Mapping risks of faecal contamination of shallow groundwater in Dakar, Senegal: an evaluation of culture-based methods and a real-time technique using tryptophan-like fluorescence

Raphaëlle Roffo

Year of submission: 2018

Supervisor: Professor Richard Taylor

This research dissertation is submitted for the MSc in Geospatial Analysis
at University College London

DISSERTATION DECLARATION

DEPARTMENT OF GEOGRAPHY

M.Sc. in Geospatial Analysis

I, Raphaëlle Roffo, hereby declare:

(a) that this M.Sc. Project is my own original work and that all source material used is acknowledged therein;

(b) that it has been prepared specially for the MSc in Geospatial Analysis of University College London;

(c) that it does not contain any material previously submitted to the Examiners of this or any other University, or any material previously submitted for any other examination.

Signed: Raphaëlle Roffo

Date: 29 August 2018

ABSTRACT

Each day, 1.8 billion individuals around the world drink water contaminated with faeces (WHO, 2017). In sub-Saharan Africa alone, this represents a leading cause of mortality, as diarrhoeal diseases killed 643,000 people in 2015 (WHO, 2016). In the coastal megacity of Dakar, Senegal, the Thiaroye shallow aquifer is a complex system in which multiple sources of pollution and a lack of sufficient sanitation infrastructure have contributed to an extreme degradation of groundwater quality. This study is an investigation of faecal contamination patterns across the aquifer. It is based on data collection conducted in the greater Dakar region under the Dakar urban observatory of the AfriWatSan project, in June-July 2018. Within the AfriWatSan framework, this study seeks to explore faecal contamination patterns across the Thiaroye aquifer, based on standard culture-based methods and tryptophan-like fluorescence (TLF). TLF is a fluorescence-based method currently being developed by the British Geological Survey for real-time screening of faecally contaminated drinking water in urban Africa. The method offers several key advantages over traditional methods as it is portable, real-time and easy to use.

97 samples were collected with 48 parameters including hydrochemical parameters and environmental risks. This study first seeks to explore the relationships between different variables, with a specific focus on TLF performance as a faecal matter detection method. It then explores spatial patterns of contamination, before adopting an unsupervised machine learning approach to classification with Agglomerative Hierarchical Clustering (HAC).

While TLF fails to accurately predict current contamination across the Thiaroye aquifer, this data exploration and modelling exercise provides additional information about the Thiaroye aquifer groundwater quality. In order to achieve a more accurate representation of the contamination, further research will need to incorporate groundwater flow modelling, and to investigate vertical contamination flows.

Research was conducted under the AfriWatSan project, funded by The Royal Society (UK) and Department for International Development (DFID), and supported by the British Geological Survey (BGS), currently developing portable, UV-based fluorimeters for real-time screening of faecally contaminated drinking water in urban Africa.

Keywords: Tryptophan-like fluorescence; Thermotolerant coliforms, Sanitation, Groundwater quality monitoring, Logistic regression, Hierarchical clustering.

(9,754 words)

ACKNOWLEDGEMENTS

I cannot thank enough the entire AfriWatSan team for giving me the opportunity to take part in this impactful project and to conduct field research in Senegal. A special thanks to the UCAD team for welcoming me so warmly and for providing all the support and logistics needed for this fieldwork. Professors, Doctors, PhD students and MSc students alike were an invaluable support and a second family, and I wish them all great success in their future research!

I would also like to warmly thank my supervisor Professor Richard Taylor for his guidance and reassurance, especially when I was a bit overwhelmed by the strange data I collected!

Many thanks as well to my lecturers at UCL for an intense year that has allowed me to use the very powerful tools of GIS software, R and Python and to draw beautiful maps.

Finally, *un grand merci* to my parents for their support, to Rebecca for always taking my calls and laughing when I needed it, and of course to Wen for being there and enthusiastically supporting me during this entire MSc!

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

AGNES: Agglomerative Nesting

BGS: British Geological Survey

CDOM: Coloured Dissolved Organic Matter

CFU/100mL: Colony-Forming Units per 100 mL

DOC: Dissolved Organic Carbon

E. coli: Escherichia Coli

GWR: Geographically Weighted Regression

HAC: Agglomerative Hierarchical Clustering

IDW: Inverse Distance Weighted Interpolation

MLSB: Membrane Lauryl Sulphate Broth

RMSE: Root Mean Square Error

SDG: Sustainable Development Goals

TLF: Tryptophan-like Fluorescence

TTC: Thermotolerant Coliforms

UCAD: Université Cheikh Anta Diop de Dakar *(Dakar University)*

UN: United Nations

UNICEF: United Nations Children's Fund

WHO: World Health Organization

Mapping risks of faecal contamination of shallow groundwater in Dakar, Senegal: an evaluation of culture-based methods and a real-time technique using tryptophan-like fluorescence

1. INTRODUCTION

1.1. Introduction

Over the past decades, megacities in Sub-Saharan Africa have undergone an unprecedented scale and pace of urbanization (Cohen, 2006; United Nations, 2014). As urban areas rapidly expand, deficiencies in urban planning result in major issues for the extension of water and sanitation networks (Criqui, 2014). In unplanned and informal settlements, the installation of pipes for water supply and the discharge of used water is impeded by the absence of rights-of-way (Criqui, 2013; Monstadt and Schramm, 2017). The extension of sanitation networks is particularly difficult and is complicated by social taboos, political gridlocks, lack of funding and the difficulty to gather coalitions to effectively finance, execute and maintain these networks (Sansom, 2006). In rapidly growing cities of the developing world, it is therefore recognized that on-site solutions to water or sanitation remain a reality, and their impact on groundwater quality and human health should be assessed (Diaw *et al.*, forthcoming).

Water supply and sanitation are indeed two closely intertwined services, with impacts cutting across most contemporary challenges: health, environment, gender, education, economic development, climate change (Acquistapace *et al.*, 2017). The importance of water has been acknowledged by the international community with the recognition in 2010 by the United Nations General Assembly of a human right to water, which entails six criteria: *"The human right to water entitles everyone to sufficient, safe, acceptable, physically accessible and affordable water for personal and domestic uses."* (United Nations *et al.*, 2010 cited in Winkler, 2014). It was further reaffirmed through the 2015 Sustainable Development Goals (SDG), with a standalone goal for safe water and sanitation (SDG 6) as well as a series of transverse targets and indicators in the Health, Urban and Ocean SDGs.

Freshwater only represents 2.8% of the planet's water resources, of which less than 1% is liquid. The Joint Monitoring Programme estimates that in 2015, 2.1 billion individuals around the world still lacked access to improved sources of water (defined in Figure 1), including 844 million across 80 countries who did not even have access to basic water sources. A large portion of this population lives in sub-Saharan Africa, where only 58% of the population have access to "at least basic" drinking water services (WHO/UNICEF Joint

Monitoring Programme for Water Supply and Sanitation, 2017). This causes serious health risks and constitutes a leading cause of mortality in developing countries. In sub-Saharan Africa alone, diarrhoeal diseases killed 643,000 people in 2015 (World Health Organization, 2016). Children under 5 are particularly at risk, with 525 000 children dying from diarrhoea each year.

SERVICE LEVEL	DEFINITION
SAFELY MANAGED	Drinking water from an improved water source that is located on premises, available when needed and free from faecal and priority chemical contamination
BASIC	Drinking water from an improved source, provided collection time is not more than 30 minutes for a round trip, including queuing
LIMITED	Drinking water from an improved source for which collection time exceeds 30 minutes for a round trip, including queuing
UNIMPROVED	Drinking water from an unprotected dug well or unprotected spring
SURFACE WATER	Drinking water directly from a river, dam, lake, pond, stream, canal or irrigation canal

Note: Improved sources include: piped water, boreholes or tubewells, protected dug wells, protected springs, and packaged or delivered water.

Fig. 11 The new JMP ladder for drinking water services

SERVICE LEVEL	DEFINITION
SAFELY MANAGED	Use of improved facilities that are not shared with other households and where excreta are safely disposed of in situ or transported and treated offsite
BASIC	Use of improved facilities that are not shared with other households
LIMITED	Use of improved facilities shared between two or more households
UNIMPROVED	Use of pit latrines without a slab or platform, hanging latrines or bucket latrines
OPEN DEFECATION	Disposal of human faeces in fields, forests, bushes, open bodies of water, beaches or other open spaces, or with solid waste

Note: improved facilities include flush/pour flush to piped sewer systems, septic tanks or pit latrines; ventilated improved pit latrines, composting toilets or pit latrines with slabs.

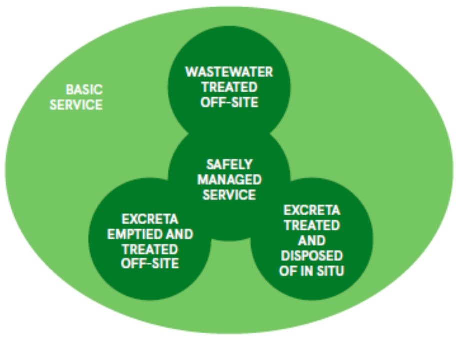


Fig. 12 The new JMP ladder for sanitation services

Figure 1: UNICEF, World Health Organization. (2017) The new JMP ladder for drinking water and sanitation services [Diagram]. In: *Progress on Drinking Water, Sanitation and Hygiene. Update and SDG Baselines*. New York/Geneva: UNICEF, WHO

The issue of drinking water is intrinsically linked to that of sanitation. The Joint Monitoring Programme estimates that in 2015, only 39% of the global population used safely managed sanitation infrastructure (defined in Figure 1), and as little as 27% were connected to a sewerage system with wastewater treatment (WHO/UNICEF Joint Monitoring Programme for Water Supply and Sanitation, 2017). In Sub-Saharan Africa, on-site sanitation represents 38% of sanitation solutions, against 8% for sewers. 72% of the population in this region do not have access to "at least basic" sanitation services (improved sanitation facilities that are not shared). But in many ways, these figures also underestimate the scale of challenges ahead. For instance, safely managed service is defined as the "population using an improved sanitation facility that is not shared with other households, and where excreta are being disposed of in-situ or transported and treated off-site", however in Senegal, 63% of rural on-site sanitation facilities, counted as "safely managed", have never been emptied. This type of practices are a threat to groundwater quality and human health, as without regular emptying, pathogens can easily infiltrate the aquifer.

Finally, the challenges facing the provision of these essential services are even more salient in coastal cities, faced with additional threats posed by climate change induced sea-level rise: floods, saline contamination of fresh groundwater and vulnerability of wastewater collection systems to heavy rainfall and higher tide levels (Gaye *et al.*, 1990; Rosenzweig *et al.*, 2011). Because it is located on a peninsula off the Atlantic coast and has undergone rapid population growth, the Senegalese capital city Dakar is an insightful case study for the investigation of groundwater faecal contamination. Home to 2.47 million inhabitants, the megacity's peri-urban area is not connected to a sewerage system and the city has faced severe groundwater quality issues in the past decades. Using two different contamination detection methods, this study explores patterns of faecal contamination in the shallow aquifer of Thiaroye, which covers the greater Dakar region.

1.2. Context of the study

1.2.1. Faecal matter detection

a. Culture-based detection methods

Faecal matter detection methods are essential for the identification of causes of water contamination. Improved detection can also improve the communication of risks to the users and support the development of adequate solutions, especially in the domain of sanitation infrastructure.

The most common method of assessing faecal contamination of water consists in measuring the presence of surrogate indicator organisms (Bartram and Ballance, 1996). Thermotolerant coliforms (TTCs) and *Escherichia coli* (*E. coli*) have been found to be good indicator organisms for faecal contamination: their presence allows the analyst to infer that harmful pathogens may be present in the sample (World Health Organization, 2016). In other words, when TTC are found, risks of developing water-borne diseases are increased (Tallon *et al.*, 2005). The World Health Organization (WHO) drinking water standards recommend that no faecal coliform or *E. coli* be found in drinking water (Snozzi, Ashbolt and Grabow, 2001).

The detection of TTCs and *E.coli*, however, relies on culture-based methods that are costly, require the use of reagents and sterile equipment and adequate logistics for the samples to reach a laboratory (Sorensen *et al.*, 2015). Moreover, the plate counts can only be carried out after a minimum of 18hours of incubation. This is especially problematic is remote areas, and considerably slows down efforts to proactively inform users of drinking water quality. In this study, TTCs are used as faecal indicator organisms; the WHO considers TTCs to be a valid alternative to *E. coli* in most circumstances (Snozzi, Ashbolt and Grabow, 2001).

b. <u>Tryptophan-like Fluorescence (TLF)</u>

Tryptophan-like Fluorescence (TLF) is a potential faecal detection method that has recently been investigated by the British Geological Survey as an alternative to culture-based method (Sorensen *et al.*, 2018a), building on Baker's investigations of protein-like fluorescence intensity (Baker and Inverarity, 2004). This method is based on UV fluorescence, detected by portable sensors such as Chelsea Group Technologies' UviLux sensors, set at specific wavelengths. When tryptophan-like particles absorb UV light, they re-emit part of this energy as longer wavelength fluorescence. The intensity of fluorescence measured is therefore proportional to the concentration of the compound measured (Chelsea Technologies Group Ltd, 2018).

Several studies have found strong, significant positive correlations between TLF and indicators of faecal contamination such as TTCs or *E. coli* (Sorensen *et al.*, 2015, 2018a; Fox *et al.*, 2017). Yet, the exact mechanisms underlying these correlations are not clear. In particular, the UviLux sensor error range is of 50nm, meaning that measured TLF excitation and emission wavelengths (respectively 280nm and 360nm) slightly overlap with other compounds such as Coloured Dissolved Organic Matter (CDOM), with 347.5 nm excitation wavelength and 450 nm emission wavelength, as shown in Figure 2.

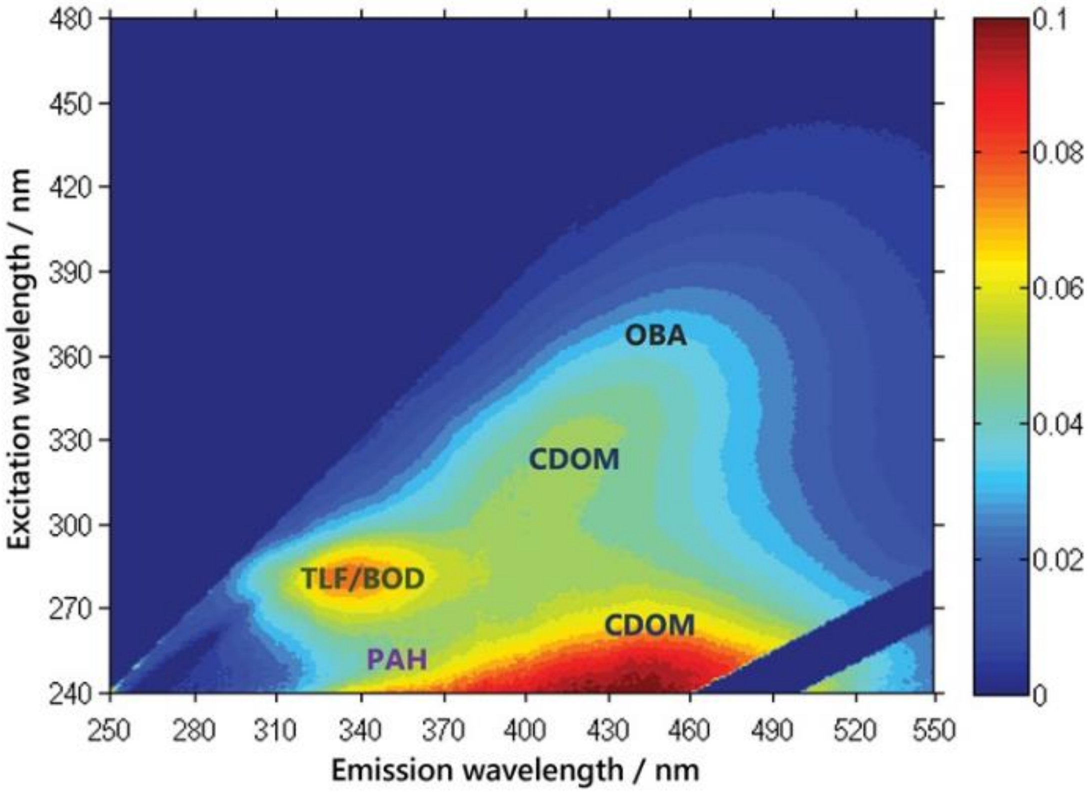


Figure 2: Chelsea Technologies Group Ltd. (2018). Fluorescence excitation-emission matrix of a natural freshwater sample, indicating PAH, TLF/BOD, CDOM and OBA [Diagram]. *UviLux Sensor datasheet* [Online]. Available from: https://d3vx6ukbh3y10k.cloudfront.net/images/Datasheets/2271-003-PD-N-UviLux.pdf [Accessed: 29/07/2018].

Nonetheless, TLF present considerable advantages that could lead to significant enhancement of water quality monitoring protocols. The key advantage of this method is that it is extremely portable (See Figure 3) and provides real-time readings (less than 10 seconds). Although the sensors currently remain costly, this method doesn't require any additional costs such as reagent costs. Finally, they are easy to manipulate, do not require any scientific expertise, and could potentially be used by community themselves to monitor water quality. In past studies such as Sorensen *et al.*, 2018, a threshold of 1.3 ppb dissolved tryptophan was found effective to infer faecal contamination while minimizing false negatives. It still yielded a significant level of 18% false positives (Sorensen *et al.*, 2018b).

