9 Scales of variation

9.1 Introduction

One of the recurring comments heard in Bangladesh when discussing the groundwater arsenic problem is of the great variability in groundwater arsenic concentrations found. It is often said that the arsenic concentrations in adjacent wells bear very little resemblance to one another. This makes mapping the spatial variation in arsenic concentrations within a village very difficult. Indeed it is usually inferred from such observations that each and every well needs to be analysed. This may indeed be the case for compliance testing but the problem with such statements is that they tend to promote the attitude that there is little or no spatial dependence in the data and that we should just accept that we are dealing with a hopelessly spatially variable environment. In reality, this is far from the case, especially on the regional scale, and we can use the spatial patterns to aid the setting of priority areas for emergency testing, and perhaps in the future, to help guide a national water resources strategy.

In this Chapter, we explore this spatial dependence quantitatively. We focus mainly on the regional hydrochemical dataset for arsenic but a similar approach could also be adopted for the other parameters. We concentrate on the regional dataset since this is at present the most complete dataset in terms of spatial coverage. However, there are many scales of variation in arsenic concentrations and with more data, it would be useful to explore other scales. In particular, the village scale of variation deserves more attention from a quantitative point of view.

These scales of variation range from the molecular to the regional (Figure 9.1) with some scales having more significance for some processes and objectives than others. For example, total arsenic concentrations in bulk sediments from Bangladesh are typically in the range 1–10 mg kg⁻¹ but within these sediments there may be minerals containing essentially no As while in other minerals in the same sediment the concentration may exceed 1000 mg kg⁻¹. Different particles of the same mineral may also have a different origin and may contain different As concentrations, and even within an individual mineral particle there may be zonation due to a changing environment during particle growth or to later diagenesis or weathering.

These spatially highly variable sediment concentrations which are seen at the microscopic scale will tend to be reflected to some extent in variations in the concentration of As in the surrounding water, and ultimately in some of the variation seen in well waters at the field scale. From a practical point of view, recognising this variability is also important when investigating the geochemical processes involved, and when reporting and interpreting the As concentration of sediments and minerals.

9.2 Country and district level

The arsenic data have been analysed using both classical statistics and geostatistics. Geostatistics is a branch of statistics that deals specifically with spatially-variable data. The mathematical background of the technique is described in standard geostatistics textbooks (Kitanidis, 1997; Webster & Oliver, 2001). The application of geostatistic involved three steps based on the assumption that the arsenic concentration could be treated as a ‘regionalised variable’: (i) computation and modelling of the variogram; (ii) prediction of concentrations by kriging, and (iii) an analysis of the errors. Kriging is superior to simple interpolation in that it uses knowledge of the spatial structure gained from the data, it interpolates exactly and provides an estimate of the error in such interpolations (kriging variance). Such an analysis can therefore lead to an insight into the spatial variance of the data. We have also used kriging to estimate the groundwater As concentration in shallow groundwaters on a 5-km grid for the whole of the sampled region of Bangladesh and to estimate the probabilities that the As concentration exceeds a specified threshold value. The following thresholds were chosen: 5 µg L⁻¹, 10 µg L⁻¹ (the WHO guideline value) and 50 µg L⁻¹ (the Bangladesh standard). A 150 µg L⁻¹ threshold was also included.

Figure 9.1. Different scales of variation and their relevance to different processes and objectives.
9.2.1 Approach

Basic statistical analysis

Basic statistical parameters were estimated for both the whole of Bangladesh and on a district basis. The frequency distributions for both the original data and the log transformed data were calculated. A major difficulty in analysing the data is that 24% of the recorded arsenic values were below the instrumental detection limit. As a result, the data distribution shows a large number of values near the lower limit that cannot be resolved further. The censored values also present a problem when attempting to calculate basic statistics. Where necessary, we have substituted half the detection limit value when calculating such statistics. Transformation of the measured arsenic concentrations to their logarithms (base 10) brings the distribution closer to normal but does not remove the peak at low concentrations. An analysis of variance (ANOVA) based on districts was carried out to estimate the within-district variance as well as the overall variance in As concentrations.

Spatial dependency

The spatial dependence of a regionalised variable is represented by the variogram. The variogram is both a theoretical function relating the magnitude of variation to the separation in space in one, two or three dimensions, and it can also be an estimate of this variation based on observed data – this is often called the experimental variogram. The variogram is calculated from the data using the formula:

\[ \gamma(h) = \frac{1}{2n(h)} \sum_{i=1}^{n(h)} (z(x_i) - z(x_i + h))^2 \]

where \( z(x) \) and \( z(x_i + h) \) are the measured values at \( x \) and \( x_i + h \) and \( h \) is the lag, \( n(h) \) is the number of pairs \( (x, x_i + h) \) separated by the vector \( h \), and \( \gamma(h) \) is the calculated semi-variance at that lag. If the variation is isotropic, then the calculated values are averages over all directions.

