Risk Analysis, Vol. 24, No. 6, 2004

Direct Data Manipulation for Local Decision Analysis as Applied to the Problem of Arsenic in Drinking Water from Tube Wells in Bangladesh

Andrew Gelman,[1]* Matilde Trevisani,[2] Hao Lu,[3] and Alexander van Geen[4]

A wide variety of tools are available, both parametric and nonparametric, for analyzing spatial data. However, it is not always clear how to translate statistical inferences into decision recommendations. This article explores the possibilities of estimating the effects of decision options using very direct manipulation of data, bypassing formal statistical analysis. We illustrate with the application that motivated this research, a study of arsenic in drinking water in nearly 5,000 wells in a small area in rural Bangladesh. We estimate the potential benefits of two possible remedial actions: (1) recommendations that people switch to nearby wells with lower arsenic levels; and (2) drilling new community wells. We use simple nonparametric clustering methods and estimate uncertainties using cross-validation.

KEY WORDS: Cluster analysis; environmental statistics; public health; spatial statistics

1. INTRODUCTION

A challenge in statistics for risk analysis is to go beyond inferences about parameters and estimate the consequences of decision options. For example, in environmental statistics, data typically are indexed spatially, and these data can be used in making local decisions. In this article, we explore how the outcomes arising from some of these decisions can be estimated very simply, using data manipulations that mimic the decisions being studied—in this case, switching of wells and drilling new wells for safe drinking water in Bangladesh, as we discuss in Section 1.2.

[1] Department of Statistics and Department of Political Science, Columbia University, NY.
[2] Department of Economic and Statistical Sciences, University of Trieste, Italy.
[3] Thales Corp., NY.
[4] Lamont-Doherty Earth Observatory, Palisades, New York.
* Address correspondence to Andrew Gelman, Department of Statistics and Department of Political Science, Columbia University, NY; gelman@stat.columbia.edu, http://www.stat.columbia.edu/∼gelman/.

1.1. Background on Arsenic in Bangladesh

This work was motivated by the immediate problem of widespread arsenic poisoning from wells used for drinking water in rural Bangladesh. The arsenic in these wells is a natural contaminant and can be extremely high, sometimes more than an order of magnitude above the Bangladesh drinking water standard of 50 μg/l and the World Health Organization guideline of 10 μg/l.[1] Ironically, most of the wells have been drilled in the past 10 years, as a response to the high levels of microbial contamination in surface water. The wells are so-called tube wells, constructed with PVC pipe sunk into holes dug in the ground, installed to draw water with a hand pump from the bottom of the tube. In our study area, most of the wells tap into sandy groundwater aquifers that are between 40 and 100 ft deep, although the depths vary from less than 30 ft to more than 300 ft. There are estimated to be over 10 million wells in the country, and the vast majority have been installed privately.[2,3]

As part of an intensive local public health study, the arsenic levels in a set of 4,827 wells within

0272-4332/04/0100-1597$22.00/1

Araihazar Upazila, were measured in 2000 (see Fig. 1). In addition, other information was gathered about each well, including the depth of the well, the year it was installed, and the number of persons who were using it for drinking water.[4] These data, which we analyze here, are an intensive local sample and are distinct from the much-analyzed British Geological Survey data.[2] In total, the wells in our study served about 55,000 people, with a median number of 11 users per well.

Arsenic is a cumulative metalloid poison causing various cancers and has no known safe threshold, and so it is reasonable to measure public health risk with total exposure (rather than, e.g., maximum exposure, or proportion of time exposed above some threshold). Is the arsenic concentration in the local well a good measure of a person's arsenic exposure? To check this, urine is being gathered from 10,000 local residents, and its arsenic level, for each person, is being compared to that in the well that the person reported using for drinking water. A previous study found strong correlation between arsenic levels in drinking water and urine,[5] confirming that it is reasonable to work with measurements of wells, which we do for the rest of this article. In addition, field and laboratory studies have confirmed the accuracy and reliability of our well-water arsenic measurements.[4] A study of 3,000 wells over all of Bangladesh estimates serious public health consequences from arsenic in drinking water.[6]

A potential concern in studying well arsenic is the stability of arsenic levels and their measurements. To our knowledge, there is no credible (i.e., with adequate quality control) evidence of large fluctuations. Groundwater arsenic appears to be remarkably constant.[2,3] This does not exclude the possibility of gradual changes in relation to well age that must be taken into account when installing new wells.[7]

In this article, we use our intensively sampled data on tube wells to assess strategies of switching to safer wells and drilling new wells at depths such that arsenic levels would be expected to be low. A variety of other strategies have been proposed to deal with the arsenic problem, including purification of surface water, rainwater harvesting, and arsenic removal from groundwater. Here we focus on well switching and new deep wells because these have been found effective, at least in the short run, to lower arsenic levels in drinking water and urine.[4,8]

1.2. Direct Data Manipulation for Decision Analysis

We develop decision recommendations for the arsenic problem in Araihazar by performing calculations on our database of 5,000 wells. Although the particular techniques we use will not be directly applicable to most decision problems, we believe that this general approach, bypassing the usual steps of statistical modeling, can be useful in a variety of problems involving large data sets that require immediate local action. For arsenic in Bangladesh, decisions are made at the village level—where to drill community wells—and by individual households, which must decide where to get their water and whether to install privately owned wells.

Initially, we approached the problem by constructing various summaries of the data in order to estimate the distribution of arsenic levels in Araihazar, as well as the relation between depth of the well and arsenic level.[3,4] We soon realized that some of the most important short-term questions could be addressed by direct computations on our well data, without the need for estimating statistical distributions. After some data exploration in Section 2, we present estimates in Section 3 of the effects of a proposed program to encourage users of dangerous wells to switch to nearby safer wells. Section 4 presents recommendations for locations to drill new wells in order to best serve people who are not near any safe wells. When drilling new wells, there is a question of how deep to drill, and here it becomes more useful to model the arsenic level of a well as a function of location and depth. We do so using nonparametric clustering methods, estimating uncertainty using cross-validation. Finally, in Section 5, we consider various decision recommendations that would be appropriate for other parts of Bangladesh, since our analysis here directly applies only to the small region of our study.

2. EXPLORATORY ANALYSIS OF THE ARSENIC DATA

We begin by summarizing the arsenic concentrations as a function of the depths of the wells. As is shown in Fig. 2, two-thirds of the wells in the area studied are between 40 and 100 ft deep (the PVC tubes come in 20-ft lengths, which explains the discreteness in the well depths, typically reported in half tube-lengths). The figure also shows that, unfortunately, water from these depths features the highest average arsenic levels.

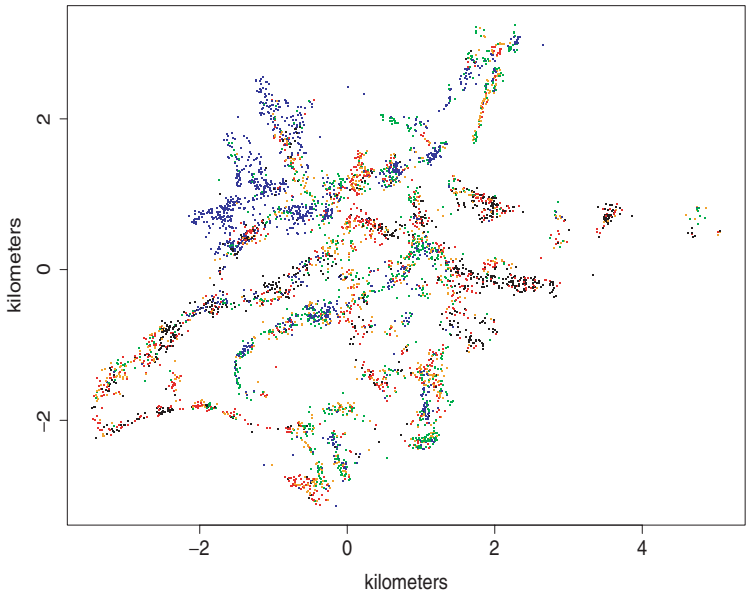

Fig. 1. Tube wells in a section of Araihazar Upazila, Bangladesh. (The (0, 0) point on this graph is at latitude 23.8° north and longitude 90.6° east.) Each dot represents a well, and these are all the wells in this area. Colors indicate arsenic levels: blue (less than 10 μg/l), green (10–50), orange (50–100), red (100–200), and black (>200). By comparison, the maximum recommended levels designated by Bangladesh and the World Health Organization are 50 and 10 μg/l, respectively.