Figure 3: Fieldwork picture of TLF and CDOM sensors, which can be easily transported in a bucket

1.2.2. Study area

a. <u>Geophysical context</u>

The study area is the Thiaroye shallow aquifer (See Figure 4), which covers most of Dakar peri-urban area, with densely populated areas such as Guediawaye and Pikine as well as rural and agricultural areas (Sangalkham). This whole area is located on the Cape Verde peninsula, continental Africa's westernmost part. This peninsula, surrounded by the Atlantic Ocean, comprises two topographic domes in the West and the East, separated by the Rufisque-Sangalkham graben. It belongs to the Senegal-Mauritanian sedimentary basin with Tertiary igneous rocks covered by Quaternary sediments (Diongue, 2018). The Quaternary Sands aquifer is part of the superficial aquifer system of the north coast, and this reservoir is based on a marly substratum of Tertiary age. It also encloses the infrabasaltic aquifer located at the head of the peninsula, that extends eastwards in the form of the Thiaroye aquifer (Cissé Faye *et al.*, 2004).

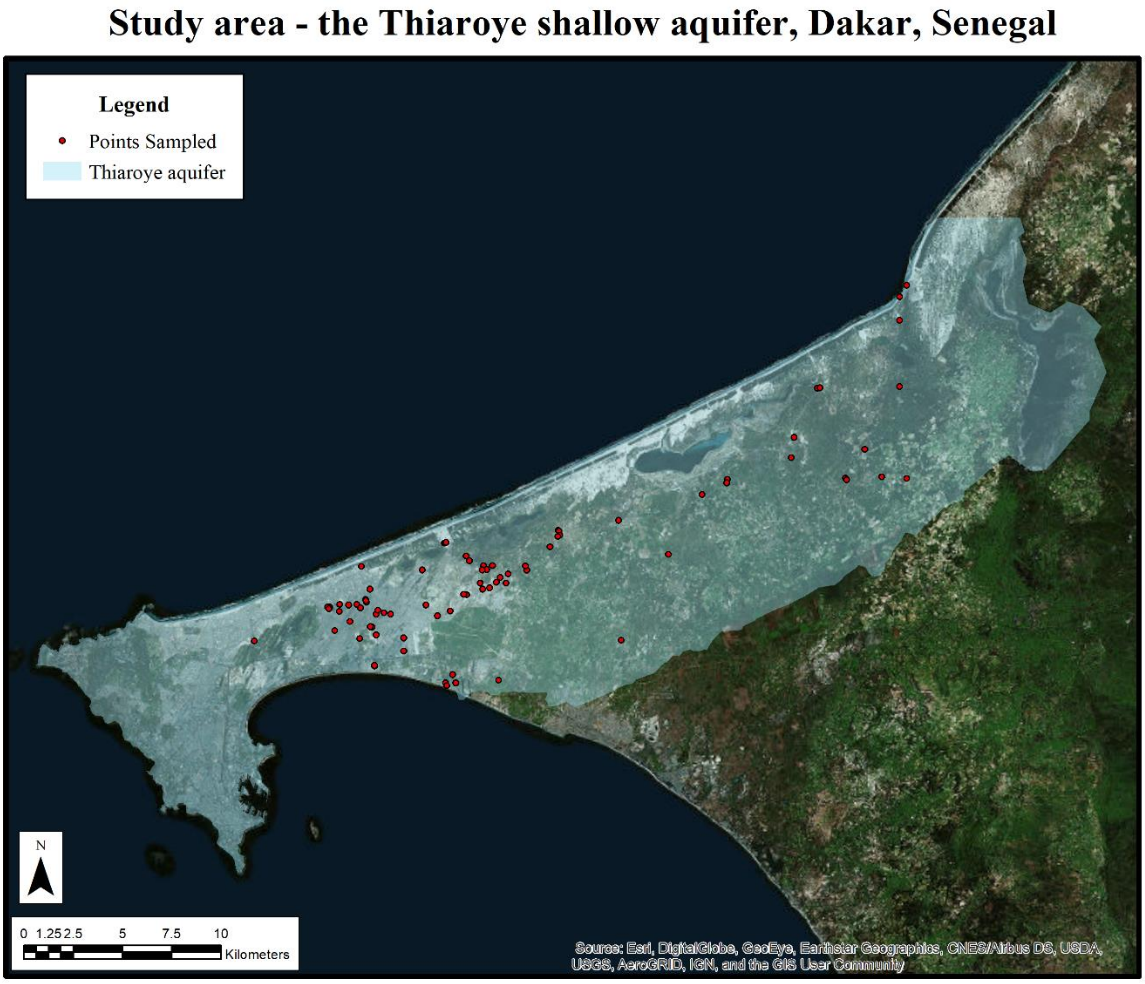


Figure 4: Map of the study area

b. <u>Pollution sources</u>

The Thiaroye aquifer is remarkable in that it has undergone multiple strong sources of anthropogenic pressure in the recent past. Exploited for decades to provide 80% of the water consumed across the Dakar region, it has also absorbed extremely high levels of pollution due to the absence of proper wastewater treatment infrastructure. Wastewater and human faeces were essentially discharged straight into the aquifer, leading to unprecedented degradation in water quality (Re *et al.*, 2011). A major issue is that only central Dakar is connected to a sewerage system, whereas the most densely populated areas such as Guediawaye rely on on-site sanitation facilities, mostly septic tanks. However, due to the high density of housing and very narrow streets, a large proportion of septic tanks in the Pikine/Guediawaye area are not accessible to desludging trucks. 52% of septic tanks in the Dakar region are essentially never emptied or are manually emptied, a practise that poses serious threats to health and the aquifer (Office National de l'Assainissement du Sénégal ONAS, 2015).

In addition to these trends, agricultural practices further increase nitrate concentration in groundwater, with levels reaching 200-800mg/L, far exceeding WHO standards of 50mg/L. These nitrate levels are a direct result of urban sewage and the use of fertilizers (Re *et al.*, 2011). Additional sources of nitrates include industrial activities (Wakida and Lerner, 2005), cemeteries (Pacheco *et al.*, 1991) and landfills (Mor *et al.*, 2006). Finally, seawater infiltration also contributes to the degradation of groundwater quality in the Thiaroye aquifer with salinization of freshwater sources. This has led local authorities to stop the exploitation of the aquifer for the production of drinking water, which had become too costly to treat. Tap water is now drawn from the Louga region, 250km North from Dakar.

c. Water consumption habits

Across the study area, groundwater abstraction is mainly carried out by the population with handpumps and dug wells. Handpumps can be easily bought and installed by households, as the aquifer is very shallow (3-8m). But often, it is installed too close to sanitation facilities, and does not respect the minimal distance of 10m recommended by WHO between septic tanks and drinking water sources (Viraraghavan, 1978; WHO, 1992; Bartram and Ballance, 1996).

Most of the population does not use groundwater as their main drinking water source (Eggleton, forthcoming). However, abstracted groundwater serves multiple purposes: construction work, house cleaning, car cleaning, cooking, hygiene, drinking water for animals, etc. Cooking and hygiene are problematic because they can lead to the ingestion of harmful pathogens. Besides, water supply being unreliable and not constant, households occasionally resort to using groundwater as backup source of drinking water. Finally, the poorest fraction of the community who cannot afford a tap water connection or buying water sachets and bottles have no choice but using groundwater as their main source of drinking water.

1.3. Research Questions

The Thiaroye aquifer is a complex system where multiple sources of pollution have mixed for decades, and where a lack of sufficient sanitation infrastructure has led to extremely high levels of nitrates and other nutrients. The AfriWatSan project has set one of its three urban observatories at the Cheikh Anta Diop University in Dakar (UCAD), where researchers in the Hydrogeology, Engineering and Health faculties work across disciplines to assess the vulnerability of this aquifer and groundwater sources to microbiological and chemical faecal pollution, as well as the impact of low-cost and on-site sanitation strategies on urban groundwater and human health. Within the AfriWatSan framework, this study seeks to explore faecal contamination patterns across the aquifer, based on standard culture-based methods and tryptophan-like fluorescence.

Overarching research question:

What can TLF and culture-based methods reveal about the patterns of faecal contamination in the Thiaroye aquifer?

Partial research questions:

RQ1: *Among the various hydrochemical and microbiological parameters collected, what are the main predictors of faecal contamination?*

RQ2: *Is the tryptophan-based, real-time detection method a significant variable when trying to model contamination of the Thiaroye aquifer?*

RQ3: *What is the predictive power of the tryptophan-based method? How do various environmental factors affect its reliability?*

RQ4: *What is the overall predictive power of a contamination model based on a selection of significant parameters?*

RQ5: *Does the faecal contamination demonstrate spatial patterns, and can it be classified?*

Hypothesis:

After decades of pollution, the Thiaroye aquifer is very rich in nutrients and debris from past contamination. This may lead to high levels of dissolved organic matter, and potential interference with the real-time TLF readings. A combination of other parameters may be used as a proxy to model contamination across the aquifer.

2. METHODS

2.1. Fieldwork

2.1.1. Overview

Data were collected over a period of 5 weeks, from May 28th to July 3rd, 2018, with 15 effective days of fieldwork. The study area covers the greater Dakar administrative region, which includes suburban, peri-urban, industrial and rural landuse.

The water sources sampled include handpumps ("*pompes diambar*"), piezometers, dug wells and one borehole (See Figure 5). At each sampling station, water was first purged for a minimum of one minute for frequently used handpumps, up to 20 minutes in the case of piezometers, until hydrochemical parameters stabilized. These parameters were recorded using a Hydrolab Quanta Multiparameter water quality probe. They include pH, Temperature, Turbidity, Salinity and Conductivity. Unfortunately, turbidity could not be properly calibrated due to a lack of calibration solutions. Geographic coordinates were also recorded in the WSG84 datum, UTM zone 28N, using a Garmin eTrex® Basic GPS.

A sanitary risk assessment was systematically conducted, based on the WHO sanitary risk assessment forms (WHO, 1997, See Appendix 2), and photos were taken to document the context (see Appendix 3). TLF and CDOM were measured and samples were collected at each sampling station.

Handpump

Dug well

Piezometer

Borehole (with electric pump)

Figure 5: Photos of the four types of sources sampled

2.1.2. TLF & CDOM

Two Chelsea Technologies Group UviLux fluorometers were used to measure Tryptophan-like fluorescence (TLF) and Coloured Dissolved Organic Matter (CDOM), respectively.

The CDOM sensor was manipulated using factory calibration, in which the manufacturer cross-correlated each calibration solution against a reference standard of quinine sulphate, measured on a bench-top spectrofluorometer. The sensor measures fluorescence in Quinine Sulphate Units (QSU), which corresponds to the fluorescence intensity recorded from a quinine sulphate concentration of 1μg/100mL at 347.5 nm excitation wavelength and 450 nm emission wavelength (Chelsea Technologies Group Ltd, 2018).

The TLF sensor also provides readings in QSU but it measures fluorescence at excitation wavelength of 280nm and emission wavelength of 360nm. Because it had been used for over a year since factory calibration, this sensor was re-calibrated prior to the fieldwork using laboratory grade L-tryptophan (Acros Organics, USA) dissolved at different concentrations in ultrapure water (Sorensen et al., 2015). A strong linear relationship ($R^2 = 0.999957$) was found between the tryptophan concentration in μg/L and the TLF sensor reading (See Figure 6). TLF concentration data in QSU were extrapolated using the calibration trendline equation (1).

$$C(Tryptophan) = 1.042437x + 1.748532 \qquad (1)$$

with $x = TLF\ sensor\ reading$

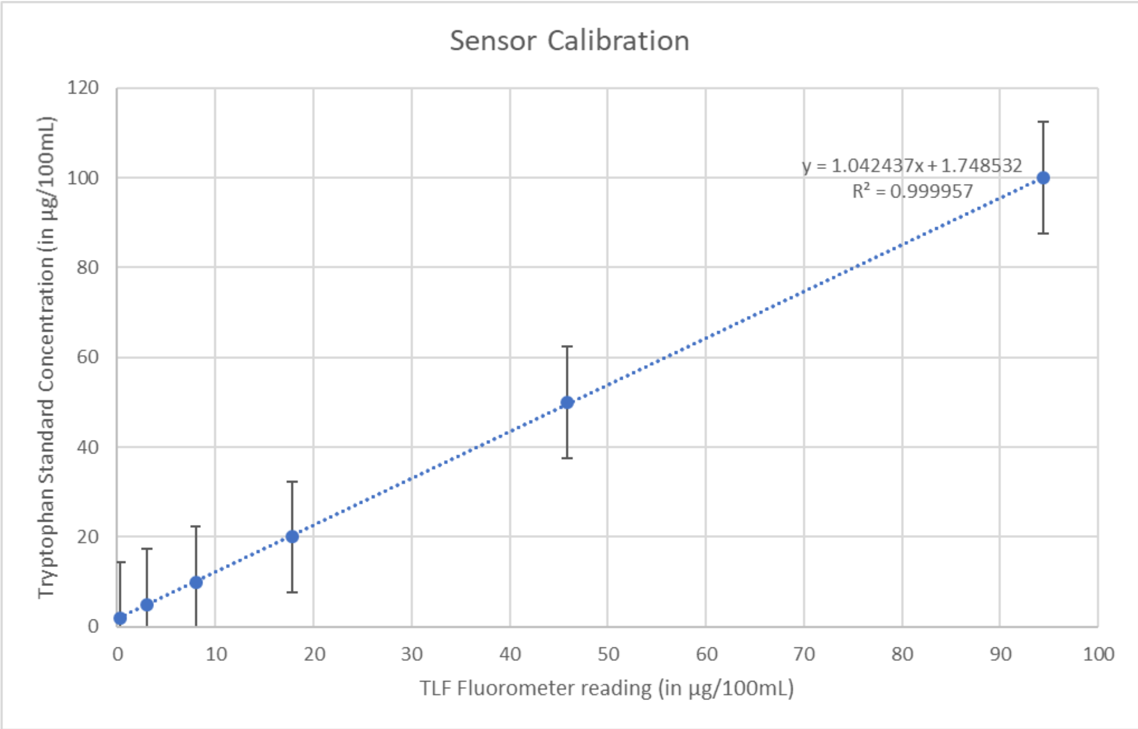


Figure 6: Calibration of the TLF sensor

At each source, a beaker was rinsed three times with the water being sampled, filled with roughly 100mL of water and positioned in a bucket. The TLF sensor was also thoroughly rinsed and introduced in the beaker, then covered for dark. As the fluorometer reading is not static, three readings were logged each time. The operation was repeated three times, or more in cases where values were significantly different (over 1.00 QSU difference), until closer readings were found. This helped control for accidental contamination of the beaker or fluorometer. The exact same protocol was followed for CDOM.

At 32 out of the 97 water sources sampled, 1L of water was filtered using a sterile, gamma irradiated PVDF Sterivex-GV pressure unit with a membrane of 0.22 µm pore diameter. TLF and CDOM were measured on the filtered water. Using this filtration method, 30mL samples were also collected to measure dissolved organic carbon (DOC) using a Thermalox™ carbon analyser after acidification and sparging at the Centre for Ecology & Hydrology in Wallingford, UK.

2.1.3. Samples

2mL samples were collected in 5mL vials using a 1mL pipette with sterile tip, and were preserved using glutaraldehyde and pluronic F68, a non-ionic surfactant with respective final concentrations of 1% and 0.01% (Marie, Rigaut-Jalabert and Vaulot, 2014). These preservatives prevent cell loss or cell proliferation during storage. 100mL samples were collected directly from the source in sterile bottles rinsed three times using water from the sampling source. Both 100mL bottles and 5mL vials were kept in a cool box containing ice in order to prevent bacteriological proliferation and transported back to the laboratory within the next 8 hours to be stored respectively in the freezer and in the fridge. Whereas the vials were subsequently shipped to the UK for flow cytometry to be conducted (Hammes *et al.*, 2008), the 100mL bottles were immediately used to conduct microbiological tests to detect thermotolerant coliforms. Flow cytometry was carried out using fluorescence excitation-emission matrix (EEM) spectroscopy by the British Geological Survey in Wallingford, UK.

Thermo-tolerant coliforms were analysed using the membrane filtration method with membrane lauryl sulphate broth (MLSB) as a reagent (Bartram and Ballance, 1996; Sartory, 2009). Blanks were made before and after all test plates to control for cross-contamination. Results were available after 18 hours of incubation at 44°C and were read by counting the number of pink colonies on the plate, giving a TTC result in CFU/100mL (colony-forming units per 100mL). Orange and yellow colonies were ignored but indicate the presence of another kind of bacteria. When the cultures on the plate were too numerous to count, the test was repeated using a smaller volume of water filtered through the membrane (Sartory, 2009).

Finally, as part of the AfriWatSan biannual aquifer monitoring campaign, 200mL sterile glass bottles were also collected on 45 sites to analyse nitrates (NO_3^-) and phosphates (PO_4^{3-}). Analysis was carried out using an Agilent Cary 60 bench spectrophotometer, with cadmium reduction method at 540 nm wavelength for nitrates and antimony-phosphomolybdate blue method at 882 nm wavelength for phosphates.

2.2. Data processing

2.2.1. Available Data

Data collected on the field with pen and paper was entered on Microsoft Excel as "SampleData.csv". Table 1 summarizes the parameters collected and their role in assessing water quality.

