A Safe Depth Forecasting Model for Insuring Tubewell Installations Against Arsenic Risk in Bangladesh

Matilde Trevisani[1(✉)], Jie Shen[2], Alexander van Geen[3], Andrew Gelman[4], Shuky Ehrenberg[5], and John Immel[6]

[1] University of Trieste, Via Tigor 22, 34123 Trieste, Italy
matildet@deams.units.it
[2] Capital One Auto Finance, Piano, USA
[3] Lamont-Doherty Earth Observatory of Columbia University, Palisades, NY, USA
[4] Department of Statistics, Columbia University, New York, NY, USA
[5] JGB Management, Boston, USA
[6] Joyful Belly Ayurveda, Asheville, USA

Abstract. Nowadays large spatial databases are available to help analysts facing a variety of environmental risk problems. Statistically accurate and computationally efficient algorithms and models are then needed to extract knowledge from these, for inference and prediction of the studied phenomenon, and, ultimately for decision both at country-wide policy and local level. Arsenic concentrations are naturally elevated in groundwater pumped from millions of shallow tubewells distributed across rural Bangladesh. Deeper tubewells often make access to groundwater with lower arsenic levels. Thereby, also thanks to a relatively low installation cost, they have proven to be an effective method to reduce arsenic exposure. Relying on a large database of well tests conducted in thousands of villages, we propose a supervised learning technique to estimate the probability that a new well will be low in arsenic based on its location and depth. For villages lacking direct information to make a local prediction, our technique, that we call the Sister-Village method, combines data from villages with similar characteristics. To further promote safe well installations and to help disseminate the information resulting from our method, we also propose and price a simple insurance model.

Keywords: Arsenic-depth pattern similarity · Probability curve · Bayesian learning · Calibration plot · Stratified cross-validation · Probability score

1 Introduction

Bacterial contamination of streams and ponds – the main source of drinking water until a few decades ago – was a leading cause of very high rates of infant mortality in rural Bangladesh. Simple tubewell technology providing access to groundwater aquifers that are typically free of human pathogens was introduced by international and non-governmental organizations (NGOs) in the mid 1970s. Although tubewells were

DOI: 10.1007/978-3-319-62404-4_1

effective in this regard, they introduced another deadly health hazard: widespread exposure to arsenic (*As*) levels exceeding the WHO guideline for *As* in drinking water of 10 μg *As* per liter (10 μg/L). A number of strategies have been suggested to help mitigate the harmful effects of groundwater *As*: water treatment, alternative water sources and tubewell based approaches. Tubewells are relatively cheap and robust, besides they may lead to positive interventions like well-switching and the digging of community wells [3, 4]. For these reasons, it is generally thought that a tubewell based approach will continue to play a major role in *As* mitigation for the near future [1].

The World Bank, UNICEF, and a number of NGOs contributed funding and data to the Bangladesh *As* Mitigation Program (BAMWSP) under which nearly five million wells were tested and identified for location, age, number of users, depth and *As* concentrations. Results of the program corroborated by other surveys indicate that there is a high degree of variability in the spatial distribution of groundwater *As*. The main reason is the spatial variability of the underlying geology. *As* is gradually flushed out of sandy aquifers over time resulting in lower concentrations in groundwater pumped from deeper (and therefore older) aquifers. However the transition to low-*As* aquifers may be quite abrupt, resulting in wells of similar depth reporting different levels of *As*, even in neighboring villages [7].

In our previous localized study of *As* in Araihazar upazilla (equivalent to a sub-district, of which there are about 500 in Bangladesh) we created a statistical model relying on village level *As* conditions – a spatial unit small enough to eliminate most of the previously mentioned variability [2, 7]. Our model was able to produce an individualized *safe depth* estimate for many villages: wells dug to a depth below the estimated threshold have a high probability of satisfying Bangladesh's standard for *As* in drinking water of 50 μg/L. Local decision analysis in Araihazar upazilla also led us to conclude that a strategy whereby individuals drinking from wells with high *As* concentrations would switch to a well low in *As* within walking distance could lead to a relevant exposure reduction.

Although previous information dissemination techniques were successful in inducing some well switching, preliminary survey information indicates that installation of wells to unsafe depths is still common, despite the occasional identification of a reliable safe depth by local drillers. Two of the most prominent reasons that have hampered the installation of safer wells are readily discernible. First, the *safe depth* identified by previous testing often requires a household to drill significantly deeper than they would otherwise, resulting in a marked increase in cost. Moreover, even at the pre-identified *safe depth* the probability of obtaining *As* free water is often well below 100%. Insurance could serve to significantly mitigate the risk component of drilling for the installation of a new well. Second, the statistical model underlying the *safe depth* method is incomplete, in the sense that there are instances where a *safe depth* is not discernible on the basis of data from existing wells. In the event that there are a very small number of *As* free wells in a village, it may not be possible to derive a *safe depth* (Fig. 1). Significantly, this occurs most frequently in areas exhibiting especially high concentrations of *As* and a dearth of safe wells – precisely the areas where new wells are most urgently needed.

In order to address the limitation of the *safe depth* method, we develop a new *Sister-Village* method for computing the probability of *As* contamination as a function of depth.

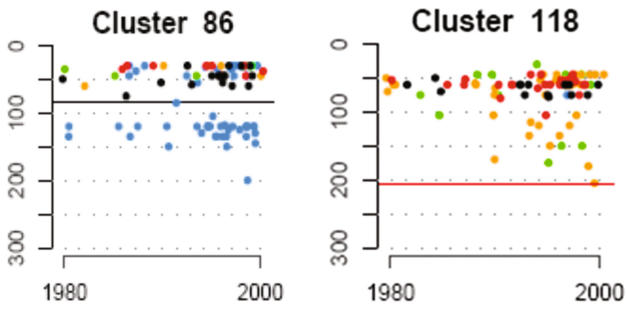

Fig. 1. Plots of *As* level as a function of depth and year of well installation in two village sized clusters. Cluster 86 allows for the computing of a *safe depth*, while cluster 118 does not (a lower bound – red line – below which the threshold cannot be derived is given instead).

