Lessons learned in working with real-life data in resource constrained settings with limited domain knowledge

Raphaëlle Roffo

Overview

← Lessons learnt →

Introduction

Motivation & Context

Research Aims & Constraints

Methods

Results

Discussion

Introduction

About me

- From turning domain knowledge and already processed data into policy guidance on a set of issues…
- … to producing the maps and analysis on a variety of topics

British Geological Survey
NATURAL ENVIRONMENT RESEARCH COUNCIL

UKaid from the British people

THE ROYAL SOCIETY

AfriWatSan

UCL

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

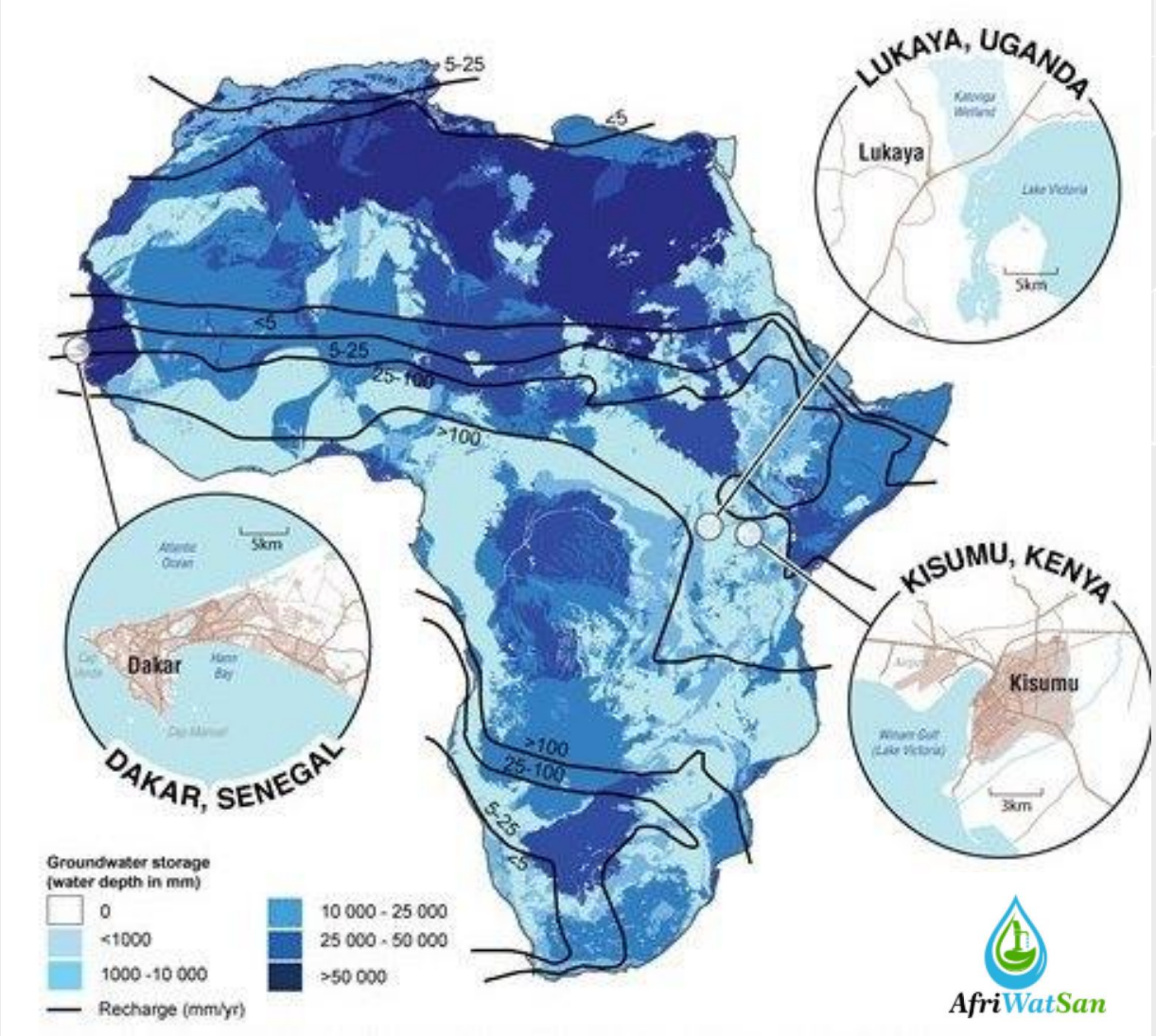
MSc Dissertation (MSc Geospatial Analysis, UCL, 2018)

Acknowledgements

- 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)

- AfriWatSan monitors a network of three Groundwater Observatories

- May-June 2018: 2 months field work to collect groundwater samples



British Geological Survey
NATURAL ENVIRONMENT RESEARCH COUNCIL

UKaid from the British people

THE ROYAL SOCIETY

AfriWatSan

UCL

Motivation & Context

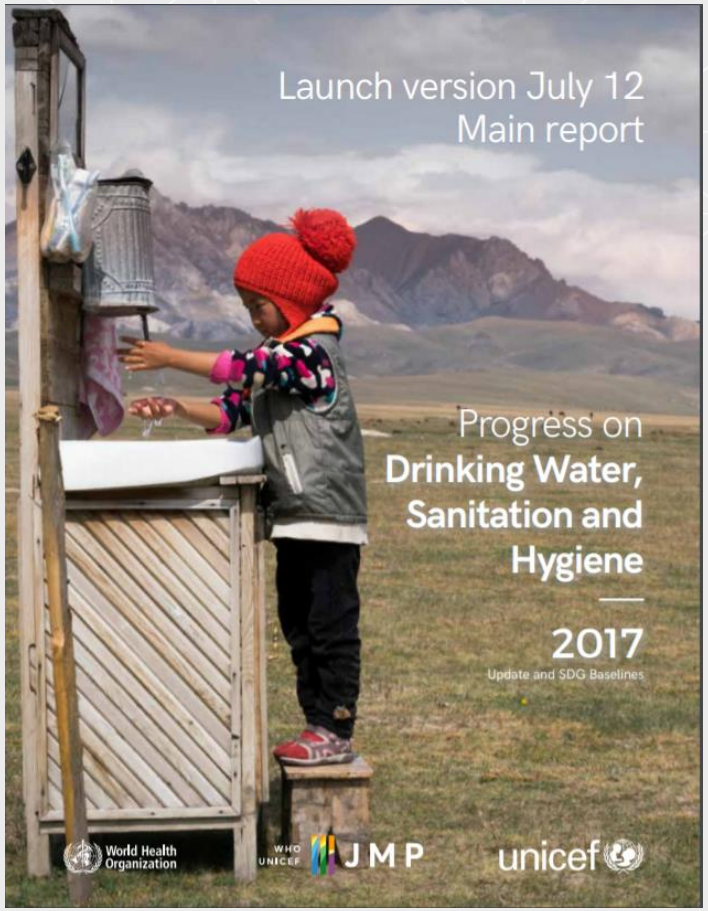
Motivation: Lack of Sanitation remains a leading cause of mortality in Sub-Saharan Africa

- Only 39% of the global population use safely managed sanitation infrastructure
- 1.8 billion people around the world still drink water that has been contaminated with faecal matter

In Sub-Saharan Africa alone:

- 72% of the population lack access to "at least basic" sanitation services
- 643,000 death from diarrhoeal diseases each year

WHO/UNICEF Joint Monitoring Programme for Water Supply and Sanitation, 2017

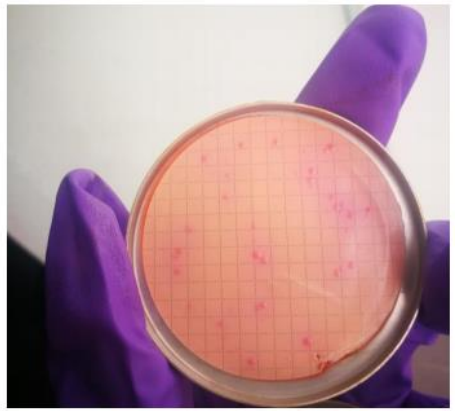


Tryptophan-like Fluorescence (TLF) as an alternative method for faecal matter detection

Thermotolerant Coliforms (TTC)
Culture-based method

+ Very reliable

– Costly, requires logistics, reagent, expertise and time (18h incubation)



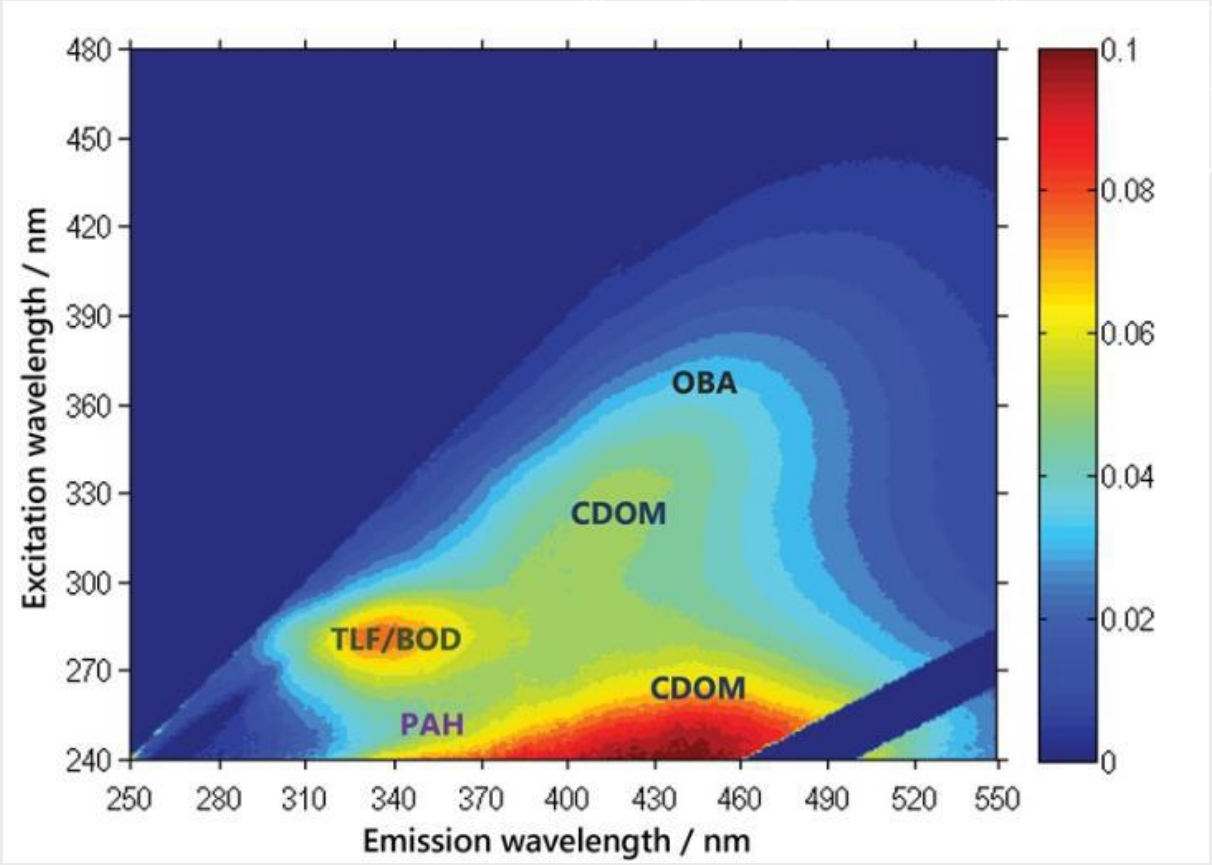
Tryptophan-like Fluorescence (TLF)
UV-fluorescence based method