Additional data were obtained from the Cheikh Anta Diop University of Dakar (UCAD) Hydrogeology Department. They include aquifer boundaries and administrative boundaries provided as ESRI shapefiles. Finally, population count by administrative units was obtained from the National Statistics and Demography Agency (ANSD).

Table 1: Description of the dataset variables

Variable	*Variable Type*	*Data transformation if any*	*Role*
ID	Ordinal	Removed for the analysis	Single identifier
NAME_4	String	Removed for the analysis in order not to work at aggregated level	Administrative Unit
TTC	Integer	Log10 Contamination (0 or 1)	Indicator organism for faecal contamination
TLF	Numeric	TLF concentration data in QSU extrapolated from the calibration trendline equation	Potential indicator of faecal contamination
Type	Categorical	Turned into numeric categories	Source type (handpump, dug well, etc.)
Rain	Binary	Ready to use	Separate points sampled before and after the rain
x	Numeric	Ready to use	Longitude
y	Numeric	Ready to use	Latitude
PopDensity	Numeric	Ready to use	Proxy for the discharge of faeces in groundwater
Conductivity	Numeric	Ready to use	Hydrochemical parameter.
pH	Numeric	Ready to use	Hydrochemical parameter
Temperature	Numeric	Ready to use	Hydrochemical parameter
Salinity	Numeric	Ready to use	Hydrochemical parameter
Turbidity	Numeric	Ready to use	Hydrochemical parameter
FC	Integer	Log10	Flow Cytometry data

CDOM	Numeric	Ready to use	Indicator of Dissolved particles, including carbon
DistanceToCemetery	Numeric	Extracted from Open Street Map	To assess influence of environmental factors
DistanceToFarm	Numeric	Extracted from Open Street Map	To assess influence of environmental factors
DistanceToIndustry	Numeric	Extracted from Open Street Map	To assess influence of environmental factors
DistanceToLandfill	Numeric	Extracted from Open Street Map	To assess influence of environmental factors
DistanceToRoads	Numeric	Extracted from Open Street Map	To assess influence of environmental factors
Sanitation	Binary	Extracted from sanitary risk form	Presence of sanitation facilities within 10m
SepticTank	Binary	Extracted from sanitary risk form	Presence of a septic tank within 10m
SoakPit	Binary	Extracted from sanitary risk form	Presence of a soak pit within 10m
Latrines	Binary	Extracted from sanitary risk form	Presence of latrines within 10m
Cattle	Binary	Extracted from sanitary risk form	Presence of cattle on the area
Trash	Binary	Extracted from sanitary risk form	Presence of trash or landfill
Cultivation	Binary	Extracted from sanitary risk form	Presence of agricultural activities
Construction	Binary	Extracted from sanitary risk form	Presence of construction works in the area
Road	Binary	Extracted from sanitary risk form	Presence of a road in the vicinity
Petrol station	Binary	Extracted from sanitary risk form	Presence of a petrol station in the vicinity
Drainage channel	Binary	Extracted from sanitary risk form	Is there a drainage channel?
Fence	Binary	Extracted from sanitary risk form	Is the source covered by a fence, when applicable?
Apron area	Binary	Extracted from sanitary risk form	Is there an apron area?
Pump insanitary	Binary	Extracted from sanitary risk form	Is the pump insanitary?
CracksLoose	Binary	Extracted from sanitary risk form	Is the pump cracked or loose at the base?
TotalRisk	Integer	Extracted from sanitary risk form	Sum of all risk indicators (/10)
TLF_filtered	Numeric	Missing data; used in a subset	TLF measured on filtered samples
CDOM_filtered	Numeric	Missing data; used in a subset	CDOM measured on filtered samples
DOC	Numeric	Missing data; used in a subset	Dissolved Organic Carbon
Nitrates	Numeric	Missing data; used in a subset	Nitrates
Phosphates	Numeric	Missing data; used in a subset	Phosphates
Repeat	Binary	Ready to use	Was this point sampled twice?
Date	Date	Removed for the analysis (irrelevant)	Date of sampling
Time	Time	Removed for the analysis (irrelevant)	Time of sampling

2.2.2. Data preparation in Excel

In order to ensure reproducibility of the method, the analysis was carried out using open-source software: QGIS 2.18.15 and R version 3.4.3 through the RStudio interface (R Development Core Team, 2013). The entire code generated for the analysis is available on GitHub (See Appendix 4).

A very first step consisted in transforming the data in the format most suited for the analysis. Subsequent logistic regression model required all data to be in a numeric format rather than string or factor. For instance, the "Sanitation type" column, which included Septic tanks, Soak Pits and Latrines, was divided into three columns "Septic Tanks", "Soak Pits" and "Latrines", filled with 0 or 1. This was done using the IF function in Excel.

2.2.3. Geographical data extraction and processing in QGIS

a. <u>Extracting distance to features of interest</u>

As a first step, the "SampleData.csv" dataset was added as a Delimited Text Layer, using x and y as the Easting and Northing coordinates in the UTM 28N coordinate reference system (WSG84 datum). The result is a "Samples" vector layer containing all sampled points and their attributes from the CSV file. The Thiaroye aquifer shapefile was also added and the map was centred around the study area, leaving additional space to the West and the South.

OpenStreetMap data were then downloaded for this map canvas and saved as an XML file (BaseMap.osm), subsequently transformed into a SpatiaLite DB file (BaseMap.osm.db). From this SpatiaLite DB file, topologies of interest were exported and added onto the map: polylines tagged as "highways", which actually encompass all roads, and "Landuse" polygons were added to the map.

Selecting polygons based on their attribute in the "Landuse" attribute table, layers were created for Landfills, Agricultural (combining Farmland, Farmyard, Greenhouse_horticulture, Orchard and Plant_nursery), Industrial and Cemeteries (Pacheco *et al.*, 1991; Wakida and Lerner, 2005; Mor *et al.*, 2006). Because OpenStreetMap data comes in the pseudo-Mercator coordinate reference systems, it needed to be reprojected to the project's coordinate reference system. To this end, the four different landuse layers and the Roads layer were saved as ESRI shapefiles in the UTM 28N coordinate reference system and loaded in the project.

Using the *NNJoin* plugin (QGIS, 2016), each of these five new layers were joined to the "Samples" layer; *NNJoin* indicates the distance of each sample point to the nearest road, industry, farm, cemetery or landfill, in meters. Attribute Tables of the joined layers were then exported as spreadsheets using the *XY Tools* plugin (Duivenvoorde, 2011), and combined with the original "SampleData.csv" file.

b. Extracting population density

The population count table provided by the Senegalese government contained a breakdown of population count by commune. However, these names did not exactly match those of the administrative units ESRI shapefile. Therefore, the shapefile was first extracted by simply copy-pasting the attribute table in the population spreadsheet. The table contained three columns: polygon geometries, name of the administrative unit and surface of that unit.

Names were matched using the Excel INDEX function, helping add a fourth column for population, as well as a population density column to the original attribute table (See Table 2)

Table 2: Example row of the Population table for the Guinaw Rail Nord commune

Geometry	Name	Surface	Population	Population Density
[Polygon coordinates]	Guinaw Rail Nord	80	40694	$\frac{Population}{Surface} = 508.675$

This "Population" table was saved as a CSV file and imported into QGIS as a Delimited Text Layer using Well-Known-Text as geometry definition, so that QGIS directly recognized the polygons. Coordinate Reference System was set to that of the project, UTM 28N.

Finally, a spatial join was operated between the "Population" layer and the "Samples" point layer. As a result, each point in the Population attribute table was assigned four new attributes: administrative units, area of this unit, population of this unit and most importantly: population density.

2.2.4. Exploratory Spatial Data Analysis in R

Conducting an Exploratory Spatial Data Analysis (ESDA) entails running a set of simple summary statistics and data visualisations in order to detect patterns and identify features of interest in a dataset (Haining, Wise and Ma, 1998). When exploring contamination patterns across the Thiaroye aquifer, ESDA enables the identification of potential issues with the data, which supports the selection of the most relevant modelling technique. This step is crucial in developing formal hypotheses regarding the data.

a. <u>Statistical Analysis</u>

The "SampleData.csv" dataset was first imported to R and the two contamination method result variables – TLF and TTC – were plotted using the **ggplot** package (Wickham, 2009). TLF data is normally distributed, although it showcases a few outliers. TTC data, however, is heavily skewed to the left and required to be normalized before any further analysis. This was done using a logarithmic scale (See Figure 7).

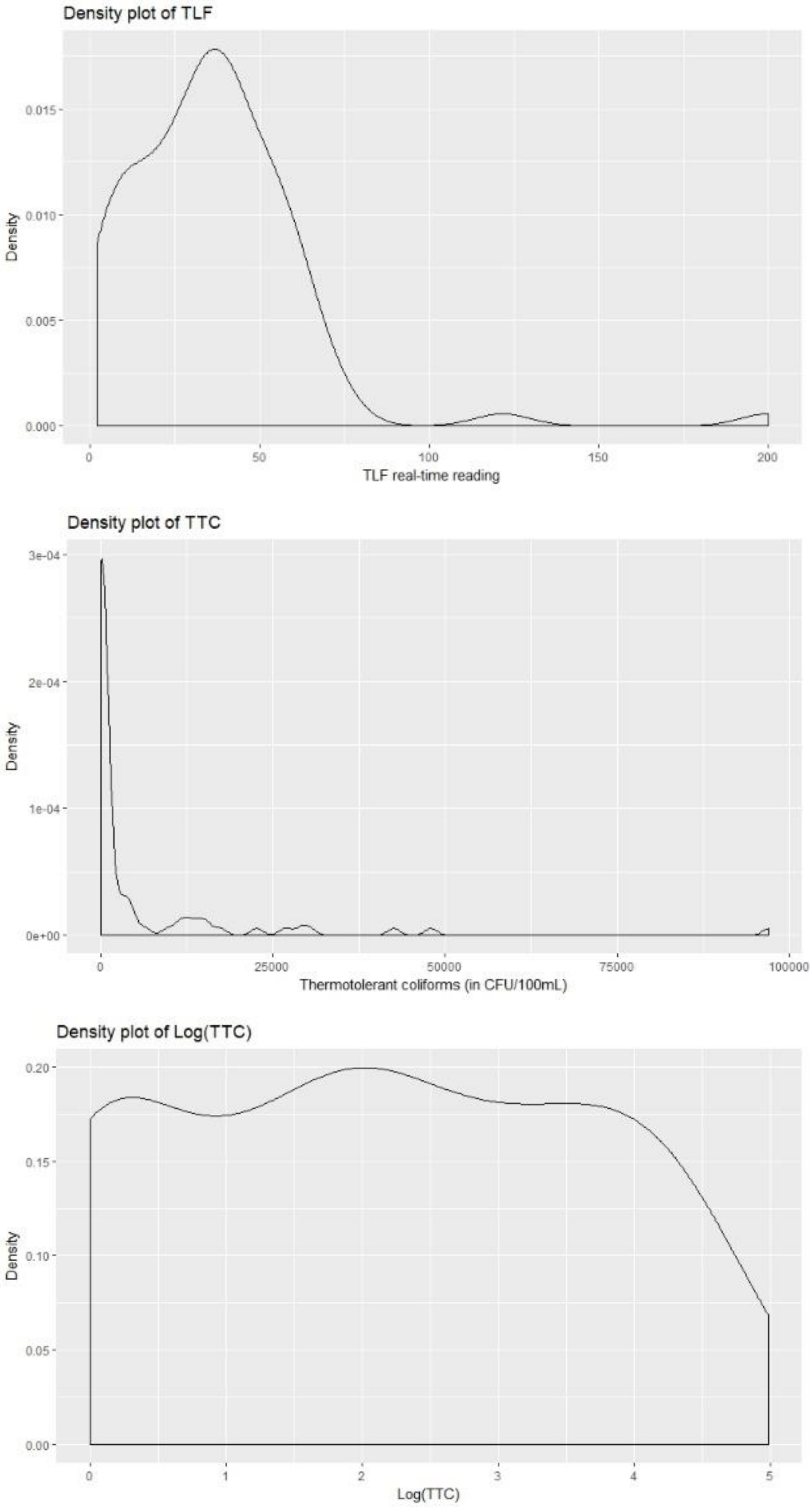


Figure 7: Density plots of TLF, TTC and LogTTC

A new column "LogTTC" was added to the SampleData data frame, containing logarithmic data for the TTC values. When TTC was null, it was replaced by 1 in order for the logarithmic scale to be applied. Another column was created to record contamination as a Boolean data. Due to the small number of strictly negative TTC results, a contamination threshold of LogTTC < 1 (or TTC <10 CFU/100mL) was set, below which the sample was considered not contaminated. For LogTTC values equal to or greater than 1, the sample was marked as contaminated.

b. <u>Data visualisation</u>

Data was then visualized geographically in QGIS. Figure 8 illustrates the study area and the location of the 97 sample points, sorted by source type and contamination status (positive or negative). Due to the scale of the map, some points are shown to overlap; transparency is therefore activated. No clear contamination pattern seems to emerge from this first visualisation, but dug wells appear to be the most contaminated source type.

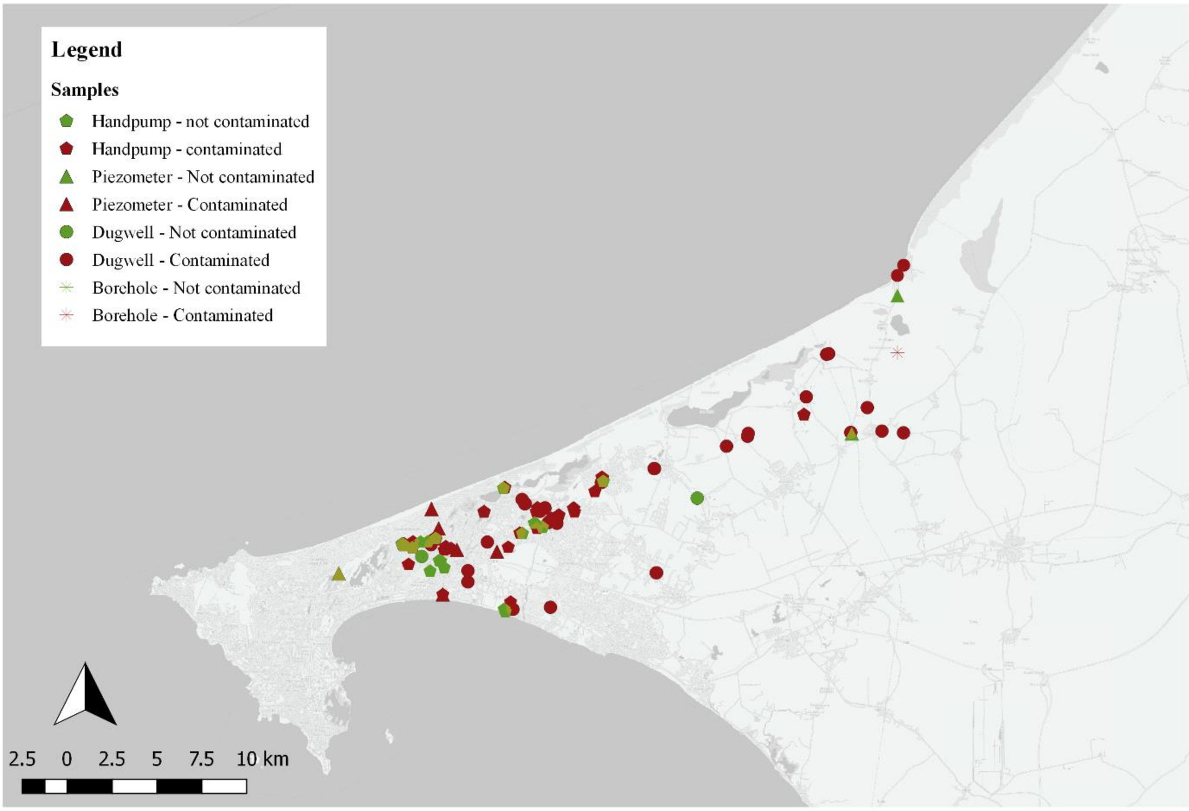


Figure 8: Contamination status of samples by source type across the study area

c. Descriptive statistics

Boxplots help visualize the distribution and range of a variable against another. Figure 9 displays TLF values against TTC counts, and clearly illustrates a lack of correlation between the two variables. On the other hand, Figure 10 shows a linear relationship between CDOM and TLF and a homogeneous distribution of values across the CDOM variable range, with the exception of two outliers.

