What are the main factors influencing the presence of faecal bacteria pollution in groundwater systems in developing countries?

Núria Ferrer[1,2], Albert Folch[1,2], Guillem Masó[3], Silvia Sanchez[1,2], Xavier Sanchez-Vila[1,2]

1- Department of Civil and Environmental Engineering, Universitat Politècnica de Catalunya, Jordi Girona 1-3, 08034 Barcelona, Spain.
2- Associated Unit: Hydrogeology Group (UPC-CSIC), Barcelona, Spain.
3- Instituto Pirenaico de Ecología (IPE-CSIC), Av. Ntra. Sra. Victoria 16, 22700 Jaca, (Huesca), Spain

Abstract

Groundwater is the major source of drinking water in most rural areas in developing countries. This resource is threatened by the potential presence of faecal bacteria coming from a variety of sources and pollution paths, the former including septic tanks, landfills, and crop irrigation with untreated, or insufficiently treated, sewage effluent. Accurately assessing the microbiological safety of water resources is essential to reduce diseases caused by waterborne faecal exposure. The objective of this study is to discern which are the most significant sanitary, hydrogeological, geochemical, and physical variables influencing the presence of faecal bacterial pollution in groundwater by means of statistical multivariate analyses. The concentration of *Escherichia coli* was measured in a number of waterpoints of different types in a rural area located in the coast of Kenya, assessing both a dry and a wet season. The results from the analyses reaffirm that the design of the well and their maintenance, the distance to latrines, and the geological structure of the waterpoints are the most significant variables affecting the presence of *E. coli*. Most notably, the presence of faecal bacteria in the study area correlates negatively with the concentration of ion Na^+ (being an indirect indicator of fast recharge in the study site), and also negatively with the length of the water column inside the well.

Keywords: *E. coli,* groundwater, Kenia, multivariate statistics, principal component analysis

1. Introduction

Worldwide, human populations rely heavily on groundwater as a drinking water source. This situation is even more significant in Asia and Africa, where groundwater is the major source of drinking water and has an important role in improving health and sustaining urban livelihoods (Adelana and MacDonald, 2008; MacDonald et al., 2012). Although groundwater is typically assumed to be free of bacterial pathogens, surveys carried out during the last decades indicate that a significant fraction of groundwater supply sources are responsible for water-borne diseases outbreaks around the world (e.g., Bhattacharjee et al., 2002). Globally, 25% of people lack access to water free from microbial contamination (Nowicki et al., 2019). In Africa, this figure doubles, to a value above 50% (Bain et al., 2014), far from compliance with the Sustainable Development Goal number 6 of the United Nations.

Hand-pumped tube-wells, being low-cost and low-tech efficient solutions, offer affordable access to shallow aquifers in many developing countries across Africa, Asia and the Pacific. These type of wells, most generally operated by families or small rural communities, are a valid alternative to private or governmentally-operated deep boreholes (Ferguson et al., 2012a). However, they are susceptible to faecal contamination, arising from a variety of sources, such as septic tanks or latrines infiltration, improper disposal of solid urban wastes, leachate from landfills, anthropogenic controlled water recharge, or crop excess irrigation with untreated or insufficiently treated sewage effluent (Charles et al., 2008; Goyal et al., 1984; Matthess et al., 1988; Oteng-Peprah et al., 2018; Yates et al., 1985).

Once bacteria reach the groundwater, and under very favourable conditions with respect to flow, geochemistry and lack of competing indigenous biomass, bacterial pathogens can eventually travel considerably long distances (Sharma and Srivastava, 2011). Groundwater transport in shallow aquifers is primarily a function of the hydrogeological setting and climate conditions (Macler and Merkle, 2000). It is known that the transport, rate of survival, and fate of microbes in the subsurface environment are directly influenced by the microbial population (both diversity and individual characteristics and concentrations, e.g., Barba et al., 2019a), the microbes physical state (dead or alive), the type and characteristics of the subsurface soil and aquifer sediment, and the hydrological conditions, such as water temperature and quality (Rao et al., 1986; Perujo et al., 2017). Therefore, in order to protect drinking water supply wells against microbial contamination, it is essential to establish safe setback distances between wastewater disposal services and water wells (Blaschke et al., 2016). Following numerous laboratory and field-based studies, these safe setback distances should be defined as a function of local soil parameters (e.g., grain sizes, Knappett et al., 2012a, and angularity, Saiers and Ryan, 2005), and general hydrogeology conditions (e.g., Knappett et al., 2008, Pang 2009).

Understanding the mechanisms of bacterial fate and transport in the subsurface is of great importance to control soil and groundwater pollution (Carles-Brangari et al., 2017,2018; Sepehrnia et al., 2018a). Recent studies focus in understanding the role of the vadose zone in the flow and transport of Escherichia coli (*E. coli*) through the soil until reaching the shallow water table in unconfined aquifers (Sepehrnia et al., 2018b; Weldeyohannes et al., 2018). Yet, this should be completed with the detailed analysis of the impact of design, construction, and maintenance of individual wells. As an example, Kilungo et al., (2018) compares the water quality of samples from wells of different designs, in order to help guiding future efforts in providing affordable and sustainable interventions to improve access to clean and safe water in rural communities without centralized supply and sewage networks. Other authors (e.g., Olajuyigbe et al., 2017) examine some relevant socio-economic characteristics of population, such as gender, age, household size, family size, employment, and average income, in order to capture information about the exposure of hand-dug wells to pollution and contamination. Devane et al. (2018) reviewed different tracking tools to recommend the suitable method to determine faecal sources in rural areas. Finally, some studies try to correlate the temporal variation in *E. coli* concentrations as a function of seasonal rainfall characteristics (Elangovan et al., 2018; Guy Howard et al., 2003; Kayembe et al., 2018), well depths, distance to a septic tank, and population density (Dayanti et al., 2018; Martínez-Santos et al., 2017; Rohmah et al., 2018).

Some authors investigated the correlation between the presence of faecal bacteria and diverse sanitary risk factors, with the objective of assessing the microbiological risk posed by groundwater sources (Ercumen et al., 2017; Godfrey et al., 2006; Lin et al., 2018). Combining together hydrogeological and non-hydrogeological variables within the same study is quite rare (Ferguson et al., 2012b; Knappett et al., 2012b; Leber et al., 2011; van Geen et al., 2011), and to our knowledge, there is no study that combines different type of variables with the goal of screening the variables that are actually correlated and ranked to eventually predict faecal pollution concentrations. Therefore, the main goal of this paper is to discern what are the hydrological, geochemical, physical, and sanitary variables potentially influencing the presence of faecal bacterial pollution in groundwater sources in rural areas. The method proposed is based on performing a number of multivariate statistics evaluations, being tested in the coastal aquifer located in Kwale (Kenia), one of the multiple zones along the African continent heavily affected by bacterial pollution (Mzuga et al., 1998; Nowicki et al., 2019; Tole, 1997). The analyses involved shallow aquifers of very different geologies and hydrochemical facies, as well as different types of waterpoints in terms of construction and maintenance. Understanding which variables are affecting, and to what degree, the presence of *E. coli* in the groundwater sources, could provide significant knowledge for an accurate management of land uses and water resources to avoid faecal contamination to population. Faecal pollution is the source of a combination of sanitary and educational problems that are perpetuating gender inequality and poverty in rural areas in developing countries.

2. Methods

2.1 Study area

The study area is located in Kwale County, a rural coastal area in South-East Kenya, near the border with Tanzania (Fig. 1). The area is populated by small communities spread from the Indian Ocean coast to the Shimba Hills range. The economy of these communities is mainly based in self-consumption livestock. There is no wastewater treatment, and the basic sanitation facilities in the area are pit latrines. The communities are supplied by diverse type of groundwater points (WP) that can be classified in four groups: (1) hand-dug wells (large-diameter wells, less than 30 meter deep, and frequently uncovered),

(2) hand-dug wells equipped with handpumps (also large-diameter wells, less than 30 meter deep), (3) handpump boreholes (small diameter, less than 30 meter deep, with a concrete cover on the surface), and (4) deep boreholes (small diameter, with depths exceeding 30 m).

The study area spans three geological units and two hydrological systems: a deep aquifer composed by quartz-feldspar sandstones with fossil wood horizons in the lower section constituting the Mazeras Formation, and a shallow aquifer composed by young geological materials, these including Pliocene Magarini sands (dominantly quarzitic, and hosting the heavy mineral-rich sands), Kilindini sands (mainly composed of limestone), and coral reef of the Pleistocene karstified limestone. The hydrochemical facies and the water isotopic composition indicate hydraulic connectivity across the materials that comprise the shallow aquifer (Ferrer et al., 2019). The same data show that the Mazeras sandstones in the Shimba Hills are hydraulically connected with the deep aquifer. Because of this inhomogeneous geological setup, the hydrochemical composition of the groundwater sampled at the wells has a distinctive signature depending on the corresponding geological formation.

The area is characterised by a bimodal rainfall pattern. In Kenya, the “long rains” generally fall from April to June, whilst the “short rains” occur between October and December (CWSB, 2013). The driest months are from January to March (Fig. 2).

