Pardo-Igúzquiza, E., Martínez-Santos, P. and Martín-Loeches, M. (2018) ‘A geostatistical protocol to optimize spatial sampling of domestic drinking water supplies in remote environments’, *Stochastic environmental research and risk assessment*, 32(8), pp. 2433–2444.

doi:10.1007/s00477-017-1499-4

A geostatistical protocol to optimize spatial sampling of domestic drinking water supplies in remote environments

Eulogio Pardo-Igúzquiza[1] · Pedro Martínez-Santos[2] · Miguel Martín-Loeches[3]

Abstract
This paper deals with the design of optimal spatial sampling of water quality variables in remote regions, where logistics are complicated and the optimization of monitoring networks may be critical to maximize the effectiveness of human and material resources. A methodology that combines the probability of exceeding some particular thresholds with a measurement of the information provided by each pair of experimental points has been introduced. This network optimization concept, where the basic unit of information is not a single spatial location but a pair of spatial locations, is used to emphasize the locations with the greatest information, which are those at the border of the phenomenon (for example contamination or a quality variable exceeding a given threshold), that is, where the variable at one of the locations in the pair is above the threshold value and the other is below the threshold. The methodology is illustrated with a case of optimizing the monitoring network by optimal selection of the subset that best describes the information provided by an exhaustive survey done at a given moment in time but which cannot be repeated systematically due to time or economic constrains.

Keywords Stochastic · Entropy · Indicator variable · Thermotolerant coliforms · Groundwater sampling · Water access · Human rights

1 Introduction

Human settlements in rural sub-Saharan Africa often rely on groundwater. Groundwater is widely accessed for a variety of reasons. For one, natural storage and replenishment capacity are often high, while water quality is generally appropriate for most uses and infrastructures tend to be affordable for poor communities (Adelana and MacDonald 2008). Moreover, groundwater is ubiquitous across most of the continent (MacDonald et al. 2012; Pavelic et al. 2012) and provides a reliable source of freshwater during droughts (Llamas and Martínez-Santos 2005; Calow et al. 2010). Thus, it is estimated that Africa's shallow aquifers underpin the daily lives of around 200 million people (Foster and Garduño 2013). The Republic of Mali is no exception. In rural areas, where groundwater is by far the main source of drinking water supply, an estimated 800,000 traditional wells and 9000 community boreholes currently exist (Barry and Obuobie 2012).

While shallow groundwater provides an accessible, affordable and reliable resource, it is also easily contaminated. Contamination in groundwater supplies usually translates into widespread gastro-intestinal disease among groundwater-dependent populations. Monitoring is required to ensure adequate water quality on a consistent basis, as well as to develop early-warning protocols and to

✉ Pedro Martínez-Santos
pemartin@geo.ucm.es

Eulogio Pardo-Igúzquiza
e.pardo@igme.es

Miguel Martín-Loeches
miguel.martin@uah.es

[1] Instituto Geológico y Minero de España, Ríos Rosas, 23, 28003 Madrid, Spain

[2] Departamento de Geodinámica, Universidad Complutense de Madrid, c/ José Antonio Novais, n° 12, Ciudad Universitaria, 28040 Madrid, Spain

[3] Departamento de Geología, Geografía y Medio Ambiente, Facultad de Ciencias Ambientales, Universidad de Alcalá, Campus Universitario. Ctra. Madrid-Barcelona, Km 33,600, 28801 Alcalá de Henares, Madrid, Spain

identify potential contamination sources. Unfortunately, monitoring is often overlooked in rural water supplies. In the case of communal boreholes this is largely due to the generalized absence of facilities and qualified technicians. Take for instance the now-obsolete Millennium Development Goals, which used the presence of "improved water sources" as a proxy for "access to safe water" because widespread testing was considered to be "prohibitively expensive and logistically complicated" (UNICEF/WHO 2012, p. 4). This has important implications for global figures, which most likely underestimate the number of people with access to safe drinking water supplies (Martínez-Santos 2017a). On the other hand, sheer unawareness and the absence of means on the part of the user explain why traditional wells are seldom monitored.

Developing effective sampling protocols is perceived as a much-needed step towards ensuring the safety of drinking water supplies. Ideally, these should allow for obtaining as much information as possible from a limited number of points. The optimal spatial design of a network of monitoring points (i.e. the establishment of a network or the optimal extension or reduction of an existing network) has given rise to numerous investigations into the problem in a stochastic framework by using geostatistics and spatial statistics. One of the most commonly used methods has been variance reduction (Rouhani 1985). This has been extended in a number of ways, for instance by using simulated annealing as the global stochastic optimization method (Pardo-Igúzquiza 1998a). These methods are based on minimizing total kriging variance, which is a measure of the uncertainty of a spatial field obtained by spatial interpolation. Additional approaches use the concept of spatial entropy (Bueso et al. 1999; Angulo et al. 2000), or a combination of variance reduction and spatial entropy (Pardo-Igúzquiza and Dowd 2005), in order to select the most informative locations and variables. These methods are preferred over other alternatives such as d-optimal design (Chen et al. 2003) for spatial problems such as the one presented in the following pages. The concept of entropy as a measure of information has been widely used in many different contexts for space–time mapping (Douaik et al. 2004; He and Kolovos 2017), combining different kinds of spatial information (Wibrin et al. 2006), adaptive sampling network (Wang and Harrison 2013), sensitivity analysis (Zeng et al. 2012) and correlated nonlinear shrinkage (Angulo et al. 2011), among other applications.

This paper develops a methodology that combines the probability of exceeding some particular thresholds with a measure of the information provided by each pair of experimental points to define the optimal locations to sample in regions where extensive field surveys may be difficult to carry out on a systematic basis. This procedure is illustrated through its application to remote rural settlements in southern Mali.

2 Methodology

In this paper we consider the stochastic framework of geostatistics in order to optimize sampling protocols during field surveys. A given number of experimental data can be seen as a realization of a random function $Z(u)$, which is only observed at a finite set of experimental locations defining a set of n random variables $\{Z(u_i); i = 1, \ldots, n\}$. We will be dealing with two-dimensional problems, where the data have spatial coordinates projected on the plane. Thus $u_i = \{x_i, y_i\}$ represents the coordinates of the i-thm datum in some adequate reference system. A common problem is to obtain a continuum map through optimal spatial interpolation by using, for example, some form of kriging (Olea 1999). Another important concept for optimizing a sampling network is entropy as defined by (Shannon 1948):

$$H(p) = -\mathrm{E}(\ln p) = -\int p(\xi)\ln(\xi)d\xi, \tag{1}$$

where $H(p)$: entropy of the distribution p. $p = N_n(\mu, \mathbf{C})$: multivariate normal distribution with mean nx1 vector μ and nxn covariance matrix C. n: is the number of experimental data.

Entropy provides a measure of uncertainty, while the purpose of sampling is to gain information, that is, to decrease uncertainty. Thus, an optimal sampling procedure is that which, for a fixed number of data points, provides the most information. In other words, an optimal sampling procedure will provide the greatest decrease in uncertainty, or, the least entropy. Assuming that the data follow a multivariate Gaussian distribution, entropy is given by (Bard 1974):

$$H(p) = \ln|\mathbf{C}| \tag{2}$$

where $\ln|\boldsymbol{C}|$ is the natural logarithm of the determinant of the covariance matrix.

