Geostatistical modeling of the spatial variability of arsenic in groundwater of southeast Michigan

P. Goovaerts,[1] G. AvRuskin,[1] J. Meliker,[2] M. Slotnick,[2] G. Jacquez,[1] and J. Nriagu[2]

Received 3 October 2004; revised 14 March 2005; accepted 6 April 2005; published 14 July 2005.

[1] During the last decade one has witnessed an increasing interest in assessing health risks caused by exposure to contaminants present in the soil, air, and water. A key component of any exposure study is a reliable model for the space-time distribution of pollutants. This paper compares the performances of multi-Gaussian and indicator kriging for modeling probabilistically the spatial distribution of arsenic concentrations in groundwater of southeast Michigan, accounting for arsenic data collected at private residential wells and the hydrogeochemistry of the area. The arsenic data set, which was provided by the Michigan Department of Environmental Quality (MDEQ), includes measurements collected between 1993 and 2002 at 8212 different wells. Factorial kriging was used to filter the short-range spatial variability in arsenic concentration, leading to a significant increase (17–65%) in the proportion of variance explained by secondary information, such as type of unconsolidated deposits and proximity to Marshall Sandstone subcrop. Cross validation of well data shows that accounting for this regional background does not improve the local prediction of arsenic, which reveals the presence of unexplained sources of variability and the importance of modeling the uncertainty attached to these predictions. Slightly more precise models of uncertainty were obtained using indicator kriging. Well data collected in 2004 were compared to the prediction model and best results were found for soft indicator kriging which has a mean absolute error of 5.6 μg/L. Although this error is large with respect to the USEPA standard of 10 μg/L, it is smaller than the average difference (12.53 μg/L) between data collected at the same well and day, as reported in the MDEQ data set. Thus the uncertainty attached to the sampled values themselves, which arises from laboratory errors and lack of information regarding the sample origin, contributes to the poor accuracy of the geostatistical predictions in southeast Michigan.

Citation: Goovaerts, P., G. AvRuskin, J. Meliker, M. Slotnick, G. Jacquez, and J. Nriagu (2005), Geostatistical modeling of the spatial variability of arsenic in groundwater of southeast Michigan, *Water Resour. Res.*, *41*, W07013, doi:10.1029/2004WR003705.

1. Introduction

[2] Assessment of the health risks associated with exposure to elevated levels of contaminants has become the subject of considerable interest in our societies. Environmental exposure assessment is frequently hampered by the existence of multiple confounding risk factors (e.g., smoking, diet, stress, ethnicity, home and occupational exposures) leading to complex exposure models, and the mobility of the population which can interact with many different sources of exposure over the lifetime. A rigorous study thus requires modeling the study subjects as spatiotemporally referenced objects that move through space and time, their cumulative exposure increasing as they come in contact with sources of contamination [*AvRuskin et al.*, 2004]. The computation of human exposure is particularly challenging for cancers because they usually take years or decades to develop, especially in presence of low level of contaminants. For example, *Steinmaus et al.* [2003] found that the latency of arsenic-caused cancer may be greater than 40 years. In this situation contaminant concentrations are rarely available for every location and time interval visited by the subjects; therefore data gaps need to be filled in through space-time interpolation.

[3] Geostatistical spatiotemporal models provide a probabilistic framework for data analysis and predictions that build on the joint spatial and temporal dependence between observations (e.g., see *Kyriakidis and Journel* [1999] for a review). Geostatistical tools have been applied to the modeling of spatiotemporal distributions in many disciplines, such as environmental sciences (e.g., deposition of atmospheric pollutants), ecology (characterization of population dynamics) and health (patterns of diseases and exposure to pollutants). These tools are increasingly coupled with GIS capabilities [*Burrough and McDonnell*, 2000] for applications that characterize space-time structures (semivariogram analysis), spatially interpolate scattered measurements to create spatially exhaustive layers of information and assess the corresponding accuracy and precision. Of critical importance when coupling GIS data and environmental/exposure models is the issue of error propagation, that is how the uncertainty in input data (e.g.,

[1] BioMedware, Inc., Ann Arbor, Michigan, USA.
[2] School of Public Health, University of Michigan, Ann Arbor, Michigan, USA.

0043-1397/05/2004WR003705

arsenic concentrations) translates into uncertainty about model outputs (e.g., risk of cancer). Methods for uncertainty propagation [*Heuvelink*, 1998; *Goovaerts*, 2001], such as Monte Carlo analysis, are critical for estimating uncertainties associated with spatially based policies in the area of environmental health, and in dealing effectively with risks [*Goodchild*, 1996].

[4] A bladder cancer case control study is underway in Michigan (11 counties) to evaluate risks associated with exposure to low levels of arsenic in drinking water (typically, 5–100 μg/L). A key part of this study is the creation of a space-time information system (STIS) to visualize and analyze the spatiotemporal mobility of study participants and their surrounding environment [*Jacquez et al.*, 2004], leading to the estimation of individual-level historical exposure to arsenic. In Michigan, many generations have depended on groundwater as their source of drinking water and they have experienced lifetime exposures to elevated concentrations of arsenic, an unwanted chemical in their water supply derived from local rocks [*Kolker et al.*, 1998]. A model of the space-time distribution of arsenic in groundwater is thus an important layer of the STIS.

[5] Geostatistics has been used recently to evaluate the spatial variability of arsenic contamination in the groundwater of the continental United States. For example, *Ryker* [2001] plotted arsenic concentrations analyzed in water samples collected from approximately 31,000 wells across the United States and developed a national-scale point map of arsenic concentrations. *Warner et al.* [2003] used cokriging to interpolate arsenic concentrations across Illinois using arsenic, iron, and manganese concentrations measured in 1449 community water supplies that utilize the glacial and alluvial aquifer. Similar geostatistical studies were conducted in other states, such as Idaho [*Welhan and Merrick*, 2003] or Michigan [*Aichele and Shortridge*, 2002]. A vast body of the literature relates to the mapping of the widespread groundwater contamination in Bangladesh [*British Geological Survey and of Department of Public Health Engineering* (*BGS and DPHE*), 2001; *Frisbie et al.*, 2002]. *Karthik et al.*'s [2001] semivariogram analysis suggested that the complex spatial distribution of high-level arsenic concentrations is a consequence of interactions among multiscale geologic and geochemical processes. *Serre et al.* [2003] showed that most of the variability in arsenic concentrations across Bangladesh occurs within a distance of 2 km, which makes spatial interpolation very challenging. *Yu et al.* [2003] found that much of the large-scale (>10 km) variability in arsenic concentration is explained by differences in geology and geomorphology, while small-scale (<3 km) variability is mainly due to variations in well depth, with lower concentrations being measured in deeper wells.

[6] This paper presents a study of the spatial variability of arsenic concentrations in groundwater of southeast Michigan, accounting for information collected at private residential wells and the hydrogeochemistry of the area. Variants of traditional semivariogram estimators and kriging systems are introduced to tackle specific features of the data set, such as the preferential sampling of high concentrations, the existence of repeated measurements at hundreds of wells, and the presence of high variability over very short distances. Cross validation is used to assess the prediction performances and quality of uncertainty models provided by multi-Gaussian and indicator approaches. Last, recently collected well data allow us to assess how well the spatial variability of arsenic is captured by the geostatistical model.

2. Data Sets

2.1. Well Data

[7] Elevated concentrations of arsenic in drinking water have been identified in groundwater supplies of 11 counties in southeastern Michigan: Genesee, Huron, Ingham, Jackson, Lapeer, Livingston, Oakland, Sanilac, Shiawassee, Tuscola, and Washtenaw [*Kim et al.*, 2002; *Kolker et al.*, 2003; *Slotnick et al.*, 2003]. The spatial distribution of arsenic over these counties will be modeled using 9188 data recorded at 8212 different wells and stored in the Michigan Department of Environmental Quality (MDEQ) database of arsenic measurements (Figure 1a). These data were collected at private wells sampled between 1993 and 2002. Graphite furnace atomic absorption spectrometry (AAS) and hydride flame (quartz tube AAS) were used to measure samples from 1989–1995; inductively coupled plasma/mass spectrometry has been used since 1996. Wells were not randomly sampled in this database, and quality control of water sampling was executed to varying degrees over time. Less than 10% of the measurements (737 observations) were below the detection limit and these were reset to half the value of the detection limit at that time; that is 0.15 μg/L for 12 wells, 0.5 μg/L for 670 wells, and 1.0 μg/L for 55 wells.

[8] The sample distribution is positively skewed, with a mean that is more than twice the median value of 6 μg/L. A lognormal transform was performed and the corresponding histogram is displayed in Figure 1b. It is noteworthy that a mere log transform does not make the distribution symmetric but it reveals the existence of two modes, the first one corresponding to half the detection limit of 1.0 μg/L. A graphical normal score transform will be introduced in section 3.2.1 to achieve a "perfectly" Gaussian univariate distribution required for the application of multi-Gaussian kriging.

[9] At 662 wells concentrations were measured multiple times (2–14 times) on the same day or up to 113 months apart, with an average time interval of 14 days. These very short time series do not allow a modeling of the temporal autocorrelation of the data, but they may provide useful information on the relative importance of spatial, temporal and laboratory variabilities. Unfortunately, the MDEQ arsenic database inconsistently reports changes in water treatment practices by homeowners, which makes it difficult to interpret apparent temporal variability. The average difference between data collected at the same well is 12.53 μg/L (median = 3.60) for same day measurements while it is 15.92 μg/L (median = 4.20) for multiple date measurements. In comparison, the average difference between noncolocated data measured the same day anywhere within the 11 counties is 17.00 μg/L (median = 8.40). Information regarding the sample origin, which could have explained the important variability observed between measurements taken the same day at the same well, is lacking. Hereafter the variability between same day measurements will be referred to as measurement errors, although it clearly

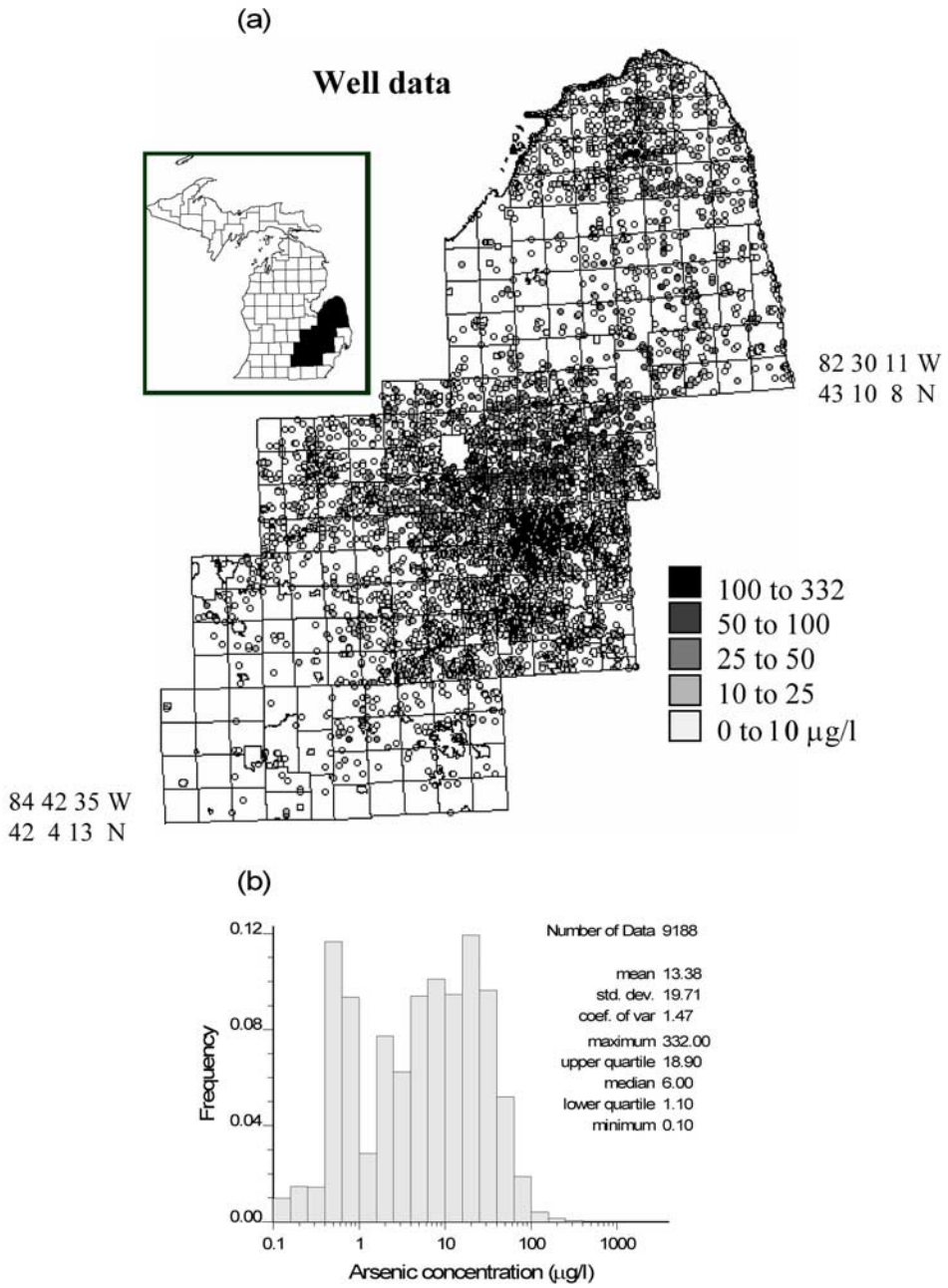

Figure 1. The 8212 well data observations (MDEQ database, 1993–2002) available for modeling. (a) Location map with the township boundaries overlaid and the location of the study area within the state of Michigan. (b) Histogram of arsenic concentration (μg/L).

includes other sources of variability. The small magnitude of temporal variation relative to the variability in space or arising from measurement error, as well as the absence of temporal trend or seasonality, led us to ignore the temporal dimension in this study. Another reason for this simplification is that a space-time model only addressing the period 1993–2002 would not help in characterizing lifetime exposure for most participants in this epidemiological study.