The purpose of the *Sister-Village* method is two-fold: firstly it will allow us to arrive at a continuum of depth and safety probabilities. Secondly it is designed specifically to help overcome the *safe depth* method's informational failing. In order to magnify local information at the village level, we select, for each village, a number of sister villages on the ground of their similarity in terms of *As*-depth trend to the target village. Information available from these sister villages is then used to impute the probability of reaching *As* free water at different depths within the target village.

The paper proceeds as follows: part 2 presents exploratory data analysis; part 3 describes the *Sister-Village* model; part 4 examines its effectiveness by means of several cross-validation criteria; part 5 applies it to an insurance model and concludes.

2 Initial Analysis of Arsenic Distribution

We conduct our exploratory analysis in a subset of 16 of Bangladesh's 520 upazillas. Within these areas, 81% and 63% of sampled tubewells have *As* concentrations above the 10 μg/l WHO standard and, respectively, the 50 μg/l drinking water standard in Bangladesh. Figure 2 shows average *As* levels in each of the study area's 214 unions, the next administrative unit down from the upazilla, which are composed in turn of 3283 villages and represent a total of 355, 846 tubewells. The western region has the lowest levels of *As* contamination, while the area to the north-east suffers from the highest. But there is also significant patchiness at the union and even village level.

Depth profiles of *As* compiled for each of the 16 upazillas show a complex relationship between depth and *As* level (Fig. 4). In the study area, approximately 60% of the wells lie within a 60–120 ft depth interval, which unfortunately corresponds to the highest *As* concentrations, with shallower and deeper wells exhibiting lower average *As* levels (Table 1 and Fig. 3). Despite this general trend, there are certain upazillas where even shallow wells are extremely unsafe (e.g. Shahrasti) or even deeper wells have relatively high average *As* concentration (e.g. Banchharampur) or almost all the wells indicate low *As* levels throughout the depth range (e.g. Damurhada). Even if in general

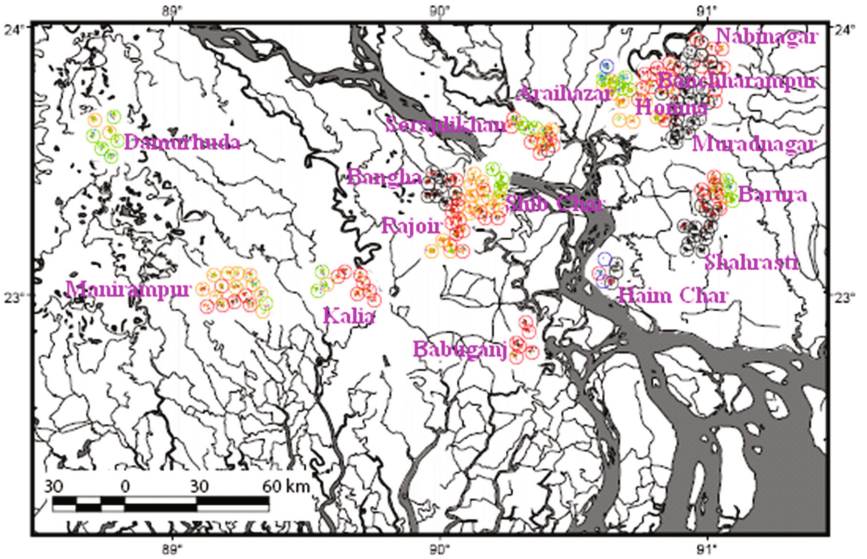

Fig. 2. Average *As* levels in unions and villages in our research area. The (0, 0) point on this graph corresponds to coordinates 23.36° North and 89.90° East. Colors indicate *As* levels: blue (less than 10 µg/l), green (10–50), orange (50–100), red (100–200), and black (>200). Each dot represents a village and each circle represents a union. (Color figure online)

Table 1. Number of wells and safe wells proportion for depth strata. The whole data (355, 846 wells) are considered.

Depth (ft)	Number of wells (%)	Safe wells proportion
$D < 60$	17.3	0.625
$60 \leq D < 75$	18.1	0.333
$75 \leq D < 100$	31.0	0.262
$100 \leq D < 125$	15.6	0.318
$125 \leq D < 175$	9.1	0.472
$175 \leq D < 225$	5.2	0.535
$225 \leq D < 275$	2.2	0.685
$275 \leq D < 325$	0.5	0.711
$325 \leq D < 400$	0.1	0.683
$400 \leq D < 500$	0.1	0.703
$D \geq 500$	0.8	0.906

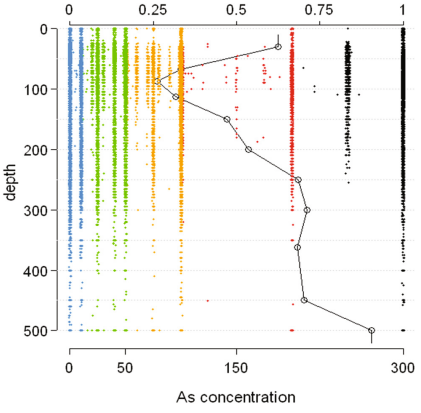

Fig. 3. Safe well proportion (top-axis) as function of well depth (*y*-axis) for the whole data. Dots are depicted as in Figs. 2 and 4. (Color figure online)