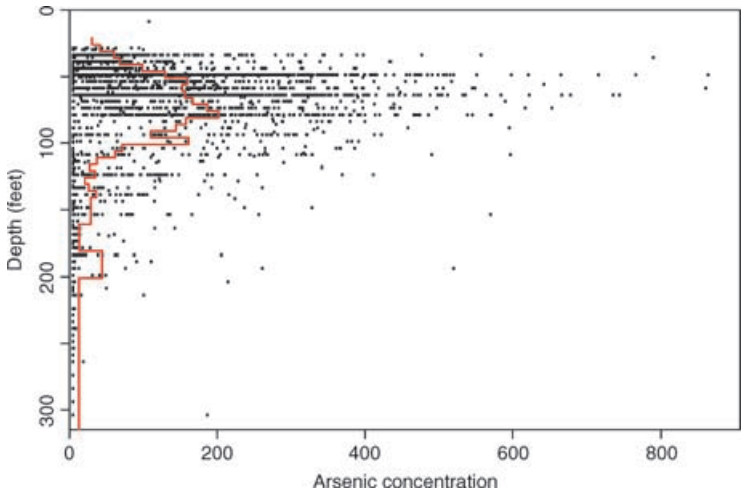

Fig. 2. Arsenic concentrations (μg/L) and depths of the 4,827 wells mapped in Fig. 1. The red line shows average arsenic concentration as a function of depth. (Although it plays the role of an explanatory variable, depth is shown on the *y*-axis because of its geographic interpretation.)

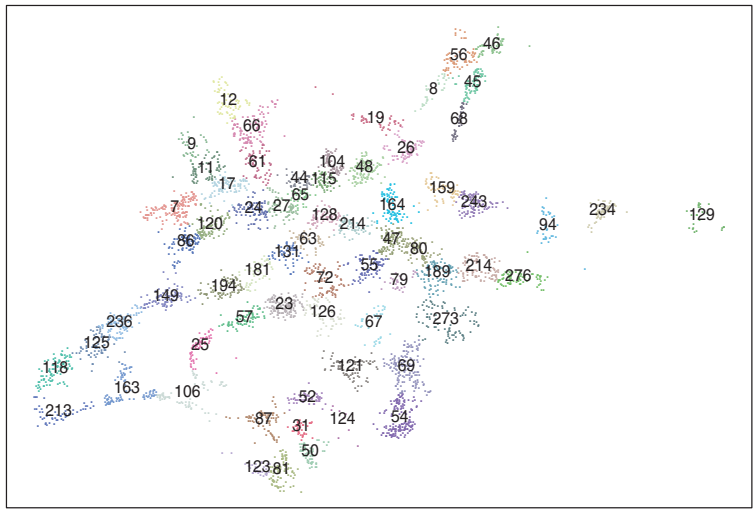

Fig. 3. The wells in Fig. 1, divided into 66 clusters based on the *k*-means clustering algorithm. The identifying numbers are the average arsenic levels of the wells in each cluster. The colors have no meaning and are simply to identify the separate clusters. The clusters were constructed using only the locations of the wells; depth and arsenic levels were *not* used in the clustering.

Some people put in the extra effort and cost to drill deeper wells (beyond 300 ft, a hydraulic pump is required that is not locally available), and these have less arsenic, on average, as shown by Fig. 2. However, it is also clear that depth alone is no guarantee of low-arsenic water. At the other extreme, the very shallow wells appear to be relatively safe, although, once again, with much variation.

The next step is to combine the spatial and depth information, which we do by dividing the area into several geographically compact clusters and then plotting arsenic level and depth for the wells in each cluster separately. The area under study includes about 60 villages, which is the level at which local decisions are made. We defined local groups of wells using the *k*-means clustering algorithm[9] as implemented in R,[10] which did a good job of grouping the wells into 66 well-defined spatial clusters, as shown in Fig. 3. Given a specified number of clusters, the algorithm minimizes the average squared distance within clusters—that is, the average squared distance of points from cluster centroids. (We do not use depth or arsenic information in computing the clusters, since our purpose in the clustering is to define localized areas for the later analysis.) The particular clustering algorithm chosen is not crucial, since our purpose here is simply to divide the wells into local groups (see, e.g., References 11 and 12 for reviews of clustering methods). Each cluster in Fig. 3 is labeled with the average arsenic concentration measured in its wells.

We chose the number of clusters with the goal of having compact clusters roughly the size of villages. In general, the choice of cluster size balances two concerns: (1) smaller clusters are more compact, and thus it is more reasonable to expect stationarity in the arsenic levels within any cluster; but (2) larger sample sizes allow more reliable inferences about arsenic levels within each cluster. Most desirable would be a model that allows the relation between arsenic concentration and depth to vary spatially in a smooth way—but before going to this effort it makes sense to perform exploratory analysis such as done in this article.

In any case, it would not make sense to attempt to "estimate" the number of clusters in the data since, fundamentally, each well is its own cluster, and the choice of number of clusters depends on a balancing of inferential goals. In our analysis, the clusters are created only to allow us to better understand and use patterns in the relation between arsenic level and well depth.

Fig. 4 shows scatterplots of arsenic level (indicated by colors) as a function of depth and year of installation, for the wells in each of the 66 clusters. The plots show that most of the wells were installed after 1995, and some of the deepest wells have been installed very recently. However, conditional on depth, the year of installation does not appear to be informative in predicting arsenic level.

A careful study of the relation between arsenic level and depth shows several patterns, only some of which were apparent in the original map (Fig. 1) and scatterplot (Fig. 2).

Most obviously, the wells in some clusters are consistently low in arsenic, whereas the wells in other clusters are all high. For simplicity we shall refer to wells as "safe" if their arsenic level is below 50 μg/l. The depths of the wells vary dramatically between clusters, and this may explain somewhat the spatial variation in arsenic levels. For example, compare Clusters 9 and 234, which are near the extremes of average arsenic levels (recall that the label of the cluster is the average arsenic concentration of its wells). The wells in Cluster 9 are all safe (as indicated by the blue and green dots, their arsenic levels are all below 50 μg/l), and all are at least 100 ft deep. Conversely, the wells in Cluster 234 are all dangerous and all less than 100 ft deep. This complete confounding makes it impossible, without further information, to know how to attribute the difference in arsenic levels between the two clusters to geography and well depths.

Some clusters show a dramatic relation between arsenic concentration and depth. In Cluster 86, for example, all the wells deeper than 100 ft are safe, and almost all the shallower wells are dangerous. Cluster 120 shows a similar threshold at a depth of 70 ft. Similar patterns appear throughout; for example, Cluster 7 may have a threshold around 40 ft, Cluster 11 between 70 and 100 ft, and so on to Cluster 243, with an apparent threshold between 110 and 170 ft.

The patterns are not consistent everywhere, however. For example, the wells in Clusters 80 and 118 exhibit a range of arsenic levels at all depths. Most strikingly, Cluster 46 shows a reverse pattern: here, all the shallow wells are safe, and most of the deep wells are dangerous. At this point it is useful to cross-reference with the maps: Figs. 1 and 3 show that Cluster 46 is separated into two geographic subregions, with the safe (and shallow) wells to the northeast and the dangerous (and deep) wells to the southeast.

Fig. 4. Plots of arsenic level (indicated by color: blue, green, orange, red, and black, as in Fig. 1) as a function of depth and year of installation of well. (The relatively few wells installed before 1980 are assigned dates of 1980 in these graphs.) Data are displayed separately for each of the 66 spatial clusters (see Fig. 3). Each cluster is labeled by the average arsenic level of the wells in the cluster. The horizontal lines indicate estimated safe-depth thresholds (or lower bounds, where thresholds cannot be estimated), as described in Section 4.1 and mapped in Fig. 13.

To summarize our exploratory analysis, the wells between 50 and 100 ft deep have, on average, the most arsenic. In many clusters there seems to be a safe depth, typically between 100 and 200 ft, below which the water is low in arsenic. However, in other places even the deepest wells are dangerous. Similarly, the shallowest wells are consistently safe in some areas but not in others.