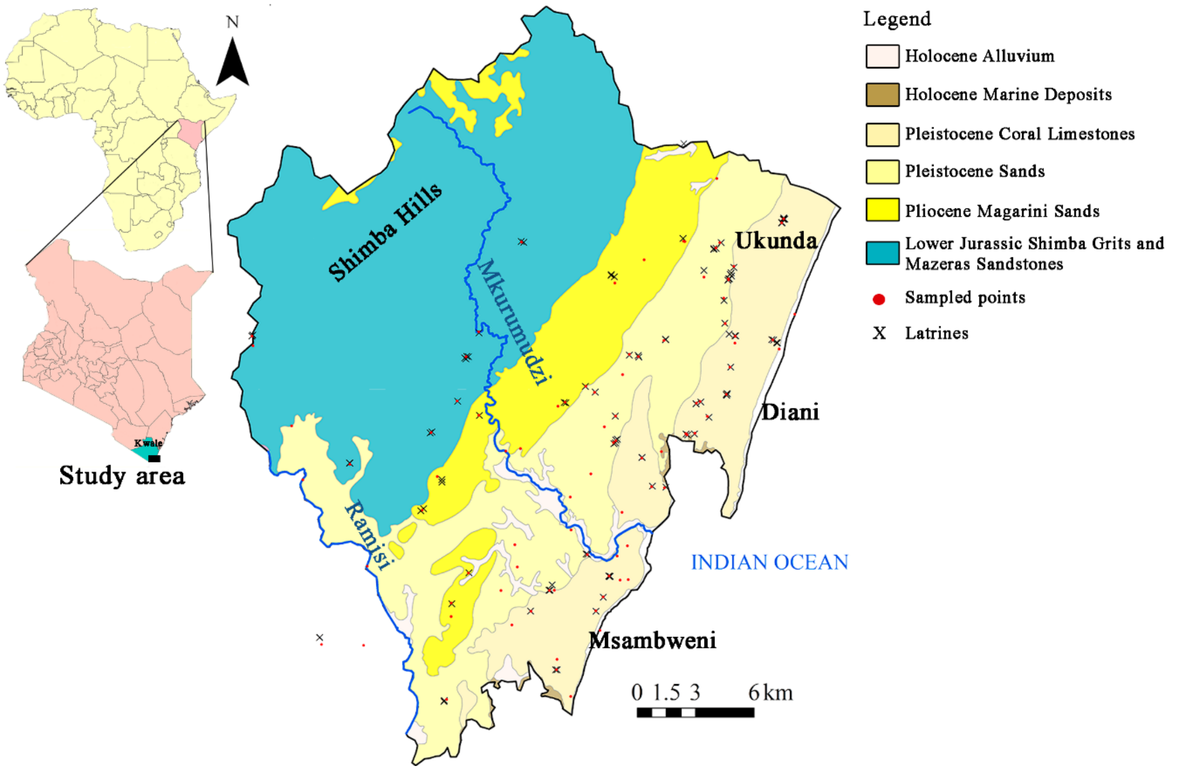

Figure 1. Plain view of the study area with the geological units outcropping. The location of the sampled points (red circles) and those of the pit latrines (black crosses) are displayed.

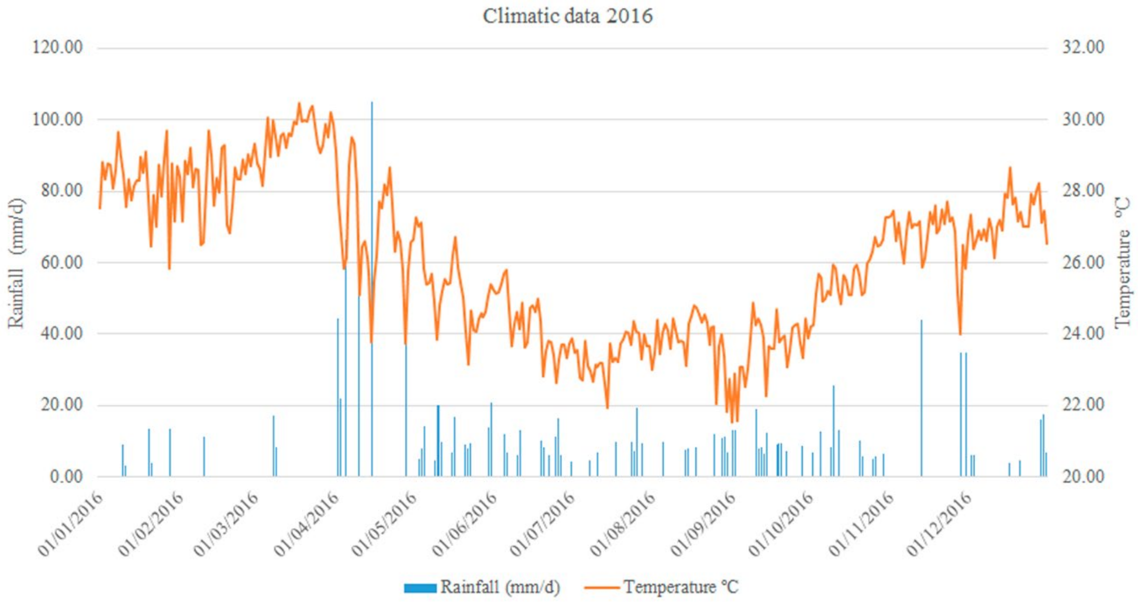

Figure 2. Rainfall and Average daily temperature data in 2016 from a weather station located in the Shimba Hills .

Despite the population density has been reported as a significant variable for faecal bacteria spatial characterization in other parts of the world (Knappett et al., 2011; van Geen et al., 2011), the information about the density of each community in the study area is missing.

2.2. Water sampling

Two sampling campaigns were carried out in March 2016 (end of the dry season) and June 2016 (end of the wet season) to measure several hydrochemical and bacteriological parameters under two very different climate conditions. During the field surveys, the number of sampling points were 78 (March) and 77 (June), here including waterpoints from all four groups presented before. In particular, all deep borehole sampled were on the range of 30 to 80 m depth. In addition, the main rivers in the study area, Mkurumudzi and Ramisi (Fig. 1), were also sampled in both campaigns.

Samples for hydrochemical analysis were taken from wells used daily by the population. The sampling protocol differed depending on the water point characteristics. In the boreholes or wells equipped with a handpump, a plastic tube was connected to the outlet, and the internal mechanism was flooded before sampling to avoid air contact. We ensured that at least three casing volumes of groundwater were removed by the handpump before sampling. Hand-dug wells were sampled using an electrical pump whenever the height of the water column allowed it, or using a plastic bucket as the last option. In completely closed deep boreholes connected to a tank, a pipe was connected to the flow cell and to the tank entry.

Bacteriological samples were taken using the same methodology just explained, except in those points in which a bucket was needed. In those cases, a stainless steel bucket previously sterilized with ethanol was used. In the waterpoints with handpumps, samples were taken at the outlet point, cleaned with ethanol before sampling was performed. All the points eventually used in the statistical analyses were sampled from 8h to 15h, the time rate that the communities in the area pump water, using an integrated sample by getting some volumes at different times of the day. Furthermore, in 4 specific points we analyzed the faecal bacteria evolution through a day, by sampling the same well at different hours; however, we could not establish a clear temporal pattern, since some points showed more pollution at early hours and other at the latest hours of the day, and so the daily evolution was not eventually explored.

2.3. Physicochemical parameters and ion analyses

The methodology to measure the physicochemical parameters and the diverse ion analyses is described in detail in Ferrer et al. (2019).

2.4. Bacteria concentration determination

Concentrations of *E. coli* were determined using Aquagenx Compartment Bag Test (CBT) (Aquagenx, 2015). CBTs allow for a quantitative assessment of *E. coli* concentration based on a most probable number (MPN) along with an upper 95% confidence interval (Foster and Willetts, 2018; Gronewold et al., 2017; Stauber et al., 2014). MPN testing involves multiple presence/absence tests on different volumes of the same sample. Samples were collected in sterile purpose-made bags, stored in a fridge during their transport, and processed within 24h, 30h or 48h after collection, depending on temperature recommendations by the manufacturer (Stauber et al., 2014). MPN was calculated with data supplied by the manufacturer, here enclosed as Table 1, and based on the World Health Organization "Guidelines for Drinking Water Quality" 4th Edition, assigning risk categories of drinking water to *E. coli* level ranges.

Table 1. E. coli risk categories of drinking water (modified from Aquagenx, 2015), and values assigned for the statistical analyses.

Sampled volume with colour changed	Risk categories	Value assigned for the statistical analyses
0/100 ml	Safe	0
1-10/100ml	Intermediate risk	1
11-100/100 ml	High risk	2
>100/100 ml	Very High risk/Unsafe	3

2.5. Sanitary risk inspections

A questionnaire on sanitary risk factors was carried out based on Wright et al. (2013) at each groundwater point. It comprised 13 questions (see Table 2) that could be answered as Y/N. The first 10 questions were answered for all points, and the last 3 for hand-dug wells only.

Table 2. Questions related to value the sanitary risk factors according to Wright et al. 2013

Question 1	Does the cement floor extend more than 1.5 m from the well?
Question 2	Is there any ponding of water on the cement floor?
Question 3	Are there cracks in the cement floor which could permit water to enter the well?
Question 4	Is the pump loose where attached to the base, allowing water to enter the casing?
Question 5	Is the drainage channel cracked, broken or in need of cleaning?
Question 6	Do animals have access to within 10 m of the well?
Question 7	Are there any latrines within 10 m of the well?
Question 8	Are there any additional latrines within 30 m of the well?
Question 9	Are there any open water sources within 20 m of the borehole?
Question 10	Are there any uncapped wells within 30 m of the borehole?
Question 11	Is there any scattered waste within 30 m of the well?
Question 12	Is the cover of the well unsanitary?
Question 13	Is there any scattered waste inside the well?

2.6. Multivariate statistics analysis

Multivariate statistics is a suitable technique to treat big datasets involving different sorts of variables, from quantitative to categorical, and thus amenable to be used to combine biochemical, hydraulic, geological and external conditions (such as design, drilling characteristics. and maintenance) of water points (Barba et al., 2019b). Principal Component Analysis (PCA) is a multivariate statistics method which involves the analysis of a number of parameters or variables, revealing associations between them, known as (vario)factors or components. Analyses were performed using the IBM-SPSS software.

The PCA analyses were subjected to Orthogonal Varimax rotation (Thompson, 2004). This implies the rotation of the original system into the directions of largest variance in the dataset. Prior to the extraction of the factors, the Kaiser-Meyer-Olkin (KMO) and the Bartlett sphericity tests were conducted to assess the suitability of the existing data for factor analysis. KMO returns values between 0 and 1, and values >0.50 are considered suitable for factor analysis (Hair et al., 1995; Tabachnik and Fidell, 2007). The Bartlett sphericity test checks if the observed correlation matrix diverges significantly from the identity matrix. It should be significant ($p<0.05$) for factor analysis to be suitable.

2.7. Selecting variables for the statistical analyses

Since the objective was to establish the variables that best could explain the *E. Coli* concentration distribution, arising from a large number of variables, we used a methodology divided in two steps. First, we removed the main variable of interest, *E. Coli* concentration, from the set, in order to reduce the number of active variables that could be used later in the final analyses; this way, information redundancy is eliminated, and the most significant variables or parameters can be elucidated. The second step involved the introduction of the variable *E. Coli* concentration into the statistical analyses.

