Analysing the determinants of Extended-Spectrum β-Lactamase-producing *Escherichia coli* and *Klebsiella pneumoniae* colonisation in the Malawian community setting

by

Mélodie Sammarro

This thesis is submitted in partial fulfilment of the requirements for the degree of

Doctor of Philosophy

in the

Faculty of Health and Medicine

Lancaster Medical School

August 2022

To my uncle Ottavio, who was my biggest supporter when I decided to pursue a PhD,
and who wanted to be the first one to see my name on a paper,
I hope you're proud..

Declaration

This thesis has not been submitted in support of an application for another degree at this or any other university. It is the result of my own work and includes nothing that is the outcome of work done in collaboration except where specifically indicated. This thesis does not exceed 80,000 words. Many of the ideas in this thesis were the product of discussion with my supervisors, Dr. Chris Jewell, Barry Rowlingson and Prof. Nicholas A. Feasey.

Mélodie Sammarro, BSc, MSc, MRes
Lancaster University, UK
December 2021

Abstract

Mélodie Sammarro - Analysing the determinants of Extended-Spectrum β-Lactamase-producing *Escherichia coli* and *Klebsiella pneumoniae* colonisation in the Malawian community setting

Antimicrobial resistance is a health issue of global concern, involving the human, food and environmental sectors. A prime example of this threat is the rapid evolution of Extended-Spectrum β-Lactamase (ESBL)-producing *E. coli* and *K. pneumoniae*[1, 2]. These bacterial species are resistant to most β-lactam antibiotics, and in sub-Saharan Africa, where last resort antimicrobials are not always available, they may render ESBL infections untreatable[3]. Interrupting transmission leading to human gut mucosal colonisation appears like an attractive strategy to prevent infections[3]. However, little is known about risk factors for human gut mucosal colonisation with ESBL-producing Enterobacteriaceae (ESBL-E) in community settings in sub-Saharan Africa. Here, we suggest the importance of within-household transmission in driving ESBL colonisation in the community by determining various risk factors for ESBL-producing *E. coli* and *K. pneumoniae*. We also highlight faecal-oral and environmental routes as potential routes of transmission, through the identification of gender and various water, sanitation and hygiene (WASH) components as risk factors. These are determined using spatial and longitudinal approaches on the data collected using the modified spatial design described in this thesis. The findings indicate that transmission is complex in this setting, with individual, household and WASH components appearing as important factors. We suggest that main transmission pathways might differ depending on the bacterial species, therefore interventions might need to vary. We also recommend that interventions aimed at preventing transmission might have the best impact when targeted at the household-level

and focused on modifying the WASH behavioural practice and/or improving the WASH infrastructure. Additionally, we show that antibiotic use is important when looking at colonisation with ESBL-producing *K. pneumoniae* and therefore infection prevention and control measures and antibiotic use and stewardship training could help in preventing transmission. Finally, we report a prevalence of ESBL-producing *E. coli* of 37% in the community setting, which is comparable to some of the highest prevalences reported in the world[4].

Acknowledgements

I would like to thank my supervisors Dr. Chris Jewell, Barry Rowlingson and Prof. Nicholas Feasey for their amazing advice, support and encouragement throughout my PhD. I especially want to thank Chris for being the best supervisor anyone could wish for and being extra supportive through the ups and downs of PhD life, Barry for joining halfway through and being there instantly any time I have needed him since and Nick for including me in the DRUM consortium, always being willing to help and giving me the chance to analyse the DRUM data and go to Uganda. Additionally, I want to thank Peter J. Diggle for the incredible experience I had working on a project with him during the first year of the program, I learned a lot and it was truly an eye-opening time for me. I also want to thank Luigi Sedda for the temporary supervision but continued advice and support, and my PhD confirmation panel, Frank Dondelinger and Rhiannon Edge, for the constructive suggestions and the positivity they showed towards my plan. I also want to thank all the members of the DRUM consortium, especially Derek Cocker, Tracy Morse and Kondwani Chidwisano for their help with the field work and WASH data. I am truly lucky that I was able to work with experts in many different fields and I am sincerely grateful for all the lessons learnt. Lastly, I am very grateful to the UK Medical Research Council for providing the funding, as well as Lancaster University and Liverpool School of Tropical Medicine for allowing me to pursue this PhD.

I am also extremely thankful to my flatmates and friends, who made the PhD life a lot easier to go through and gave me the most amazing memories of my time in Lancaster and Liverpool. I am grateful to Mi, Ella, Abbie, Hannah and especially Silviya for taking care of me when my health was not at its best, to Genny and Kazi for the much-needed distractions, to Josh for being my own PhD support group, there is truly no one I would

rather have gone through PhD life with, and to Diana for being the best and most supportive friend and for inspiring me every day to be my best self. Finally, I want to thank my (immediate and second) family for the support and patience, more specifically my sister for the various moves and the food, my father for trying to understand what I do and especially my mum for the extra support (and gifts) without which it would have been a lot harder to get through the writing up period.

Table of contents

List of Figures

List of Tables

Acronyms

AIC Akaike Information Criterion

AMR Antimicrobial Resistance

CPT Cotrimoxazole Preventive Therapy

DRUM Drivers of Resistance in Uganda and Malawi

ESBL Extended-Spectrum β-Lactamase

ESBL-E Extended-Spectrum β-Lactamase-producing Enterobacteriaceae

ESBL-Ec Extended-Spectrum β-Lactamase-producing *Escherichia coli*

ESBL-K Extended-Spectrum β-Lactamase-producing *Klebsiella pneumoniae*

FAO Food and Agriculture Organisation

GAM Generalised Additive Models

HIC High-Income Countries

HMC Hamiltonian Monte Carlo

ICC Intraclass Correlation Coefficient

ICP Inhibitory design with Close Pairs

IPC Infection Prevention and Control

LMIC Low- and Middle-Income Countries

MCMC Markov Chain Monte Carlo

MDR Multidrug-Resistant

MRC Medical Research Council

NUTS No U-Turn Sampler

OIE World Organisation for Animal Health

OR Odds Ratio

OSM OpenStreetMap

PCR Polymerase Chain Reaction

sSA sub-Saharan Africa

STRATAA Strategic Typhoid alliance across Africa and Asia

WASH Water, Sanitation and Hygiene

WHO World Health Organisation

Chapter 1

Introduction

1.1 Thesis overview

This thesis focuses on the transmission patterns of ESBL-producing *E. coli* and *K. pneumoniae* leading to human gut mucosal colonisation in the community in Southern Malawi. This involves a spatial and temporal investigation of risk factors for colonisation, accompanied by a new extended spatial sampling design adapted to various settings.

1.2 Antimicrobial resistance

1.2.1 A global threat

Recently declared as one of the top 10 global public health threats facing humanity by the World Health Organisation (WHO)[5], antimicrobial resistance (AMR) has become a potential impediment to the achievement of the sustainable development goals[6, 7]. Antimicrobial resistance takes place in situations where infection-causing micro-organisms such as bacteria or fungi evolve over time and develop the ability to survive exposure to their respective antimicrobial drugs[8]. These drugs, which are commonly used to prevent and treat infections in humans, animals and plants, therefore become ineffec-

tive in the fight against infections and the risk of disease spread, severe illness or death increases[9].

1.2.2 Consequences of antimicrobial resistance

Now that many microorganisms causing common human diseases, including tuberculosis and malaria, have become resistant to a wide range of antimicrobial medicines, the use of "last-resort" medicines, which are most costly, potentially more dangerous and often unavailable or unaffordable in low- and middle-income countries, is becoming more frequent[5]. For instance, due to a lack of rapid diagnostic test for gonorrhoea and the rapid development of resistant strains, there will be no options left once the bacteria develops resistance to the current "last-resort" available antibiotic[8]. In 2014, the World Health Organisation described the post-antibiotic era we were heading towards, an era in which the global population could die of simple infections or injuries, due to the loss of effectiveness of the antibiotics that have been protecting us for the past 70 years[10]. This will slowly result in older techniques being used to control or treat infections, which will impact the treatment duration, invasiveness and success[5, 11]. Not only life-threatening infectious diseases will become more lethal, but complex procedures, such as surgery and chemotherapy, will become more risky without effective antibiotics to prevent infections[5, 8, 11].

Consequently, duration of illness and mortality will increase, hospital stays will become longer and more expensive medicines will be required[5]. This will also impact the economy at the societal and individual level, with loss of productivity in the workforce for longer periods of time, higher costs of treatment and longer hospital stays[11, 12]. Livelihoods will also be damaged through the loss of effective antimicrobials to treat animals in the food production sector and the additional burden of caring for the people infected by resistant bacteria[5, 11]. The economy will also be affected by the need to prevent other infections within or outside healthcare facilities[12]. Deaths caused by resistant infections are already estimated at 700,000 deaths a year, and without any sustainable action, they could reach 10 million deaths a year by 2050[8]. As it stands, the World Bank has also estimated that by 2030, there could be an additional 24.1 million people suffering from extreme poverty due to AMR, of whom 18.7 million would

be in low-income countries[13].

The reason behind the rise in concern over antimicrobial resistance in recent years is the lack of new antimicrobial drugs to counter the resistant strains, as much as the rapidity at which it is developing and spreading[8]. New antimicrobials are urgently needed, and they need to be regulated to ensure a more responsible and sustainable use than in the past, in order to slow down the development of antimicrobial resistance[8, 9, 14]. However, to this day, only few innovative antimicrobials are currently in clinical development to address the list of priority pathogens drawn up by the World Health Organisation[8, 9]. According to the latest report of WHO on data from 2020, only 26 out of the 43 antibiotics in the clinical antibacterial pipeline are active against the WHO priority pathogens. Among these, only seven fulfil at least one of the WHO innovation criteria and only two are active against the critical Gram-negative bacteria[15]. Moreover, over 80% (9/11) of the new antibiotics approved since 2017 are from existing classes where resistance mechanisms are well recognised and where rapid emergence of resistance is to be expected[15].

In the last three decades, investments in research and development of new antimicrobials have significantly dropped and pharmaceutical companies have shut down antibiotic discovery programs[16]. This is related to the inability for antimicrobials to be a good source of income for pharmaceutical companies. Due to their role as the 'last-line defence', even newly developed antibiotics would not be used frequently[17]. Additionally, promising drugs would also be abandoned when commercial or global health priorities change[17]. Competition between pharmaceutical companies also has an impact, as many drugs in development remain inaccessible to the public or other companies once the company has filed a patent[17]. The WHO estimated that out of the 10 antibiotics in Phase 1 that are possibly active or active against the critical Gram-negative bacteria, only one will likely make it to market in the next 10 years[15]. In 2020, the International Federation of Pharmaceutical Manufacturers & Associations launched, in collaboration with many companies and the WHO, the AMR Action Fund[18], whose goal is to help accelerate the research and development of new antimicrobials[15].

1.2.3 A complex process

Antimicrobial resistance is a normal evolutionary process for microorganisms that has been observed since before the first antibiotic was introduced. In fact, the first penicillin-resistant strain of bacteria was found in 1940, a year before it was first used on the public[8, 19]. Since then, resistance to an antibacterial drug has consistently been detected following the development of said antibacterial drug[10, 20, 21]. Antimicrobial-resistant bacterial strains have even been recently found in permafrost sediments that were around 40,000 years old[22]. However, the increasing widespread use of antibiotics in the world and the lack of AMR surveillance, infection prevention and control have allowed it to become a problem of global concern in the last decades[8, 10, 23]. Gradually, resistance has increased and stabilised in bacteria leading to a new phenomenon called multidrug-resistant bacteria (MDR)[8]. Multidrug resistance is defined as "acquired non-susceptibility to at least one agent in three or more antimicrobial categories"[24]. MDR bacteria are now responsible for causing infections that are difficult to treat or not treatable with existing antimicrobial medicines such as antibiotics[9]. Bacterial species can become resistant or multidrug-resistant in many ways, such as through the accumulation of resistance genes or through a single resistance mechanism that gives resistance to more than one antibacterial agent[25]. Antibiotic resistance can spread through bacteria populations when they inherit antibiotic resistance genes, and when they share or exchange sections of genetic material with other bacteria. This latter process, called horizontal gene transfer, occurs in three main ways: conjugation, where two bacteria connect and transfer plasmids containing resistance genes from one to the other; transduction, where bacteriophages bring along genes that they picked up during infection of another bacterium and finally transformation, where bacteria can take up pieces of DNA directly from the environment around the cell[20]. Through these various processes, sharing of genetic material can even occur between different bacterial species, therefore affecting a wider range of infections[5]. Antibiotics remove drug-susceptible bacteria, leaving resistant bacteria behind to reproduce as a result of natural selection[21]. Antimicrobial-resistant organisms can be found in humans, animals, food and in the environment and can circulate between them[5, 9].

1.2.4 Influencing factors

1.2.4.1 Misuse and overuse of antibiotics

Since the initial discovery of penicillin in 1928, antibiotics have changed the practice of modern medicine and saved countless lives. However, the main drawback from this breakthrough is the arising of a general belief among the population that antimicrobials are universally efficacious and should be applied to cure most infections[5, 11]. As a result, their use became more widespread in both hospital settings and community settings. Overuse in hospital can present itself in life-threatening cases where the clinician lacks time to accurately diagnose the infectious disease or its pathogen, such as when patients meet the definition of suspected sepsis and are therefore commonly prescribed broad spectrum antibiotics[11]. The overuse of antimicrobials prescribed by general practitioners can also occur, due to patient pressure or a lack of time and/or resources to identify the infection[5, 11]. The lack of availability of rapid diagnostic test for infectious diseases can cause the practitioner to rely too often on educated guesses and prescribe inappropriate antimicrobials for the patient's actual disease. For instance, although many respiratory infections are caused by viruses, about half of the patients receive unnecessary antibiotics[16, 26]. Patients might also misuse antimicrobial drugs by re-using courses that were not previously completed or by continuing them for longer than needed[16].

In 2015, a "country situation analysis" conducted by the World Health Organisation in Member States in each of the six WHO regions showed that the sale of antimicrobial drugs is still widely unmonitored and regulations on the sale of prescription-only drugs were not enforced in multiple regions[27]. Many countries additionally lacked standard treatment guidelines for health care workers[23, 27]. Furthermore, in low- and middle- income countries (LMICs), the risk of inadequate dosage is higher, influenced either by drug sharing or by using counterfeited antimicrobials[16, 28]. Although there has been an increasing trend in the implementation of national monitoring systems of antimicrobial consumption in the recent years, over half of the low-income countries that took part in the country self-assessment survey by the Tripartite (WHO, Food and Agriculture Organization of the United Nations (FAO), World Organisation for Animal Health (OIE)) in 2020 reported not having a national plan or system to monitor the use

of antimicrobials in human health[29].

Furthermore, little variation exists between the classes of antibiotics used in animals and in humans, increasing the risk of resistant bacteria spreading between animal species and humans[10, 23]. The use of antimicrobials in animals varies from disease treatment to disease prevention, but mainly poses a threat when used as growth promoters through mass administration to herds[5, 30]. Although the addition of any antibiotic as a growth promoter in animal feed has been banned in the European Union since 2006, such use is still common practice in many low- and middle-income countries[5, 30, 31].

1.2.4.2 Water, sanitation and hygiene (WASH) and infection prevention and control (IPC)

Enteric bacteria, whether antimicrobial-resistant or not, can either directly shed into the environment, whether through open defecation, or to the sewerage system, through faecal matter[32, 33, 34]. They can directly contaminate rivers or spread through sewer sludge, which can be used as a fertiliser[33, 35]. In the same way, antimicrobial use in animals increases the risk of contaminating the environment when treating the herds directly on pasture or in aquaculture[32, 33, 36]. Animal faeces is also commonly used as manure, indirectly spreading antimicrobial resistance further into the environment[5, 32, 36, 37]. The use of sludge and/or manure as a fertiliser also jeopardises surface or ground waters[33, 34, 37].

The natural environment, which encompasses water, soil and air, has been identified as a potential pathway for transmission of antimicrobial resistance[20, 37, 38]. However, due to its complexity, the interaction between environment and antimicrobial resistance is still not well understood[36, 38]. In the environment, similarly to inside a human or animal host body, the addition of antimicrobials (and other agents such as pesticides) exerts a selective pressure on the existing environmental bacteria[16, 39, 40, 41]. This leads to the disappearance of the antimicrobial-susceptible bacteria and the development of ever more resistant strains, allowing for the antimicrobial-resistant agents to be shared with more ease across various bacterial populations and pathogens[32]. Antimicrobial-resistant bacteria are usually introduced into the environment through

human or animal faecal matter, however, humans can also find themselves exposed (or re-exposed) to these bacteria through contact with the environment[37]. In fact, environmental exposure can occur through different types of contaminated water sources, such as drinking water, recreational water or irrigation water[34, 37, 38, 40]. Due to the importance of water in our daily life, it has become an unavoidable pathway for antimicrobial resistance to spread[32, 36].

Globally, over 80% of wastewater is released into the environment without adequate treatment[42]. In the developing world, up to 90% of all wastewater from piped sewage is released, without treatment, into rivers, lakes and oceans[43]. By unknowingly using contaminated manure or water to fertilise or irrigate crops, the fresh produce also becomes at risk of being colonised with antimicrobial-resistant bacteria. Likewise, the proximity of food-producing animals to contaminated sources of water such as rivers or wastewater puts entire herds at risk of becoming colonised, therefore additionally putting at risk the food produced[32, 37]. Moreover, it has recently been shown that direct inhalation or deposition of contaminated dust onto skin, food or water from livestock facilities expose populations to antimicrobial residues or antimicrobial-resistant bacteria[36, 44]. Intensive farming produce a larger quantity of waste than small or medium farms, yet due to the lack of appropriate waste removal systems, these farms also find themselves at increased risk through the contamination of the environment by the waste disposal of these intensive facilities[36, 45]. Subsequently, farm workers can themselves spread resistant bacteria or genes into their respective families and community[36, 46]. Smaller farms produce less waste, however, due to their size, regulations surrounding waste disposal are also less tightly regulated. Therefore, these farms might inappropriately dispose of their waste, by throwing it into nearby water sources or onto nearby land, potentially polluting the environment[36].

Due to the various interactions at the interface between humans, animals and environment through which antimicrobial resistance can occur (see Figure 1.1), the importance of looking at antimicrobial resistance in a One Health setting is unequivocally necessary[8, 16, 47, 48]. As of December 2021, One Health has been defined by the One Health High Level Expert Panel as "an integrated, unifying approach that aims to sustainably balance and optimise the health of people, animals and ecosystems"[49, 50]. It recognises that "the health of humans, domestic and wild animals, plants, and the wider

environment (including ecosystems) are closely linked and inter-dependent"[50]. These pathways are usually bi-directional, e.g. from humans to animals and animals to humans, and there is still a vast lack of knowledge surrounding their respective importance, especially in LMICs[36].

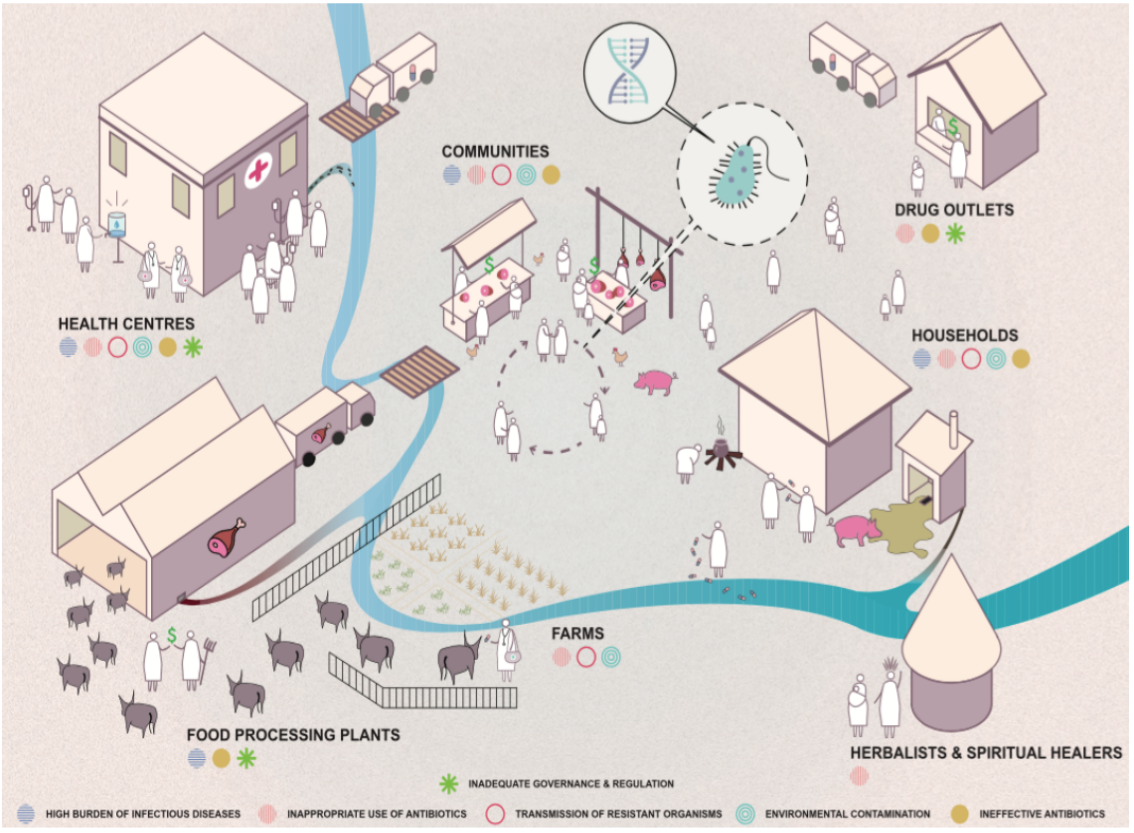


Figure 1.1: Hypothetical model of behaviours and movement of AMR determinants and bacteria in Uganda and Malawi. Created by Design Without Borders Uganda. ©DRUM Consortium

In 2021, a joint report from the World Health Organisation and UNICEF estimated that almost half of the global population, approximately 3.6 billion people, still lack safe sanitation while one in four people lack safe drinking water[51]. In fact, at least 494 million people still practise open defecation and at least 2 billion people around the world use a drinking source that is contaminated with faeces[51]. Furthermore, one in three people around the world lack basic hand-washing facilities[51].

