number of wells available in the Southern Sierra study unit is substantially less than the number of wells available in the Central Eastside study unit.

[65] Six constituents were detected above health-based benchmarks in the grid wells and in the CDPH database (Table 3a). Two additional constituents were detected above benchmarks only in the CDPH database (Table 3b). Overall, the constituents detected above benchmarks in the Southern Sierra study unit were detected more frequently than those in the Central Eastside; the average spatially weighted proportion for the constituents in Table 3 (Southern Sierras) was 9% as compared to 2% in Table 1 (Central Eastside).

[66] For all eight constituents detected above benchmarks in the Southern Sierra study unit, the spatially weighted and the raw detection frequencies are within the 90% confidence intervals computed for the grid wells. For the eight constituents as a group, the MAD (equation (9)) was 1.4% for the spatially weighted values and 3.2% for the raw detection frequencies; spatial weighting shifts the aquifer scale proportion toward the grid-based estimate.

Table 3. Case Study for Concentrations Above Health-Based Benchmarks Listed in Table 1 in the Southern Sierras[a]

	Grid Data (One Sample per Cell)				Grid + Additional Data (Many Samples per Cell)					
	Num. Wells		90% CI		Num.	Raw Det.	Spatially Weighted	Effective Num.	90% CI	
Constituent	(Cells)	Prop.	Lower	Upper	Wells	Freq.	Prop.	Wells	Lower	Upper
Constituents With High Concentrations in Grid Wells										
Arsenic	29	20.7%	11.0%	35.0%	173	14.5%	17.7%	45	10.0%	28.5%
Gross alpha	23	17.4%	8.1%	34.0%	143	13.3%	17.9%	40	9.7%	29.3%
Fluoride	30	10.0%	4.1%	22.0%	168	6.5%	6.7%	49	2.6%	14.5%
Uranium	21	9.5%	3.2%	25.0%	95	7.4%	10.8%	28	4.0%	23.4%
Boron	13	7.7%	1.7%	28.0%	80	3.8%	6.4%	21	1.5%	19.5%
Nitrate	33	3.0%	0.7%	12.5%	204	5.4%	3.5%	104	1.4%	7.5%
Constituents With High Concentrations in Additional (CDPH) Data Set but not in Grid Wells										
Radium	10	0.0%	0.0%	21.0%	43	2.3%	1.1%	33	0.1%	7.9%
Antimony	27	0.0%	0.0%	9.1%	163	0.6%	0.3%	115	0.0%	2.3%

[a]For constituents with high concentrations in grid wells. For constituents with high concentrations in additional data set but not in grid wells. The lower and upper confidence limits are computed using the Jeffreys interval. For nonzero proportions, the confidence interval is computed as a two-sided interval; for zero, a one-sided interval is computed. CDPH, California Department of Public Health; CI, confidence interval; Num, number; Prop., proportion; Det. Freq., detection frequency.

Table 4. Case Study for Detections of Organic Constituents at Any Concentration, Southern Sierras[a]

Compound	Compound Type	USGS MDL (μg/L)	CDPH MDL (μg/L)	Aquifer Scale Proportion		
				GAMA Grid Based	GAMA Spatially Weighted	CDPH Spatially Weighted
Chloroform	Trihalomethane	0.012	0.5	17.1%	16.2%	5.8%
Deethylatrazine	Herbicide degradate	0.007	na	14.3%	15.1%	na
PCE	Solvent	0.015	0.5	14.3%	14.8%	0.5%
Atrazine	Herbicide	0.0035	0.3	14.3%	14.7%	0.0%
Simazine	Herbicide	0.0025	0.3	11.4%	12.4%	0.0%
Prometon	Herbicide	0.005	0.5	5.7%	5.7%	0.0%
CFC-11	Refrigerant	0.04	0.5	2.9%	3.3%	0.0%
Carbon tetrachloride	Solvent	0.03	0.5	2.9%	2.9%	0.0%
CFC-113	Refrigerant	0.019	0.5	2.9%	2.9%	0.0%
TCE	Solvent	0.02	0.5	2.9%	1.4%	0.0%
MTBE	Gasoline oxygenate	0.05	0.5	2.9%	1.4%	0.0%
1,2-Dichlorobenzene	Solvent	0.024	0.5	0.0%	1.0%	0.0%
cis-1,2-Dichloroethene	Solvent	0.012	0.5	0.0%	1.0%	0.0%
1,2-Dichloropropane	Fumigant	0.015	0.5	0.0%	0.4%	0.0%