The covariance matrix of the n experimental data is the nxn matrix whose generic ij element is given by $\mathrm{C}_{ij} = C(u_i, u_j)$, and the matrix itself is given by:

$$\mathbf{C} = \begin{bmatrix} C(u_1, u_1) & \cdots & C(u_n, u_1) \\ \vdots & \ddots & \vdots \\ C(u_1, u_n) & \cdots & C(u_n, u_n) \end{bmatrix} \tag{3}$$

The covariance matrix can be completed by using a covariance model that represents the phenomenon under study. Usually that covariance model is unknown and must be estimated using the experimental data. Thus, the covariance given in Eq. (3) can be calculated and the criteria of minimizing the entropy in Eq. (2) can be applied. The assumption of a multivariate normal distribution is not a requirement for the problem that we are dealing with. If the data are not multivariate normal but multivariate lognormal, the logarithmic transformation converts them to normal. Even if the data are neither normal nor lognormal, the assumption of the multivariate Gaussian distribution still provides appropriate results in many problems that require the assumption of a multivariate Gaussian distribution (Kitanidis 1997). Furthermore, if the variable is transformed to an indicator variable by using a specified threshold t:

$$I(u;t) = \begin{cases} 1 & if \quad Z(u) > t \\ 0 & \text{otherwise} \end{cases} \quad (4)$$

Then, although the indicator variable is clearly non-Gaussian, the multivariate normal distribution can still be assumed for indicator covariance estimation (Pardo-Igúzquiza 1998a, b) and for evaluating the uncertainty of estimated indicators (Pardo-Igúzquiza et al. 2006).

However, the criteria of variance reduction (Rouhani 1985) and maximizing entropy (Bueso et al. 1999) are a function of the spatial locations of the experimental data only, and not of the actual values of the variable. In this case, we are interested in obtaining a spatial sampling scheme that provides as much information as possible with regard to detecting changes: for example, the deterioration of water quality. This can be achieved by applying one or several thresholds to the original variable, like in Eq. (4), and by looking for neighbor couples of samples which have a different value. Thus, without loss of generality, one looks for the pairs of samples (u_i, u_j) such that:

$$\{I(u_i;t_k) \neq I(u_j;t_k)\} \quad (5)$$

This implies that the value of the variable Z at one of the locations is above the threshold and in the other location is below the threshold; thus, the border of the process defined by the threshold must be somewhere between the two locations. If the threshold t_k is chosen adequately, sampling at these two locations is very informative because it will be more likely to detect changes in contamination than if the two locations were simultaneously above or below the threshold. Thus the new information unit is the pair of neighbor locations with different indicator values. We apply the entropy method to these new information units. In order to include the gradient of the variable between the two locations there is the possibility of using multiple

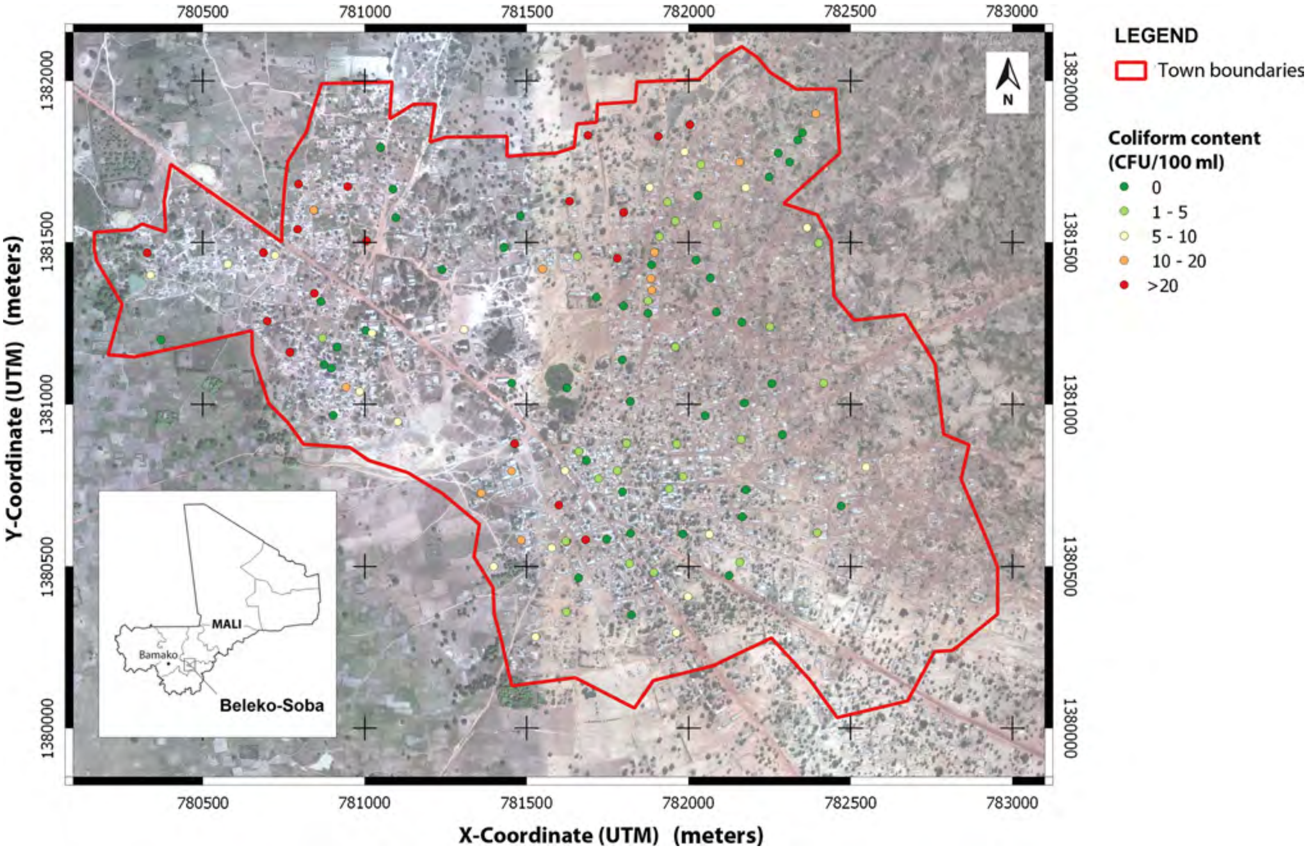
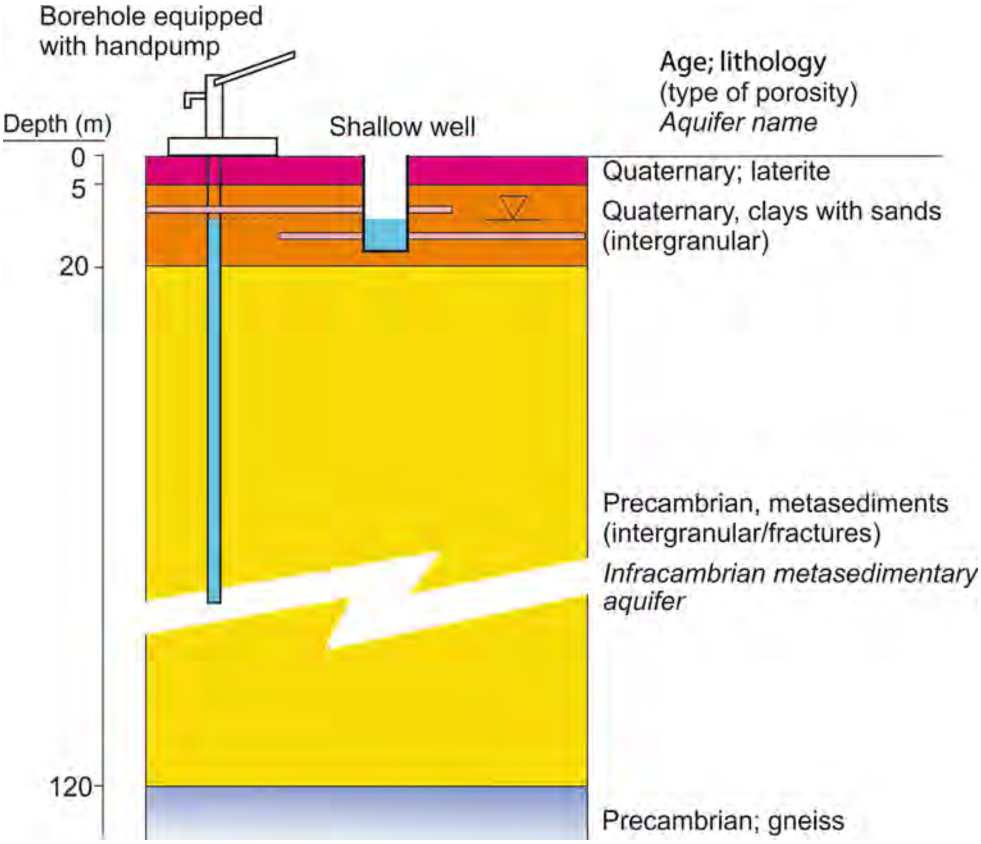