The experimental variogram consists of an ordered set of estimates for a chosen set of lags. The true variogram is a continuous function of \( h \). It can be approximated from the experimental variogram by fitting a theoretically acceptable equation or model. Depending on the complexity of the variation, this can be a single model such as an exponential or spherical model which is bounded, or a power function or a more complex model in which two or more simple models are combined and which may or may not be bounded, e.g. bounded linear, double spherical and double exponential models. The more complex models usually fit better than simple models and the Akaike Information Criterion (Webster and McBratney, 1989) based on the residuals can be used to determine the most appropriate, i.e. the best compromise between simplicity and goodness of fit.

Predictions using kriging

Kriging is a technique that enables estimates of the As concentration to be made at points where no observations are available. It can also be applied to blocks of land. At its simplest, kriging is a method of local-weighted averaging of the observed values. Weights are allocated to the sample data with their magnitude depending on the correlation structure of the data, i.e. the variogram of the variable, the configuration of the sampling sites and its relation to the target point or area. Using the variogram and the data, the concentration of As and the kriging variance can be estimated using ordinary kriging. However, one is often interested not only in the estimated As value at a certain point, but also in the probability, given the data, that the As concentration exceeds a certain threshold value at that point, e.g. a drinking water standard. The technique of disjunctive kriging is appropriate for such situations and was applied to our dataset. Rivoirard (1994) describes disjunctive kriging in detail, and an application of this technique is described in Webster (1991).

With disjunctive kriging, the original data are transformed to a Gaussian (normal) distribution using Hermite polynomials and then the variogram of the transformed variable is calculated and modelled. Using the variogram, the concentration of As and the indicators for each of the thresholds can then be estimated. Estimates for these indicators are given by the estimated probabilities that the concentrations exceed the value associated with each indicator (5 µg L\(^{-1}\), 10 µg L\(^{-1}\), 50 µg L\(^{-1}\) and 150 µg L\(^{-1}\), respectively, in our case).

9.2.2 Basic statistics and the data distribution

The data were split into two sets: those for shallow wells and those for deep wells. A depth of 150 m was used for making this split. These two datasets were treated separately. Data for the shallow wells (3208 sites) were analysed and those for deep wells. A depth of 150 m was used for making this split. These two datasets were treated separately. The data were split into two sets: those for shallow wells and those for deep wells. A depth of 150 m was used for making this split. These two datasets were treated separately. Data for the shallow wells (3208 sites) were analysed using both basic statistics and also geostatistics. Only the basic statistical analysis was carried out on the deep wells.

Summary statistics for As for both datasets are given in Table 9.1. The probability distributions of the As concentration exceed a certain threshold value at that point, e.g. a drinking water standard. The technique of disjunctive kriging is appropriate for such situations and was applied to our dataset. Rivoirard (1994) describes disjunctive kriging in detail, and an application of this technique is described in Webster (1991).

### Table 9.1. Summary statistics for deep wells (>150 m) and shallow wells (<150 m)

<table>
<thead>
<tr>
<th></th>
<th>Shallow wells (&lt;150 m)</th>
<th>Deep wells (&gt;150 m)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of observations</td>
<td>3208</td>
<td>326</td>
</tr>
<tr>
<td>Mean (µg L(^{-1}))</td>
<td>60.52</td>
<td>2.95</td>
</tr>
<tr>
<td>Standard deviation (µg L(^{-1}))</td>
<td>123.11</td>
<td>8.29</td>
</tr>
<tr>
<td>Geometric mean (µg L(^{-1}))</td>
<td>1515.59</td>
<td>68.7</td>
</tr>
<tr>
<td>Variance (µg L(^{-1}))^2</td>
<td>2.47</td>
<td>1.06</td>
</tr>
<tr>
<td>Geometric variance (µg L(^{-1}))^2</td>
<td>22.77</td>
<td>2.08</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Percentiles (µg L(^{-1}))</th>
<th>Shallow wells (&lt;150 m)</th>
<th>Deep wells (&gt;150 m)</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>0.1</td>
<td>0.25</td>
</tr>
<tr>
<td>20</td>
<td>0.2</td>
<td>0.25</td>
</tr>
<tr>
<td>30</td>
<td>0.6</td>
<td>0.25</td>
</tr>
<tr>
<td>40</td>
<td>2.0</td>
<td>0.75</td>
</tr>
<tr>
<td>50 (Median)</td>
<td>6.1</td>
<td>1.04</td>
</tr>
<tr>
<td>60</td>
<td>16.6</td>
<td>1.56</td>
</tr>
<tr>
<td>70</td>
<td>40.4</td>
<td>2.07</td>
</tr>
<tr>
<td>80</td>
<td>83.4</td>
<td>3.17</td>
</tr>
<tr>
<td>90</td>
<td>199.7</td>
<td>5.29</td>
</tr>
<tr>
<td>95</td>
<td>320.0</td>
<td>9.39</td>
</tr>
<tr>
<td>99</td>
<td>570.7</td>
<td>42.85</td>
</tr>
</tbody>
</table>
trations, and the log As concentrations of the shallow wells, are given in Figure 9.2.