+ Real-time, portable and easy to use

– Early stages of the development, poorly understood mechanisms

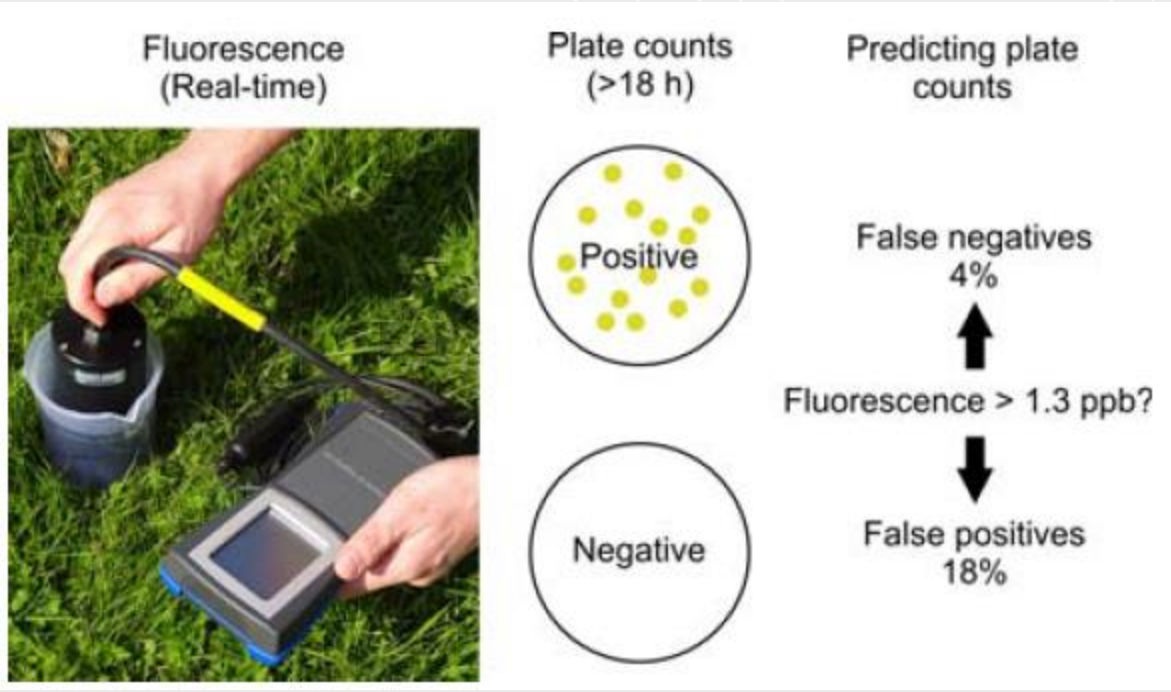
Tryptophan-like Fluorescence (TLF) as an alternative method for faecal matter detection

- Portable fluorometers measure concentration of organic dissolved matter at given wavelengths (280nm and 360nm).



Tryptophan-like Fluorescence (TLF) as an alternative method for faecal matter detection

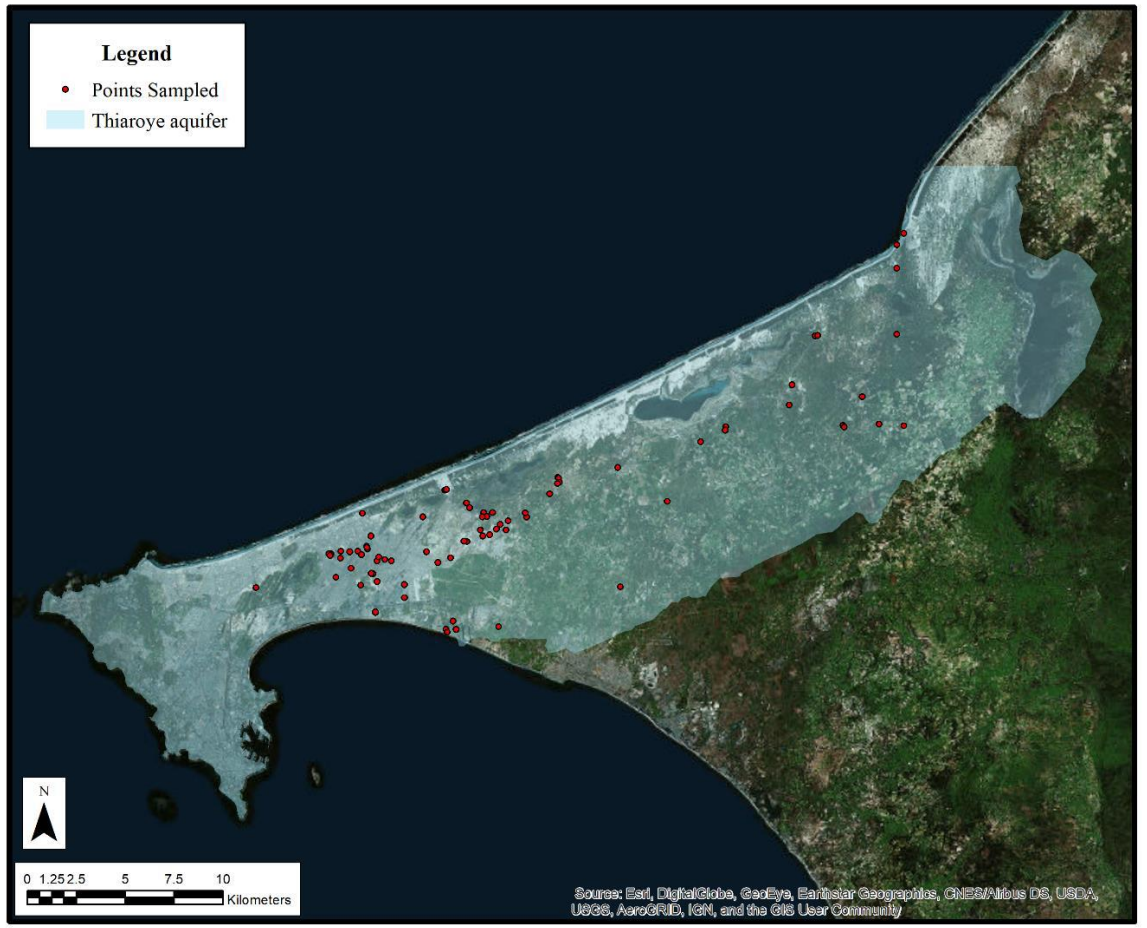
- In previous work lead by the BGS in Malawi, India, Zambia and South Africa, a logistic regression model found TLF to be a strong indicator of the presence of thermo-tolerant coliforms (TTC).



Sorensen et al, 2017

Study Area: the Thiaroye shallow aquifer, Dakar, Senegal

- 2.47 million inhabitants (mostly peri-urban)
- 52% of septic tanks are never emptied or are manually emptied
- Decades of multi-causal groundwater pollution (excreta, wastewater, fertilizers, industry, etc.)
- Coastal city (saline water infiltrations)



British Geological Survey
NATURAL ENVIRONMENT RESEARCH COUNCIL

UKaid from the British people

THE ROYAL SOCIETY

AfriWatSan

UCL

Research Aims & Constraints

Research Aims

1. Explore Tryptophan-Like Fluorescence (TLF) robustness and reliability as a predictor of faecal contamination in a highly polluted and densely populated area.
2. Explore contamination of the Thiaroye aquifer in relation to other available environmental variables. Understand whether environmental factors and hydrochemical parameters can predict actual faecal contamination across the aquifer.

Constraints

- Time! (4 months total)
- Internet connexion
- Logistics / Equipment
- Sampling pattern based on accessible sources → entire areas were left out

British Geological Survey
NATURAL ENVIRONMENT RESEARCH COUNCIL

UKaid from the British people

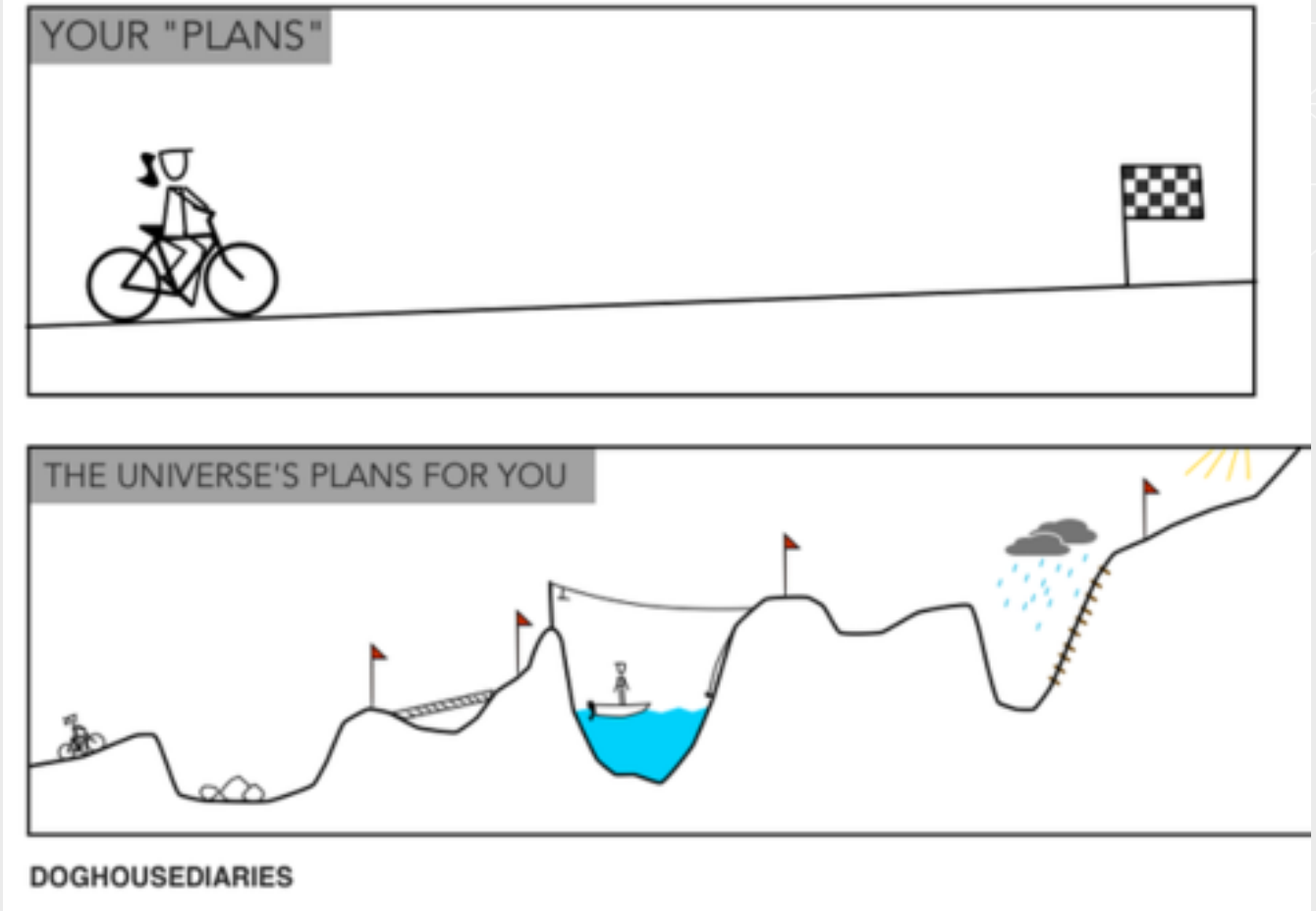
THE ROYAL SOCIETY

AfriWatSan

UCL

Lesson 1:

Plan for the unexpected



British Geological Survey
NATURAL ENVIRONMENT RESEARCH COUNCIL

UKaid from the British people

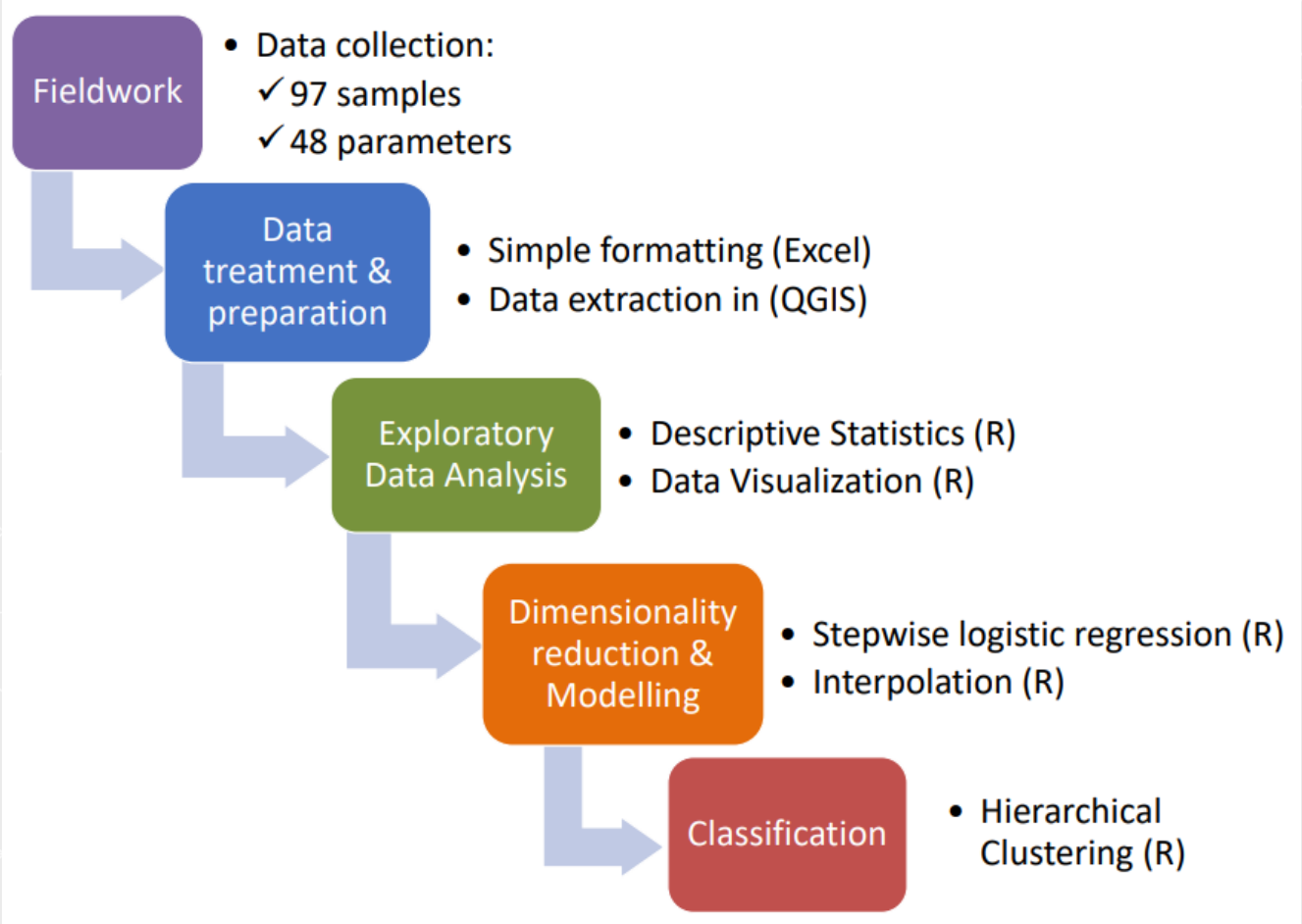
THE ROYAL SOCIETY

AfriWatSan

UCL

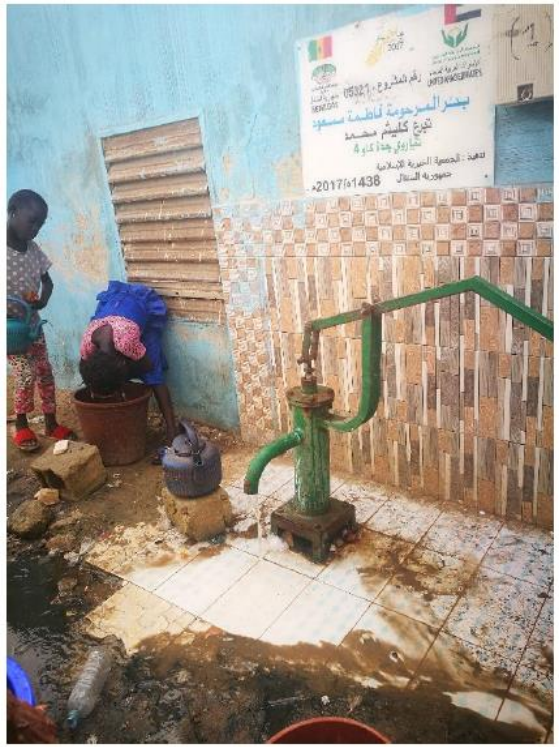
Methods

Methods


Fieldwork
• Data collection:
✓97 samples
✓48 parameters
Data treatment & preparation
• Simple formatting (Excel)
• Data extraction in (QGIS)
Exploratory Data Analysis
• Descriptive Statistics (R)
• Data Visualization (R)
Dimensionality reduction & Modelling
• Stepwise logistic regression (R)
• Interpolation (R)
Classification
• Hierarchical Clustering (R)

Data Collection:

97 samples across the Dakar region, 48 parameters



Handpump

Dug well

Piezometer

Borehole (with electric pump)

British Geological Survey
NATURAL ENVIRONMENT RESEARCH COUNCIL

UKaid from the British people

THE ROYAL SOCIETY

AfriWatSan

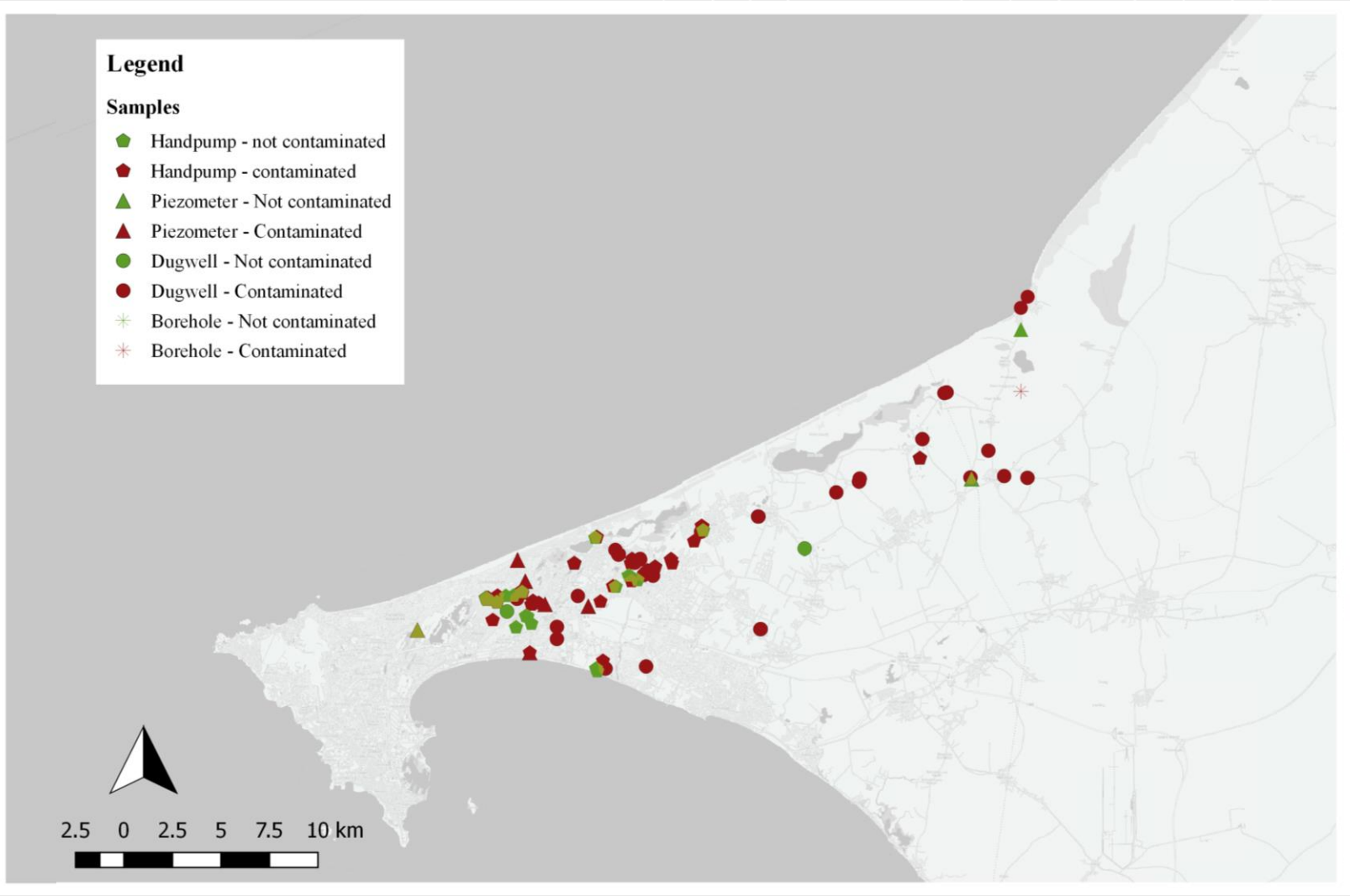
UCL

Data Collection:

97 samples across the Dakar region, 48 parameters

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

Exploratory Data Analysis



The Unexpected strikes, part 1: TLF doesn't… work?