Improving WASH infrastructure and behaviour will reduce transmission and help to curb the spread of AMR[52]. WASH in the healthcare environment forms part of

infection prevention and control (IPC) and IPC in healthcare facilities is a well-recognised means to limit the spread of AMR bacteria[47, 48]. However, it is still not well established or not established at all in many countries, especially in low- and middle-income countries where healthcare is further compromised by a lack of access to clean water, sanitation and hygiene infrastructure but also a lack of access to affordable and quality antimicrobials, vaccines and diagnostics[5, 52]. This is tragic as the Organisation for Economic Co-operation and Development (OECD) recently estimated that promoting simple infection prevention and control measures such as hand hygiene could reduce by about 40% the AMR health burden[53, 54].

1.2.4.3 International travel and trade

The growing ease of international travel and trade in the last few decades (excepting recent interruptions consequent upon the COVID-19 pandemic) has become an additional factor contributing to the dissemination of antimicrobial resistance [55]. In 2018, there were 1.4 billion tourist arrivals worldwide[56]. The ability to travel to anywhere in the world in a lapse of two days has made it very difficult for the human population to control the spread of infectious diseases. Furthermore, this ability is shared between humans and microbes as they can travel within various hosts, without even being detected, and therefore are able to cross the globe within a few days' time[11]. This has rendered all populations accessible to the threat of existing microbes from all environments that a few decades ago, would not have been able to reach them[5, 11]. Not only are those travellers at risk of exposure to resistant bacteria in endemic areas, they also are likely to return colonised[55, 57]. A key example of this is the rapid global spread of New Delhi metallo-β-lactamase-1–positive Enterobacteriaceae, first identified in Sweden in 2008 in an Indian patient previously hospitalised in New Delhi and subsequently detected on all continents within two years[58]. In addition to human travellers, international travel and trade has allowed other carriers, such as livestock animals or food products of animal origin, to spread resistant bacteria to other parts of the world [59, 60].

1.2.5 Antimicrobial resistance in low- and middle- income countries

Due to a higher burden of infectious diseases and a poor regulation of antimicrobial use in both humans and animals, low- and middle- income countries (LMIC) are especially vulnerable to the emergence of antimicrobial resistance[52, 55, 61]. The availability of non-prescription drugs and lack of healthcare access in some regions drives the misuse of the only available products[53]. Additionally, these drugs available without prescription are also commonly associated with poor quality or counterfeiting. The World Bank estimated that 60% of antimicrobials used in Africa and Asia contain little to no active ingredients[13]. Without appropriate access to healthcare and effective antimicrobials, infections that are usually treatable, such as respiratory infections or diarrhoea, account for an estimated 12% of deaths in LMICs[62]. While twenty years ago, LMICs had much lower antibiotic consumption than high-income countries (HIC) mainly due to a lack of affordability and access, between 2000 and 2015, antibiotic consumption in LMICs had a 114% increase, rapidly converging towards the same consumption as in HICs[63]. However, the lack of affordable and appropriate antimicrobials in LMICs increases the widespread use of the same antimicrobials, which are then commonly misused, thus driving resistance for these specific products. This increase in consumption of the same antimicrobials, the lack of resources to diagnose the considerable presence of infectious diseases in these areas, the unavailability of appropriate antimicrobials to treat them and the inadequate water, sanitation and hygiene infrastructure makes LMICs an ideal landscape for antimicrobial resistance to emerge and spread[52, 55].

In particular, between 2000 and 2015, consumption of the cephalosporin β-lactam antibiotic class, especially third-generation cephalosporins, increased rapidly in LMICs. Their use was rapidly accompanied by the emergence of extended spectrum β-lactamase (ESBL)-producing bacteria[63, 64]. ESBLs are enzymes produced by bacteria that break down and render ineffective a broad range of β-lactam antibiotics[65]. Genes coding for ESBLs are often situated on plasmids, and therefore propagate through bacteria by horizontal gene transfer [66, 67]. Often, these genes are located on the same plasmid as other AMR genes, which can lead to MDR, and complicate further the treatment of infections [66]. Plasmids are usually present in multiple copies on a bacterium,

causing the resistance genes they carry to rapidly evolve and consequently allowing the plasmids to act as an evolutionary catalyst to accelerate the evolution of new forms of resistance [68]. Due to the widespread use of β-lactam antimicrobials to treat bacterial infections, the frequent exposure of bacteria to such antibiotics has caused a continuous production and mutation of ESBLs[64, 65]. Since before 1990, approximately a thousand resistance-related β-lactamases inactivating β-lactam antibiotics have been found[23]. This rapid evolution of bacterial resistance causing such a sudden increase in multiple drug resistant organisms is of important public health concern[65]. In 2017, cephalosporin-resistant Enterobacteriaceae were included by the World Health Organisation in the list of priority 1 critical pathogens for which research and development of new antibiotic options are urgently needed[69]. ESBL-producing bacteria have been found all over the world, however whilst surveillance studies are present in Europe, North America and Asia, there is a substantial lack of research published on the situation in Africa[70]. In sub-Saharan Africa (sSA), where cephalosporins have increasingly become the antimicrobial of choice due to the unavailability of other reserve antibiotics, ESBL-producing bacteria are spreading rapidly[2, 71].

1.3 ESBL-producing Enterobacteriaceae

1.3.1 ESBL-producing *Escherichia coli* and ESBL-producing *Klebsiella pneumoniae*

Escherichia coli is a diverse bacterial species typically found in the intestines of animals and humans[72]. Although most strains of *E. coli* are harmless, some strains can cause various types of infection such as serious diarrhoea or urinary tract infections[72]. They are also the leading cause of bloodstream infections in the community and the second leading cause of these infections in hospitals globally[53]. *Klebsiella pneumoniae* is another bacterium often found in the human or animal intestines and the environment and is the archetypal nosocomial pathogen[73]. Similarly to *E. coli*, many infections, such as urinary tract infections, soft tissue and respiratory infections, can be caused by *K. pneumoniae*[71, 74]. Standard treatments for infections caused by either bacteria

include cephalosporins[53]. ESBL-producing Enterobacteriaceae are now a large problem in African healthcare institutions and communities, where *E. coli* and *K. pneumoniae* are predominant[70, 75, 76, 77, 78]. *E. coli* and *K. pneumoniae* are the two Gram-negative bacteria from the Enterobacteriaceae family in which ESBLs have been found most often[79]. In particular, the spread of drug-resistant *E. coli* is due mostly to the global dissemination of one specific sequence type 131 (ST131)[80]. Whilst ESBLs used to be found mainly in hospitals as the cause of hospital-acquired infections, they are now regularly detected in both hospitals and communities, although usually at a lesser extent in communities[70, 77, 81]. Combined with a lack of resources, this caused a focus on hospital-based research and resulted in an important lack of estimates for community-based ESBL carriage, especially in East Africa[77, 82].

1.3.2 ESBL-producing Enterobacteriaceae in East Africa

In East Africa, information on the prevalence of ESBL-producing Enterobacteriaceae is still scarce and most available studies are from hospitals, and many times from target populations within the hospital[10, 70, 77, 83]. A review from 2016 showed that in East African hospitals between 2003 and 2014, the overall pooled proportion of ESBL-producing Enterobacteriaceae was 42%, with wide differences between countries and between studies[77]. The lowest pooled ESBL-producing Enterobacteriaceae proportion was 0.30 (95 % CI: 0.21–0.38) in Ethiopia, while the highest was 0.62 (95% CI: 0.38–0.87) in Uganda[77]. Proportions in Tanzania and Kenya were respectively estimated at 0.39 (95 % CI: 0.30–0.48) and 0.47 (95 % CI: 0.23–0.71)[77]. Previous estimates from East African hospital studies ranged from 0.7% to 22.8%, with ESBL *E. coli* ranging from 14 to 25% and ESBL *K. pneumoniae* from 13 to 17.3%[76]. A more recent review of ESBL-producing Enterobacteriaceae prevalence in humans, animals and the environment in Tanzania showed an overall prevalence of 22.6%. The author suggested that these differences could be explained by the fact that the ESBL samples from the 2016 review were taken from hospitals where multidrug-resistant bacteria is very common[78]. A separate review from 2019 on the prevalence of third-generation cephalosporin-resistant Enterobacteriaceae found a median prevalence of resistance of 14.3% (10.0–24.3%) in *E. coli* and 46.7% (17.3–84.5%) in *Klebsiella* spp. in Eastern Africa[84]. These varying estimates

highlight the need of a comprehensive study of the ESBL-producing Enterobacteriaceae prevalence in settings outside the hospital or healthcare facilities in East Africa[70].

1.3.3 ESBL-producing Enterobacteriaceae in Uganda and Malawi

In Uganda and Malawi, antibiotics are widely available, either through over-the-counter sale, as is allowed in Uganda[85], or through the lack of adherence to the prescription-only sale of antibiotics in Malawi[86]. Furthermore, in Uganda, approximately 20 million people do not have access to clean water and 37 million people do not have a decent toilet[87]. In Malawi, 33% of the population does not have access to improved sanitation and 13% do not have access to a clean water facility[88]. These conditions make Uganda and Malawi ideal places for AMR to transmit and persist. In recent years, prevalence of ESBL-producing Enterobacteriaceae has been estimated in multiple hospital studies in Uganda, ranging widely from 13.4% to 89%[89, 90, 91, 92]. Among the ESBL-producing Enterobacteriaceae, the prevalence of ESBL *E. coli* varied between 34% and 58.1% and the prevalence of ESBL *K. pneumoniae* varied between 12.7% to 72.7%[89, 90, 91, 92]. However, community surveillance of ESBL-producing Enterobacteriaceae is much less common in Uganda although ESBLs were detected in over 80% of the screened isolates in clients attending outpatient clinics in Kampala and rural districts of Uganda[93]. In Malawi, ESBL resistance in *E. coli* has increased from 0.7% to 30.3% and ESBL resistance in *K. pneumoniae* from 11.8% to 90.5% between 2003 and 2016[2]. Since then, community ESBL carriage has been estimated at 16.7% in a study focusing on community health centers outpatients[94]. The most common bacteria was *E. coli* with 66%, followed by *K. pneumoniae* with 8%[94]. Studies on the prevalence of ESBL-producing Enterobacteriaceae in Malawi are extremely limited, showing a lack of ESBL surveillance even more significant than in Uganda.

1.3.4 ESBL colonisation and risk factors

Many risk factors have been described for infections due to the acquisition of an ESBL-producing bacteria, both globally[95, 96, 97] and in sub-Saharan Africa[75, 98]. These risk factors have mostly been explored because of the infection the patients were suffering from and not because of their gut mucosal colonisation with ESBL-producing bacteria. However, recent studies have shown that gut mucosal colonisation with ESBL-producing bacteria is itself a risk factor of ESBL related infections[99, 100]. Therefore, interrupting the transmission of ESBL-producing bacteria by studying the risk factors of ESBL asymptomatic colonisation within the community seems like an interesting strategy to prevent infection[3]. Only four studies were identified in a global review from 2016 as investigating ESBL-producing Enterobacteriaceae fecal colonisation and risk factors among healthy individuals in Africa [4]. Among these, no studies described risk factors and none of the studies took place in East Africa. In a more recent review describing gut mucosal colonisation with ESBL-producing Enterobacteriaceae in sub-Saharan Africa, among twelve community studies, only three did not refer to a special population, and only one described risk factors[3]. This study reported 16.5% of ESBL carriage among healthy people in a community in Tanzania. Among the ESBL-producing bacteria, 76.3% were ESBL *E. coli* and among all the sampled isolates, 15.5% were ESBL *E. coli* and 3.8% were ESBL *K. pneumoniae*. Age, history of antibiotic use and history of admission were found to be risk factors of ESBL carriage [101]. Considering the One Health aspect of antimicrobial resistance and the recent global increase of ESBL-producing *E. coli* and *K. pneumoniae*, gaining a better understanding of the transmission of ESBL-producing Enterobacteriaceae is essential to slow their spread and prevent infections in low- and middle- income countries where WASH infrastructure is lacking and reserve antibiotics might not be available.

1.4 Drivers of resistance in Uganda and Malawi (DRUM)

The Drivers of resistance in Uganda and Malawi (DRUM) consortium is a trans-disciplinary collaboration, funded by the Medical Research Council, between UK universities and African institutes working across urban and rural sites in Uganda and Malawi. Its aim is to study AMR transmission in a One Health setting: areas with different human and animal population densities, and different levels of affluence and infrastructure[102]. The consortium wants to answer the following questions: What are the drivers of AMR transmission? How can we maximise the impact our efforts to interrupt human AMR acquisition are likely to have? Which strategies are likely to be most affordable and feasible? Its goal is to investigate the transmission patterns of ESBL-producing *E. coli* and *K. pneumoniae* by exploring demographic, WASH, behavioural, spatial and longitudinal risk factors for ESBL asymptomatic colonisation. It aims to look for the best course of action to prevent transmission leading to asymptomatic colonisation, which is itself a risk factor for ESBL infections. These bacteria were selected as they belong to the same family and often share AMR phenotypes, however *E. coli* is typically considered to be both community-acquired and nosocomial, whereas *K. pneumoniae* is more often judged to be the archetypal nosocomial AMR pathogens[73, 102]. Stool samples from the participants were initially cultured for growth of ESBL-producing bacteria. Bacterial colonies were then classified by color into categories, and speciation took place for ESBL-producing *K. pneumoniae* using polymerase chain reaction (PCR)[102]. Throughout this thesis, colonisation refers to asymptomatic carriage of either of these pathogens.

1.5 Aim and objectives of the thesis

The aim of this thesis is to investigate individual, household and WASH risk factors, across space and time, of ESBL-producing *E. coli* and *K. pneumoniae* transmission leading to human gut mucosal colonisation in Southern Malawi. The DRUM consortium provided an ideal framework to address this aim through three main objectives :

1. Determination of a suitable sampling design for all DRUM study areas, accounting for population density and socioeconomic stratification

2. Spatial investigation of household and individual risk factors of ESBL-producing *E. coli* and *K. pneumoniae* colonisation in the community

3. Temporal investigation of individual, household and WASH risk factors of ESBL-producing *E. coli* and *K. pneumoniae* colonisation in the community

Throughout this thesis, we group our candidate risk factors into different categories. Individual characteristics (age, gender, HIV status, antibiotic use) and household characteristics (household density, income, sharing the household with ESBL-colonised individuals) are explored in chapter 3 and 4. Older age has been found as a risk factor in multiple studies of ESBL colonisation [101, 103] as well as current or recent antibiotic use [104, 105]. While we focus on recent antibiotic use in this study, in Malawi, testing positive for HIV usually requires the patient to be put on cotrimoxazole preventive therapy thus looking at long-term antibiotic therapy. Furthermore, due to the need for more frequent healthcare utilisation, HIV-positive individuals are at increased risk of infections with resistant bacteria[106]. In low- and middle income countries, income is complex as it can influence many other covariates, such as the WASH infrastructure and/or antibiotics available to the household. In previous studies in similar contexts, household income has been shown as a risk factor for ESBL colonisation however opposite findings resulted from those studies [107, 108]. In order to look at the importance of within-household transmission, which has been assessed in other contexts [109, 110, 111], we also look at the household density, defined as the number of people living in the household at time of enrolment, and whether sharing the household with ESBL-colonised individuals increases your risk of being colonised with ESBL-producing bacteria. Additionally, we look at seasonality to see if it impacts the prevalence of ESBL colonisation in the community. In Malawi, the wet season spans from November to April and the dry season from May to October. During the wet season, accumulation of rainwater and flooding can occur, potentially increasing the risk of contamination of the environment. Such weather might also affect the behaviour of the population causing more indoor crowding and therefore potentially increasing the transmission within the household. Additionally, the seasonality of infectious diseases causing increased levels of antibiotic use might

also result in an association between seasonality and higher AMR levels [112].

In low- and middle-income countries, the lack of appropriate wash infrastructure combined with poor hygiene practices facilitates interactions between people and both human and animal waste in the environment. Water, sanitation and hygiene factors are thought to play a main role in the transmission of AMR [32] therefore in chapter 4, we include WASH risk factors to look at their effect on human colonisation with ESBL bacteria. These risk factors can be categorised in various groups: sanitation factors (toilet present in the household, type of toilet, presence of a drop hole cover, presence of cleaning materials in the toilet, reported open human defecation, available disposal mechanism for animal waste), hygiene factors (hand washing facilities, presence of soap), food factors (eating street food, eating from shared plates), water factors (drinking water source, storage method, alternative water used for cleaning utensils), animal factors (animal ownership, animal inside the house, visible animal faeces) and broader environmental factors (accumulation of wastewater, interaction with river water, interaction with drains).

Poor sanitation infrastructure has been associated with higher levels of antimicrobial resistance [113] while the use of improved infrastructure like drop hole covers is typically used to prevent flies from accessing faecal matter, reducing the ability of flies in transporting and transmitting bacteria [114, 115]. Water factors such as private access to drinking water and using water from a borehole have been associated with a lower prevalence of ESBL colonisation [3, 116]. In fact, environmental exposure can occur through different types of contaminated water sources, such as drinking water, recreational water or irrigation water [34, 37, 38], therefore it is crucial to look at the effect of various water sources (drinking water, wastewater, river water) on ESBL colonisation. Furthermore, in LMICs, animal husbandry is frequently a primary source of income [117], and permitting animals inside the home is common practice. However, this practice increases the risk of faecal contamination of the soil by enteric pathogens like *E. coli* [118] and therefore puts household members, especially young children, at higher risk for exposure to faecal pathogens and enteric infections [119]. Thus it is essential to consider animal factors when looking at colonisation with enteric bacteria. Overall, all the risk factors categorised under WASH risk factors were included in order to detect any association arising from these human-animal-environment interactions.

1.6 Methods overview

This section outlines the methodological approaches taken throughout this thesis. The information in this section is primarily from Diggle and Ribeiro[120] and Diggle and Giorgi[121]. For more detailed information, please refer to these books[120, 121].

1.6.1 Bayesian geostatistical methods for disease data

To answer their questions, the DRUM study proposed to run a detailed, longitudinal microbiological survey of humans, animals and the environment in order to discover social, demographic, and spatial predictors of AMR carriage. Due to the interactions between human, animal and environment surrounding AMR, spatial variation in prevalence of AMR is complex. Many factors could be responsible for such variations such as antimicrobial consumption levels, level of environmental contamination, household transmission and socio-economic factors[122]. Studies have shown that the prevalence of AMR varies depending on the global region [123, 124] and that differences in antimicrobial consumption have been linked to the emergence of spatial clusters of AMR in the community [125]. However, little is known about more localised spatial clustering of AMR in LMICs. The presence of ESBL-producing Enterobacteriaceae clusters at the community level using geographic mapping tools was also detected in previous studies [126, 127]. Spatial studies of AMR genes in animal populations have also detected small-scale clusters [128, 129]. In previous studies of other bacterial diseases[130, 131, 132], significant spatial clustering of prevalence was found across similar spatial extents to those used in DRUM.Generally, the transmission of infectious diseases tends to occur over short distances, therefore their spatial structure usually acts in three main ways: individuals far from sources of infection are at little risk, local transmission and depletion of susceptible hosts impact majorly the epidemic growth and at-risk individuals can be targeted using spatial proximity as a local control measure[133]. Although SARS-CoV-2 is a viral disease, the principles of infections by contact or proximity are similar; several authors have spatial clustering of cases as a result[134, 135, 136]. In fact, disease outcomes often display some spatial structure, with neighbouring values being correlated due to shared characteristics and transmission, which infers that information from one site can pro-

vide information about neighbouring sites[137]. Thus an analytic approach capable of detecting, and accounting for, residual spatial correlation is required in order to ensure accurate inferences to be drawn. In such a setting, geostatistical models present a natural and powerful method of analysis[121].