[10] Since this database contains arsenic measurements requested by homeowners, we hypothesized that sampling would be more dense in areas where higher pollutant concentrations were initially reported. The existence of such a preferential sampling was investigated by plotting the average arsenic concentration measured for pairs of wells as a function of their separation distance (Figure 2a). The average concentration at multisampled wells (i.e., distance = 0) is 24.75 μg/L, and it still exceeds 19.00 μg/L for separation distances smaller than 150 meters. Both values are substantially higher than the global mean of 13.38 μg/L, and this curve clearly indicates the existence of preferential sampling at high-valued wells. Because of the positive relationship between the local mean and variance of the data, known as direct proportional effect [*Goovaerts*, 1997], the clustering of high values entails a bias in the estimation of short-range variability. In other words, most data pairs that contribute to small separation distances come from these high-valued

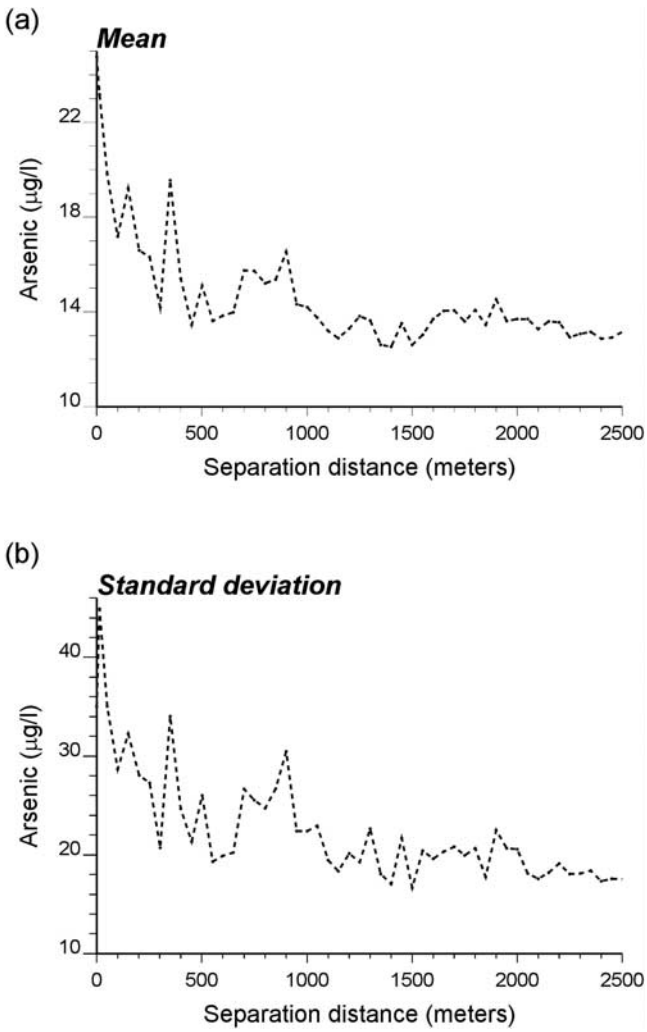

Figure 2. Plots of (a) mean and (b) standard deviation of arsenic concentrations measured at increasingly distant wells. The higher concentrations observed for closer wells reflects the preferential sampling of areas where elevated concentrations of arsenic are anticipated.

areas, leading to an overestimation of the variability of arsenic concentration for short distances (Figure 2b). This effect will have to be taken into account during the modeling of the spatial variability, see section 3.1.

[11] Another consequence of the preferential sampling is that sample statistics, such as the mean and standard deviation displayed in Figure 1b, are not representative of the whole study area. The uneven sampling of MDEQ data was corrected using the cell-declustering technique [*Deutsch and Journel*, 1998] which calls for dividing the study area into rectangular cells; then each observation within a cell is assigned a weight inversely proportional to the number of data within that cell. These declustering weights are used, instead of equal weights, in the computation of summary statistics. This correction gives more importance to isolated wells which tend to be located in areas with low levels of arsenic in groundwater. Square cells of 2.5 km were used since they lead to the smallest declustered mean, which is the target because of the preferential sampling of high-valued areas. The declustered mean and standard deviation are, respectively, 10.97 and 15.22 μg/L, which is much smaller than the estimates obtained without regard to the clustering of high-valued wells. This declustered distribution will be used hereafter for the normal score transform of arsenic concentrations and the interpolation/extrapolation of probability values estimated by indicator kriging.

2.2. Secondary Information

[12] Elevated levels of naturally occurring arsenic have been identified in regional patterns within the United States and are attributed to geochemistry, geology, climate, and glacial history [*Welch et al.*, 2000]. In the Michigan thumb region, arsenian pyrite (up to 7% As by weight) has been identified in the bedrock of the Marshall Sandstone aquifer, one of the region's most productive aquifers [*Westjohn et al.*, 1998]. The mechanisms responsible for arsenic mobilization into groundwater supplies in Michigan, however, are not well understood. Geochemical analyses reveal that arsenic is not likely to be oxidized out of the bedrock since the groundwater is reducing; suggesting there must be another mechanism to explain the elevated arsenic levels in the water [*Kolker et al.*, 1998]. Arsenian pyrite grains have also been identified in the glacial till, where the conditions are favorable for the oxidation of arsenic into the water [*Kolker et al.*, 2003]. In addition to arsenian pyrite, arseniferrous iron oxy-hydroxides have been found in Marshall Sandstone till fragments [*Kolker et al.*, 2003]; these arsenic-rich iron oxy-hydroxides may be undergoing reductive dissolution.

[13] Literature suggests that geochemical properties of the aquifer, as well as characteristics of the wells (i.e., well depth, casing depth, and depth of bedrock-unconsolidated interface), might explain part of the spatial variability of arsenic concentrations. In this paper, the focus has been on variables that can be retrieved easily at each location across the eleven counties; hence well characteristics, although potentially important factors, have not been considered since this information is available only at recently drilled wells. Maps of unconsolidated deposits and bedrock subcrops, however, were retrieved from the Michigan Center for Geographic Information, Geographic Data Library (http://www.mcgi.state.mi.us/mgdl/). The following layers of secondary information have been built using a Geographical Information System: type of bedrock and unconsolidated deposits, and proximity of well to the Marshall Sandstone subcrop, where the highest concentrations of arsenic were found [*Kim*, 1999]. Multivariate regression was conducted using the aforementioned attributes and a quadratic function of the spatial coordinates as explanatory variables. A very small R^2 of 17.3% was obtained, which is likely caused by the large variability of arsenic concentration over very short distances and the magnitude of measurement and data entry errors.

2.3. Validation Data Set (2004 Campaign)

[14] Additional well data were collected in 2004 at the homes of 73 participants of the cancer case control study. Water samples were collected from the kitchen tap, or primary source of drinking water, for participants with private wells. If an in-line softener or filter was present at the tap, a second sample was collected prior to the treatment; this sample was often taken from an outside or basement spigot. Only samples

taken in the absence of a softener or filter will be used for validation purposes. To flush out standing water in the pipes, the faucet was run for at least a minute prior to sample collection. Water samples were collected in 60 ml acid-washed polyethylene bottles, stored on ice, acidified with 0.2% trace metal grade nitric acid, and refrigerated until analysis. Field blanks and replicates were collected for 10% of the samples. Laboratory analysis for arsenic was done using an inductively coupled plasma mass spectrometer (ICP-MS, Argilent Technologies Model 7500c).

[15] Figure 3 shows the location of the 73 validation wells and the histogram of measured arsenic concentration. Unlike the MDEQ sampling campaign, wells with high concentrations have not been sampled preferentially; hence the sample mean is half the average concentrations of data used for the spatial modeling.

3. Geostatistical Approach

[16] Geostatistics is used to model the uncertainty about the arsenic concentration in the drinking water at any location $\mathbf{u}$ in the study area $\mathcal{A}$. This model takes the form of a conditional cumulative distribution function (ccdf) $F(\boldsymbol{u};z|(\text{Info}))$ which gives the probability that the concentration is no greater than any given threshold z. The conditioning information, "Info", consists of the set of n = 9,188 well concentrations $\{z(\boldsymbol{u}_\alpha);\ a = 1,\ldots,n\}$ plus the L secondary attributes available across the study area $\{z_l(\boldsymbol{u});\ l = 1,\ldots,L,\ \forall\ \boldsymbol{u} \subset \mathcal{A}\}$.

3.1. Semivariogram Estimation and Modeling

[17] All techniques described in section 3 capitalize on the presence of spatial correlation between either the raw arsenic concentrations or their transforms to make predictions (using kriging) at unsampled locations. Although all kriging systems are written in terms of covariances, common practice consists of computing and modeling the semivariogram rather than the covariance function. The experimental semivariogram for a given lag vector $\mathbf{h}$ is estimated as

$$\hat{\gamma}(\mathbf{h}) = \frac{1}{2N(\mathbf{h})} \sum_{\alpha=1}^{N(\mathbf{h})} [z(\mathbf{u}_\alpha) - z(\mathbf{u}_\alpha + \mathbf{h})]^2 \quad (1)$$

where $N(\mathbf{h})$ is the number of data pairs within the class of distance and direction used for the lag vector $\mathbf{h}$. To correct for the preferential sampling of high values, the following rescaled semivariogram estimator implemented in Variowin [*Pannatier*, 1996] was used:

$$\hat{\gamma}(\mathbf{h}) = \frac{1}{2N(\mathbf{h})\sigma^2(\mathbf{h})} \sum_{\alpha=1}^{N(\mathbf{h})} [z(\mathbf{u}_\alpha) - z(\mathbf{u}_\alpha + \mathbf{h})]^2 \quad (2)$$

where $\sigma^2(\mathbf{h})$ is the variance of the $2N(\mathbf{h})$ data used for estimation at lag $\mathbf{h}$. The rescaling accounts for the large change in variance from one lag to the next (recall Figure 2b), leading to a semivariogram with much less erratic fluctuations and a more accurate estimate of the short-range variability (i.e., nugget effect).

[18] A continuous function must be fitted to $\hat{\gamma}(\mathbf{h})$ in order to compute semivariogram values for any possible lag $\mathbf{h}$ required by prediction algorithms, and also to smooth out sample fluctuations. In this paper, the semivariograms were modeled using least squares regression [*Pardo-Iguzquiza*, 1999] under the constraint of reproduction of the nugget effect estimated from colocated well measurements. All semivariogram models were bounded (i.e., reached a sill), and the covariance models were derived by subtracting the semivariogram model from the sill value.

3.2. Multi-Gaussian Approach

[19] Under the multi-Gaussian (MG) model, the ccdf at any location $\mathbf{u}$ is Gaussian and fully defined by its mean and variance which can be estimated by kriging [*Goovaerts*, 2001]. The approach typically requires a prior normal score transform of data to ensure that at least the univariate distribution (histogram) is normal. The normal score ccdf then undergoes a back transform to yield the ccdf of the original variable.

3.2.1. Normal Score Transform

[20] Normal score transform is a graphical procedure that normalizes any distribution, regardless of its shape. It can be seen as a correspondence table between equal p quantiles z_p and y_p of the z cdf $F(z)$ (cumulative distribution function) and the standard Gaussian cdf $G(y)$. In practice, the normal score transform proceeds in three steps.

[21] 1. The n original data $z(\boldsymbol{u}_\alpha)$ are ranked in ascending order. Since the normal score transform must be monotonic, ties in z values must be broken, which may be a problem in presence of a large proportion of censored data (i.e., non detects). In this paper, such untying or despiking has been done randomly as implemented in GSLIB software [*Deutsch and Journel*, 1998].

[22] 2. The sample cumulative frequency of the datum $z(\boldsymbol{u}_\alpha)$, denoted p_k^*, in the declustered sample distribution is computed.

[23] 3. The normal score transform of the z datum with rank k is matched to the p_k^* quantile of the standard normal cdf:

$$y(\mathbf{u}_\alpha) = \phi(z(\mathbf{u}_\alpha)) = G^{-1}[F(z(\mathbf{u}_\alpha))] = G^{-1}[p_k^*] \quad (3)$$

3.2.2. Multi-Gaussian Kriging

[24] The probability distribution of the normal score variable Y at location $\mathbf{u}$ is Gaussian. Its mean and standard deviation are the ordinary kriging (OK) estimate $y_{OK}^*(\mathbf{u})$ and simple kriging (SK) standard deviation $\sigma_{SK}^*(\mathbf{u})$ computed from the normal score data:

$$F_Y(\mathbf{u};y|(\text{Info})) = G\big[\big(y - y_{OK}^*(\mathbf{u})\big)/\sigma_{SK}^*(\mathbf{u})\big] \quad (4)$$

The OK estimate is computed as a linear combination of $n(\boldsymbol{u})$ surrounding normal score data:

$$y_{OK}^*(\mathbf{u}) = \sum_{\alpha=1}^{n(\mathbf{u})} \lambda_\alpha y(\mathbf{u}_\alpha) \quad (5)$$

The kriging weights λ_α are calculated by solving the following system of linear equations:

$$\sum_{\beta=1}^{n(\mathbf{u})} \lambda_\beta \big[C\big(\mathbf{u}_\alpha - \mathbf{u}_\beta\big) - \delta_{\alpha,\beta} b_0\big] + \mu = C(\mathbf{u}_\alpha - \mathbf{u}) \quad \alpha = 1,\ldots,n(\mathbf{u})$$
$$\sum_{\beta=1}^{n(\mathbf{u})} \lambda_\beta = 1 \quad (6)$$

where $C(\boldsymbol{u}_\alpha - \boldsymbol{u}_\beta)$ is the covariance function of the normal score variable Y for the separation vector $\mathbf{h}_{\alpha\beta} = \boldsymbol{u}_\alpha - \boldsymbol{u}_\beta$,

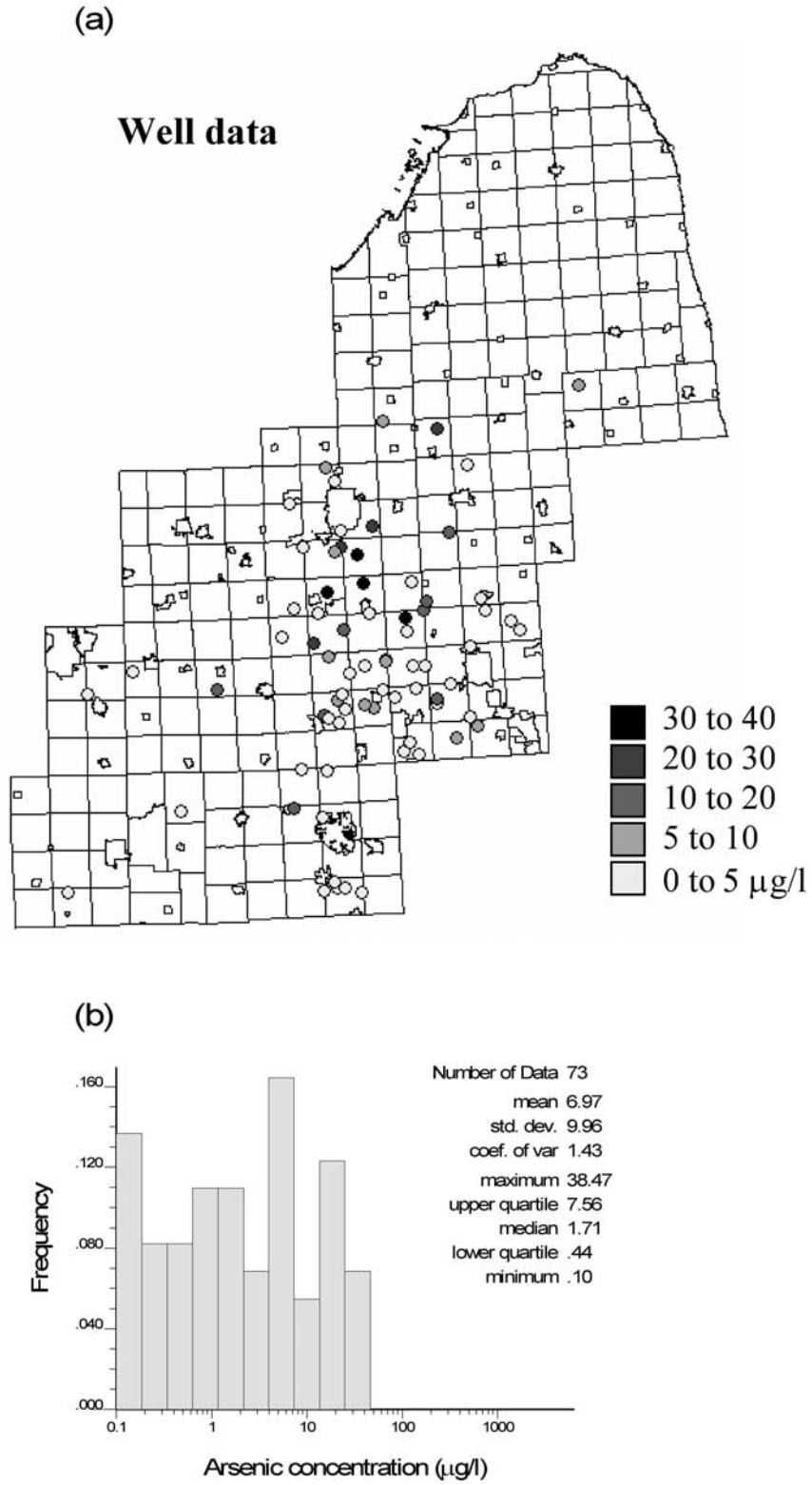

Figure 3. The 73 validation wells sampled in 2004. (a) Location map with the township boundaries overlaid. (b) Histogram of arsenic concentration (µg/L).