[a]Grid-based proportion based on 35 grid wells. GAMA spatially weighted proportion based on 50 wells total (in 35 cells). CDPH spatially weighted proportion based on 71–143 wells. na, not available; USGS, United States Geological Survey; CDPH, California Department of Public Health; MDL, method detection limit; μg/L, micrograms per liter.

[67] In the Southern Sierras, as in the Central Eastside, it is illustrative to calculate the size of a small target. The median number of grid wells for the constituents in Table 3 is 25 (the average is 23). At a 90% confidence level, a target is unlikely to be detected if it is present in less than 0.4% of the aquifer; all of the constituents in Table 3a are above this threshold. At a 90% confidence level, a target is unlikely to be missed if it is present in more than 8.8% of the aquifer; all of the constituents in Table 3b are below this threshold. In the Southern Sierras, as in the Central Eastside, a target that is too small was not detected using grid-based sampling, and a target that is sufficiently large was not missed. In the Southern Sierras, a small target is larger than a small target in the Central Eastside, because there are fewer grid wells available in the Southern Sierras.

9.3. Case Study 3: Detections of Organic Compounds in the Southern Sierra Study Unit

[68] The presence of anthropogenic organic constituents at low concentrations in aquifers used for public supply can provide an indication of the extent to which human activities influence groundwater quality. In the Southern Sierra study unit, fourteen organic compounds were detected using low-level analytical methods, but only two were detected using the less sensitive methods required for regulatory compliance (Table 4). None of the detections were above a health-based benchmark; most were at concentrations less than 1/100 of the benchmark [*Fram and Belitz*, 2007]. Five of the organic compounds were prevalent (detection frequency $\geq$ 10%) using low-level analytical methods, but none were prevalent using the less sensitive analytical methods. If one were to rely only on the CDPH data, one would underestimate the extent to which anthropogenic compounds are present in the public supply aquifer system.

[69] Three of the organic compounds were detected only in the understanding wells. The spatially weighted aquifer scale proportion is correspondingly low. If one were to rely only on the grid wells for computation of aquifer scale proportions, then one would not be able to quantify the occurrence of the three compounds detected only in the understanding wells. Cell declustering provides a basis for estimating these aquifer scale proportions. The list of compounds that are prevalent is identical whether one uses only the grid wells or if one uses spatial weighting with all of the available GAMA data. Overall, the aquifer scale proportions estimated using the grid wells is similar to the proportions estimated using spatial weighting of grid and understanding wells (MAD = 0.7%).

10. Conclusions

[70] Aquifer scale proportion can be viewed as a nondimensional measure of regional scale groundwater quality. From that perspective, it can be used as a criterion for determining which constituents in an aquifer are more noteworthy and which are less so. For example, a constituent that is high in 10% of an aquifer could be considered more noteworthy than one that is high in 2% of an aquifer. Aquifer scale proportion can also be used as a criterion for comparing different aquifers: an aquifer with a smaller proportion of high concentrations could be considered to have better water quality than an aquifer with a larger proportion of high concentrations. If one were to use aquifer scale proportion as a measure of water quality, one would clearly want to define what is meant by high concentrations.