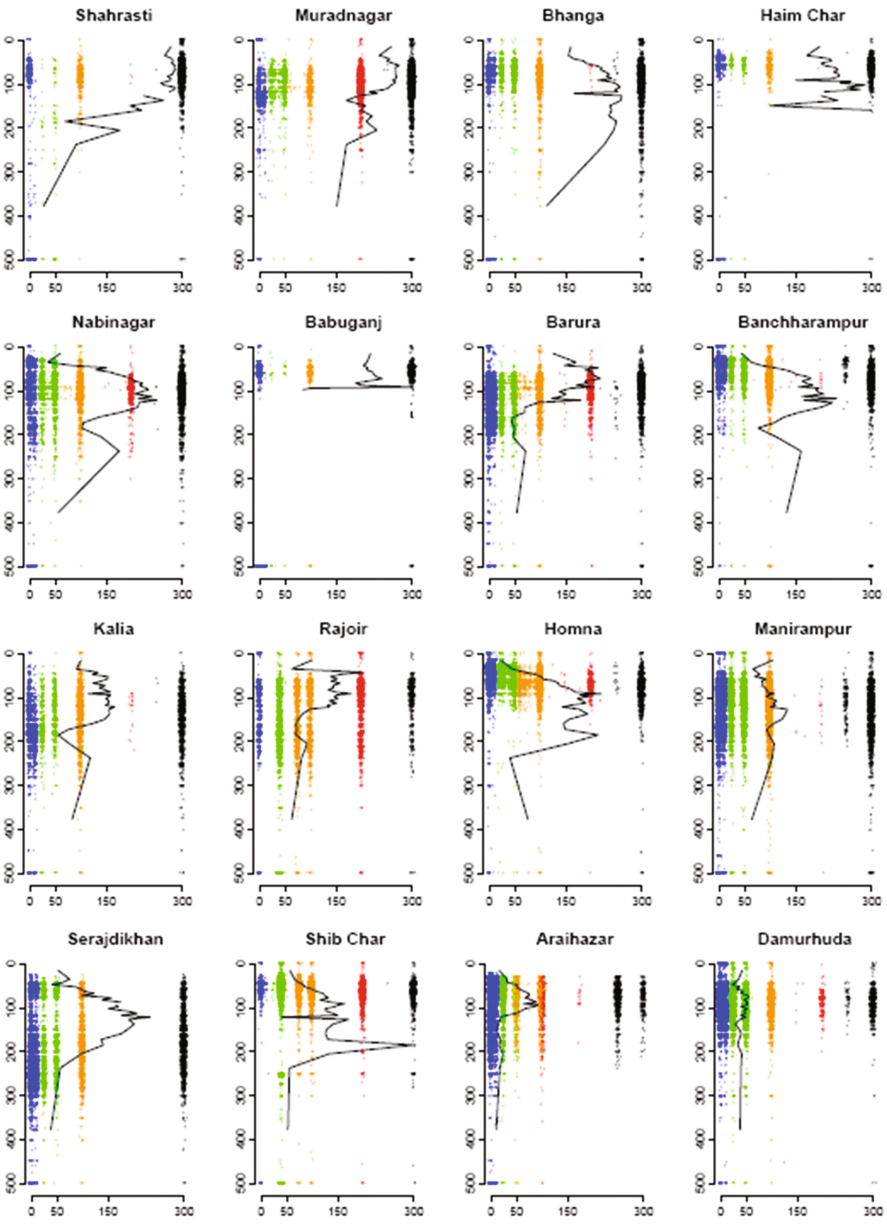

Fig. 4. *As* concentrations (*x*-axis) shown as function of well depth (*y*-axis) in 16 upazillas with each dot referring to one well and colors using a scale identical to the one of Fig. 2. The panels are arranged in decreasing order of average *As* concentration for each upazilla. The black line indicates average *As* concentration as a function of depth within each upazilla.

very shallow wells can contain less *As*, reliance on these wells is not a long-term strategy because they are vulnerable to bacterial and industrial contamination.

The origin of the three-dimensional geographic variability of *As* levels in groundwater of Bangladesh is only partially understood. Whereas the available data show considerable spatial variability, it is reasonable for the present analysis to assume that *As* concentrations are essentially invariant in time, even if the occasional mechanical failure of a deep well has been documented [8].

3 Sister-Village Method

The purpose of the *Sister-Village* method is to provide a continuous menu of depths and associated *As* contamination probabilities, especially in regions where there is a dearth of safe wells causing estimation methods relying solely on local data to fail. In order to achieve these objectives we make use of information available from sister-villages – villages whose depositional setting is assumed to have been similar to the one of the target village. The assumption underlying the *Sister-Village* method is that there exist a limited number of depositional patterns, of which *As*-depth distribution is one phenomenon, so that data from diverse geographic regions can be used to make predictions where data is not available. A similar pattern is very often to be found in nearby regions, but it also could exist in discontinuous locations. Moreover, our previous work has shown that disaggregating the data to the village level is justified from both a geological and a practical perspective.

3.1 You Can Choose Your Family – Selecting Sister Villages

The reliability of the *Sister-Village* method rests, given a query village, on a precise identification of best matching villages in the dataset. We use the village of Patershari in the Peruli union within the Kalia upazilla as a sample target village to illustrate the process of sister-village selection and probability menu computation. The outcome of the sister-village selection is a set of scores measuring the degree of similarity between the target village (A) and all other villages (B_i. $i = 1, 2, \ldots, N - 1$, being N the total number of villages and $B_i \neq A$).

Quantifying Local Village Characteristics. We begin by characterizing local *As* patterns in each village. Initially we divide the range of well depths in each village into nine mutually exclusive strata: 0–75, 75–125, 125–175, 175–225, 225–275, 275–325, 325–400, 400–500, ≥ 500 ft. The number of total wells and safe wells sunk into stratum k are denoted by n_k and y_k respectively, with k ($1 \leq k \leq 9$) indexing the nine different depth strata. In each stratum k, we obtain the estimated probabilities of safe depth in village A and B_i, $\widehat{\pi}_k^A$ and $\widehat{\pi}_k^B$ whose prior values are set to 0.5, as indicated by equations

$$\hat{\pi}_k^A = \frac{y_k^A + 1}{n_k^A + 2} \qquad \hat{\pi}_k^B = \frac{y_k^B + 1}{n_k^B + 2}. \tag{1}$$