A geostatistical model assumes a generalised linear modelling framework in which a response variable Y_i, $i = 1,\ldots,n$ is observed at a discrete set of sampling locations x_i, $i = 1,\ldots,n$.

Y_i is modelled using a noise distribution conditional on mean μ_i such that

$$g(\mu_i) = \alpha + z_i^T \beta + S(x_i) \tag{1.1}$$

where $g(\mu_i)$ is a link function that relates the mean to a linear combination of variables, z_i is a vector of covariates, and $S(x_i)$ is a realisation of a spatial Gaussian process with covariance function

$$\Sigma_{ij} = \sigma^2 \rho(||x_i - x_j||;\phi) + \tau^2 \mathbf{1}_{\{i=j\}} \tag{1.2}$$

for locations x_i and x_j. ϕ is the length scale of the spatial correlation, σ^2 is the variance of the spatial process, τ is the nugget effect and $||x_i - x_j||$ is the Euclidean distance between x_i and x_j.

1.6.2 Bayesian inference and MCMC methods

Whilst much of the statistical analysis in this thesis is performed in a frequentist approach, we employ Bayesian methods at several points to work with more flexible probability models. In Bayesian inference, the model parameters are treated as an unobserved random variable and must be assigned a probability distribution, called the prior distribution $p(\theta)$. The prior distribution represents the uncertainty about the parameter before data collection. Inference about the parameter is carried out through the distribution of the parameter given the data, called the posterior distribution $p(\theta|Y)$[121]. They can be described using Bayes' Theorem:

$$p(\theta|Y) = \frac{p(Y|\theta)p(\theta)}{p(Y)} \tag{1.3}$$

where θ are the model parameters and Y the data. The likelihood $p(Y|\theta)$ represents the probability of the data give the parameters. Inference is usually performed ignoring the normalizing constant $p(Y)$, thus the equation can be rewritten as:

$$p(\theta|Y) \propto p(Y|\theta)p(\theta) \tag{1.4}$$

where $\propto$ means "is proportional to". Because the posterior is a joint distribution over multiple parameters, it can be hard to interpret and is high-dimensional. Markov Chain Monte Carlo (MCMC) methods are used to sample from probability distributions using Markov Chains[120]. We used a standard implementation of the No-U-Turn Sampler (NUTS), an adaptation to the Hamiltonian Monte Carlo algorithm, for fitting the Bayesian posteriors. For an introduction to MCMC algorithms, please see Gilks, W.R., Richardson, S. and Spiegelhalter, D. (1995)[138].

1.6.3 Spatial data analysis

A spatial study has to take in an added consideration compared to other types of studies, which is to focus on where to sample the locations, as well as how many locations to sample from. Contrarily to the number of locations to sample from, the sampling locations are usually not constrained by cost since it is assumed that the researcher is free to go to any location within their study area[120]. However, that is not always the case. In prevalence studies, sampling locations have to be restricted to inhabited locations within the study area, in order to sample from the population. In the case of the DRUM study, we also had an a priori expectation that households close to each other might have more similar ESBL colonisation status than households further apart. For geostatistical models such as the model described in Equation 1.1, an efficient sampling design must consider not only maximising variability in terms of the covariates z, but also ensure that variability exists between sampling locations x so as to inform on the spatial parameters σ^2, ϕ, and τ^2 [139]. The design concept is developed in Chapter 2.

1.6.4 Bayesian models for longitudinal disease data

For disease-related states such as asymptomatic colonisation with ESBL *E. coli* and *K. pneumoniae* that are known to be a risk factor for ESBL infections[99, 100], there is an interest in finding out what the temporal trends are, not only seasonally over time, but also for specific individuals or households. Gaining a better understanding of how the colonisation evolves at the individual-level or household-level could offer some additional insights into the transmission patterns of ESBL *E. coli* and *K. pneumoniae*. The DRUM study initially planned to compare visits that were fixed at baseline, 1 month, 3 months and 6 months but due to the very large variation in dates for each specific visit group (1,2,3,4) and in days between visits (caused by the COVID-19 pandemic), it was no longer possible to compare these specific groups. For example, when comparing between individuals and/or households, some had a month between 2 specific visit groups while others had four months between the same two visit groups. In order to account for the gap in samples created by the pandemic and to appropriately look at the evolution of colonisation status over time, we decided to look at time as a continuous variable instead of using the visit groups. We also use a personalised covariance structure for our households while keeping the model at the individual-level. In order to fit the household-level temporal random effect, we made the pragmatic decision to move to a Bayesian framework that offers more flexibility over how we design our probability model. In fact, at various points in this thesis, we moved from a frequentist framework to a Bayesian framework to fit models that would have been difficult to fit otherwise, i.e. the Gaussian process in chapter 3 and the temporal model in chapter 4.

1.7 Impact of COVID-19 on the thesis

The COVID-19 pandemic impacted the DRUM study in various ways, consequently impacting this thesis. It mainly affected it by stopping the data collection and the microbiological testing, therefore delaying the results for a couple of months. Considering the DRUM study is a longitudinal study, all visit groups were affected by this. Additionally, a substantial amount of samples already sampled were also unavailable for analysis due to the scientists inability to go into work and perform the microbiological analysis be-

cause of restrictions. The effect on results will be mentioned throughout the chapters and resumed in the discussion.

Chapter 2

Spatial design for the DRUM study areas

2.1 Introduction

Antimicrobial resistance is deemed one of the biggest global health threats facing humanity by the World Health Organisation[9]. Specifically, Enterobacteriaceae producing extended-spectrum β-lactamases (ESBL) have become a large problem in African healthcare institutions and communities in the last decades. In sub-Saharan countries, where AMR research is still scarce[70], gaining a better understanding of the spatial distribution of ESBL-producing Enterobacteriaceae carriage is essential in order to investigate the dynamics of transmission of these pathogens and to tailor appropriate interventions to reduce their dissemination[5].

When investigating phenomena characterized by spatial variation, finding optimal sample locations in the study area is essential[140, 141, 142]. These sample locations need to be as representative as possible of the entire population, taking into account the spatial variability of the area [143]. In the context of disease, under sampling in some areas could prevent the spatial variability from being captured, therefore not correctly portraying the prevalence of the disease. While acquiring information at every possible location is necessary in terms of representation, avoiding oversampling of areas with low

density population, such as rural areas where households are inexistent or more spaced out, is crucial in terms of efficiency.

Common issues in spatial sampling design relate to a lack of information regarding the geographical structure of the study area. If no information about the area is available and the sampling is done at random, the likelihood of sampling in low-density areas increases, therefore complicating the field data collection, e.g. field teams arriving at a location at which nobody lives. Moreover, this can lead to a potentially important cluster of households being ignored, causing a misrepresentation of the population in that area. The choice of design can heavily impact the resulting sampling and subsequent results in many ways; if the data used to construct the design is not precise, some households will not be selected if they do not appear in the data and some might be misplaced due to infrastructure changes or a lack of recent information. Furthermore, typically some households that have been pre-selected will decline to take part, so in-field decisions about how to approach the next nearest household must be taken, which may affect the results and should be considered in advance.

Two main classes of geostatistical designs exist: adaptive and non-adaptive designs. The first requires the sampling locations to be chosen sequentially during data collection whilst the latter requires the sampling locations to be chosen in advance of any data collection[139]. Henceforth, the latter will be discussed. Two examples of non-adaptive designs are completely random and lattice designs. In a completely random design, the design points form an independent random sample from a uniform distribution on the study area[120]. In a completely regular design, the design points form a regular lattice (usually square, at times triangular) over the study area. In theory, both these designs are used for classical statistical sampling, with a preference towards random designs to protect against bias. From a spatial point of view, these designs can be inefficient due to their uneven coverage of the study area. Contrarily, regular designs offer a very even coverage of the study area[121].

Generally, when constructing a sampling design, an important condition is to compromise between efficient parameter estimation and efficient prediction given the values of the covariance model parameters (see Chapter 1, Equation 1.2)[139, 144, 145, 146]. Lattice designs lead to efficient spatial prediction provided model parameters are

known, whilst a completely random design has the advantage if the model parameters are unknown, due to the assumption of a wider range in distances between points and consequently that more information will be provided on the parameters of the covariance function[139, 147]. Whereas regular designs are susceptible to bias in the presence of periodicity across space, more-regular-than-random designs can improve spatial prediction, while maximising coverage across the study region[148]. Therefore, random sampling of points conditional on inhibitory distance compromises between a regular grid and complete spatial randomness to ensure points are randomly located, but distributed evenly across the study region[139, 147]. Inhibitory designs guarantee a minimum distance m between points, meaning that short-range spatial correlation cannot be measured. Therefore, additional points may be introduced at random within a circle of small radius close to randomly selected primary points – "close-pair points".

In the context of non-adaptive designs, random sampling has been shown to be sufficiently efficient for parameter estimation [121]. However, when looking at prediction over a new set of points throughout a specific study region, it can create large empty spaces with no samples within said study region [145]. Diggle and Ribeiro did a simulation study contrasting a regular lattice design with a completely random design, each with 100 location points. They considered three design criteria: the spatial maximum of mean square prediction error, the spatial average of mean square prediction error and the scaled mean square error. They generated data from replicated simulations of a stationary Gaussian process with exponential correlation function and varied the correlation function parameter and the noise-to-signal variance ratio in order to evaluate the design criteria. They have shown that when comparing random and regular designs, the regular design consistently outperformed the random design [120]. Similarly, Diggle and Giorgi carried out a simulation to look at the performance of three different designs applied to realisations of two stationary Gaussian processes. They estimated the average squared prediction errors for each design and each model and found that inhibitory designs, both with and without the addition of close pairs, outperformed random designs in terms of predictive performance. Additionally they have shown that when a nugget effect is included in the model, inhibitory design with close pairs outperform a simple inhibitory design [121]. Diggle and Lophaven looked at comparing the performance of a simple regular lattice design when integrated with close pairs or in-fill. Fixing the total number

of locations at 64, they compared an 8×8 lattice design with a 7×7 lattice design with 15 added close pairs located uniformly at random within a disc whose centre is randomly selected within the lattice and a 7×7 lattice design with three added 3x3 in-fill. Using a linear Gaussian model with constant mean and an exponential function, they averaged the design criterion over five independent replicates due to the randomness of the secondary locations. They varied the covariance function parameter and the noise-to-signal variance ratio and computed the design criterion, the average prediction variance, for each using direct simulation of 1000 independent draws from the posterior. They found that when varying the covariance function parameter and the noise-to-signal variance ratio, a lattice plus close pairs design consistently outperformed a regular and a lattice plus in-fill design [147]. Chipeta and Diggle performed a simulation study using a linear Gaussian model including a Gaussian process with Matérn correlation. They fixed the value of the correlation shape parameter but varied the correlation range parameter as well as the nugget variance. Then they approximated the criterion by simulating data at 150 sampling locations and evaluating it using 1500 independent simulations of measurement data. They showed that, when the nugget effect is non-negligible, designs with a ratio of 10 to 30% of close pairs have the best predictive performance with consistent lower average prediction variances[139]. Furthermore, previous studies have been consistent in showing that in order to compromise between prediction accuracy and efficient parameter estimation, optimal geostatistical designs should include a small proportion of close pairs (between 10% and 30%) in an otherwise spatially regular design[147, 149, 150].

In the context of DRUM, study areas could have been selected by focusing on homogeneous areas where participants already took part in similar studies. However, this could have created a selection bias in our study. By selecting study areas with varying population densities and accessibility, we avoided this bias but included, in return, an added difficulty in sampling in rural areas where no household is present. Due to this difference in population density within/between the study areas and the difficulty in accessing some of the more rural areas, complete spatial randomness could not work for the DRUM study areas. We suggest that using geographical information or pre-existing location data can improve the location accuracy of the households, therefore increasing the time efficiency of the field teams. Therefore, we propose to address this issue by using different types of geographical information: pre-existing census data, population density

rasters and OpenStreetMap data.

The aim of this chapter is to extend the spatial inhibitory design with close pairs principle[139] to allow sampling within sites with spatially heterogeneous populations. This addresses the practical problem of efficiently identifying households on the ground, reducing the number of locations sampled where in fact no household exists. In fact, we wished to address the practical problem of efficiently identifying households on the ground, by implementing different versions of the design algorithm depending on the level of a priori knowledge.

2.2 Data

Prior to starting the DRUM study, the consortium was established and consortium members identified study sites representing urban, peri-urban and rural settings, in order to enable variations in WASH behaviours, animal practices, antimicrobial usage and ESBL-producing bacterial contamination. Sites were identified based on acceptability of research to the communities through long term public involvement, site safety profiling, logistical constraints and existing research capacity[102]. In total, five sites were selected: three in the Southern Region of Malawi and two in Uganda. Their general location can be seen in Figure 2.1. These five sites can be observed in Figures 2.5, 2.6, 2.7, 2.8 and 2.9.

2.2.1 Malawi

The Malawian sites are classified as follows:

- Chileka (peri-urban)

  Chileka is an area of Malawi situated 10-15 km north of Blantyre. Chileka, with an area of 14 km^2, is classified as peri-urban. It was included because of the need for a peri-urban region in Blantyre and due to local prior knowledge and land topology.

- Ndirande (urban)

  Ndirande, a high-density informal settlement, is located three kilometres from the Central Business District of Blantyre, the second city and commercial capital of

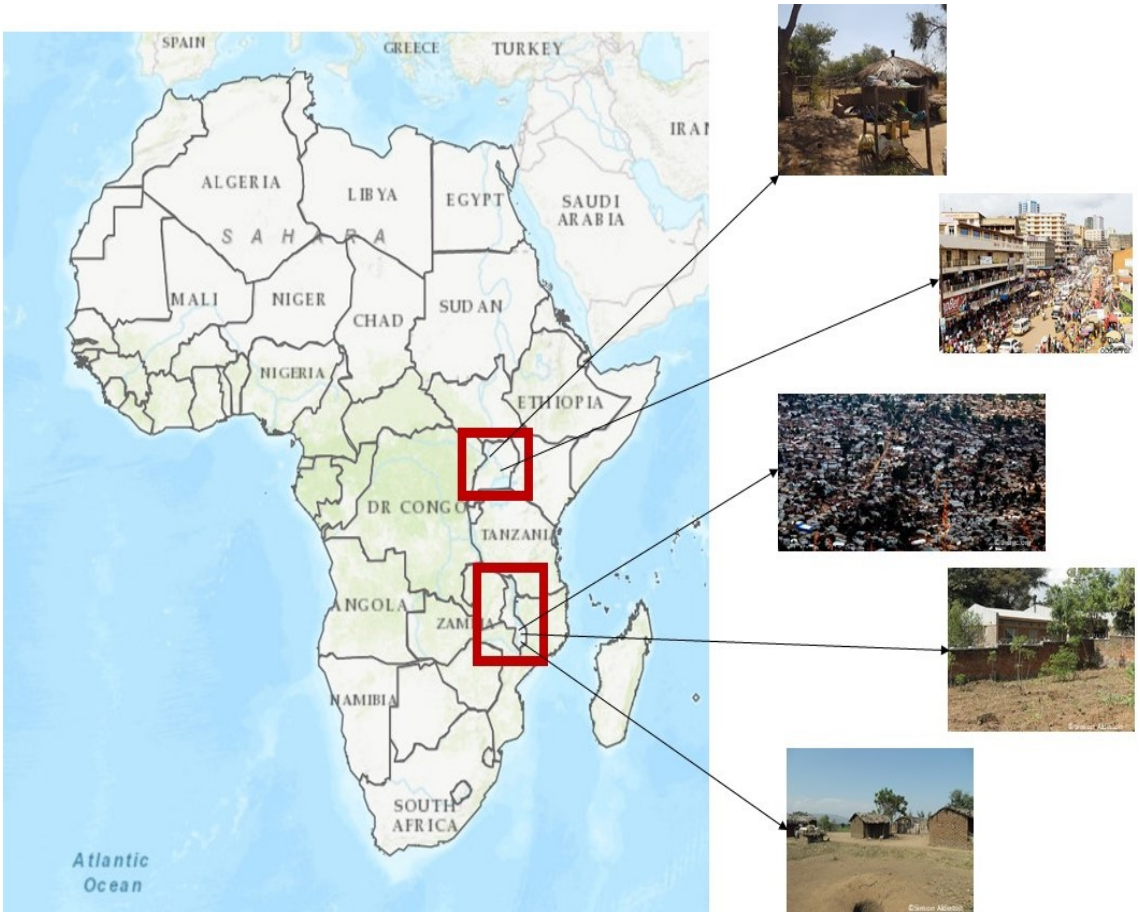


Figure 2.1: Sites (top to bottom): Hoima, Kampala, Ndirande, Chileka and Chikwawa

Malawi. Ndirande is the smallest study area in Malawi with an area of 3 km^2 and is classified as the urban study area due to its proximity to the city and its dense population. Ndirande was chosen as a pre-existing study area because of the possibility to pull in data from other previous studies and because of the ease of access to inhabitants used to participating in studies.

- Chikwawa (rural)

  The study area classified as rural in Malawi, with an area of 71 km^2, is part of the Chikwawa District. Chikwawa is also the name of the main town and administrative capital of the district with a population of approximately twelve thousand people. Chikwawa lies almost thirty miles south of Blantyre. Similarly to Ndirande, Chikwawa was chosen as a pre-existing study area following the same strategy.

2.2.2 Uganda

The Ugandan sites are classified as follows:

- Kampala (urban)

  Kampala is the capital and largest city of Uganda. Defined as the urban study area in Uganda, it is divided into three separate but consecutive polygons drawn in wedge shape and ranging from 1 to 3.4 km^2. Because of a much smaller sample size for the microbiological sampling, socioeconomic status stratification is applied to classify these three polygons into medium/high/low socioeconomic status determined by local prior knowledge. The smallest polygon closest to the center is classified as low, whilst the one further away from the center is classified as medium and finally the middle one as high socioeconomic status.

- Hoima (peri-urban and rural)

  The area classified as peri-urban and rural is Hoima, with two non-consecutive polygons of 3.6km^2 and 7.6km^2. Hoima is the main municipal, administrative, and commercial center of Hoima District, in the Western Region of Uganda. Hoima is approximately 200 kilometres northwest of Kampala, Uganda's capital. The two polygons in Hoima were classified as "peri-urban" and "rural" areas due to local prior knowledge such as the awareness that people are less likely to keep animals in an urban setting. The "peri-urban" polygon is situated mostly within Hoima city while the "rural" polygon is situated outside of Hoima city but still within Hoima district.

  In Uganda, both Kampala and Hoima were decided thanks to local prior knowledge in order to achieve a gradient across socioeconomic status and urban/rural composition.

In rural areas such as Chikwawa, due to the varying population density, using spatial tools such as WorldPop population rasters allowed us to preferentially propose sampling sites in areas of high population density. However, in other locations such as Hoima and Ndirande, the resolution of these population rasters was low, therefore data sources such as OpenStreetMap and research-collected census data provided spatial point information on building locations to inform our sampling. In Kampala, where the population density

was high and homogeneous, a stratified spatial design was required, stratifying by socioeconomic status. Since the polygons were contiguous, this created a restriction on the number of households in each polygon that can be dealt with using a rejection sampling, rejecting the point if it lies outside the chosen polygon or if the polygon already has the required number of points.

2.3 Methods

2.3.1 Sampling strategy

To study the transmission of AMR in the Malawian and Ugandan study areas chosen for DRUM, our sampling strategy took the household as the spatial sample unit. Random households needed to be selected within urban, peri-urban and rural settings in Uganda and Malawi. This allowed the different field teams to capture the microbiological, antimicrobial usage/consumption and WASH surveillance data necessary to inform the model, and the economics of antimicrobial resistance. In order to determine a suitable spatial sampling design for all study sites in Uganda and Malawi, the following design was developed. The methodology is based on the spatial inhibitory design with close pairs (ICP)[139] but was modified to allow sampling within sites with spatially heterogeneous populations, and to allow us to stratify our population by socioeconomic status.

2.3.2 Spatial inhibitory design with close pairs

Inhibitory designs are random designs that generate spatially regular configurations of design points.[139] Chipeta and Diggle proposed two specific classes of inhibitory design: inhibitory designs and inhibitory designs with close pairs. An example can be found in Figure 2.2.