Statistical parametric methods (such as PCA) perform best when data follows a unimodal symmetric distribution (Paliy and Shankar, 2016). For this reason, some variables from the initial dataset were grouped, transformed and/or eliminated. Following Barba et al., (2019b), non-Gaussian hydrochemical variables were transformed to log concentrations, these being Alkalinity, Eh (a proxy for redox conditions), and the concentrations of SO_4^{2-}, Na^+, Cl^-, and SiO_2. On the other hand, TOC (Total Organic Carbon), DO (Dissolved Oxygen), and the concentrations of NO_3^- and NH_4^+, were added to the analysis as raw data without any transformation (mostly based on a trial and error basis). The most redundant, and therefore less informative geochemical variables, such as Mg^{2+}, Ca^{2+} and K^+, were disregarded due to the strong correlation with other hydrochemical elements. The discrete (also termed categorical) variables were transformed to continuous ones based on a logical structure, as indicated in Table 3. Due to the zero variability in the response in the questions 4 and 5 of the questionnaire (Table 2), these two questions were not included in the analysis, being statistically insignificant.

Table 3. Assigning categorical data to quantitative values to be included in the statistical analysis.

Variable	Weights	Value assigned	Justification
Geology	Pliocene sands Pleistocene sands Pleistocene sands /corals Pleistocene corals Sandstones.	1 2 3 4 5	According to the aquifer units composition based on the conceptual model described in Ferrer et al., 2019b.
Aquifer unit	Shallow aquifer Deep aquifer	0 1	
Type of well	Hand-dug well Hand-dug wells w/handpump Handpump borehole Deep borehole	1 2 3 4	Increasing from the simplest structure to the most complex one.
Sanitary risk factors (questionnaire. Table 2)	No Yes	0 1	Binary answer for each one of the questions

2.8. The use of generalized mixed models

To assess the variables that most significantly influence the presence of *E. coli*, generalized mixed models with Poisson error distribution (Bates et al., 2015) were used. First, an additional PCA was performed, including only the variables found significant in the final PCA, but removing the *E.coli* concentration variable. Then, the scores for the main variofactors were extracted for each observation. Finally, a generalized mixed models analysis was performed, where concentration of *E. coli* was included as a dependent variable, and the covariates included correspond to the main principal variofactors of the PCA runs. As repeated measures were taken at each waterpoint, "Sample ID" was modelled as a random factor. For all tests, the significance level was set at α = 0.05 (two-tailed test). Overdispersion was tested and eventually corrected by including the number of observation as a random factor (Broström and

Holmberg, 2011). The output of the generalized mixed models correspond to the significant correlation between *E.coli* and the variofactor. All runs were performed using R 3.5.1 (Team, 2018).

3. Results

3.1. *E. coli* quantification

33 of the 78 waterpoints sampled in March 2016 showed low-risk, meaning no *E. coli* colonies were detected; 5 waterpoints were classified as intermediate-risk, 12 as high-risk, and 28 were in the range very high-risk/unsafe. Samples from surface bodies (rivers) were classified as very high risk. In the June 2016 campaign, *E. coli* risk was measured in 77 waterpoints; 34 showed low-risk, 3 intermediate-risk, 15 high-risk, and 25 very high-risk/unsafe (see the Supplementary material).

From the 72 waterpoints sampled in both campaigns, in 9 (13%) *E. coli* risk reduced from the March to the June campaign; contrarily, in 5 (7%) of the points, the risk increased in that same period.

3.2. PCA results

Five different PCAs were carried out to evaluate all the information available and considering all type of wells/boreholes present in the study area (Table 4). The variables included in the first sets of PCAs to value to correlation of *E. coli* with all type of water points were: geology; aquifer unit, type of well, sanitary risk factors (Questions 1, 2, 6, 7, 8, 9, 10), field parameters (conductivity, pH, TOC, alkalinity, DO, Eh), hydrochemical parameters (NH_4, Cl, SO_4, NO_3, Na, Si), and seasonality (March or June campaigns). The closest latrine was considered for every waterpoint, even if several were found nearby. Surface water samples were not included in the analyses.

All the analyses in this first set are compiled in Table 4. Each analysis involves a different number of variables, as this was a consequence of several trials to find the number of variables that were significant, yet leading to a large value of the KMO test, meaning that the results were statistically significant. For each one of the final PCAs presented in Table 4, we indicate the extracted components, the variables involved in each component, the proportion of variance represented by each component, and the measure of sampling adequacy (KMO test value). A brief explanation about what means each group of correlated variables that form each component is also included for later discussion.

Table 4. Components extracted from the first set of PCA analyses (bold indicates negative correlation);–log indicates that log transformation was performed (geochemical variables); Q stands for "question" (from Table 2). For all PCA's the Bartlett sphericity test was significant (p<0.001). The variables involved in each component, the proportion of variance represented by each component, and the measure of sampling adequacy (KMO test value) are included.

Type of variables	# of variables	PCA number	Extracted components	% of variance	Total of variance	KMO Test Value	Indication of each component
Physicochemical parameters	14	PCA$_1$	C1: Geology, Log Cl, Log EC, Log Na, Log SO_4	25.65	72.74	0.69	Major ions and geological setup
			C2: Aquifer unit, Geology, Log Si	15.18			Aquifer unit
			C3:NH4, **Log Eh,** Log Alkalinity	11.38			Redox state
			C4: Date, DO	11.07			Oxygen as function of seasonality
			C5: NO3, TOC	9.46			Nitrate correlated with TOC
Sanitary risk factor+latrine data	12	PCA$_2$	C1: Q6, **Type of well**	17.41	63.62	0.51	Deep boreholes mainly from the industries have a fence
			C2: Q1, Q9, Q10	13.34			Unknown explanation
			C3: **Num. Latrines,** distance latrines	12.44			Presence and distance from latrines
			C4: **Q7**	10.49			Isolated variable representing a statistical component
			C5: Q2	9.94			Isolated variable representing a statistical component
Physicochemical + *E. coli*	7	PCA$_3$	C1: **Aquifer unit**, *E. coli*	21.32	62.44	0.50	Highest presence of *E. coli* in the shallow aquifer
			C2: Date, DO	21.18			Oxygen as function of seasonality
			C3: **log Eh**, Na	19.94			Redox state
Sanitary risk + *E. coli* +latrine data	8	PCA$_4$	C1: Q1, **Q8**, Type of well, ***E. coli***	26.31	59.55	0.60	*E. coli* concentration increases with presence of latrines nearby and poor well construction
			C2: Q10, Distance latrines	17.5			Unknown explanation
			C3:Q2, **Q7**	15.74			Unknown explanation
Physicochemical data + *E. coli*+Sanitary risk factor +latrine data	10	PCA$_5$	C1: Type of well, Aquifer unit, **Q8, *E. coli***	25.39	69.05	0.63	Main variables related to presence of *E. coli*
			C2: Log Na, Q1	15.74			Waterpoints located in the coastline have cemented floor
			C3: DO, log Eh, Num. Latrines	15.08			Eh partially depends on DO content
			C4: Q2	12.84			Isolated variable representing a statistical component
PCA$_5$ without *E. coli*	9	PCA$_{5.1}$	C1: Type of well, Aquifer unit, **Q8**	23.32	69.99	0.61	Main variables related to presence of *E. coli*
			C2: Log Na, Q1	17.49			Waterpoints located in the coastline have cemented floor
			C3: DO, log Eh, Num. Latrines	15.26			Eh partially depends on DO content
			C4: Q2	13.92			Isolated variable representing a statistical component

All type of groundwater points (wells/boreholes)

A first analysis, PCA_1 (see Table 4), was conducted in order to observe which physicochemical variables displayed high correlation and to exclude those which would make the subsequent PCAs (2 to 5) redundant or masked (thus reducing reliability). A significant result from this analysis is that dissolved oxygen changes with seasonality (component #4); this could be attributed to an increment in recharge of oxygenated water, with high DO values, during the wet season. PCA_2 was conducted in order to exclude the sanitary risk factors (from Table 2) that add no significant information in the subsequent analyses.

Once the two firsts PCAs were conducted, the most redundant variables were detected, then removed from the list of variables, and additional PCAs were performed adding the variable representing the concentration of *E. coli*. PCA_3 and PCA_4 (Table 4) thus contain the most relevant hydrochemical variables extracted respectively from PCA_1 and PCA_2, plus *E. coli* concentrations. In PCA_4, the first component indicates that *E. coli* concentrations correlate positively with the presence of latrines nearby (Q8), but negatively with the presence of cement floor (Q1, indicating improper construction). Notice that type of well and presence of cement floor were positively correlated, as virtually all handpumps are cemented.

The variables most correlated with the presence of *E. coli*, after PCA_3 and PCA_4, were selected to conduct PCA_5. It includes hydrochemical parameters, sanitary risk factors, latrine data and *E. coli* quantification, for a total of ten variables. Results indicate that all deep boreholes have pumps, and that the probability of faecal bacterial pollution increased with the presence of nearby latrines and with uncapped wells. Component #2 merges Na concentration and the presence of cemented floor around the well, as the latter is common in waterpoints situated near the coast line, with sea water intrusion influence. PCA_5 is represented in Figure 3 in a projection on the plane corresponding to variofactors 1 and 2 allowing a graphical depiction of some of the variables.