**Shallow wells**

*Whole country.* The mean As concentration in the shallow wells is 60.5 µg L\(^{-1}\) with a maximum of 1670 µg L\(^{-1}\) and a minimum of less than the detection limit (usually less than 0.5 or 0.25 µg L\(^{-1}\)). The distribution is strongly skewed with almost 80% of the samples in the lowest class. Only three samples were greater than 1000 µg L\(^{-1}\). The median concentration of 6.1 µg L\(^{-1}\) is much smaller than the mean value which reflects the skewed nature of the distribution. The variance of the distribution is large leading to a standard deviation of 123 µg L\(^{-1}\) which is 203% of the mean value. The As concentrations were transformed to their logarithms (base 10) in order to attempt to normalise this long-tailed distribution. The skewness is reduced, but the fact that many samples were close to or below the detection limit leaves many values in the lower classes (−1.0 to −0.6). Also, after log transformation the distribution is still far from normal.

*Districts.* The analysis of variance (ANOVA) based on the As concentrations in each of the 61 sampled districts are given in Table 9.2. The variance ratio from the mean squares is approximately 30 in the case of the untransformed data and 41 in the case of the log transformed data. This points to a large variation in the means between districts, i.e. it confirms a significant regional pattern.

**Deep wells**

For the deep wells the As concentrations are much lower. The mean concentration is about 3 µg L\(^{-1}\), the median concentration is 1 µg L\(^{-1}\), the minimum is less than the detection limit and the maximum is 108 µg L\(^{-1}\). The standard deviation is also very large, 8.3 µg L\(^{-1}\), which is 277% of the mean value. These statistics are strongly influenced by the large number of censored data values.

Table 9.2. District-wise analysis of variance (ANOVA) for the As measurements from the shallow wells

<table>
<thead>
<tr>
<th>Source of variation</th>
<th>Degrees of freedom</th>
<th>Sum of squares</th>
<th>Mean square</th>
<th>Variance ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Untransformed As measurements</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Between districts</td>
<td>60</td>
<td>18016773</td>
<td>300280</td>
<td>30.88</td>
</tr>
<tr>
<td>Within districts</td>
<td>3146</td>
<td>30587459</td>
<td>9273</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>3206</td>
<td>48604232</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Log transformed As measurements</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Between districts</td>
<td>60</td>
<td>1918.98</td>
<td>31.98</td>
<td>41.35</td>
</tr>
<tr>
<td>Within districts</td>
<td>3146</td>
<td>2433.50</td>
<td>0.77</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>3206</td>
<td>4352.48</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 9.2. Histograms of the arsenic and log arsenic data (n=3534)
9.2.3 Spatial analysis

**Whole country**

The variogram for the log_{10}-transformed As concentrations in the shallow wells (Figure 9.3) was computed up to a maximum lag distance of 150 km at intervals of 2 km. The horizontal line on the graph indicates the overall variance. The experimental variogram was fitted to an isotropic, exponential model:

\[
\gamma(h) = c_0 + c(1 - e^{-ah})
\]

where \(c_0\) represents the nugget variance, \(c\) is the sill variance and \(a\) is the distance parameter. The optimal parameter values based on a minimisation of the sum of squares of the residuals were: \(c_0=0.5335 \text{ (µg L}^{-1})^2\), \(c=0.724 \text{ (µg L}^{-1})^2\) and \(a=48100\) m.

The nugget variance, which is the intercept of the function on the ordinate, is large. It represents the variation in As concentrations over distances smaller than a couple of km interval plus any measurement error. This nugget variance is equivalent to some 42% of the overall variance observed in the data for the whole country. The sampling strategy deliberately avoided closely-spaced samples and so it is not surprising that little is known about the spatial dependence over short distances.

It was decided that a maximum of 20 data points would be used for making the kriged estimates and so a maximum lag distance of 150 km was chosen. Given the sample density of the data, it is unlikely that the maximum lag distance required to achieve this would often exceed 150 km. With increasing lag distances, the variance increased steadily. Even at the maximum lag distance, the sill value (being a constant variance not dependent on the lag distance) was not reached. This reflects the long-range component in the As concentrations. Figure 9.4 shows that in general As
concentrations in the south of Bangladesh tend to be much greater than in the north. District-mean As concentrations vary by more than two orders of magnitude.

**Districts**

In order to see if the spatial variation depended on the mean As concentration in a district, a distinction was made between districts having a mean As concentration of less than 50 µg L⁻¹ (SET 1) and districts with a mean As concentration greater than 50 µg L⁻¹ (SET 2).