μ is a Lagrange multiplier that results from minimizing the estimation variance subject to the unbiasedness constraint on the estimator, $\delta_{\alpha\beta} = 1$ if $\boldsymbol{u}_\alpha = \boldsymbol{u}_\beta$ with $\alpha \neq \beta$, and 0 otherwise. The parameter b_0 is the nugget effect (i.e., variability for a distance of zero) which was estimated using the set of colocated observations at 662 wells. This system is modified from the traditional OK system so that data with the same spatial coordinates can be incorporated

without making the kriging matrix noninvertible. The standard deviation of the probability distribution is computed as

$$\sigma^*_{SK}(\mathbf{u}) = \left[1 - \sum_{\alpha=1}^{n(\mathbf{u})} \lambda'_\alpha C(\mathbf{u}_\alpha - \mathbf{u})\right]^{1/2} \quad (7)$$

where the kriging weights λ'_α are obtained by solving a system similar to equation (6) except that no constraint is imposed on the weights (simple kriging).

3.2.3. Accounting for Secondary Information

[25] Several approaches are available for incorporating secondary data in the estimation procedure [e.g., see *Goovaerts*, 1997]. In this study the secondary data, which are available everywhere, were used to compute the local mean of the normal score variable Y, $m^*_Y(\mathbf{u})$, at any node of the interpolation grid. The ccdf means were then estimated by simple kriging with local means (SKlm) as

$$y^*_{SKlm}(\mathbf{u}) = \sum_{\alpha=1}^{n(\mathbf{u})} \lambda_\alpha \left[y(\mathbf{u}_\alpha) - m^*_Y(\mathbf{u}_\alpha)\right] + m^*_Y(\mathbf{u}) \quad (8)$$

The weights λ_α are the solution of the following system:

$$\sum_{\beta=1}^{n(\mathbf{u})} \lambda_\beta \left[C_R(\mathbf{u}_\alpha - \mathbf{u}_\beta) - \delta_{\alpha,\beta} b_{0R}\right] = C_R(\mathbf{u}_\alpha - \mathbf{u}) \;\; \alpha = 1, \ldots, n(\mathbf{u}) \quad (9)$$

where $C_R(\boldsymbol{u}_\alpha - \boldsymbol{u}_\beta)$ is the covariance function of the normal score residual variable $R(\boldsymbol{u}) = Y(\boldsymbol{u}) - m^*_Y(\boldsymbol{u})$ for the separation vector $\boldsymbol{h}_{\alpha\beta} = \boldsymbol{u}_\alpha - \boldsymbol{u}_\beta$, and b_{0R} is the corresponding nugget effect. The ccdf variance is computed as the sum of the SKlm variance (equation (7) using the residual covariance $C_R(\boldsymbol{h})$) and the variance of the local mean estimator.

[26] In this paper the local means $m^*_Y(\mathbf{u})$ were predicted by multivariate regression on the secondary variables Z_l and a quadratic function of the spatial coordinates, see section 4 for more details. The independent variable was the local mean of normal score data estimated by factorial kriging [*Matheron*, 1982; *Goovaerts et al.*, 1993] using the following estimator:

$$m^*_Y(\mathbf{u}_\alpha) = \sum_{\alpha=1}^{n(\mathbf{u})} \lambda_\alpha y(\mathbf{u}_\alpha) \quad (10)$$

The weights λ_α are the solution of an ordinary kriging system (equation (6)) with the right-hand side covariance terms, $C(\boldsymbol{u}_\alpha - \boldsymbol{u})$, set to zero [*Goovaerts*, 1997, p. 135].

3.2.4. Back Transform of Results

[27] The probability of nonexceeding any arsenic concentration z can be easily computed as

$$\text{Prob}\{Z(\mathbf{u}) \le z|(\text{Info})\} = F(\mathbf{u}; z|(\text{Info})) = F_Y(\mathbf{u}; \phi(z)|(\text{Info})) \quad (11)$$

where $\phi(z) = y$ is the normal score transform of the threshold of interest, and the function $F_Y(.)$ is defined in equation (4). The mean and variance of the probability distribution of the original variable Z are estimated using the following empirical expressions [*Saito and Goovaerts*, 2000]:

$$z^*_{MG}(\mathbf{u}) = \frac{1}{100}\sum_{j=1}^{100} z_p(\mathbf{u}) = \frac{1}{100}\sum_{j=1}^{100} \phi^{-1}\left(y_p(\mathbf{u})\right) \quad \text{with} \quad p = 0.01 \times (j - 0.5) \quad (12)$$

$$\sigma^{2*}_{MG}(\mathbf{u}) = \frac{1}{100}\sum_{j=1}^{100} \left[z_p(\mathbf{u}) - z^*_{MG}(\mathbf{u})\right]^2 \quad \text{with} \quad p = 0.01 \times (j - 0.5) \quad (13)$$

where $z_p(\boldsymbol{u})$ are p quantiles of the z ccdf obtained by a normal score back transform of the corresponding p quantiles of the y ccdf, $y_p(\boldsymbol{u})$.

[28] An implicit assumption of multi-Gaussian kriging is that the multipoint cdf of the random function $Z(\boldsymbol{u})$ is Gaussian. Unfortunately, the normality of the one-point cdf (histogram), which is achieved by the normal score transform, is a necessary but not sufficient condition to ensure the normality of the multipoint cdf [*Goovaerts*, 1997]. Although graphical procedures exist for checking the appropriateness of the normality assumption for the two-point cdf [*Deutsch and Journel*, 1998], there is no formal statistical test. Furthermore, to be complete one should also check the normality of the three-point, four-point,. . ., N-point cumulative distribution functions, which is unfeasible in practice. For all these reasons the MG approach is usually adopted with little regard to the underlying assumptions.

3.3. Indicator Approach

[29] Although the normal score transform makes the sample histogram perfectly symmetric it is not well suited to censored data since it requires a necessarily subjective ordering of all equally valued observations. For example, in this study 670 data below the same detection limit need to be artificially untied (first mode in the sample histogram of Figure 1b). Indicator kriging [*Journel*, 1983] is an alternative to the use of multi-Gaussian kriging to infer the ccdf $F(\boldsymbol{u};z|(\text{Info}))$ and the corresponding E-type estimate:

$$z^*_{IK}(\mathbf{u}) = \frac{1}{100}\sum_{j=1}^{100} z_p(\mathbf{u}) \quad \text{with} \quad p = 0.01 \times (j - 0.5) \quad (14)$$

Estimates (12) and (14) differ in the way the ccdf is modeled and so in how the series of quantiles $z_p(\boldsymbol{u})$ are computed. Instead of assuming that the ccdf is Gaussian and fully characterized by two parameters (parametric approach), the indicator approach estimates the conditional probabilities for a series of thresholds z_k discretizing the range of variation of z, and the complete function is obtained by interpolation/extrapolation of the estimated probabilities.

3.3.1. Indicator Transform

[30] The first step is to transform each observation $z(\boldsymbol{u}_\alpha)$ into a vector of K indicators defined as

$$i(\mathbf{u}_\alpha; z_k) = \begin{cases} 1 & \text{if } z(\mathbf{u}_\alpha) \le z_k \\ 0 & \text{otherwise} \end{cases} \qquad k = 1, 2, \ldots, K \quad (15)$$

In this paper K = 22 threshold values were selected, including the 19 0.05 quantiles of the sample histogram to cover uniformly the range of variation of arsenic concentration and three thresholds (47, 70, and 100 μg/L) in the upper tail of the distribution.

3.3.2. Indicator Kriging (IK)

[31] Ccdf values at unsampled location $\boldsymbol{u}$ are estimated as a linear combination of $n(\boldsymbol{u})$ surrounding indicator data:

$$F_{IK}(\mathbf{u}; z_k|(\text{Info})) = \sum_{\alpha=1}^{n(\mathbf{u})} \lambda_{\alpha k} i(\mathbf{u}_\alpha; z_k) \tag{16}$$

The kriging weights are computed by solving a kriging system similar to equation (6) where the normal score covariance terms are replaced by covariances of indicator variables $I(\boldsymbol{u};z_k)$, $C_I(\boldsymbol{h};z_k)$. Indicator covariance functions are derived from indicator semivariogram models, $\gamma_I(\boldsymbol{h};z_k)$, fitted to experimental values which were computed according to equation (1) where the z data are replaced by the corresponding indicator transforms.

[32] As for the multi-Gaussian approach secondary information was incorporated using simple kriging with local means (SKlm). The estimator is the following:

$$F_{SKlm}(\mathbf{u}; z_k|(\text{Info})) = j(\mathbf{u}; z_k) + \sum_{\alpha=1}^{n(\mathbf{u})} \lambda'_{\alpha k}[i(\mathbf{u}_\alpha; z_k) - j(\mathbf{u}_\alpha; z_k)] \tag{17}$$

The probabilities $j(\boldsymbol{u}_\alpha;z_k)$ are referred to as "soft" indicators since they are valued between zero and one, unlike the "hard" indicators defined in equation (15). They were computed from the soft information as:

$$j(\mathbf{u}; z_k) = \text{Prob}\{Z(\mathbf{u}) \le z_k|(\text{Info})\} = G[(\phi(z_k) - m_Y^*(\mathbf{u}))/\sigma^*(\mathbf{u})] \tag{18}$$

where $m_Y^*(\boldsymbol{u})$ and $\sigma^*(\boldsymbol{u})$ are the mean and standard error of the multivariate regression prediction at location $\boldsymbol{u}$. In other words, the distribution of the regression estimator is used as a prior probability distribution for arsenic concentration at that location, and the soft probabilities are retrieved directly for normal score transforms of the target thresholds z_k. The weights $\lambda'_{\alpha k}$ in expression (17) are computed by solving a simple kriging system (equation (9)), where the residual variable is now the difference between hard and soft indicator variables $R(\boldsymbol{u};z_k) = I(\boldsymbol{u};z_k) - j(\boldsymbol{u};z_k)$.

3.3.3. Modeling of the ccdf

[33] Because the K probabilities are estimated individually (i.e., K indicator kriging systems are solved at each location) the following constraints, which are implicit to any probability distribution, might not be satisfied by all sets of K estimates:

$$0 \le F_{IK}(\mathbf{u}; z_k|(\text{Info})) \le 1 \qquad \forall\ \text{k} \tag{19}$$

$$F_{IK}(\mathbf{u}; z_{k'}|(\text{Info})) \le F_{IK}(\mathbf{u}; z_k|(\text{Info})) \quad \text{if} \quad z_{k'} \le z_k \tag{20}$$

All probabilities that are not within [0,1] are first reset to the closest bound, 0 or 1. Then, condition (20) is ensured by averaging the results of an upward and downward correction of ccdf values [*Deutsch and Journel*, 1998].

[34] Once conditional probabilities were estimated and corrected for potential order relation deviations, the set of K probabilities must be interpolated within each class $(z_k, z_{k+1}]$ and extrapolated beyond the smallest and the largest thresholds to build a continuous model for the conditional cdf. In this paper, the resolution of the discrete ccdf was increased by performing a linear interpolation between tabulated bounds provided by the sample histogram of arsenic concentration [*Deutsch and Journel*, 1998].

3.4. Validation of the Prediction Models

[35] Relative performances of the multi-Gaussian versus indicator approaches, as well as the benefit of incorporating secondary information in the prediction, were assessed first using cross validation. One observation of the MDEQ data set is removed at a time and reestimated using neighboring noncolocated well data; i.e., repeated measurements in time were not used when estimating arsenic concentration. The second step was to compare the predictions with the recently collected data which were not used during the modeling itself (jack knife approach).

3.4.1. Prediction Errors

[36] The ability of the different techniques to estimate arsenic concentration was quantified using the mean absolute error of prediction (MAE) defined as

$$MAE = \frac{1}{n}\sum_{\alpha=1}^{n} |z(\mathbf{u}_\alpha) - z^*(\mathbf{u}_\alpha)| \tag{21}$$

where n is the number (8212 or 73) of individual wells. Colocated data were averaged to compute the reference arsenic concentration $z(\boldsymbol{u}_\alpha)$.