- TFL and TTC do not correlate

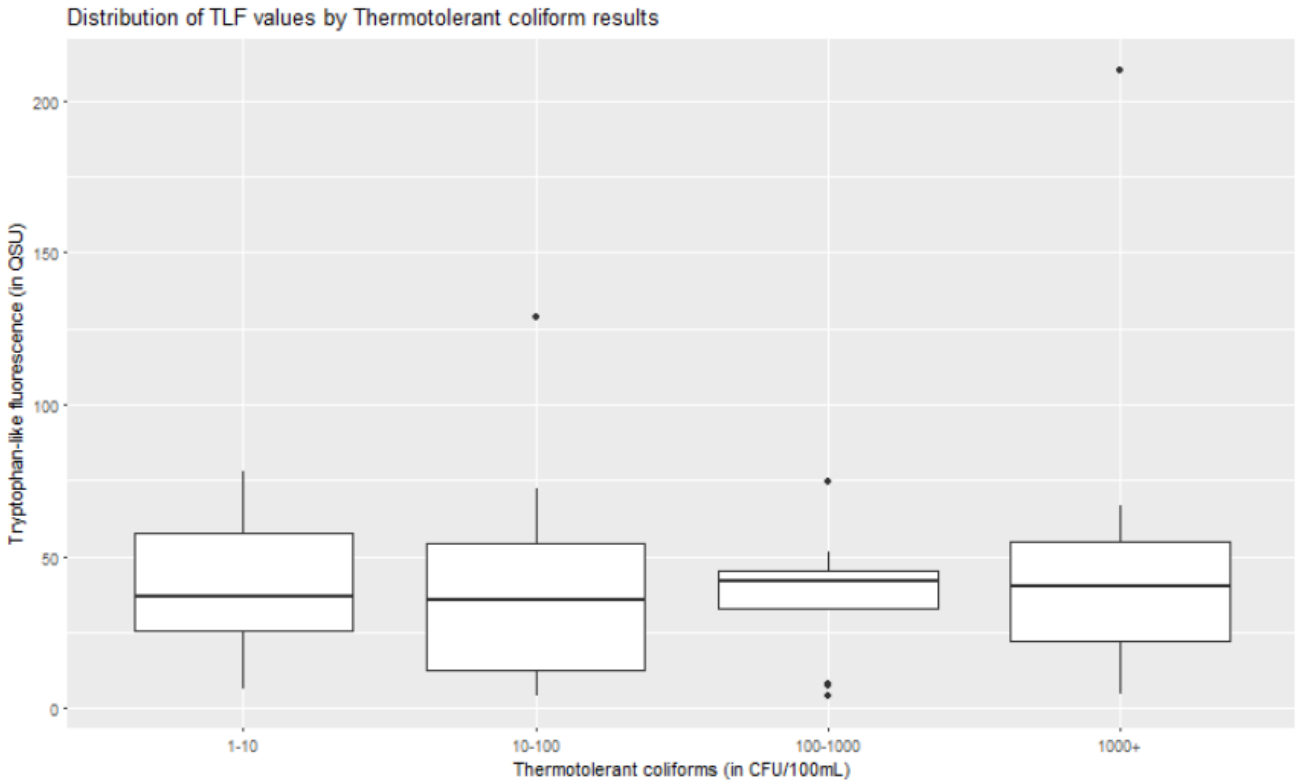
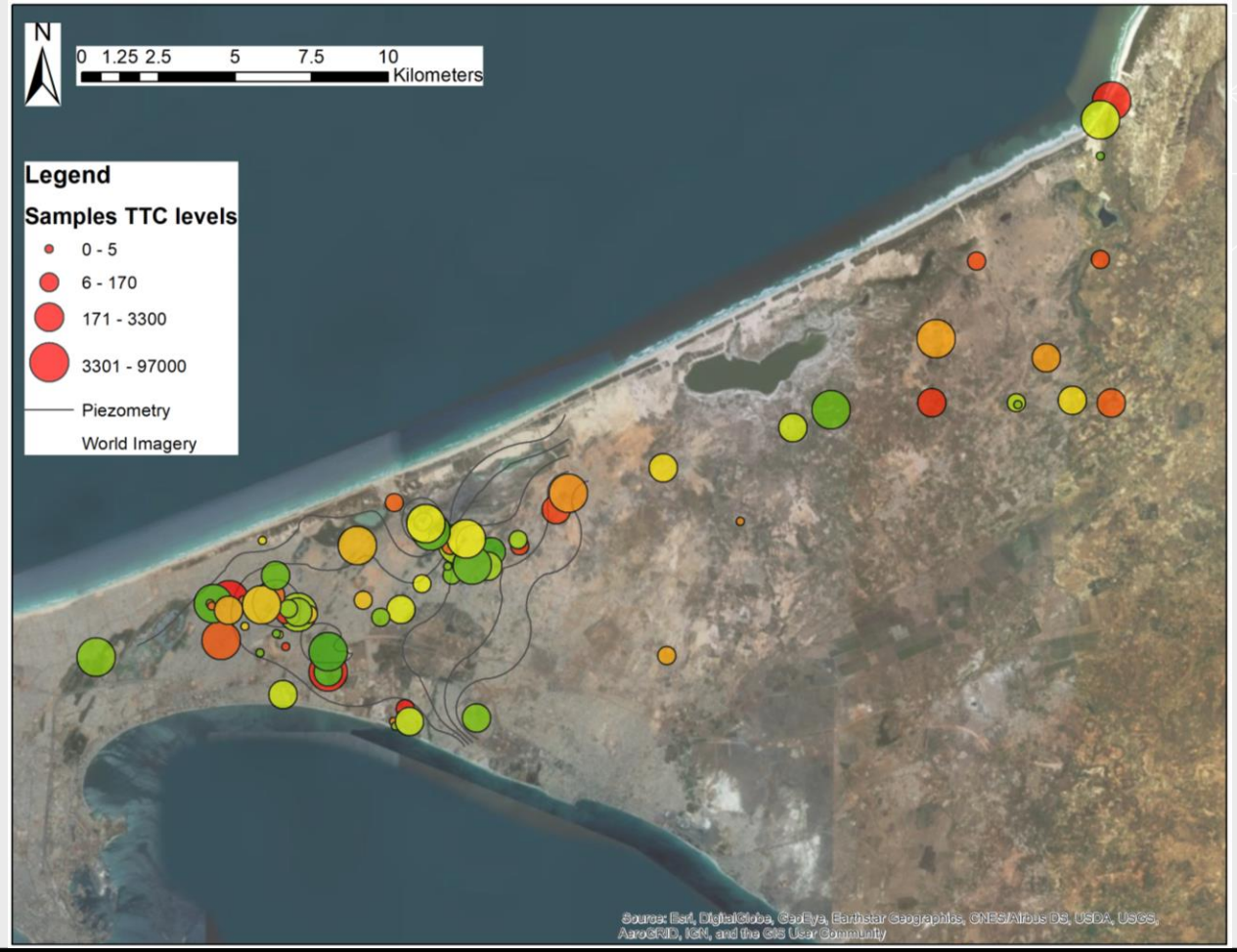


Figure 9: Boxplot of TLF by TTC count

The Unexpected strikes, part 1:TLF doesn't… work?

- Colour: green = low TLF , red = high TLF
- If there were a strong relationship, we'd only observe small green bubbles of low actual contamination and low TLF, and large red bubbles of high actual contamination and high TLF



Lesson 2:

Stay truthful. No results is better than made-up results!