[71] Equal area grids and the binomial distribution provide a basis for obtaining a spatially unbiased estimate of the aquifer scale proportion and a confidence interval for that estimate. If one water quality sample is obtained from each grid cell (grid-based approach), the aquifer scale proportion is equal to the observed detection frequency, and the observed number of water quality samples (number of cells) is used to estimate a confidence interval. The Jeffreys confidence interval is identified as the preferred method among a number of methods considered. If one were to use a grid-based approach, one would want to insure that the equal area grid is representative of the area under consideration.

[72] If many wells are available per cell, then one needs to account for the potential effects of spatial correlation. One simple approach is cell declustering, whereby each well is assigned a weight proportional to the number of cells in the grid and the number of wells in the cell containing that well. The aquifer scale proportion is then computed as a weighted sum. Confidence intervals for the spatially weighted estimate

can be obtained using Kish's design effect to compute an effective sample size. If the cell declustering method is viewed as a clustered design, the effective sample size ranges from the number of cells in the grid to the total number of wells. Given this range, the confidence interval for the spatially weighted proportion is equal to or narrower than the confidence interval for the grid-based proportion. If the cell declustering method is viewed as a stratified design, then the effective sample size could be larger than the total number of wells. This could lead to a substantial narrowing of the computed confidence interval. Identification of an appropriate design effect remains an important issue to be addressed.

[73] The binomial distribution is used to evaluate the adequacy of the grid-based approach for identifying a small target, which is defined as a constituent with high concentrations in a small proportion of the aquifer (p_s). At a 90% (one-sided) confidence level, a small target is unlikely to be detected if $p_s \leq 0.1/n$ (where n is the number of cells) and is unlikely to be missed if $p_s > 2.2/n$. These bounds provide perspective for the interpretation of results from a grid-based sampling program. For example, if a grid consists of 50 cells, a target is unlikely to be detected if it is present in less than 0.2% of the aquifer, and it is unlikely to be missed if it is present in more than 4.4%. These bounds can also be used to design a grid network. For example, if one wants to identify constituents that are present at high concentrations in 5% of an aquifer system (at a 90% confidence level), then one would need to design a grid with 44 cells. If one can subsequently bring additional data into the analysis, then one can identify smaller targets than if one relies only on one sample per grid cell.

[74] The methods presented in this paper are applied to three case studies in California: (1) concentrations above health-based benchmarks in the Central Eastside of the San Joaquin Valley, (2) concentrations above health-based benchmarks in the Southern Sierras, and (3) detections of volatile organic compounds and pesticides above analytical limits in the Southern Sierras. The first two case studies demonstrate a consistency between grid-based and spatially weighted estimates of aquifer scale proportion. The first two case studies also illustrate the benefit of having more data available: the width of the confidence interval is reduced and smaller targets are identified. The third case study indicates the value of better quality data: lower detection limits provide a more accurate assessment of the presence of anthropogenic compounds in an aquifer system.

[75] The use of aquifer scale proportion as a measure of regional scale groundwater quality provides an objective basis for comparing different constituents (or groups of constituents) to one another, for comparing one aquifer to another, and for obtaining a better understanding of the factors affecting groundwater quality [*Broers*, 2002, 2004]. The approach presented in this paper is particularly useful when one is evaluating a large number of constituents in a large number of wells in a large number of aquifers. The approach presented in this paper can be generalized to contaminants in media other than groundwater.

[76] **Acknowledgments.** The authors appreciate the financial support of the California State Water Resources Control Board, the cooperation of well owners in California, the efforts of USGS colleagues who participated in the sampling program, and the reviewers at the USGS and WRR who helped to improve the paper.

References

Agresti, A., and B. A. Coull (1998), Approximate is better than "exact" for interval estimation of binomial proportions, *Am. Stat.*, *52*(2), 119–126.

Alley, W. (Ed.) (1993), *Regional Ground-Water Quality*, 634 p., Van Nostrand Reinhold, New York.

Backman, B., D. Bodis, P. Lahermo, S. Rapant, and T. Tarvainen (1998), Application of a groundwater contamination index in Finland and Slovakia, *Environ. Geol.*, *36*(1–2), 55–64.