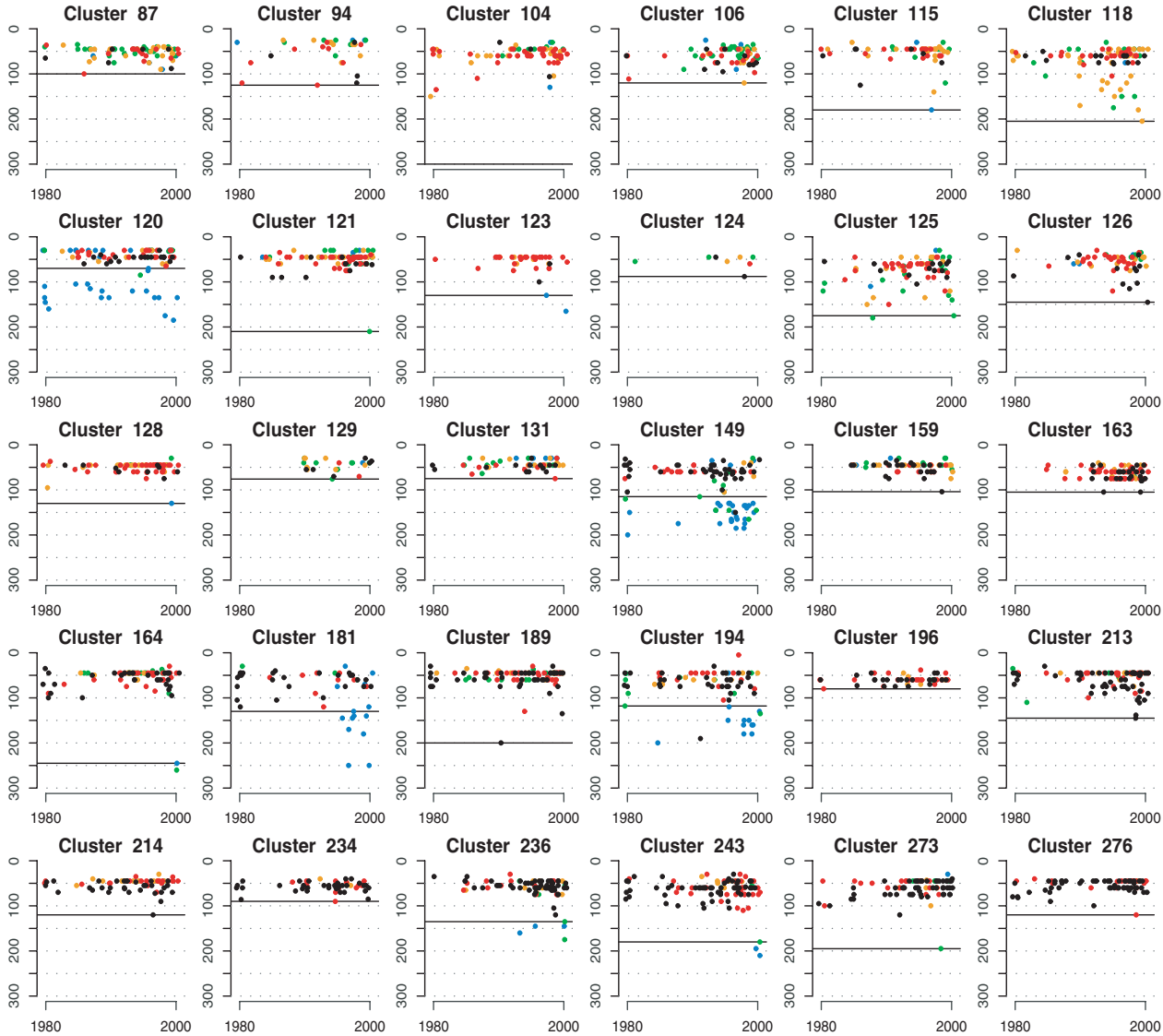

Fig. 4. Continued.

We now move to more focused analyses to answer the following applied questions: (1) How effective would be a strategy of encouraging users of dangerous wells to switch to nearby wells that are low in arsenic? (2) Where should new wells be located to maximize the number of people served? (3) How deep should the wells be drilled to ensure they will be safe? We can answer the first two questions using direct data manipulations; in attempting to answer the third question, we shall augment our exploratory analysis with some modeling and estimation of probabilities.

3. RECOMMENDATIONS FOR SWITCHING WELLS

As can be seen from Fig. 1, low- and high-arsenic wells are mixed throughout the region, and there appears to be no simple spatial pattern or rule that would allow one to reliably identify a well as low or high in arsenic without actually measuring it. This is unfortunate because accurately measuring a well's arsenic level requires equipment that is not readily available in rural Bangladesh. The cost of the field test is \$0.50, but this does not include the salary of the

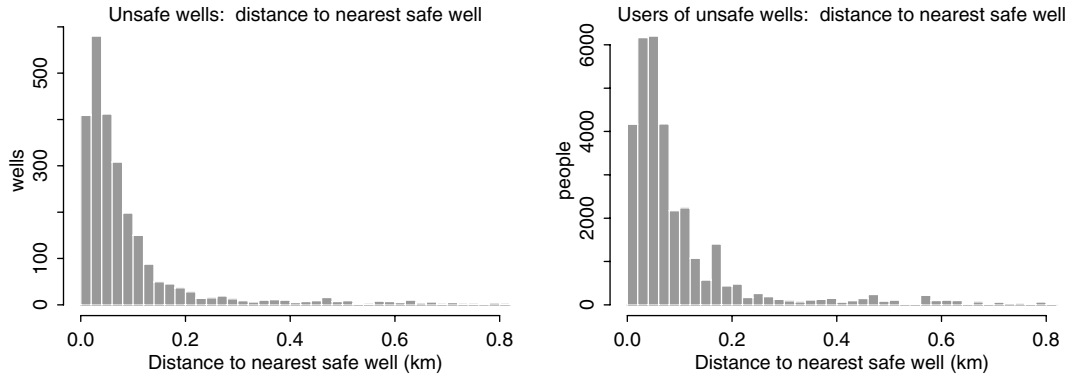

Fig. 5. Distribution of distance from the nearest "safe" well (arsenic concentration less than 50 μg/l), for (a) unsafe wells and (b) people drinking from unsafe wells (i.e., weighting each well by the number of users). These graphs exclude the 48% of wells that are already safe.

tester, the infrastructure needed for training, and the mapping that would be needed to make the most of the data. (By way of reference, it costs about $1 per foot to install a tube well, and the per-capita GNP of Bangladesh is about $400). The good news, however, is that most of the people in this area—including those currently drinking water high in arsenic—live close to a low-arsenic well.

Fig. 5 summarizes the distance to the nearest safe well for users of unsafe wells. Almost all the unsafe wells—and almost all the current users of unsafe wells—are within 200 m of an existing safe well.

Given this information, a reasonable short-term arsenic reduction strategy might be to recommend that people who are currently drinking from high-arsenic wells switch to nearby low-arsenic wells. Preliminary results from a survey of local residents suggest that about three-fourth of the people who are drinking from unsafe wells will be willing to walk to obtain safe water from a nearby well, and that owners of safe wells will generally be willing to share their water with neighbors (A. Pfaff, private communication).

For the region under study, the well-switching strategy is feasible since the arsenic levels in all the wells were measured. We began by creating a list, for each of the dangerous wells (those with arsenic measurements exceeding 50 μg/l), of the locations of the 10 nearest wells, along with their arsenic levels. Investigators took these lists into the field to guide people in well switching.[3] To estimate the effectiveness of the well-switching strategy, we compute the expected reduction in total arsenic exposure under various assumptions about switching behavior.

Suppose that people drinking from wells with arsenic levels higher than X were to switch to the nearest safe well, if there is a safe well less than D m away, or, if there are no safe nearby wells, to the lowest-arsenic well within a distance of D. Fig. 6a shows the proportion of people who would be switching under this recommendation, and Fig. 6b displays the average arsenic exposure for all the residents in the area (not just the residents who switch). Both graphs are plotted as a function of D, for several values of X. For each plot, the curves start at $D = 0$ with zero people switching and the current mean level of 97 μg/l.

With a simple well-switching strategy, a few wells will be overburdened—the isolated low-arsenic wells that are in high-arsenic areas. To avoid overusing any well, we assume in Fig. 6 that users from no more than 10 "dangerous" wells are allowed to switch to any existing "safe" well. When a well is full-up, our algorithm switches users to the nearest safe well within a distance D that is still free.

We also evaluate the effects of the recommendation if it is only partially followed: Fig. 7 displays the proportion of people who switch and the average arsenic exposure from well water, assuming that only half the residents switch (which we believe is a conservative assumption, given our preliminary survey findings). Again, these results are shown as a function of the arsenic threshold X and the distance threshold D, with the assumption that no more than 10 existing wells are referred to any single "safe" well.