3.4.2. Model of Uncertainty

[37] At any location $\boldsymbol{u}$ knowledge of the ccdf $F(\boldsymbol{u};z|(\text{Info}))$ allows the computation of a series of symmetric p probability intervals (PI) bounded by the $(1 - p)/2$ and $(1 + p)/2$ quantiles of that ccdf. For example, the 0.5 PI is bounded by the lower and upper quartiles $[F^{-1}(\boldsymbol{u};0.25|(\text{Info})), F^{-1}(\boldsymbol{u};0.75|(\text{Info}))]$. A correct modeling of local uncertainty would entail that there is a 0.5 probability that the actual z value at $\boldsymbol{u}$ falls into that interval or, equivalently, that over the study area 50% of the 0.5 PI include the true value. Cross validation or jack knife yields a set of z measurements and independently derived ccdfs at the n locations $\boldsymbol{u}_\alpha$, allowing the fraction of true values falling into the symmetric p PI to be computed as:

$$\bar{\zeta}(p) = \frac{1}{n}\sum_{\alpha=1}^{n} \zeta(\mathbf{u}_\alpha; p) \tag{22}$$

where $\zeta(\mathbf{u}_\alpha;p)$ equals 1 if $z(\boldsymbol{u}_\alpha)$ lies between the $(1 - p)/2$ and $(1 + p)/2$ quantiles of the ccdf, and zero otherwise. The scattergram of the estimated, $\bar{\zeta}(p)$, versus expected, p, fractions is called the "accuracy plot". *Deutsch* [1997] proposed to assess the closeness of the estimated and theoretical fractions using the following "goodness" statistic:

$$G = 1 - \frac{1}{K}\sum_{k=1}^{K} w(p_k)|\bar{\zeta}(p_k) - p_k| \tag{23}$$

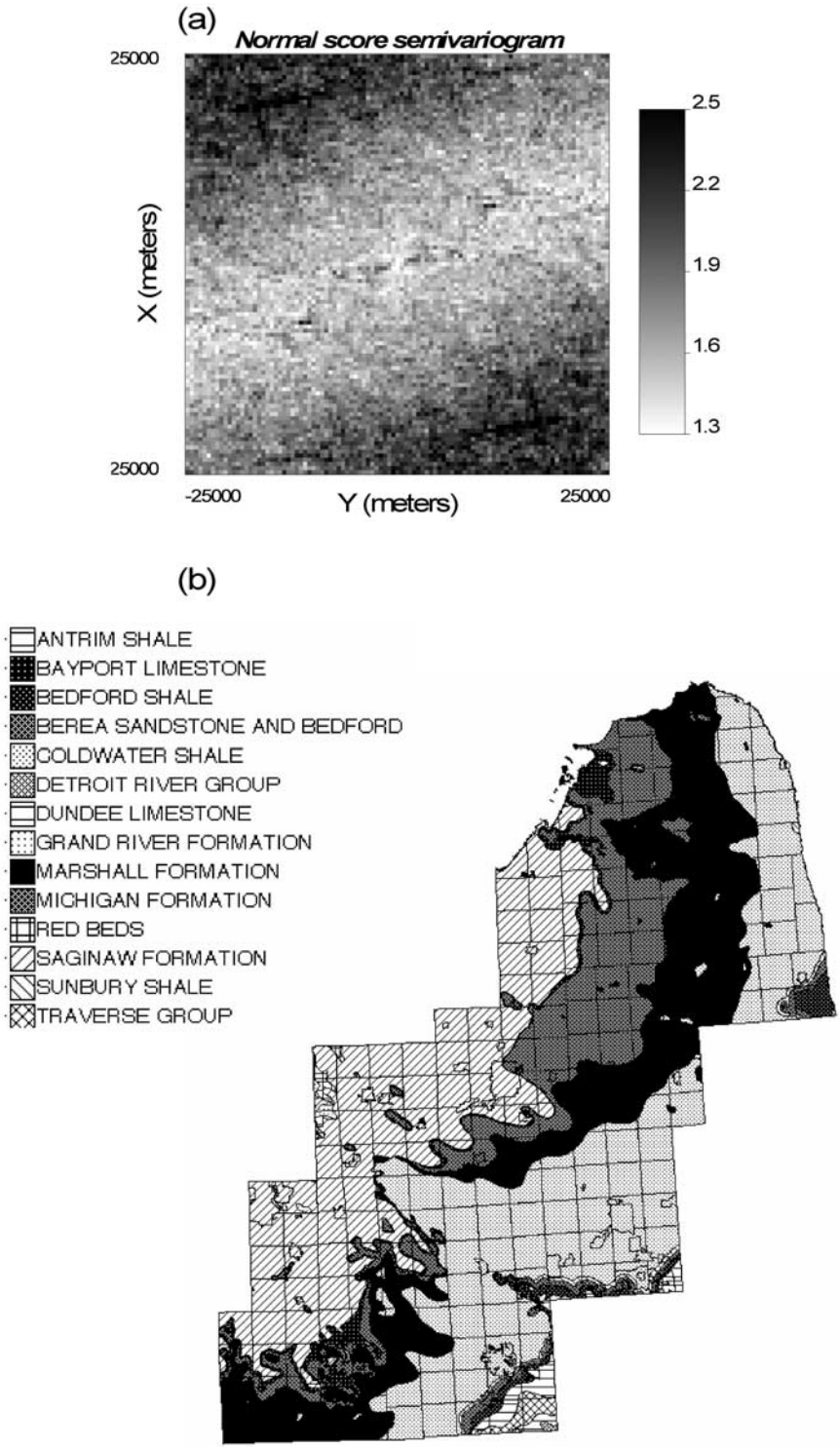

Figure 4. Spatial variability of normal score transforms. (a) The semivariogram map. (b) The map of bedrock with the location of the Marshall Sandstone subcrop where the highest concentrations of arsenic were found. Township boundaries are overlaid on the bedrock map. See color version of this figure at back of this issue.

where $w(p_k) = 1$ if $\bar{\zeta}(p_k) > p_k$, and 2 otherwise. Twice more importance is given to deviations when $\bar{\zeta}(p_k) < p_k$ (inaccurate case), i.e., the case where the fraction of true values falling into the p PI is smaller than expected.

[38] Not only should the true value fall into the PI according to the expected probability p, but this interval should be as narrow as possible to reduce the uncertainty about that value. In other words, among two probabilistic models with similar goodness statistics one would prefer the one with the smallest spread (less uncertain). Different measures of ccdf spread can be used: variance, interquartile range, and entropy. Following *Goovaerts* [2001], the

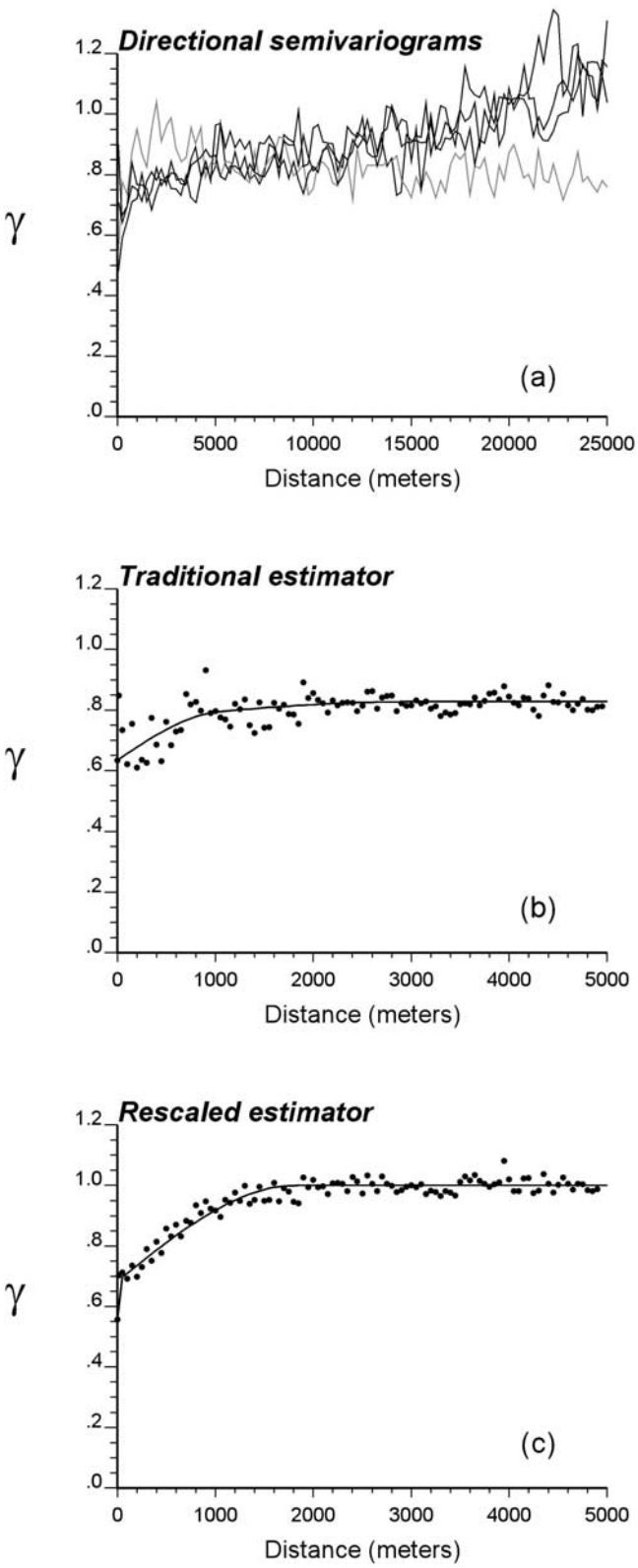

Figure 5. Spatial variability of normal score transforms. (a) Directional semivariograms (shaded line is NE-SW direction). (b) Omnidirectional semivariogram (traditional estimator). (c) Omnidirectional semivariogram (rescaled estimator).

average width of the PIs that include the true value are plotted for a series of probabilities p. For a probability p the average width is computed as

$$\bar{W}(p) = \frac{1}{n\bar{\zeta}(p)} \sum_{\alpha=1}^{n} \zeta(\mathbf{u}_\alpha; p) \left[F^{-1}(\mathbf{u}_\alpha; (1+p)/2 | (Info)) - F^{-1}(\mathbf{u}_\alpha; (1-p)/2 | (Info)) \right] \quad (24)$$

4. Results and Discussion

4.1. Mapping Arsenic Concentration Using Multi-Gaussian Kriging

[39] The study of the spatial variability of arsenic concentration began with the computation of the semivariogram map of normal score transforms which plots the experimental $\gamma(\boldsymbol{h})$ values in the system of coordinates (h_x,h_y). This map (Figure 4a) shows that as the distance between observations increases (i.e., as we go away from the center of the semivariogram map) the variability increases more slowly in the NE-SW direction (azimuth = 45° as measured in degrees clockwise from the N-S axis). This anisotropy reflects the impact of bedrock on the spatial distribution of arsenic concentrations, since NE-SW corresponds to the preferential orientation of bedrock layers (Figure 4b).

[40] Directional semivariograms (equation (1)) were then computed along four directions of azimuth: 0, 45, 90 and 135 (Figure 5a). These graphs indicate that the anisotropy occurs mainly for large distances (large-scale variability), and the variability is essentially isotropic within the radius of 5 km used for prediction later. Thus the subsequent modeling of the variability was limited to the omnidirectional semivariogram computed up to 5 km (Figures 5b and 5c). Two semivariogram estimators were used: the traditional one (equation (1)) and a rescaled semivariogram where each value is divided by the variance of the data used for that lag (equation (2)). The rescaling clearly attenuates the erratic fluctuations displayed by the traditional estimator which are caused by the preferential sampling of high-valued wells. On these graphs, the solid line represents the model fitted using least squares regression under the constraint of reproduction of the nugget effect inferred from colocated well measurements.

[41] Regardless of the estimator, the nugget effect represents more than 50% of the total variance and the well data are spatially independent for a separation distance larger than 2 km. This short-range variability of arsenic concentrations, which was reported by other authors [e.g., *BGS and DPHE*, 2001; *Serre et al.*, 2003; *Yu et al.*, 2003], suggests that spatial interpolation would benefit from secondary information to complement the arsenic data. Multivariate regression on the secondary variables described in section 2 resulted in a very small R^2 of 17.3%, which is caused by the large variability of arsenic concentration over very short distances and the magnitude of measurement errors (i.e., nugget effect). Once these short-range fluctuations were filtered using the factorial kriging procedure described in section 3.2.3 (equation (10)), the secondary information explained 65.2% of the variance. The regression model was then used to predict the local mean of arsenic concentrations across the study area (Figure 6a). Except in the southern part of the study area, the map of local means illustrates the

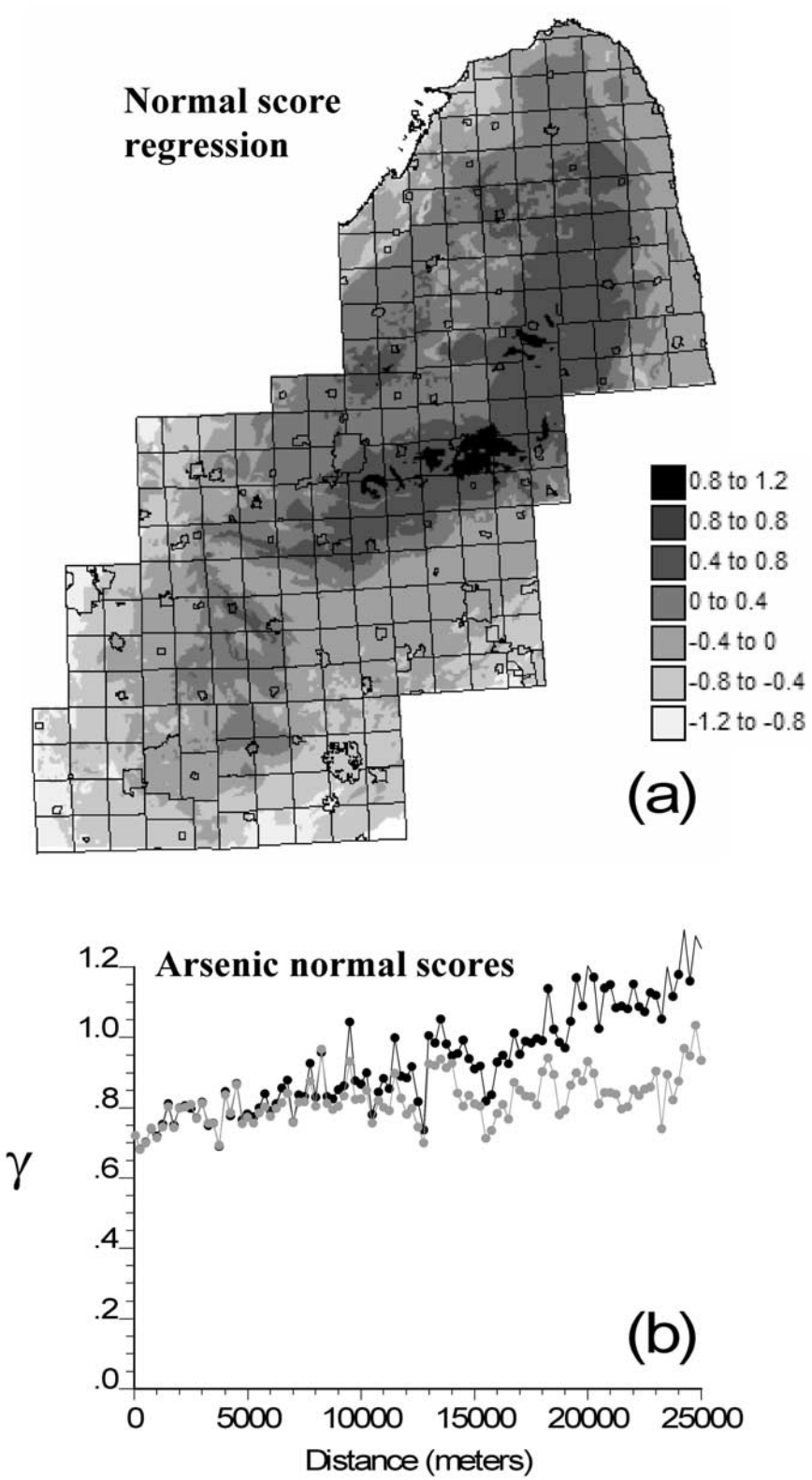

Figure 6. Incorporation of secondary information in the spatial prediction of arsenic concentration. (a) Map of normal score local means obtained using multiple linear regression and the bedrock map of Figure 4 as one of the explanatory variables. (b) Omnidirectional semivariogram of normal score transforms before (black dots) and after subtracting the local means (gray dots). See color version of this figure at back of this issue.