Fig. 1 Geographical location of the study area

Fig. 2 Geological profile of a typical well



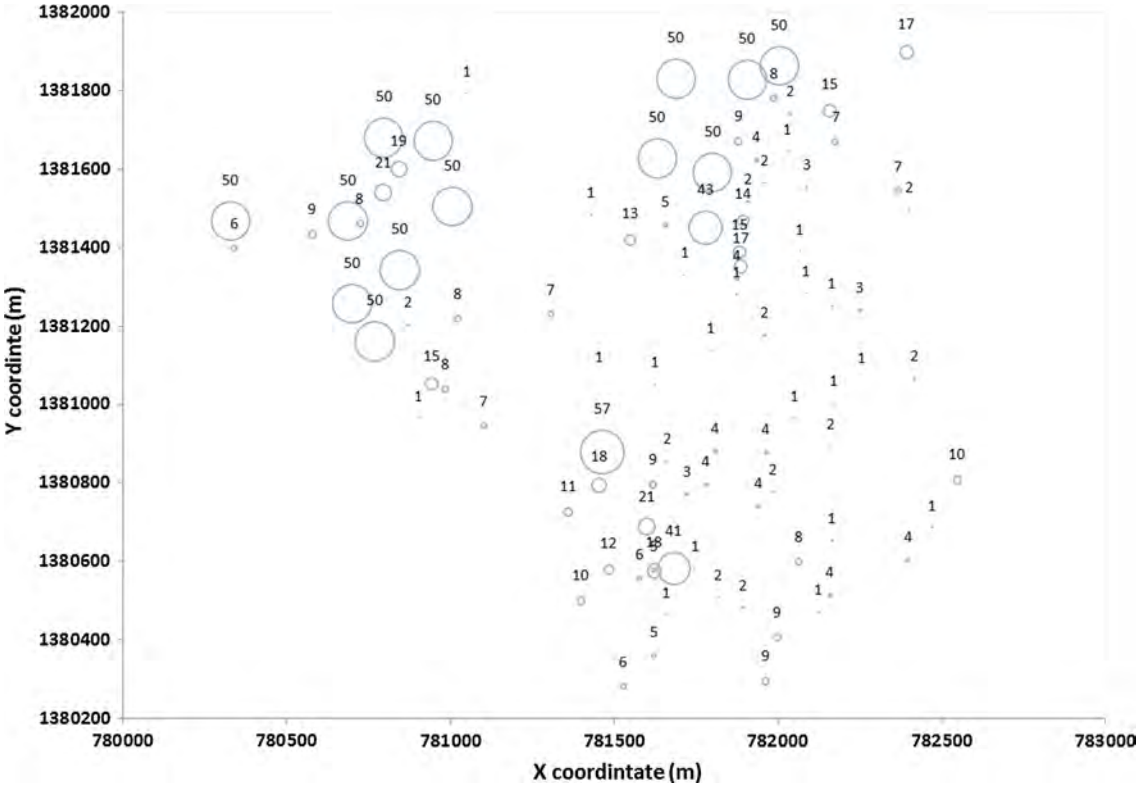


Fig. 3 Bubble plot of the experimental values of the thermotolerant coliforms variable

thresholds. This will become clear in the case study that is presented in the following sections.

Using indicators is a better alternative than working with the raw variable that has a skewed and censored distribution. This approach provides several advantages in variogram analysis, in working with the thresholds of interest and in avoiding transformation of the skewed and censored raw variable.

3 Case study

3.1 Study site

Fieldwork was carried out in Beleko-Soba, the main village in the rural commune of Djedougou, southern Mali (Fig. 1). Beleko-Soba is located approximately 200 km to the East of Bamako and is home to 6000 people. The region features a hot tropical climate, with the average yearly temperature standing at 26 °C. Rainfall patterns are typical of the West African monsoon. Average precipitation amounts to 800 mm/yr, taking place almost exclusively between June and September. This configuration presents two practical implications for the purpose of this research. The first one has to do with physical accessibility. Beleko-Soba can only be reached by a dirt track that is left in poor condition during the wet season. This means that monitoring drinking water supplies is complicated and that there is a need to devise sampling protocols to ensure that, if conditions allow, the process can be carried out as quickly and effectively as possible. The second implication is the absence of permanent surface water courses during roughly two-thirds of the year, which results in the local population relying exclusively on groundwater.

Groundwater is accessed either through community boreholes equipped with hand pumps, public standpipes served by gravity distribution networks, or domestic wells. Domestic wells are particularly widespread, with over 80% of households owning at least one (Martínez-Santos 2017b; Martínez-Santos et al. 2017). These consist of shallow pits excavated using picks and shovels. In the study area, wells are typically less than 15 meters deep and their diameter usually ranges from one to two meters. Despite the relative abundance of community water sources, which are theoretically safer, wells are preferred by many people because they are cheap to construct and allow users to avoid potentially long trips to collect water (Martínez-Santos et al. 2017).