British Geological Survey
NATURAL ENVIRONMENT RESEARCH COUNCIL

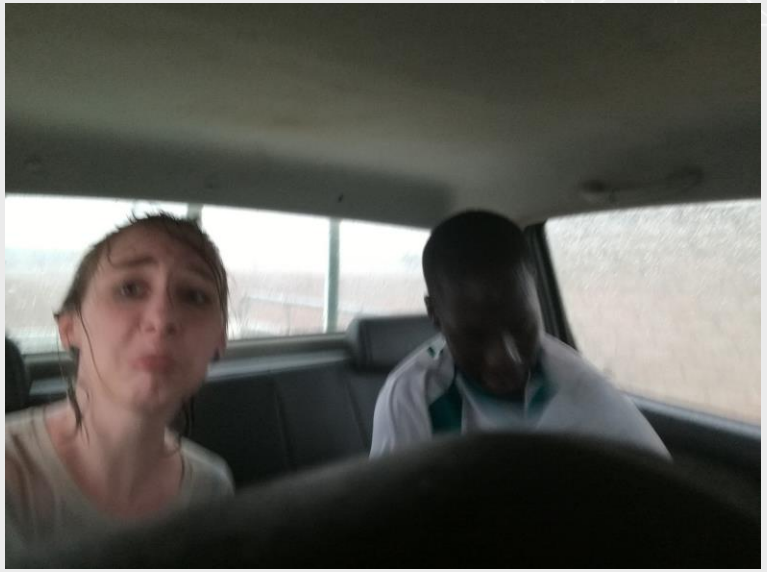
UKaid from the British people

THE ROYAL SOCIETY

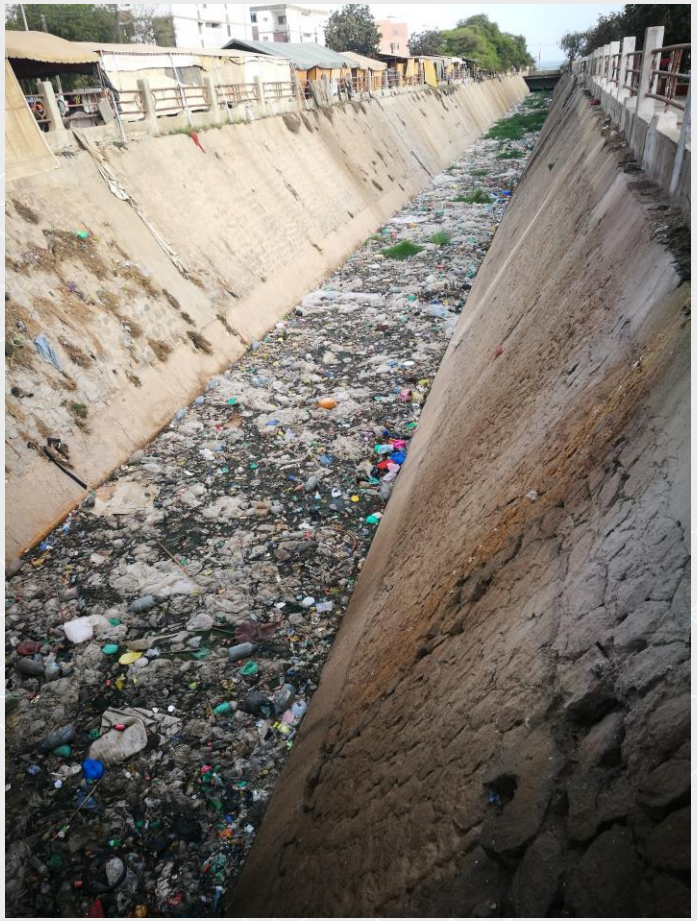
AfriWatSan

UCL

The Unexpected strikes, part 2: the RAIN



The Unexpected strikes, part 2: the RAIN



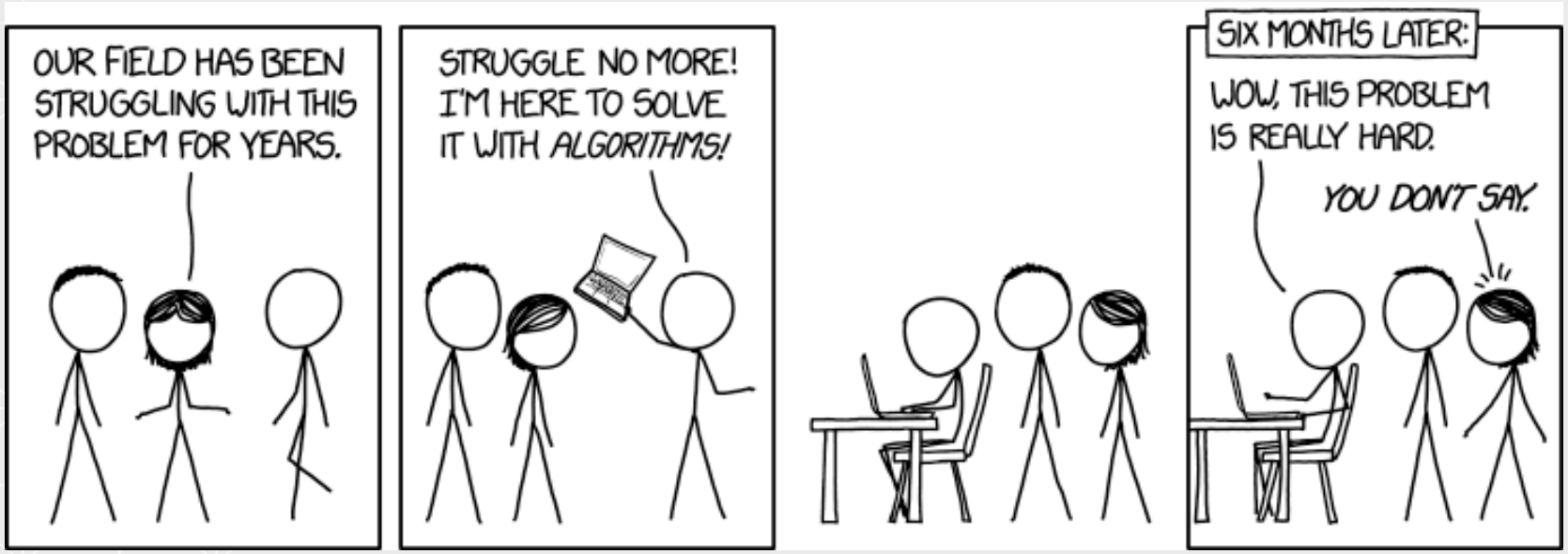
The Unexpected strikes, part 2: the RAIN

Re-samples after the first rain show that:

- TLF rates were systematically lower than the first time, suggesting a dilution effect
- BUT TTC rates were higher (as much as 10x) than before the rain, which points to an additional contamination of the shallow aquifer.

Lesson 3:

Ask the experts! You probably don't have enough domain knowledge ¯_(ツ)_/¯



Working 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.

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?

British Geological Survey
NATURAL ENVIRONMENT RESEARCH COUNCIL

UKaid from the British people

THE ROYAL SOCIETY

AfriWatSan

Results

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

- No! Almost perfect absence of correlation: TLF/TTC spearman rank ρ = - 0.01190626

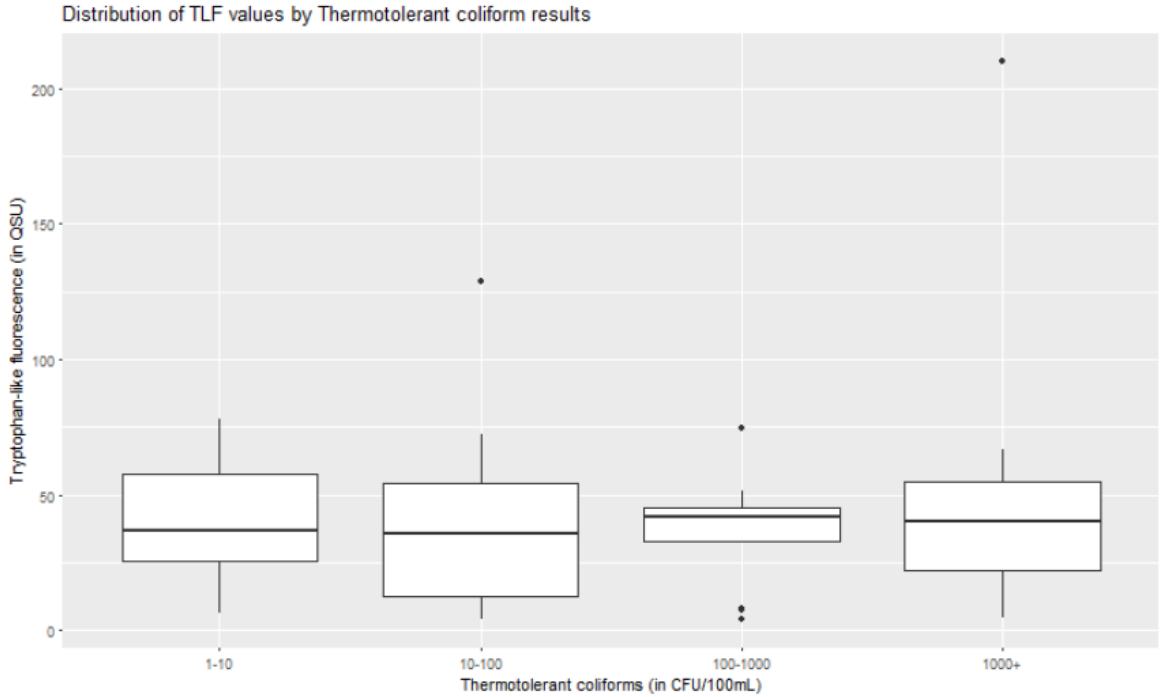
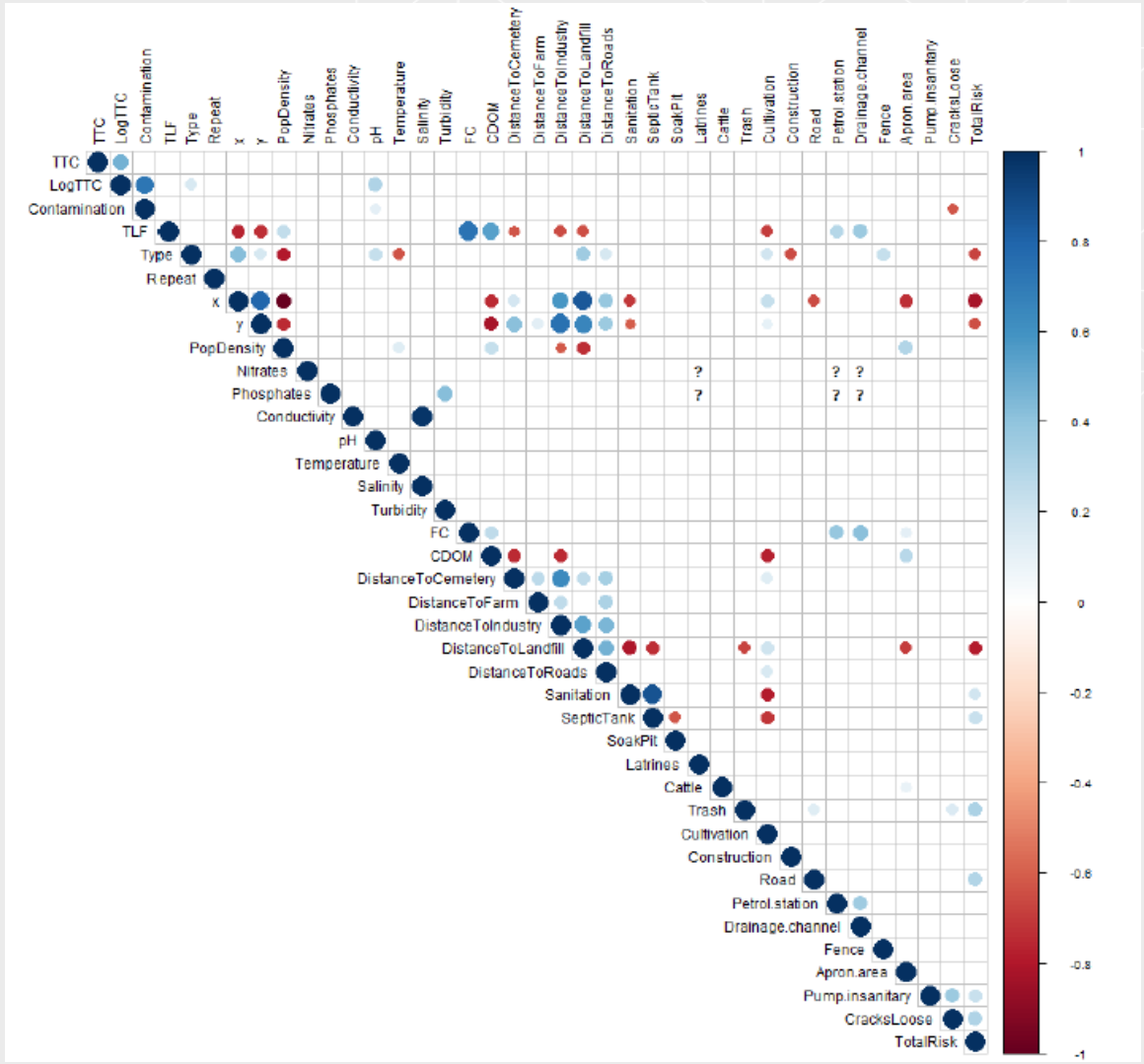


Figure 9: Boxplot of TLF by TTC count

Lesson 4:
Be flexible!