In a simple inhibitory design with n points to be sampled, the distance between each point must be at least a prearranged inhibition distance d. In an inhibitory plus close pairs design with n points to be sampled, only $n-k$ points are sampled using

the inhibitory design. k points are then sampled using the close pairs technique, which consist of randomly selecting k points among the $n-k$ inhibitory points just sampled and place a new point for each of them using a uniform distribution within a disc of radius ζ. Each of these k close pairs are then located close to an inhibitory point previously sampled. This allows us to improve predictions and take into account a more localised variation, mostly useful for the microbiological sampling.

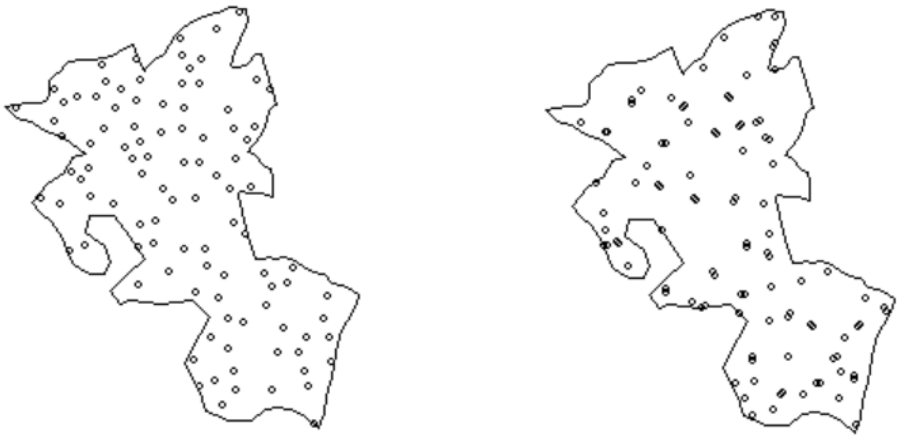


Figure 2.2: Example of inhibitory designs in Chileka (left: simple inhibitory design / right: inhibitory design with close pairs)

2.3.3 The `geosample` package

In order to construct such design efficiently, we needed to use a software that has shown to rapidly implement similar designs. Among existing available designs, some can be easily implemented using R packages. Packages such as `sp`,`sf` and `BiodiversityR` all include functions to perform spatial sampling. However, they only focus on random or regular sampling methods while other packages such as `spcosa`, `SamplingStrata` and `GridSample` focus on stratified sampling designs.

The `geosample` package[151] is an R Package used to construct geostatistical sampling designs. It allows the user to determine sampling locations within a set of spatial constraints and information from existing sampling locations. The package focuses on geostatistical sampling designs that compromise between efficient parameter estima-

tion and efficient prediction given the values of model parameters. Figure 2.3 gives a visual representation of the package workflow.

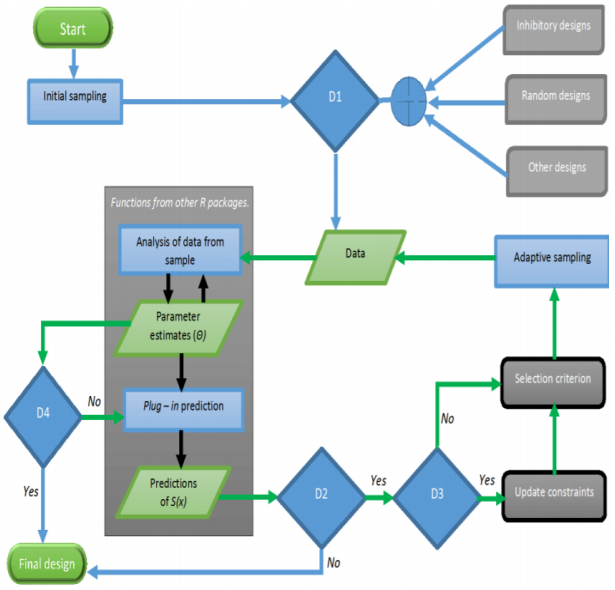


Figure 2.3: Geostatistical sampling workflow within geosample package. D1: user decision for initial design. D2: user decision whether to sample additional samples, in which case adaptive sample will be generated. D3: user decision to update sampling constraints. D4: user decision to stop further sampling[151].

In the `geosample` package, inhibitory designs for a finite set of points are implemented by the function `discrete.inhibit.sample`, and for points in a continuum, by the function `contin.inhibit.sample`[151]. In each of these implementations, the `geosample` package can generate simple inhibitory or inhibitory-with-close-pairs samples. The package uses novel and computationally efficient algorithms for constructing adaptive and non-adaptive geostatistical designs, including traditional random sampling. It also provides automatic visualisation of the results by plotting the sampled locations[151]. The `osmgeosample` package[152], which builds on `geosample`, is a recent R Package used to create spatially continuous or discrete random samples from a predefined spatial border using OSM data.

However, these packages do not allow the inclusion of supplementary geographical information into the design (except for OpenStreetMap data) and therefore, do not work well for heterogeneous populations. Our code allows for a more inclusive

design and could be a good addition to the `geosample` package.

2.3.4 Extended spatial inhibitory design with close pairs: the algorithm

The general algorithm takes in as input: the total number of points, the number of close pairs, the inhibitory distance, the radius for close pairs, the polygon and the sampling proposal ("density","census"..). It returns a set of locations points.

The general construction of our ICP design is as follows:

1. Calculate the bounding box of the polygon P (minimum and maximum for coordinates x and y)
2. Sample initial point and set $i = 1$
3. Draw $x_i \sim G$, a proposal distribution chosen by the user
4. If $x_i \in P$, go to step 5 else go to step 2
5. Calculate the minimum d of the distances from x_i to all other x_j in the current sample;
6. If $d \geqslant m$, store x_i, increase i by 1 and return to step 3 if $i < n - k$, otherwise stop;
7. If $d < m$, return to step 3.

To obtain k close pairs, we follow these steps for $n - k$ points and then proceed as follows :

1. Sample k from $x_i, \ldots, x_{n-k}$ without replacement and call them x_j, $j = 1, \ldots, k$
2. For $j = 1, \ldots, k$, x_{n-k+j} is uniformly distributed on a disk with center x_j and radius ζ
3. If $x_{n-k+j} \in P$, go to step 4 else go to step 2
4. Store x_{n+k-j}, increase j by 1 and return to step 2

In the case of DRUM, these sampling locations were then turned into actual households by the field teams. If no household was present at the location, or if a household refused to take part in the study, a random direction was chosen by the field team, and the closest compliant household was selected.

The initial set distance m chosen to be the minimum distance between each inhibitory point was 100 meters and the radius for each close pair was 30 meters. These values were chosen because of recent work done on Typhoidal Salmonella that showed a spatial correlation up to approximately 150 meters[131]. The ratio for the number of inhibitory points and the number of close pairs in our design for all areas was 70% to 30% due to the general convention because of a lack of information about the expected correlation[121].

2.3.5 The implementation

The design was written using different functions and implementations described below. A UML class diagram showing the relationship between the different sampling options and the general algorithm can be seen in Figure 2.4.

- `samplePoint`

  The `samplePoint` function takes in a polygon as argument and samples a new point depending on the information available on the polygon. This point is then turned into a spatial point using the `st_point` function from the `sf` package. Using a rejection sampling, if the resulting point is within the polygon, then the point is stored and, if not, it is rejected and the function runs again until a point is accepted. The function returns the new spatial point.

  1. If no information is available on the households: Determine its bounding box (box with the smallest area within which all the points lie). Then coordinates for a spatial point are sampled using a uniform distribution on the minimum and maximum coordinates of the box.
  2. If information is available through a census or OpenStreetMap data : Sample from the existing data using a simple sample function over the rows of the

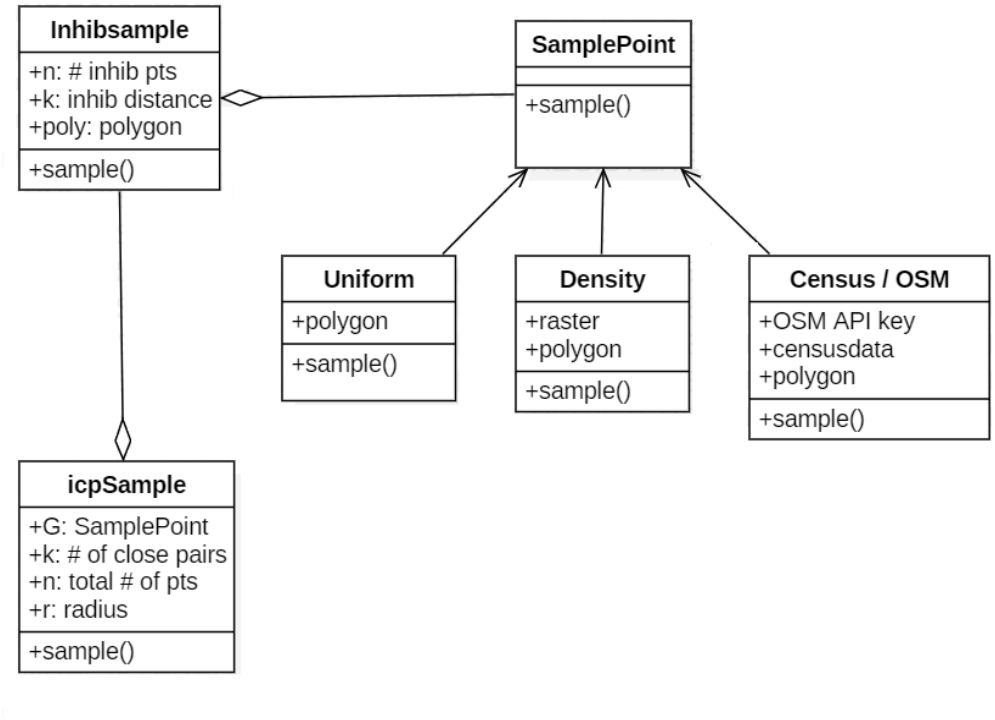


Figure 2.4: UML class diagram on the relationship between the sampling modules and the general algorithm

data frame containing the GPS coordinates of the existing households.

3. If the population density is too heterogeneous: A population raster is used to allow for a population density-weighted sampling. The function first samples a pixel from the population raster with probability proportional to pixel value with probability κ, and complete spatial randomness with probability $1 - \kappa$. It then draws a point within the pixel with complete spatial randomness.

- `minDistance`

  The `minDistance` function takes in as arguments a new spatial point and a matrix of already determined spatial points. The minimum distance is calculated between all spatial points. The function returns the minimum distance.

- `inhibSample`

  The `inhibSample` function takes in as arguments a number of points n, a set distance k and a polygon. This function is used to sample inhibitory points and return them.

1. If the polygon is not split into contiguous polygons: Initially, a first point is sampled using the appropriate version of the `samplePoint` function. For each new point sampled using the same function, a rejection sampling is used. If the minimum distance between each spatial point is greater than k, the point is stored, and, if not, the function keeps iterating until n points are sampled.
2. If the polygon is split into contiguous polygons: Before implementing the `inhibSample` function, a new function called `whichpol` is created, the purpose of which is to determine which consecutive polygon a point belongs to. A first point is sampled using the appropriate version of the `samplePoint` function. For this initial point and every new point, the `whichpol` function is used to determine which polygon the point belongs to. A count of the number of points in each polygon is kept to restrict each polygon to an equal number of inhibitory points. For each new point sampled using the `inhibSample` function, a rejection sampling is used. If the minimum distance between each spatial point is greater than k and the number of inhibitory points in that polygon is less than our chosen value, the point is stored and the number of points in that polygon gets updated, and, if not, the function keeps iterating until n points are sampled.

- `samplePtInRadius`

  The `samplePtInRadius` function takes in as arguments a spatial point and a radius ζ. An angle θ is determined randomly using a uniform distribution. Then a new point is created from the spatial point adding to each coordinate a constant determined by the radius ζ and the cosinus and sinus of the angle θ. Each new point falls under the restriction that it has to belong to the polygon. This function is used to sample the close pairs and return the close pair for the point entered as argument.

- `icpSample`

  The `icpSample` function takes in as arguments the number of total points needed, the number of close pairs k, the minimum set distance between each inhibitory point, the radius for the close pairs and the polygon. Using the previous functions, this function is the function that allows to create all inhibitory points and close pairs.

1. If the polygon is not split into contiguous polygons: Using the appropriate version of the `inhibSample` function, the inhibitory points are sampled. Then a sample of size k is taken from these inhibitory points and the `samplePtinRadius` function is applied to each point to create each close pair. The rejection sampling restricts all close pairs to fall into the polygon, else the point is rejected. The function returns a dataframe with the inhibitory points and close pairs.
2. If the polygon is split into contiguous polygons: Using the appropriate version of the `inhibSample` function, the inhibitory points are sampled. We know each point belongs to a particular polygon. A sample of size k divided by the number of consecutive polygons is taken from the inhibitory points belonging to each consecutive polygon and the `samplePtinRadius` function is applied to each point to create a new close pair. The rejection sampling restricts all close pairs to fall into the polygon, else the point is rejected. The function returns a dataframe with the inhibitory points and close pairs.

The code for implementing the design can be found in Appendix A.1.

2.3.6 Implementational decisions and issues

The `Simple Features for R (sf)` package is a recent package whose purpose is to be much more convenient and flexible than the pre-existing `Classes and Methods for Spatial Data (sp)` package. The latter was the first general package to provide classes and methods for spatial data types to create and work with geometries such as points, lines, polygons in the early 2000s[153]. Both packages allow the representation of spatial data in R by defining their own classes of objects to store spatial data.

`Sf` objects consist of rows of features which have both non-spatial and spatial data, therefore can be treated as data frames also containing spatial data. The spatial part of an `sf` object is contained in a geometry column of class `sfc`. This column contains the common types of spatial data: the CRS (coordinate reference system), coordinates, and type of geometric object. The `sfc` class has different subclasses to denote them. The possible geometric objects are `point, linestring, polygon, multipoint,`

multilinestring, multipolygon and geometrycollection for any combination of the other types.

Two different ways exist in the sf package to identify a specific coordinate reference system (CRS) and to transform objects from one CRS to another: the PROJ library and EPSG codes. The EPSG library supplies codes for well-known CRSs, and thus provides an easier method to identify a subset of the CRSs available through the PROJ library. A data frame of over 5,000 EPSG codes is available in R through the rgdal package. The most widely used EPSG code is 4326, which is the geographic reference system that uses units of longitude and latitude on the World Geodetic System 1984 (WGS84) ellipsoid[154].

It should be noted that the sf package does not allow for any integration with many other packages, including raster (as of March 2018)[155]. Thus, the sp package is used to work with both vector and raster data for this study area.

The sf package has almost all of the capabilities of sp, but it uses objects that are easier to work with than sp objects. Moreover, sf allows an easier visualisation of the contents of an object, which makes it much easier to get an overview of the data that you are working with.

The recent development of the sf package has modernized the implementation of spatial data in R and made it possible to integrate spatial data into the tidyverse and ggplot2 plotting system. sf has made it easier to work with spatial data in R by minimizing the distinction between spatial data and other forms of data[154].

Contrary to sp, sf objects don't change class when you apply spatial operations to them (and they keep their associated data). Moreover, spatial indexing allows us to massively speed up spatial queries, like intersecting polygons, especially on large datasets.

The rgdal and tmap packages were used to visualise the results on R. The tmap leaflet and google earth were used to show the proposed geolocations on the map.

2.4 Results

Having outlined our methods for creating ICP designs, we now describe how this was used in each of our study areas. The study was designed to centre on the household level. We initially aimed to enrol a total of 562 households, with 262 recruited in Uganda and 300 recruited in Malawi. In order to collect longitudinal microbiological data and WASH data, the households were followed up three to four times over the course of the study. Microbiological sampling took place to determine the presence of ESBLs from various samples and WASH behaviors such as toilet use, washing practices and water supply were explored using the Risks, Attitudes, Norms, Abilities and Self Regulation (RANAS) model[156, 157].

2.4.1 Malawi - Chileka

In each Malawian study area, 100 households were sampled. Considering the ratio of inhibitory points (households) to close pairs for our design is 70% to 30%[121] and the number of households that needed to be sampled was a hundred, we expected to sample 70 inhibitory points and 30 close pairs. Chileka is the only area where we increased the minimum distance between each inhibitory point to 200 meters and the radius for the close pairs to 50 meters due to the population within the village being sparser than in the other areas and causing difficulties in finding close pairs.

Since visual inspection of OpenStreetMap building location data plotted over satellite imagery of Chileka showed that the former was incomplete, we decided to sample at random without restrictions except being inside the Chileka polygon. The sampling design for Chileka was the simplest one of all the study areas as it required no supplementary restrictions. The results of the sampling on Figure 2.5 showed the points evenly spread out across the polygon, which confirmed the random aspect of the algorithm combined with a minimum distance of two hundred meters between each inhibitory point.

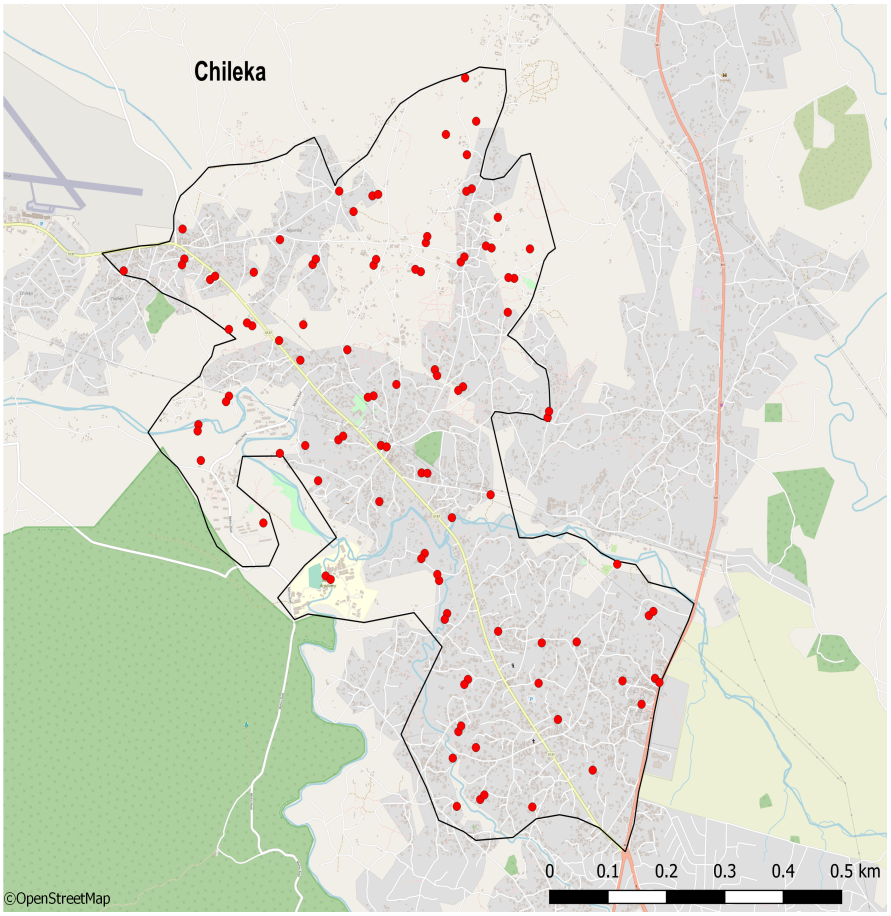


Figure 2.5: Resulting location points for Chileka

2.4.2 Malawi - Ndirande

For Ndirande, the Strategic Typhoid alliance across Africa and Asia (STRATAA) census gave us access to existing households that took part in that study over two different censuses. The census took part twice over 2 years in order to accurately calculate an incidence rate of typhoid fever for the area[158]. Therefore we conditioned our sampling locations on these existing households. Households with missing gps coordinates were removed. The algorithm was modified to allow sampling from the existing coordinates instead of random sampling, but keeping the restriction of belonging to the polygon.

The sampling results in Figure 2.6 also showed an even distribution of the points within the polygon that can be explained by the absence of restrictions in the algorithm for that area as in Chileka. Even using existing households to sample from, it was difficult to see on the map which household is chosen due to the really high density of population. We

note that using pre-existing data increased the risk of introducing a systematic bias in the analysis because of the aging of the dataset. However, the STRATAA study was carried out recently and therefore we assumed that this uncertainty remained low.