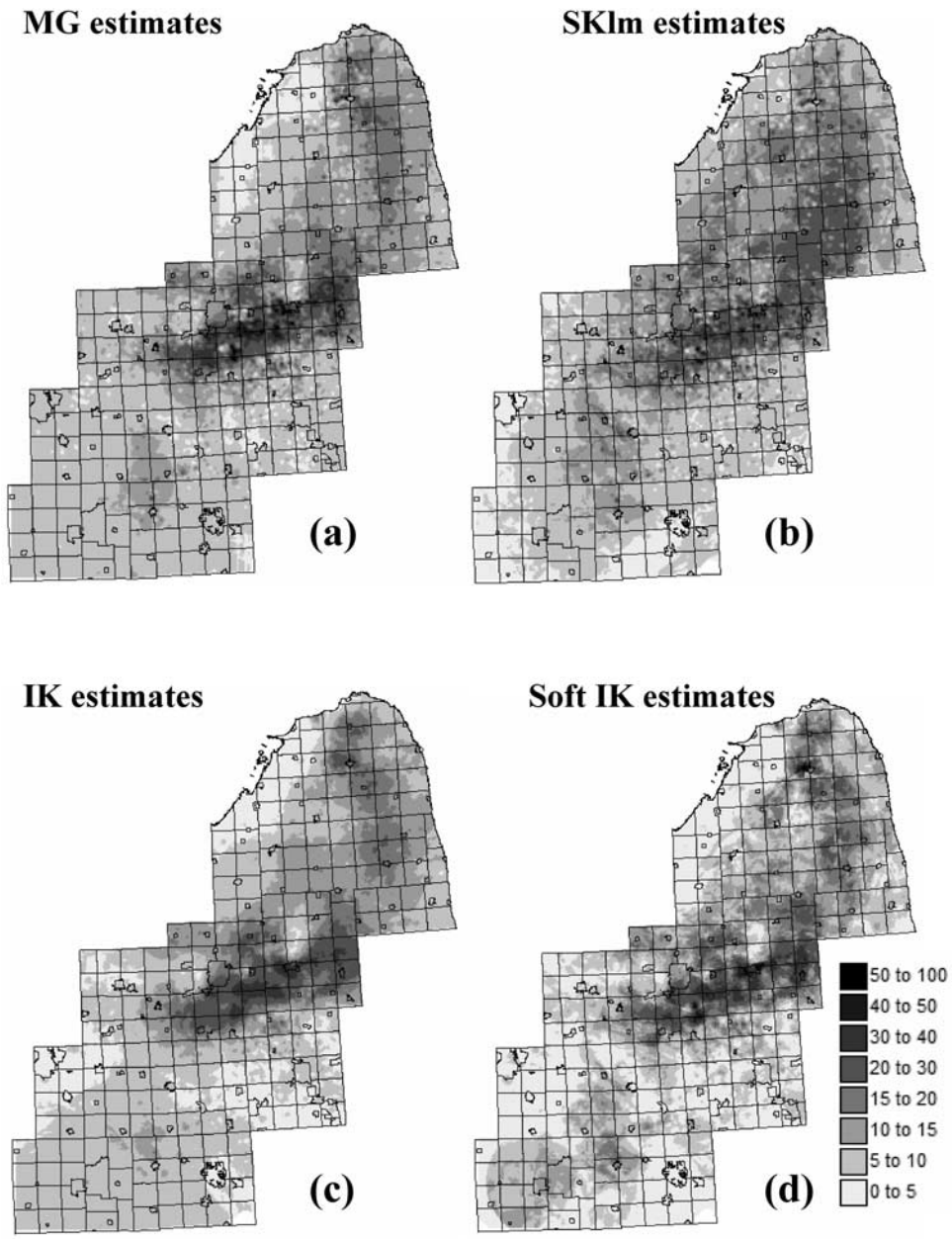

Figure 7. Alternative methods for spatial prediction of arsenic concentration. (a) Multi-Gaussian kriging. (b) Simple kriging of normal scores using the map in Figure 6a as local means. (c) Indicator kriging. (d) Soft indicator kriging using the same secondary information as for Figure 7b. Township boundaries are overlaid on each map. See color version of this figure at back of this issue.

impact of Marshall Sandstone on high arsenic concentrations. Interestingly, the semivariogram of residuals (gray dots in Figure 6b) is very close to the original normal score semivariogram (black dots) up to 15km, which indicates that secondary data explain mainly large-scale fluctuations. Similar results and scale were found by *Yu et al.* [2003] in Bangladesh.

[42] Figure 7 shows the maps of arsenic concentration obtained after back transform of estimates produced by multi-Gaussian kriging using only well data (Figure 7a) or incorporating the local means displayed in Figure 6 (SKlm estimate, Figure 7b). The estimates were computed at the nodes of a 500 meter spacing grid. Accounting for the secondary information allows one to capture better the bedrock regional variability, leading to fewer pixels with concentration exceeding 35 μg/L, but more pixels in the middle range 20–30 μg/L. Both maps display strong discontinuities in the vicinity of well locations; this salt-and-pepper effect is caused by the high nugget effect and short-range variability that lead to smooth estimates as one moves a few hundred meters away from sampled wells. Figures 8a and 8b show the corresponding maps of the probability of exceeding the USEPA standard of 10 μg/L. Probability maps show similar patterns as the

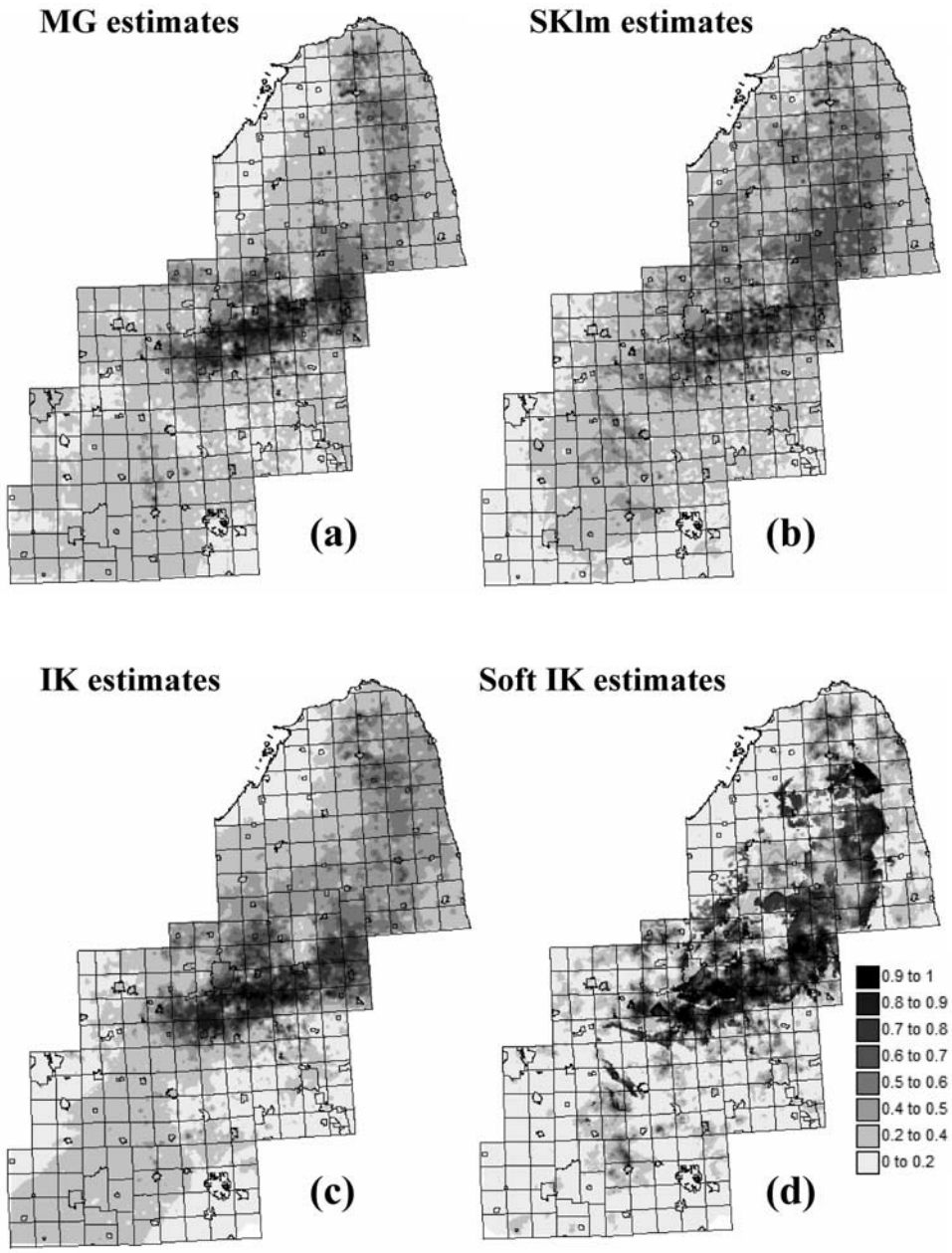

Figure 8. Alternative methods for spatial prediction of the probability of exceeding the USEPA standard of 10 μg/L. (a) Multi-Gaussian kriging. (b) Simple kriging of normal scores using the map in Figure 6a as local means. (c) Indicator kriging. (d) Soft indicator kriging using the same secondary information as Figure 8b. Township boundaries are overlaid on each map. See color version of this figure at back of this issue.

arsenic maps, with the secondary information (in particular the location of Marshall Sandstone) leading to higher probability of exceeding USEPA standard in the northern part of the study area.

4.2. Mapping Arsenic Concentration Using Indicator Kriging

[43] Indicator kriging was performed using $K = 22$ thresholds described in section 3.3.1. For each threshold, the omnidirectional indicator semivariogram was estimated up to 5 km, and then a model was fitted using least squares regression. Figure 9 shows the experimental and model semivariograms for five different thresholds. As the threshold increases, the short-range variability becomes more important, which indicates that small arsenic concentrations are better connected in space than large concentrations. This effect has been observed for other contaminants, such as soil Cd concentrations [*Goovaerts*, 1997].

[44] The 22 semivariogram models were used in ordinary kriging of indicators to estimate the set of local probabilities of nonexceedence at each grid node. After order relation correction and interpolation/extrapolation, the mean of each local ccdf was estimated using expression (14), and they are mapped in Figure 7c. Except for a less pronounced salt-and-pepper effect and fewer pixels with concentration exceeding 40 μg/L, this map is very

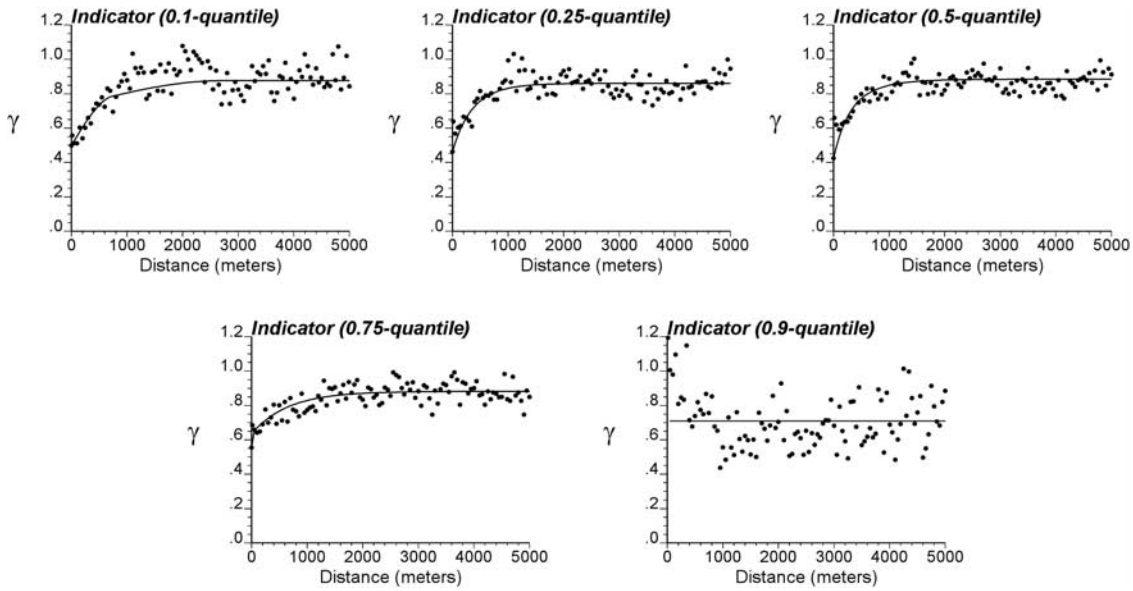

Figure 9. Sample omnidirectional indicator semivariograms with the model fitted using least squares regression. Note the larger short-range variability for the upper quartile and ninth decile of the sample histogram, which reflects the smaller spatial connectivity of high arsenic concentrations.