The raw discrepancy between each village pair (A, B_i) in stratum k can be calculated as follows:

$$\hat{\theta}_k^{\text{raw}} = (\hat{\pi}_k^A - \hat{\pi}_k^B)^2 \quad \text{with } \widehat{\text{bias}} = \frac{\hat{\pi}_k^A(1 - \hat{\pi}_k^A)}{n_k^A + 2} + \frac{\hat{\pi}_k^B(1 - \hat{\pi}_k^B)}{n_k^B + 2}. \tag{2}$$

Combining the information from all k strata, we obtain:

$$\text{discrepancy}(A, B_i) = \sum_k (\widehat{\theta}_k^{\text{raw}} - \widehat{\text{bias}}_k) \times \frac{n_k^A}{n_k^A + 5} \tag{3}$$

with an adjusting factor $n_k^A/(n_k^A + 5)$ assigning more weight to those strata of village A containing a higher concentration of wells. The smaller the value of the discrepancy – as described by Eq. (3) – the more similar villages A and B_i will be on average.

Categorizing Spatial Information. As mentioned previously, we assume that nearby geographical regions have similar geological properties. That being said, it is possible that similar geological properties can be observed in more distant geographic regions (i.e. in non-contiguous areas). We represent geographical variation amongst villages through a category difference function,

$$\text{Category difference } (A, B_i) = \begin{cases} 1, \text{ in the same union} \\ 2, \text{ in the same thana, but different unions} \\ 3, \text{ in the same division, but different thanas} \\ 4, \text{ in different divisions} \end{cases}$$

Two fundamental assumptions underlie our categorization method. Firstly, we assume that the administrative partition of villages reflects proximity. Secondly, we assume that proximity is related to some extent to the underlying geology that produced the observed patterns. Moreover, we assume that the distance score is a linear function of the administrative level. Figure 5 shows a scatter plot of discrepancy values as a function of category differences between each pair of villages (A, B_i).

Initially we make use of the geographic coordinates to measure the distance between any two villages. However, since only 10% of all sampled wells include this information, we are forced to make use of administrative partitions. Fortunately, a comparative analysis of the wells with geographic coordinates indicates that administrative partitions offer very good approximations for geographic distances. Furthermore, as Fig. 2 suggests, nearby villages are likely to have similar average *As* distributions.

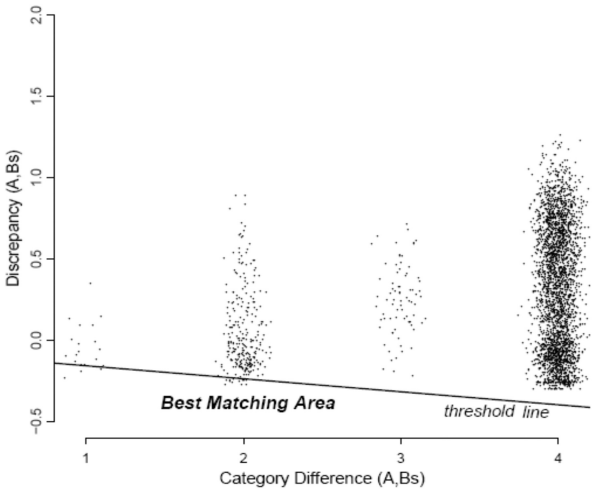

Fig. 5. Plot of discrepancy values as function of category differences for each village pair (A, B_i) with target village Patershari (A). Each dot represents a different village B_i (with a horizontal random jitter). Dots that fall below the threshold line are selected as the sister villages for A.

Selecting Sister Villages. We represent the total difference in *As* patterns within a village pair (A, B_i) as a weighted sum of local discrepancy and category difference score, to form a total score = discrepancy + $\alpha \cdot$ (category difference), with α indicating the relative weights given to non-spatial and spatial dissimilarity. This equation can be rewritten as

$$\text{discrepancy} \;=\; \text{total score} \;-\alpha \cdot (\text{category difference}) \tag{5}$$

Representing each village pair (A,B_i) as a dot on a two dimensional graph, we can use Eq. (5) to draw a threshold line defining the best matching area (Fig. 5).

Any two of the three parameters (α, total score, γ), with γ being the number of sister villages, are sufficient to determine a threshold line which in turn will determine village A's sister villages. The choice of γ depends on a tradeoff between two factors: on the one hand we need to have a sufficient γ for effective estimation; on the other, a high degree of similarity between the target village and its sisters must be maintained. The optimization of the floating parameters is discussed in Sect. 4.

Estimating the Probability that a Depth Interval is Safe. Next we estimate the probability that a well will be safe in a given depth stratum k for a village A, $\widehat{p}_k^A$. First we pool the wells from village A' sister villages, with n_k^s and y_k^s indicating the total number of wells and number of safe wells in depth stratum k, respectively. Now, $\widehat{p}_k^A$ is computed as a weighted sum of the information from village A's wells and from the wells in its sister villages,

$$\tilde{p}_k^A = \frac{y_k^A + \lambda / f(n_k^A \times y_k^s + 1)}{n_k^A + \lambda / f(n_k^A) \times n_k^s + (1 / p_k^{\text{prior}})}, \tag{6}$$

where λ is a constant and $f(n_k^A) = \sqrt{(n_k^A + 2)}$ Thus the weight $\lambda / f(n_k^A)$ associated with information from the sister villages is inversely proportionate to n_k^A, the total number of wells in the target village (consistent with the *borrowing strength* rule).