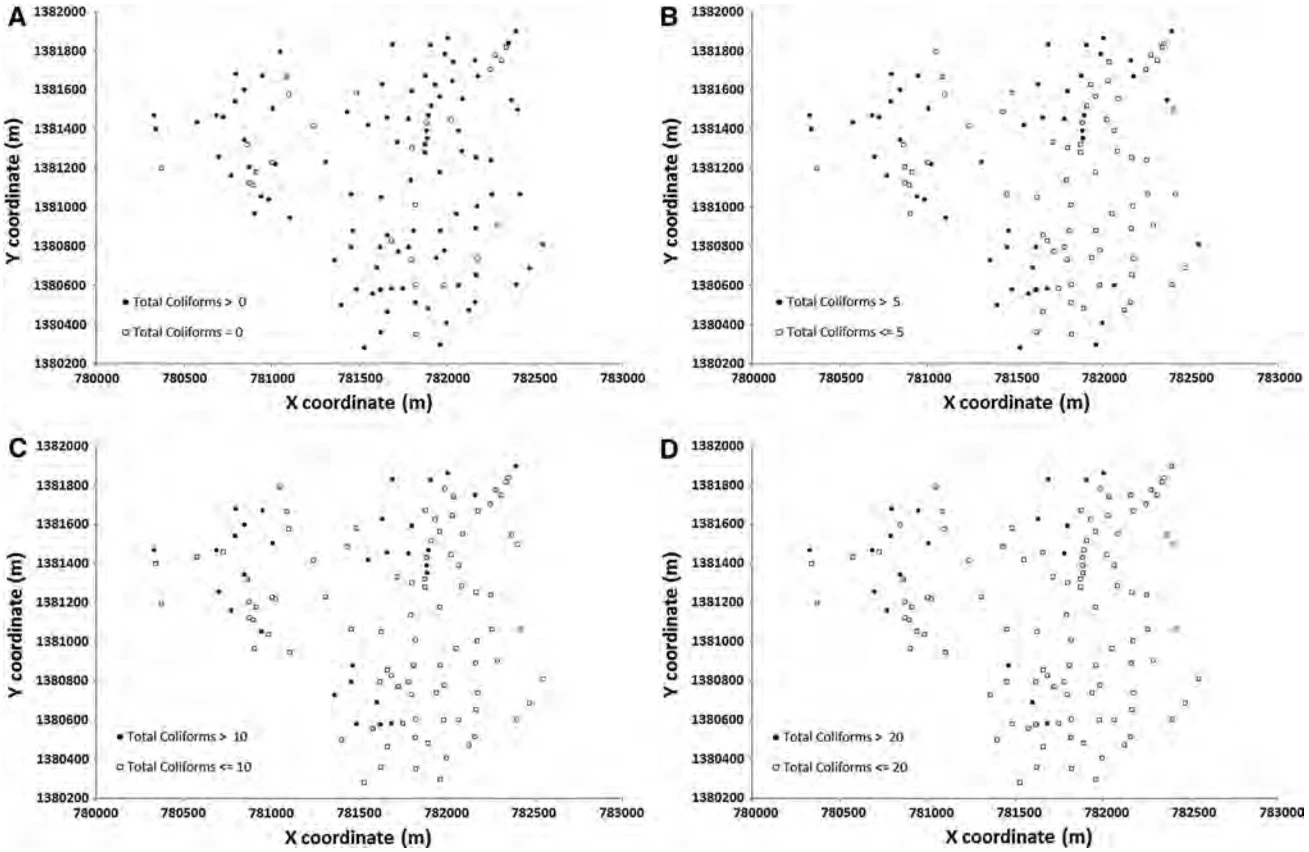


Fig. 4 Indicator variables (1/0) for the threshold of thermotolerant coliforms equal to **a** 0, **b** 5, **c** 10 and **d** 20 CFU/100

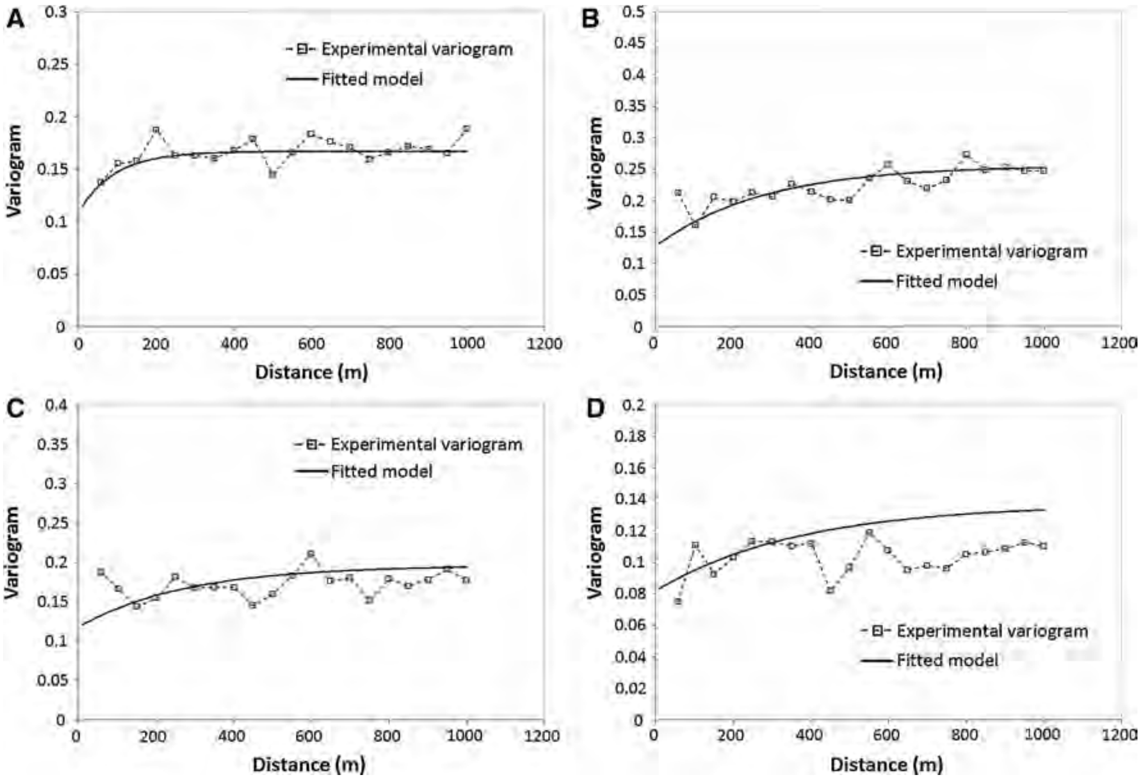


Fig. 5 Experimental indicator variograms and theoretical model fitted to them for the thresholds of thermotolerant coliforms equal to **a** 0, **b** 5, **c** 10 and **d** 20 CFU/100