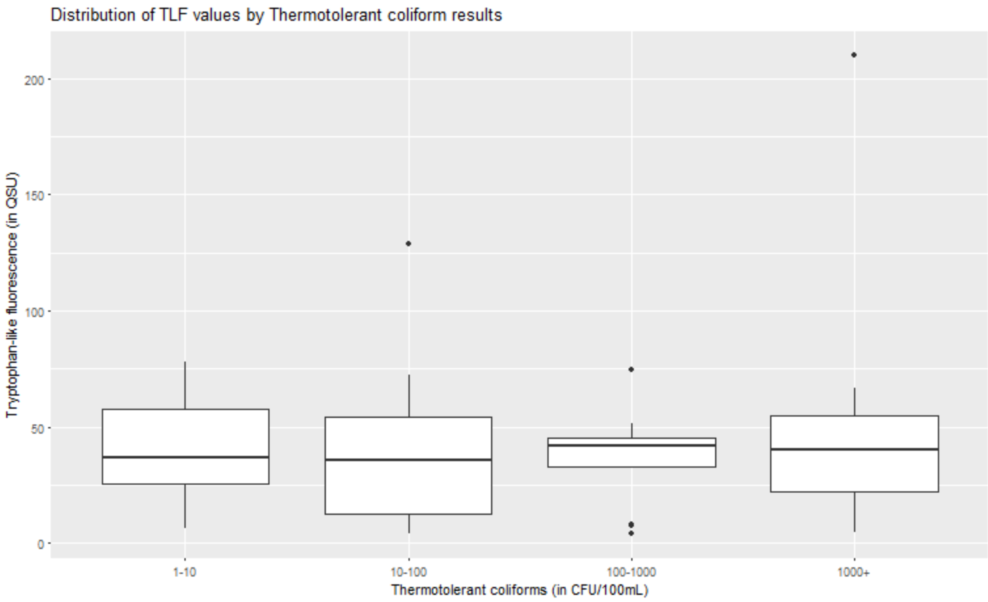


Figure 9: Boxplot of TLF by TTC count

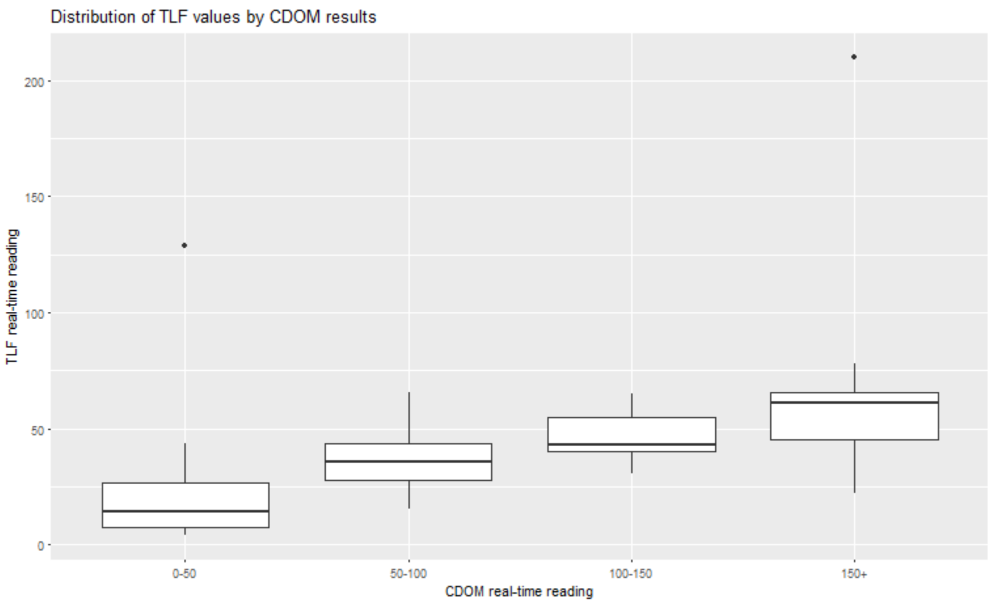
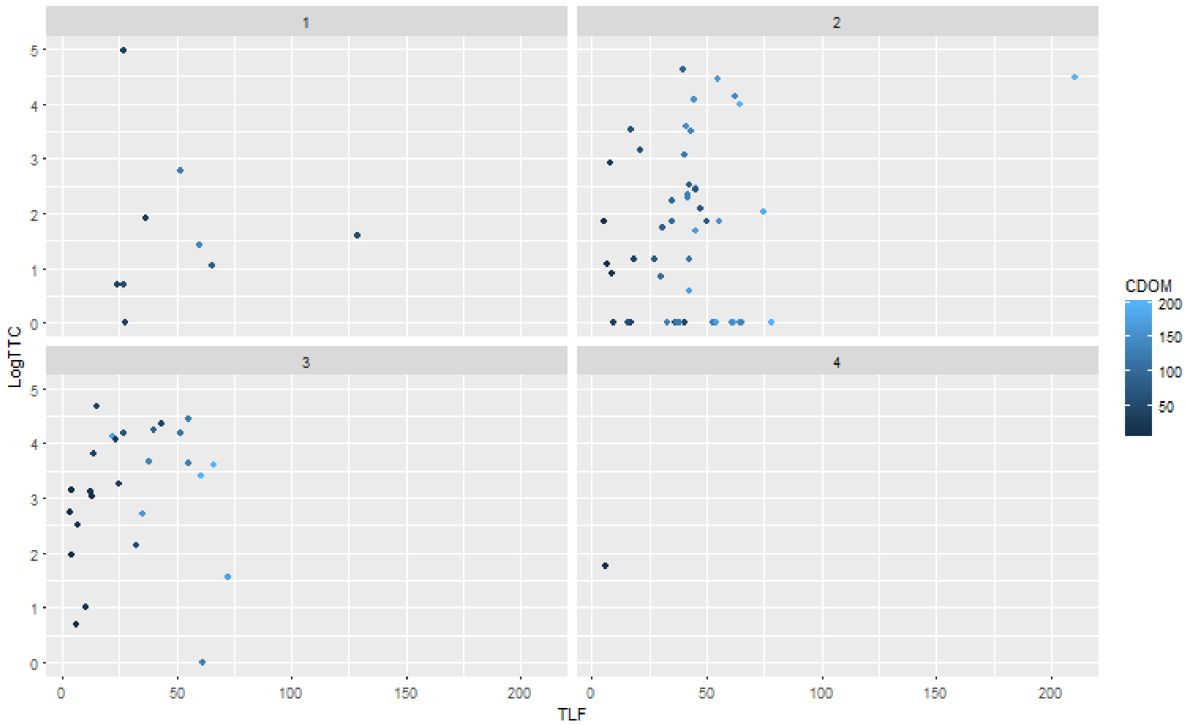


Figure 10: Boxplot of TLF values by CDOM reading

Next, Figure 11 represents LogTTC, TLF and CDOM values plotted by source type. This graph demonstrates that handpumps and dug wells present significantly higher levels of thermotolerant coliforms, but whereas dug wells are almost systematically contaminated, many handpumps are not; in other words, one can expect a dug well to almost certainly be contaminated, but handpumps present more variability in their results. This result is not surprising as dug wells are exposed to many additional sources of pollution (atmosphere, bucket, animals, etc.). But unexpectedly, dug wells and the borehole also present overall lower CDOM and TLF levels than other source types; TLF is not a reliable method to detect TTCs in this dataset.



Legend: 1 = piezometer, 2 = handpump, 3 = dug well, 4 = borehole

Figure 11: Values of LogTTC, TLF & CDOM by source type

d. <u>Correlation matrix</u>

Links between each variable in the dataset are investigated using a correlation matrix with p-values. Because the data does not follow a perfectly gaussian distribution, a non-parametric spearman correlation is run between all pairs of variables in the SampleData data frame, using the **Hmisc** and **corrplot** R packages (Harrell *et al*., 2018; Wei and Simko, 2018).

Figure 12 only displays significant correlations (p-value > 0.01), ranging from strong negative correlation in dark red to strong positive correlation in dark blue. Question marks signify variables for which correlation wasn't computed due to missing values.

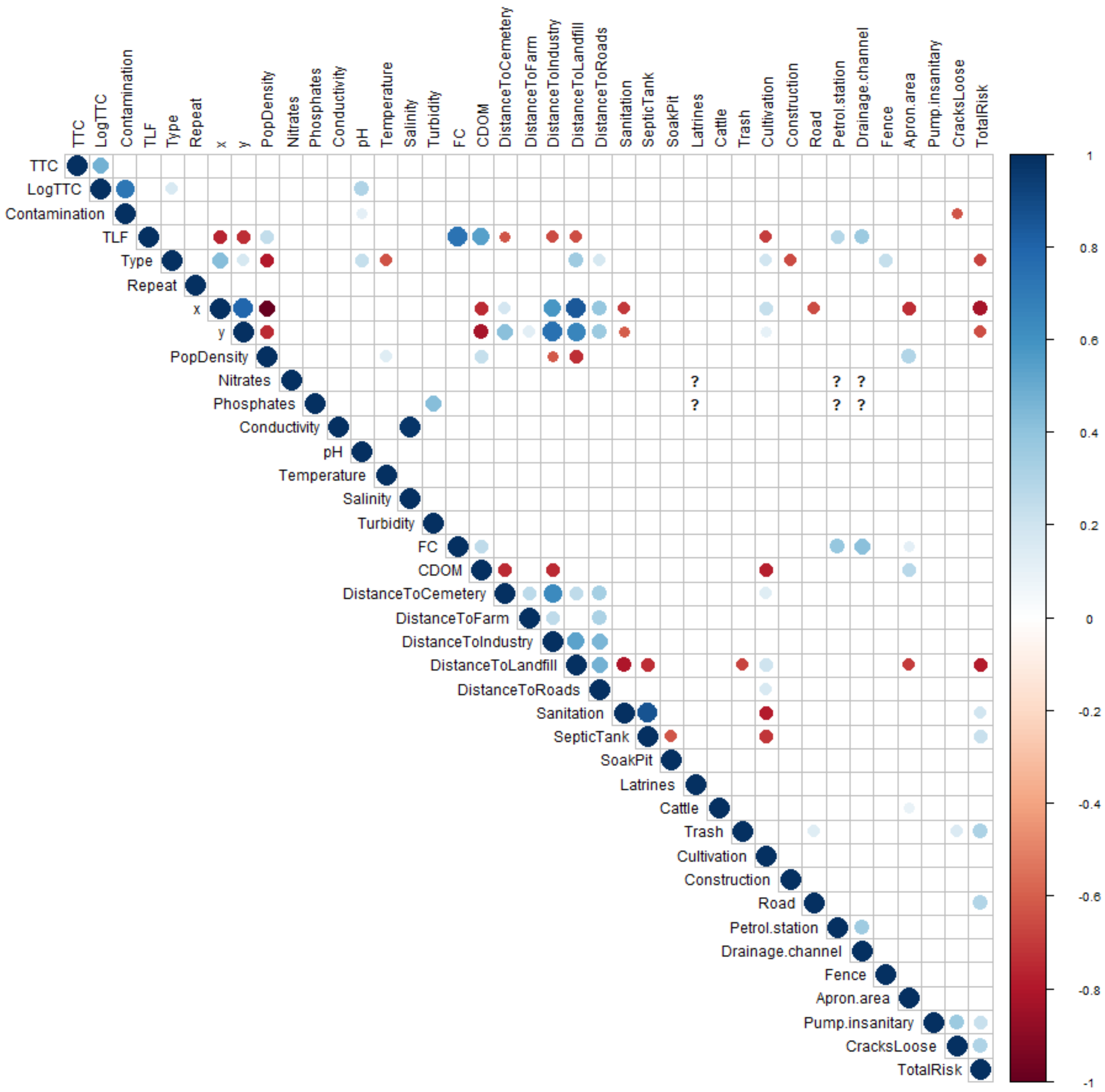


Figure 12: Correlation matrix for each pair of columns in the SampleData data frame

Many correlations are a simple result of collinearity (e.g. TTC and contamination, or Distance to certain features and geographic coordinates), but others are more interesting for the analysis: there is a very strong positive correlation between TLF and CDOM, TLF and Flow

Cytometry results, and between conductivity and salinity. There is also a negative correlation between TLF and distance to cemeteries, industries and landfills, meaning that TLF are higher as proximity to these features increases.

This leads to answering **RQ2:** *Is the tryptophan-based, real-time detection method a significant variable when trying to model contamination of the Thiaroye aquifer?* and **RQ3***:* *What is the predictive power of the tryptophan-based method? How do various environmental factors affect its reliability?*

- ➔ **TLF is not a good predictor of faecal contamination in the Thiaroye aquifer (Spearman rank of $\boldsymbol{\rho = -0.01190626}$ between TLF and TTC).**
- ➔ **TLF is correlated with Flow Cytometry count and CDOM reading, which suggests that the sensor is measuring other compounds.**
- ➔ **TLF is negatively correlated to the presence of cultivation activities, and TLF levels decrease with distance to cemeteries, industries and landfills. This could be due to specific compounds present around these facilities, but it is impossible to conclude with this dataset (further tests and controls would be needed).**

e. <u>Subset analysis</u>

Many statistical techniques cannot be computed with missing values. Subsets of the main dataset were created to look at parameters that were largely incomplete, such as the TLF and CDOM results on filtrated water or the nitrates and phosphates (See Table 3).

Table 3: Data subsets for analysing variables with missing data

Subset	*Variables of interest*	*Aim*
DOCSample	DOC	Look at the DOC-CDOM relationship
PreRainSample	Points 1-60, all variables included	Inspect faecal contamination load at the end of the dry season
PostRainSample	Points 61-97, all parameters included	Inspect faecal contamination load after a heavy rain event
FilteredSample	Filtered TLF and CDOM	Investigate characteristics of TLF and CDOM in relation to the bacteriological load
Nitrates & Phosphates	Points 49-93, on which nitrate and phosphate analysis was performed	Investigate characteristics of nitrate and phosphate pollution in relation to the bacteriological load

DOC Sample:

As expected, a strong positive correlation is found between DOC and CDOM, with a 0.8613518 correlation coefficient. This correlation can be visually plotted (Figure 12).

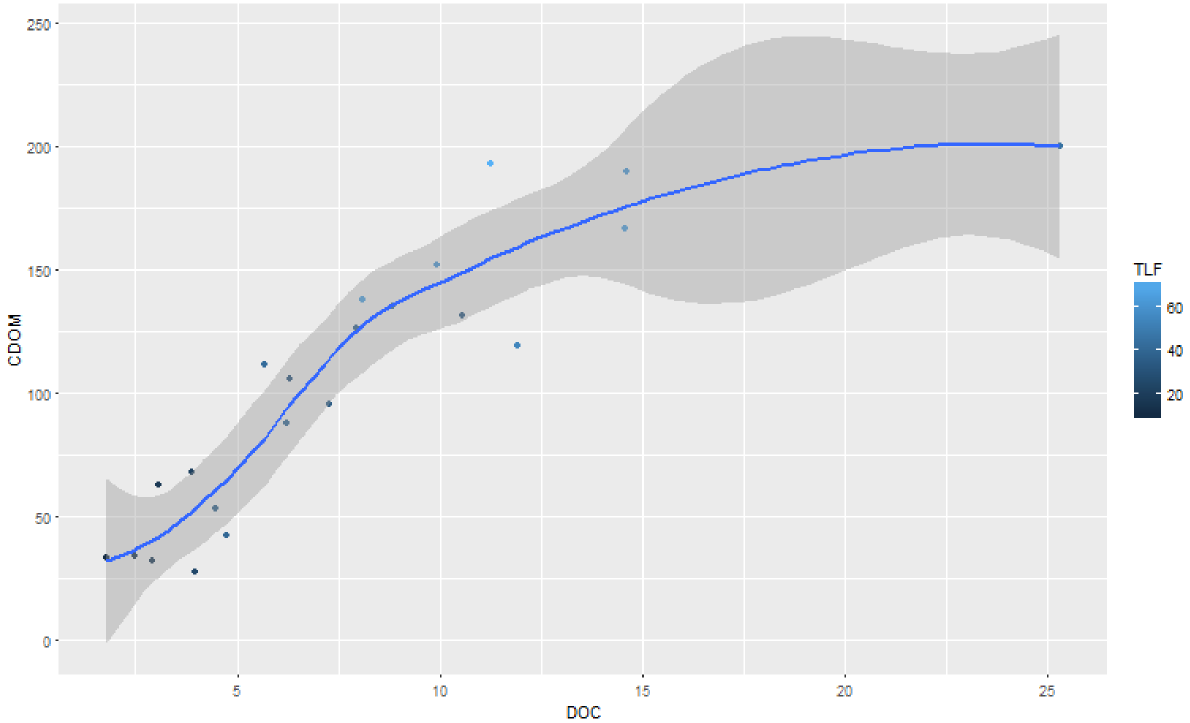


Figure 13: Relationship between CDOM and DOC

Pre-rain / post-rain:

The first rain of the season was an unexpected event; it constitutes a structural break in the data but due to the small number of observations collected, it is impossible to control for it based on a statistical model. Besides, the samples were collected in different parts of the study area, and differences in the results could be caused by many other factors. Table 4 explores summary statistics for the full dataset against pre- and post-rain events, for the variables TLF, TTC and Log(TTC). Significant variability is observed, notably with a narrower range of values for TLF after the rain. It also appears that TTC and Log(TTC) values are significantly higher after the rain both in terms of mean and median, and the highest TTC value of 97,000 was recorded after the rain. However, it is impossible to conclude on any form of causality.

Table 4: Summary statistics of the data before and after a rain event

Variable		*Min*	*Max*	*Mean*	*Median*	*Standard Dev.*
Full dataset	TLF	3.81047	210.23593	27.65824	35.58187	28.86355
	TTC	1	97000	116.2023	120	13779.83
	Log(TTC)	0	4.986772	2.13	2.0603	1.539508
PreRainSample	TLF	3.810	210	37.52	34.844	32.13961
	TTC	1	42600	3106	56	8034.084
	Log(TTC)	0	4.6294	1.802	1.748	1.511206
PostRainSample	TLF	8.518	62.309	35.74	37.46	16.85628
	TTC	0	97000	8738	855	19561.22
	Log(TTC)	0	4.9868	2.653	2.913	1.51530

<u>Filtered Samples:</u>

A very strong positive correlation is found between filtered and unfiltered TLF values ($\rho = 0.9901478$) and filtered/unfiltered CDOM values ($\rho = 0.9394089$). Figures 13 and 14 illustrate these quasi-linear relationships. This supports the hypothesis that the TLF method measures soluble particles, that are not filtered by the bacteriological filter.