Let p_k^{prior} represent our prior information about the probability that a well in stratum k in village A is safe. We construct p_k^{prior} from *As*-depth distribution characterizing the upazilla of the target village, that is

$$p_k^{\text{prior}} = \frac{\text{number of safe wells in stratum } k \text{ in upazilla } + 1}{\text{number of wells in stratum } k \text{ in upazilla } + 2}$$

Besides, since the general geology of Bangladesh leads us to believe that deeper wells are on average safer, we adjust the probability curves for all villages to be monotone for deeper strata. For strata 5–9 (wells deeper than 250 ft) we draw an upper boundary curve requiring that the estimated probabilities of stratum k are equal to or greater than the probabilities of stratum $k - 1$. We also draw a lower boundary curve that works symmetrically to the upper curve, requiring that the estimated probabilities for stratum

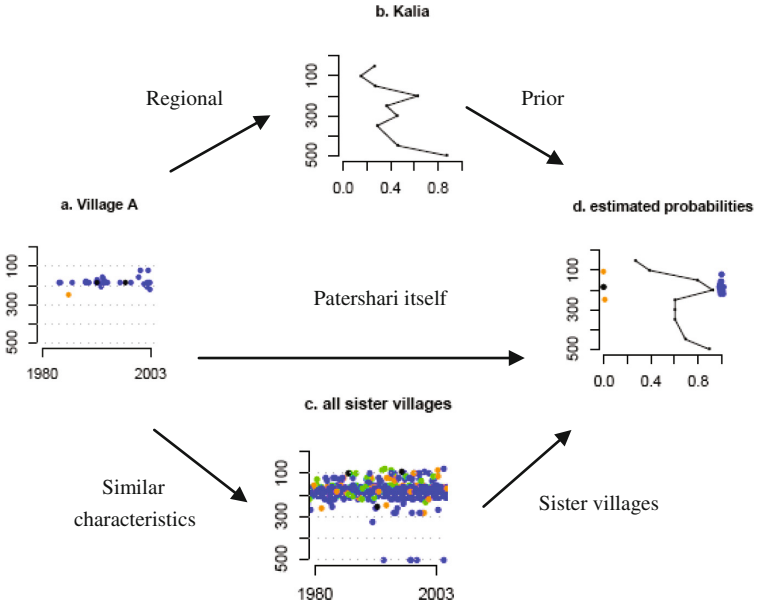

Fig. 6. Estimation of probability curve for Patershari (target village *A*) in Kalia upazilla. *As* level as function of depth and installation year for wells in *A* (plot a) and in the γ sister villages (c). Average safety probability curve for Kalia (b). Estimated probability curve for *A* (d).

k do not exceed those estimated for stratum $k + 1$. Thus, we produce an adjusted estimated probability curve as the average of the upper and lower bounds.

Figure 6 represents a scheme of the probability estimation process for a target village A starting from the well distribution (left plot) and ending with the estimated probability curve (right). The probability of a well being safe increases with depth, peaking at approximately 0.90 correspondent to a depth of 200 ft. Information derived from sister villages (bottom) and geographical trends (top) further supports tentative observations based on local information, indicating that 200 ft deep wells are very likely to be safe. Combining this information with monotonic adjustments then allows us to extend the probability curve below 200 ft, despite a lack of wells dug below this depth in village A.

4 Quantifying Uncertainty

Since predictive accuracy is important in the context of insurance, we use both calibration curves and probability-scores yielded by implementing three different cross validation (CV) criteria, to measure the level of uncertainty underlying our estimated probability curves. We also use probability scores as a criterion in the optimization of floating parameters when selecting sister villages.

4.1 Cross-Validation

We design three types of CV criteria: first we use a stratified k-fold CV method, then we reevaluate after truncating deeper wells, finally we leave out recently dug wells. We explain the rationale underlying each method in context.

To test the credibility of our method, we apply a k-fold CV albeit making it suitable to our specific analysis. That is, in partitioning data (the wells) into k parts we consider that wells are grouped into villages, moreover, that they are unevenly distributed across depth. Hence, before randomly allocating data to folds, we carry out a two-way stratification, by village as well as by depth strata, and, either way, proportionally to the stratum size (the number of wells therein). Once data are divided into k parts (k is set to 5), the k-fold CV proceeds as customarily.

The k-fold CV is aimed at evaluating the overall predictive ability of our methodology. On the other hand, to test how effective it is for prediction of specific data groups, we implement two further validation criteria. The *leave-deep-out* method removes for each village, considered the target village in turn, all deeper wells – those below 150 ft – and estimates the probability curve on the basis of its remaining shallower wells. The estimated probabilities and actual safety values for the holdout deeper wells of every village, making up the test set, are then recorded. Similarly, the *leave-recent-out* method eliminates more recent wells – those dug after 2000 – with the purpose of testing how effective our model is in predicting the safety of a new well on the basis of past data.

4.2 Calibration Plots and Probability Scores

CV results are summarized both graphically, by drawing calibration plots, and quantitatively, by means of probability scores.

The probability score (p–score) consists in the average log-likelihood of the set y_i, $i = 1,\ldots,n$, constituting the test set of the undergoing CV method. More in detail, each well i of the test sample receives a score of either $\log(p_i)$ if it is safe ($y = 1$), or $\log(1 - p_i)$ if it is unsafe ($y = 0$) – with p_i indicating the estimated safety probability derived from fitting the model to the training sample as defined in the undergoing CV analysis – so that, by averaging, the $\text{p}-\text{score} = \sum_i \{y_i \log(p_i) + (1 - y_i) \log(1 - p_i)$ is obtained. The closer the score is to 0, the more reliable is our method.

Analyses of prediction performance are typically accompanied by calibration plots in which the relative frequencies of an event are plotted against the respective estimated probabilities. For example, Fig. 7 displays the calibration curves for the stratified *k*-fold, the *leave-deep-out* and *leave-recent-out* CVs (plot, respectively, at top, bottom left and right): the dots represent actual proportions of safe wells and corresponding estimated safety probabilities as averaged over a number – set to 20 – of equal probability intervals into which the range of estimated probabilities has been divided. All three calibration curves are very close to the 45 degree line, suggesting that the *Sister Village* method is highly reliable on aggregate. A more careful examination shows that there is