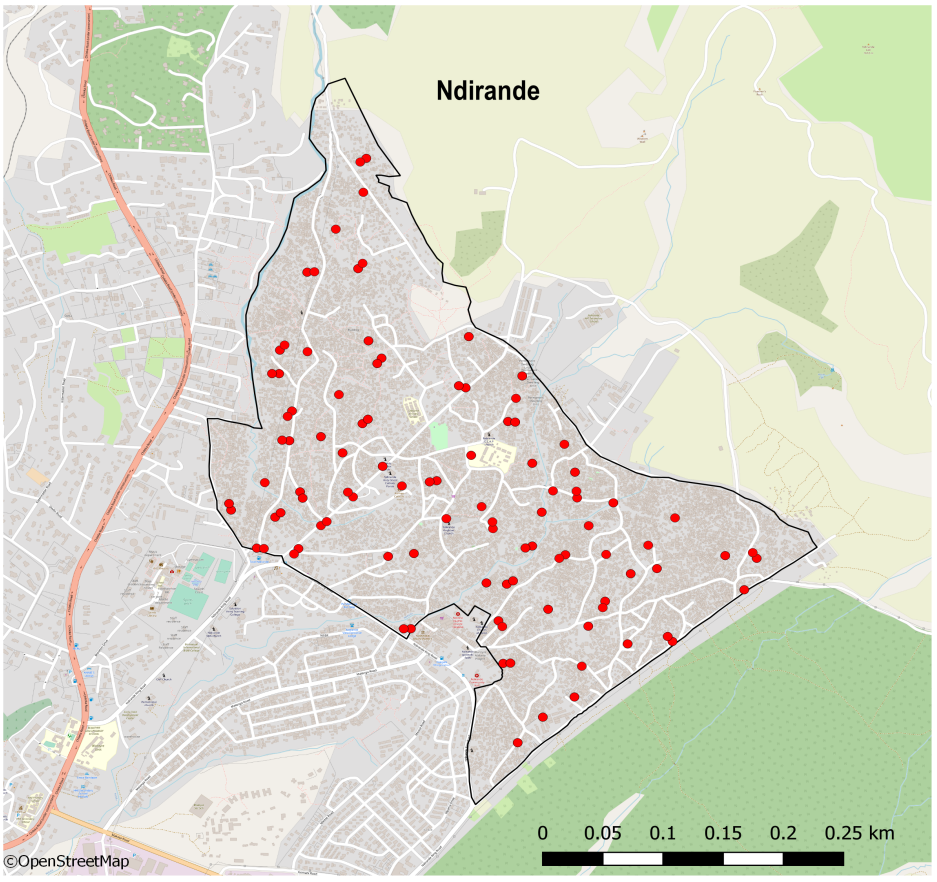


Figure 2.6: Resulting location points for Ndirande

2.4.3 Malawi - Chikwawa

Due to the extreme heterogeneity of the population density in the Chikwawa polygon and the lack of accuracy of OpenStreetMap data in that area, the assumption of an homogeneous population could not be assumed. Therefore, a population raster was used to allow for a population density-weighted sampling in Chikwawa. The sampling results for Chikwawa in Figure 2.7 show a highly noticeable distinction between the extremities and the center of the polygon. By modifying the algorithm to include population density-weighted sampling, the majority of the points turned out to be located close to the extremities of the polygon and only few in the middle of the polygon.

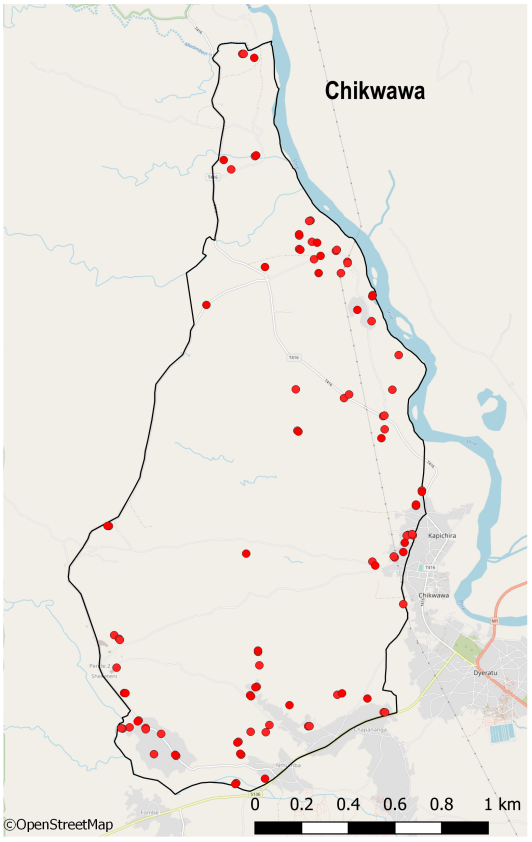


Figure 2.7: Resulting location points for Chikwawa

2.4.4 Uganda - Kampala

In Kampala, the number of households needed for sampling was 102. Due to the high population density in Kampala and its homogeneity, OpenStreetMap data was unnecessary and the sampling was carried out without prior knowledge on the household locations.

Considering the separation of the area in three consecutive polygons, an equal number of thirty-four points was sampled in each polygon. This strategy was deemed reasonable even though their size differs in order to look at heterogeneity between the polygons. A new restriction was added to sample exactly twenty-four inhibitory points in each polygon with the help of a new function determining which polygon a point belongs to, to avoid oversampling in any of the polygons. Once the inhibitory points were selected, we randomly selected ten inhibitory points in each of the three polygons, instead of using the three polygons as one to sample from. Then, the close pairs were added from

these selected inhibitory points in each of the three polygons. The sampling results for Kampala in Figure 2.8 confirmed the same number of households was sampled in each polygon and ten close pairs were present in each polygon.

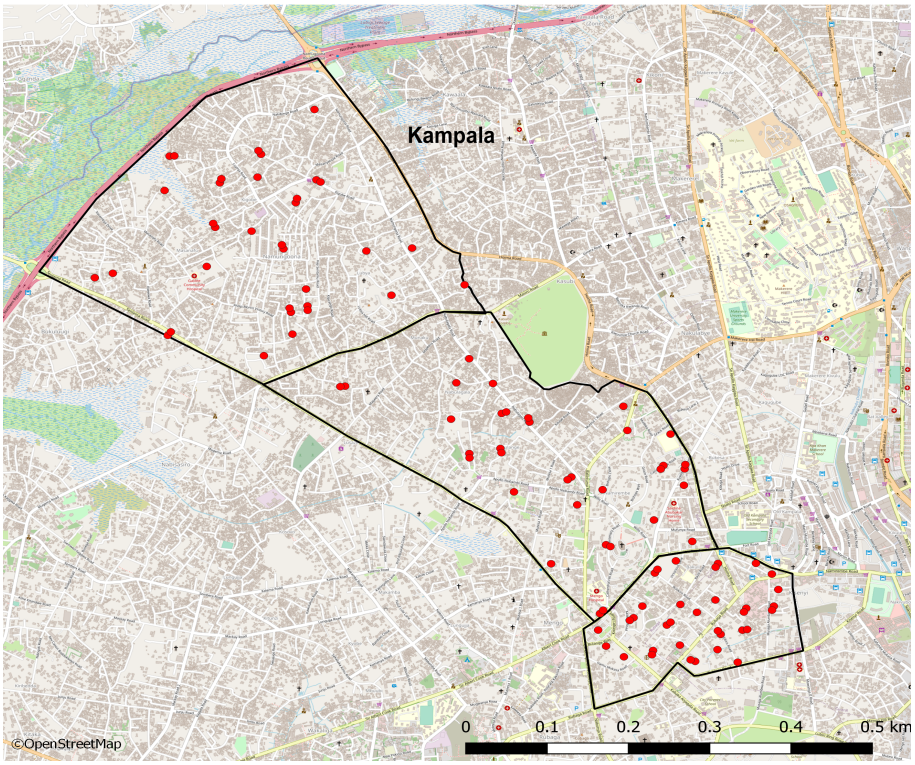


Figure 2.8: Resulting location points for Kampala

2.4.5 Uganda - Hoima

In Hoima, the total number of households to be sampled was 160. These households were equally divided between the two polygons. Using the general random spatial inhibitory design used for the Chileka area showed that the design worked well for the peri-urban area but not for the rural area due to the heterogeneity of its population. Moreover, using a world population raster to perform a similar design to the one chosen for Chikwawa was dismissed due to the low resolution of the raster after visual inspection.

However, using OpenStreetMap (OSM) data showed a high level of coverage after visual inspection and was selected to perform the design. The "osmdata" package was used to perform queries on the OpenStreetMap data and obtain the gps coordinates for all build-

ings in both Hoima study areas. We note that not all buildings represented by OSM were residential buildings but within our study site, we expected the number of residential buildings to be greater than the number of commercial buildings. The sampling was then carried out using these gps coordinates as households to sample from, as was carried out in Ndirande with the STRATAA data. The sampling results for Hoima in Figure 2.9 showed that using the OpenStreetMap data allowed us to avoid sampling in the forest area in one of the polygons or in an area with low density in the other.

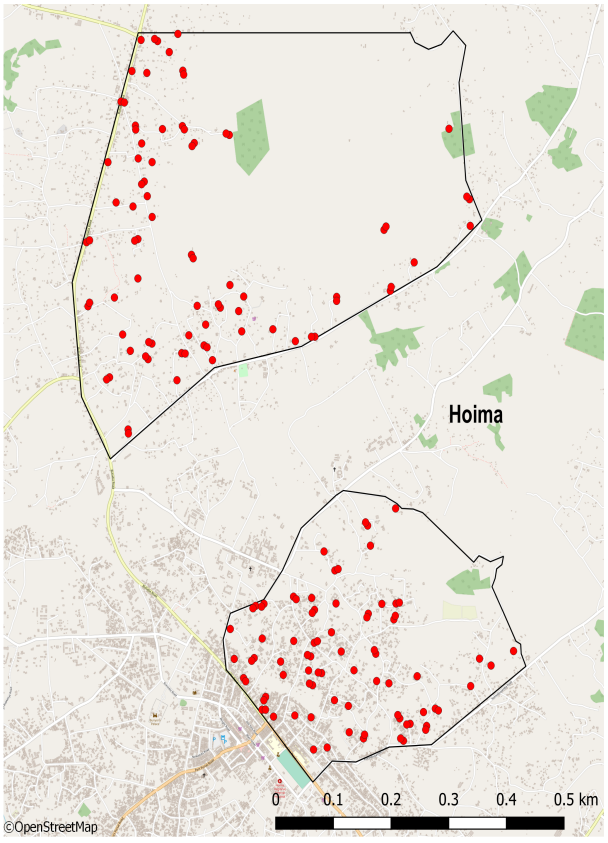


Figure 2.9: Resulting location points for Hoima

2.5 Discussion

In this chapter, we aimed to determine a suitable spatial sampling design for all study areas of the DRUM study. The study took place in five separate areas: three in the South of Malawi (Chileka, Chikwawa and Ndirande), Hoima in the North of Uganda and Kampala in the South of Uganda. These study areas allowed for a variety of urban, peri-urban

and rural settings to be investigated.

Here, we aimed to select our households by adapting recent improved spatial designs that allow for a mixture of randomness and regularity in the household sampling. We based our spatial design on the inhibitory-with-close-pairs design described by Chipeta and Diggle[139]. Using this design, seventy percent of the points were sampled using an inhibitory distance that restricts them to be at a minimum distance from all other points. The rest of the points were sampled using one inhibitory point at a time and randomly selecting a new point, called a close pair, within a circle with a pre-determined radius. Whilst the inhibitory points allow for a dispersed evaluation of antimicrobial resistance within a study area, the close pairs allow for a more localised comparison of samples, important when looking at the prevalence of antimicrobial-resistant bacteria within the same neighbourhood. Depending on the study area, the socioeconomic status and the information available on the houses, we implemented different versions of this algorithm to personalise the design to each unique area.

In Chileka, varying the arguments of the design such as the inhibitory distance and the close pairs radius were sufficient to outweigh the issue of sparsity of population within villages. In Chikwawa, using spatial tools such as WorldPop population rasters allowed us to preferentially propose sampling sites in areas of population concentration and avoid open farming land. It gave us a less precise sample of households than census data but allows for a more efficient sample of households in terms of field work efficiency than a normal inhibitory-with-close-pairs design where the density of the population is not taken into account. It helped us to avoid a massive area with little to none population due to the land topography of the study area. In Ndirande, the use of research-collected census data provided a more precise sample of households using spatial point information on existing households in the area but increased the chance of systematic bias in the study.

For more rural locations such as Hoima in Uganda, the resolution of population rasters was low, and data sources such as OpenStreetMap humanitarian project provided spatial point information on building locations and hence, a more precise selection of households to sample from. However, it did introduce a potential bias as we were not sure whether the buildings were residential or commercial. In urban regions such as

Kampala, a simple design was required due to the homogeneity and high density of the population but restrictions applied to keep the same number of points in each polygon without overlapping as they were conditioned on spatial stratification of socioeconomic status. This stratification allowed for a more precise investigation of how the inhabitants living conditions can affect the WASH practices and sociological behaviors.

These results show that varying the implementations of the algorithm to account for the density of the population in different socioeconomical settings allows for a more comprehensive sampling of households. Using spatial tools to account for the difference in density between urban and rural areas allows for an easier and faster access to households on the field. Avoiding areas with little to no population is crucial in increasing the efficiency of the fieldwork teams, as is being able to access previously existing and willing-to-participate households. However, it also shows how local factors can affect the nature of a study design and how important local prior knowledge of the areas is in order to have an efficient design.

Due to the lack of information surrounding the expected spatial correlation in these areas, we decided to follow the general convention of placing the ratio of inhibitory points to close pairs at 70% to 30%[121]. Moreover, the values chosen for the inhibitory distance and radius for the close pairs parameters were also inspired by the work done by Jillian Gauld on Typhoid[131] that showed a spatial correlation up to approximately 150 meters. This study will render results that will be essential in helping design future studies and will provide prior information to optimise the design parameters. Follow-up studies will be able to conjecture values for the length scale and variance of the Gaussian Processes and use bootstrap methodology to optimise the number of inhibitory points, close pairs and radius for the close pairs[139]. It is currently being discussed that the implementation of this design will either be integrated into the `geosample` package or will be made into its own package.

Chapter 3

Transmission patterns of ESBL-producing *E. coli* and *K. pneumoniae* leading to human gut mucosal colonisation in the community in Southern Malawi

3.1 Introduction

Infections caused by extended-spectrum β-lactamase (ESBL)-producing Enterobacteriaceae are a large problem in African healthcare institutions and communities[70]. The prevalence of ESBL-producing Enterobacteriaceae varies widely, however, depending on the specimen studied, the setting, the country and the type of sample processed[70]. The emergence of ESBL-producing strains of *Escherichia coli* and *Klebsiella pneumoniae* is of great concern in Africa, because the antimicrobials required to treat them are typically not available and thus therapeutic options for these infections are severely limited [159]. Between 2003 and 2016, among hospital patients with fever or suspicion of sepsis in Malawi, ESBL resistance in *Escherichia coli* rose from 0·7% to 30·3% while ESBL resis-

tance in *Klebsiella spp* rose from 11·8% to 90·5% [2]. In order to develop symptomatic infection, one typically has to swallow these pathogens and be colonised with them first. Due to paucity of diagnostic microbiology in Africa, most of the available studies focused on symptomatic infections in the hospital setting, and left a knowledge gap regarding carriage in the community settings in sub-Saharan Africa. Consequently, the spread of these ESBL-producing Enterobacteriaceae in low- and middle- income countries like Malawi is poorly understood.

Gaining a better understanding of the spatial distribution of ESBL-producing Enterobacteriaceae is essential in order to assess the dynamics of transmission and tailor appropriate interventions to reduce their increasing dissemination. To the best of our knowledge, no study has investigated the spatial distribution of ESBL-producing Enterobacteriaceae in various community settings in sub-Saharan Africa. However, studies of other bacterial diseases have found clustering of prevalence across space[130, 160]. Specifically, one of these studies explored spatial patterns of typhoid fever in Blantyre in Malawi and found small-scale spatial correlation of approximately 200 meters[160]. Considering the spatial closeness of our studies and the nature of *Salmonella* Typhi as an enteric pathogen closely related to *E. coli*, we hypothesised we might expect to find similar spatial clustering in the prevalence of ESBL-producing Enterobacteriaceae colonisation in the area.

The dynamics of ESBL-producing Enterobacteriaceae in the community are still unclear. Therefore, there is a lack of knowledge on which transmission pathways are more important. Moreover, ESBL-producing Enterobacteriaceae is a global problem, with varying prevalences depending on the country and region, but present in both high-income and low- and middle-income countries[5]. This highlights the need to determine whether the prevalence and transmission of ESBL-producing Enterobacteriaceae colonisation is dependent on the socioeconomic context of an individual or on their individual or household characteristics. While several studies have found antibiotic use as a risk factor for colonisation[3, 4], not much is known about the role of demographics and socioeconomic status in the prevalence of human gut colonisation. Some studies have found older age as a risk factor for ESBL-producing Enterobacteriaceae colonisation[101, 103]. Few studies have found socioeconomic status to be a risk factor for colonisation. However, their association is not well defined as these studies

have found opposite results[107, 108].

The aim of this chapter is to investigate individual-level and household-level risk factors for ESBL-producing Enterobacteriaceae colonisation in different socioeconomic settings in Malawi, accounting for spatial variability. Herein, we use microbiological data on sampled participants and individual surveys to explore potential risk factors. Additionally, the study was designed to include geolocalisation of the households to allow for a spatial investigation and compare the risk factors for ESBL-producing *E. coli* and *K. pneumoniae* between rural, urban and peri-urban areas. We first fit generalised linear mixed models on all combined areas and for each specific area to identify individual-level and household-level risk factors and to quantify their effect on ESBL colonisation. Subsequently, we use a Bernoulli geostatistical model to explore eventual spatial clustering of the prevalence of ESBL colonisation.

3.2 Methods

3.2.1 DRUM database

During the course of the DRUM study, due to its trans-disciplinary nature, many different types of data were collected. The different teams captured longitudinal human microbiological data, environmental samples (food, animal, river..), case report forms which contain both individual and household survey data, and WASH surveillance data. This data was pulled daily from the local data centres in Uganda and Malawi and subsequently transferred to Lancaster University servers[102]. There, the data was managed and linked by Barry Rowlingson (DRUM Data Manager) and formalised into an SQL database. The outline of the DRUM database is divided into five categories: individual, household, laboratory results, human and non-human samples which are illustrated in Figure 3.1. Each of these categories also consists of multiple datasets, reaching a total of forty-five datasets for the entirety of the DRUM database. The complete DRUM database schematic can be found in Appendix B.1.

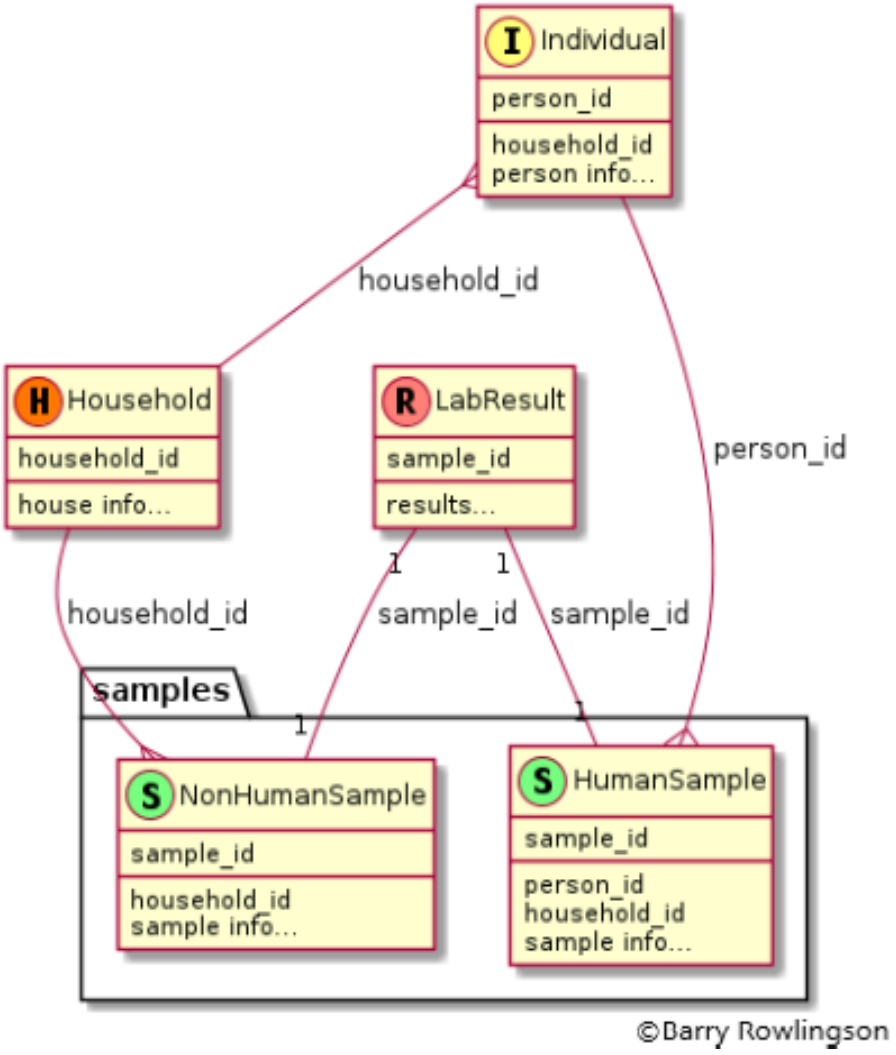


Figure 3.1: DRUM database outline

3.2.2 Data

Using the DRUM database at the time of analysis (April 2021), a chosen subset of variables was extracted from the Individual dataset, following a screening of the questionnaire for variables having the most importance for this analysis. This was a pragmatic decision made in response to a very wide dataset to focus on personal characteristics and antibiotic use. The resulting dataset contained 182 variables such as household ID, patient ID, age, gender, healthcare worker status, HIV-related variables, tuberculosis-related variables and antibiotic-related variables.