Based on our formal survey and informal conversations with local residents, we think it is reasonable to

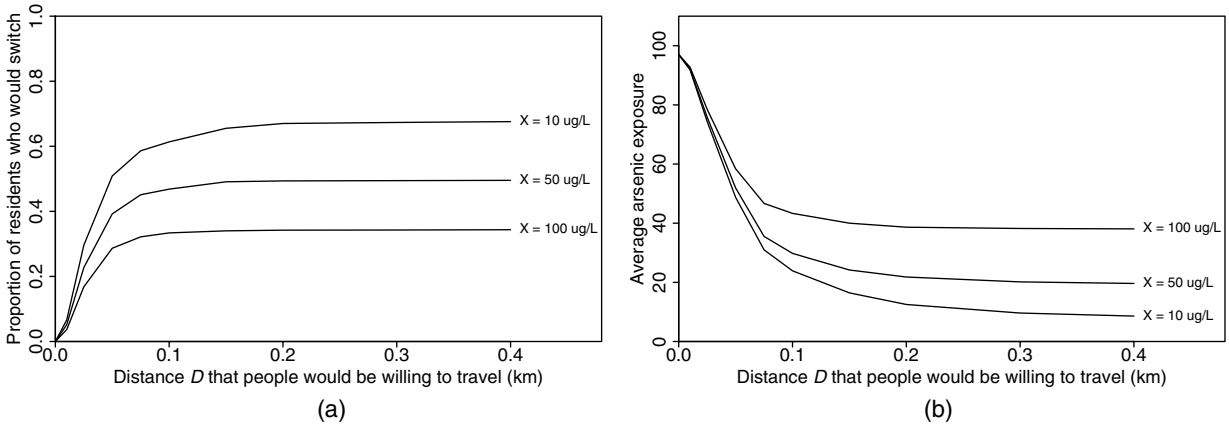

Fig. 6. (a) Proportion of residents of the studied area who would switch wells and (b) estimated average arsenic exposure (μg/L) of all the area residents if the following recommendation is followed: all users with arsenic levels exceeding X should switch to the nearest well with arsenic level below X, if such a well is less than D m away, or else to the safest well within a distance of D. Switching is restricted so that no well is used by the previous users of more than 10 other wells.

suppose that people in the area will be willing to walk up to 200 m for safe water.[7] By means of flow meters and interviews with villagers carrying water from six newly-installed community wells, a study was made of the extent to which these were used during one year. The results were compared with household and well data obtained during a previous survey in the same area. Many women walked hundreds of meters each day to fetch water from the wells. On average, 2,200 l were hand-pumped daily from each community well, regardless of the season.

Assuming 50 μg/l as the safe-water threshold, Fig. 7 shows that this recommendation would result in 26% of the local residents switching, and a new average arsenic exposure of 60 μg/l (a 38% reduction in total exposure compared to the existing mean level).

4. RECOMMENDATIONS FOR DRILLING NEW WELLS

As discussed in Section 3, switching wells is a cheap, immediate, and effective method that could realistically reduce arsenic exposure by nearly 40% at the cost of having one-quarter of the local residents having to travel distances of less than 200 m for drinking water. However, this does not help the people who live more than 200 m from a safe well, and so we would like to supplement the switching strategy with the drilling of some new wells. In addition, people who are currently drinking from high-arsenic wells would generally like to use their own safe wells or community wells, rather than a neighbor's private well. We are thus led to two decision questions: (1) where to drill new wells to serve the

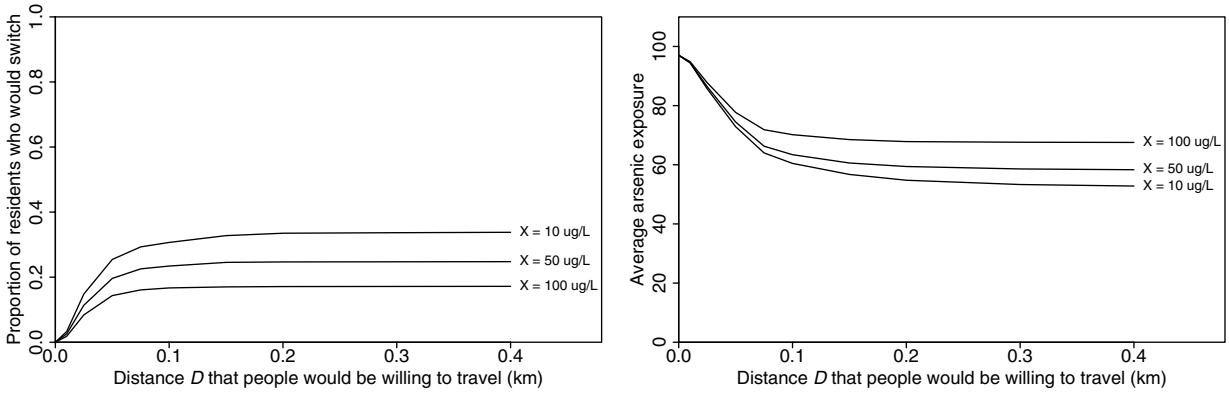

Fig. 7. Replication of Fig. 6, but under the assumption that only half the people follow the recommendation to switch.

immediate needs of people who are currently drinking high-arsenic water, and (2) how deep to drill new wells at these and other locations.

4.1. Where to Drill

We would like to drill new wells so as to serve the maximum number of people currently drinking high-arsenic water. If we assume that people will walk up to 200 m to a well, it is a computational problem to identify the optimum locations. We solve the problem using a stepwise optimization algorithm, first determining the best place to put the first well, then the best location for the second well, and so forth. This heuristic approach will not in general find the optimal locations for a set of n wells but is simpler than a general optimization.[13] At each step of our algorithm, there is an infinite range of locations to put any given well, but to figure out the optimal location we need only evaluate at a finite set of points corresponding to the intersections of the arcs of circles centered at each well (see Fig. 8). As the number of wells n in a data set grows larger, with the density of wells per square kilometer held constant, the number of such points of intersection grows linearly with the number of wells, so this is a feasible computation even for large data sets. At each point of intersection, we can quickly evaluate the number of people within 200 m who are currently drinking water with arsenic concentrations above 50 μg/l.

Fig. 9a shows where to drill 30 new wells sequentially to maximize the number of people within 200 m who are currently drinking from high-arsenic wells.

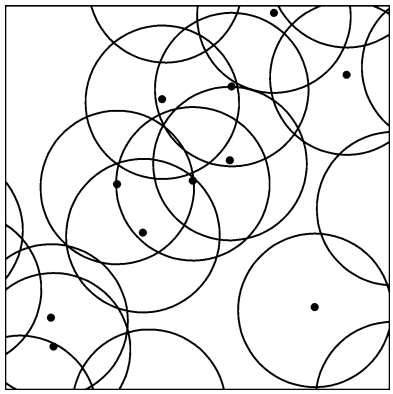

Fig. 8. Diagram illustrating the search for the optimal location for a new well. The circles are centered at each existing well with radii 200 m, the assumed maximal distance a person will travel to get water. We need only evaluate the points at the intersections of the circles.

Fig. 9b shows where to drill the new wells, assuming that half the users of high-arsenic wells have already switched, as described in Section 3. We can then evaluate the effectiveness of the new-well-drilling strategy on average arsenic exposure.

What if these new wells are drilled with, let us assume, arsenic levels of 5 μg/l? Fig. 10 shows the proportion of people affected and the average reduction in arsenic exposures. A glance at the y-axes of the figures shows that the proportion of people affected and the average reduction is impressive: considering there are about 5,000 existing wells, these numbers represent good value from installing no more than 30 new wells.

Fig. 11 shows the consequences if half the users of high-arsenic wells have already switched to nearby wells with lower arsenic concentration, as described in Section 3. The benefits of the community wells is slightly less than if no users had switched (compare to Fig. 10) but is still a substantial benefit for drilling only 30 wells.

4.2. How Deep to Drill

If new wells are to be drilled, it is crucial to have an idea of how deep to drill them. As indicated in Figs. 2 and 4, shallow or deep wells appear safe in some areas but not others. We cannot hope for certainty, but we would like to give better recommendations than simply, "Drill as deep as necessary." To this end, we attempt to estimate a "safe-depth" threshold—a depth below which the arsenic level will be less than 50 μg/l—in each of the 66 clusters into which we have divided our data.

In estimating the safe-depth thresholds, we do not set up a full statistical model of arsenic levels, but we construct an inferential procedure based on the patterns we see in the data in Fig. 4: most notably, that in many clusters there appears to be a sharp threshold, below which all the wells are safe—blue and green dots in the graphs. Sometimes, however, there is a single exception—a dangerous well mixed with the safe deep wells, as in Cluster 194—which is consistent with there being an imperfect safe-depth threshold or with a depth that was reported in error.

In the next section, we present two statistical techniques for estimating such safe-depth thresholds: a *search algorithm* that we first used to identify thresholds, and a *matching algorithm* that yields similar results. We developed the search algorithm with the decision problem in mind, whereas the matching algorithm is closer in spirit to statistical hypothesis testing.