Table 1 Variogram parameters of an exponential model fitted to the experimental variograms of different indicator variables according to different thresholds CFU/100 ml

	Nugget variance	Partial variance	Total variance	Range (m)
$I(u; t_1 = 0)$	0.109	0.058	0.167	92.60
$I(u; t_2 = 5)$	0.128	0.128	0.256	277.81
$I(u; t_3 = 10)$	0.118	0.078	0.196	277.81
$I(u; t_4 = 20)$	0.082	0.055	0.137	370.41

The parameters were estimated by maximum likelihood

From a hydrogeological standpoint, the study area is located within the metasedimentary Precambrian region of southern Mali. The available borehole information suggests a four-layer geological profile (Fig. 2). The uppermost layer is made up of a hard laterite crust whose thickness can exceed five meters in some areas. Immediately beneath, there is an unconsolidated layer made up of clays with intercalations of fine sand. This formation, which typically ranges between 10 and 15 m, lies on top of the regional sandstone aquifer, whose depth exceeds one hundred meters and which is underlain by a gneiss basement. At the village scale, groundwater flows from the south to the north, following the regional flow pattern. The water table depth ranges from five to 15 meters at the end of the dry season, and remains closer to the surface, i.e. one to three meters, during the rainy months.

Approximately 85% of households own pit latrines. These are built much like domestic wells and are generally unlined. This allows feces to come directly into contact with groundwater, particularly during the rainy season, when the water table rises close to the ground level. As demonstrated by the results of the field survey, fecal contamination in drinking water sources poses a major health threat to the population (Martínez-Santos et al. 2017).

3.2 Indicators and sampling procedures

Indicator bacteria such as coliforms are frequently used to estimate the microbial quality of drinking water supplies. The existence of these bacteria in water is not necessarily dangerous to human health, but it hints at the presence of viruses, protozoa and other parasites, which are more cumbersome to test and require more complex laboratory

Table 2 Group of data pairs (P1, P2) that are closest neighbors, fulfill condition (5) for the threshold 0 CFU/100 ml, and have the greatest difference between their values

P1	P2	Dif	
2	115	50	**
8	63	50	**
13	25	4	
14	53	15	*
15	50	4	
16	91	7	
17	18	15	*
24	112	43	**
27	23	14	*
46	99	4	
49	48	8	
51	50	4	
54	56	2	
57	74	50	**
67	66	9	
77	74	50	**
79	75	50	**
80	115	50	**
81	115	50	**
98	41	4	
100	45	10	
103	11	9	
106	53	15	*
107	74	50	**
119	117	50	**
121	3	8	

The values with differences greater than 30 (**) are the ones selected as candidates

Table 3 Group of data pairs (P1, P2) that are closest neighbors, fulfill condition (5) for the threshold 5 CFU/100 ml, and have the greatest difference between their values

P1	P2	Dif	
1	10	56	**
4	112	42	**
5	65	8	
7	92	17	*
20	19	7	
21	114	5	
22	113	48	**
25	87	13	*
28	86	14	*
30	91	5	
32	31	36	**
37	45	8	
44	45	9	
47	48	7	
52	12	18	*
56	63	48	**
68	31	40	**
69	31	40	**
70	71	1	
76	113	45	**
78	75	49	**
83	18	13	*
88	82	4	
97	10	55	**
101	48	4	
102	48	6	
104	53	14	*
108	65	7	
120	118	12	

The values with differences greater than 30 (**) are the ones selected as candidates

equipment. Hence, indicator bacteria such as thermotolerant coliforms are considered a suitable proxy for the likely presence of pathogens. Thermotolerant coliforms may occur naturally in the environment, but originate mostly in the digestive tract of warm-blooded animals. Among thermotolerant coliforms, *Escherichia coli* is a common indicator of fecal contamination, as it is universally present in large numbers in feces and does not grow in natural waters (Paruch and Mæhlum 2012). In most circumstances, populations of thermotolerant coliforms are composed predominantly of *E. coli* (WHO 2011; Sphere 2011; Hachich et al. 2012). Hence, thermotolerant coliforms as a whole are regarded as an acceptable indicator of fecal pollution (WHO 2011).

The presence of thermotolerant coliforms is measured in terms of the quantity of colony forming units (CFU) per 100 ml of water. International standards suggest that 0 CFU/100 ml is the only acceptable coliform count for drinking purposes (WHO 2011), though a concentration of 1–10 CFU of *E. coli* is sometimes considered tolerable (WHO 2002; UNICEF/WHO 2013).

Water was sampled at the 121 wells and public water sources in late May and early June 2016, during a 10-day survey (Fig. 3). As per standard procedure, all collected samples were refrigerated, kept in the dark and filtered prior to microbiological analyses. An Oxfam-Delagua portable kit was used to analyze the samples (Oxfam 2009). Cultures were prepared in petri dishes within 8 h of collection and then incubated at a constant temperature of 44 °C for 18 h. Colony forming units were counted within 10 min of retrieval from the incubator. Coliform contents in excess of 50 CFU were considered too numerous to count.

Table 4 Group of data pairs (P1, P2) that are closest neighbors, fulfill condition (5) for the threshold 10 CFU/100 ml, and have the greatest difference between their values

P1	P2	Dif	
9	62	42	**
11	10	48	**
19	115	42	**
33	94	6	
34	94	2	
55	95	4	
58	59	44	**
60	61	41	**
82	18	8	
91	85	10	
105	53	7	
114	113	41	**

The values with differences greater than 30 (**) are the ones selected as candidates

Table 5 Group of data pairs (P1, P2) that are closest neighbors, fulfill condition (5) for the threshold 20 CFU/100 ml, and have the greatest difference between their values

P1	P2	Dif	
23	113	36	**
86	112	28	*
92	10	39	**
93	31	23	*
94	12	9	
110	111	31	**

The values with differences greater than 30 (**) are the ones selected as candidates

4 Results

Intensive monitoring surveys such as this one are expensive, cumbersome to carry out and logistically complex, particularly during the wet season. Thus, the purpose of this study is to define the optimal locations to sample if field work is limited to fewer days. The distribution of coliforms is heavily skewed, with a minimum value of 0 and a maximum of 57. The percentiles 10, 25, 50, 75 and 90% are 0, 1, 3, 10 and 50, respectively. Taking into account that ideally drinking water should have 0 CFU and no more than 10 (WHO 2002, 2011; UNICEF/WHO 2013), the thresholds that have been considered were 0, 5, 10 and 20. The 0 and 10 marks are established according to these guidelines, while 5 has an intermediate value between 0 and 10, and 20 doubles the maximum tolerable limit. This value is used to characterize very high coliform counts. The indicator variables (0/1) using the thresholds 0, 5, 10 and 20 may be observed in Fig. 4. The distribution of black dots (values larger than the threshold) and white dots (values smaller than or equal to the threshold) shows the spatial variability of the variable and neighbor values with different indicator values are the locations close to the border that provide more information. The experimental variograms and the theoretical models fitted are shown in Fig. 5 and Table 1 shows where the models were estimated by maximum likelihood (Pardo-Igúzquiza 1997) and which model parameters were used.