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

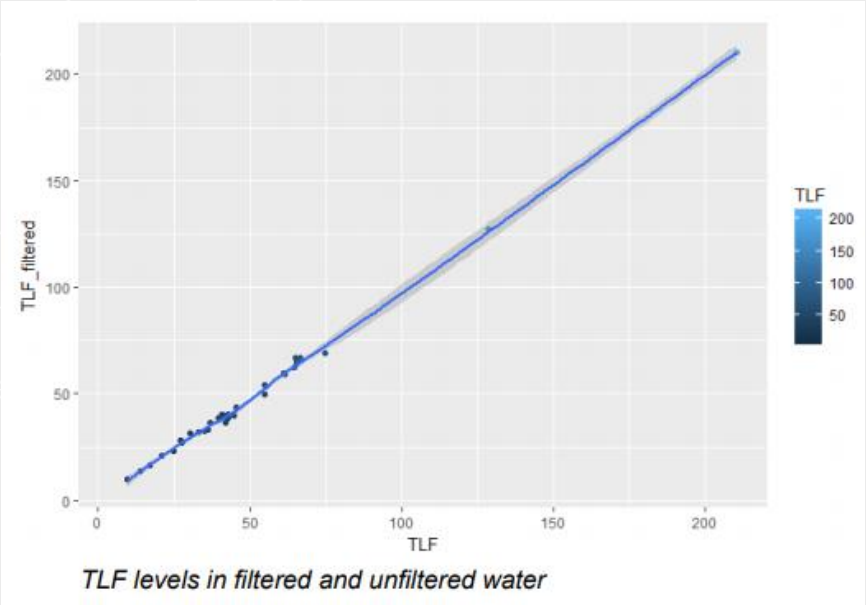
- Correlation Matrix
- Stepwise Logistic Regression, with 9 parameters retained as significant:
  - Latitude & longitude
  - Nearby presence of a septic tank / latrines
  - Nearby presence of a landfill
  - pH
  - Temperature
  - Turbidity
  - Flow cytometry count



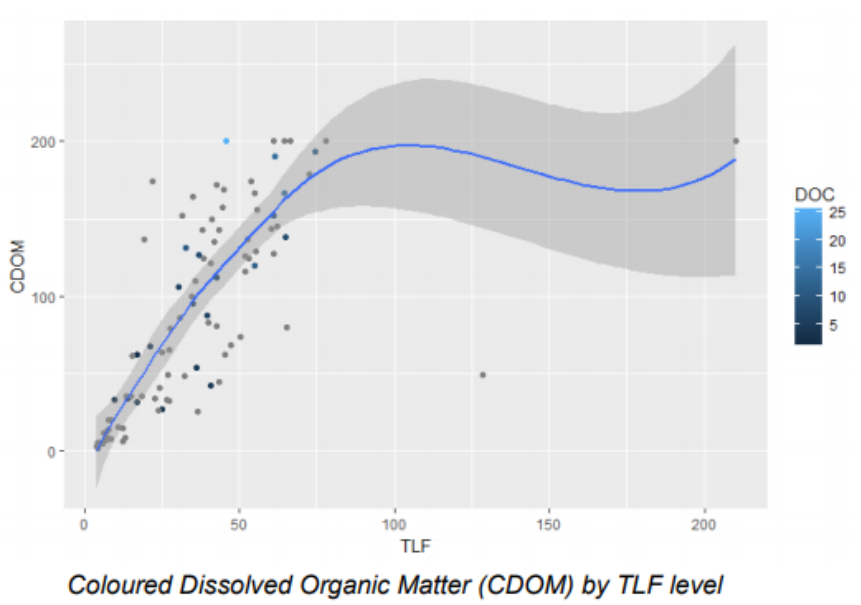
RQ3: What is the predictive power of the tryptophan-based method? How do various environmental factors affect its reliability?

- TLF is strongly correlated with Flow Cytometry and Dissolved Organic Carbon

→ It appears that TLF is extra-cellular and the fluorometers are measuring debris of past pollution



TLF levels in filtered and unfiltered water



Coloured Dissolved Organic Matter (CDOM) by TLF level

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

- TLF is negatively correlated to the presence of cultivation activities
- TLF levels are higher near 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).

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

- Our logistic regression based on the 9 selected parameters correctly classifies unknown data as contaminated or not contaminated **72.22%** of the time.

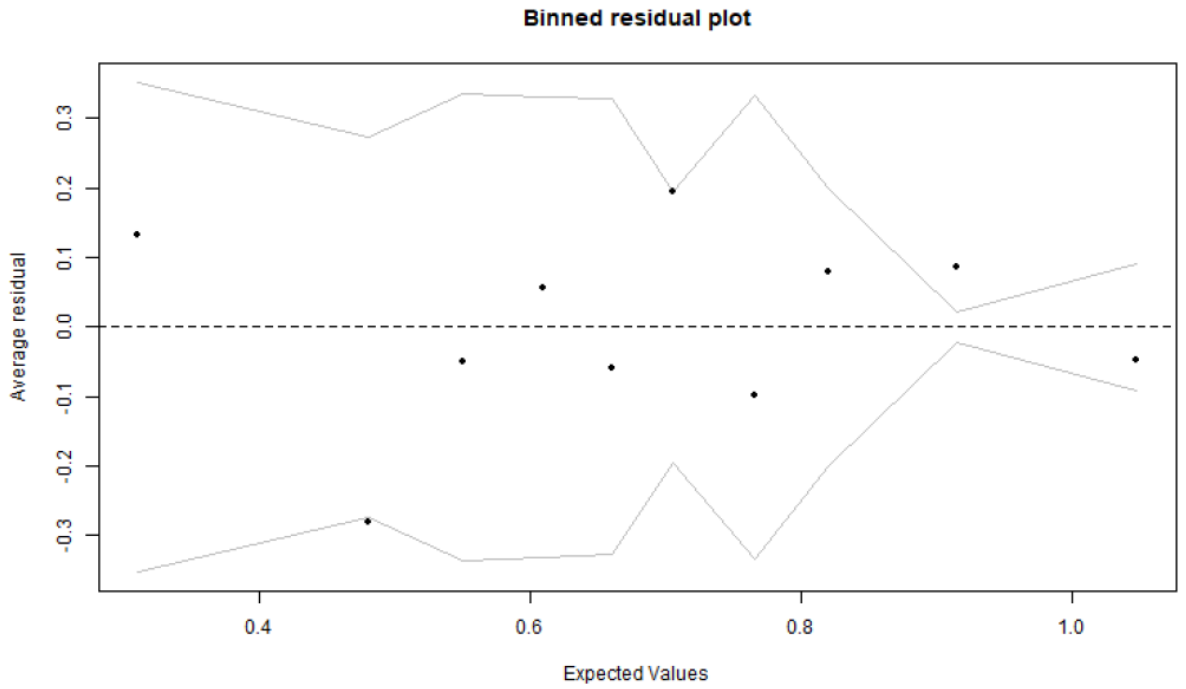
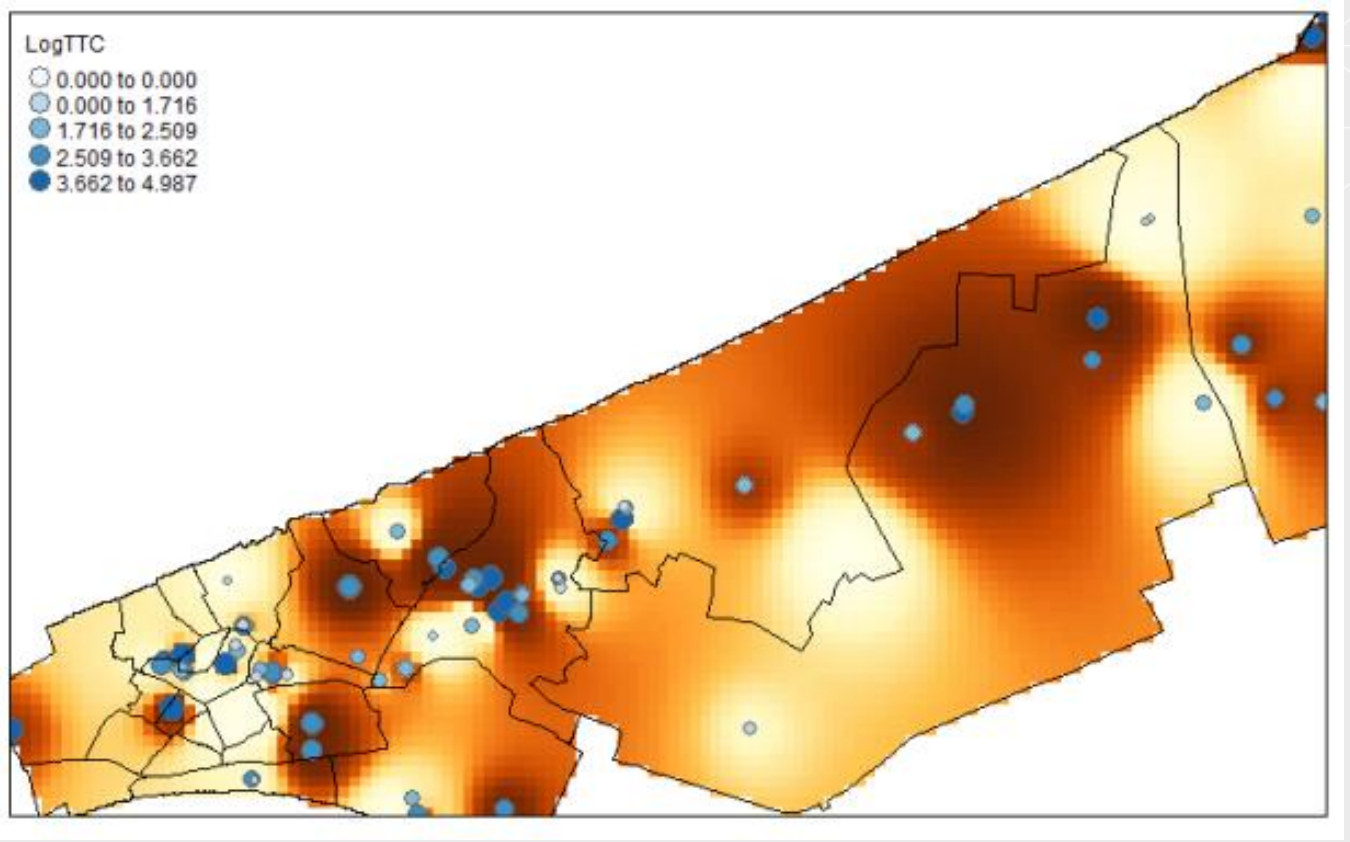