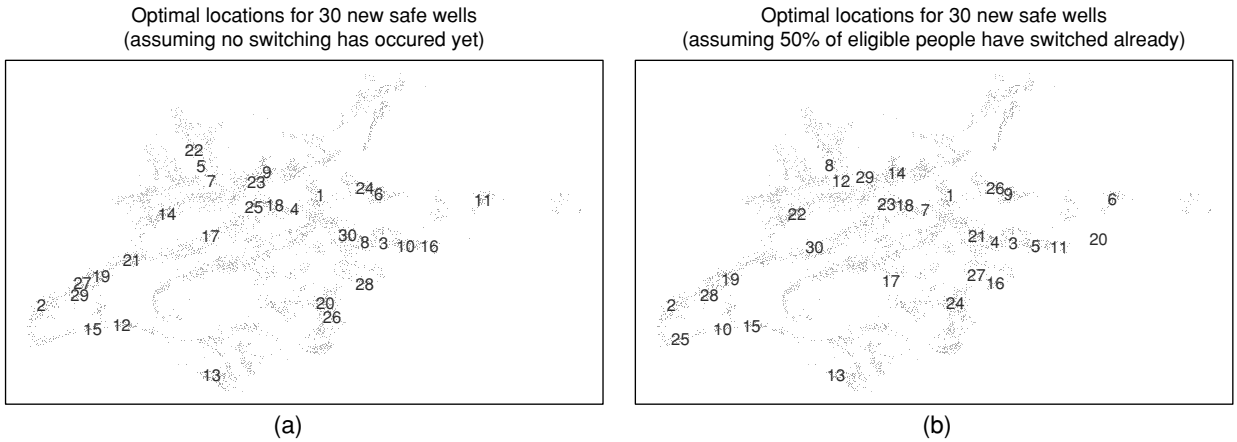

Fig. 9. Optimal locations for 30 new wells to be drilled in sequence, with the goal of maximizing the number of people who can be within 200 m of a safe well (arsenic less than 50 μg/l), under two different assumptions: (a) assuming that no one has switched wells yet; (b) assuming that half the users of high-arsenic wells have already switched to the nearest safe well, if it is within 200 m. In each map, the locations are numbered in decreasing order of priority. Light and dark dots show existing safe and unsafe wells, respectively.

4.2.1. Estimating Safe-Depth Thresholds

To allow for the possibility of aberrations or outliers, we construct the following *search algorithm* to estimate safe-depth thresholds D. For each cluster of wells, we start with the deepest wells (the bottom of each of the graphs in Fig. 4) and move up until we identify the deepest unsafe well (i.e., with arsenic level exceeding 50 μg/l). We label its depth as U_1 and denote the depth of the shallowest safe well that is (strictly) deeper than U_1—if such a well exists in this cluster—as S_1. We then move up to the next-deepest unsafe well, labeling its depth as U_2, and correspondingly look for the shallowest safe well that is (strictly) deeper than U_2, denoting its depth as S_2. The left panels of Fig. 12 illustrate the application to two of the 66 clusters. In Cluster 189, the deepest unsafe well is at $U_1 = 200$ ft but we cannot find any safe well deeper than 200 ft; the next-deepest unsafe well (also indicated by a black dot) is at $U_2 = 135$ ft but again no safe well exists below it. Turning to Cluster 194, $U_1 = 190$ ft, and here we can find a deeper safe well (a blue dot on the graph) at $S_1 = 200$ ft, the next-deepest unsafe wells (two dots, one red and the other black) are at $U_2 = 105$ ft, and the shallowest safe well below them is at $S_2 = 118$ ft (a green dot). We plot the wells identified so far in the central panels of the figure, with x-axis corresponding to the distance of the wells from the cluster centroid. In these central figures, the bottom left arrows indicate the N–S/E–W directions of the wells relative to the cluster centroid. Information on distance and direction is not used by the search algorithm so far implemented. However, when putting some results into practice, such information

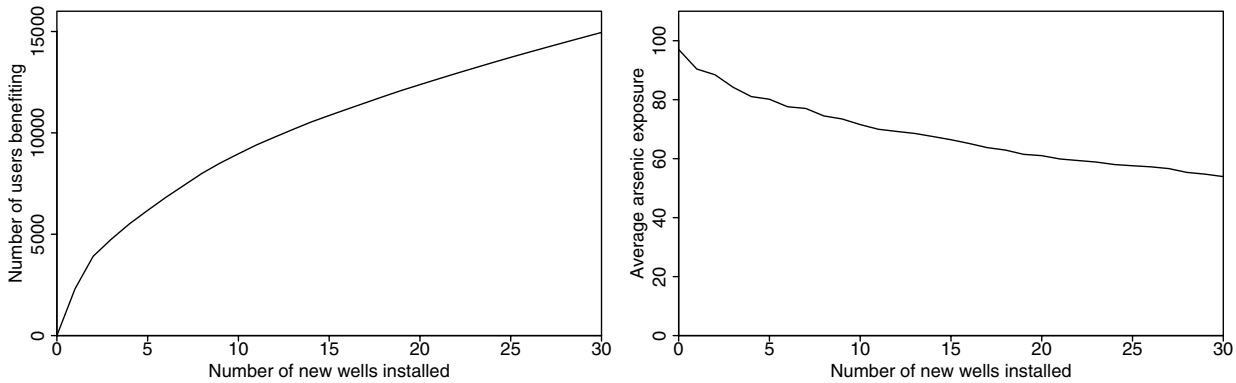

Fig. 10. Consequences of adding up to 30 new safe wells at the locations indicated in Fig. 9a. The graphs show the number of people who would benefit and the average arsenic (μg/L) levels among all 55,000 people in the area (including those not affected by the new wells).

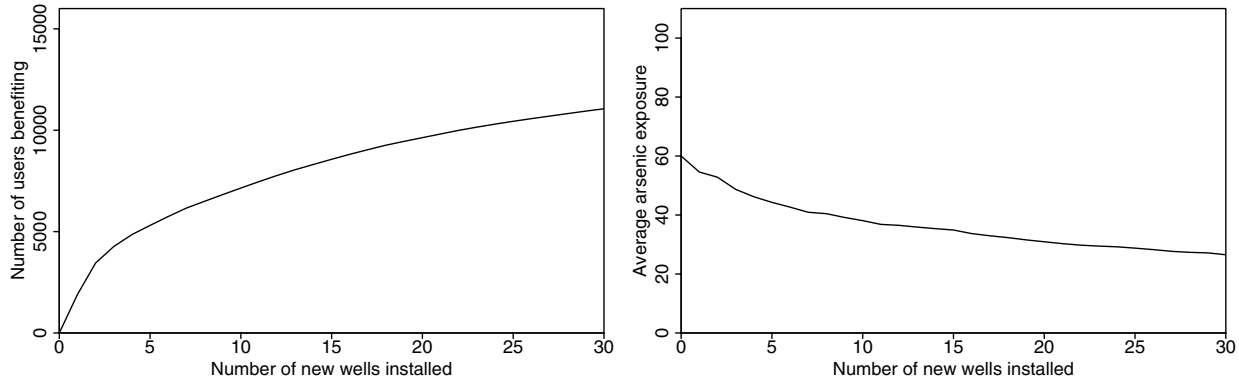

Fig. 11. Replication of Fig. 10, but assuming that the new community wells are assigned in the locations shown in Fig. 9b, and under the assumption that half the people have already switched wells.

may help in assessing well location with respect to the cluster: if the well is for some reason considered nonrepresentative of the area we are about to drill, it can be eventually discarded from the search procedure.

Having identified U_1 and U_2, we consider two possible threshold configurations, which are defined: (1) as any depth within $(U_1, S_1]$ and (2) as any depth within the $(U_2, S_2]$ interval (in practice, we will set D at S_1 or S_2—the shallow endpoint of the finally selected interval). We evaluate these choices on the basis of the associated probability that a well drilled deeper than the estimated threshold actually has arsenic concentration less than 50 μg/l.

To estimate such a probability, we adopt an approximate Bayesian approach.[14] Our prior information consists of the evidence (see Fig. 2) that there are on average lower arsenic levels both in the shallowest and deepest strata. In addition, related discussion in Reference 4 suggests synthesizing a general pattern by dividing the range of well depths into three strata: $\leq$30, 30–100, and >100 ft, which have proportions of safe wells in our data of 0.76, 0.35, and 0.83, respectively. From these premises, we set up a Beta (α_j, β_j) prior distribution for the probability θ_j that a well sunk into stratum j is safe, with $j = 1, 2, 3$ indexing (in the same order) our depth strata, and (α_j, β_j) set at (3, 2), (1, 1), (3, 2), respectively. These hyperparameters imply prior probabilities of a well being safe as 3/5, 1/2, and 3/5 in stratum 1, 2, and 3, thus roughly reflecting the general pattern in the area being studied, with low degrees of freedom so that data from a reasonable number of wells will dominate the inference in any cluster.

The data we have at hand are for each cluster the number of safe and total wells sunk into stratum j and below U_k—that is, y_{jk} and n_{jk}, with sums $y_k = \sum_j y_{jk}$ and $n_k = \sum_j n_{jk}$.