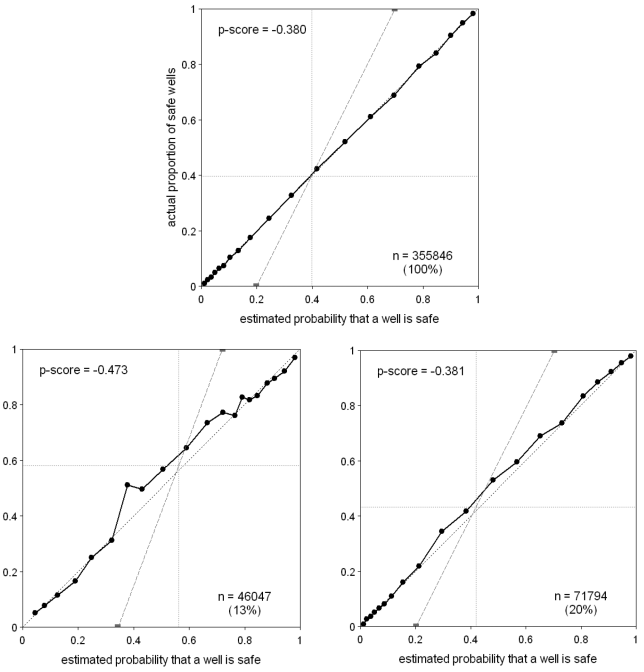

Fig. 7. Calibration of safety probability estimates using stratified *k*-fold/*leave-deep-out*/*leave recent-out* CV (top, bottom left and right). p–score and number of tested wells (plus % over the total) are added to each plot. The 45° line shows the curve of perfect calibration.

a slight underestimation in the 0.3–0.8 probability range, both in predicting deeper wells and – less relevantly – recent wells safety. However, given the nature of the problem, it is at least reasonable if not prudent to err on the side of caution in our estimations. On the insurance side, it implies an overpricing, that is a lesser risk of under-coverage for the insurer.

In order to further corroborate the results for the overall data, we create additional calibration curves in which cross-validated data-after 2000 are disaggregated at both upazilla and depth-stratum levels. Our estimations hold up well at the upazilla level for the most part. Inspection by depth stratum (Fig. 8) indicates that the *Sister Village* method generally enjoys very high predictive accuracy at depth shallower than 275 ft, with some biases existing for deeper wells. However, these biases are not of especial concern since the vast majority of well installations occur above the 300 ft level.

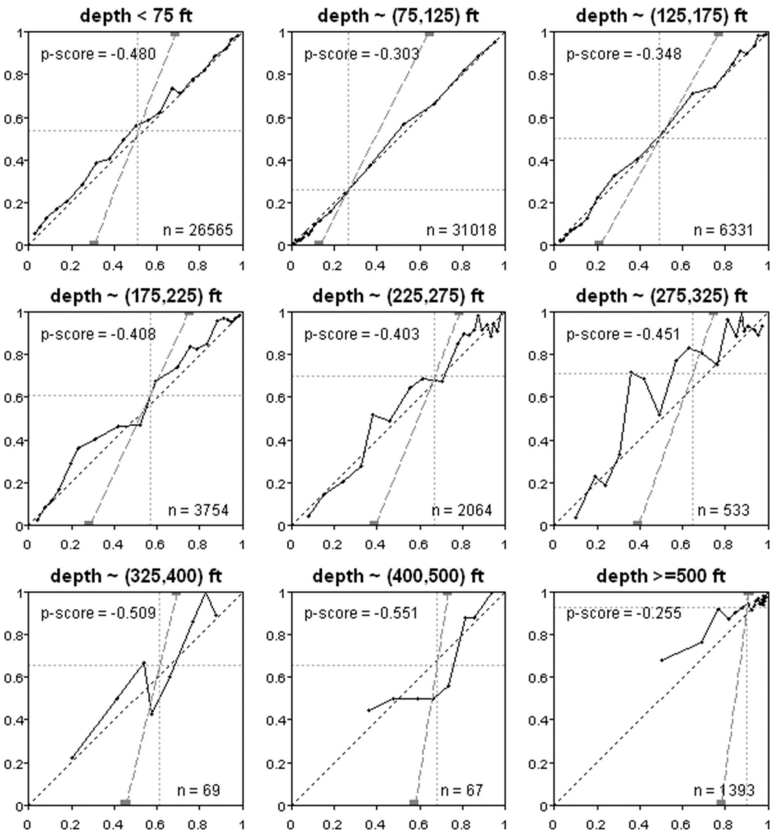

Fig. 8. Calibration of safety probability estimates at nine different depth ranges. Test data of *leave-recent-out* CV method are used. Axes notation is as in Fig. 7.

4.3 Optimal Determination of Floating Parameters

The *Sister-Village* model has three floating parameters: α and γ determine sisters villages' selection whereas λ controls the relative weight of sister and target villages. We select the parameter vector $\left(\widehat{\alpha}, \widehat{\gamma}, \widehat{\lambda}\right)$ which maximizes the p–score under the stratified *k*-fold CV.

Contours of p–score across a wide span of parameter values (Fig. 9) shows that information from sister villages (quantified by both γ and λ) is relevant to improve the method's predictive ability. Moreover, integration of geographic information