Figure 17: Binned residual plot of the logistic regression model

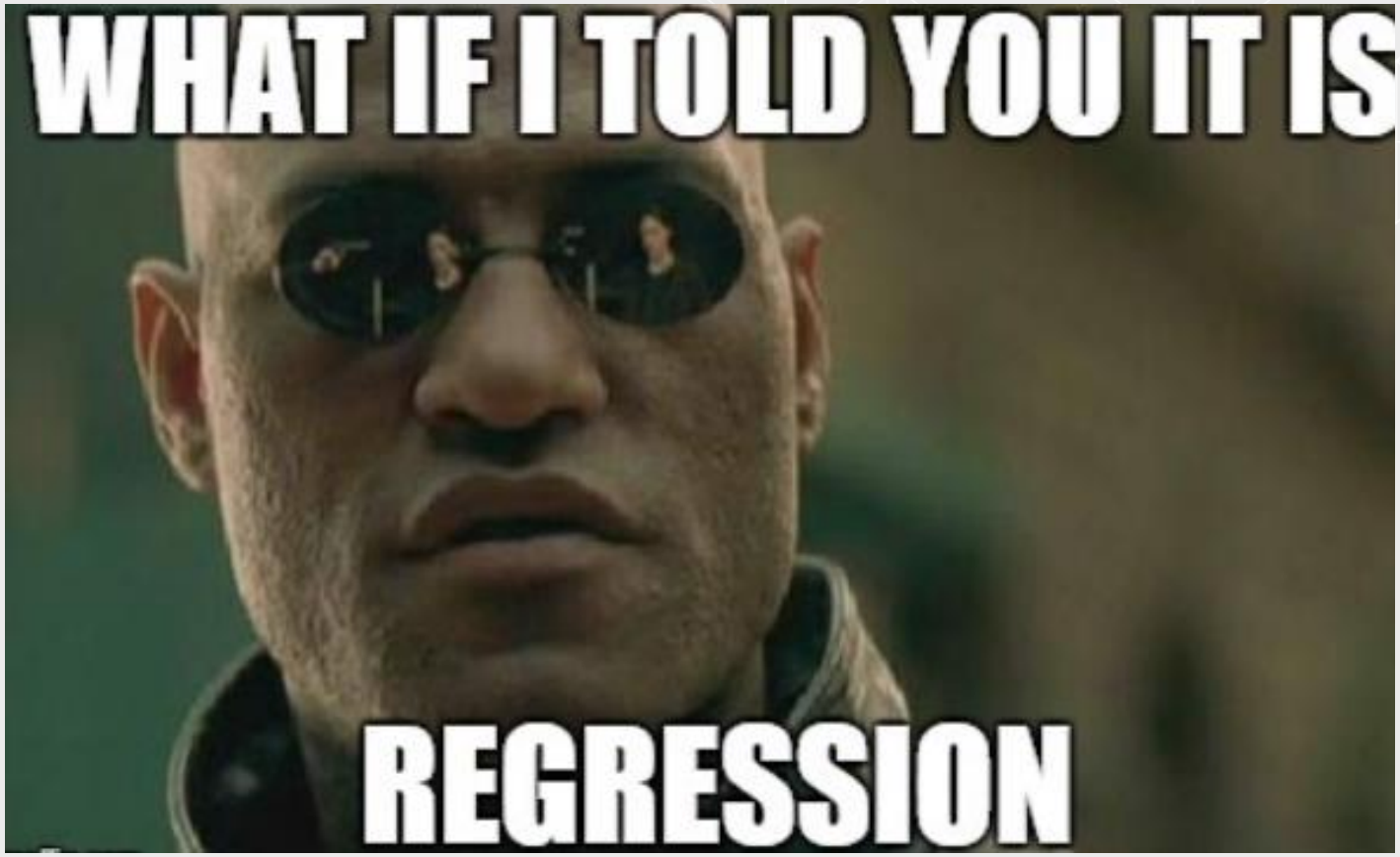
RQ5: Does faecal contamination demonstrate spatial patterns?

- Interpolation is not reliable in this case due to low degree of spatial autocorrelation in the contamination.

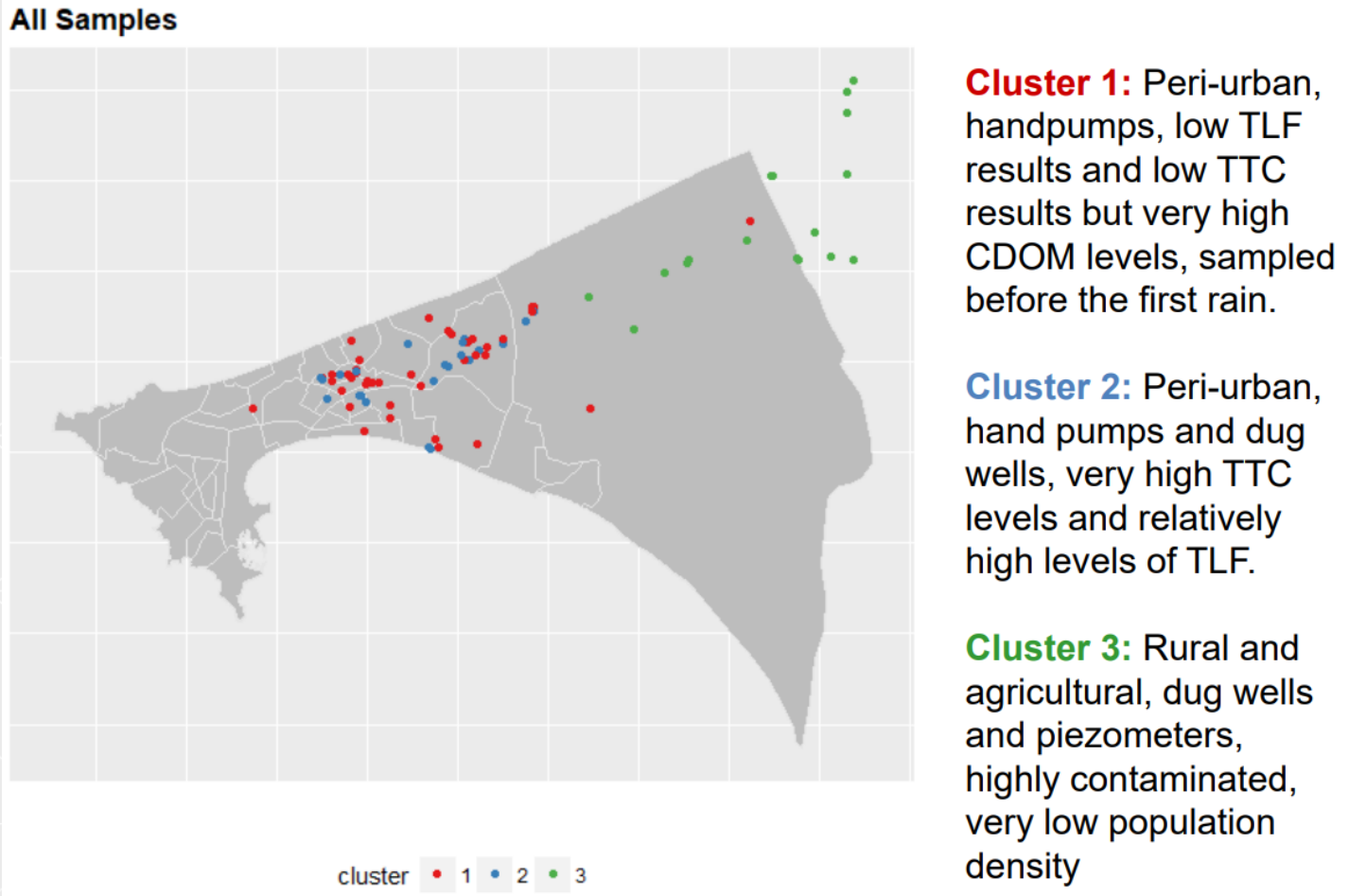


Lesson 5:

It's ok not to use fancy models/tools



RQ5: Can faecal contamination be classified?



Cluster 1: Peri-urban, handpumps, low TLF results and low TTC results but very high CDOM levels, sampled before the first rain.

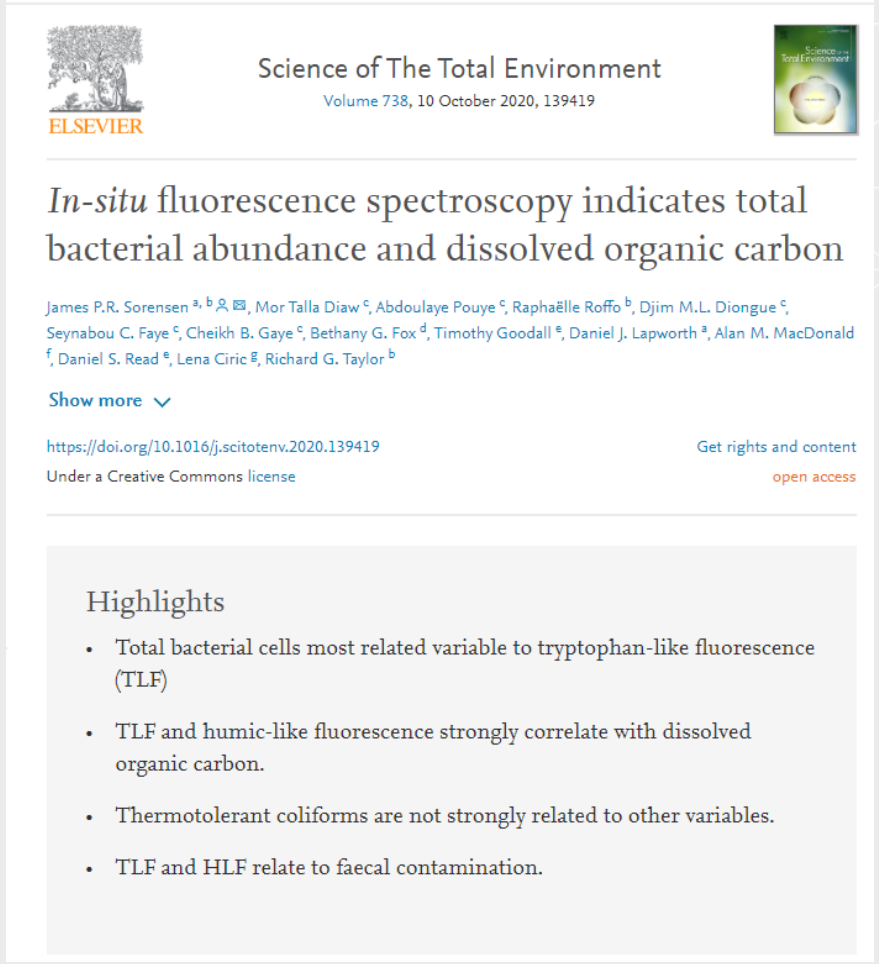
Cluster 2: Peri-urban, hand pumps and dug wells, very high TTC levels and relatively high levels of TLF.

Cluster 3: Rural and agricultural, dug wells and piezometers, highly contaminated, very low population density

Discussion

Why this research matters

- Further our understanding of TLF
- Flag the limitations of TLF
- Skills exchange across the AfriWatSan network



ELSEVIER

Science of The Total Environment

Volume 738, 10 October 2020, 139419

In-situ fluorescence spectroscopy indicates total bacterial abundance and dissolved organic carbon

James P.R. Sorensen [a, b], Mor Talla Diaw [c], Abdoulaye Pouye [c], Raphaëlle Roffo [b], Djim M.L. Diongue [c], Seynabou C. Faye [c], Cheikh B. Gaye [c], Bethany G. Fox [d], Timothy Goodall [e], Daniel J. Lapworth [a], Alan M. MacDonald [f], Daniel S. Read [e], Lena Ciric [g], Richard G. Taylor [b]

Show more

https://doi.org/10.1016/j.scitotenv.2020.139419

Highlights

- Total bacterial cells most related variable to tryptophan-like fluorescence (TLF)
- TLF and humic-like fluorescence strongly correlate with dissolved organic carbon.
- Thermotolerant coliforms are not strongly related to other variables.
- TLF and HLF relate to faecal contamination.

Next Steps

- Incorporate groundwater flows (vertical and horizontal) into groundwater pollution modelling
- Improve sampling scheme (very difficult!)
- With access to historical pollution and land use data, investigate links between historical loads of faecal bacteria and current TLF & CDOM rates

British Geological Survey
NATURAL ENVIRONMENT RESEARCH COUNCIL

UKaid from the British people

THE ROYAL SOCIETY

AfriWatSan

UCL

Lessons Learnt: recap!