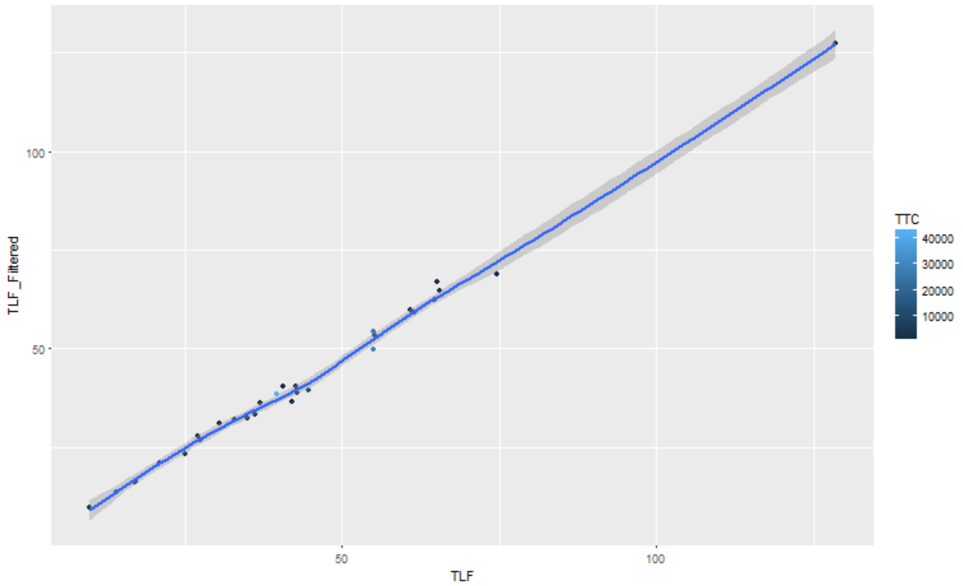


Figure 14: TLF values for filtered / unfiltered samples

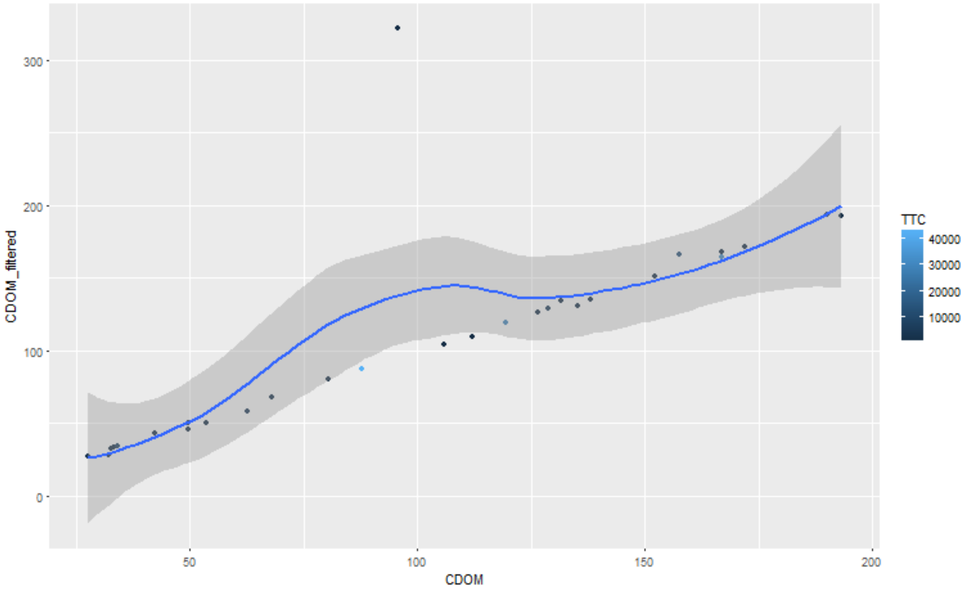


Figure 15: CDOM values for filtered/unfiltered samples

Nitrates and phosphates

Nitrates are directly linked to faecal contamination, since the degradation of nitrogen compounds present in excreta leads to the formation of nitrates, which can infiltrate groundwater (Yates, 1985). Phosphates can be the result of natural mineral decay, stormwater runoff, agricultural runoff or industrial discharges. For both nitrates and phosphates, no significant correlation is found with other variables.

The Exploratory Spatial Data Analysis phase served to explore key characteristics of the dataset and after several iterations through a first set of hypotheses, led to the main research questions defined in section 1.3. The next section seeks to effectively model and map out risks of contamination in the Thiaroye aquifer.

3. DATA ANALYSIS AND RESULTS

3.1. Overview

Before building a spatial model, a logistic model is first considered. Global and local autocorrelation are then observed to determine the most adequate geostatistical modelling approach. Finally, an unsupervised machine learning approach to classification is adopted to identify contamination clusters.

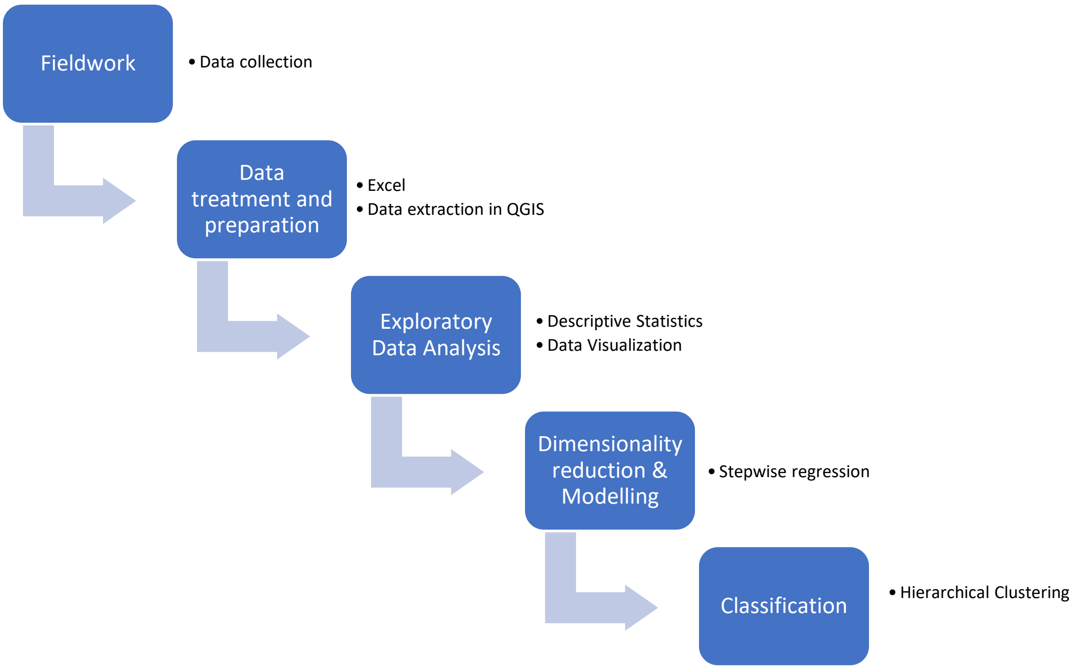


Figure 16: Methods flowchart

3.2. Dimensionality reduction and modelling of contamination status: stepwise logistic regression

Subsequent regression steps require the removal of missing values and NAs in the full dataset; this can be done using averaging, imputing, creating new categories or simply dropping the NAs. In this instance, columns ID, NAME_4, TLF_filtered, CDOM_filtered, DOC, Nitrates, Phosphates, Date and Time are dropped because they are single identifiers, redundant information or largely incomplete data. Rows with missing FC values are also dropped, leaving 84 observations for the full SampleData dataset.

A logistic regression is the most straightforward approach towards answering **RQ2:** "*Among the various hydrochemical and microbiological parameters collected, what are the main predictors of faecal contamination?"*. However, working with high dimensional data implies that parameters are likely to be prone to multicollinearity: certain independent variables are highly intercorrelated (e.g. TotalRisk is the sum of all risks such as Sanitation, Cattle, Trash, etc.). When the number of observations is limited with regards to the number of predictors, this can lead to an unstable estimation of parameter values in the regression model. A stepwise logistic regression is therefore performed to reduce the number of significant parameters included in the regression model.

Because of the range of factors influencing the levels of faecal contamination, logistic regression can yield far more interpretable results in this case than linear regression. Logistic regression considers a categorical variable, in this case: contaminated (1) or not (0).

Data is first scaled, apart from the response variable (contamination) and categorical variables, using the base R `scale()` function. The **caret** package (Max *et al.*, 2015) is then used to perform a five-fold stepwise regression using the `trainControl()` function to divide the data into five training data subsets, and the `train()` function to iterate through them, leaving one out as testing data. A backward strategy to stepwise regression is adopted, whereby the iteration starts with all predictors in the model and iteratively removes the least significant variables until all predictors in the regression are statistically significant (Chen, Goo and Shen, 2014; Bruce and Bruce, 2017). The final set of parameters is such that it minimizes the Root Mean Square Error (RMSE) of the logistic regression, meaning that it lowers the model's prediction error (Kassambara, 2018).

The results lead to answering **RQ1:** *Among the various hydrochemical and microbiological parameters collected, what are the main predictors of faecal contamination?*

- ➔ **The stepwise logistic regression retains 9 parameters: latitude and longitude, presence of a septic tank or latrines in the vicinity of the water source, pH, temperature and turbidity of the sample, flow cytometry count and distance to a landfill.**

And **RQ4:** *What is the overall predictive power of a contamination model based on a selection of significant parameters?*

➔ **This model performs relatively well; when leaving out an observation and trying to predict its contamination status based on the rest of the dataset, it reaches a correct classification 72.22% of the time.**

➔ **Figure 17 visualizes these residuals; 9 out of 10 points fall within the grey lines.**

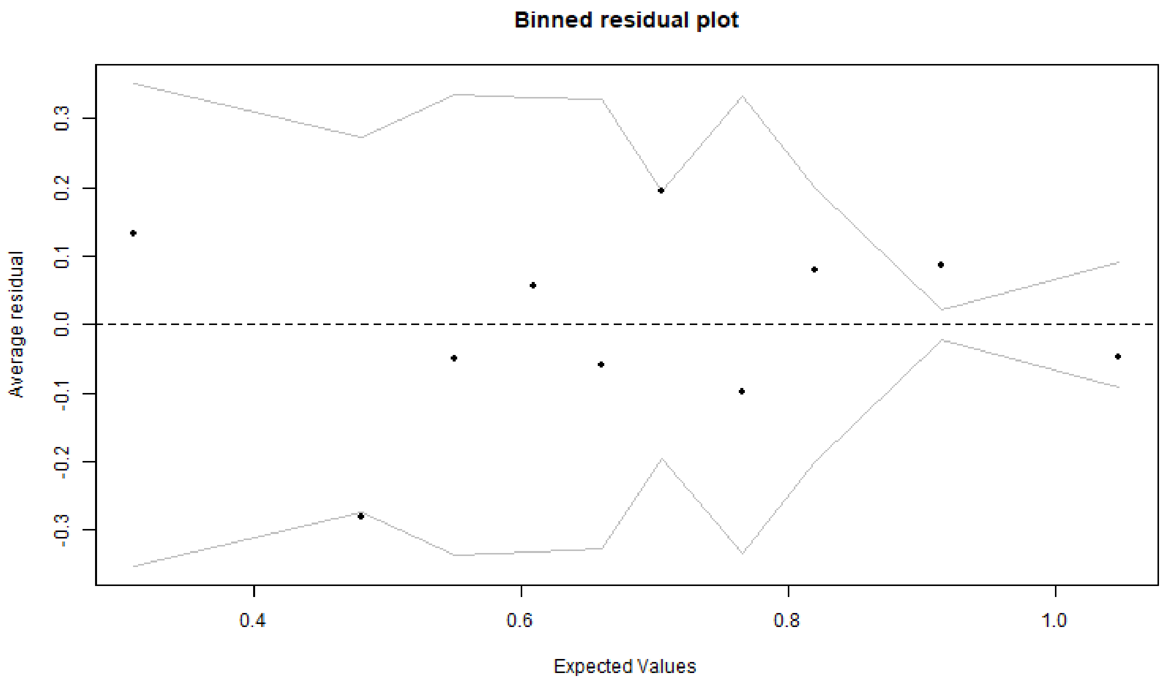


Figure 17: Binned residual plot of the logistic regression model

3.3. Spatial autocorrelation investigation and geostatistical modelling

A geostatistical modelling approach overcomes some of the main drawbacks of regular regressions, that tend to produce too much global smoothing and ignore local effects. In order to avoid the Modifiable Areal Unit Problem (Fotheringham and Wong, 1991), data is not aggregated by administrative unit or other form of spatial unit but rather, analysed as point data. The sample point dataset only provides information about 97 sample events across the study area, including 14 repeats. When possible, handpumps were sampled every 200m (Guediawaye and Pikine districts) but this sampling strategy could not be adopted throughout the study area. In less densely populated areas, the sampling pattern was considerably sparser due to the lack of access to any groundwater source. Interpolation is therefore needed to estimate contamination values at unsampled point locations. Spatial interpolation is the "*procedure of predicting the value of attributes at unsampled sites from measurements made at point locations within the same area*" (Burrough and McDonnell, 1998), and can be used to create continuous surfaces from point data. To determine the best interpolation approach, spatial autocorrelation is first tested with three different methods.

Moran's I test is performed using the `Moran.I()` function in the **ape** package Mantel test using the `mantel.rtest()` in the **ade4** package. Both tests demonstrate that the residuals of the logistic regression, TTC and TLF values are arranged independently across the study area. Finally, a semi-variogram is performed using the **gstat** R package (Pebesma, 2004; Pebesma and Heuvelink, 2016) and displays in the three cases a flat line. Hence, these three tests demonstrate that no spatial autocorrelation pattern emerges (See Figure 18).

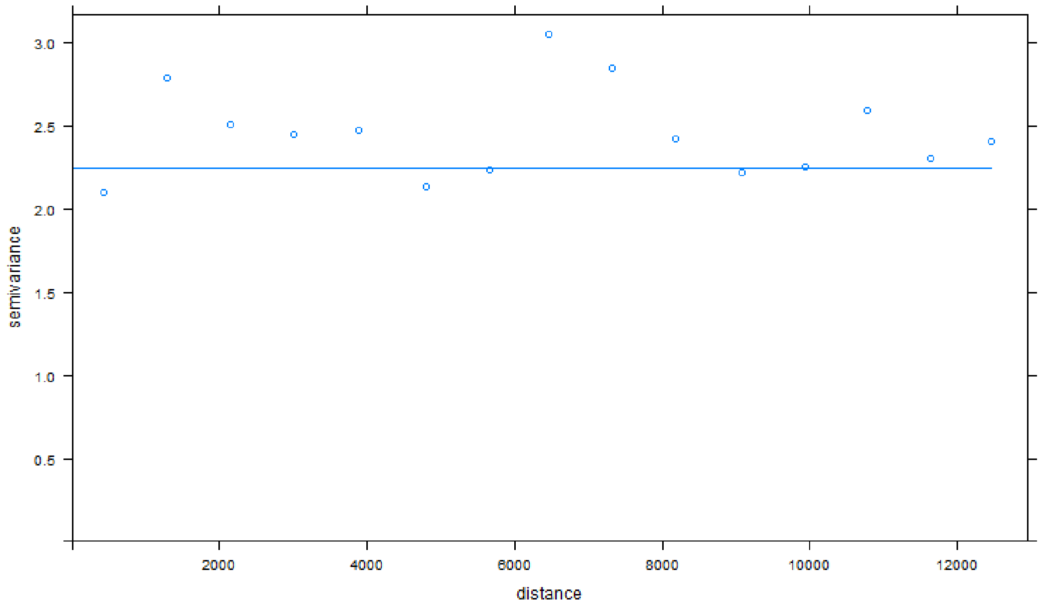


Figure 18: Semivariogram of Log(TTC)

Due to the absence of spatial autocorrelation in the stepwise regression residuals, the TTC variable and the TLF variable, fitting a global variance model such as the Kriging interpolation method or running a Geographically Weighted Regression (GWR) is not relevant (Brundson, Fotheringham and Charlton, 2002). In this instance, Inverse Distance Weighting interpolation (IDW), based on a simpler premise, could produce more simple results that can serve as a visualisation if not a spatial model. IDW is a deterministic, local interpolation method that relies on Tobler's First Law of Geography: "everything is related to everything else, but near things are more related than distant things" (Tobler, 1970 cited in Miller, 2004). IDW is an exact interpolator, meaning that sampled values are preserved. In IDW, the interpolated value z is an average of all sampled values x, weighted by the inverse square of their distance d to the unknown value (Longley *et al.*, 2015). This can be expressed as equation (2).

$$z(x) = \frac{\sum w_i z_i}{\sum w_i} \tag{2}$$

$$\text{With } w_i = \frac{1}{d_i^2}$$

A tessellated surface, in the form of Thiessen polygons (See Figure 19), is created from the sample points using the `spatstat()` function in the R package **spatstat**. A grid is then defined based on the extent of the sample dataset, and IDW is performed with the **spatstat** `idw()` function to interpolate the values of LogTTC and TLF variables across this grid. The result is a raster layer in which every pixel indicates a predicted value of the interpolated variable (See Figure 20).