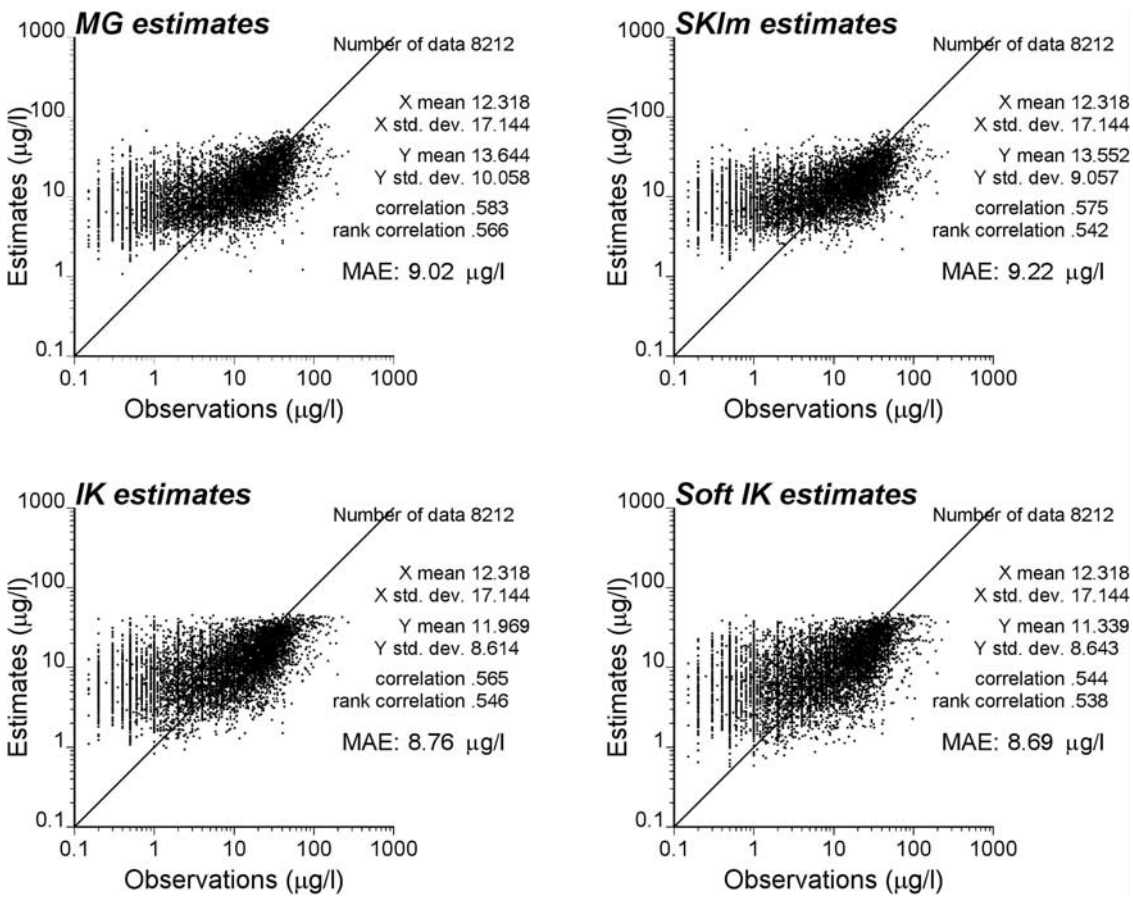

Figure 10. Scatterplots of estimated versus observed arsenic concentrations (MDEQ data set) for multi-Gaussian and indicator kriging approaches with and without secondary information. The mean absolute error of prediction (MAE) is also reported.

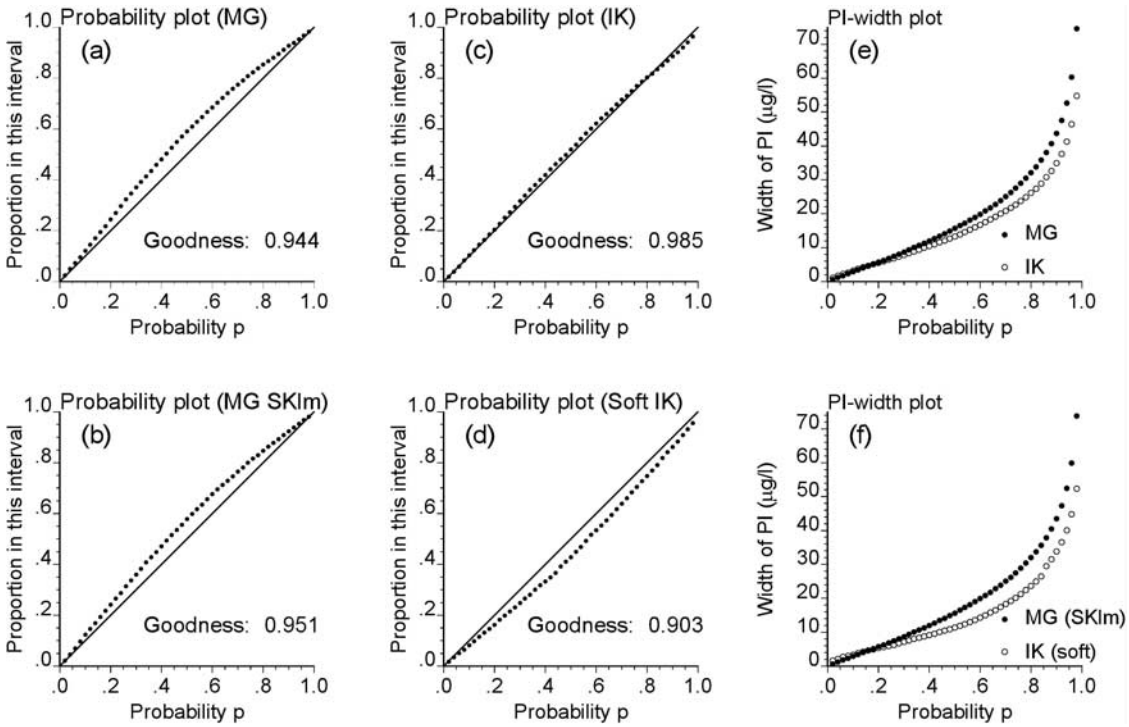

Figure 11. Plots of the proportion of observed arsenic data falling within probability intervals (accuracy plot) and the width of these intervals versus the probability p. The goodness statistics measure the similarity between the expected and observed proportions in the accuracy plots.

similar to the one obtained using multi-Gaussian kriging. In both cases, artifacts can be seen in the lower left part of the map and these discontinuities are caused by the sparse density of sampled wells in that region.

[45] The map of local means displayed in Figure 6a was combined with the regression standard error to derive prior probability distributions at each grid node. The prior information provided by secondary data was then updated with hard well data using the SKlm algorithm described in section 3.3.2. The resulting estimates of arsenic concentration are mapped in Figure 7d. This map strongly reproduces the spatial pattern of bedrock formation (Figure 4b), yielding smaller estimates in the Northwestern part of the study area. In general, the indicator kriging estimates are lower than the ones obtained using the multi-Gaussian approach (mean = 9.17 μg/L for soft IK versus 11.53 μg/L for SKlm). Although the SKlm mean is closer to the declustered sample mean of 10.97 μg/L, cross-validation analysis below indicates that at sampled wells the indicator approach leads to smaller bias. Also, the artifact noticed in the lower left part of the MG map is now clearly apparent, and is caused by the presence of a few high concentrations in this sparsely sampled region which are then spread through the circular search window.

[46] As with the multi-Gaussian approach, the maps of the probability of exceeding the USEPA standard show patterns similar to the arsenic maps (Figures 8c and 8d). The impact of the secondary information is however much more pronounced, with clear discontinuities in the soft IK probability map which coincide with the boundaries of the bedrock map. One of the thresholds for the indicator coding is 10 μg/L, hence this probability is directly estimated and does not need to be retrieved from the local probability distribution as for the MG approach.

4.3. Performance Comparison: The MDEQ Data Set

[47] The visual comparison of maps in Figure 7 does not indicate which technique produces the most accurate estimates. Cross validation of the MDEQ data was used to compare the prediction performances of the four interpolation algorithms. Figure 10 shows the scatterplots of observed concentrations versus estimates at each of the 8212 individual wells. They all reveal an underestimation of large concentrations and an overestimation of low concentrations, which is common for least squares estimators such as the mean of local probability distributions (overestimation of low concentrations is also visually enhanced by the use of a log-log scale). The balancing of these two effects results however in a somewhat global unbiasedness. The smallest bias (−0.4 μg/L) is observed for indicator kriging, while both types of multi-Gaussian kriging lead to an average overestimation of 1.2 μg/L. The same two algorithms also yield the largest mean absolute errors of prediction (MAE), however differences between all four algorithms are quite small. Although the magnitude of mean absolute errors is large with respect to the USEPA standard of 10 μg/L, it is worth remembering that the average difference between data collected at the same well is 12.53 μg/L in the MDEQ arsenic data set. Thus the uncertainty attached to the sampled values themselves contributes to the poor accuracy of the geostatistical predictions.

[48] For comparison purposes, cross-validation was also performed for lognormal kriging and the inverse square distance method. For lognormal kriging, the MAE of

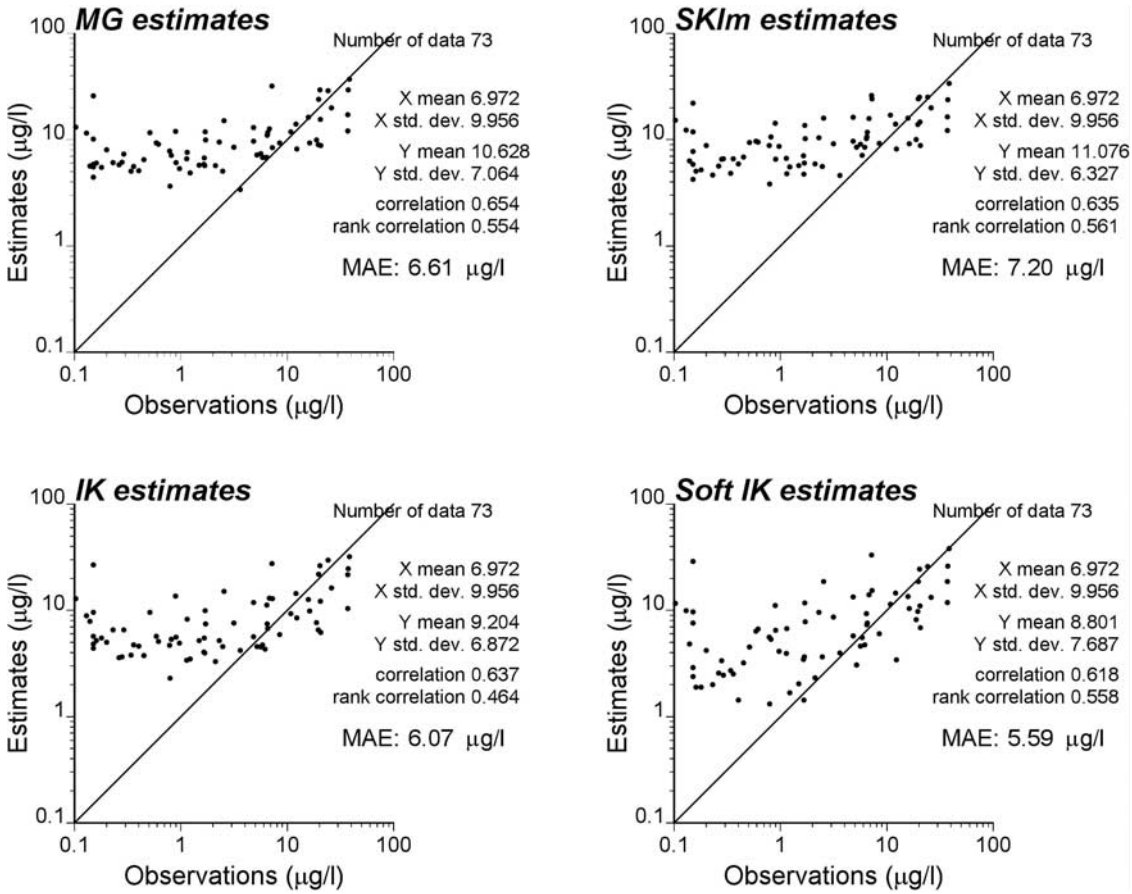

Figure 12. Validation of the prediction models using recently collected well data. Scatterplots of estimated versus observed arsenic concentrations (2004 campaign) for multi-Gaussian and indicator kriging approaches with and without secondary information. The mean absolute error of prediction (MAE) is also reported.

11.49 μg/L is substantially larger than the results obtained in this study, which emphasizes the risk of using this type of kriging because of the strong influence of semivariogram modeling on the lognormal back transform of estimated values. The inverse square distance method yields a MAE of 9.66 μg/L, which is slightly larger than the errors obtained by any of the four geostatistical methods presented in Figure 10. The benefit of kriging might not be as large as intuitively expected, but this result is in agreement with previous studies [e.g., *Goovaerts*, 2000] that showed that the gain of using ordinary kriging versus the inverse square distance method decreases as the correlation between observations weakens.

[49] The quality of the models of uncertainty for each technique was assessed using the accuracy plots, described in section 3.4.2 and displayed in Figures 11a–11d. For the two MG models of uncertainty, the probability intervals (PI) contain a higher than expected proportion of true values, a property referred as to accuracy by *Deutsch* [1997]. Best results are obtained again for univariate indicator kriging where expected and empirical proportions are very close, yielding a goodness statistic close to unity. Not only should the true value fall into the PI according to the expected probability p, but this interval should be as narrow as possible to reduce the uncertainty about that value. Figures 11e and 11f indicate that the best model of uncertainty is obtained using indicator kriging in that the probability intervals are narrower (larger precision) while including the expected proportions of true values (large goodness statistic).

4.4. Performance Comparison: The 2004 Campaign

[50] Because cross-validation in section 4.3 relies on the MDEQ data both for building the prediction model and assessing its quality, it may tend to provide optimistic assessments of prediction performances. Interpolation errors might also be underestimated since a single observation is removed at a time, leaving a high sampling density for prediction. Well data collected at the homes of 73 participants in the cancer case control study are here used to validate the prediction models obtained using the multi-Gaussian and indicator approaches. This data set was not utilized in the preliminary analysis and so qualifies as an independent validation set for quantifying the accuracy of the prediction at unmonitored locations and times. To mimic the future use of the layer of arsenic concentration estimates in the space-time information system, the concentration at validation wells was estimated using the value of the 500 × 500 meter pixel within which it falls, thereby eliminating the need to solve a kriging

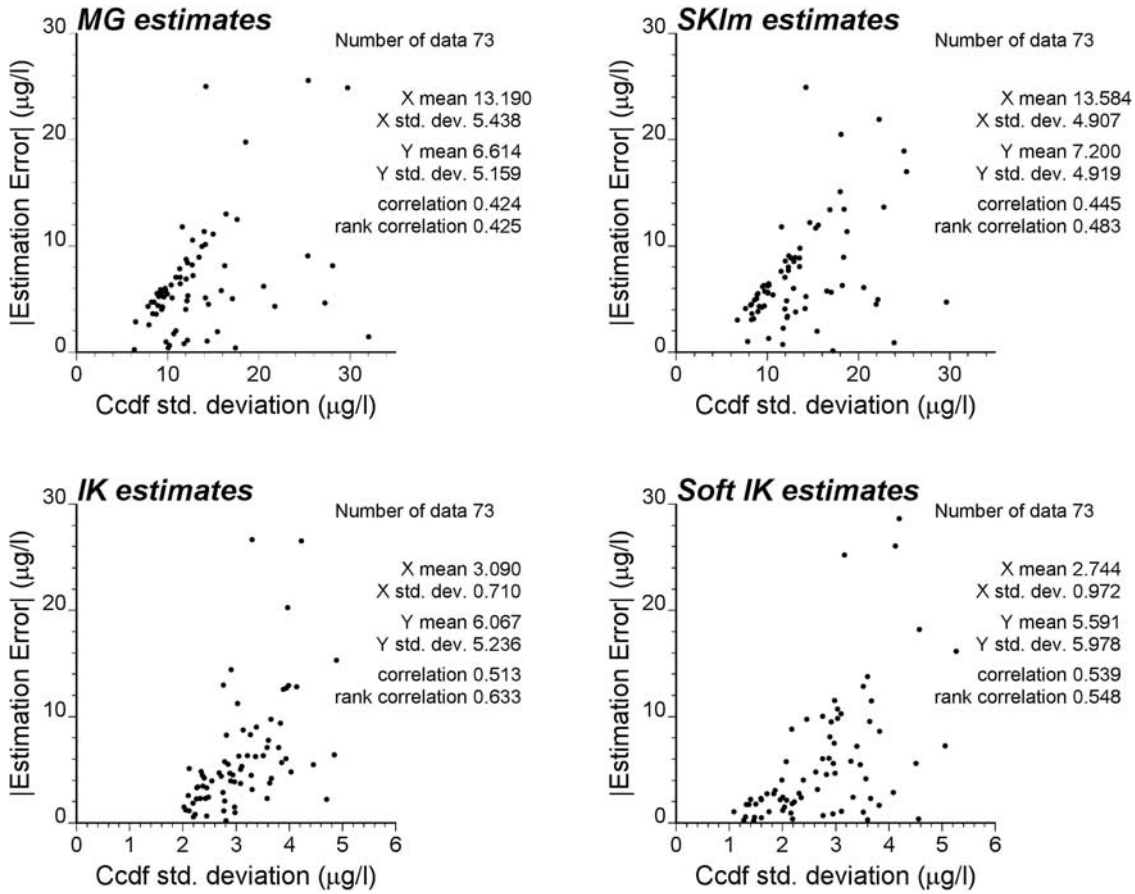

Figure 13. Validation of the models of local uncertainty using recently collected well data. Scatterplots of absolute prediction errors versus standard deviation of the local probability distributions (ccdf) modeled using multi-Gaussian and indicator kriging approaches with and without secondary information.

system every time a study subject moves. Sensitivity analysis indicated that such a simplification leads only to a marginal increase in the prediction error (e.g., MAE = 6.61 μg/L versus 6.58 μg/L using the exact spatial coordinates in multi-Gaussian kriging).