After looking at HIV status, whether people were on therapy for HIV, and

whether they were on cotrimoxazole preventative therapy (CPT), all three variables showed a very high correlation to each other and two were therefore removed. Only the HIV status variable was kept for analysis since over 90% of the HIV-positive individuals were on CPT. The tuberculosis-related variables showed that only twelve individuals had previously been treated for tuberculosis and none of the individuals were currently on therapy for tuberculosis, thus both variables were removed due to lack of information. Only four individuals worked as healthcare workers among all the individuals, considering seven hundred and forty of them were students or unemployed, this variable was also removed, due to the sample size.

The variables relating to antibiotic use all required the participant to indicate whether or not they had received antibiotics in the recent past: in the last four weeks, last three months, last six months and currently. For each of these variables, we had twenty-seven variables detailing possible choices of antibiotics that we reduced to four: amoxicillin, amoxicillin-clavulanic acid (co-amoxiclav), cotrimoxazole and others. This was decided after key informant interviews showed that amoxicillin and cotrimoxazole were the most used antibiotics in Malawi and after interviews with policy makers revealed that cotrimoxazole is viewed as a likely driver of resistance. Due to the relatively small sample of people who received antibiotics for a majority of the antibiotic-related variables, these variables were all merged to create a new variable that indicated whether the individual had taken amoxicillin, co-amoxiclav, cotrimoxazole or other antibiotics within the last six months. The co-amoxiclav variable was taken out for the modelling after noticing that no individual was given the combination.

The household ID, household size and household income were extracted from the HouseholdSummary dataset in the DRUM database in order to be merged with our dataset by household ID. The household size in this chapter refers to the number of people living in the house. A new sample date variable was created by determining for each sample the number of days that passed since the first ever sample for all individuals was taken on 30th of April 2019.

Using the HumanSample dataset and the StoolLab dataset, diagnotics functions were used to determine whether samples that had already been analysed were positive for ESBL *E. coli* or ESBL *K. pneumoniae*. These results were then merged with our

dataset by patient ID to allow access to sample, patient and household information in a unique dataset. After removing various duplicates, samples with missing individual information and the households that were located more than 200 meters away from a polygon limit, the dataset contained 2015 samples from 650 individuals in 187 households and thirty-six variables. The list of variables with the rationale for including them can be found in Table 3.1. Around 900 individuals were initially expected to participate in the study, however due to the delay caused by the COVID-19 pandemic, there was a lack of availability of lab-returned samples at the time that explains the reduction in the number of individuals. Consequently, for this chapter, we chose to focus on a subset of this dataset, which contained baseline extended samples. We defined a baseline extended sample as the first available lab-returned sample for each of the six-hundred and fifty individuals. This also allowed us to focus on the spatial aspect of the data, before exploring its longitudinal aspect in Chapter 4.

Table 3.1: Rationale for the covariates

Variable	**Rationale**
Age	Risk factor: Infants are more likely to enter in contact with potentially contaminated environments. Older adults require access to healthcare more frequently, increasing the risk of exposure to contaminated healthcare settings.
Sex	Protective: Women are more likely to get into contact with contaminated environment while doing the housework and taking care of the children.
HIV status	Risk factor: Immunosuppression due to HIV infection can lead to gut dysbiosis and long term antibiotic prophylactic therapy and consequently susceptibility to acquisition and long-term carriage of resistant bacteria
Use of amoxicilline in the last 6 months	Risk factor: Antibiotic use leads to the depletion of drug-susceptible gut-commensal bacteria, providing a selective advantage to resistant bacteria behind to reproduce.
Use of cotrimoxazole in the last 6 months	
Use of other antibiotics in the last 6 months	
Number of people living in the household	Risk factor: More people living in the household, more contact and proximity, more opportunities for within-household transmission.

Average household monthly income	Other: Income allows for better access to WASH infrastructure (protective) but also to antibiotics (risk factor).
In-house prevalence	Risk factor: Sharing a household with other colonised individuals is likely to increase the risk.
Number of colonised people in the house	Risk factor: Sharing a household with other colonised individuals is likely to increase the risk.
Harmonic terms (sinday, cosday, sinday2, cosday2)	Risk factor / Protective: Higher risk during the wet season than the dry season.
Study area	Risk factor: Chikwawa is the rural area, with a higher presence of animals and potentially antibiotic use for animals. Ndirande is a high-density settlement, with potential for higher within-household transmission.

3.2.3 Exploratory analysis

The processed dataset contained the first available lab-returned sample for each individual. As Figure 3.2 shows, 57.2% of the individuals were under the age of 20 years old, 34.6% were between 21 and 50 years old, 6.9% were between 51 and 70 years old and the last 1.2% were over 70 years old. It should be noted that adults were considered to be over 15 years old (52.2% of the individuals), and school age was considered to be older than 5 years old and up to 15 years old (28.6%). Among the individuals, 57.7% were female with slightly varying proportions in each age group. 66.1% of the adults (>15) were female and 48.4% of the school age children were female.

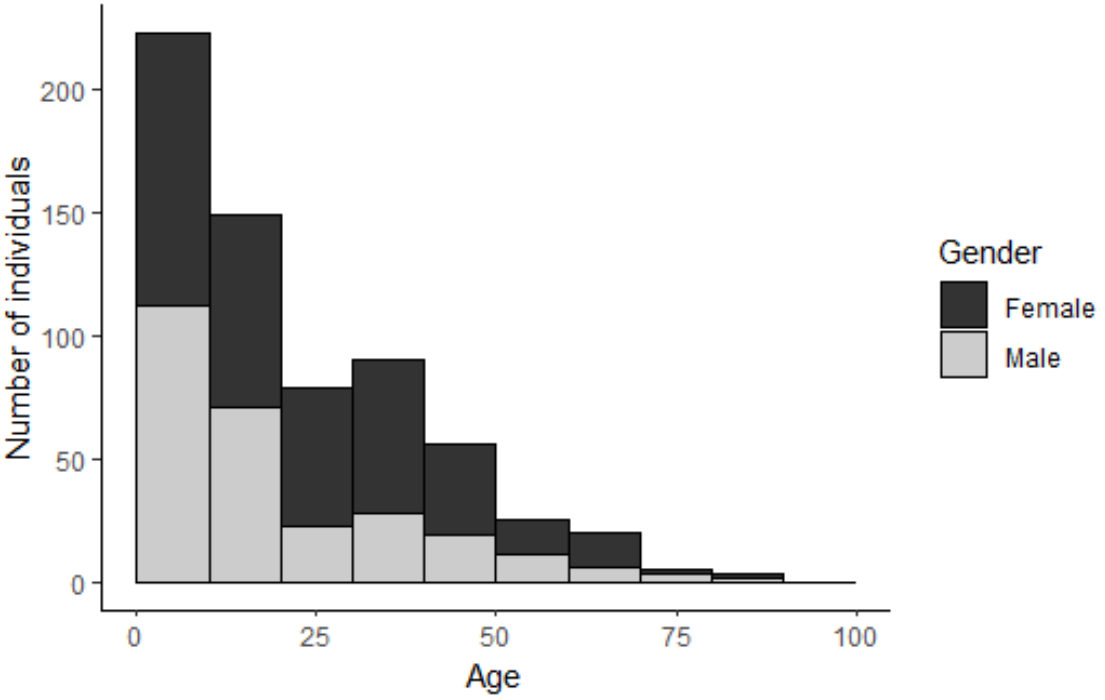


Figure 3.2: Distribution of age and gender for the 650 individuals

As can be seen in Figure 3.3, over three-quarter of the households had between three and six individuals. Fifteen households had two individuals while twelve households had seven individuals. Only twelve households had eight or more individuals. The median number of individuals in one household over all households was four individuals. The distribution of individuals per household in each Malawian study area is shown in Figure 3.4. Ndirande was the area with the highest number of households with eighty-six households and a median number of five individuals per household. Chileka

and Chikwawa respectively had fifty-six and forty-six households with a median of four people per household. Chikwawa had the smallest distribution out of the three areas. The two largest households with ten and thirteen individuals were situated in Ndirande. Spatial maps of household size by study area can be found in Figure 3.5. There was no sign of a correlation between spatial location of the households and household size.

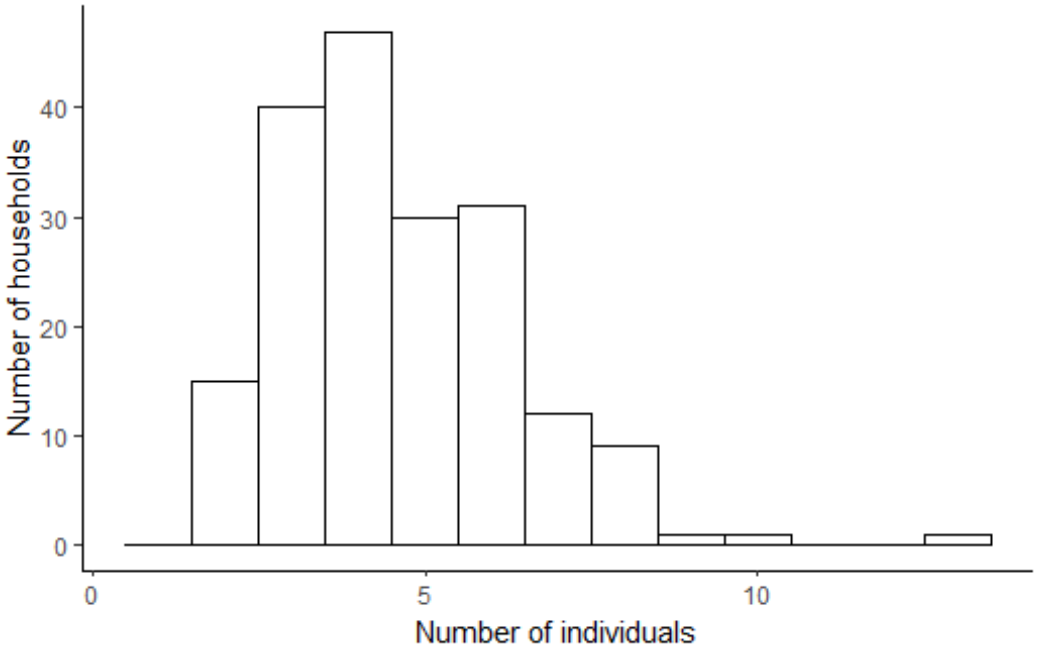


Figure 3.3: Distribution of individuals per household

The distribution of the monthly income of the households can be observed in Figure 3.6. More than half of the households had a monthly income between 15000 and 50000 Malawi Kwacha (mwk). Among twenty-eight households with a monthly income of less than 15000 mwk, fifteen had an income equal to or above 5000 mwk while thirteen had an income lower than 5000 mwk. Forty-six households had a monthly income of over 50000 mwk, including fifteen households that had over 100000 mwk. The median monthly household income over all areas was 30000 mwk.

The boxplots in Figure 3.7 indicated a slight difference in the distribution of income among the different study areas. Chikwawa was the area with the lowest monthly income with a median income of only 20000 mwk compared to 32500 mwk in Chileka

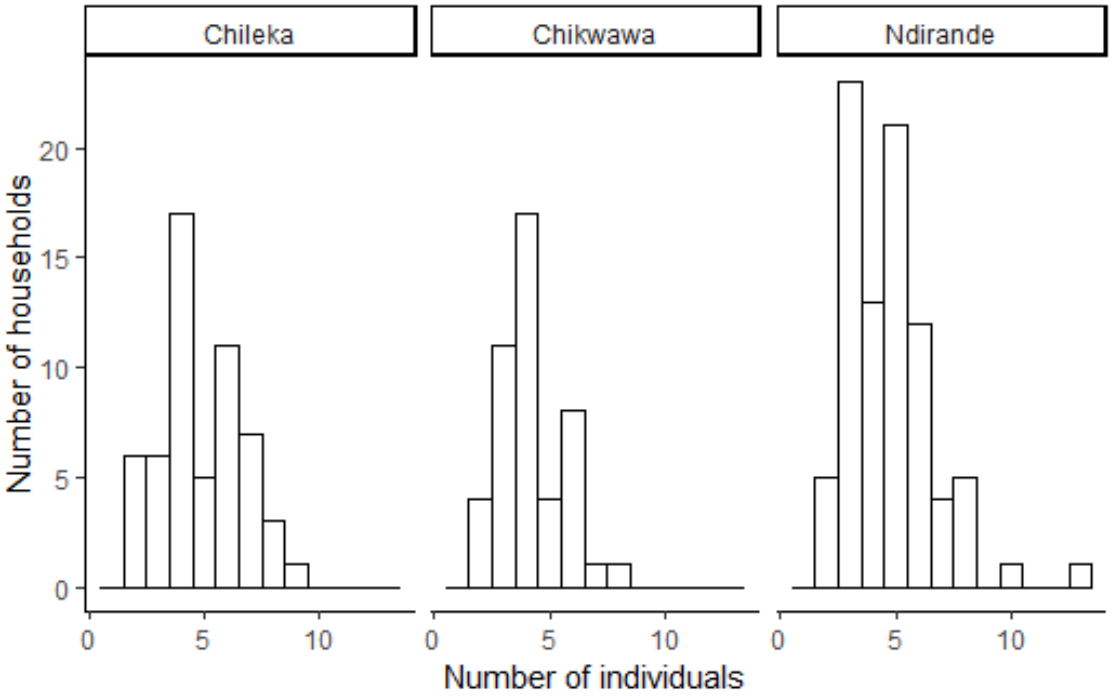


Figure 3.4: Distribution of individuals per household in each study area

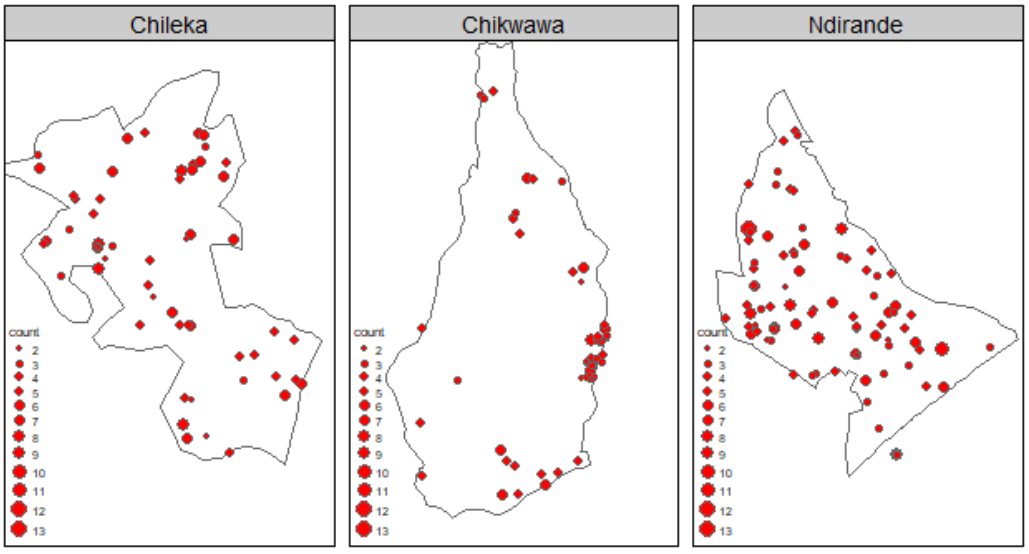


Figure 3.5: Map of individuals per household in each study area

and 50000 mwk in Ndirande. There appeared to be less variation in the monthly income within the Chikwawa study area with a range of around 50000 mwk. In comparison, the ranges of Chileka and Ndirande were almost twice the one of Chikwawa with values close to 100000 mwk. Each area appeared to have a couple of outlier houses with a higher income than average.

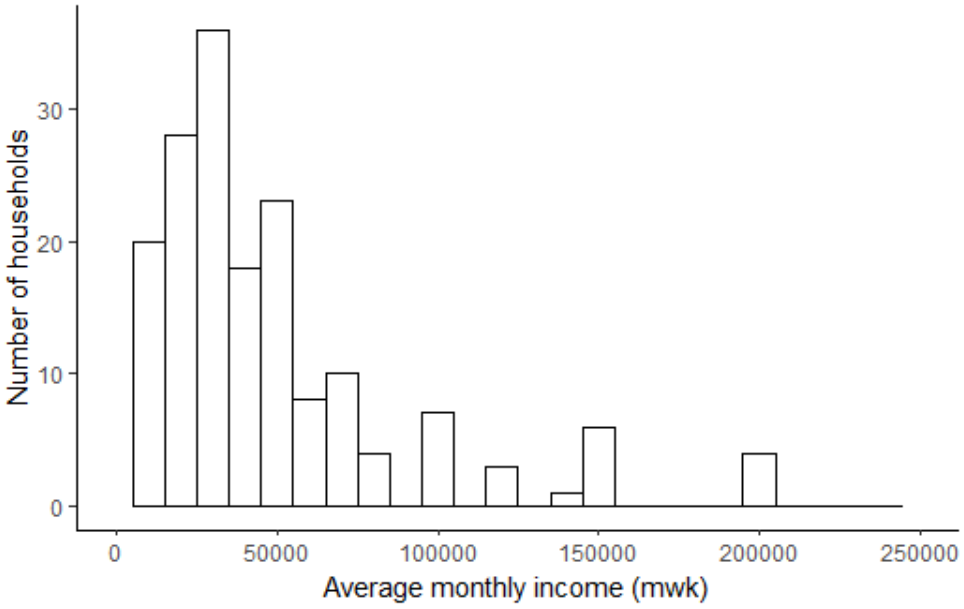


Figure 3.6: Distribution of household monthly income

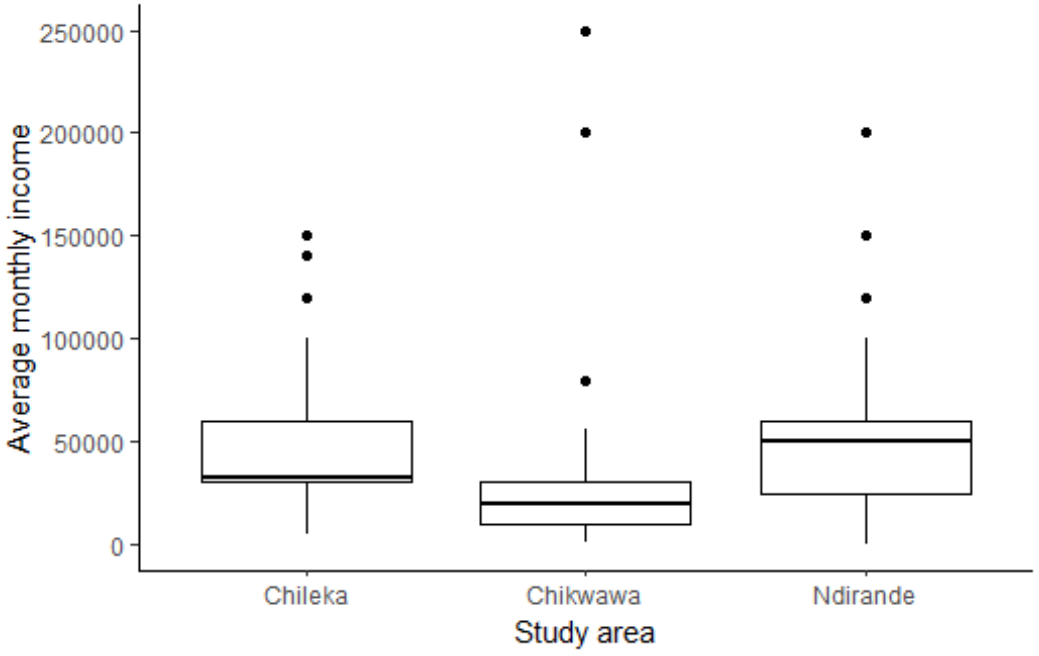


Figure 3.7: Distribution of household monthly income in each study area

Among all the study areas, Chikwawa was the only one where the monthly income appeared to be higher in a specific area of the polygon, on the east side. The distribution of the monthly income in Chileka was relatively sparse with a small area of richer houses in the north-west area of the polygon. In Ndirande, it appeared to be spatially homo-

geneous throughout the polygon. These spatial maps of income can be found in Figure 3.8.