To choose one threshold configuration $(U_k, S_k]$ out of the two possibilities $k = 1, 2$, we calculate the posterior probability that a new well drilled deeper than U_k in the given cluster is safe. To this end, we first calculate the predictive probability that a new well $\tilde{y}$ at depth $\tilde{d} U_k$ is safe: $\Pr(\tilde{y} = 1 \mid \tilde{d} > U_k) = \int_{\boldsymbol{\theta}} \Pr(\tilde{y} = 1 \mid \tilde{d} > U_k, \boldsymbol{\theta}) \Pr(\tilde{d} > U_k \mid \boldsymbol{\theta}) p(\boldsymbol{\theta}) d\boldsymbol{\theta}$. We suppose that d can (uniformly) assume only values corresponding to depths at which the surveyed wells were actually drilled. The reasoning underneath such assumption is that so far we have not tested whether drilling at other depths than those of already existing wells is feasible: it might not be so because of some (still unknown) limiting geological characteristics. Hence $\Pr(\tilde{y} = 1 \mid \tilde{d} > U_k) = \frac{1}{n_k} \sum_{d_y : \exists y : d_y > U_k} \int_{\boldsymbol{\theta}} \Pr(\tilde{y} = 1 \mid \tilde{d} = d_y, \boldsymbol{\theta}) p(\boldsymbol{\theta}) d\boldsymbol{\theta}$, and, having modeled y having d within stratum j as a Bernoulli (θ_j), we obtain

$$\Pr(\tilde{y} = 1 \mid \tilde{d} > U_k) = \sum_j w_{jk} \mathrm{E}(\theta) = \sum_j w_{jk} \frac{\alpha_j}{\alpha_j + \beta_j}, \tag{1}$$

with $w_{jk} = n_{jk}/n_k$. At this point, we can specify a Beta (α_k, β_k) prior distribution for the probability θ_k that a well below U_k is safe, with α_k and β_k solving $\Pr(\tilde{y} = 1 \mid \tilde{d} > U_k) = \alpha_k/(\alpha_k + \beta_k)$. (We fix $\alpha_k = 1$ if expression (1) is less than 1/2, and set $\alpha_k = 2$ otherwise: such a choice is made in order that prior opinion has more weight if more of the n_k wells are in Strata 1 and 3 rather than Stratum 2; that was implicitly assumed also for the α_j's defined above.)

We are now able, after setting a binomial $(\theta_k | n_k)$ model for y_k, to update Equation (1) to the posterior probability $\Pr(\tilde{y} = 1 \mid \tilde{d} > U_k, y_k)$, which yields

$$p_k = \frac{\alpha_k + y_k}{\alpha_k + \beta_k + n_k}, \tag{2}$$

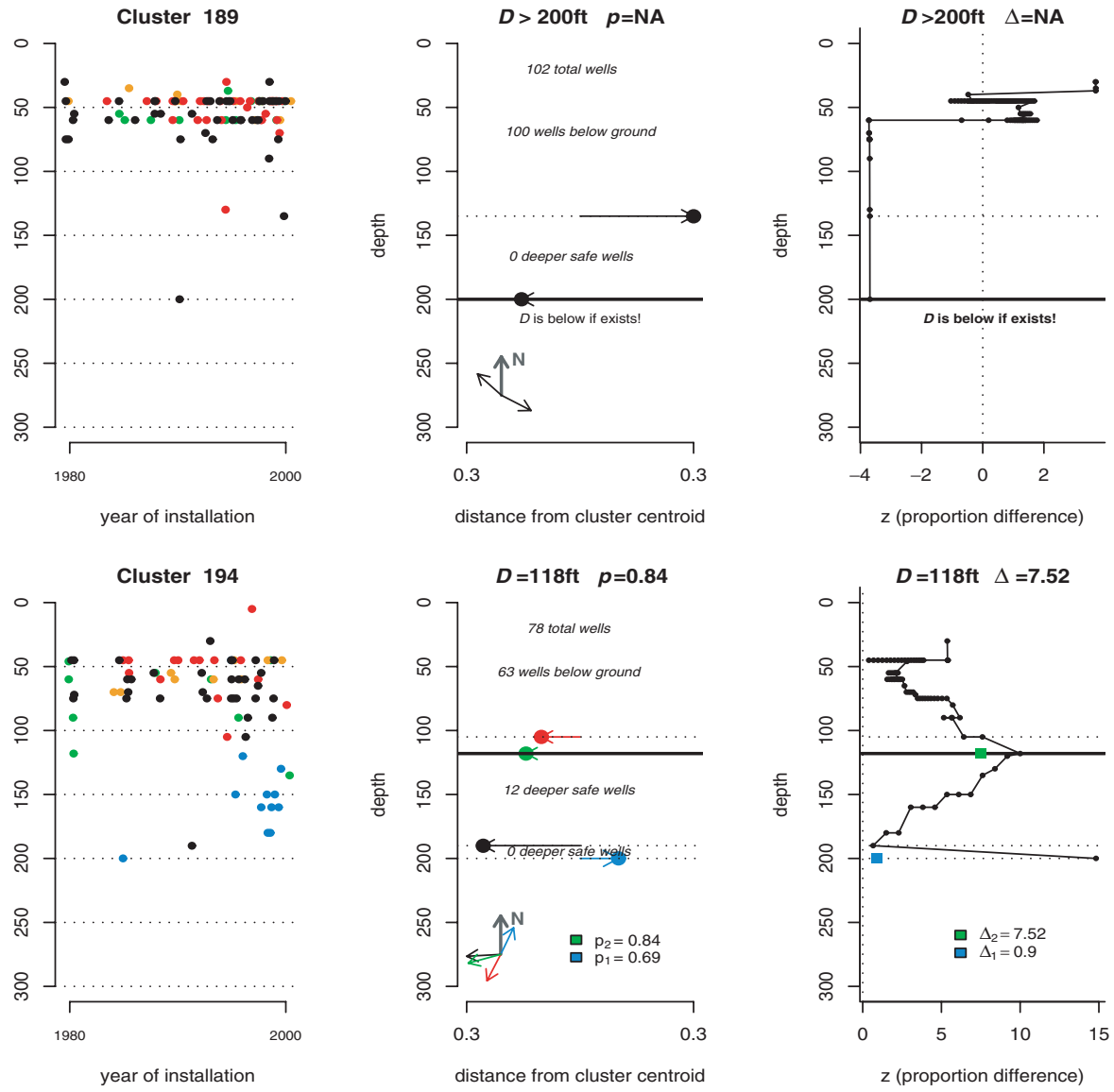

Fig. 12. Diagrams showing how we estimate "safe-depth" thresholds for each of the 66 clusters in the data set. The central plots show the search: the dots represent the deepest wells, the horizontal lines show the estimated thresholds (the bottom left arrows indicate the N–S/E–W directions of the wells with respect to the cluster centroid, and the distances from the center are the abscissas of the dots). The rightmost plots represent the *matching algorithm* explained in the text. In Cluster 189 no safe-depth threshold can be estimated, so we assign a lower bound. In Cluster 194 the threshold is set at S_2, the depth of the first safe well below the second deepest unsafe well, since $p_2 > p_1$.

a well-known result of the beta-binomial model (see, for instance, Reference 20, ch. 2). Finally, we estimate $D = S_1$ if $p_1 \geq p_2$, otherwise $D = S_2$.

The search algorithm can be described formally in terms of steps **s** and exit points **e**; for each cluster,

s0—Search for U_1: if U_1 does not exist then go to **e1**, otherwise go to **s1**;

s1—Search for S_1: if S_1 does not exist (i.e., the deepest well is unsafe) then go to **s2**, otherwise go to **s3**;

s2—Search for U_2: if does not exist then go to **e1**, otherwise if $S_2 < U_1$ exists then $D = S_2$ else go to **e2**;

s3—Search for U_2: if U_2 does not exist then S_2 is the shallowest depth else search for S_2 ($\leq S_1$); go to **e3**;

e1—D is the shallowest depth;
e2—D is censored (i.e., is estimated to be deeper than U_1 but that is all);
e3—$D = S_1$ if $p_1 \geq p_2$, otherwise $D = S_2$.

Turning to our example, Cluster 189 is a case of censored safe threshold (in fact, $y_2 = 0$); in Cluster 194 the safe threshold has been set at $S_2 = 118$ ft as $p_2 = 0.84$ exceeds $p_1 = 0.69$.

The rightmost graphs of Fig. 12 show the application of another method, which we call the *matching algorithm*, to estimate D. Sorting the n well depths of each cluster from the deepest, d_1, to the shallowest, d_n, we calculate for each i, $i = 1, \ldots, n$, the normalized difference z_i between the proportions of safe wells below and above d_i given by $z_i = (\bar{y}_i - \bar{y}_i^c)/s_i$, where $\bar{y}_i = y_i/n_i$ and y_i and n_i denote the number of safe and total wells counted from 1 to i (with depth $\geq d_i$). Similarly, $\bar{y}_i^c = y_i^c/n_i^c$ with y_i^c and n_i^c indicating the complementary quantities, that is the number of safe and total wells counted from $i+1$ to n (with depth $\leq d_i$). Moreover, $s_i = \sqrt{\bar{y}_i(1-\bar{y}_i)/n_i + y_i^c(1-\bar{y}_i^c)/n_i^c}$ is an estimate of the standard error of the difference.