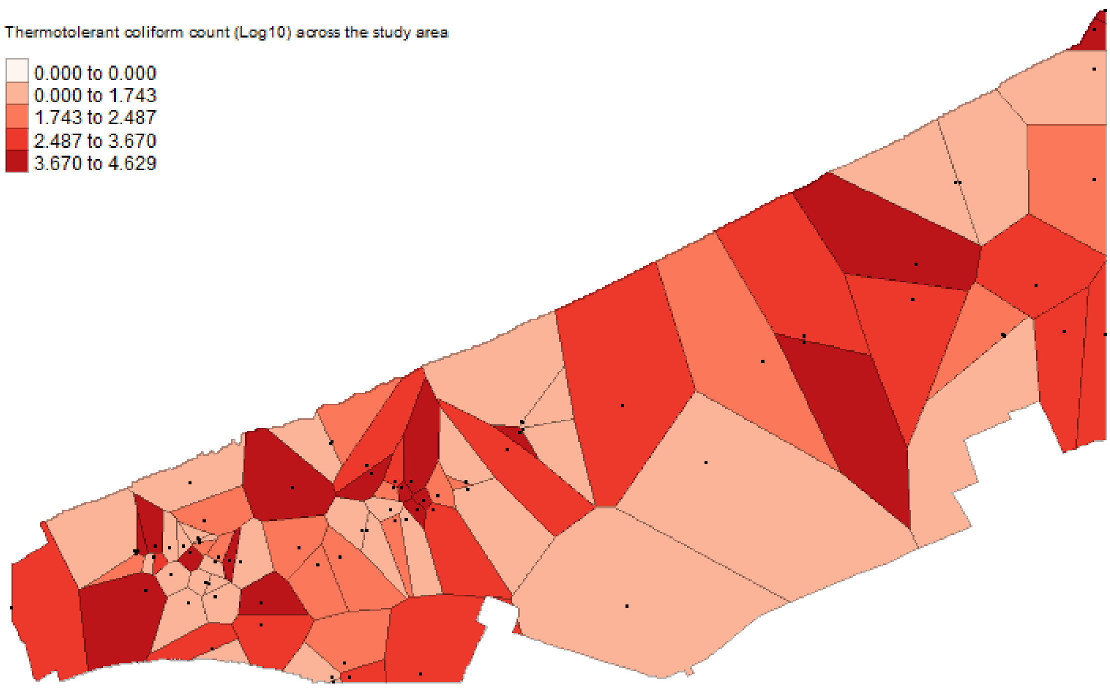


Figure 19: Thiessen Polygons for Log(TTC) values across the study area

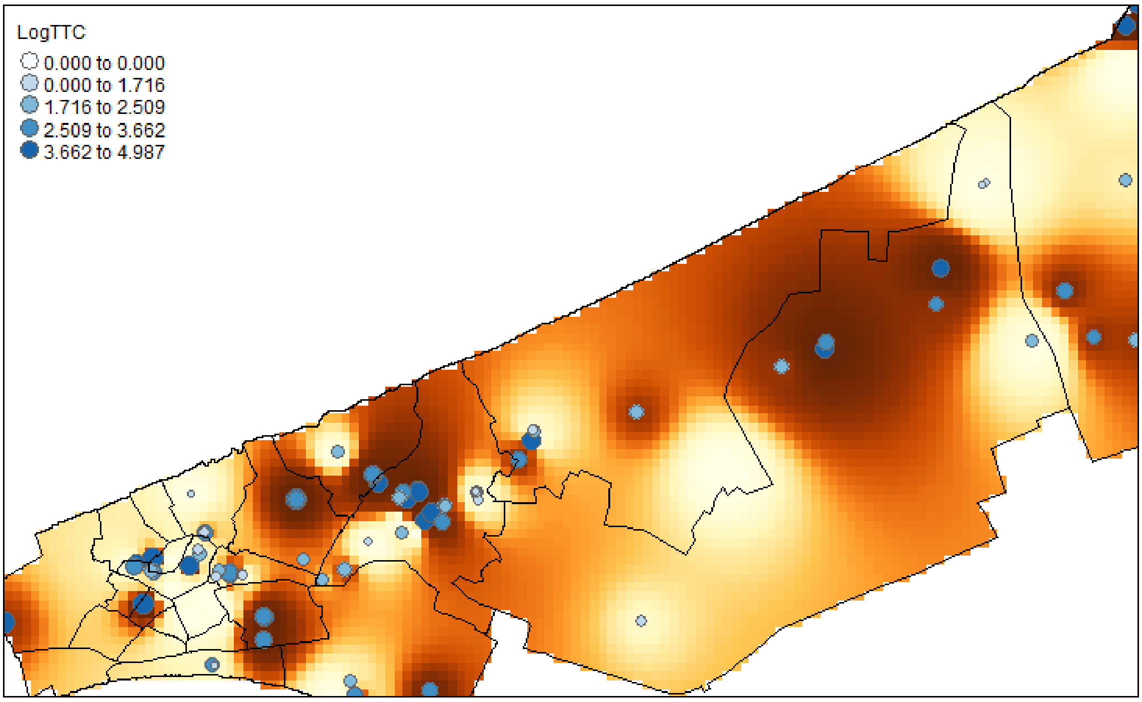


Figure 20: IDW interpolation of Log(TTC) across the study area

This interpolation offers a partial answer to ***RQ5:*** *Does the faecal contamination demonstrate spatial patterns, and can it be classified?*

- ➔ **Faecal contamination does not exhibit significant spatial autocorrelation, therefore any data smoothing effort is not representative of the processes underlying faecal matter pollution. IDW in this case can be used to visualize the data that has been collected but is not a good option for modelling the data.**

3.4. Unsupervised Machine Learning: Hierarchical Clustering

Finally, an unsupervised approach to classification is adopted, with Agglomerative Hierarchical Clustering (HAC), also called Agglomerative Nesting (AGNES). This method starts from n clusters formed by n individual data points, that are progressively merged until a single cluster containing all n data points is formed. This creates a dendogram, and metrics are then available for the user to determine the optimal number of clusters. Unlike other clustering techniques such as k-means or DBSCAN, the clustering is performed without specifying a priori the number of clusters or a density function. Therefore, it is a powerful technique to reveal hidden data patterns (Berkhin, 2006). Figure 21 illustrates the workflow adopted for this analysis.

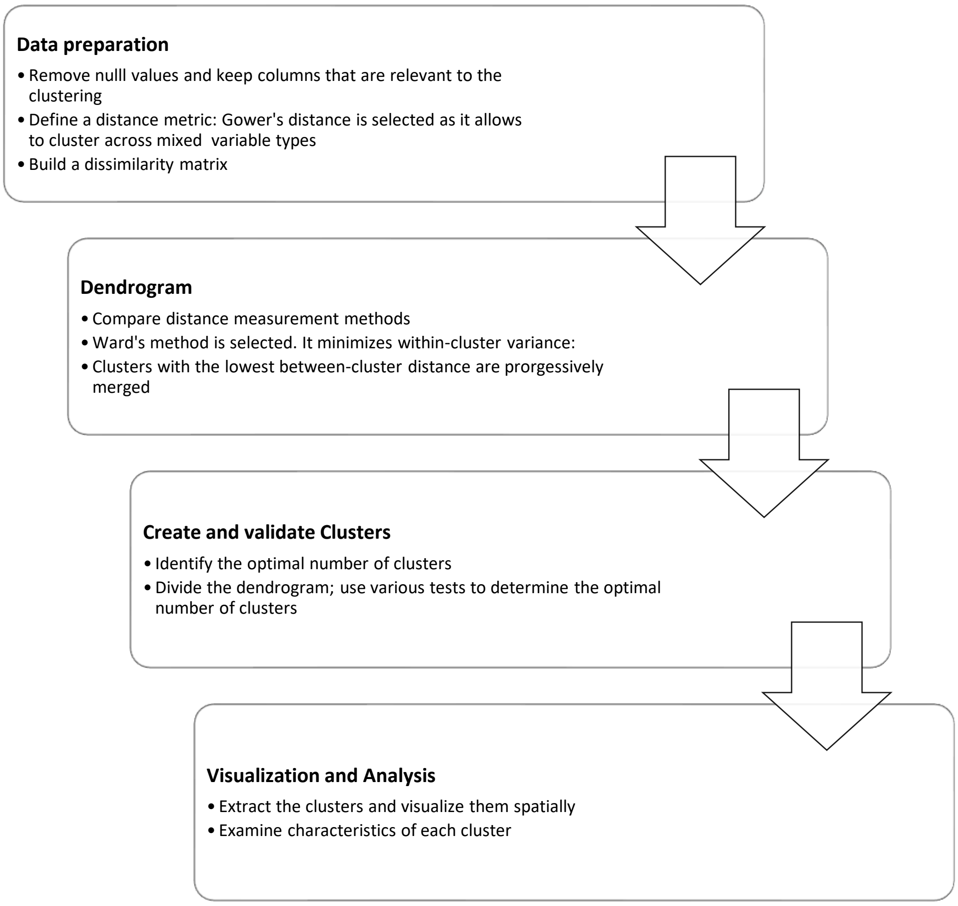


Figure 21: Agglomerative Hierarchical Clustering Workflow

A Gower's distance matrix is built using the function `daisy()` in the R package **cluster** (Maechler *et al.*, 2018). The best method for agglomerating more similar points as a cluster are assessed using an iteration of the `agnes()` function, and determines that Ward is the most relevant choice. In Ward's method, clusters with the lowest between-cluster distance are merged progressively, thereby minimizing the total within-cluster variance (Chessel, Thioulouse and Dufour, 2004). Next, a dendogram is built on the basis of the Gower dissimilarity matrix, using `agnes()` again.

Several functions are available to find the best possible cut for this tree. Comparing a few methods allows to determine which will best fit the purposes of the analysis. The loss of within-cluster cohesion produces an inertia plot (See Figure 22), on which the points that undergo the largest loss of inertia are identified as good potential cuts for the dendogram.

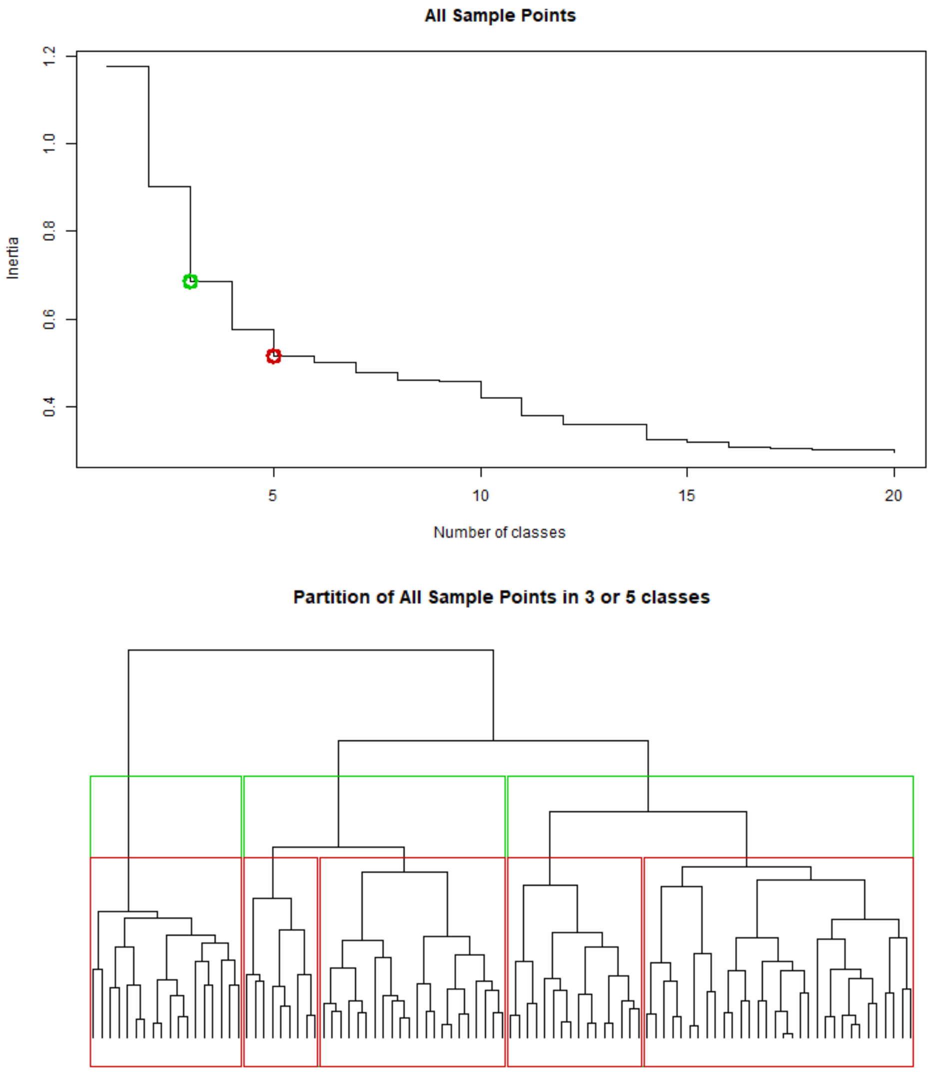


Figure 22: Inertia plot and corresponding cuts (at k=3 and k=5) on the dendogram

These suggested two clustering options can be validated with a silhouette plot; silhouette width is a measure of within-cluster similarity opposed to between-cluster distance and its values range from -1 (poor within cluster cohesion) to 1 (perfect cohesion), so the higher the value of S, the better. Figure 23 demonstrates that k = 3 is the optimal number of clusters for this dataset, as it maximizes the silhouette width.

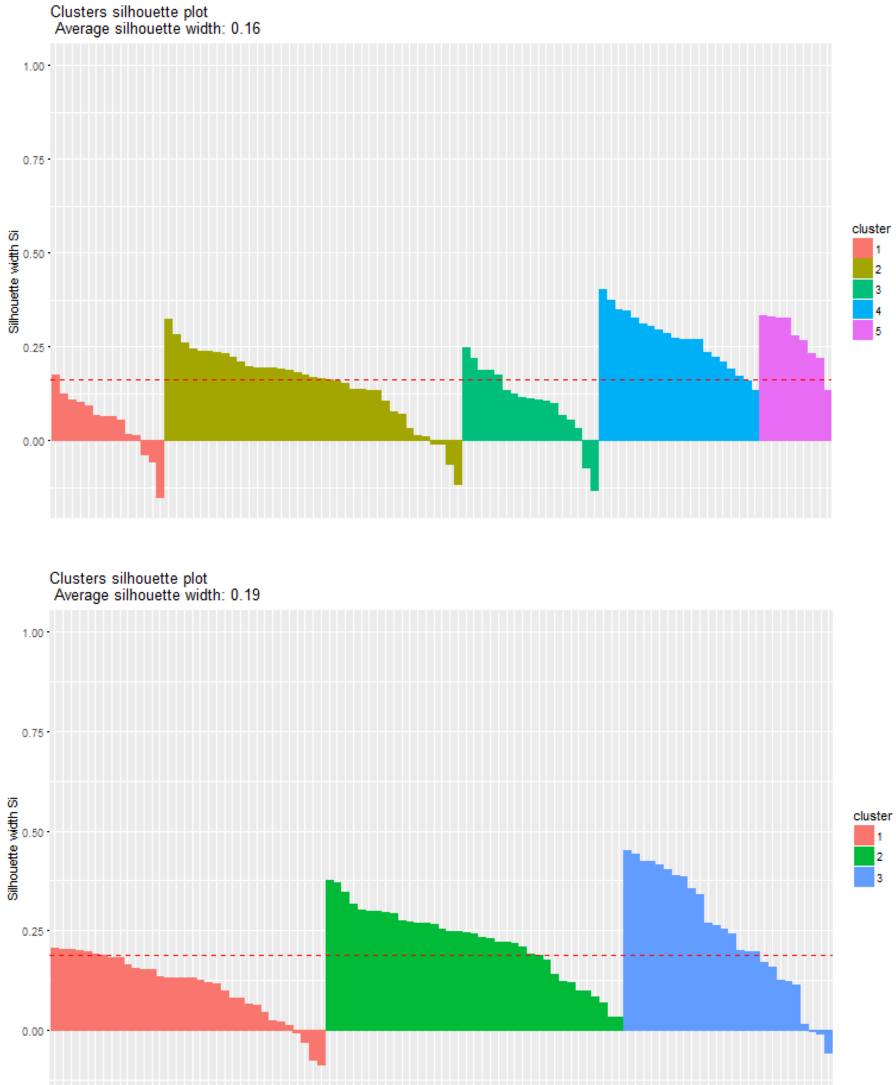


Figure 23: Silhouette Width for k=3 and k=5

Finally, using the **dplyr** package, the SampleData dataset is mutated; a new column is added, which assigns each point to a cluster. The results can be mapped on the study area (See Figure 24). Looking at the characteristics of each cluster, it appears that these clusters can be broadly characterized as:

- **Cluster 1:** Located in the peri-urban area, from Pikine to Keur Massar. Mostly hand pumps, with low TLF results and low TTC results but very high CDOM levels, sampled before the first rain.
- **Cluster 2:** Located in the peri-urban area, from Pikine to Keur Massar. Mostly hand pumps and dug wells, contaminated, with very high TTC levels and relatively high levels of TLF.