[51] Figure 12 shows the scatterplots of observed concentrations versus estimates at each of the 73 validation wells. As for the cross validation results, the low concentrations are overestimated by the four algorithms. However, this conditional bias is clearly reduced for the concentration range 1 to 5 μg/L when using soft indicator kriging. It is noteworthy that accounting for secondary information actually leads to larger prediction errors for the multi-Gaussian approach. In general, indicator kriging outperforms multi-Gaussian kriging, as exemplified by the smaller mean absolute error of prediction obtained for both hard and soft indicator kriging. The magnitude of these errors is also smaller than the ones found in cross validation, which is likely due to the smaller concentrations measured at validation wells versus the MDEQ data set (no preferential sampling of high-valued areas).

[52] The uncertainty attached to the concentration estimate at any particular location can be assessed using the spread of the local distribution of probability at that location. The standard deviation of the ccdf is here used as a measure of uncertainty and plotted against the magnitude of the actual prediction error in Figure 13. The standard deviation is clearly much smaller for the IK-based ccdfs and exhibits a stronger correlation with the actual prediction error, which confirms the ability of the nonparametric approach to account for data values in uncertainty modeling [*Goovaerts*, 2001].

5. Conclusions

[53] This paper described several approaches for spatial interpolation of arsenic concentration in southeast Michigan groundwater. It is the first step toward the assessment of the risk associated with exposure to low levels of arsenic in drinking water (typically, 5–100 μg/L), in particular for the development of bladder cancer. This study confirmed results in the literature that reported intense spatial nonhomogeneity of As concentration, resulting in samples that vary greatly even when located only a few meters apart [*Serre et al.*, 2003]. However, the short-range variability in this data set was likely inflated by the combination of water samples of different origins, which could explain the magnitude of fluctuations observed between observations collected the same day at the same wells. Indicator semivariograms showed a better spatial connectivity of low concentrations while values exceeding 32 μg/L (10% of wells) are spatially uncorrelated. Secondary information, such as proximity to Marshall Sandstone, helped only the prediction at a regional scale (i.e., beyond 15 km), leaving the short-range variability largely unexplained.

[54] Several geostatistical tools were tailored to the features of the MDEQ data set: (1) semivariogram values were standardized by the lag variance to correct for the preferential sampling of wells with high arsenic concentrations, (2) semivariogram modeling was conducted under the constraint of reproduction of the nugget effect inferred from colocated well measurements, (3) kriging systems were modified to account for repeated measurements at a series of wells while avoiding noninvertible kriging matrices, (4) kriging-based smoothing was combined with multivariate regression to predict the regional background of arsenic concentration across the study area. Cross validation indicated the little benefit of secondary information in local prediction of arsenic concentration. Slightly better results were obtained using soft indicator kriging that generated the smallest mean absolute error of prediction, while the most precise and accurate models of uncertainty are produced by univariate indicator kriging.

[55] All predictions in this study were conducted at the nodes of a 500 meter spaced grid using punctual kriging algorithms. Analysis of the 2004 validation set indicated that reasonable estimates are obtained by assigning these punctual estimates to all the wells located within a 500×500 meter square centered on this node. The prediction support could, however, be expanded to provide in theory more meaningful values for exposure assessment. Future research will implement geostatistical simulation to perform an upscaling of the model of uncertainty, yielding empirical probability distribution of the arsenic concentration over 2.5 km^2 blocks. This probabilistic model will then be combined with the spatiotemporal mobility and water consumption of study participants, leading to the estimation of individual-level historical exposure to arsenic and the attached uncertainty [*Meliker et al.*, 2005].

[56] **Acknowledgments.** This study was supported by grant R01 CA96002-10, Geographic-Based Research in Cancer Control and Epidemiology, from the National Cancer Institute. Development of the STIS software was funded by grants R43 ES10220 from the National Institutes of Environmental Health Sciences and R01 CA92669 from the National Cancer Institute.

References

Aichele, S. S., and A. M. Shortridge (2002), Assessing spatial impact of proposed drinking-water arsenic standards in southeast Michigan using geostatistics, paper presented at the 98th Annual Association of American Geographers Meeting, Los Angeles, Calif., 19–23 March.

AvRuskin, G. A., G. M. Jacquez, J. R. Meliker, M. J. Slotnick, A. M. Kaufmann, and J. O. Nriagu (2004), Visualization and exploratory analysis of epidemiologic data using a novel space time information system, *Int. J. Health Geogr.*, *3*, 26, doi:10.1186/1476-072X-3-26.

British Geological Survey and of Department of Public Health Engineering (BGS and DPHE) (2001), Arsenic contamination of groundwater in Bangladesh, edited by D. G. Kinniburgh and P. L. Smedley, *Rep. WC/00/19*, vol. 1–4, Br. Geol. Surv., Keyworth, U. K. (Available at http://www.bgs.ac.uk/arsenic/Bangladesh)

Burrough, P. A., and R. A. McDonnell (2000), *Principles of Geographical Information Systems*, Oxford Univ. Press, New York.

Deutsch, C. V. (1997), Direct assessment of local accuracy and precision, in *Geostatistics Wollongong '96*, edited by E. Y. Baafi and N. A. Schofield, pp. 115–125, Springer, New York.

Deutsch, C. V., and A. G. Journel (1998), *GSLIB: Geostatistical Software Library and User's Guide*, 2nd ed., Oxford Univ. Press, New York.

Frisbie, S. H., R. Ortega, D. M. Maynard, and B. Sarkar (2002), The concentrations of arsenic and other toxic elements in Bangladesh's drinking water, *Environ. Health Perspect.*, *110*(11), 1147–1153.

Goodchild, M. F. (1996), The application of advanced information technology in assessing environmental impacts, in *Application of GIS to the Modeling of Non-point Source Pollutants in the Vadose Zone*, edited by D. L. Corwin and K. Loague, *SSSA Spec. Publ.*, *48*, 1–17.

Goovaerts, P. (1997), *Geostatistics for Natural Resources Evaluation*, Oxford Univ. Press, New York.

Goovaerts, P. (2000), Geostatistical approaches for incorporating elevation into the spatial interpolation of rainfall, *J. Hydrol.*, *228*, 113–129.

Goovaerts, P. (2001), Geostatistical modelling of uncertainty in soil science, *Geoderma*, *103*, 3–26.

Goovaerts, P., P. Sonnet, and A. Navarre (1993), Factorial kriging analysis of springwater contents in the Dyle river basin, Belgium, *Water Resour. Res.*, *29*(7), 2115–2125.

Heuvelink, G. B. M. (1998), *Error Propagation in Environmental Modeling With GIS*, Taylor and Francis, Philadelphia, Pa.

Jacquez, G. M., G. AvRuskin, E. Do, H. Durbeck, D. A. Greiling, P. Goovaerts, A. Kaufmann, and B. Rommel (2004), Complex systems analysis using space-time information systems and model transition sensitivity analysis, in *Accuracy 2004: Proceedings of the 6th International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences* [CD-ROM], edited by H. T. Mowrer, R. McRoberts, and P. C. VanDeusen, Int. Environmetrics Soc., St. Lucia, Queensl., Australia.

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

Karthik, B., S. Islam, and C. F. Harvey (2001), On the spatial variability of arsenic contamination in the groundwater of Bangladesh, *Eos. Trans. AGU*, *82*(20), Spring Meet. Suppl., Abstract H61C-01.

Kim, M. J. (1999), Arsenic dissolution and speciation in groundwater of southeast Michigan, Ph.D. thesis, Univ. of Mich., Ann Arbor.

Kim, M. J., J. Nriagu, and S. Haack (2002), Arsenic species and chemistry in groundwater of southeast Michigan, *Environ. Pollut.*, *120*, 379–390.

Kolker, A., W. F. Cannon, D. B. Westjohn, and L. G. Woodruff (1998), Arsenic-rich pyrite in the Mississippian Marchall Sandstone: Source of anomalous arsenic in southeastern Michigan ground water, paper presented at the 1998 National Meeting of the Geological Society of America, Toronto, Ont., Canada, 25–29 Oct.

Kolker, A., S. K. Haack, W. F. Cannon, D. B. Westjohn, M. J. Kim, J. Nriagu, and L. G. Woodruff (2003), Arsenic in southeastern Michigan, in *Arsenic in Groundwater: Geochemistry and Occurrence*, edited by A. H. Welch and K. G. Stollenwerk, pp. 281–294, Springer, New York.

Kyriakidis, P. C., and A. G. Journel (1999), Geostatistical space-time models: A review, *Math. Geol.*, *31*(6), 651–684.

Matheron, G. (1982), Pour une Analyse Krigeante de Données Régionalisées, *Internal Note N-732*, Cent. de Géostat., Fontainebleau, France.

Meliker, J. R., M. J. Slotnick, G. A. AvRuskin, A. M. Kaufmann, G. M. Jacquez, and J. O. Nriagu (2005), Improving exposure assessment in environmental epidemiology: Application of spatio-temporal visualization tools, *J. Geogr. Syst.*, *7*, 49–66.

Pannatier, I. (1996), *VARIOWIN: Software for Spatial Data Analysis in 2D*, Springer, New York.

Pardo-Iguzquiza, E. (1999), VARFIT: A Fortran-77 program for fitting variogram models by weighted least squares, *Comput. Geosci.*, *25*, 251–261.

Ryker, S. J. (2001), Mapping arsenic in ground water—A real need, but a hard problem, *Geotimes*, *46*(11), 34–36.

Saito, H., and P. Goovaerts (2000), Geostatistical interpolation of positively skewed and censored data in a dioxin contaminated site, *Environ. Sci. Technol.*, *34*(19), 4228–4235.

Serre, M. L., A. Kolovos, G. Christakos, and K. Modis (2003), An application of the holistochastic human exposure methodology to naturally occurring arsenic in Bangladesh drinking water, *Risk. Anal.*, *23*(3), 515–528.

Slotnick, M. J., J. Meliker, and J. Nriagu (2003), Natural sources of arsenic in southeastern Michigan groundwater, *J. Phys. IV*, *107*, 1247–1250.

Steinmaus, C., Y. Yuan, M. N. Bates, and A. H. Smith (2003), Case-control study of bladder cancer and drinking water arsenic in the western United States, *Am. J. Epidemiol.*, *158*, 1193–1201.

Warner, K. L., A. Martin Jr., and T. L. Arnold (2003), Arsenic in Illinois ground water—Community and private supplies, *U.S. Geol. Surv. Water Resour. Invest. Rep.*, *03-4103*. (Available at http://il.water.usgs.gov/pubs/wrir03_4103.pdf)

Welch, A. H., D. B. Westjohn, D. R. Helsel, and R. B. Wanty (2000), Arsenic in ground water of the United States: Occurrence and geochemistry, *Ground Water*, *38*, 589–604.

Welhan, J., and M. Merrick (2003), Statewide network data analysis and kriging project, final report, Idaho Geol. Surv. (Available at http://www.idwr.idaho.gov/hydrologic/info/pubs/gwq/IGS_Kriging_Project-Final_Report.pdf)

Westjohn, D. B., A. Kolker, W. F. Cannon, and D. F. Sibley (1998), Arsenic in groundwater in the "Thumb Area" of Michigan: The Mississipian Marshall sandstone revisited, paper presented at Michigan: Its Geology and Geological Resources, 5th Symposium, Mich. State Univ., East Lansing, 9–10 April.

Yu, W. H., C. M. Harvey, and C. F. Harvey (2003), Arsenic in groundwater in Bangladesh: A geostatistical and epidemiological framework for evaluating health effects and potential remedies, *Water Resour. Res.*, *39*(6), 1146, doi:10.1029/2002WR001327.