In the rightmost panels of Fig. 12, z for each cluster is represented by a black line. Starting from the bottom d_1, the line drifts to the left or to the right, each time that an unsafe well or a safe well, respectively, is found as the computation moves toward shallower depths. In Cluster 189, $U_1 = d_1$ so that $z_1 < 0$, correspondingly; in Cluster 194, z sharply turns to the left (i.e., decreases) both at U_1 and U_2.

With the same reasoning as before, we consider the two depths, S_1 and S_2 (if they exist) strictly below U_1 and U_2 as candidates for D. But this time we use as decision criterion the posterior estimate Δ_k of the normalized difference of proportions of safe wells below and above U_k—that is, $(\theta_k - \theta_k^c)/\sqrt{\text{var}(\theta_k - \theta_k^c)}$, for $k = 1, 2$. We do not adopt a full model but resort to an empirical Bayesian approach to obtain Δ_k. In particular, we estimate it as

$$\Delta_k = \frac{p_k - p_k^c}{s_k^p}, \quad k = 1, 2, \tag{3}$$

where p_k and p_k^c are the posterior probabilities that a well below and, respectively, above U_k is safe. Equation (2) gives p_k while p_k^c is computed as p_k but by a "mirror" procedure (i.e., considering in Equations (1) and (2) the wells with depths less than U_k). Moreover, s_k^p is produced by substituting p_k and p_k^c to $\bar{y}_k$ and $\bar{y}_k^c$ in the formula above for s_i with $i = k$. Again, we estimate the safe-depth threshold as S_1 if $\Delta_1 \geq \Delta_2$ or S_2 otherwise.

In the formal description of the search algorithm above, only **e3** item has to be changed (by substituting p with Δ). After that, the matching algorithm yields the same estimates for D over all the clusters. In particular, turning to Fig. 12, D is censored in Cluster 189 while $\Delta_2 > \Delta_1$ then $D = S_2$ in Cluster 194 (Δ_k, $k = 1, 2$, is plotted as a square point at the respective S_k depth).

We have displayed the estimated safe-depth thresholds (or bounds) in Fig. 4, overlain with the well data, for each of the 66 clusters. Fig. 13 shows how the safe-depth thresholds vary spatially. The different shades of green on the map indicate where safe depths are estimated to be less than 100, 100–150, 150–200, and deeper than 200 ft, and the different shades of brown show where the safe-depth estimates are censored. This graph shows some patterns that were not apparent in the map of arsenic concentrations (see Fig. 1)—the areas with lowest safe-depth thresholds do not always coincide with the areas with the safest wells.

The relation between arsenic, depth, and geographic location is complex, as can be seen by comparing the map of arsenic levels in Fig. 1 with the estimated thresholds in Fig. 4. For example, consider Cluster 120 (in the northwest part of the sampled area): it has one of the higher average arsenic levels, but its estimated safe-depth threshold is only 70 ft. By comparison, the wells in nearby Cluster 9 have very low arsenic, but perhaps only because they are almost all below 100 ft deep. In contrast, Cluster 118 in the southwest provides evidence that drilling below 100 ft is no guarantee of low arsenic. Finally, when considering high-arsenic areas, it might be useful to distinguish between clusters such as 276 (in the east), where no wells have been drilled below 150 ft, and nearby clusters such as 189, where even very deep wells have been tried but with no success.

4.2.2. Estimating the Probability that a Deep Well Is Safe

Before using the estimated safe-depth thresholds in decision making, it is useful to have some sense of our confidence in them. To this end we estimate, in any cluster, the probability p_D that a well drilled deeper than the estimated threshold D actually has arsenic concentration less than 50 μg/l, as max (p_1, p_2), following the estimation procedure described in Section 4.2.1. The estimated probability for each cluster is shown in Fig. 13.

We then perform a cross-validation, removing each well from the data set and reestimating the

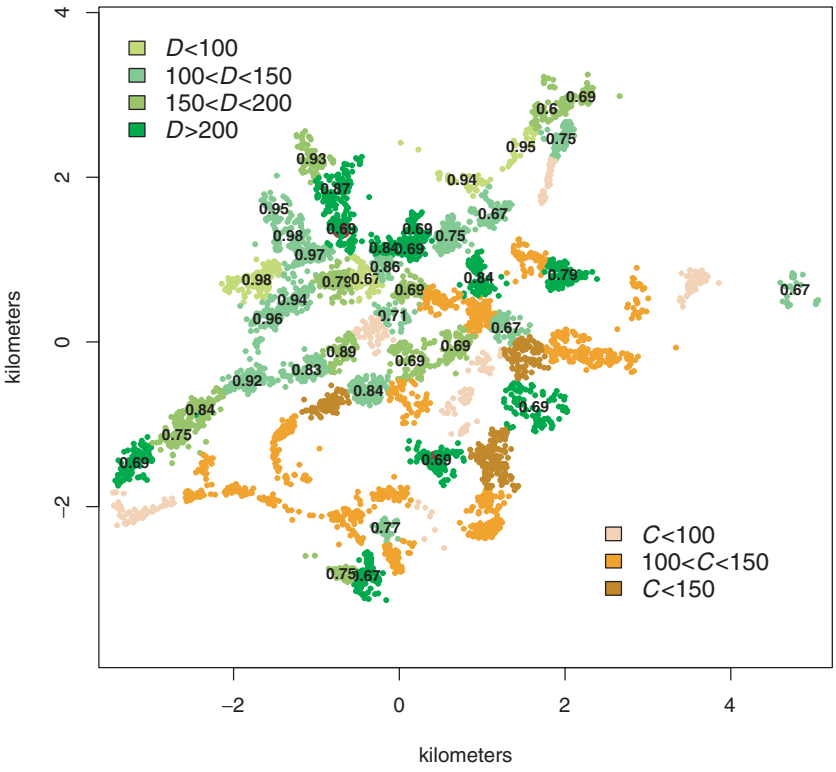

Fig. 13. Estimated safe-depth thresholds, and estimated probabilities that a new well will be safe, if it is deeper than the estimated safe depth, within each of the 66 clusters. Estimated depth thresholds D are indicated by different colors of green. For clusters with bounds on depth thresholds, the censored bounds C are indicated by different colors of brown. Within each cluster with an estimated safe depth is shown the estimated probability that a well will be safe if it is drilled deeper than that estimated threshold.

safe-depth threshold for its cluster (for simplicity, we keep the regions fixed). We then consider the excluded well: Is it deeper than the estimated safe-depth threshold, and is it actually safe? Counting these for each well in turn yields a cross-validated probability for each cluster. We create a calibration curve by binning these estimated probabilities into nine intervals (from (0.6, 0.7] to (0.975, 1], with wider ranges in the sparser lower-probability classes) and calculating the average empirical probability within each bin. Fig. 14 plots these as a function of the estimated probabilities (the midpoints of the above intervals, that is 0.65 up to 0.9825).

So, how deep should new wells be drilled? Fig. 13 gives minimum depths, and in some places we are quite confident that wells deeper than this will be safe. But in areas where we only have a minimum for the safe depth, or where the estimated probability of encountering a dangerous well is high, community wells can be drilled under expert supervision and the sediment monitored until, for geological reasons, it is plausible that the water will be safe, and then it can be tested. In Araihazar, it has been necessary to drill between 200 and 500 ft in some places to obtain safe water.[7]

5. DISCUSSION

5.1. Further Study of Arsenic in Bangladesh

We have analyzed data on nearly 5,000 wells to make some recommendations about switching wells

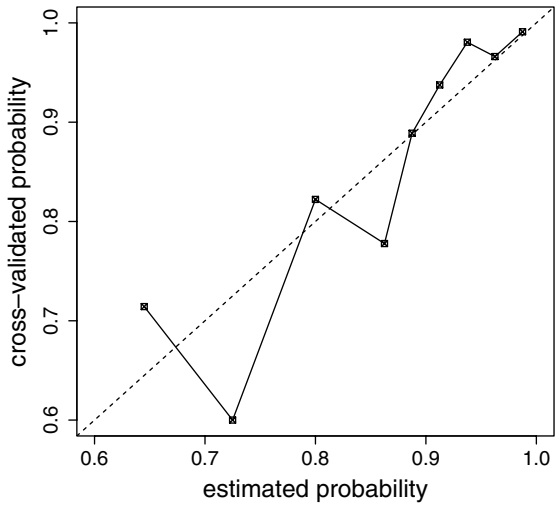

Fig. 14. Calibration of the probability estimates for the safe-depth thresholds, using cross-validation. The 45° line shows the ideal of perfect calibration.

and drilling new wells in a region in Araihazar, Bangladesh. What can be done about the rest of the country? A national program supported by the World Bank claims that all wells in approximately half the country (about 5 million wells) have been tested. The actual response of villagers to this information, communicated by painting the spout of a well red or green (with paint that remains visible for only one year) is largely unknown. We hope this article will lead to a closer examination of patterns in the existing data to direct mitigation activities at the village scale.