The practical procedure is:

1. For each spatial location $\{u_i; i = 1, \ldots, n\}$, the four nearest neighbors (one for each quadrant) are calculated.
2. For each data pair (datum i and each nearest neighbor) Eq. (5) is verified for threshold $t_k = 0$.
3. From the pairs that fulfill Eq. (5), the one with highest difference between the variable (thermotolerant coliforms) is selected. If there are several data pairs with the same difference in the variable, the pair with the smallest distance (i.e. high gradient of the variable) is selected.
4. Repeat the previous steps for the other thresholds $t_k = 5, 10, 20$.
5. From all the data pairs selected, keep the ones of most interest, which are those that have a difference between the variables larger than a given amount (in our case 30, as explained below). This value can be set to zero if the gradient is not considered to be important.
6. The entropy of the number of final pairs selected is calculated and if the desired number of samples is smaller, that optimal network is obtained by sequentially eliminating the least informative sample so the remaining samples provide the greatest amount of information.

The results from applying the previous algorithm may be seen in Tables 2, 3, 4 and 5 and Fig. 6. A further restriction is that, of all the data pairs selected previously, only those where the difference between the data pair is larger than 30 are selected. With this restriction, obtained using Tables 2, 3, 4 and 5, by determining the difference that would leave enough data pairs, there are 29 data pairs of locations that will provide the most information and from which one can select the best network. This is done by sequentially eliminating the worst data pair, which is the one that results in the remaining data having the maximum entropy. The optimal entropy for data pairs 2–29 is shown in Fig. 7. This figure demonstrates how the uncertainty is reduced (or information is gained) as the number of data pairs increases. It is a relative measure that could be used to see how much data is required for the information to double. Thus, for example, the information provided by 10 data pairs is doubled by considering 14 data pairs. Finally, Fig. 8 gives the optimal sampling networks for 4, 10 and 16 locations (2, 5 and 8 data pairs), respectively. Their entropy can be seen in Fig. 7. The optimal sampling network could be obtained for any other number of locations.

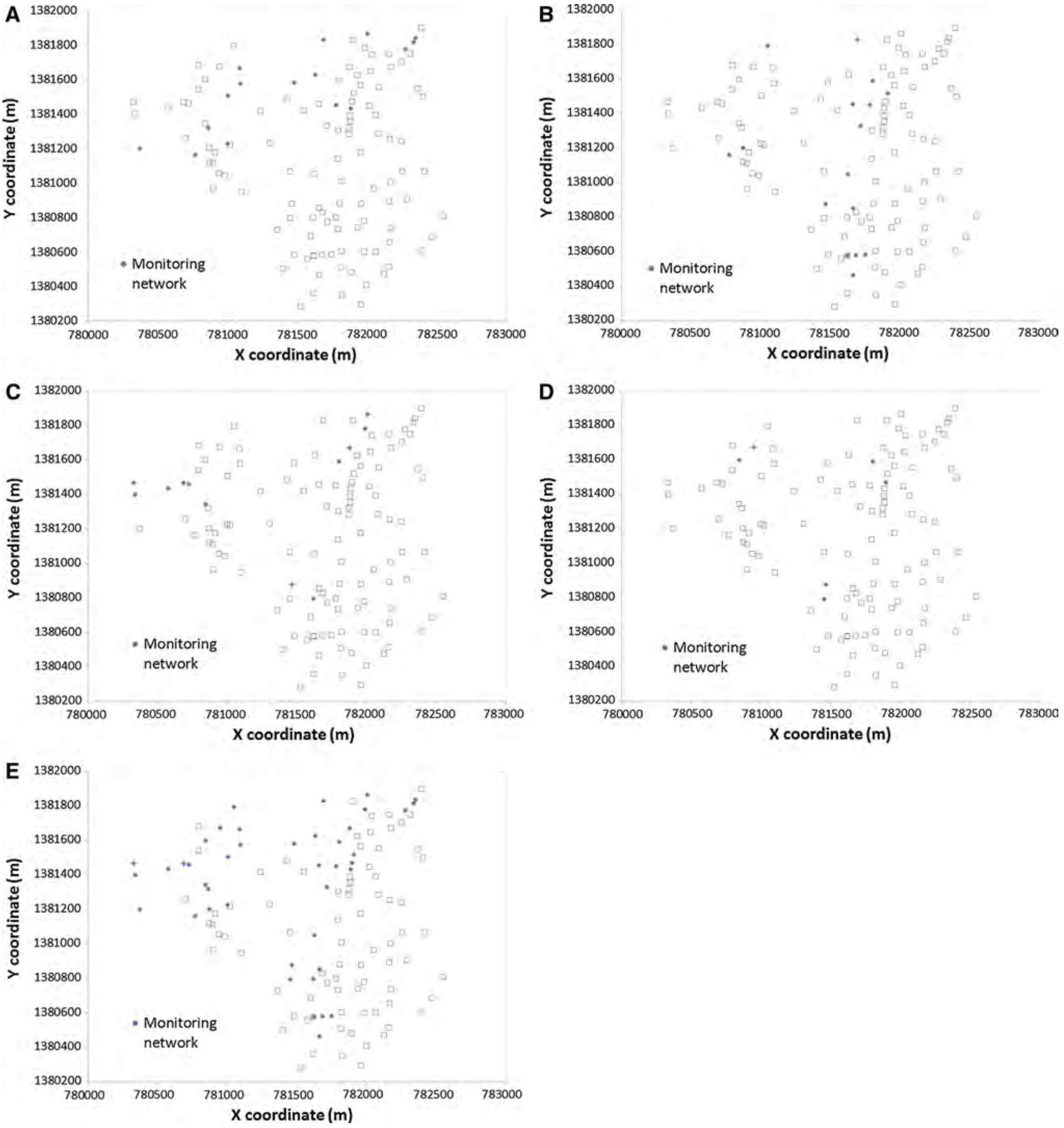


Fig. 6 **a** Most informative monitoring network (16 locations) according to the first threshold. **b** Most informative monitoring network (16 locations) according to the second threshold. **c** Most informative monitoring network (12 locations) according to the third threshold. **d** Most informative monitoring network (6 locations) according to the fourth threshold. **e** Most informative monitoring network (42 locations) taking into account the four thresholds

5 Discussion

The purpose of this method is to select the optimal subset from a number of experimental locations where the variable of interest was measured in an extensive survey that cannot be repeated on a consistent basis. It should be clear that the purpose of optimization is to obtain as much information as possible from a limited number of samples *m*, smaller than the complete network with *n* samples. The basic unit of information is a data pair (nearest neighbor at the quadrant level) and these data pairs are chosen from the border of the phenomenon defined by four thresholds: 0, 5,