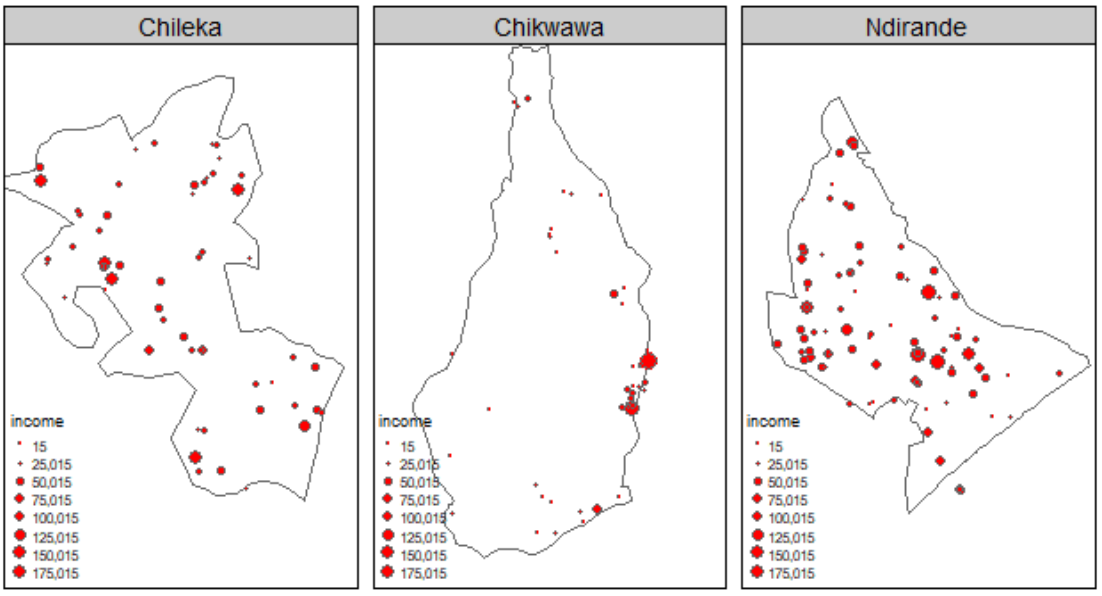


Figure 3.8: Map of income per household in each study area

As can be seen in Figure 3.9, approximately half of the individuals had an unknown HIV status, especially in Chileka. Of the remainder, the majority were HIV negative. Chikwawa and Ndirande showed a higher number of individuals with HIV non-reactive status than individuals with unknown status. Per area, the lowest proportion of individuals with unknown status was found in Chikwawa (25%). We can conjecture that a potential correlation between low income and known HIV status might be caused by the increased accessibility of local testing by organisations such as non-governmental organisations that are more present in areas with lower income. The distribution of HIV status per study area can be found in Figure 3.10.

In the last six months, forty-two individuals (6.5%) were given amoxicillin, forty-three were given cotrimoxazole (6.6%) and thirty-seven were given other antibiotics (5.7%). These proportions can be found in Table 3.3.

Out of the 650 samples, 249 were positive for ESBL-producing *E. coli* (38.3%), 84 were positive for ESBL-producing *K. pneumoniae* (12.9%) and 48 were positive for both (7.4%). The ESBL prevalence can be found in Table 3.4.

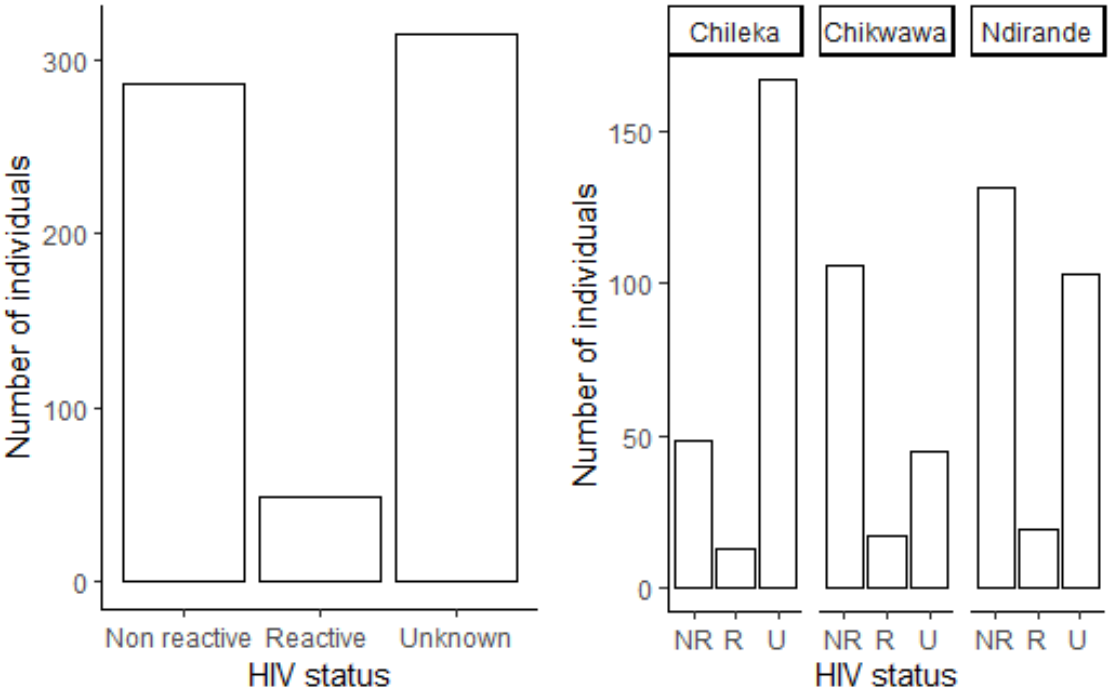


Figure 3.9: Distribution of HIV status (NR: Non-reactive, R: Reactive, U: Unknown)

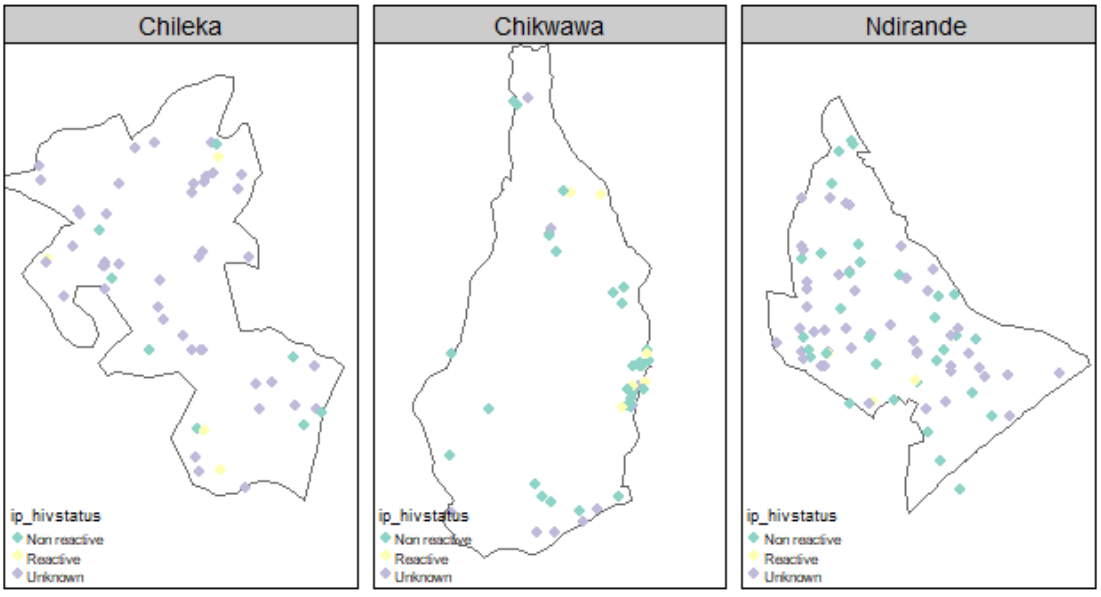


Figure 3.10: Map of HIV status per polygon

Table 3.3: Antibiotic use in the last six months

Antibiotics	Yes	Total
Amoxicillin	42	650
Cotrimoxazole	43	650
Others	37	650

Table 3.4: Prevalence of ESBL *E. coli* and ESBL *K. pneumoniae*

Polygon	ESBL	Positive (P)	Total (T)	(P/T)*100
Ndirande	*E. coli*	112	254	44.1%
	K. pneumoniae	32	254	12.6%
Chikwawa	*E. coli*	54	168	32.1%
	K. pneumoniae	15	168	8.9%
Chileka	*E. coli*	83	228	36.4%
	K. pneumoniae	37	228	16.2%
All	*E. coli*	249	650	38.3%
	K. pneumoniae	84	650	12.9%

According to the latest census in Malawi [161], the sex ratio for the Southern Region of Malawi in 2018 was 92.6 men per 100 women. The census results showed that infants aged less than 1 year represented about 3% of the population, under five years old children about 15%, adults (between 18 and 65 years old) about 49% and seniors (65 years or older) represented a further 4%. The median age of the population was 17 years old. The average household size in the Southern Region was 4.3 individuals. According to the World Bank, more than 70% of Malawians were living below the international poverty line of $1.90/day in 2016[162]. We note that the vast majority of households in our study were below that line. Our age distribution data is highly representative of Malawi's population structure, with a similar repartition and a median of 18 years old. The sex ratio in our data was approximately 74 men to 100 women (73.3 in chapter 3 and 75.3 in chapter 4). Traditionally in this context, men go to work while women take care of the children and the housework, therefore this could explain the lower ratio of men being sampled during our visits. However, considering the very high prevalence of ESBL-producing *E. coli* we found in this study, we believe that a more balanced ratio of men to women would not have made a difference. The average household size reflected well the Malawian household structure with 4 to 5 individuals. Overall, our study sample is very representative of the study population.

3.2.4 Identifying individual and household-level risk factors

3.2.4.1 Modelling framework

Generalised additive models (GAM)[163] were used as an exploratory tool to look at the relationship between the response and our continuous explanatory variables such as age and date. Generalised additive models allow for an estimation of the shape of a potentially non-linear relationship directly from the data and can be used to determine whether a linear relationship can be assumed between these variables and the response or if a non-linear relationship is more adequate.

In order to explore the effect of the different covariates on the ESBL-producing Enterobacteriaceae colonisation status, a logistic model was fitted using the data obtained after preprocessing. This model can be expressed as

$$\text{logit}(p_{ij}) = \alpha + x_{ij}^T\beta + v_i \tag{3.1}$$

where Y_i follows a Bernoulli distribution with probability p_{ij} i.e.

$$Y_{ij} \sim \text{Bernoulli}(p_{ij})$$

α is the intercept, x_{ij} are the explanatory variables, β are the regression coefficients for the fixed effects and v_i is the household-level residual following a Normal distribution

$$v_i \sim \text{Normal}(0, \sigma^2)$$

This model returned coefficients that can be interpreted as log(odds). In order to return odds ratio (OR) for each variable, we exponentiated the resulting coefficient and calculated the 95% confidence intervals (CI).

3.2.4.2 Variable selection

The variables used for the logistic models are presented in Table 3.5 and 3.6 with their description and summary. The study area was included as a variable for the models on all three areas combined. In order to investigate whether the individual's colonisation was

associated with sharing the house with other colonised individuals, the number of positive people in the house and the in-house prevalence (excluding the sampled individual) were included as potential variables. Harmonic regression terms (annual and biannual) were also created using the sample date to look at seasonal effect over the year. They are constructed using the sample date, with $T = (1, \frac{1}{2})$ the period in years, as follows :

$$\text{sinday} = \sin(\frac{2\pi * \text{sampledate}}{365T}) \qquad \text{cosday} = \cos(\frac{2\pi * \text{sampledate}}{365T})$$

Table 3.5: Variables for the models (numerical)

Description	**Min-Max**	**Median**	**1st-3rd quantile**
Age	0-87	17	7.25-34
Household size	2-13	5	4-6
Household monthly income	40-250000	30000	20000-50000
Number of colonised people in the house (excluding sample)	0-6 (*E. coli*)/0-4 (*K. pneumoniae*)	1/0	0-2/0-1
In-house prevalence (excluding sample)	0-10 (*E. coli*)/0-10 (*K. pneumoniae*)	3.33/0	0-6.67/0-2
Sample date (Days since 1st sample on 30/04/2019)	0-539	210	154-301
Harmonic terms	[-1,1]	N/A	N/A

The following variables: age, antibiotic use variables, household size, household income and sample date were standardised. First, we subtracted the mean from the value of that variable for each sample, and then we divided this difference by the standard deviation, resulting in a mean of zero and a standard deviation of one. When interpreting model results for standardised variables, the OR for each standardised variable was interpreted as a change with every increase of one standard deviation, and not as one increase in the variable unit such as for the non-standardised variables. In the case of the in-house prevalence variable, we multiplied by 10 the value before running the models, in order to get an increase of 0.1 instead of 1.

Table 3.6: Variables for the models (categorical)

Description	Categories	Distribution
Gender	Female/male	375/275
HIV status	Reactive/non reactive/unknown	49/286/315
Amoxicillin in the last 6 months	1/0	42/608
Cotrimoxazole in the last 6 months	1/0	43/607
Other antibiotics in the last 6 months	1/0	37/613
Study area/polygon	Ndirande/Chikwawa/Chileka	254/168/228

3.2.4.2.1 Forwards selection algorithm

A forwards selection algorithm was used to select the most important variables in the model and determine if a mixed-effects model is more appropriate than a fixed-effects model. This algorithm takes in a dataset, a response variable and a list of explanatory variables (fixed or random). It acts as a loop, starting with a null model, running every potential model adding a single variable, calculating its respective Akaike information criterion (AIC)[164] and selecting the variable with the lowest AIC at the end of each step to be added to the final model. This algorithm allowed us to compare models with and without random effects in a more efficient manner and to select for each area which one is more suitable. The code for implementation of this algorithm can be found in Appendix B.2.

3.2.4.2.2 Model selection

Initially, all areas were considered as one combined entity and subsequently, models were run on each specific study area for ESBL-producing *E. coli* and *K. pneumoniae*. For each area and each bacterial species, the following approach was undertaken:

1. A full model including all variables in Tables 3.5 and 3.6 (except for the number of colonised people in the house and in-house prevalence) was run. This allows for an estimation of all variables and an estimate of the full model's AIC.

2. After using the forwards selection algorithm on the full model, a new model using these selected variables was run. This helped us highlight the importance of specific variables.

3. Finally, using the initial full model, we replaced the household size by the number of colonised people in the house and ran the selection algorithm again to compare the results. Then, we did the same replacing it by the in-house prevalence and compared the AIC of the two models. The one with the lowest AIC was selected to be the third model. This approach was chosen in order to avoid correlation issues between household size, number of colonised people in the house and in-house prevalence. The inter-class correlation (ICC) was calculated for the random effects.

The different models fitted for *E. coli* and *K. pneumoniae* are described in Table 3.7. Analyses were conducted using the statistical software R, with `lme4`[165] and `lmerTest`[166] packages.

Table 3.7: Description of the models

Polygon	ESBL	Model	Fixed effects	Household random effect	AIC	ICC
Malawi	*E. coli*	ME1	Full model	Present	819.5	0.212
		ME2	Sample date, harmonic terms	Present	803.7	0.231
		ME3	In-house prevalence, sample date, harmonic terms	Absent	773.3	-
	K. pneumoniae	MK1	Full model	Present	509.8	0.234
		MK2	None	Present	493.2	0.262
		MK3	Colonised people in the house	Absent	488.39	-
Ndirande	*E. coli*	NE1	Full model	Present	344.4	0.204
		NE2	Sample date, harmonic terms	Present	326.9	0.200
		NE3	In-house prevalence, sample date, harmonic terms	Absent	316.8	-
	K. pneumoniae	NK1	Full model	Present	211.7	0.212
		NK2	None	Present	193.1	0.242
		NK3	Colonised people in the house	Absent	191.3	-
Chikwawa	*E. coli*	KE1	Full model	Present	199.1	0.189
		KE2	Cotrimoxazole in the last 6 months, harmonic terms	Present	187.8	0.254
		KE3	In-house prevalence, cotrimoxazole, harmonic terms	Absent	174.2	-
	K. pneumoniae	KK1	Full model	Present	104.2	0.879
		KK2	Cotrimoxazole in the last 6 months	Present	93.6	0.849
		KK3	Colonised people in the house	Absent	90.6	-
Chileka	*E. coli*	CE1	Full model	Present	308.2	0.146
		CE2	Sample date	Present	292.0	0.180
		CE3	In-house prevalence, sample date	Absent	285.7	-
	K. pneumoniae	CK1	Full model	Present	211.6	0.056
		CK2	HIV status, sample date	Absent	198.4	-
		CK3	HIV status, sample date	Absent	198.4	-

3.2.5 Prevalence of ESBL-producing *E. coli* and *K. pneumoniae* across space

In order to investigate potential spatial correlation within our areas, a Bernoulli geostatistical model was fitted under the following assumptions:

- S(x) is a stationary and isotropic Gaussian process[121] with covariance matrix

$$\Sigma_{ij} = \sigma^2 e^{-||x_i - x_j||^2/\phi} + \tau \tag{3.2}$$

  for locations x_i and x_j. ϕ is the scale of spatial correlation, σ^2 is the variance of the spatial process, τ is the nugget effect.

- Y_i are mutually independent and identically distributed, such that:

$$Y_i(p(x_i))$$

  and

$$\text{logit}(p(x_i)) = \alpha + {d_i}^T \beta + S(x_i)$$

  where d_i is a vector of covariates and p(x_i) is the probability of colonisation, at location x_i.

Here, we used a non-centered parameterisation therefore

$$\text{logit}(p) = \alpha + d^T \beta + Lu$$

where $u \sim N(0,1)$ and $LL^T = \Sigma$ (Cholesky decomposition)

Priors :

$$\alpha \sim \text{Normal}(0,100)\ \beta \sim \text{Normal}(0,100)$$

$$\sigma^2 \sim \text{Gamma}(2,1)$$

$$\tau \sim \text{Gamma}(2,1)$$

$$\phi \sim \text{InverseGamma}(1.12, 683.56)^{-1}$$

We explored different priors for the scale of the spatial correlation ϕ, starting with normal and half-normal distributions, to a Gamma distribution to finally, an inverse Gamma distribution. The normal distributions were too informative, therefore impacted the

model too significantly. In order to select a prior that constrained our inferences to a more reasonable length scale, we then looked at a Gamma distribution. We wanted 98% of the mass of our prior to stay between a chosen lower and upper length scale, but the Gamma distribution tail strongly heads towards zero so we decided to use an inverse Gamma distribution that has a lighter tail towards zero and more strongly constrains the posterior distribution as we wished. We learned from similar research on Typhoid [160] that spatial correlation in Malawi was detected at approximately 200 meters and we designed the spatial sampling according to that information. Therefore, we set our spatial correlation range at 30 meters and 200 meters and used the correlation structure to calculate an estimated ϕ for both values, resulting in 0.0003 and 0.084. Using these values, we determined the shape and scale of the respective inverse Gamma distribution.

The model was fitted using the Stan modelling language [167] on R through the package RStan [168]. A No-U-Turn Sampler (NUTS), an extension to the Hamiltonian Monte Carlo (HMC) algorithm[169], was used to simulate the samples required for Monte Carlo maximum likelihood estimation. Whilst HMC is highly sensitive to the desired number of steps given by the user, NUTS uses a recursive algorithm that automatically selects an appropriate number of steps in each iteration, removing the need for the user to set this parameter and avoiding a u-turn of the algorithm[170]. NUTS has been shown to perform at least as efficiently, if not better than HMC[170]. The model was run with 10,000 iterations for each of the three Markov chains. Throughout this thesis, the Bayesian models were run on CentOS 7.6 Linux servers through the High End Computing facility at Lancaster University.

In order to avoid the model getting stuck due to the step size of the sampler being too high, we set the target acceptance rate parameter delta of the sampler at 0.95 instead of 0.8, pushing it to take smaller steps. This caused the sampling to become slower but more efficient and allowed us to ensure better validity of the model estimates.[171] Convergence was evaluated by inspection of traceplots and by making sure the Gelman-Rubin statistic[172] was close to 1 (<1.01).