The two sets of data were then analysed separately. The frequency distributions for both sets are given in Figure 9.5 and their variograms given in Figure 9.6. In both sets of data, there are many values close to or less than the detection limit although the influence of these values on the resulting distribution is less important for the districts with high As concentrations. The variograms for the two sets of data are remarkably similar. The nugget variance is only slightly smaller for SET 1 than for SET 2 and both values are similar to the nugget variance of the combined data. Also, the overall variances of the two data-sets are similar. Both experimental variograms fitted an exponential model well. The similarity between the variograms implies that the spatial variation of As concentrations in SET 2 is not essentially different from that of SET 1 and that the variance does not correlate with the As concentrations. This means that for the kriging, the overall variogram can be used for the whole of Bangladesh and that no distinction has to be made between areas with small and large As concentrations.

**9.2.4 Predicting arsenic concentrations and the probability that a threshold level has been exceeded**

**Kriging and disjunctive kriging calculations**

In the first instance, ordinary punctual kriging was carried out with the log-transformed As concentrations. Kriged estimates and kriging variances were calculated on a 5-km grid for the whole of Bangladesh. The kriged estimates were calculated for the centre point of each grid cell. The parameters used for the kriging are given in Table 9.3. Since the kriging was carried out on the log-transformed As concentrations, the estimates derived must be transformed back to the original scale taking into account the estimated kriging variance:

$$\xi = \exp(y \ln 10 + 0.5\sigma^2 (\ln 10)^2)$$
where $\gamma$ is the estimated As concentration in $\mu g L^{-1}$, $y$ is the estimated $\log_{10}$ As concentration and $\sigma^2$ is the kriging variance in terms of $\{\log_{10}(\mu g L^{-1})\}^2$. In order to check the reliability of the results, both ordinary punctual kriging and block kriging using the original variable were also carried out.

Subsequently, disjunctive kriging (von Steiger et al., 1996; Webster, 1991) was carried out on the original As concentrations. The Hermite transformation of the original data to a standard normal variable is presented in Figure 9.7(a). The large deviation from a linear function for the transformation function indicates the strong skewness of the original data. If the original data were normally distributed then the transformation function would be linear. Looking at the distribution of the Hermite-transformed As concentrations (Figure 9.7(b)), it is clear that although the transformed variable is closer to a normal distribution than the original variable, the distribution is still not normal. It is assumed that the large proportion of values close to the detection limit makes it impossible for a Hermite transformation to give a normal distribution. Therefore one of the conditions for disjunctive kriging is not fulfilled. This means that the disjunctive-kriged estimates, their variances and the estimated threshold probabilities have to be treated with caution. An alternative approach to overcome the non-normality of the distribution would be to apply indicator kriging (Goovaerts, 1997). No assumptions concerning the distribution of the kriged variable are made with indicator kriging and the deviation from normality would not pose a problem. However, indicator kriging is time-consuming and could not be carried out within this project.

The variogram for the Hermite-transformed As concentrations is shown in Figure 9.7(c). It was fitted to an exponential model. Disjunctive punctual kriging was used to calculate arsenic concentrations and their variances and also the probabilities that the As concentration exceeded the 5, 10, 50 and 150 $\mu g L^{-1}$ thresholds. These estimates, variances and probabilities were calculated for a 5 km grid. The parameters used in these calculations are given in Table 9.3.

The representative sample interval (which is the area for which one sample is representative – in this case, this could be interpreted as the area of the depression cone of a sampled well) is in this case smaller than the grid area used for the estimates (25 km$^2$ for a 5 km grid). In geostatistical terms, the estimates have been calculated for a scale greater than the scale of the ‘support’. In order to overcome this problem, which might affect the estimated probabilities, Webster (1991) describes how disjunctive block kriging can be used to increase the scale of the support. However, because of time constraints, this approach could not be tested.

<table>
<thead>
<tr>
<th>Table 9.3. Parameters used for ordinary and disjunctive kriging</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parameter</td>
</tr>
<tr>
<td>-----------</td>
</tr>
<tr>
<td>Type of kriging</td>
</tr>
<tr>
<td>Max number of data for any one estimate</td>
</tr>
<tr>
<td>Interval between estimates in the X and Y-direction</td>
</tr>
<tr>
<td>Variogram model</td>
</tr>
<tr>
<td>Variogram (m)</td>
</tr>
<tr>
<td>Nugget variance (${\log_{10}(\mu g L^{-1})}^2$)</td>
</tr>
<tr>
<td>Sill variance (${\log_{10}(\mu g L^{-1})}^2$)</td>
</tr>
<tr>
<td>Number of terms in the Hermite expansion</td>
</tr>
<tr>
<td>Number of terms in the least squares fit</td>
</tr>
<tr>
<td>Number of terms in the Hermite integration</td>
</tr>
</tbody>
</table>

Figure 9.7. Behaviour of the Hermite-transformed variable for disjunctive kriging of arsenic concentrations. (a) The transform function between the original and the Hermite-transformed variable; (b) the distribution of the Hermite-transformed concentrations, and (c) the variogram of the Hermite-transformed variable.
9.2.5 Analysis of the results

Comparison of disjunctive kriging and ordinary kriging

In order to compare the As estimates obtained by disjunctive kriging and ordinary kriging, the two estimates have been plotted against each other for each point on the 5 km grid. In Figure 9.8(a), the disjunctive kriged estimates are compared with the ordinary kriged estimates using the log-transformed variable, while in Figure 9.8(b) the disjunctive kriged estimates are compared with the ordinary kriged estimates using the original variable.