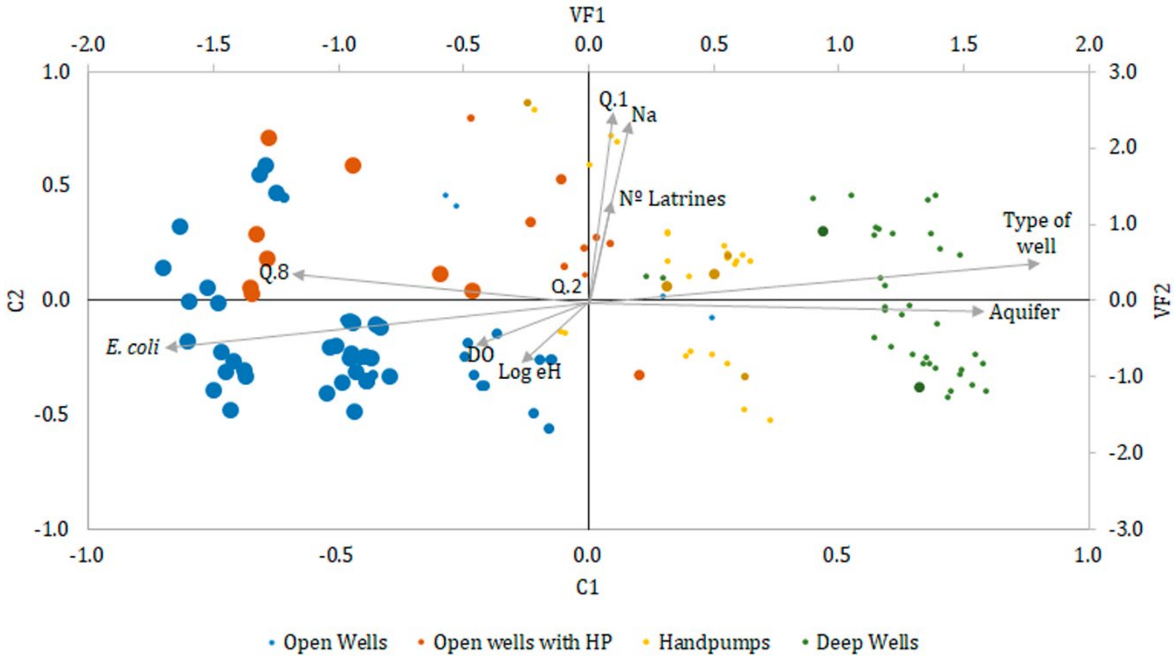

Figure 3. Graphical depiction of the results of PCA$_5$ projected in variofactor space (VF1 and VF2 axes). Position of samples is scaled for visualization purposes. Size of the points increase with E. coli measurements. Grey arrows represent the contribution of each variable projected into the variofactor plane, so that components can be easily identified.

Once the final correlation between hydrogeological and non-hydrogeological parameters with the presence of *E. coli* was obtained, a new and final PCA$_{5.1}$ (see Table 4) was performed including the same variables as PCA$_5$, except for *E. coli* concentration, that was removed from the set. The main goal was to assess the variables that most significantly influence the presence of *E. coli*. Afterwards, a generalised mixed model with Poisson error distribution was performed, including principal component variofactors as covariates. The covariates affecting significantly the presence of *E. coli* were only C1 (Figure 4a; χ^2_1= 63.379; $p < 0.001$; β =-1.04) and C2 (Figure 4b; χ^2_1= 3.852; p = 0.049; β =-0.19), while C3 (χ^2_1= 2.655; p = 0.103) and C4 (χ^2_1= 0.199; p = 0.655) were not significant (p values exceeded 0.05), thus making the conclusions of PCA$_5$ even more robust.

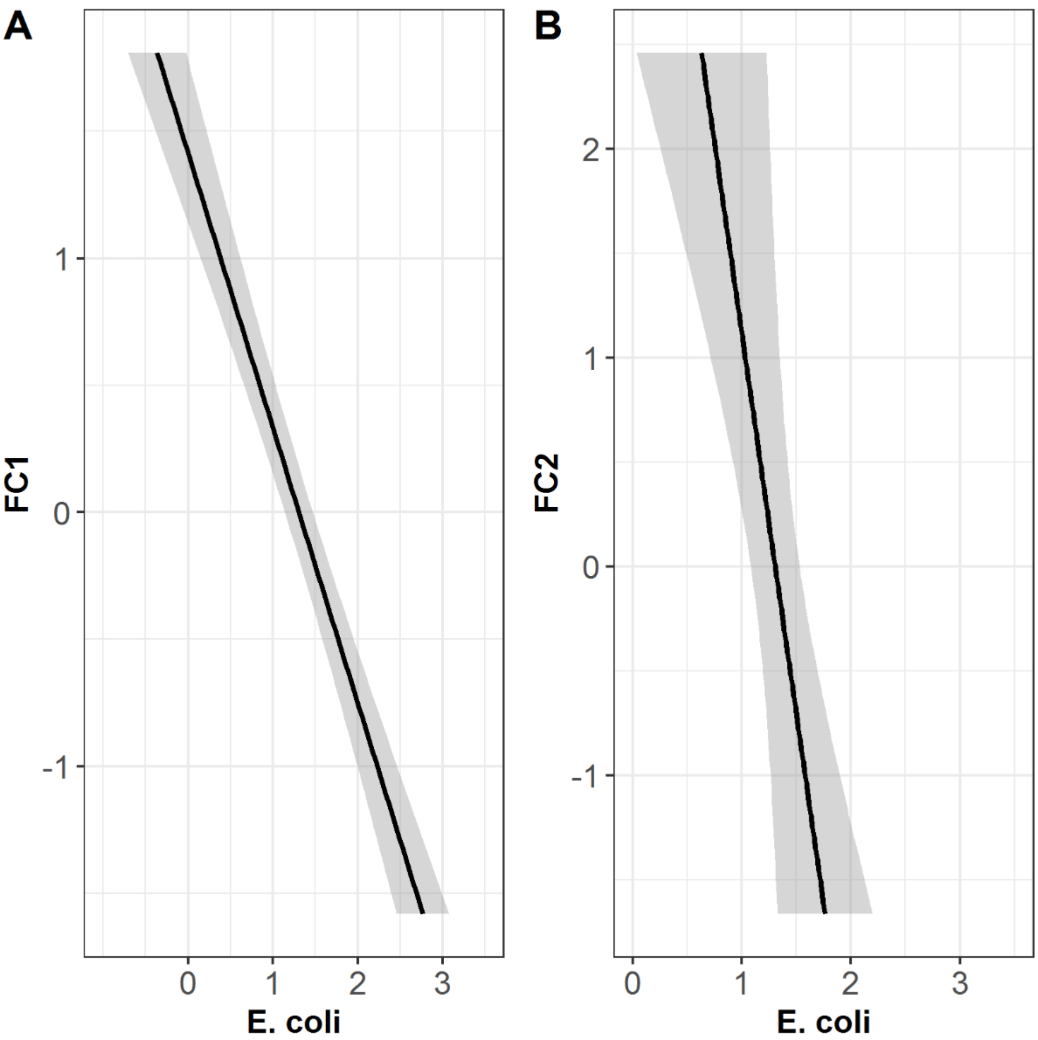

Figure 4. Significant relation between E. coli and variofactor 1 (A) and variofactor 2 (B) from PCA$_{5.1}$ considering all types of wells. E. coli ranges from safe (0) to unsafe (3) (Table 1). Regarding the PCA$_{5.1}$ results: FC1 is positively correlated with Type of well and Aquifer unit, and negatively with the distance to latrines within 30 m (Q8). FC2 is positively correlated with Na concentration and the extension of the cement floor (Q1).

Hand-dug wells and Hand-dug wells with handpumps

Another way of reading Figure 3a is by noticing that hand-dug wells (regardless whether they are equipped with handpumps) are the well types with the highest presence of *E. coli*. This is in agreement with previous studies (Dayanti et al., 2018; Kilungo et al., 2018; Mzuga et al., 1998; Ugochukwu and Ojike, 2019). In order to find which variables are affecting the presence of *E. coli* in these most polluted points, five new PCAs were performed now only including data from hand-dug wells*.* Therefore, we included here the variables such as groundwater depth (GWL), groundwater column height within the well, and some specific sanitary risk factors related only to this type of waterpoints. Like in the previous sets of PCAs, both sampling surveys were included (Table 5).

Table 5. Components extracted from the second set of PCA analyses (bold indicates negative correlation). For all PCAs the Bartlett sphericity test was significant ($p<0.001$). The variables involved in each component, the proportion of variance represented by each component, and the measure of sampling adequacy (KMO test value) are included

Type of variables	# of variables	PCA number	Extracted components	% of variance	Total variance	KMO Test Value	Indication of each component
Physicochemical parameters	15	PCA$_a$	C1: Geology, Log EC, Log Na, Log Cl, Log SO_4	23.02	86.72	0.540	Major ions and geological setup
			C2:Geology, Log EC, Log Alkalinity, GWL	15.32			GWL related to geology and chemical properties
			C3:Date, DO	11.57			Oxygen as function of seasonality
			C4:**NO_3,** Log Si	10.42			Relevant minor geochemical species
			C5: GWL, Water Column	9.86			Water levels
			C6: Log Eh, **NH_4**	8.96			Redox state
			C7: TOC	7.57			Isolated variable representing a statistical component
Sanitary risk data	12	PCA$_b$	C1:Q6, **Num. Latrines**	23.02	71.13	0.564	Latrines located inside the main villages in the coast, where animals have no physical access
			C2:Q1, Q3, Q10	15.32			Well construction and maintenance parameters
			C3: Q8, Q12, Q13	11.57			Pollution and sanitary conditions from presence of latrines
			C4: Q11	11.04			Isolated variable representing a statistical component
			C5: Q2, **Distance Latrines**	10.18			Unknown explanation
Physicochemical + *E. coli*	8	PCA$_c$	C1: Log Eh, ***E. coli,*** Water Column, DO, TOC, Geology	55.09	72.16	0.805	*E. coli* concentrations correlate negatively with water columns, redox potential (Eh, DO) and organic carbon
			C2: NO_3, GWL	17.07			The largest villages with NO_3 pollution are located near the coast where the groundwater level is shallow
Sanitary risk + *E. coli*	9	PCA$_d$	C1: **Q1**, Q12, Q13, *E. coli*	26.97	69.2	0.573	*E. coli* positively correlated with unsanitary practices, and inversely with the presence of cement floor
			C2: Q6, **Q8**	14.4			No animal physical access to latrines located in the main villages.
			C3: Q11	13.97			Isolated variable representing a statistical component
			C4: Q2, Q3	13.86			Sanitary conditions caused by direct water infiltration
	9	PCA$_e$	C1: **Q1**, Q12, Q13	20.64	68.91	0.517	Unsanitary practices correlated inversely to presence of cement floor

Physicochemical + *E. coli*+Sanitary risk data			C2:Log Eh, ***E. coli,*** Water column, GWL	19.45			*E. coli* inverse correlated with depth to groundwater level, water column and Eh
			C3:Q8, Q13, Water column	15.93			Latrines and scattered waste inside open well
			C4: Geology	12.89			Isolated variable representing a statistical component
PCA$_e$ without *E. coli*	8	PCA$_{e.1}$	C1: **Q1**, Q12, Q13	24.41	60.71	0.497	Unsanitary practices correlated inversely to presence of cement floor
			C2:Log Eh, Water column, GWL	18.71			*E. coli* inverse correlated with depth to groundwater level, water column and Eh
			C3:Q8, Water column	17.59			Latrines and scattered waste inside open well

As in the previous section, the first PCA$_a$ performed (Table 5) was a preliminary screening of variables to select the ones being significant in terms of information, thus allowing eliminating those that were redundant or irrelevant. A second analysis was conducted (PCA$_b$) (Table 5), considering only the sanitary risk factors from the questionnaire for this particular subset of waterpoints (thus, without the need to include here the variable "type of well"). Following the scheme reported previously, these two PCAs were followed by two more where the variable *E. coli* was added. PCA$_c$ included selected hydrochemical variable and *E. coli*. PCA$_d$ included sanitary risk factors (selected from the results of PCA$_b$) and the variable *E. coli*, for a total of nine variables.