- **Cluster 3:** Located in the rural area in the East. Mostly dug wells and piezometers, highly contaminated, quite high TLF and TTC levels, very low population density, sampled before the rain for points in the East, and after the rain for points in the Pikine/Guediawaye area.

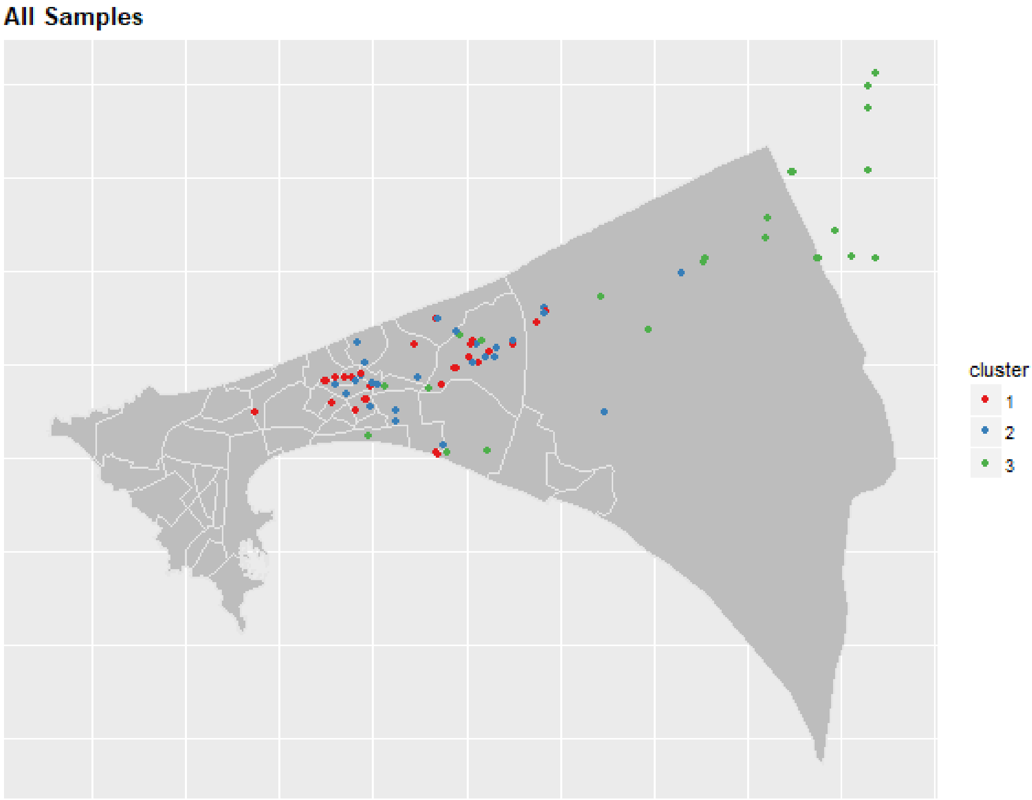


Figure 24: Mapping of the dataset classification on the study area

This classification finally answers the last research question ***RQ5:*** *Does the faecal contamination demonstrate spatial patterns, <u>and can it be classified?</u>*

➔ **Points sampled for this study suffer from a poor sampling pattern, which impedes possibilities of geostatistical modelling. However, hierarchical clustering provides insight into patterns that cannot be detected using supervised learning techniques. This classification opens new hypotheses and can inform future sampling strategies.**

4. DISCUSSION

4.1. TLF as a faecal matter detection method

This study found that TLF fails to predict TTC levels and is therefore currently not an adequate method for real-time faecal matter detection in a nutrient-rich aquifer such as the Thiaroye aquifer. However, a positive correlation between TLF data and flow cytometry data was found (Spearman rank $\rho = 0.6230972$), which supports the hypothesis that a large variety of particles are being detected by the TLF sensor and should not necessarily be interpreted as TTCs. The comparison of TLF readings on filtered and unfiltered samples further supports this hypothesis: TTCs were blocked by the bacteriological filter, yet TLF readings were quasi-similar on filtered and non-filtered samples.

A strong correlation was also found between TLF and CDOM; as there is a slight overlap in their excitation-emission wavelength ranges, it is possible that the TLF sensor also detects parts of the CDOM compounds. Unlike in other contexts (Sorensen *et al.*, 2018a), in the Thiaroye aquifer the TLF fluorometer is a poor indicator of *current* contamination status but may well be reflecting *past* contamination, in the form of microbial debris.

The same sampling protocol for measuring TTC, CDOM and TLF on Kisumu, Kenya and Lukaya, Uganda, yielded dramatically different results (See Table 5). This raises a number question about which environmental variables influence TLF's performance as a faecal matter detection method.

Table 5: Spearman rank for TLF & TTC in the three AfriWatSan urban observatories

Case study	*Spearman rank correlation coefficient for TLF/TTC*
Lukaya, Uganda (Carr, 2018)	0.721
Kisumu, Kenya (Van der Marel, 2018)	0.799
Dakar, Senegal	0.3195

4.2. Predictors of faecal contamination

The stepwise regression pointed to a set of predictors that, taken together, offer decent predictions of faecal coliforms. However, no single parameter emerged as a reliable proxy for estimating TTC levels. TLF and CDOM were not significant in the model and, surprisingly, neither were nitrates and population density. This is unexpected, as population density was chosen as a proxy for the volume of human excreta produced in a given area.

It could be argued that TTCs and even *E. coli* are not a perfect indicator of faecal contamination. (Leclerc *et al.*, 2001) points to the limitations of using these bacteria to assess health risk. The TTC group is comprised of *E. coli* but also other unharmful bacteria (WHO, 2011 cited in Sorensen *et al.*, 2015). As *E. coli* test results could not be interpreted due to logistics issues, TTC remains a solid alternative widely used in the literature. (Howard *et al.*, 2003) found that 99% of TTCs in groundwater polluted with sanitation effluents were actually *E. coli*. Chances that negative TTC results were false positives can therefore be considered very low.

4.3. Sampling Strategy

The sampling pattern suffered critical limitations to producing a proper geostatistical model of contamination across the aquifer. Data was under-sampled, because sampling opportunities were tied to physical limitations: in certain areas no groundwater source existed, the source was unusable (e.g. piezometers blocked by sand infiltration) or inaccessible (e.g. access to a private source denied by the owner). Moreover, in more well-off areas, no handpumps could be found as most residents were connected to the water supply network. The installation of additional piezometers in strategic areas could help overcome this obstacle and improve the sampling strategy to improve quality of the interpolation.

Still, the most important flaw of this sampling scheme is that points are spaced out, when water contamination is a very local phenomenon, for which any spatial correlation is at very short range. Future research could focus on a smaller subset of the aquifer to extensively map out and sample all groundwater sources and obtain a ore closely knitted network of sampling sources.

4.4. Other limitations

Other limitations of this study concern mostly the data collected. First, Turbidity and Dissolved Oxygen probes were not calibrated, meaning these two parameters are not reliable. Second, the first rain created a structural break in the data, which was impossible to control for due to the insufficient number of observations with regard to the high dimensionality of the dataset. This raises the issue of comparability between pre- and post-rain samples. Finally, flow cytometry data was conducted, after being shipped from Dakar to Wallingford, on samples collected from 3 days to 28 days earlier. One may not exclude the possibility that in spite of the preservatives, the older samples underwent a higher modification of the cell count than more recent samples, raising again the question of comparability of a variable across the dataset.

CONCLUSION

What can TLF and culture-based methods reveal about the patterns of faecal contamination in the Thiaroye aquifer?

Due to the incredibly complex interplay of multiple factors in the contamination of the aquifer, and because hydrodynamic processes mean that faecal bacteria are transported by groundwater flows, basic regression and interpolation techniques fail to accurately model faecal contamination of groundwater in the Thiaroye aquifer. Unsupervised machine learning is useful in developing a simple classification of these multi-dimensional observations collected in the study area. But in order to achieve a more accurate representation of the contamination, further research will need to incorporate groundwater flow modelling, and to investigate vertical contamination flows (Pouye, forthcoming).

While TLF is a poor predictor of current contamination across the Thiaroye aquifer, it provides additional information beyond faecal contamination and seems to indicate the presence of soluble particles. Empirically, it appeared that more recently urbanized areas such as Keur Massar, where houses and on-site sanitation facilities have been installed for less than five years, displayed significantly lower TLF rates, independently of the actual contamination level. This supports the hypothesis that the soluble particles detected by the TLF and CDOM fluorometers are debris of past pollution. Access to historical data of pollution and historical landuse data was not granted for this study, but this hypothesis could be tested with a spatio-temporal investigation of the link between historical loads of faecal bacteria and current TLF and CDOM rates. STARIMA could be a valid approach to treating such data (Deng *et al.*, 2018).

ORIGINAL DISSERTATION PROPOSAL

Each day, 1.8 billion individuals around the world drink water contaminated with faeces (WHO, 2017). In sub-Saharan Africa alone, this represents a leading cause of mortality, with diarrhoeal diseases killing 643,000 people in 2015 (WHO, 2016). Improved faecal matter detection methods are crucial in identifying causes of water contamination, communicating risks to users more efficiently, and developing adequate solutions, especially in the domain of sanitation infrastructure.

This MSc dissertation project seeks to map using GIS tools the relationships between environmental and social characteristics of Senegal's capital city Dakar and faecal contamination of shallow groundwater using evidence from both standard culture-based methods and a new, real-time technique using tryptophan-like fluorescence. Further, using spatial autocorrelation and geographically weighted regressions, locations returning the highest levels of discrepancies between TLF and culture-based methods can reveal the conditions under which TLF tests are less reliable. Preliminary evidence suggests that TLF false positives may be induced by the presence of gasoline yet this thesis will also explore whether other factors come into play. Depending on distance to existing labs, financial resources of a given area, the degree of contamination and the reliability of TLF technology in any given location in Dakar, we can determine which one of the lab tests or TLF real-time tests will be more relevant. A final map will display the relevance of using TLF over bacteriological tests across the city of Dakar.

Research will be conducted under the AfriWatSan project, funded by The Royal Society (UK) and Department for International Development (DFID), and supported by the British Geological Survey (BGS), currently developing portable, UV-based fluorimeters for real-time screening of faecally contaminated drinking water in urban Africa.

AUTO-CRITIQUE

Having worked in research on water policies and tariff schemes, including in the context of the SDGs, I had a great interest in the AfriWatSan project advertised in the Geography department. I chose to apply for the Dakar-based research because I am a native French speaker and have a background in urban governance in developing countries. This funded research opportunity was a fantastic chance to get fieldwork experience and collect my own data. This allowed me to be fully aware of the dataset strengths and limitations.

Weaknesses:

My proposal was probably a bit naïve as to how the data acquisition would go. Until fieldwork started, I never questioned the assumption that the correlation between the results of fluorescence-based and culture-based methods would be very strong. However, the data collected in Dakar was very different from previous case studies on TLF and no clear correlation emerged between TLF and TTC. Had I known that, I would certainly have done my homework better prior to the field and gotten up-to-date with basic concepts of hydrogeology and microbiology. I had anticipated that the main difficulty would be to get access to additional sources of urban, demographic and environmental data, but I assumed that a good understanding of geo-statistics and GIS would be sufficient to conduct this research. In reality, I really struggled to get a grasp of all the complex phenomena that impact the quality of groundwater subject to such high levels of pollution as the Thiaroye aquifer.

Another difficulty I faced relates to the timeframe of the research, tied to the agenda of the AfriWatSan project. A number of parameters required some time for the analysis to be conducted, and this led to a tight schedule to run the analysis, make sense of the results and develop a solid discussion.

Last but not least, the first rain of the year hit Dakar two weeks earlier than expected and three days before the end of fieldwork, which forced me to add this dimension to my sampling and to accept that not all data points would be fully comparable. This just proved that fieldwork usually implies a certain degree of unpredictability and requires anticipation!

Strengths:

I was extremely lucky to work within a multi-disciplinary team at UCAD and to be able to ask questions and obtain textbook references whenever I needed to understand a parameter or a phenomenon better. This is invaluable and I must thank again all the AfriWatSan team for this! The length of my stay (7 weeks) was also sufficient for me to get familiar with the study area, to cover the entirety of the UCAD monitoring network, and to conduct 14 re-samplings after the first rain.

Finally, being involved in the 4th AfriWatSan Consortium Workshop at the end of my stay was a brilliant opportunity to share experiences with Kenyan and Ugandan teams and to hear an institutional perspective with Senegalese decision-makers and stakeholders. This provided me with precious insights from other disciplines that, I am sure, add considerable value to my dissertation.

(503 words)

REFERENCES

Acquistapace, A. *et al.* (2017) *Baromètre 2017 de l'eau, de l'hygiène et de l'assainissement*. Available at: https://www.solidarites.org/wp-content/uploads/2017/05/Barometre-de-leau-lhygiene-et-lassainissement-2017.pdf (Accessed: 29 August 2018).

Baker, A. and Inverarity, R. (2004) 'Protein-like fluorescence intensity as a possible tool for determining river water quality', *Process*, 18, pp. 2927–2945. doi: 10.1002/hyp.5597.

Bartram, J. and Ballance, R. (1996) *Water Quality Monitoring-A Practical Guide to the Design and Implementation of Freshwater Quality Studies and Monitoring Programmes Edited by Chapter 10-MICROBIOLOGICAL ANALYSES*. Available at: http://www.who.int/water_sanitation_health/resourcesquality/wqmchap10.pdf (Accessed: 2 August 2018).

Berkhin, P. (2006) 'A Survey of Clustering Data Mining Techniques', in *Grouping Multidimensional Data*. Berlin/Heidelberg: Springer-Verlag, pp. 25–71. doi: 10.1007/3-540-28349-8_2.

Bruce, P. and Bruce, A. (2017) *Practical Statistics for Data Scientists: 50 Essential Concepts*. ' O'Reilly Media, Inc.'

Brundson, C., Fotheringham, A. S. and Charlton, M. E. (2002) 'Geographically Weighted Regression: The Analysis of Spatially Varying Relationship'. John Wiley & Sons Ltd, England.

Burrough, P. A. and McDonnell, R. A. (1998) 'Creating continuous surfaces from point data', *Principles of Geographic Information Systems. Oxford University Press, Oxford, UK*.

Carr, A. (2018) *Monitoring the relationship between heavy rainfall and faecal contamination of groundwater in Lukaya, Uganda*. University College London.

Chelsea Technologies Group Ltd (2018) *UviLux Sensor - Real-time detection of aromatic hydrocarbons, CDOM, Tryptophan-like fluorescence, BOD or Optical Brighteners*. Available at: https://www.chelsea.co.uk/uvilux-sensor#specification (Accessed: 2 August 2018).

Chen, S., Goo, Y.-J. J. and Shen, Z.-D. (2014) 'A hybrid approach of stepwise regression, logistic regression, support vector machine, and decision tree for forecasting fraudulent financial statements', *The Scientific World Journal*. Hindawi, 2014.

Chessel, D., Thioulouse, J. and Dufour, A. B. (2004) *Introduction à la classification hiérarchique - Exemples sur R*. Available at: https://fr.scribd.com/document/38089749/2004-Introduction-a-la-classification-hierarchique-Exemples-sur-R (Accessed: 10 August 2018).

Cissé Faye, S. *et al.* (2004) 'An assessment of the risk associated with urban development in the Thiaroye area (Senegal)', *Environmental Geology*, 45(3), pp. 312–322. doi: 10.1007/s00254-003-0887-x.

Cohen, B. (2006) 'Urbanization in developing countries: Current trends, future projections, and key challenges for sustainability', *Technology in society*. Elsevier, 28(1–2), pp. 63–80.

Criqui, L. (2013) 'Pathways for progressive planning through extending water and electricity networks in the irregular settlements of Lima'. Wilson Center.

Criqui, L. (2014) *Attention! Travaux en cours: l'extension des réseaux de services essentiels dans les quartiers irréguliers de Delhi et Lima*. Université Paris-Est.

Deng, M. *et al.* (2018) 'Heterogeneous Space--Time Artificial Neural Networks for Space--Time Series Prediction', *Transactions in GIS*. Wiley Online Library, 22(1), pp. 183–201.

Diaw, M. T. *et al.* (no date) 'The relationship between on-site sanitation density and shallow groundwater quality: evidence from the Thiaroye aquifer of Dakar, Senegal', *Water Research*.

Diongue, D. M. L. (2018) *Estimation des taux de recharge à partir d'un modèle de fluctuation piézométrique dans le bassin de Thiaroye*. Université Cheikh Anta Diop de Dakar, Sénégal.

Duivenvoorde, R. (2011) *QGIS Python Plugins Repository*. Available at: https://plugins.qgis.org/plugins/xytools/ (Accessed: 2 August 2018).

Eggleton, F. (2018) *Factors influencing the supply and demand of alternative sanitation facilities in the Thiaroye area of Dakar*. University College London.

Fotheringham, A. S. and Wong, D. W. S. (1991) 'The modifiable areal unit problem in multivariate statistical analysis', *Environment and planning A*. SAGE Publications Sage UK: London, England, 23(7), pp. 1025–1044.

Fox, B. G. *et al.* (2017) 'The in situ bacterial production of fluorescent organic matter; an investigation at a species level', *Water Research*. Pergamon, 125, pp. 350–359. doi:

10.1016/J.WATRES.2017.08.040.