Ideally the points should plot on a straight line through the origin in both graphs. This is certainly not the case for ordinary kriging using the log-transformed variable (Figure 9.8(a)). The estimates obtained with ordinary kriging using the log-transformed As concentrations lead to far greater estimates (up to 6000 µg L⁻¹ while the maximum value in the data was 1670 µg L⁻¹) than the estimates obtained with disjunctive kriging. The reason for this large discrepancy is not understood, and needs further investigation. The estimates obtained with ordinary kriging on the original variable agree well with the disjunctive kriging estimates. Most of the points which deviate considerably from the diagonal have a large kriging variance (and lie outside the Bangladesh borders) as can be seen from Figure 9.8(c) where points with a disjunctive kriging variance greater than 11,000 (µg L⁻¹)² have been eliminated.

The large kriging variances lead to a great deal of scatter. This means that the estimates from ordinary kriging of the log-transformed variable are of dubious value. Because the distribution of the original arsenic concentration, on which ordinary kriging is based, was so skewed, preference was given to the disjunctively-kriged estimates which were based on Hermite-transformed concentrations. These are therefore the estimates which have been mapped.

Kriged arsenic concentrations and kriging variances

The map of the disjunctively-kriged estimates is shown in Figure 9.9. An arbitrary colour scale (a ‘temperature’ scale) was chosen to distinguish the concentrations. Kriged estimates having a kriging variance greater than 11,000 (µg L⁻¹)² are not displayed because the error of the estimate would be too large. The maximum kriged estimate was 417 µg L⁻¹ which is smaller than the maximum measured As concentration of 1670 µg L⁻¹. This illustrates the smoothing effect of kriging. Due to the high sampling density the kriging variances are not very informative. The kriging variances depend principally on the sampling density and since the sampling density was fairly homogeneous over the whole of Bangladesh (with the exception of the south-east and the islands), the map of the kriging variances (not shown) shows little variation within Bangladesh.

Probabilities for exceeding a threshold

A map of the probabilities was produced for each of the specified threshold concentrations (Figure 9.10). As before, a linear scale was used and probabilities with a kriging variance greater than 11000 (µg L⁻¹)² were not dis-
played because of the poor reliability of these estimates.

The estimated probabilities for each of the points has been plotted against the disjunctively-kriged estimates (Figure 9.11(a-d)). In most cases, the larger the estimated As concentration the greater the probability that the threshold will be exceeded. The farther away that the threshold was from the detection limit and the closer it was to the mean of the distribution, the more nearly linear was the relation between the estimates and their probabilities. This is clearly illustrated in Figure 9.11(c-d). For the 150 µg L⁻¹ threshold, an almost linear correlation exists between the two variables. Also the results for the 50 µg L⁻¹ threshold also suggests, but with more scatter, a linear relation. For the 10 µg L⁻¹ threshold and certainly for the 5 µg L⁻¹ threshold, the relationship seems to curve towards the probability axis indicating a greater range of probabilities for the same estimated As concentration. For these graphs, where the threshold value is much less than the mean value, many observations exceed the threshold value. The estimates might not be as reliable as the probabilities found for thresholds closer to the mean of the distribution.

An estimate of the number of people exposed to an As concentration greater than 50 µg L⁻¹ was calculated by combining the arsenic map with the upazila-based population map. The following approach was taken: (i) the 1991-based upazila populations were increased by 2.1% per year to give estimates of the 1999 upazila populations, i.e. the factor (1+0.021)^8=1.181 was used. This gave an estimated 1999 population of 125.5 million for the whole of Bangladesh; (ii) the population densities were calculated for each upazila; (iii) the population of each 5 km grid cell was estimated based on the area-weighted contribution of each upazila to the grid cell; (iv) this value was multiplied by the estimated probability for As exceedance of each grid cell to give the exposed population for that cell (a zero probability was assumed for unsampled upazilas principally from the Chittagong Hill Tracts), and (v) the total population was estimated by summing the exposed populations for all of the cells.

The following three factors were also taken into account in this calculation:

1. the upazila “Dhaka Metro” was excluded because the drinking water for most people in Dhaka city is derived from the deeper, uncontaminated aquifer;
2. the south east coastal part of the country, some of the islands and the Chittagong Hill Tracts were excluded because the kriging variances exceeded 13000 (µg L⁻¹)² and the estimated probabilities were therefore deemed too unreliable (in practice, these are all believed to be areas of relatively low arsenic risk in part because the shallow groundwater is not primarily used for drinking water);
3. it was assumed that 97% of the population used well water for drinking (and cooking) water.