PCA$_e$ involved nine variables including hydrochemical, selected risk factor variables and presence of *E. coli*. In component two, *E. coli* showed inverse correlation with depth to groundwater level, water column and Eh. In general, despite the uniformity in the physical and chemical properties in the water column, a prominent stratification of microbial groups was observed (consistent with Karlov et al., 2008). The inverse correlation between *E. coli* and the water column suggested preferential presence of faecal bacteria when the water column was small. Results of PCA$_e$ are represented in Figure 5 as a projection on the plane corresponding to variofactors 1 and 2.

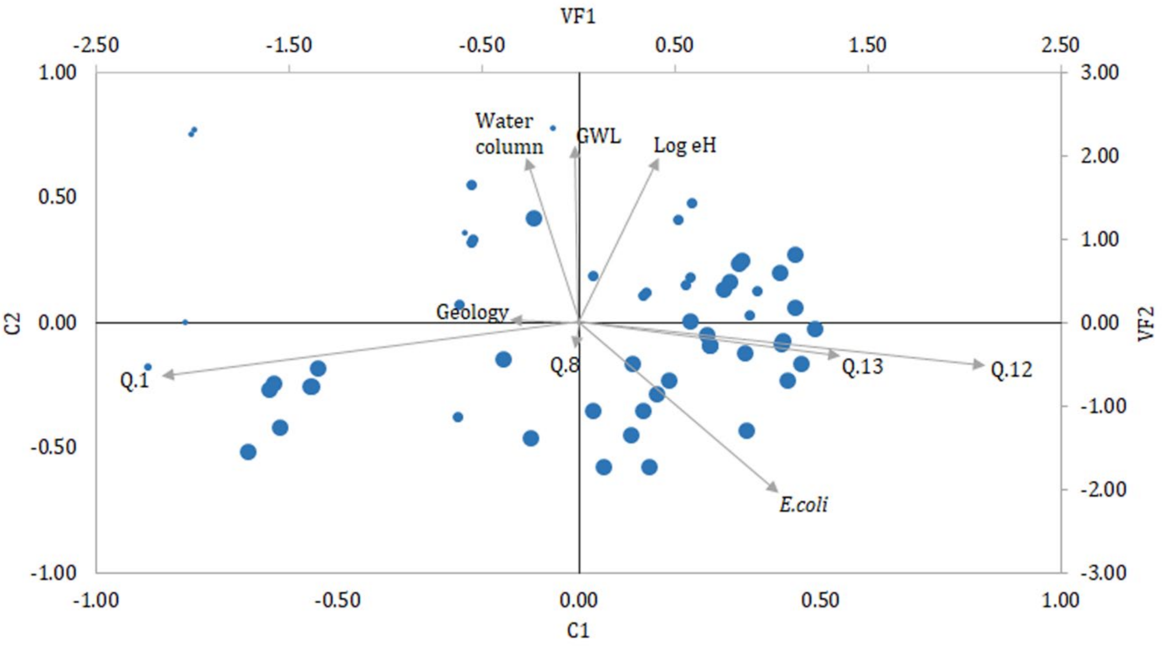

Figure 5. Main results of PCA$_e$. Samples are projected to variofactor space (VF1 and VF2 axes), and position of samples are scaled for visualization purposes. Size of the points increase with E. coli measurements. Grey arrows represent the contribution of each variable projected into the variofactor plane, so that some components can be easily identified.

Finally, PCA$_{e.1}$ was performed including the same variables as PCA$_e$, after excluding *E. coli*. In short, the components obtained were very similar to those from PCA$_e$, thus indicating robustness in the analysis.

From the results of PCA$_{e.1}$, a generalised mixed model with Poisson error distribution was performed including principal components variofactors as covariates. The analysis indicates that the covariates affecting significantly the presence of *E. coli* only in hand-dug wells were C1 (Figure65a; χ^2_1= 7.399; p = 0.006; β =0.15) -sanitary issues- and C2 (Figure 6b; χ^2_1= 4.496; p = 0.033; β=-0.15) -redox state related to GW levels-, while C3 (χ^2_1= 1.388; p = 0.238) and subsequent components, were not found significant.

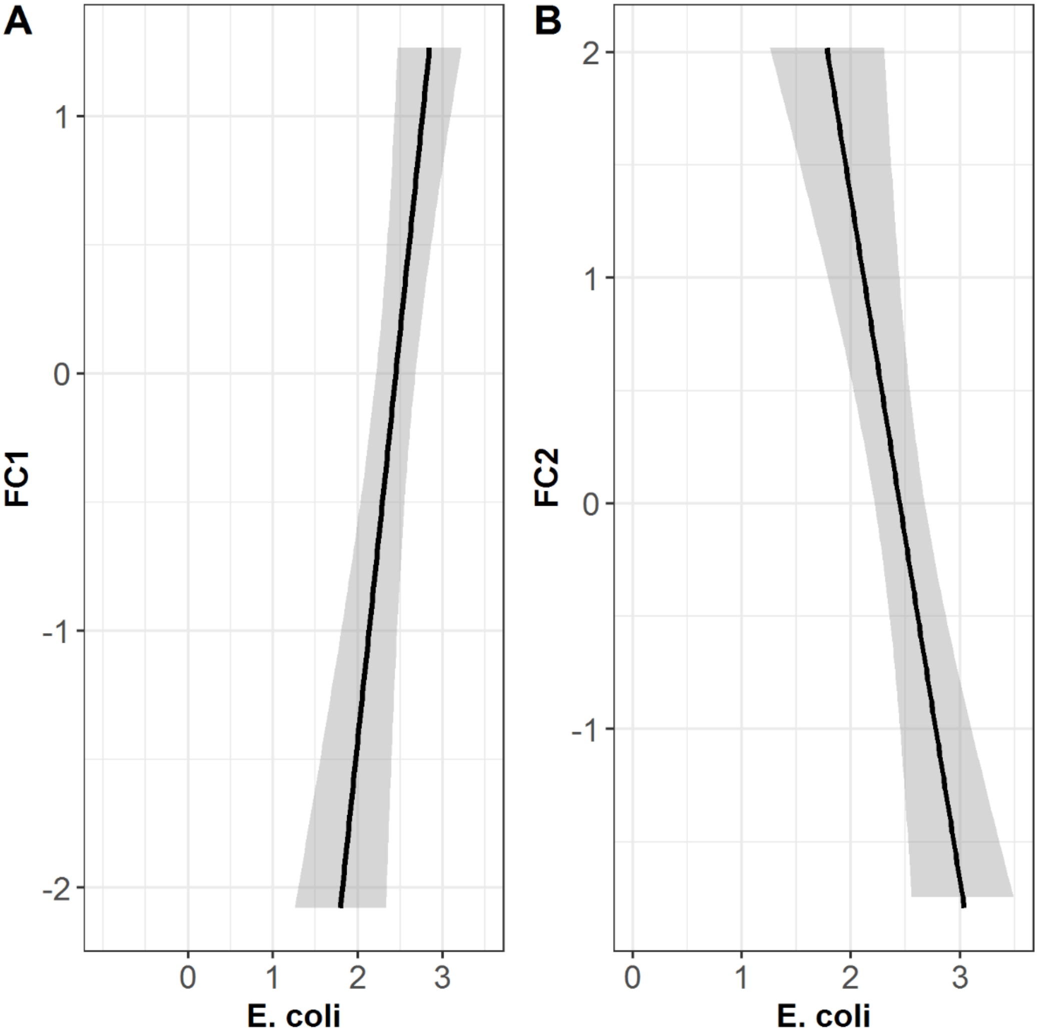

Figure 6. Significant relation between E. coli and variofactors V1 (A) and V2 (B) from PCA$_{e.1}$ considering only hand-dug wells and hand-dug wells with handpumps. E. coli ranges from Safe (0) to Unsafe (3). FC1 is positively correlated with the extension of the cement floor (Q1), the cover of the hand-dug well (Q12) and the presence of waste inside the well (Q13). FC2 is positively correlated with log Eh, GWL depth and water column.

4. **Discussion**

In the study area, a coastal rural area in South-East Kenia, microbiological pollution levels exceeded the WHO drinking water quality recommendations in almost all the waterpoints analysed in two campaigns in 2016. We could not find any direct relation of geology to *E. coli* pollution, although in other cases we believe that geology could be a significant factor, as it might drive fast/slow recharge.

Most bacteriological problems in groundwater supply points can be associated to improper well design, bad construction, and/or insufficient maintenance practices. Lutterodt et al. (2018) already points out that shallow hand-dug wells have more pollution and sanitary issues as compared to boreholes. Actually, in this study, well type is the primary variable controlling the presence of *E. coli.* Inadequate maintenance of hand pumps, improper sanitation and unhygienic conditions around the waterpoints, are factors that may contribute to faecal contamination, in line with the existing literature (Sukumaran et al., 2015; Ercumen et al., 2017; Godfrey et al., 2006; Lin et al., 2018), since unsanitary covers and litter scattered inside (or around) the well strongly result in the presence of *E. coli* in hand-dug wells, regardless of the presence of handpumps. Furthermore, the extension of the cement floor around the waterpoints and its maintenance state, significantly affects *E. coli* presence, since a small protection by cementation could imply short transit times and direct injection of bacteria through the non-saturated zones.