Gaye, C. B. *et al.* (1990) ‘Analyse de l’intrusion saline dans les aquifères de la presqu’ile du Cap-Vert: analyse des processus de minéralisation et de dégradation de la qualité de l’eau dans les nappes infrabasaltiques et des sables quaternaires’. Université Cheikh Anta Diop, Dakar, SN.

Haining, R., Wise, S. and Ma, J. (1998) ‘Exploratory Spatial Data Analysis in a Geographic Information System Environment’, *Journal of the Royal Statistical Society. Series D (The Statistician)*. WileyRoyal Statistical Society, pp. 457–469. doi: 10.2307/2988627.

Hammes, F. *et al.* (2008) ‘Flow-cytometric total bacterial cell counts as a descriptive microbiological parameter for drinking water treatment processes’, *Water Research*. Pergamon, 42(1–2), pp. 269–277. doi: 10.1016/J.WATRES.2007.07.009.

Harrell, F. (2018) ‘Hmisc: Harrell Miscellaneous. R package version 4.1-1.’ Available at: https://cran.r-project.org/package=Hmisc.

Howard, G. *et al.* (2003) ‘Risk factors contributing to microbiological contamination of shallow groundwater in Kampala, Uganda’, *Water Research*, 37(14), pp. 3421–3429. doi: 10.1016/S0043-1354(03)00235-5.

Kassambara, A. (2018) *Machine Learning Essentials: Practical Guide in R*. STHDA.

Leclerc, H. *et al.* (2001) ‘Advances in the Bacteriology of the Coliform Group: Their Suitability as Markers of Microbial Water Safety’, *Annual Review of Microbiology*. Annual Reviews 4139 El Camino Way, P.O. Box 10139, Palo Alto, CA 94303-0139, USA , 55(1), pp. 201–234. doi: 10.1146/annurev.micro.55.1.201.

Longley, P. A. *et al.* (2015) *Geographic Information Science and Systems*.

Maechler, M. *et al.* (2018) ‘cluster: Cluster Analysis Basics and Extensions’.

Van der Marel, L. (2018) *Groundwater contamination and its vulnerability - a comparative analysis of culture based and real-time methods in Kisumu, Kenya*. University College London.

Marie, D., Rigaut-Jalabert, F. and Vaulot, D. (2014) ‘An improved protocol for flow cytometry analysis of phytoplankton cultures and natural samples’, *Cytometry Part A*. Wiley-Blackwell, 85(11), pp. 962–968. doi: 10.1002/cyto.a.22517.

Max, A. *et al.* (2015) ‘Classification and Regression Training (CARET)’. Available at: https://cran.r-project.org/package=caret.

Miller, H. J. (2004) ‘Tobler ’ s First Law and Spatial Analysis’, *Annals of the Association of American Geographers*, 94(2), pp. 284–289. doi: 10.1111/j.1467-8306.2004.09402005.x.

Monstadt, J. and Schramm, S. (2017) ‘Toward The Networked City? Translating Technological ideals and Planning Models in Water and Sanitation Systems in Dar es Salaam’, *International Journal of Urban and Regional Research*. Wiley/Blackwell (10.1111), 41(1), pp. 104–125. doi: 10.1111/1468-2427.12436.

Mor, S. *et al.* (2006) ‘Leachate Characterization and Assessment of Groundwater Pollution Near Municipal Solid Waste Landfill Site’, *Environmental Monitoring and Assessment*. Kluwer Academic Publishers, 118(1–3), pp. 435–456. doi: 10.1007/s10661-006-1505-7.

Office National de l’Assainissement du Sénégal ONAS (2015) *Market Structuring of Fecal Sludge Management for the Benefit of Poor Households in Dakar*. Available at: http://www.susana.org/images/documents/07-cap-dev/b-conferences/15-FSM3/Day-1/Rm-1/1-1-3-1-Mbeguere.pdf (Accessed: 8 August 2018).

Pacheco, A. *et al.* (1991) ‘Cemeteries - A Potential Risk to Groundwater’, *Water Science and Technology*. IWA Publishing, 24(11), pp. 97–104. doi: 10.2166/wst.1991.0341.

Pebesma, E. and Heuvelink, G. (2016) ‘Spatio-temporal interpolation using gstat’, *RFID Journal*, 8(1), pp. 204–218.

Pebesma, E. J. (2004) ‘Multivariable geostatistics in S: the gstat package’, *Computers & Geosciences*. Elsevier, 30(7), pp. 683–691.

Pouye, A. (no date) ‘Assessment of groundwater vulnerability to Nitrate contamination using GIS-based hydrogeological methods at a catchment scale: Case study of the Thiaroye aquifer, Dakar, Senegal’.

QGIS (2016) *The QGIS NNJoin Plugin — NNJoin 3.0.7 documentation*. Available at: http://arken.nmbu.no/~havatv/gis/qgisplugins/NNJoin/# (Accessed: 2 August 2018).

R Development Core Team (2013) ‘R: A language and environment for statistical computing.’, *R Foundation for Statistical Computing, Vienna, Austria*. Citeseer, 3. doi: 10.1007/978-3-540-74686-7.

Re, V. *et al.* (2011) 'Water quality decline in coastal aquifers under anthropic pressure: the case of a suburban area of Dakar (Senegal)', *Environmental Monitoring and Assessment*. Springer Netherlands, 172(1–4), pp. 605–622. doi: 10.1007/s10661-010-1359-x.

Rosenzweig, C. *et al.* (2011) *Climate change and cities: First assessment report of the urban climate change research network*. Cambridge University Press.

Sansom, K. (2006) 'Government engagement with non-state providers of water and sanitation services', *Public Administration and Development*. Wiley-Blackwell, 26(3), pp. 207–217. doi: 10.1002/pad.419.

Sartory, D. (2009) *The Microbiology of Drinking Water - Methods for the isolation and enumeration of coliform bacteria and Escherichia coli (including E. coli O157:H7) Methods for the Examination of Waters and Associated Materials*. Available at: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/316786/MoDWPart4-223MAYh.pdf (Accessed: 2 August 2018).

Snozzi, M., Ashbolt, N. J. ; and Grabow, W. O. (2001) 'Indicators of microbial water quality', in *Water Quality: Guidelines, Standards and Health. Edited by Lorna Fewtrell and Jamie Bartram.* London: IWA Publishing.

Sorensen, J. P. R. *et al.* (2015) 'In-situ tryptophan-like fluorescence: A real-time indicator of faecal contamination in drinking water supplies', *Water Research*. Elsevier Ltd, 81, pp. 38–46. doi: 10.1016/j.watres.2015.05.035.

Sorensen, J. P. R. *et al.* (2018a) 'Real-time detection of faecally contaminated drinking water with tryptophan-like fluorescence: defining threshold values', *Science of the Total Environment*. Elsevier B.V., 622–623, pp. 1250–1257. doi: 10.1016/j.scitotenv.2017.11.162.

Sorensen, J. P. R. *et al.* (2018b) 'Real-time detection of faecally contaminated drinking water with tryptophan-like fluorescence: defining threshold values', *Science of the Total Environment*. Elsevier B.V., 622–623, pp. 1250–1257. doi: 10.1016/j.scitotenv.2017.11.162.

Tallon, P. *et al.* (2005) 'Microbial Indicators of Faecal Contamination in Water: A Current Perspective', *Water, Air, and Soil Pollution*. Kluwer Academic Publishers, 166(1–4), pp. 139–166. doi: 10.1007/s11270-005-7905-4.

Tobler, W. R. (1970) 'A Computer Movie Simulating Urban Growth in the Detroit Region', *Economic Geography*, 46, p. 234. doi: 10.2307/143141.

United Nations *et al.* (2010) *The Right to Water*. Available at: https://www.ohchr.org/Documents/Publications/FactSheet35en.pdf (Accessed: 29 August 2018).

United Nations, D. of E. and S. A. (2014) 'World urbanization prospects, the 2011 revision', *Population Division, Department of Economic and Social Affairs, United Nations Secretariat*.

Wakida, F. and Lerner, D. (2005) 'Non-agricultural sources of groundwater nitrate: a review and case study', *Water Research*. Pergamon, 39(1), pp. 3–16. doi: 10.1016/J.WATRES.2004.07.026.

Wei, T. and Simko, V. (2018) 'R package "corrplot": Visualization of a Correlation Matrix (Version 0.85)'. Available at: https://github.com/taiyun/corrplot.

WHO/UNICEF Joint Monitoring Programme for Water Supply and Sanitation (2017) 'Progress on Drinking Water, Sanitation and Hygiene', *Unicef*, pp. 1–66. doi: 10.1111 / tmi.12329.

WHO (1997) *Appendix C Sanitary Inspection Forms. In: Guidelines for drinking-water quality, 2nd edition: Volume 3 - Surveillance and control of community supplies*. Available at: http://www.who.int/water_sanitation_health/publications/wsp170805AppC.pdf (Accessed: 28 August 2018).

WHO (2011) 'Guidelines for drinking-water quality - 4th edition', *WHO chronicle*, 38(4), pp. 104–108.

Wickham, H. (2009) *Ggplot2 : elegant graphics for data analysis*. Springer.

Winkler, I. (2014) *The human right to water: significance, legal status and implications for water allocation*. Bloomsbury Publishing.

World Health Organization (2016) *World health statistics 2016: monitoring health for the SDGs sustainable development goals*. World Health Organization.

Yates, M. V (1985) 'Septic tank density and ground-water contamination', *Groundwater*. Wiley Online Library, 23(5), pp. 586–591.

APPENDICES

1. **Dataset of Sampled Points**

Following the MSc Dissertation handbook guidelines, this dataset is available upon request / on GitHub

https://github.com/raphaelleroffo/MScDissertation

2. WHO Sanitary risk forms (WHO, 1997)

I. Type of Facility BOREHOLE WITH HANDPUMP

1. General Information : Zone
   : Location
2. Code Number
3. Date of Visit
4. Water sample taken? …….. Sample No. ……… FC/100ml ………..

II Specific Diagnostic Information for Assessment

Risk

1. Is there a latrine within 10m of the borehole? Y/N
2. Is there a latrine uphill of the borehole? Y/N
3. Are there any other sources of pollution within 10m of borehole? Y/N
(e.g. animal breeding, cultivation, roads, industry etc)
4. Is the drainage faulty allowing ponding within 2m of the borehole? Y/N
5. Is the drainage channel cracked, broken or need cleaning? Y/N
6. Is the fence missing or faulty? Y/N
7. Is the apron less than 1m in radius? Y/N
8. Does spilt water collect in the apron area? Y/N
9. Is the apron cracked or damaged? Y/N
10. Is the handpump loose at the point of attachment to apron? Y/N

Total Score of Risks …./10

Risk score: 9-10 = Very high; 6-8 = High; 3-5 = Medium; 0-3 = Low

III Results and Recommendations:
The following important points of risk were noted:
(list nos. 1-10)

Signature of Health Inspector/Assistant:

Comments:

Form used for handpumps

I. Type of Facility DUG WELL WITH HANDPUMP / WINDLASS

1. General Information : Zone:
   : Location
2. Code Number
3. Date of Visit
4. Water sample taken? …….. Sample No. ……… FC/100ml ………..

II Specific Diagnostic Information for Assessment

	Risk
1. Is there a latrine within 10m of the well?	Y/N
2. Is the nearest latrine uphill of the well?	Y/N
3. Is there any other source of pollution within 10m of well? (e.g. animal breeding, cultivation, roads, industry etc)	Y/N
4. Is the drainage faulty allowing ponding within 2m of the well?	Y/N
5. Is the drainage channel cracked, broken or need cleaning?	Y/N
6. Is the fence missing or faulty?	Y/N
7. Is the cement less than 1m in radius around the top of the well?	Y/N
8. Does spilt water collect in the apron area?	Y/N
9. Are there cracks in the cement floor?	Y/N
10. Is the handpump loose at the point of attachment to well head?	Y/N
11. Is the well-cover insanity?	Y/N

Total Score of Risks …./11

Risk score: 9-11 = Very high; 6-8 = High; 3-5 = Medium; 0-3 = Low

III Results and Recommendations:

The following important points of risk were noted:
(list nos. 1-11)

Signature of Health Inspector/Assistant:

Comments:

Form used for dug wells

I. **Type of Facility** DEEP BOREHOLE WITH MECHANISED PUMPING

1. General Information : Supply zone
   : Location:
2. Code Number
3. Date of Visit
4. Water sample taken? …….. Sample No. ……… FC/100ml ………..

II **Specific Diagnostic Information for Assessment**

	Risk	
1. Is there a latrine or sewer within 100m of pumphouse?		Y/N
2. Is the nearest latrine unsewered?	Y/N	
3. Is there any source of other pollution within 50m?	Y/N	
4. Is there an uncapped well within 100m?	Y/N	
5. Is the drainage around pumphouse faulty?	Y/N	
6. Is the fencing damaged allowing animal entry?		Y/N
7. Is the floor of the pumphouse permeable to water?	Y/N	
8. Does water forms pools in the pumphouse?	Y/N	
9. Is the well seal insanitary?		Y/N

Total Score of Risks …./9

Risk score: 7-9 = High; 3-6 = Medium; 0-2 = Low

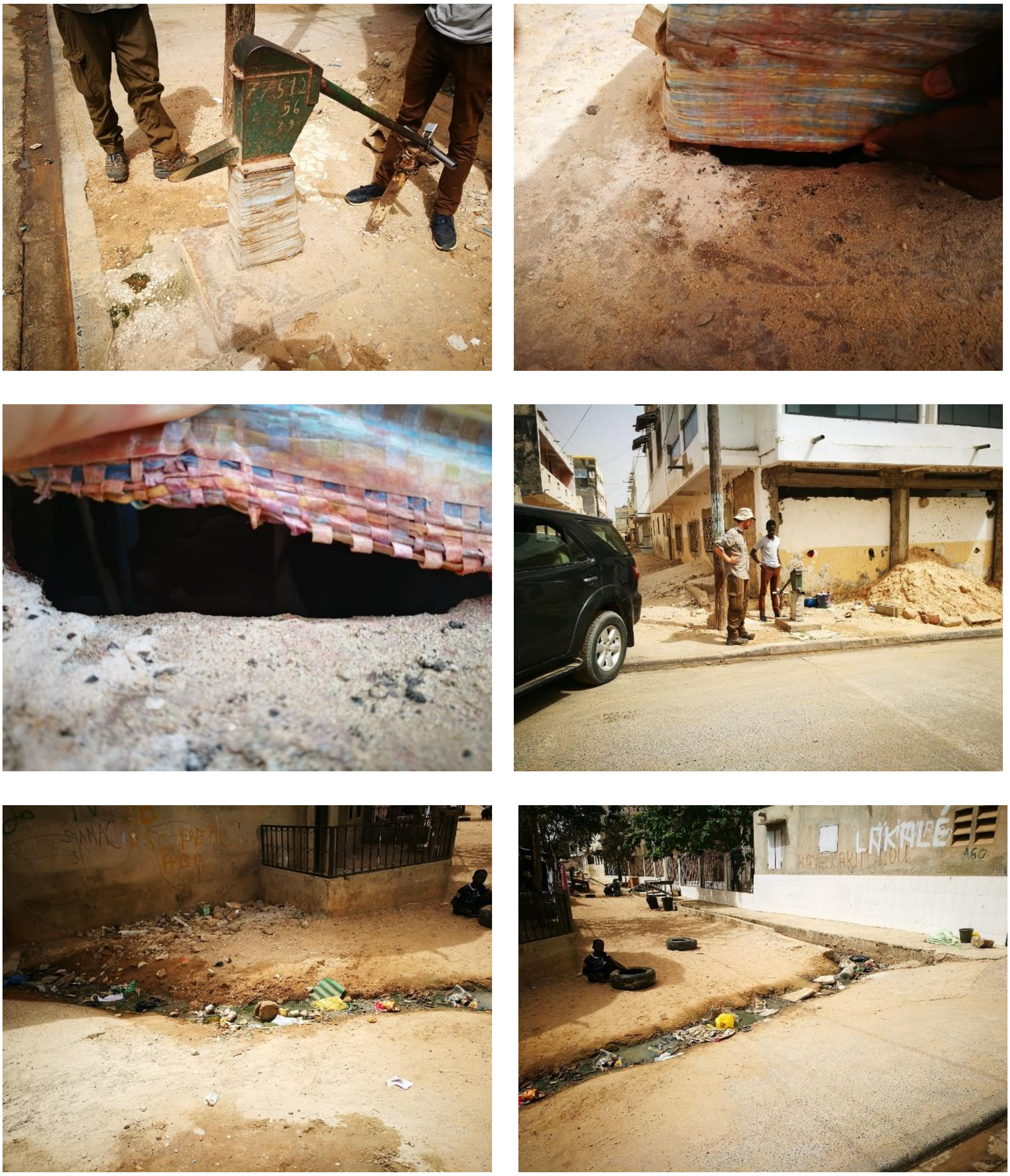
III **Results and Recommendations**:
The following important points of risk were noted:
(list nos. 1-9)

Signature of Health Inspector/Assistant:

Comments:

Form used for piezometers and the borehole

3. Example of photos taken to record sanitary risk and context



Pictures taken at sample point number 17 (HP7), on June 1st, 2018. This stream is a septic tank effluent, and was located 6m away, uphill from the handpump.

4. R code – GitHub repository

R code used for this project is available on this GitHub repository:

https://github.com/raphaelleroffo/MScDissertation