No explicit ‘corrections’ were made for large population centres such as Chittagong and Khulna where drinking water is supplied from a relatively small number of municipal wells. These are usually set deeper than many of the surrounding shallow tubewells and in the case of Khulna and Chittagong are believed to be low in As.

Bearing in mind these assumptions, this calculation leads to an estimated total number of people exposed to an As concentration of greater than 50 µg L⁻¹ of 35.2 million. Excluding the city of Khulna would reduce this number by approximately 0.7 million people. A map showing the exposed population densities obtained is given in Figure 9.12. Making the same calculation to estimate the number of people exposed to an As concentration exceeding 10 µg L⁻¹ gave an estimate of 56.7 million people.

Simpler calculations of the exposed population were also made using entirely upazila-based statistics. The percentage of wells in each upazila exceeding a given threshold was multiplied by the population of that upazila and then by the factor 0.97 since 97% of the rural population are assumed to use groundwater. UNSampled upazilas were assumed to contribute no exceedances. This gave estimates of the population exposed to greater than 50 µg L⁻¹ and 10 µg L⁻¹ as 28.1 million and 46.4 million people, respectively. The kriged estimates are significantly greater than the purely upazila-based estimates and in the absence of data to the contrary are believed to be more accurate. Both sets of calculations could be refined with higher resolution and more up-to-date population maps and with a greater density of sampled wells. Nevertheless, the figures clearly demonstrate the massive scale of the problem. Changing from the Bangladesh standard to the WHO guideline value would add approximately another 20 million people to the population ‘exposed’.

Figure 9.9. Smoothed map showing the estimated arsenic concentrations in shallow wells (<150 m) based on disjunctive kriging.
Figure 9.10. Probabilities, calculated using disjunctive kriging, that the arsenic-concentration exceeds specified thresholds.
(a) Threshold = 5 µg L\(^{-1}\)  (b) Threshold = 10 µg L\(^{-1}\)  (c) Threshold = 50 µg L\(^{-1}\)  (d) Threshold = 150 µg L\(^{-1}\).
It is unfortunate that many of the most-contaminated areas south-east of Dhaka are also areas of high population density. The percentage of Bangladesh by area that exceeds the 50 µg L⁻¹ Bangladesh arsenic standard has also been calculated as a function of the associated probability (Table 9.4). For example, there is a 90% probability of finding a well to be contaminated in only 3% of Bangladesh and a 50:50 chance of finding a well to be contaminated in 17%, or about one-sixth, of Bangladesh.

### Table 9.4. Percentage of Bangladesh by area that exceeds a probability limit with respect to the 50 µg L⁻¹ Bangladesh arsenic standard

<table>
<thead>
<tr>
<th>Probability limit with respect to the Bangladesh Arsenic Standard (50 µg L⁻¹)</th>
<th>0.1</th>
<th>0.2</th>
<th>0.3</th>
<th>0.4</th>
<th>0.5</th>
<th>0.6</th>
<th>0.7</th>
<th>0.8</th>
<th>0.9</th>
</tr>
</thead>
<tbody>
<tr>
<td>Percentage of the total area of Bangladesh exceeding the probability limit (in%)</td>
<td>62</td>
<td>47</td>
<td>35</td>
<td>25</td>
<td>17</td>
<td>13</td>
<td>9</td>
<td>6</td>
<td>3</td>
</tr>
</tbody>
</table>

Percentage of the total area of Bangladesh for which the probability was calculated = 92%.

9.3 **LOCAL VARIATION**

**Lakshmipur upazila**

The Lakshmipur dataset from the Special Study Area survey (Chapter 7) (59 observations, Table 9.5) was selected for further statistical analysis. All samples selected were obtained from a well depth of less than 150 m. The mean As concentration of 159 µg L⁻¹ is considerably higher than the national average since Lakshmipur is located in the
high-As region of Bangladesh. The maximum concentration is 986 µg L\(^{-1}\) and the standard deviation is 181.5 µg L\(^{-1}\), 114% of the mean concentration. The frequency distributions are discontinuous because of the small number of samples (Figure 9.13(a,b)).

The experimental variogram of the log As concentrations is shown in Fig 9.13(c). The interval between the calculated points is 500 m with a maximum lag of 20 km. Although the experimental variogram is scattered, a trend can be observed and a fit to an isotropic, exponential model was obtained:

\[ \gamma(\delta) = \epsilon_0 + \epsilon \left( 1 - e^{-\delta/a} \right) \]

where \(\epsilon_0\) is the nugget variance, \(\epsilon\) is the sill variance and \(a\) is the distance parameter. The optimal parameters based on a minimisation of the sum of squares of the residuals were: \(\epsilon_0=0.0909\), \(\epsilon=0.2951\), and \(a=15800\) m. The variogram indicates a significant reduction in the nugget variance for the Lakshmipur dataset compared with the nugget variance calculated from the national dataset (0.53). All variances have units of log (µg L\(^{-1}\))^2.

**Mandari village**

The Mandari dataset contains 337 observations (Table 9.6).