The highest counts of faecal bacteria were observed near human settlements. Unlike other studies that suggest that groundwater faecal pollution is highly variable in a monthly basis (Knappett et al., 2012b; van Geen et al., 2011), influenced by seasonal changes, and being significantly largest during the wet season (Howard et al., 2003; Kayembe et al., 2018), the present study does not show any difference in E. coli quantification between seasons. This could be explained due to the low precipitation during the wet season in 2016, when the study area was affected by La Niña event, with an estimated 69% reduction in recharge compared to average values (Ferrer et al., 2019). For this reason, more sampling campaigns would be needed in order to study the effect of seasonality on the faecal bacteria in the study area. Actually, it was observed that *E. Coli* concentration values increased with low groundwater levels mostly in the dry season, most probably related to direct input of bacteria (either for well construction or maintenance conditions) into small volumes of water. Future research is needed to understand the actual causality of this correlation, since longer and more recurrent droughts will be expected under future climate change conditions that in sub-Saharan Africa might imply the lowering of groundwater levels, causing a potential cascading effect on water availability and quality.

Some geochemical variables displayed a strong correlation with the registered concentrations of *E. coli.* Yet, in some cases it is only due to some external factor that explains both variables together. An example, is Na^+ concentrations. In the study area, Na^+ and *E. coli* concentrations display a significant negative correlation, mainly related to low Na^+ and high *E. coli* concentration values found in the wells located in the Margarini and Kilindini sands. These geological formations show low transit time through the unsaturated zone (Ferrer et al. 2019) and thus less attenuation capacity of the soil, with high *E. coli* counts reaching the shallow aquifer. These observations are in line with Howard et al., (2003), who suggested that fast recharge is the major cause of microbiological contamination, and underpinned that hydrochemical and isotopical data, routinely used to evaluate transit times in aquifer systems, might also be used as indicators of the presence or absence of faecal pollution in other study areas in similar realities. Furthermore, these results are in line with the studies of Leber et al., (2011) and van Geen et al., (2011), where low permeability layers and large residence times of groundwater were suggested as the cause of the little to none *E. coli* presence.

Regarding the risk factors affecting all type of waterpoints, this study confirms that the presence of leaching pit latrines in the vicinity of supply wells is a clear driver of faecal pollution, causing serious concerns for the public, as already shown in Howard et al., 2003; Graham and Polizzotto, 2013; Martínez-Santos et al., 2017; Prüss-Ustün et al., 2016; Schmoll et al., 2006. This effect increases whenever there is a general lack of physical barriers (e.g., concrete) in the latrines between stored excreta and soil and/or groundwater (Van Ryneveld and Fourie, 1997). It is relevant to note that despite the presence of *E. coli* in the study area is correlated to the presence of pit latrines within 30 m from the well, it is not correlated to the actual number of latrines in the vicinity. Thus, it seems that the presence of just one single latrine is enough to cause pollution at the well, while the actual number of latrine just becomes irrelevant.

Redox condition shows a positive correlation with dissolved oxygen, number of latrines and *E. coli* concentration. Latrines are an obvious source of oxygenated water with organic matter and bacteria loads and so, despite the presence of organic matter usually leads to a fast depletion of oxygen, they can coexist for short times in particular in heterogeneous soils (see e.g., Freixa et al., 2015). Low values of Eh result in enhanced transport of bacteria in groundwater. *E. coli* is also correlated again to water levels; thick non-saturated zones increase water transit times from the surface to the aquifer, reducing aquifer vulnerability to pollution. As Weldeyohannes et al. (2018) show, the levels of *E. coli* decrease dramatically (below detection limits) when the vadose zone is more than 0.9 m thick. This could be due to the additional mechanisms in the unsaturated zone favouring colloid/bacterial retention at the solid-water interfaces (Sepehrnia et al., 2018a). This effect might counteract that of bacteria increasing with reducing water levels mentioned before; a reduction of the saturated thickness also results in a large retention of faecal bacteria in the unsaturated zone.

A management strategy to reduce sanitary risks related with groundwater supply should focus on the correct construction of the wells to improve the isolation of the waterpoints to external sources of pollution. One possible solution would imply drilling of shallow boreholes equipped with handpumps and totally protected on the top, and also using vertical seals consisting of cement or expanding bentonite clay along the annulus between the casing and the borehole. Notice that vertical seals were not explored in this study because of the Kenia reality. Furthermore, despite they are less affordable, drilling deep boreholes seem to be the safest solution, but could result in groundwater in anaerobic conditions, with the need for additional treatment. Well maintenance, protection of waterpoints (preventing water ponding around, and with sanitation coverage implementation), and sanitary practices are a must, and should be emphasized; as a consequence, awareness and sensitization campaigns to eradicate malpractices should be carried out.

5. **Conclusions**

While the presence of faecal bacteria in domestic supply wells has been acknowledged for decades, no study until the present discriminate and quantify how the combination of hydrogeological and non-hydrogeological parameters correlate with the presence of *E. coli* as a proxy of faecal pollution. Therefore, a number of qualitative and quantitative variables combining geological, hydrological, geochemical, sanitary risk factors, well types, and maintenance variables have been statistical analysed for correlations with *E. coli* concentrations in a coastal area of Sub-Saharan Africa, with high presence of faecal bacteria in the groundwater used to supply the population.

This study demonstrates that including in a PCA such an interdisciplinary set of variables can be a useful methodology to obtain precise information of the relations between different types of variables, most times separated in analysis (e.g., in modelling efforts). Furthermore, this study goes a step forward when trying to assess which variables are related to faecal bacteria pollution; that is, including PCA variofactors as a covariate in mixed models might become a useful tool to assess the factors influencing significantly the presence of pathogenic organisms. Despite the geological formation itself did not show a direct relation with *E. coli* pollution, some hydrogeologically related parameters and variables (flow velocity, redox condition, water column, etc) were found significant in the analyses. Therefore, geology and hydrogeology can be combined when explaining risk pollution in shallow aquifer wells.

This methodology has confirmed in a quantitative way that the well constructive characteristics are most important to avoid the presence of pathogenic bacteria in groundwater. Extended cement floor would reduce the presence of faecal bacteria pollution, being more important in those areas where the water infiltrates fast through the unsaturated zone. Furthermore, knowing the geochemical elements, indicators of transit time, and groundwater depth, could just be simple and good indicators of the presence of faecal bacteria. Hence, easy to identify and measure physical and geochemical measurements, such as water column and Eh, may be used to assess a priori faecal pollution. Actually, Eh can be related to the presence of input water with high organic matter load (indicative of the presence of latrines nearby), while variations in the water column are driven by climate and well operation conditions.

Acknowledgements

The authors gratefully acknowledge the support of Kenya's Water Resource Authority (formerly WRA), the Kwale Country Government, Base Titanium Ltd., Kwale International Sugar Company Ltd. and Rural Focus Ltd (RFL). This research was funded by the UK Government via NERC, ESRC and DFID as part of the Gro for GooD project (UPGro Consortium Grant: NE/M008894/1).

Bibliography

Adelana, S.M.A., MacDonald, A.M., 2008. Groundwater research issues in Africa. IAH Sel. Pap. Hydrogeol. 13, 1–7.

Aquagenx, 2015. Compartment Bag Test (CBT) Instructions for Use: Drinking Water. [WWW Document]. URL www.aquagenx.com

Bain, R., Cronk, R., Hossain, R., Bonjour, S., Onda, K., Wright, J., Yang, H., Slaymaker, T., Hunter, P., Prüss-Ustün, A., Bartram, J., 2014. Global assessment of exposure to faecal contamination through drinking water based on a systematic review. Trop. Med. Int. Heal. 19, 917–927.

Barba, C., Folch, A., Gaju, N., Sanchez-Vila, X., Carrasquilla, M., Grau-Martínez, A., Martínez-Alonso, M., 2019a. Microbial community changes induced by Managed Aquifer Recharge activities: linking hydrogeological and biological processes. Hydrol. Earth Syst. Sci 23, 139–154.

Barba, C., Folch, A., Sanchez-Vila, X., Martínez-Alonso, M., Gaju, N., 2019b. Are dominant microbial sub-surface communities affected by water quality and soil characteristics? J. Environ. Manage. 237, 332–343.

Bates, D., Mächler, M., Bolker, B., Walker, S., 2015. Fitting Linear Mixed-Effects Models Using lme4. J. Stat. Softw. 67, 1–48.

Bhattacharjee, S., Ryan, J.N., Elimelech, M., 2002. Virus transport in physically and geochemically heterogeneous subsurface porous media. J. Contam. Hydrol. 57, 161–187.

Blaschke, A.P., Derx, J., Zessner, M., Kirnbauer, R., Kavka, G., Strelec, H., Farnleitner, A.H., Pang, L., 2016. Setback distances between small biological wastewater treatment systems and drinking water wells against virus contamination in alluvial aquifers. Sci. Total Environ. 573, 278–289.

Broström, G., Holmberg, H., 2011. Generalized linear models with clustered data: Fixed and random effects models. Comput. Stat. Data Anal. 55, 3123–3134.

Carles-Brangari, A., Sanchez-Vila, X., Freixa, A., Romani, A., Rubol, S., Fernandez, D., 2017. A mechanistic model (BCC-PSSICO) to predict changes in the hydraulic properties for bio-amended variably saturated soils. Water Resources Research, 53(1), 93-109.

Carles-Brangari, A., Fernandez, D., Sanchez-Vila, X., Manzoni, S., 2018. Ecological and soil hydraulic implications of microbial responses to stress: a modeling analysis. Advances in Water Resources, 116, 178-194.

Charles, K.J., Souter, F.C., Baker, D.L., Davies, C.M., Schijven, J.F., Roser, D.J., Deere, D.A., Priscott, P.K., Ashbolt, N.J., 2008. Fate and transport of viruses during sewage treatment in a mound system. Water Res. 42, 3047–3056.

CWSB, 2013. Coastal Water Services Board-Water Point Mapping Report: Kwale County [WWW Document]. URL www.cwsb.go.ke

Dayanti, M.P., Fachrul, M.F., Wijayanti, A., 2018. Escherichia coli as bioindicator of the groundwater quality in Palmerah District, West Jakarta, Indonesia, in: IOP Conference Series: Earth and Environmental Science. 106, 1-7.

Devane, M.L., Weaver, L., Singh, S.K., Gilpin, B.J., 2018. Fecal source tracking methods to elucidate critical sources of pathogens and contaminant microbial transport through New Zealand agricultural watersheds – A review. J. Environ. Manage. 222, 293–303.