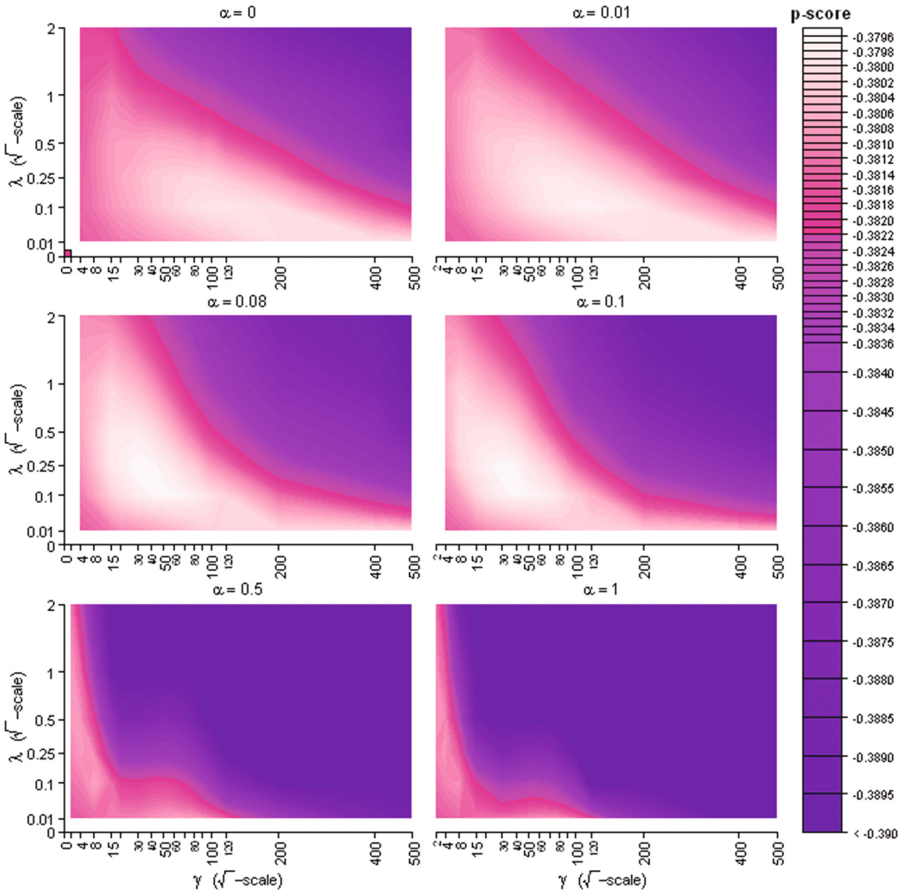

Fig. 9. Filled contour plots of p–score as function of the three floating parameters: λ versus γ (represented on *y* and *x* axes after being square-root transformed) conditional on α. p–score of case $\lambda = \gamma = 0$ (−0.3822) depicted as a point (red-colored) in the top-left panel ($\alpha = 0$) is the boundary level between parameter combinations in favor of the *Sister-Village* method (higher p–scores, increasingly lighter reds) and those against it (lower p–scores, increasingly darker violets). The maximum p–score (−0.3796) is achieved at $\lambda = 0.25$, $\gamma = 30$, $\alpha = 0.08$. (Color figure online)

(reflected by α) improves on performance as well. In fact, p–score obtained by using the sole local information ($\gamma = \lambda = 0$; see the left top panel, $\alpha = 0$) is only -0.3822. Besides, note that sister villages are more efficiently selected when spatial similarity is taken into account: a relatively small number of sister villages (γ from 30 to 50 when α is around 0.1; see mid-panels) enables the method to achieve even better results than those produced by increasing γ solely (when α is about 0 the best scores correspond to γ from 100 up to 400). Lastly, as expected, there is a trade-off between γ and λ (contours are negatively oriented everywhere) so that an optimal balance between number of sister villages and weight of non-local information need to be found. Concluding, p–score reaches its maximum of -0.3796 at $\alpha = 0.08$, $\gamma = 30$, $\lambda = 0.25$ (left-mid panel). Floating parameters of our final model are then set at these values.

We have also explored to what extent the monotone adjustment affects p–scores (Fig. 10). Almost everywhere p–score is relatively stable when the monotonic transformation is limited to deepest strata (k = 9, 8, 7) – practically at the level associated with having not transformed ($k = 0$) – while drops when is applied below 300 ($k = 6$), or, even more steeply, below 250 ft (k = 5). The maximum occurs at k = 7, i.e. 350 ft.

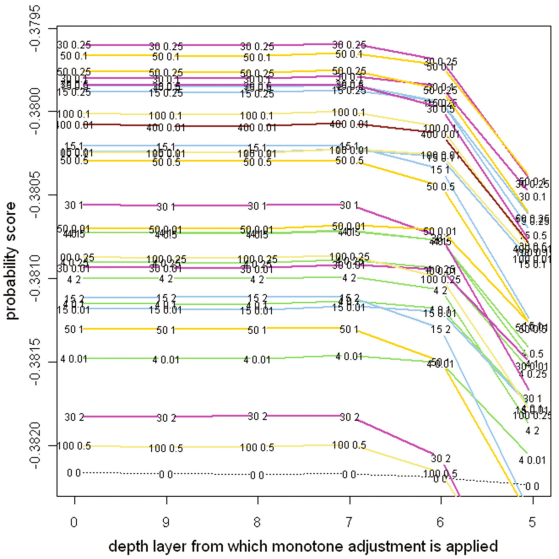

Fig. 10. p–score as function of depth stratum below which the probability curve is made monotonic, i.e. below stratum: 5 (250 ft), 6 (300 ft), 7 (350 ft), 8 (450 ft), 9 (500 ft), 0 (no monotonic transformation). p–scores superior to -0.3822 associated with reference case $\gamma = \lambda = 0$ (broken line) are only considered. The plot shows only a sample of parameter combinations, i.e. $\gamma = 4$, 15, 30, 50, 100, 400, coupled with $\lambda = 0.01$, 0.1, 0.25, 0, 5, 1, 2, all conditional on $\alpha = 0.08$.

5 Insurance and Final Remarks

In order to mitigate the risks inherent in installing a new well with unknown ex post *As* properties, we devise an insurance plan. The insurance plan also serves as a useful tool for disseminating depth safety information, which is incorporated into the insurance model through the premium's pricing.

Individuals can choose to opt in to the insurance plan – essentially a money-back guarantee, by paying a premium. In the event that an insured well fails, the insurance plan will refund most of the costs associated with digging the well. We make use of the probability curves derived from the *Sister-Village* method in order to effectively price the insurance plan. Since the plan purpose is to reduce *As* exposure, and since our method provides somewhat conservative results, we assume a risk neutral insurer.