Table 9.5. Summary statistics for the Lakshmipur wells

<table>
<thead>
<tr>
<th>Shallow Wells (&lt;150 m)*</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of observations</td>
<td>59</td>
</tr>
<tr>
<td>Minimum (µg L(^{-1}))</td>
<td>1.3</td>
</tr>
<tr>
<td>Mean (µg L(^{-1}))</td>
<td>159.0</td>
</tr>
<tr>
<td>Maximum (µg L(^{-1}))</td>
<td>986.0</td>
</tr>
<tr>
<td>Standard deviation (µg L(^{-1}))</td>
<td>181.5</td>
</tr>
<tr>
<td>Variance (µg L(^{-1}))^2</td>
<td>32945.2</td>
</tr>
<tr>
<td>Geometric mean (µg L(^{-1}))</td>
<td>91.8</td>
</tr>
<tr>
<td>Geometric variance (µg L(^{-1}))^2</td>
<td>1.9</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Percentiles (µg L(^{-1}))</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>21.8</td>
</tr>
<tr>
<td>20</td>
<td>46.2</td>
</tr>
<tr>
<td>30</td>
<td>64.5</td>
</tr>
<tr>
<td>40</td>
<td>81.8</td>
</tr>
<tr>
<td>50 (Median)</td>
<td>88.7</td>
</tr>
<tr>
<td>60</td>
<td>114.0</td>
</tr>
<tr>
<td>70</td>
<td>165.0</td>
</tr>
<tr>
<td>80</td>
<td>256.0</td>
</tr>
<tr>
<td>90</td>
<td>390.0</td>
</tr>
<tr>
<td>95</td>
<td>669.5</td>
</tr>
<tr>
<td>99</td>
<td>986.0</td>
</tr>
</tbody>
</table>

*All wells are shallow wells

\(\gamma(\delta)\) = \epsilon_0 + \epsilon \left( 1 - e^{-\delta/a} \right)
of which 227 are for tubewells shallower than 150 m. The mean As concentration of the shallow tubewells is 122.4 µg L–1 with a maximum of 707 µg L –1. The mean concentration is again considerably larger than the overall mean As concentration in the country. The standard deviation is 132.6 µg L–1, which is 106% of the mean. The frequency distribution (Figure 9.14(a,b)) of the As concentrations is strongly skewed. Taking logarithms eliminates much of the skewness.

The experimental variogram of the log-As concentrations of the shallow wells is shown in Figure 9.14(c). The interval between the calculated points is 100 m. The calculated values show a large scatter and are situated close to the overall sample variance. This indicates that the experimental variogram consists of pure nugget variance indicating that there is no spatial dependence in the dataset. The accuracy of the GPS measurements of the map coordinates will affect the variogram because of the small distance between many of the wells. The nugget variance of the Mandari variogram is approximately the same as the nugget variance for the Lakshmipur data (0.091) and about 5 times less than for the national dataset.

### Table 9.6. Summary statistics for the deep wells (>150 m) and the shallow wells (<150 m) in Mandari

<table>
<thead>
<tr>
<th></th>
<th>Shallow Wells (&lt;150 m)</th>
<th>Deep Wells (&gt;150 m)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of observations</td>
<td>228</td>
<td>10</td>
</tr>
<tr>
<td>Minimum (µg L–1)</td>
<td>11.71</td>
<td>0.03</td>
</tr>
<tr>
<td>Mean (µg L–1)</td>
<td>125.04</td>
<td>5.43</td>
</tr>
<tr>
<td>Maximum (µg L–1)</td>
<td>701.95</td>
<td>16.00</td>
</tr>
<tr>
<td>Standard deviation (µg L–1)</td>
<td>132.56</td>
<td>6.10</td>
</tr>
<tr>
<td>Variance (µg L–1)²</td>
<td>17572.00</td>
<td>37.26</td>
</tr>
<tr>
<td>Geometric mean (µg L–1)</td>
<td>86.31</td>
<td>1.51</td>
</tr>
<tr>
<td>Geometric variance (µg L–1)²</td>
<td>1.34</td>
<td>9.84</td>
</tr>
<tr>
<td>Percentiles (µg L–1)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>37.79</td>
<td>0.04</td>
</tr>
<tr>
<td>20</td>
<td>45.14</td>
<td>0.40</td>
</tr>
<tr>
<td>30</td>
<td>60.25</td>
<td>0.92</td>
</tr>
<tr>
<td>40</td>
<td>67.68</td>
<td>1.69</td>
</tr>
<tr>
<td>50 (median)</td>
<td>77.92</td>
<td>2.53</td>
</tr>
<tr>
<td>60</td>
<td>92.25</td>
<td>4.45</td>
</tr>
<tr>
<td>70</td>
<td>110.07</td>
<td>7.93</td>
</tr>
<tr>
<td>80</td>
<td>138.33</td>
<td>11.54</td>
</tr>
<tr>
<td>90</td>
<td>300.18</td>
<td>14.20</td>
</tr>
<tr>
<td>95</td>
<td>465.75</td>
<td>15.10</td>
</tr>
<tr>
<td>99</td>
<td>610.77</td>
<td>15.82</td>
</tr>
</tbody>
</table>