Elangovan, N.S., Lavanya, V., Arunthathi, S., 2018. Assessment of groundwater contamination in a suburban area of Chennai, Tamil Nadu, India. Environ. Dev. Sustain. 20, 2609–2621.

Ferguson, A.S., Layton, A.C., Mailloux, B.J., Culligan, P.J., Williams, D.E., Smartt, A.E., Sayler, G.S., Feighery, J., McKay, L.D., Knappett, P.S.K., Alexandrova, E., Arbit, T., Emch, M., Escamilla, V., Ahmed, K.M., Alam, M.J., Streatfield, P.K., Yunus, M., van Geen, A., 2012a. Comparison of fecal indicators with pathogenic bacteria and rotavirus in groundwater. Sci. Total Environ. 431, 314–322.

Ferguson, A.S., Layton, A.C., Mailloux, B.J., Culligan, P.J., Williams, D.E., Smartt, A.E., Sayler, G.S., Feighery, J., McKay, L.D., Knappett, P.S.K., Alexandrova, E., Arbit, T., Emch, M., Escamilla, V., Ahmed, K.M., Alam, M.J., Streatfield, P.K., Yunus, M., van Geen, A., 2012b. Comparison of fecal indicators with pathogenic bacteria and rotavirus in groundwater. Sci. Total Environ. 431, 314–322.

Ferrer, N., Folch, A., Lane, M., Olago, D., Odida, J., Custodio, E., 2019. Groundwater hydrodynamics of an Eastern Africa coastal aquifer, including La Niña 2016–17 drought. Sci. Total Environ. 661, 575–597.

Foster, T., Willetts, J., 2018. Multiple water source use in rural Vanuatu: are households choosing the safest option for drinking? Int. J. Environ. Health Res. 28, 579–589.

Freixa, A., Rubol, S., Carles, A., Fernandez, D., Butturini, A., Sanchez-Vila, X., Romani, A., 2015. The effects of sediment depth and oxygen concentration on the use of organic matter: An experimental study using an infiltration sediment tank. Science of the total environment, 540, 20-31.

Goyal, S.M., Keswick, B.H., Gerba, C.P., 1984. Viruses in groundwater beneath sewage irrigated cropland. Water Res. 18, 299–302.

Graham, J.P., Polizzotto, M.L., 2013. Pit Latrines and Their Impacts on Groundwater Quality: A Systematic Review. Environ. Health Perspect. 121, 521–530.

Gronewold, A.D., Sobsey, M.D., Mcmahan, L., 2017. The compartment bag test (CBT) for enumerating fecal indicator bacteria: Basis for design and interpretation of results. Sci. Total Environ. 587–588, 102–107.

Hair, J., Anderson, R., Tatham, R., Black, W., 1995. Multivariate Data Analysis, 4th ed, Technometrics. Prentice-Hall Inc, New Jersey.

Howard, G., Pedley, S., Barrett, M., Nalubega, M., Johal, K., 2003. Risk factors contributing to microbiological contamination of shallow groundwater in Kampala, Uganda. Water Res. 37, 3421–3429.

Howard, G., Pedley, S., Barrett, M., Nalubega, M., Johal, K., 2003. Risk factors contributing to microbiological contamination of shallow groundwater in Kampala, Uganda. Water Res. 37, 3421–3429.

Karlov, D.S., Marie, D., Danil, •, Sumbatyan, A., Chuvochina, M.S., Kulichevskaya, I.S., Alekhina, I.A., Sergey, •, Bulat, A., 2008. Microbial communities within the water column of freshwater Lake Radok, East Antarctica: predominant 16S rDNA phylotypes and bacterial cultures. Polar Biol. 40.

Kayembe, J.M., Thevenon, F., Laffite, A., Sivalingam, P., Ngelinkoto, P., Mulaji, C.K., Otamonga, J.P., Mubedi, J.I., Poté, J., 2018. High levels of faecal contamination in drinking groundwater and recreational water due to poor sanitation, in the sub-rural neighbourhoods of Kinshasa, Democratic Republic of the Congo. Int. J. Hyg. Environ. Health 221, 400–408.

Kilungo, A., Powers, L., Arnold, N., Whelan, K., Paterson, K., Young, D., 2018. Evaluation of well designs to improve access to safe and clean water in rural Tanzania. Int. J. Environ. Res. Public Health 15, 1–11.

Knappett, P.S.K., Emelko, M.B., Zhuang, J., McKay, L.D., 2008. Transport and retention of a bacteriophage and microspheres in saturated, angular porous media: Effects of ionic strength and grain size. Water Res. 42, 4368–4378.

Knappett, P.S.K., McKay, L.D., Layton, A., Williams, D.E., Alam, M.J., Huq, M.R., Mey, J., Feighery, J.E., Culligan, P.J., Mailloux, B.J., Zhuang, J., Escamilla, V., Emch, M., Perfect, E., Sayler, G.S., Ahmed, K.M., Van Geen, A., 2012a. Implications of fecal bacteria input from latrine-polluted ponds for wells in sandy aquifers. Environ. Sci. Technol. 46, 1361–1370.

Knappett, P.S.K., Mckay, L.D., Layton, A., Williams, D.E., Alam, M.J., Mailloux, B.J., Ferguson, A.S., Culligan, P.J., Serre, M.L., Emch, M., Ahmed, K.M., Sayler, G.S., Geen, A. Van, 2012b. Unsealed tubewells lead to increased fecal contamination of drinking water. J. Water Health 10, 565–578.

Leber, J., Rahman, M.M., Ahmed, K.M., Mailloux, B., van Geen, A., 2011. Contrasting Influence of Geology on E. coli and Arsenic in Aquifers of Bangladesh. Ground Water 49, 111–123.

MacDonald, A.M., Bonsor, H.C., Dochartaigh, B.É.Ó., Taylor, R.G., 2012. Quantitative maps of groundwater resources in Africa. Environ. Res. Lett. 7, 1–7.

Macler, B.A., Merkle, J.C., 2000. Current knowledge on groundwater microbial pathogens and their control, Hydrogeology Journal. 8, 29-40.

Martínez-Santos, P., Martín-Loeches, M., García-Castro, N., Solera, D., Díaz-Alcaide, S., Montero, E., García-Rincón, J., 2017. A survey of domestic wells and pit latrines in rural settlements of Mali: Implications of on-site sanitation on the quality of water supplies. Int. J. Hyg. Environ. Health 220, 1179–1189.

Matthess, G., Pekdeger, A., Schroeter, J., 1988. Persistence and transport of bacteria and viruses in groundwater - a conceptual evaluation. J. Contam. Hydrol. 2, 171–188.

Mzuga, J.M., Tole, M.P., Ucakuwun, E.K., 1998. The impact of geology and pit latrines on groundwater quality in Kwale District, Dunes, groundwater, mangroves and birdlife in coastal Kenya. Chapter 6, 85-96

Nowicki, S., Lapworth, D.J., Ward, J.S.T., Thomson, P., Charles, K., 2019. Tryptophan-like fluorescence as a measure of microbial contamination risk in groundwater. Sci. Total Environ. 646, 782–791.

Olajuyigbe, A.E., Olamiju, I.O., Ola-Omole, C.M., 2017. Vulnerability of hand-dug wells in the core area of Akure, Nigeria. Urban Water J. 14, 797–803.

Oteng-Peprah, M., de Vries, N.K., Acheampong, M.A., 2018. Greywater characterization and generation rates in a peri urban municipality of a developing country. J. Environ. Manage. 206, 498–506.

Paliy, O., Shankar, V., 2016. Application of multivariate statistical techniques in microbial ecology. Mol. Ecol. 25, 1032–1057.

Perujo, N., Sanchez-Vila, X., Proia, L., Romani, A., 2017. Interaction between physical heterogeneity and microbial processes in subsurface sediments: a laboratory-scale column experiment. Environmental Science and Technology", 51 (11), 6110-6119.

Rao, V.C., Metcalf, T.G., Melnick, J.L., 1986. Articles in the Update series Human viruses in sediments, sludges, and soils*, Bulletin of the World Health Organization.

Prüss-Ustün, A., Wolf, J., Corvalán, C., Bos, R., Neira, M., 2016. Global Burden of Diseases From Environmental Risks.

Rohmah, Y., Rinanti, A., Hendrawan, D.I., 2018. The determination of ground water quality based on the presence of Escherichia coli on populated area (a case study: Pasar Minggu, South Jakarta). IOP Conf. Ser. Earth Environ. Sci. 106.

Saiers, J.E., Ryan, J.N., 2005. Colloid deposition on non-ideal porous media: The influences of collector shape and roughness on the single-collector efficiency. Geophys. Res. Lett. 32, 1–5.

Schmoll, O., Howard, G., Chilton, J., Chorus, I., 2006. Protecting Groundwater for Health: Managing the Quality of Drinking-water Sources, Protecting Groundwater for Health: Managing the Quality of Drinking-water Sources.

Sepehrnia, N., Bachmann, J., Hajabbasi, M., Afyuni, M., Horn, M., 2018a. Modeling Escherichia coli and Rhodococcus erythropolis transport through wettable and water repellent porous media. Colloids Surfaces B Biointerfaces 172, 280–287.

Sepehrnia, N., Memarianfard, L., Moosavi, A.A., Bachmann, J., Rezanezhad, F., Sepehri, M., 2018b. Retention modes of manure-fecal coliforms in soil under saturated hydraulic condition. J. Environ. Manage. 227, 209–215.

Sharma, P.K., Srivastava, R., 2011. Numerical analysis of virus transport through heterogeneous porous media. J. Hydro-Environment Res. 5, 93–99.

Stauber, C., Miller, C., Cantrell, B., Kroell, K., 2014. Evaluation of the compartment bag test for the detection of Escherichia coli in water. J. Microbiol. Methods 99, 66–70.

Tabachnik, B., Fidell, L., 2007. Using multivariate statistics. Pearson Education Inc, Boston, MC.

Team, 2018. R: A Language and Environment for Statistical Computing.

Thompson, B., 2004. Exploratory and Confirmatory Factor Analysis: Understanding Concepts and Applications.

Tole, M.P., 1997. Pollution of groundwater in the coastal Kwale Distric, Kenya. Sustain. Water Resour. under Increasing Uncertain. 287–297.