Belitz, K., N. M. Dubrovsky, K. Burow, B. Jurgens, and T. Johnson (2003), Framework for a ground-water quality monitoring and assessment program for California, *U. S. Geol. Surv. Water Resour. Invest. Rep. 03-4166*, 78 p.

Berger, J. O. (1985), *Statistical Decision Theory and Bayesian Analysis*, 2nd ed., Springer, New York.

Blyth, C. R., and H. A. Still (1983), Binomial confidence intervals, *J. Am. Stat. Assoc.*, *78*, 108–116.

Broers, H. P. (2002), Strategies for regional groundwater quality monitoring, Ph.D. thesis, 231 pp., Utrecht University, Utrecht.

Broers, H. P. (2004), The spatial distribution of groundwater age for different geohydrological situations in the Netherlands: Implications for groundwater quality monitoring at the regional scale, *J. Hydrol.*, *299*, 84–106.

Brown, L. D., T. T. Cai, and A. DasGupta (2001), Interval estimation for a binomial proportion, *Stat. Sci.*, *16*(2), 101–133.

Brown, L. D., T. T. Cai, and A. DasGupta (2002), Confidence intervals for a binomial proportion and asymptomatic expansions, *Ann. Stat.*, *30*(1), 160–201.

Cai, T. T. (2005), One-sided confidence intervals in discrete distributions, *J. Stat. Plan. Inf.*, *131*, 63–88.

Campbell, M. J., A. Donner, and N. Klar (2007), Developments in cluster randomized trials and Statistics in Medicine, *Stat. Med.*, *26*, 2–19.

Clopper, C. J., and E. S. Pearson (1934), The use of confidence intervals or fiducial limits illustrated in the case of the binomial, *Biometrika*, *26*(4), 404–413.

Cochran, W. G. (1977), *Sampling Techniques*, 3rd ed., 428 p., John Wiley, New York.

Fram, M. S., and K. Belitz (2007), Ground-water quality data in the Southern Sierra Study Unit, 2006-results from the California GAMA program, *U.S. Geol. Surv. Data Ser. 301*.

Gilbert, R. O. (1987), *Statistical Methods for Environmental Pollution Monitoring*, 320 p., Van Nostrand Reinhold, New York.

Grath, J., R. Ward, A. Scheidleder, and P. Quevauviller (2007), Report on EU guidance on groundwater monitoring developed under the common implementation strategy of the water framework directive, *J. Environ. Monit.*, *9*, 1162–1175.

Helsel, D. R., and R. M. Hirsch (2002), Statistical methods in water resources, *U.S. Geol. Surv., Tech. Water-Resour. Invest. Book 4*, Chap. A3, 348 p.

Isaaks, E. H., and R. M. Srivastava (1989), *An Introduction to Applied Geostatistics*, 561 p., Oxford Univ. Press, New York.

Journel, A. G. (1983), Nonparametric estimation of spatial distributions, *Math. Geol.*, *15*(3), 445–468.

Journel, A. G., and C. J. Huijbregts (1978), *Mining Geostatistics*, 600 p., Academic, London.

Kish, L. (1965), *Survey Sampling*, 641 p., John Wiley, New York.

Kish, L. (1995), Methods for design effects, *J. Official Stat.*, *11*, 55–77.

Landon, M. K., and K. Belitz (2008), Ground-water quality data in the Central Eastside San Joaquin Basin 2006: Results from the California GAMA program, *U.S. Geol. Surv. Data Ser. 325*, 88 p.

Lapham, W. W., P. A. Hamilton, and D. N. Myers (2005), National water quality assessment program—Cycle II: Regional assessment of aquifers, *U.S. Geol. Surv. Fact Sheet 2005-3013*, 4 p.

Lesage, S. (2004), Groundwater quality in Canada, a national overview, in *Bringing Groundwater Quality Research to the Watershed Scale*, *IAHS Publication 297*, edited by N. R. Thomson, pp. 11–18, IAHS Press, Wallingford, UK.