G. AvRuskin, P. Goovaerts, and G. Jacquez, BioMedware, Inc., 516 North State Street, Ann Arbor, MI 48104, USA. (goovaerts@biomedware.com)

J. Meliker, J. Nriagu, and M. Slotnick, School of Public Health, University of Michigan, Ann Arbor, MI 48109-2029, USA.

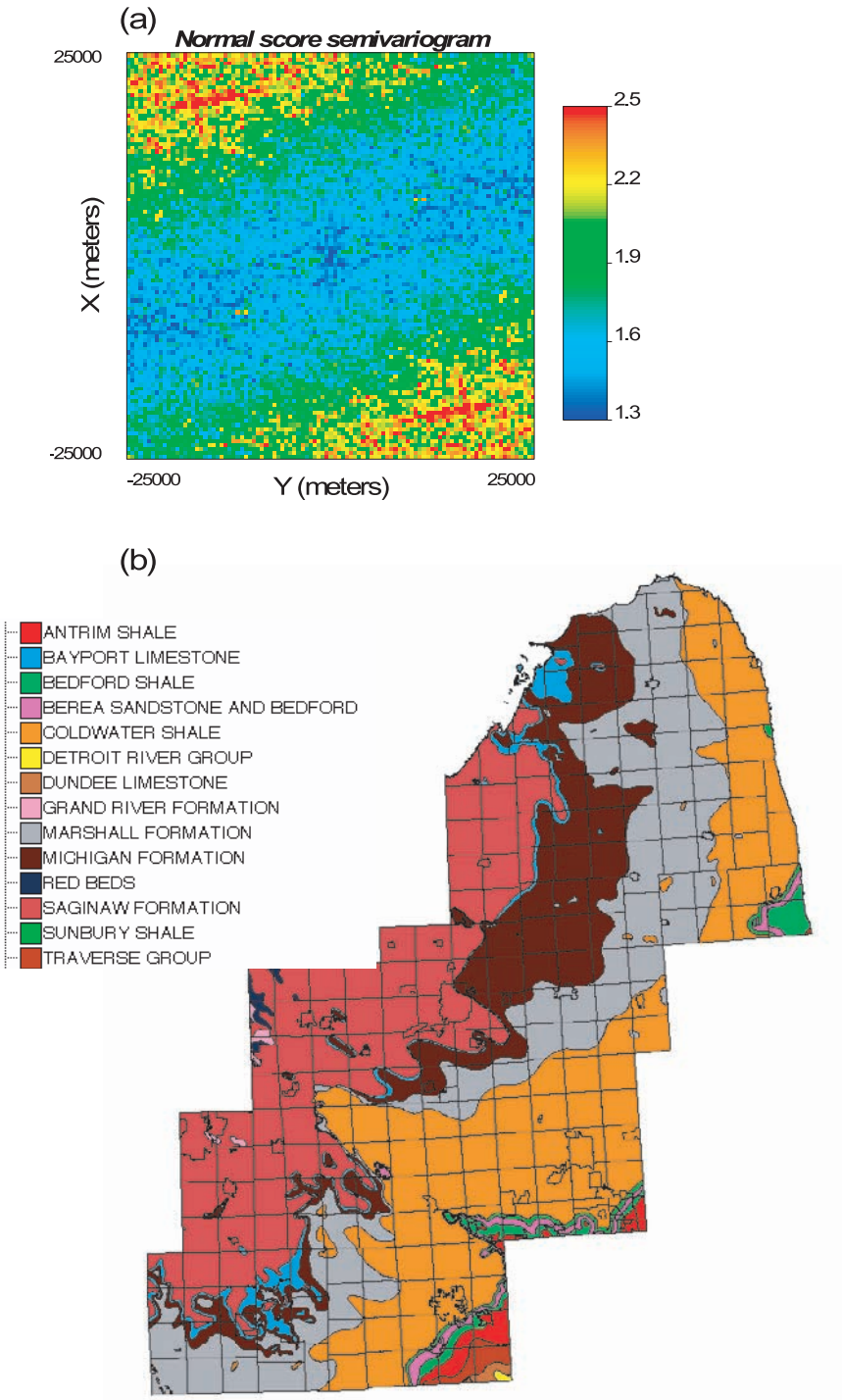

Figure 4. Spatial variability of normal score transforms. (a) The semivariogram map. (b) The map of bedrock with the location of the Marshall Sandstone subcrop where the highest concentrations of arsenic were found. Township boundaries are overlaid on the bedrock map.

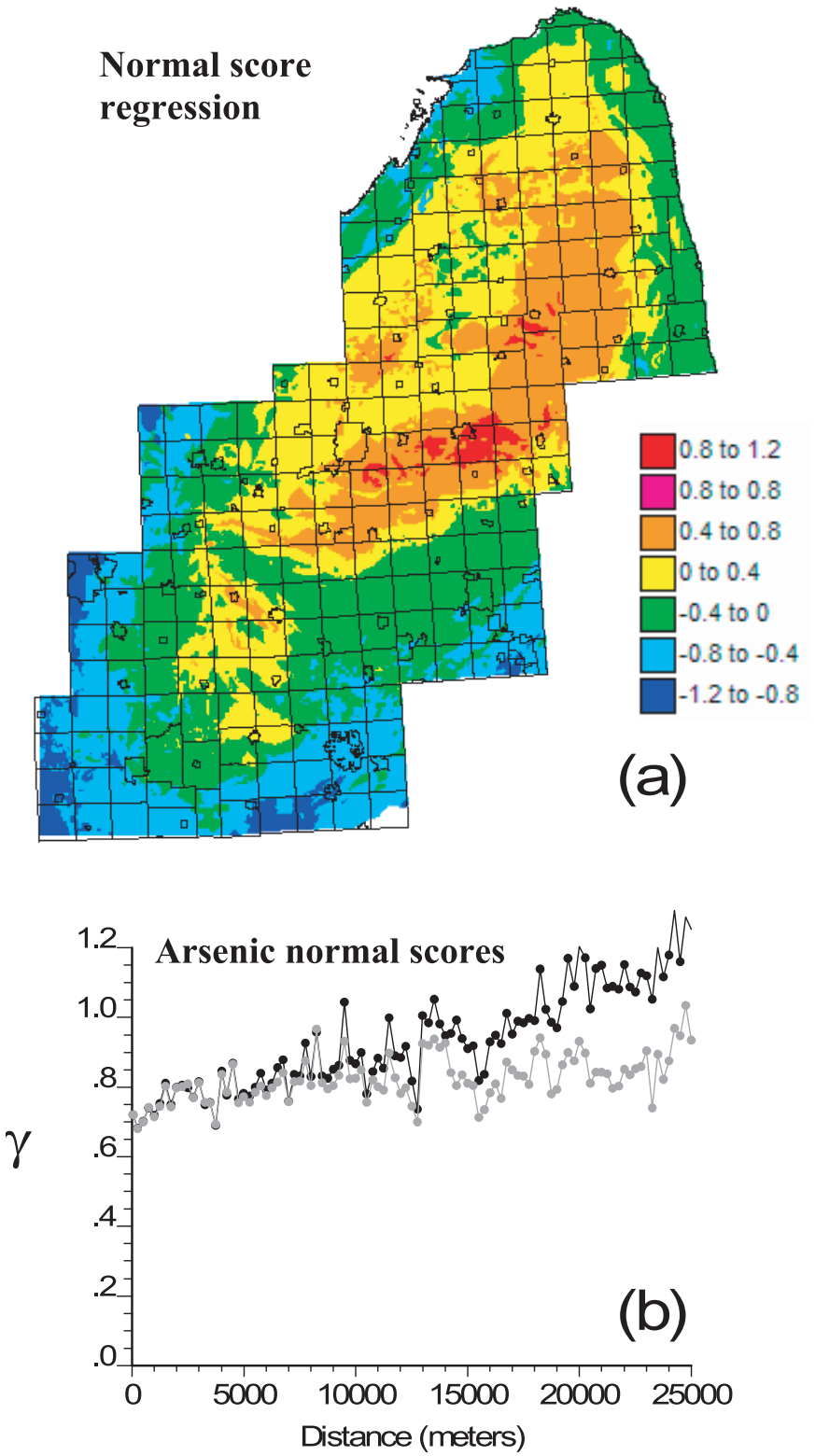

Figure 6. Incorporation of secondary information in the spatial prediction of arsenic concentration. (a) Map of normal score local means obtained using multiple linear regression and the bedrock map of Figure 4 as one of the explanatory variables. (b) Omnidirectional semivariogram of normal score transforms before (black dots) and after subtracting the local means (gray dots).

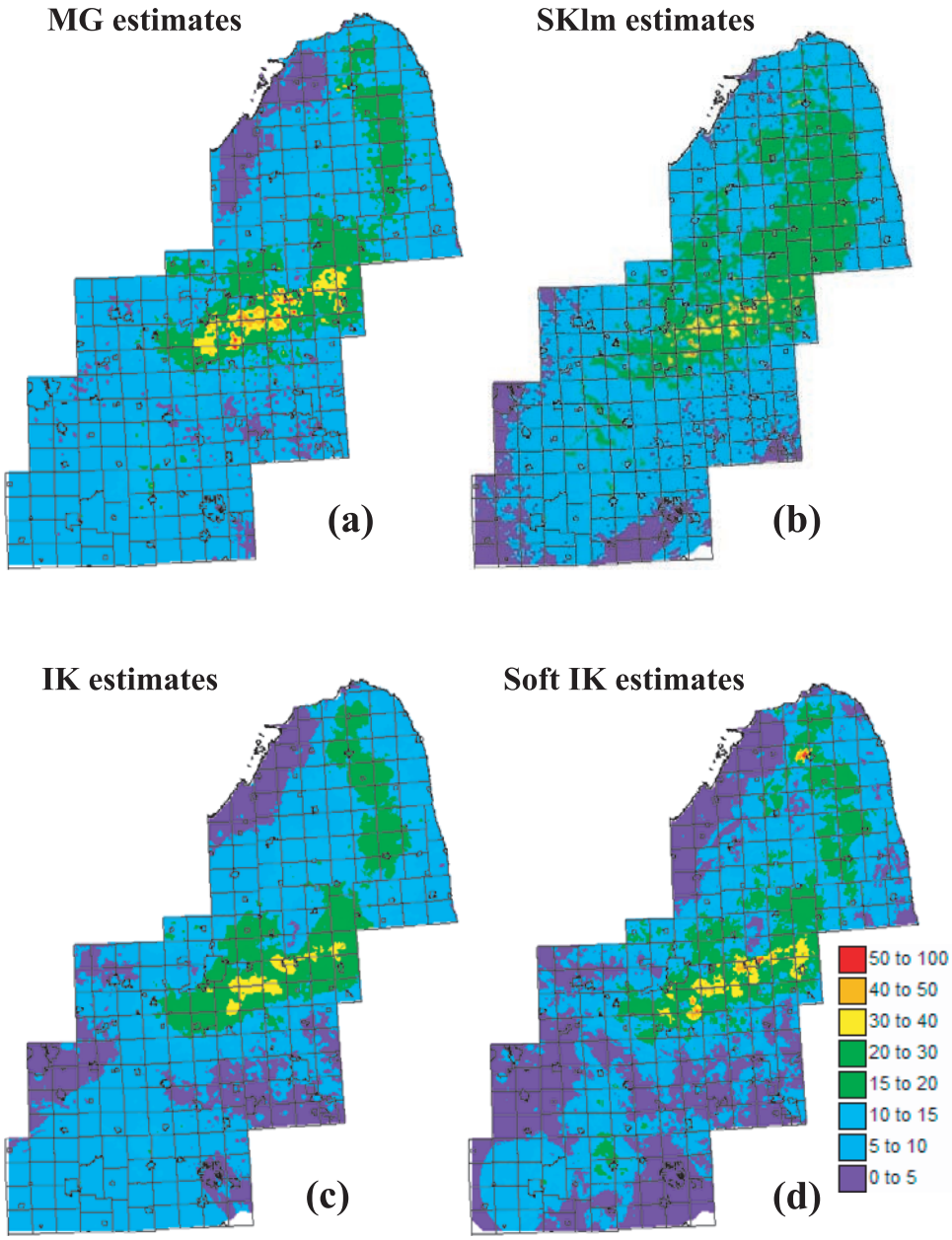

Figure 7. Alternative methods for spatial prediction of arsenic concentration. (a) Multi-Gaussian kriging. (b) Simple kriging of normal scores using the map in Figure 6a as local means. (c) Indicator kriging. (d) Soft indicator kriging using the same secondary information as for Figure 7b. Township boundaries are overlaid on each map.

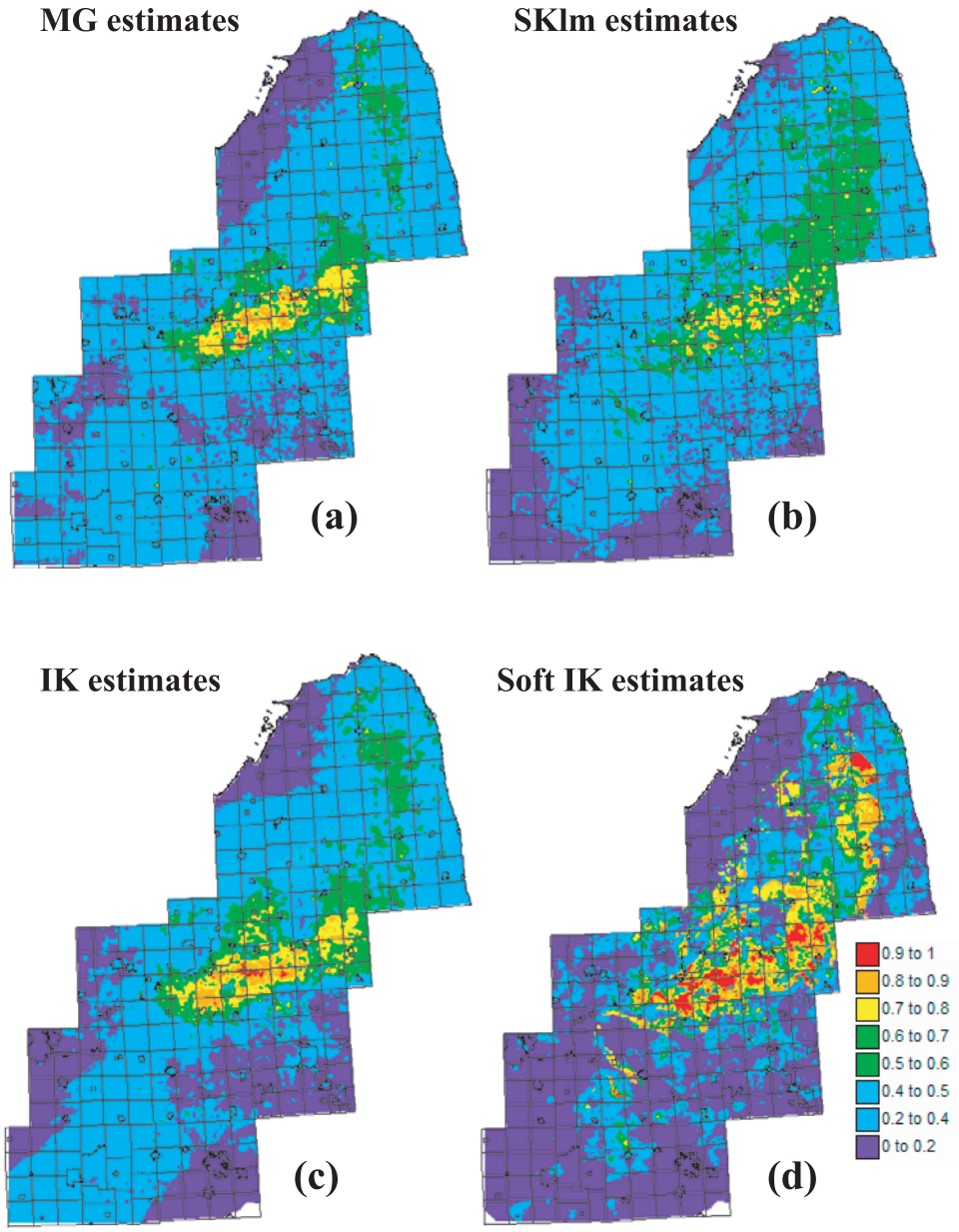

Figure 8. Alternative methods for spatial prediction of the probability of exceeding the USEPA standard of 10 μg/L. (a) Multi-Gaussian kriging. (b) Simple kriging of normal scores using the map in Figure 6a as local means. (c) Indicator kriging. (d) Soft indicator kriging using the same secondary information as Figure 8b. Township boundaries are overlaid on each map.