5.1 Rationale for Insurance: Risk Mitigation, Dissemination of Information

Classic economic theory suggests that the convexity of individual preferences results in risk aversion [5]. As a corollary, an individual exhibiting risk aversion will always choose to fully insure when faced with a fairly priced insurance contract. In the context of well digging, a newly dug well can be seen as a bet with expectation $E(V(w)) = C\Pr(As)$ where $V(w)$ is equal to the value of the well, C its cost, and $\Pr(As)$ indicating the probability of the well being contaminated. Assuming consumers are strictly risk-adverse will allow us to provide insurance at a fair price (plus transaction costs), while maintaining the long run solvency of the plan. Our hope is that the availability of insurance will result in individuals digging more and more expensive (deeper) wells, especially in high risk areas, since they are able to contract away their risk.

In addition to its risk mitigation properties, insurance can be used to disseminate information regarding safe depth. Firstly, we convey safety information through premium price menus – a safer depth will result in a cheaper insurance premium per foot. Secondly, we convey the information directly when possible by recommending that individuals dig to the optimal *safe depth*. As our previous analysis [2] has shown, a strategy based solely on information dissemination and well switching will reduce *As* exposure by 38%. Additional informed digging of a small number of wells based on the insurance model will serve to further reduce exposure significantly.

5.2 Pricing and Imperfect Credit Market

The basic price of an insurance contract for a risk neutral insurer is equal to the probability of the event being insured against occurring multiplied by the cost of restitution if it occurs. Transaction specific costs are then added to the basic price.

The cost of a well can be decomposed into two parts: the fixed costs of the head and filter, f, and the varying cost of a foot of depth, v. The total price of an uninsured well is therefore $p = f + dv$, with d indicating depth in feet. Likewise the price of an insurance

contract can be decomposed into the fixed transaction costs (load, L) and the varying costs of providing insurance per foot of depth, $P = L + \Pr(As)\ dv$.

There is no need to insure most of the fixed costs, since they are largely recoverable. In general it is our goal to insure all or nearly all unrecoverable costs.

The introduction of a load factor to the price of the insurance is likely to reduce its attractiveness. However, the fact that the load is fixed means that the optimal quantity of insurance purchased is not likely to change dramatically which involves that those who do choose to insure will choose to do so fully.

Due to the possibility of imperfect credit markets, we also make allowances for a price menu of partial insurance. An individual can choose how many feet to insure, while paying the per foot price of the deepest foot.

5.3 Adverse Selection, Moral Hazard and Contract Design

The structure of the *Sister Village* model may lend itself to inadvertent undesirable selection effects. If wells very close to a target well are better predictors than the model, and if individuals who have just had a well fail are more likely to want to insure than individuals digging a well for other reasons, they may select into our insurance plan heavily. These adversely selected wells would in fact be riskier than the model predicts, resulting in a higher frequency of claims. We are able to partially address these concerns by introducing more detailed geographic information. Moreover, we implement a "one shot" policy: a single individual can only insure one well.

A serious concern is that of collusion between the well owner and driller. They may contract to dig a shallower well that indicated to the insurance agent (at a lower cost), splitting the difference. In order to avoid this problem all wells will be measured before a claim is paid. Furthermore, all wells will be destroyed upon the payment of a claim so that people will not have an incentive to temporarily poison their well.

5.4 Conclusions

This work use spatial databases to help analyse a major environmental risk: soil As contaminating groundwater. In continuation of previous works, we develop a supervised learning technique for estimating a probability curve over depth of *As* water safety, also for villages lacking direct information by borrowing strength from sister villages. Our initial analysis is based on information about each well's depth, *As* level, year of installation and administrative location. There are a number of other variables which could potentially become available in the future, and so used to improve the reliability of our method. The first possibility is adding detailed geographic information about each well. This type of expansion may be relevant also to the insurance model as discussed above.

References

1. Ahmed, M.F., Ahuja, S., Alauddin, M., Hug, S.J., Lloyd, J.R., Pfaff, A., Pichler, T., Saltikov, C., Stute, M., van Geen, A.: Ensuring safe drinking water in Bangladesh. Science **314**, 1687–1688 (2006)
2. Gelman, A., Trevisani, M., Lu, H., van Geen, A.: Direct data manipulation for local decision analysis, as applied to the problem of arsenic drinking water from tube wells in Bangladesh. Risk Anal. **24**, 1597–1612 (2004)
3. Madajewicz, M., Pfaff, A., van Geen, A., Graziano, J., Hussein, I., Momotaj, H., Sylvi, R., Ahsan, H.: Can information alone both improve awareness and change behavior? Response to arsenic contamination of groundwater in Bangladesh. J. Dev. Econ. **84**, 731–754 (2007)
4. Opar, A., Pfaff, A., Seddique, A.A., Ahmed, K.M., Graziano, J.H., van Geen, A.: Responses of 6500 households to arsenic mitigation in Araihazar, Bangladesh. Health Place **13**, 164–172 (2007)
5. Rabin, M.: Risk aversion and expected-utility theory: a calibration theorem. Econometrica **68** (5), 1281–1292 (2000)
6. van Geen, A., Zheng, Y., Versteeg, R., Stute, M., Horneman, A., Dhar, R., Steckler, M., Gelman, A., Small, C., Ahsan, H., Graziano, J., Hussein, I., Ahmed, K.M.: Spatial variability of arsenic in 6000 tube wells in a 25 km^2 area of Bangladesh. Water Resour. Res. **39**(5), 1140 (2003). doi:10.1029/2002WR001617
7. van Geen, A., Trevisani, M., Immel, J., Jakariya, M., Osman, N., Cheng, Z., Gelman, A., Ahmed, K.M.: Targeting low-arsenic groundwater with mobile-phone technology in Araihazar. J. Health Popul. Nutr. **24**, 282–297 (2006)
8. van Geen, A., Cheng, Z., Jia, Q., Seddique, A.A., Rahman, M.W., Rahman, M.M., Ahmed, K.M.: Monitoring 51 deep community wells in Araihazar, Bangladesh, for up to 5 years: implications for arsenic mitigation. J. Environ. Sci. Health **42**, 1729–1740 (2007)