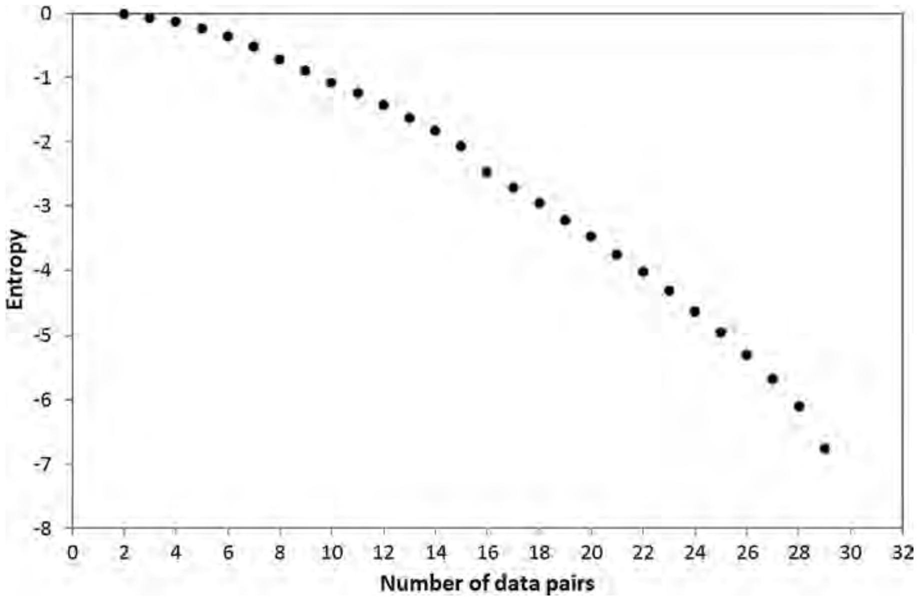


Fig. 7 Entropy as a function of the number of data pairs

10 and 20 CFU. The concept of entropy is then used to develop a network with the desired number of sampling locations. The nugget effect that can be seen in the variograms of the indicators in Fig. 5 implies that the border of the phenomenon is abrupt and that it cannot be perfectly delimited with a small number of samples. However, the proposed methodology guarantees that the number of samples selected will be taken at the optimal locations for providing the maximum information.

This approach has been demonstrated through a case study that deals with the quality of drinking water supplies in remote regions. Within this context, it is recognized that thermotolerant coliforms are just one of many relevant water quality indicators. Coliforms have been considered sufficiently representative to illustrate the method because their presence in drinking water supplies suggests an immediate threat to public health. Nonetheless, further work would be needed to develop alternatives for optimizing the number of sampling points based on several relevant quality variables.

While a thorough initial survey will always be required, this method allows for the optimization of ongoing monitoring efforts, thus leading to an effective use of human and material resources. Moreover, it can be extrapolated to a variety of engineering and environmental applications. In cases such as the one at hand, where a field laboratory can carry out up to sixteen tests per day, the outcomes can be easily translated into time and cost estimates. This makes the method immediately useful for planning purposes.

In practice, the choice of variables will be case-specific and will need to be informed both by technical judgement and common sense. To continue with the example at hand, fecal contamination of drinking water supplies is a highly sensitive issue because of its potential influence on human health. This means that the optimization of field surveys by statistical means may not be acceptable under certain circumstances. Consider, for instance, the case of multiple isolated populations within a large rural area. In such a situation, restricting the number of samples may not be a sensible course of action because contamination could be constrained by local factors. In other words, the risk of overlooking potential threats to human health would be too significant. On the other hand, optimizing the number of samples in large neighborhoods where sampling every well on a systematic basis is physically unfeasible may provide an appropriate alternative to delineate the sectors where fecal contamination is more persistent. This information could be used, for instance, to prioritize where to carry out activities to raise awareness among the local population or where to build improved water supplies.

6 Conclusions

Access to safe water supplies remains an important challenge for local communities in developing countries, where the absence of facilities and qualified technicians hampers adequate water quality monitoring. Within this context, geostatistical approaches may provide cost-effective ways to optimize sampling protocols, thus delivering a practical solution to practitioners and contributing to improving people’s living conditions.

A new method for the design of optimal sampling networks has been introduced in this paper. The method uses the new concept of the data pair as a basic information unit, where each datum in the data pair has a different indicator

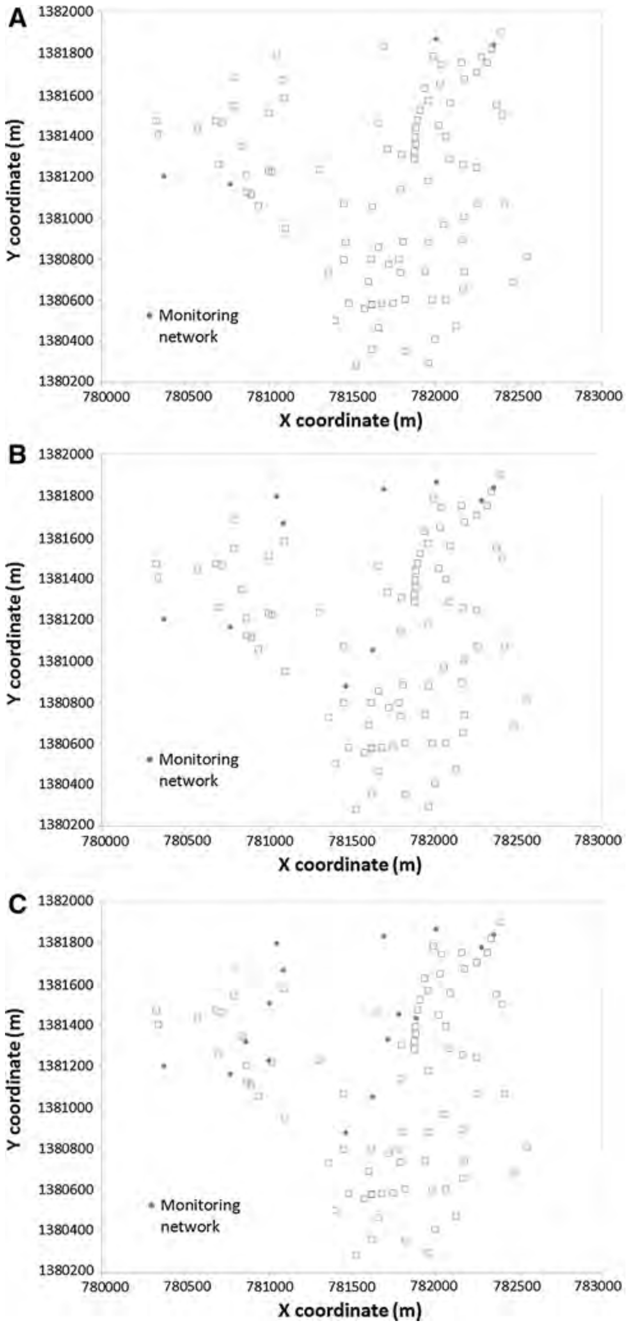


Fig. 8 **a** Optimal network with 4 sampling points. **b** Optimal network with 10 sampling points. **c** Optimal network with 16 sampling points

value according to a series of selected thresholds. A second restriction is that the difference between the variable of interest (thermotolerant coliforms in this case) has a value larger than a given threshold. This is done in order to restrict the data pairs to the most informative ones. The border is between the data pair and the large gradient implies a more probable expected change in the future. Finally, the concept of entropy was used to ensure that the data pairs from the most informative locations could be selected for any given number of desired data.