British Geological Survey
NATURAL ENVIRONMENT RESEARCH COUNCIL

UKaid from the British people

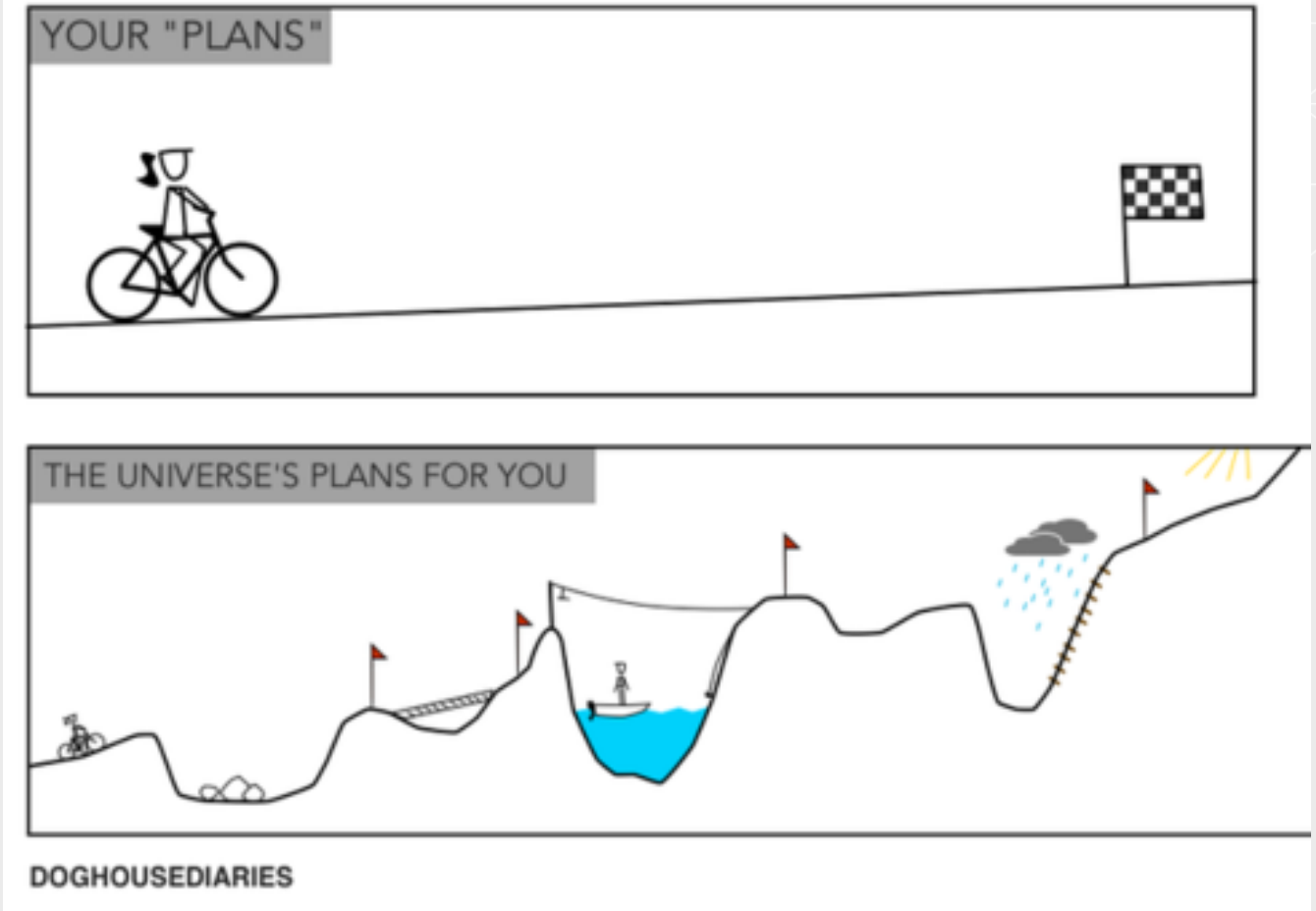
THE ROYAL SOCIETY

AfriWatSan

UCL

Lesson 1:

Plan for the unexpected

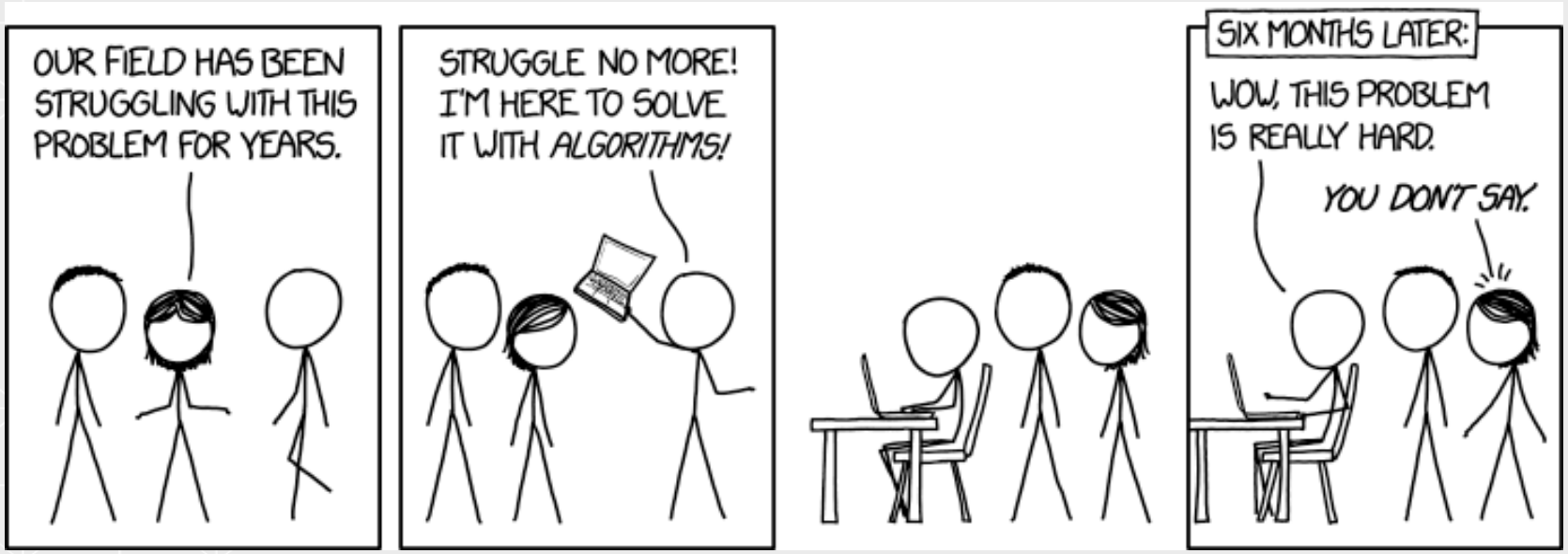


Lesson 2:

Stay truthful. No results is better than made-up results!

Lesson 3:

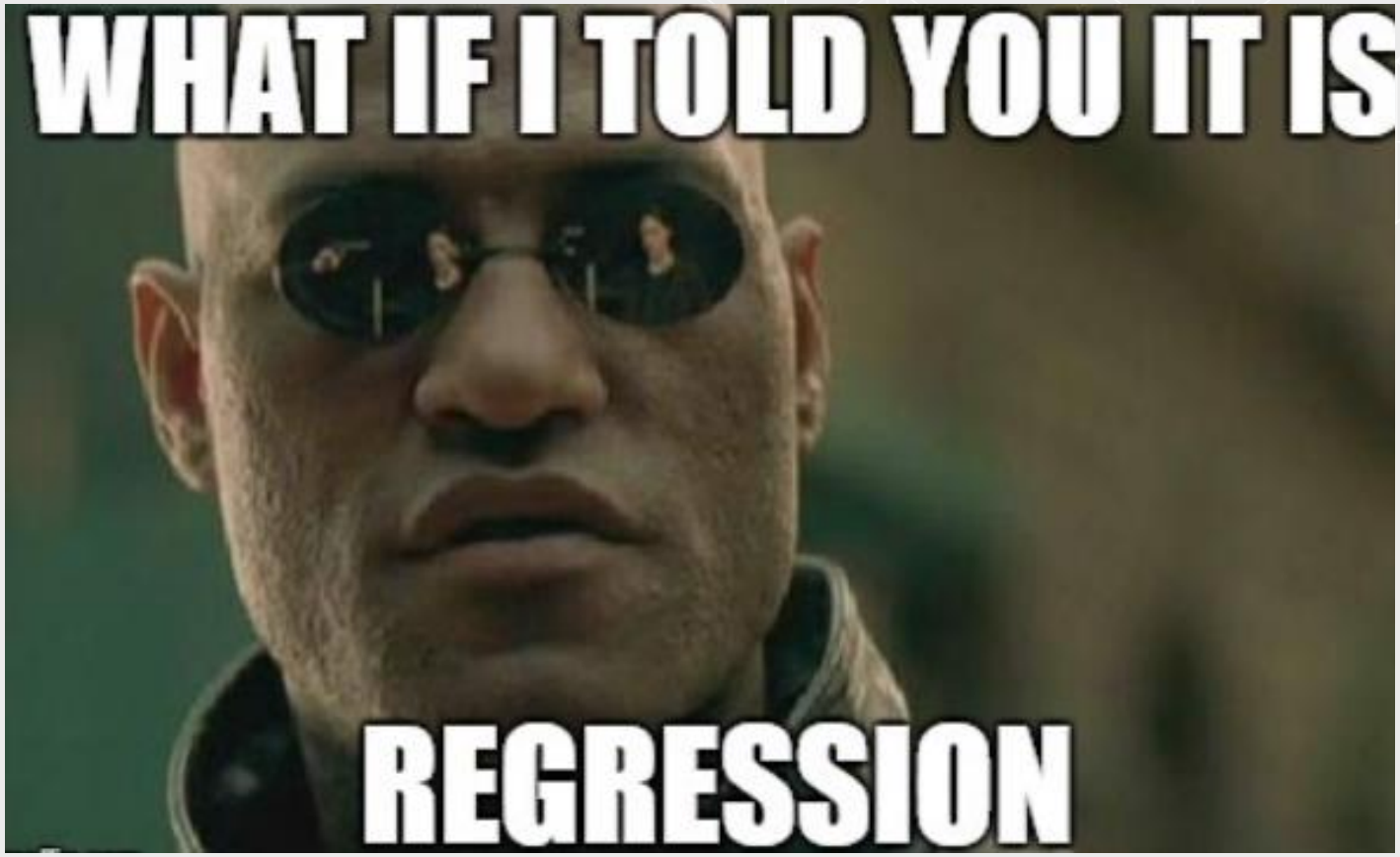
Ask the experts! You probably don't have enough domain knowledge ¯_(ツ)_/¯



Lesson 4:
Be flexible!

Lesson 5:

It's ok not to use fancy models/tools



British Geological Survey
NATURAL ENVIRONMENT RESEARCH COUNCIL

UKaid from the British people

THE ROYAL SOCIETY

AfriWatSan

Thank you for listening!

Q&A