The experimental variogram of the log-As concentrations of the shallow wells is shown in Figure 9.14(c). The interval between the calculated points is 100 m. The calculated values show a large scatter and are situated close to the overall sample variance. This indicates that the experimental variogram consists of pure nugget variance indicating that there is no spatial dependence in the dataset. The accuracy of the GPS measurements of the map coordinates will affect the variogram because of the small distance between many of the wells. The nugget variance of the Mandari variogram is approximately the same as the nugget variance for the Lakshmipur data (0.091) and about 5 times less than for the national dataset.

### 9.4 CONCLUSIONS

Classical statistics and geostatistics have been used to analyse the spatial variability of some of the water quality data collected during this project, principally the arsenic data from the National Hydrochemical Survey.

The NHS groundwater data were divided into those from shallow wells (<150 m) and those from deep wells (>150 m) since there was a clear difference in their chemistries, particularly in terms of their arsenic concentrations. Shallow wells (n=3208) had a mean arsenic concentration of 60 µg L–1 whereas deep wells (n=326) had a mean concentration of 3.0 µg L–1. The frequency distribution in both shallow and deep wells was strongly skewed. For example, the coefficient of variation of arsenic concentrations in the shallow wells was 203%. More than 30% of the samples were below 1 µg L–1, and 24% were below the instrumental detection limit (normally 0.25 or 0.5 µg L–1). The maximum As concentration was 1670 µg L–1.

Log transforming the data reduced the skewness in the distribution to some extent but there are indications of two (or more) populations. Again the large number of samples below the instrumental detection limit makes a detailed analysis difficult. The geometric mean arsenic concentration in the shallow groundwaters was 5.5 µg L–1, an order of magnitude lower than the arithmetic mean.

Varioigram analysis was performed on the log-transformed data. The data fitted well to an isotropic exponential model. There was no evidence for significant anisotropy. The nugget variance, representing the variance for separations of less than about 2 km, was large – about 40% of the variance (in log terms) over the whole country. This reflects the local experience. It has been found to be extremely difficult to predict the concentration of arsenic in unsampled wells at the village scale.

There is also a long range component reflecting the variation in the major geological units at the national scale. Even at separations of 150 km, the variance had not reached the sample variance.

Figure 9.14. Histogram of (a) normal (b) log-transformed (b) data from Mandari. (c) shows the variogram for arsenic.
The average As concentration was calculated for each of the 61 sampled districts. The districts were divided into two sets: those with an average As concentration of <50 µg L⁻¹ and those >50 µg L⁻¹. Separate variogram analysis of the two sets indicated that there was no significant difference between them. Therefore subsequent variogram analysis was carried out for the whole of Bangladesh using the single national data set.

Ordinary punctual kriging based on the raw data and the log-transformed data, and disjunctive kriging based on Hermite-transformed data, showed that the estimates from disjunctive kriging agreed best with those of the raw data. Back-transforming the log-transformed ordinary kriged estimates to the original scale gave some very high, and probably unrealistic, concentrations (greater than 5000 µg L⁻¹).

Disjunctive kriging was used to produce a smoothed As map based on a 5 km grid and it also gave a series of maps which indicated the probability that a particular threshold As concentration (5, 10, 50, 150 µg L⁻¹) would be exceeded. These maps highlighted the high arsenic areas in the south of Bangladesh and the low arsenic areas in much of northern Bangladesh including the Pleistocene Tracts.

Estimates of the population exposed to arsenic contaminated water were made by overlaying the maps of the arsenic distribution and the estimated upazila-based populations. Two methods of estimation were used depending on how the discretisation was accomplished. When the arsenic statistics were based on average percentage exceedances (at the 50 µg L⁻¹ level) over whole upazilas, an exposed population of 28 million was estimated. When the arsenic statistics were based on interpolated estimates on a 5-km grid using disjunctive kriging (the ‘smoothed arsenic map’), the estimate was 35 million. If the more stringent WHO guideline value of 10 µg L⁻¹ is used, then these figures increase to 46 million and 57 million, respectively. In the absence of any data to the contrary, we assume that the kriged estimates (larger figures) are more reliable. It is uncertain what the errors associated with these figures are but clearly the exposed population is very large. There is a 50:50 chance of a shallow well being contaminated in 17%, or about one-sixth, of Bangladesh.

We had insufficient short-range data (i.e. pairs of wells with separations of less than 2 km) to calculate a reliable variogram for short distances. However, an analysis of the available shallow tubewell data from Lakshimipur upazila (n=59) and Mandari village (n=228) suggests that there is an approximately 5-fold reduction in the nugget variance in the more localised data sets compared with the national data set. We found no significant spatial structure in the Mandari data set.