In the arsenic study, one must balance several costs: most obviously, the health effects of arsenic exposure, but also the inconvenience of regularly carrying drinking water a distance of perhaps several hundred meters, and the financial cost of drilling new wells. Going beyond the studied region to elsewhere in Bangladesh, a key issue is the cost of measuring arsenic in drinking water. If all the estimated 10 million wells in the country can be measured, then it should be possible to generalize the switching and drilling strategies in this article to the other high-arsenic areas in the country. However, if it is infeasible or expensive to measure all the wells, then a more sophisticated strategy of sampling would be appropriate.

It is perhaps a concern that, if people rush to switch to "safe" wells, that a few wells will be overburdened, and they may, in fact, draw out arsenic that otherwise would have gone into an existing "dangerous" well. We suspect this will not be a problem in most wells, because most of the well water in this area is in fact used for irrigation, not drinking. However, we plan to follow up the well-switching recommendations with arsenic measurements of a sample of wells that we expect to have a sharp increase in use. In the short term, it might also be useful to identify areas such as Clusters 45 and 121 where shallow wells are safe.

Perhaps most importantly in the long run is the research goal of understanding where the arsenic is coming from and what characteristics of a well can block it from entering the drinking water. A limited number of sediment cores collected in the study area, and previous work elsewhere in Bangladesh, show a fairly consistent relation between local geology and the distribution of groundwater arsenic.[2,4] Deeper wells that are low in arsenic typically tap into the so-called Dupi Tila formation, which is recognizable from the orange-brown coatings of the sand grains. In contrast, sediment cuttings recovered during drilling of shallower wells, which are often high in arsenic, are typically gray. These two types of deposits are often, but not always, vertically separated by a thick impermeable clay layer. Within our relatively small study area, the depth of this transition ranges from about 100 to 500 ft. Such spatial variability indicates the futility of establishing a single "safe" depth for arsenic at anything beyond the village level. From Fig. 4, we see gaps in well depths, below which wells are safe, in several clusters (e.g., those labeled 55, 243, and 273), and we plan to take more core samples to study this.

Unfortunately, geological factors cannot directly be used to determine good locations for wells. Geological and hydrological characteristics are at least as spatially variable as well-arsenic concentrations and much more difficult to document by drilling and mapping using various geophysical tools.[2] Our group is very active in this respect, but the information is available only in a handful of test areas at this point—nowhere near the spatial resolution of the well-arsenic data.

For other parts of the country, more data must be gathered, at enough of a sampling density to identify areas where many people are at risk for high arsenic exposure.[6] The appropriate sampling density will depend on the relative costs of measurement and drilling new wells. Setting up a sampling and decision plan for the rest of Bangladesh is an important next step in this research.

5.2. General Comments on Decision Analysis with Spatial Data

Decisions in public health and social policy typically are made at both aggregate and local levels. For example, in Bangladesh a national policy might be developed to encourage measurement of existing wells and drilling of new safe wells, possibly with some purification of surface water. At the same time, even with outside help, a local resident must decide whether to invest money and labor in drilling a deep tube well. In such a situation in which decisions are locally dispersed, one of the most important things a government or international organization can do is to provide information, both on the dangers of arsenic and on the potential benefits of remediation strategies.

There is a well-developed and longstanding theory of decision analysis using Bayesian inference.[14,15] A special feature of spatial data, compared with other sorts of information that can be used for decision analysis, is that they can be directly adapted for local decisions. In a Bayesian inferential setup, this leads to hierarchical modeling—as illustrated in

Reference 16; assigning a parameter to each local area allows for local decision recommendations. The implicit spatial character of hierarchical modeling is captured in the term "small area estimation,"[17] which is used for hierarchical modeling of survey data. Here, we have explored the strategy of going straight to the decisions without formally modeling, but whatever statistical method is used, it is important for it to capture the spatial structure of the data.

From a statistical point of view, the methods used in this article are not very sophisticated—but we think there is something new here, in that we are applying data analytic techniques directly to the decision problem. A more standard approach would be to estimate distributions, correlation functions, variograms, and so forth, in order to understand the spatial structure.[18,19] Our approach is almost a spatial version of a "spreadsheet" analysis in business, using the data to directly draw conclusions about potential outcomes.

ACKNOWLEDGMENTS

We thank several reviewers and Shaw-Hwa Lo for helpful suggestions and Xin Feng for calculations. Partial support for this research was provided by the National Science Foundation and the National Institute of Health Grant 1 P42 ES10349. The data were obtained with support from the NIEHS Superfund Basic Research Program (superfund.ciesin.columbia.edu).

REFERENCES

1. Smith, A. H., Lingas, E. O., & Rahman, A. (2000). Contamination of drinking-water by arsenic in Bangladesh: A public-health emergency. *Bulletin of the World Health Organization*, *78*, 1093–1103.
2. BGS and DPHE. (2001). Arsenic contamination of groundwater in Bangladesh. In D. G. Kinniburgh & P. L. Smedley (Eds.), *BGS Technical Report WC/00/19*, vol. 2, Final Report. Keyworth, UK: British Geological Survey.
3. van Geen, A., Ahsan, H., Horneman, A., Dhar, R. K., Zheng, Y., Hussain, A. Z. M. I., Ahmed, K. M., Gelman, A., Stute, M., Simpson, H. J., Wallace, S., Small, C., Parvez, M. F., Slavkovich, V., Loiacono, N. J., Becker, M., Cheng, Z., Momotaj, H., Shahnewaz, M., Seddique, A. A., & Graziano, J. (2002). Promotion of well-switching to mitigate the arsenic crisis in Bangladesh. *Bulletin of the World Health Organization*, *80*, 732–737.
4. van Geen, A., Zheng, Y., Versteeg, R., Stute, M., Horneman, A., Dhar, R., Steckler, M., Gelman, A., Small, C., Ahsan, H., Graziano, J., Hussain, I., & Ahmed, K. M. (2003). Spatial variability of arsenic in 6000 contiguous tube wells in a 25 km^2 area of Bangladesh. *Water Resources Research*, *39*, 1140.
5. Ahsan, H., Perrin, M., Rahman, A., Parvez, F., Stute, M., Zheng, Y., Milton, A. H., Brandt-Rauf, P., van Geen, A., & Graziano, J. (2000). Associations between drinking water and urinary arsenic levels and skin lesions in Bangladesh. *Journal of Occupational and Environmental Medicine*, *12*, 1195–1201.
6. Yu, W. H., Harvey, C. M., & Harvey, C. F. (2003). Arsenic in groundwater in Bangladesh: A geostatistical and epidemiological framework for evaluating health effects and potential remedies. *Water Resources Research*, *39*, 1146.
7. van Geen, A., Ahmed, K. M., Seddique, A. A., & Shamsudduha, M. (2003). Community wells to mitigate the current arsenic crisis in Bangladesh. *Bulletin of the World Health Organization*, *82*, 632–638.
8. BRAC. (2000). Combating a deadly menace: Early experiences with a community-based arsenic mitigation project in Bangladesh. In *Research Monograph Series*, vol. 16. Dhaka: BRAC.
9. Hartigan, J. A., & Wong, M. A. (1979). A *K*-means clustering algorithm. *Applied Statistics*, *28*, 100–108.
10. R Project (2003). *The R Project for Statistical Computing*. Available at http://www.R-project.org/.
11. Everitt, B. S. (1993). *Cluster Analysis*, 3rd ed. London: Edward Arnold.
12. Lawson, A. B., & Denison, D. G. T. (Ed.) (2002). *Spatial Cluster Modelling*. New York: CRC Press.
13. Church, R. L. (2001). Spatial optimization. In *International Encyclopedia of the Social and Behavioral Sciences*. Pergamon.
14. DeGroot, M. H. (1970). *Optimal Statistical Decisions*. New York: McGraw-Hill.
15. Clemen, R. (1996). *Making Hard Decisions*, 2nd ed. Duxbury.
16. Lin, C. Y., Gelman, A., Price, P. N., & Krantz, D. H. (1999). Analysis of local decisions using hierarchical modeling, applied to home radon measurement and remediation (with discussion). *Statistical Science*, *14*, 305–337.
17. Fay, R. E., & Herriot, R. A. (1979). Estimates of income for small places: An application of James-Stein procedures to census data. *Journal of the American Statistical Association*, *74*, 269–277.
18. Cressie, N. A. C. (1993). *Statistics for Spatial Data*, 2nd ed. New York: Wiley.
19. Chiles, J. P., & Delfiner, P. (1999). *Modeling Spatial Uncertainty*. New York: Wiley.
20. Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2003). *Bayesian Data Analysis*. London: Chapman & Hall.