Ugochukwu, U.C., Ojike, C., 2019. Assessment of the groundwater quality of a highly populated district in Enugu State of Nigeria. Environ. Dev. Sustain. Online ISSN 1573-2975.

Van Geen, A., Ahmed, K.M., Akita, Y., Alam, M.J., Culligan, P.J., Emch, M., Escamilla, V., Feighery, J., Ferguson, A.S., Knappett, P., Layton, A.C., Mailloux, B.J., McKay, L.D., Mey, J.L., Serre, M.L., Streatfield, P.K., Wu, J., Yunus, M., 2011. Fecal contamination of shallow tubewells in Bangladesh inversely related to arsenic. Environ. Sci. Technol. 45, 1199–1205.

Van Ryneveld, M., Fourie, A., 1997. A strategy for evaluating the environmental impact of on-site sanitation systems. Water SA 23, 279–291.

Weldeyohannes, A.O., Kachanoski, G., Dyck, M., 2018. Wastewater Flow and Pathogen Transport from At-Grade Line Sources to Shallow Groundwater. J. Environ. Qual. 47, 1051.

Wright, J.A., Cronin, A., Okotto-Okotto, J., Yang, H., Pedley, S., Gundry, S.W., 2013. A spatial analysis of pit latrine density and groundwater source contamination. Environ. Monit. Assess. 185, 4261–4272.

Yates, M. V, Gerba, C.P., Kelley, L.M., 1985. Virus persistence in groundwater. Appl. Environ. Microbiol. 49, 778–781.

Supplementary material

Table 1S. E. coli quantification results from CBT in March 2016. Green colour means safe, yellow means intermediate risk, orange means high risk and red means unsafe.

Code	Aquagenx (bags) *MPN/100 ml*	Code	Aquagenx (bags) *MPN/100 ml*	Code	Aquagenx (bags) *MPN/100 ml*
Footprints School	0,0	**A/14/10**	0,0	**Z2-112**	48,3
Z4-11	48,3	**Z3-87**	1,5	**Z1-140**	0,0
Z4-09	>100	**Z3-98**	>100	**Z2-104**	2,6
Z4-01	>100	**Z3-90**	>100	**Z1-110**	>100
A/04/12	0,0	**A/05/11**	>100	**DB/FI/HP**	0,0
Z4-18	>100	**HOTSPRING**	0,0	**Z3-96**	>100
A/06/12	1,2	**C108HWL**	>100	**E/29/01**	13,6
Z4-78B	>100	**3KD01**	48,3	**A/09/11**	0,0
Z4-08	48,3	**S1-3KD06**	>100	**MIVUMONI**	0,0
Z4-06	>100	**GD31**	0,0	**C/15/10**	0,0
D/100/16	13,6	**MUK DAM**	>100	**C/109/21**	0,0
Z4-04	>100	**MUK DWS**	>100	**C/12/12**	0,0
Z4-MS	>100	**Z1-122**	13,6	**C/06/12**	0,0
D/82/14	0,0	**Z1-125**	>100	**C/19/10**	0,0
Z4-85	>100	**Z1-124**	>100	**D/129/19**	0,0
Z4-24	>100	**D/16/10**	1,2	**DB/MH/CO**	0,0
Z3-25	>100	**Z1-121B**	>100	**Z1-141**	>100
D/63/13	0,0	**Z1-116**	13,6	**UK-WL**	0,0
D/68/13	0,0	**C/07/09**	0,0	**A/06/13**	0,0
Z3-30	48,3	**A/01/11**	0,0	**D/103/16**	0,0
Z3-29	13,6	**Z2-103**	>100	**LUKORE-SH**	48,3
DB/BM/HP	0,0	**D/203/27**	1,1	**Z1-118**	>100
BH310	13,6	**DB/MS/LST**	0,0	**VIN-WL**	0,0
BH402	0,0	**Z1-135**	>100	**Base_BH_1**	0,0
NK-03	0,0	**DB/KI/ST**	0,0	**Base_BH_3**	0,0

Z1-70	>100	**Z1-33**	>100	**Base_BH_7**	0,0

Table 2S. E. coli quantification results from CBT in June 2016. Green colour means safe, yellow means intermediate risk, orange means high risk and red means unsafe.

Code	**Aquagenx (bags)** *MPN/100 ml*	**Code**	**Aquagenx (bags)** *MPN/100 ml*	**Code**	**Aquagenx (bags)** *MPN/100 ml*
Footprints School	0,0	**A/05/11**	>100	**Z1-110**	>100
Z4-11	48,3	**HOTSPRING**	0,0	**DB/FI/HP**	0,0
Z4-01	>100	**C108HWL**	>100	**Z3-96**	48,3
A/04/12	0,0	**3KD01**	48,3	**E/29/01**	9,6
Z4-18	48,3	**MUACHEMA**	>100	**A/09/11**	0,0
A/06/12	0,0	**S1-3KD06**	>100	**MIVUMONI**	0,0
Z4-78B	>100	**GD31**	0,0	**C/15/10**	0,0
Z4-08	>100	**MUK DAM**	>100	**C/109/21**	0,0
Z4-06	>100	**MUK DWS**	>100	**C/12/12**	0,0
D/100/16	48,3	**Z1-122**	13,6	**C/06/12**	0,0
Z4-04	>100	**Z1-125**	>100	**C/19/10**	0,0
Z4-MS	48,3	**Z1-124**	>100	**D/129/19**	0,0
D/82/14	0,0	**D/16/10**	0,0	**DB/MH/CO**	0,0
Z4-85	48,3	**Z1-121B**	>100	**Z1-141**	>100
Z4-24	>100	**Z1-116**	13,6	**UK-WL**	0,0
D/63/13	0,0	**C/07/09**	0,0	**D/103/16**	0,0
D/68/13	0,0	**A/01/11**	0,0	**LUKORE- SH**	0,0
Z3-30	>100	**Z2-103**	>100	**Z1-118**	>100
Z3-29	>100	**D/203/27**	13,6	**VIN-WL**	0,0
DB/BM/HP	0,0	**DB/MS/LST**	0,0	**Base_BH_3**	0,0
BH310	0,0	**Z1-135**	>100	**Base_BH_7**	0,0
Z1-70	>100	**Z2-112**	48,3	**DB/KI/ST**	0,0
Z1-33	48,3	**Z1-140**	1,5	**Z3-102B**	13.6
A/14/10	0,0	**Z2-104**	4,7	**BH302**	0,0
Z3-87	>100	**Z3-98**	>100	**C/05/09**	48.3
Z3-90	13,6	**C/03/09**	0,0		

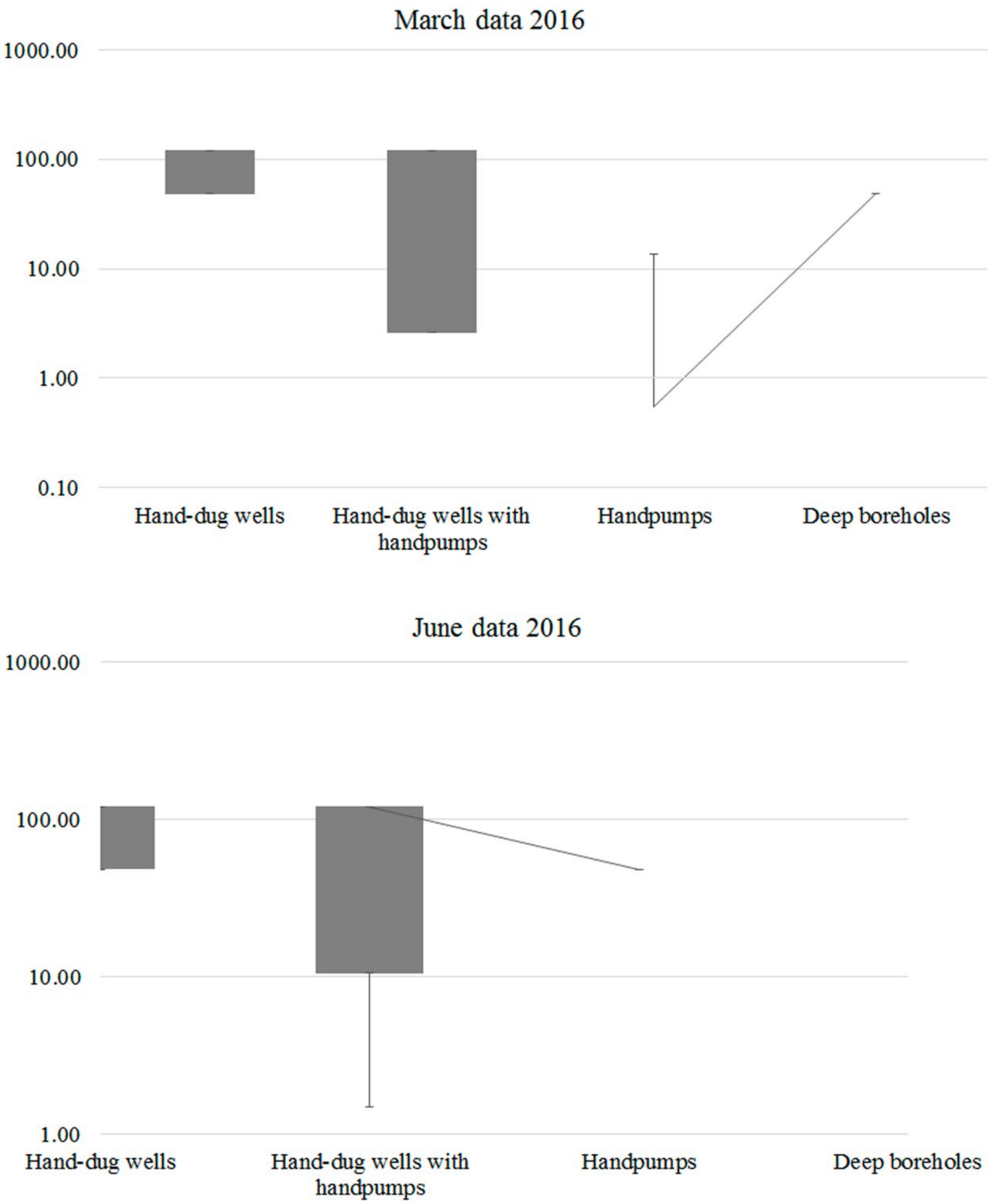

Figure SM. E. coli presence for each type of each water point in each field survey.