Acknowledgements This research has been funded by the Agencia Española de Cooperación al Desarrollo (AECID), Under Grants 2014/ACDE/005226 and 2016/ACDE/001953. For their time and support, the authors would like to thank Geólogos Sin Fronteras (Geologists Without Borders). We would like to thank the reviewers for their comments that have helped to improve the final version of this paper.

References

Adelana SMA, MacDonald AM (2008) Groundwater research issues in Africa. In: Applied groundwater studies in Africa. IAH selected papers on hydrogeology, vol 13. CRC Press/Balkema, Leiden

Angulo JM, Bueso MC, Alonso FJ (2000) A study on sampling design for optimal prediction of space–time stochastic processes. Stoch Environ Res Risk Assess 14:412–427

Angulo JM, Madrid AE, Ruiz-Medina MD (2011) Entropy-based correlated shrinkage of spatial random processes. Stoch Environ Res Risk Assess 25:389–402

Bard Y (1974) Nonlinear parameter estimation. Academic Press, New York

Barry B, Obuobie E (2012) Mali. In: Pavelic P, Giordano M, Keraita B, Ramesh V, Rao T (eds) Groundwater availability and use in Sub-Saharan Africa: a review of 15 countries. International Water Management Institute (IWMI), Colombo. https://doi.org/10.5337/2012.213

Bueso MC, Angulo JM, Cruz-Sanjulián J, García-Aróstegui JL (1999) Optimal spatial sampling design in a multivariate framework. Math Geosci 31(5):507–525

Calow RC, MacDonald AM, Nicol AL, Robins NS (2010) Ground water security and drought in Africa: linking availability, access, and demand. Ground Water 48(2):246–256

Chen S, Hong X, Harris CJ (2003) Sparse kernel regression modeling using combined locally regularized orthogonal least squares and D-optimality experimental design. IEEE Trans Autom Control 48(6):1029–1036

Douaik A, van Meirvenne M, Tóth T, Serre M (2004) Space–time mapping of soil salinity using probabilistic Bayesian maximum entropy. Stoch Environ Res Risk Assess 18:219–227

Foster S, Garduño H (2013) Groundwater-resource governance: are governments and stakeholders responding to the challenge? Hydrogeol J 21(2):317–320

Hachich EM, Di Bari M, Christ APG, Lamparelli CC, Ramos SS, Sato MIZ (2012) Comparison of thermotolerant coliforms and *Escherichia coli* densities in freshwater bodies. Braz J Microbiol 43(2):675–681

He J, Kolovos A (2017) Bayesian maximum entropy approach and its applications: a review. Stoch Environ Res Risk Assess. https://doi.org/10.1007/s00477-017-1419-7

Kitanidis PK (1997) Introduction to geostatistics: applications to hydrogeology. Cambridge University Press, Cambridge

Llamas MR, Martínez-Santos P (2005) Intensive groundwater use: silent revolution and potential source of social conflict. ASCE J Water Resour Plan Manag 131(5):337–341

MacDonald AM, Bonsor HC, Dochartaigh BEÓ, Taylor RG (2012) Quantitative maps of groundwater resources in Africa. Environ Res Lett 7:024009

Martínez-Santos P (2017a) Does 91% of the world’s population really have “sustainable access to safe drinking water”? Int J Water

Resour Dev 33(4):514–533. https://doi.org/10.1080/07900627.2017.1298517

Martínez-Santos P (2017b) Determinants for water consumption from improved sources in rural villages of southern Mali. Appl Geogr 85:113–125. https://doi.org/10.1016/j.apgeog.2017.06.006

Martínez-Santos P, Martín-Loeches M, García-Castro N, Solera D, Díaz-Alcaide S, Coulibaly B, García-Rincón J, Montero E (2017) Pit latrines, shallow wells and domestic-scale water treatment: a delicate balance in rural settlements of Mali. Int J Hyg Environ Health 220(7):1179–1189. https://doi.org/10.1016/j.ijheh.2017.08.001

Olea RA (1999) Geostatistics for engineers and earth scientists. Springer, New York

Oxfam (2009) Oxfam Delagua portable water testing kit. User manual. University of Surrey, Guildford

Pardo-Igúzquiza E (1997) MLREML: a computer program for the inference of spatial covariance parameters by maximum likelihood and restricted maximum likelihood. Comput Geosci 23(2):153–162

Pardo-Igúzquiza E (1998a) Optimal selection of number and location of rainfall gauges for areal rainfall estimation using geostatistics and simulated annealing. J Hydrol 210(1–4):206–220

Pardo-Igúzquiza E (1998b) Inference of spatial indicator covariance parameters by maximum likelihood using MLREML. Comput Geosci 24(5):453–464

Pardo-Igúzquiza E, Dowd PA (2005) Multiple indicator cokriging with application to optimal sampling for environmental monitoring. Comput Geosci 31(1):1–13

Pardo-Igúzquiza E, Grimes DIF, Teo C (2006) Assessing the uncertainty associated with intermittent rainfall fields. Water Resour Res 42(W01412):1–13

Paruch AM, Mæhlum T (2012) Specific features of *Escherichia coli* that distinguish it from coliform and thermotolerant coliform bacteria and define it as the most accurate indicator of faecal contamination in the environment. Ecol Ind 23(2012):140–142

Pavelic P, Giordano M, Keraita B, Ramesh V, Rao T (2012) Groundwater availability and use in Sub-Saharan Africa: a review of 15 countries. International Water Management Institute (IWMI), Colombo. https://doi.org/10.5337/2012.213

Rouhani S (1985) Variance reduction analysis. Water Resour Res 21(6):837–846

Shannon CE (1948) A mathematical theory of communication. Bell Syst Tech J 27(3):379–423

Sphere (2011) Humanitarian Charter and minimum standards in humanitarian response. The Sphere Project, Rugby. ISBN 978-1-908176-00-4

UNICEF/WHO (2012) Progress on drinking water and sanitation: 2012 update (Report). UNICEF and World Health Organization, New York

UNICEF/WHO (2013) WASH targets and indicators post-2015: outcomes of an expert consultation. UNICEF and World Health Organization, New York

Wang H, Harrison KW (2013) Bayesian approach to contaminant source characterization in water distribution systems: adapative sampling framework. Stoch Environ Res Risk Assess 27:1921–1928

WHO (2002) Environmental health in emergencies and disasters: a practical guide. World Health Organization, Geneva. ISBN 92-4-154541-0

WHO (2011) Guidelines for drinking-water quality. Technical report. World Health Organization, Geneva. ISBN 978-92-4-154815-1

Wibrin MA, Bogaert P, Fasbender D (2006) Combining categorical and continuous spatial information within the Bayesian maximum entropy paradigm. Stoch Environ Res Risk Assess 20:423–433

Zeng X, Wang D, Wu J (2012) Sensitivity analysis of the probability distribution of groundwater level based on information entropy. Stoch Environ Res Risk Assess 26:345–356