Mendizabal, I., and P. J. Stuyfzand (2009), Guidelines for interpreting hydrochemical patterns in data from public supply well fields and their value for natural background groundwater quality determination, *J. Hydrol.*, *379*, 151–163.

McBride, G. B. (2003), Confidence of compliance: Parametric versus nonparametric approaches, *Water Res.*, *37*(15), 3666–3671.

McBride, G. B., and J. C. Ellis (2001), Confidence of compliance: A Bayesian approach for percentile standards, *Water Res.*, *35*(5), 1117–1124.

Ott, R. L., and M. Longnecker (2001), *An Introduction to Statistical Methods and Data Analysis*, 5th ed., 1152 p., Duxbury, Pacific Grove.

Reijnders, H. F. R., G. Van Drecht, H. F. Prins, and L. J. M. Boumans (1998), The quality of the groundwater in the Netherlands, *J. Hydrol.*, *207*(3–4), 179–188.

Rentier, C., F. Delloye, Brouyere, and A. Dassargues (2006), A framework for an optimized groundwater monitoring network and aggregated indicators, *Environ. Geol.*, *50*, 194–201.

Rosen, M. R., and W. W. Lapham (2008), Introduction to the U.S. Geological Survey National Water quality Assessment (NAWQA) of ground-water quality trends and comparison to other national programs, *J. Environ. Qual.*, *37*, S190–S198.

Scott, J. C. (1990), Computerized stratified random site-selection approaches for design of a ground-water quality sampling network, *U.S. Geol. Surv. Water Resour. Invest. Rep., 90-4101*, 109 p.

Shelton, J. L., K. R. Burow, K. Belitz, N. M. Dubrovsky, M. Land, and J. Gronberg (2001), Low-level volatile organic compounds in active public supply wells as groundwater tracers in the Los Angeles Physiographic Basin, California, 2000, *U.S. Geol. Surv. Water Resour. Invest. Rep. 01-4188*, 29 p.

Smith, E. P., Y. E. Keying, C. Hughes, and L. Shabman (2001), Statistical assessment of violations of water quality standards under section 303(d) of the Clean Water Act, *Environ. Sci. Technol.*, *35*, 606–612.

Stigter, T. Y., L. Ribeiro, and A. M. M. Carvaho Dill (2006), Application of a groundwater quality index as an assessment and communication tool in agro-environmental policies—Two Portuguese case studies, *J. Hydrol.*, *327*, 578–591.

Toccalino, P., and J. F. Norman (2006), Health-based screening levels to evaluate U.S. Geological Survey ground water quality data, *Risk Analysis*, *25*(5), 1339–1348.

Vollset, S. E. (1993), Confidence intervals for a binomial proportion, *Stat. Med.*, *12*, 809–824.

Ward, R. S., M. J. Streetly, A. J. Singleton, and R. Sears (2004), A framework for monitoring regional groundwater quality, *Q. J. Eng. Geol. Hydrogeol.*, *37*, 271–281.

Wendland, F., et al. (2008), European aquifer typology: A practical framework for an overview of majr groundwater composition at European scale, *Environ. Geol.*, *55*, 77–85.

Worrall, F., and T. Besien (2005), The vulnerability of groundwater to pesticide contamination estimated directly from observations of presence or absence in wells, *J. Hydrol.*, *303*, 92–107.

Worrall, F., and D. W. Kolpin (2003), Direct assessment of groundwater vulnerability from single observations of multiple contaminants, *Water Resour. Res.*, *39*(12), 1345, doi:10.1029/2002WR001212.

K. Belitz, T. Johnson, and M. K. Landon, USGS California Water Science Center, 4165 Spruance Rd., Ste. 200, San Diego, CA 92101, USA. (kbelitz@usgs.gov)

M. S. Fram and B. Jurgens, USGS California Water Science Center, Placer Hall, 6000 J St., Sacramento, CA 95819, USA.