We projected the results of the Gaussian process over a regular grid of the respective areas by using the covariance matrix:

$$S^* = k(x^*, x)[k(x, x^*)^{-1} S]$$

with S the value of the Gaussian process determined at observed locations x, x^* the projected locations, S^* the Gaussian process at projected locations x^* and k the covariance structure in Equation 3.2.

3.3 Results

3.3.1 Risk factors for ESBL-producing *E. coli* and *K. pneumoniae* colonisation in various community settings

3.3.1.1 Malawi

Looking at Malawi, as can be seen in Figure 3.11, a marked seasonality was detected for ESBL-producing *E. coli*, with higher prevalence in the wet season (November to April) compared to the dry season (May to October). This is consistent with other research on different cohorts in the same region of Malawi[173]. By area, we noticed it in Ndirande and Chikwawa so we assumed the same seasonality applies to Chileka considering the closeness of the areas in terms of seasons. This seasonality was included into the models through the use of harmonic terms as potential covariates, as we expected this may further explain apparent spatial clustering, and field teams tend to visit local clusters of households together.

For ESBL *K. pneumoniae*, no seasonality was detected. Therefore, the harmonic terms were not included as potential covariates into the following models. The results for ESBL *E. coli* and ESBL *K. pneumoniae* also indicated that integrating age as a simple linear variable would be sufficient.

3.3.1.1.1 ESBL-producing *E. coli* in Malawi

Logistic regression was applied first over all the polygons in Malawi combined (Ndirande, Chikwawa and Chileka). Looking at the full model (Table 3.8) using all the variables we pre-selected as important variables to investigate, only temporal variables came

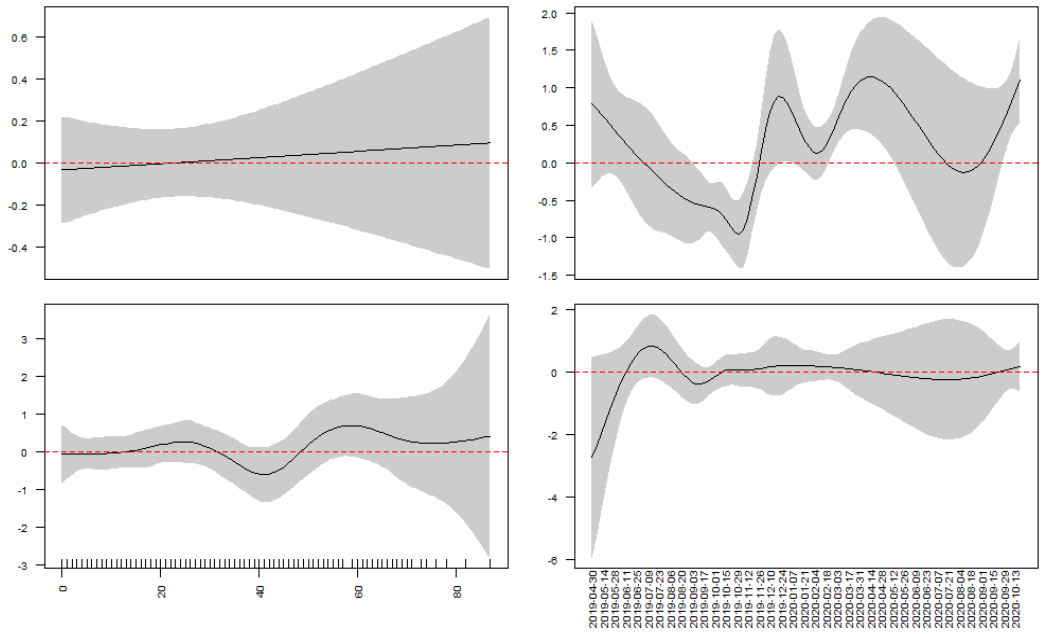


Figure 3.11: GAM results on age (left) and date (right) for ESBL *E. coli* (top) and ESBL *K. pneumoniae* (bottom) over all the study areas

out as significant, such as the sample date (OR 1.719 [CI:1.290-2.291]) and the harmonic term cosday (OR 2.403 [CI:1.483-3.894]). The household income was also identified as slightly significant but had an odds ratio of 1. The full model was a mixed-effects model and included the household as a random effect with a variance of 1.066.

Throughout this chapter, the ICC for the models showed that 15% to 26% of the variability is accounted for by the household random effect, depending on the model, as can be seen in Table 3.7. Therefore whilst variability is still dominated by the individual level, the household has a marked intrinsic effect on an individual's chance of colonisation. The only variations in the ICC are when looking at ESBL-producing *K. pneumoniae* in Chikwawa and Chileka. Respectively, the ICC is approximately 0.85 and 0.05, which can be explained by the distribution of the individuals colonised in these areas. In Chikwawa, out of the fifteen individuals colonised with ESBL-producing *K. pneumoniae*, eight of them live in the same three households. Contrarily, in Chileka, the thirty-eight individuals colonised with ESBL-producing *K. pneumoniae* live in twenty-seven households.

After using our selection algorithm to select the most important variables among the full model, the previously significant variables were still chosen as significant (Table

3.9). Their odds ratio were extremely similar to the previous model. The model also included the second harmonic term cosday2 as a non-significant variable. The household random effect variance had slightly increased, going from 1.066 to 1.142.

Finally, we looked at adding the number of positive people in the house or the already existing prevalence (excluding the individual) in the house as potential variables. This allowed us to investigate if a dilutional effect can be found in the house or if the population density in the house prevailed. For both these models, we included the new variable and removed the household size to avoid correlation issues. We then used the forwards selection algorithm on the variables to choose the best possible model. Throughout this chapter, we defined the best model as the model with the lowest AIC.

Looking at the AIC, we selected the best model out of the three models. For ESBL *E. coli* in Malawi, the best model was the prevalence model (Table 3.10), which identified three important risk factors, the in-house prevalence (OR 1.199 [CI:1.142-1.258]), the sample date (OR 1.328 [CI:1.112-1.586]) and the harmonic term cosday (OR 1.464 [CI:1.134-1.891]). We also noted that the household random effect had not been selected by the algorithm in this model, and we assumed that the in-house prevalence explains a big proportion of that previously existing random effect. This was confirmed by the AIC slightly increasing and the variance of the random effect becoming null when adding the random effect back into the best model. This model showed that for ESBL *E. coli*, living in a house with a high prevalence puts individuals at higher risk of colonisation and that that risk is increased during the wet season.

Table 3.8: Model ME1 : ESBL *E. coli* in Malawi

	Log-odds	**P-value**	**Odds ratio (95% CI)**
Intercept	0.174	0.625	1.189 (0.593-2.385)
Age	0.029	0.797	1.029 (0.827-1.281)
Being male (vs female)	-0.106	0.608	0.900 (0.601-1.347)
Reactive to HIV testing (vs non-reactive)	-0.430	0.310	0.650 (0.283-1.493)
Unknown HIV status (vs non-reactive)	-0.259	0.292	0.772 (0.476-1.250)
Use of amoxicilline in the last 6 months	-0.007	0.944	0.993 (0.812-1.214)
Use of cotrimoxazole in the last 6 months	-0.106	0.363	0.899 (0.716-1.130)
Use of other antibiotics in the last 6 months	0.066	0.539	1.068 (0.866-1.317)
Number of people living in the household	-0.134	0.322	0.875 (0.672-1.140)
Average household monthly income	-0.217	0.076	0.805 (0.633-1.023)
Living in Chikwawa (vs Chileka)	-0.605	0.105	0.546 (0.263-1.134)
Living in Ndirande (vs Chileka)	0.138	0.669	1.148 (0.610-2.163)
Number of days since the first sample*	0.542	**0.0002**	1.719 (1.290-2.291)
Harmonic term (sinday)	-0.090	0.717	0.914 (0.563-1.484)
Harmonic term (sinday2)	0.134	0.549	1.143 (0.738-1.772)
Harmonic term (cosday)	0.877	**0.0004**	2.403 (1.483-3.894)
Harmonic term (cosday2)	0.285	0.193	1.329 (0.866-2.041)

*Significant variables highlighted in bold.
Household random effect variance: 1.066

Table 3.9: Model ME2 : ESBL *E. coli* in Malawi

	Log-odds	**P-value**	**Odds ratio (95% CI)**
Intercept	-0.150	0.305	0.861 (0.647-1.146)
Number of days since the first sample*	0.574	**8.3e-06**	1.775 (1.379-2.284)
Harmonic term (cosday)	0.883	**0.0001**	2.418 (1.544-3.786)
Harmonic term (cosday2)	0.335	0.111	1.398 (0.926-2.112)

*Significant variables highlighted in bold.
Household random effect variance: 1.142

Table 3.10: Model ME3 : ESBL *E. coli* in Malawi

	Log-odds	**P-value**	**Odds ratio (95% CI)**
Intercept	-0.997	7.2e-11	0.369 (0.274-0.498)
In-house prevalence*	0.181	**2.9e-13**	1.199 (1.142-1.258)
Number of days since the first sample	0.284	**0.002**	1.328 (1.112-1.586)
Harmonic term (cosday)	0.381	**0.004**	1.464 (1.134-1.891)

*Significant variables highlighted in bold

3.3.1.1.2 ESBL-producing *K. pneumoniae* in Malawi

Using the same strategy for ESBL *K. pneumoniae*, we initially looked at the full model using all the variables we pre-selected as important variables to investigate (Table 3.11). For ESBL *K. pneumoniae*, we noted that no seasonal effect was previously detected therefore we removed the harmonic terms and only kept the sample date as a temporal variable. The full model did not find any significant risk factors for ESBL *K. pneumoniae* and found a household random effect variance of 1.053.

After using our selection algorithm to select the most important variables in the full model, none of the variables were chosen as significant and the step model became a null model with only the household random effect still present with a slightly higher variance of 1.169 (Table 3.12). The small difference can be explained by the removal of

all the fixed effects in the model.

Subsequently, we looked at models including the colonised people in the house and the prevalence. Looking at the AIC, we selected the best model out of the three models. For ESBL *K. pneumoniae* in Malawi, the best model was the last model (Table 3.13), which identified a unique risk factor, the number of positive people in the house (OR 1.697 [CI:1.325-2.174]). We also note that the household random effect has again not been selected by the algorithm in this model, and we can assume that the number of positive people in the house explains a big proportion of the household random effect. This model showed that for ESBL *K. pneumoniae*, living in a house with colonised people is the main risk factor.

Table 3.11: Model MK1 : ESBL *K. pneumoniae* in Malawi

	Log-odds	**P-value**	**Odds ratio (95% CI)**
Intercept	-2.047	2.9e-07	0.129 (0.059-0.282)
Age	0.107	0.453	1.112 (0.842-1.470)
Being male (vs female)	-0.186	0.494	0.830 (0.487-1.415)
Reactive to HIV testing (vs non-reactive)	0.110	0.836	1.116 (0.393-3.169)
Unknown HIV status (vs non-reactive)	0.308	0.340	1.360 (0.723-2.559)
Use of amoxicilline in the last 6 months	0.098	0.445	1.103 (0.858-1.418)
Use of cotrimoxazole in the last 6 months	-0.019	0.895	0.981 (0.734-1.310)
Use of other antibiotics in the last 6 months	-0.023	0.880	0.978 (0.731-1.308)
Number of people living in the household	0.035	0.825	1.036 (0.759-1.412)
Average household monthly income	0.100	0.469	1.105 (0.843-1.447)
Living in Chikwawa (vs Chileka)	-0.631	0.176	0.532 (0.213-1.326)
Living in Ndirande (vs Chileka)	-0.377	0.321	0.686 (0.326-1.444)
Number of days since the first sample	0.144	0.368	1.154 (0.844-1.578)

*Significant variables highlighted in bold.
Household random effect variance: 1.053

Table 3.12: Model MK2 : ESBL *K. pneumoniae* in Malawi

	Log-odds	**P-value**	**Odds ratio (95% CI)**
Intercept	-2.319	<2e-16	0.098 (0.064-0.151)

Household random effect variance: 1.169

Table 3.13: Model MK3: ESBL *K. pneumoniae* in Malawi

	Log-odds	**P-value**	**Odds ratio (95% CI)**
Intercept	-2.182	< 2e-16	0.113 (0.085-0.149)
Number of colonised people in the house*	0.529	**2.9e-05**	1.697 (1.325-2.174)

*Significant variables highlighted in bold

3.3.1.2 Ndirande

3.3.1.2.1 ESBL-producing *E. coli* in Ndirande

Focusing on ESBL *E. coli* in Ndirande, the full model (Table 3.14) identified similar significant variables to the model for all of the Malawian areas, such as the sample date (OR 1.625 [CI:1.141-2.316]) and the harmonic terms cosday (OR 3.105 [CI:1.383-6.971]) and cosday2 (OR 2.328 [CI:1.119-4.845]). The seasonality appeared even more important in this more localised area of Malawi. This might be partially explained by the fractional differences between seasons in each study area. All three areas are fairly close but Chileka is lower and hotter and Chikwawa is even lower down and drier than Ndirande. After using our selection algorithm, only the previously significant variables were kept and were still significant (Table 3.15). Their odds ratio were extremely similar to the previous model. The household random effect variance had slightly decreased, going from 1.037 to 0.980.

Afterwards, we looked at the third possible model. Looking at the AIC, we selected the best model out of the three models. For ESBL *E. coli* in Ndirande, the best model was the prevalence model (Table 3.16), which identified four important risk factors, the in-house prevalence (OR 1.160 [CI:1.082-1.244]), the harmonic terms cosday (OR 2.133 [CI:1.228-3.708]) and cosday2 (OR 1.749 [CI:1.026-2.980]) and the sample

date (OR 1.284 [CI:1.017-1.620]). Similarly to Malawi as a whole area, this model showed that for ESBL *E. coli* in Ndirande, living in a house with a high prevalence puts individuals at a higher risk of colonisation and that risk is increased during the wet season. Throughout this analysis for all areas, the household random effect variance was constantly captured by the in-house prevalence or the number of colonised people in the house.

Table 3.14: Model NE1 : ESBL *E. coli* in Ndirande

	Log-odds	**P-value**	**Odds ratio (95% CI)**
Intercept	0.335	0.343	1.398 (0.699-2.798)
Age	0.097	0.606	1.101 (0.763-1.590)
Being male (vs female)	-0.245	0.474	0.782 (0.400-1.531)
Reactive to HIV testing (vs non-reactive)	-0.904	0.180	0.405 (0.108-1.520)
Unknown HIV status (vs non-reactive)	-0.383	0.319	0.682 (0.321-1.449)
Use of amoxicilline in the last 6 months	0.089	0.575	1.093 (0.801-1.490)
Use of cotrimoxazole in the last 6 months	0.024	0.904	1.024 (0.695-1.510)
Use of other antibiotics in the last 6 months	-0.037	0.836	0.964 (0.680-1.367)
Number of people living in the household	-0.093	0.635	0.912 (0.622-1.336)
Average household monthly income	-0.104	0.567	0.901 (0.631-1.287)
Number of days since the first sample*	0.486	**0.007**	1.625 (1.141-2.316)
Harmonic term (sinday)	-0.059	0.879	0.943 (0.444-2.006)
Harmonic term (sinday2)	0.042	0.904	1.043 (0.529-2.057)
Harmonic term (cosday)	1.133	**0.006**	3.105 (1.383-6.971)
Harmonic term (cosday2)	0.845	**0.024**	2.328 (1.119-4.845)

*Significant variables highlighted in bold.
Household random effect variance: 1.037

Table 3.15: Model NE2: ESBL *E. coli* in Ndirande

	Log-odds	**P-value**	**Odds ratio (95% CI)**
Intercept	0.030	0.899	1.030 (0.650-1.632)
Harmonic term (cosday)*	1.228	**0.002**	3.415 (1.600-7.287)
Number of days since the first sample	0.461	**0.004**	1.586 (1.155-2.178)
Harmonic term (cosday2)	0.845	**0.019**	2.329 (1.151-4.711)

*Significant variables highlighted in bold.
Household random effect variance: 0.980

Table 3.16: Model NE3 : ESBL *E. coli* in Ndirande

	Log-odds	**P-value**	**Odds ratio (95% CI)**
Intercept	-0.725	0.003	0.484 (0.298-0.787)
In-house prevalence*	0.149	**3.3e-05**	1.160 (1.082-1.244)
Harmonic term (cosday)	0.758	**0.007**	2.133 (1.228-3.708)
Harmonic term (cosday2)	0.559	**0.040**	1.749 (1.026-2.980)
Number of days since the first sample	0.250	**0.035**	1.284 (1.017-1.620)

*Significant variables highlighted in bold

3.3.1.2.2 ESBL-producing *K. pneumoniae* in Ndirande

For ESBL *K. pneumoniae* in Ndirande, none of the risk factors were found significant by the full model (Table 3.17). The household random effect had a variance of 0.904. After using our selection algorithm to select the most important variables among the full model, none of the variables were chosen as significant and the step model became a null model with only the household random effect still present with a slightly higher variance of 1.052 (Table 3.18).

After looking at the AIC for a potential model with the number of colonised people or in-house prevalence, we selected the best model. For ESBL *K. pneumoniae* in Malawi, the best model was the third model (Table 3.19), which identified a unique risk

factor, the number of positive people in the house (OR 1.630 [CI:1.089-2.441]). This model showed that for ESBL *K. pneumoniae*, living in a house with colonised people is the main risk factor. These results are consistent with what we found when looking at all the Malawian areas as one area.

Table 3.17: Model NK1 : ESBL *K. pneumoniae* in Ndirande

	Log-odds	**P-value**	**Odds ratio (95% CI)**
Intercept	-2.370	2.3e-07	0.094 (0.038-0.229)
Age	0.047	0.847	1.048 (0.652-1.683)
Being male (vs female)	-0.040	0.929	0.961 (0.399-2.316)
Reactive to HIV testing (vs non-reactive)	0.533	0.475	1.705 (0.395-7.350)
Unknown HIV status	0.109	0.825	1.115 (0.425-2.930)
Use of amoxicilline in the last 6 months	-0.025	0.906	0.975 (0.640-1.486)
Use of cotrimoxazole in the last 6 months	0.095	0.702	1.100 (0.675-1.791)
Use of other antibiotics in the last 6 months	-0.182	0.525	0.834 (0.477-1.459)
Number of people living in the household	0.081	0.714	1.084 (0.704-1.671)
Average household monthly income	0.026	0.904	1.026 (0.675-1.560)
Number of days since the first sample	-0.001	0.998	1.000 (0.691-1.446)

*Significant variables highlighted in bold.
Household random effect variance: 0.904

Table 3.18: Model NK2 : ESBL *K. pneumoniae* in Ndirande

	Log-odds	**P-value**	**Odds ratio (95% CI)**
Intercept	-2.327	8.8e-11	0.098 (0.048-0.197)

Household random effect variance: 1.052

Table 3.19: Model NK3 : ESBL *K. pneumoniae* in Ndirande

	Log-odds	**P-value**	**Odds ratio (95% CI)**
Intercept	-2.146	<2e-16	0.117 (0.076-0.180)
Number of colonised people in the house*	0.489	**0.018**	1.630 (1.089-2.441)

*Significant variables highlighted in bold

3.3.1.3 Chikwawa

3.3.1.3.1 ESBL-producing *E. coli* in Chikwawa

In Chikwawa, the full model for ESBL *E. coli* (Table 3.20) identified harmonic terms cosday (OR 381.97 [CI:2.508-58179.19]) and sinday2 (OR 31.516 [CI:1.026-968.56]). It also showed a slight effect for people that have had antibiotics in the last six months (excluding amoxicillin and cotrimoxazole). The most frequently used antibiotics in this grouped category were metronidazole and gentamicin. After an increase in positive samples during the wet season in Chikwawa up to March 2020, there was a 5 month break before six more returned samples became available in August 2020 (COVID-19 pandemic). This caused the model to overestimate and create unrealistic estimates for the harmonic terms. After using our selection algorithm, only the previously significant harmonic term cosday (OR 4.868 [CI:1.927-12.294] was still significant (Table 3.21). The only other selected variable was having had cotrimoxazole in the last six months but it was not significant.

Finally, after looking at a third model for ESBL *E. coli* in Chikwawa, the best model was the prevalence model (Table 3.22), which identified two important risk factors, the in-house prevalence (OR 1.279 [CI:1.142-1.432]) and the harmonic term cosday (OR 2.286 [CI:1.198-4.359]). Similarly to previous findings, this model showed that for ESBL *E. coli* in Chikwawa, living in a house with a high prevalence puts individuals at a higher risk of colonisation and that risk is increased during the wet season. As for Malawi and Ndirande, the household random effect was no longer present in this model.

Table 3.20: Model KE1: ESBL *E. coli* in Chikwawa

	Log-odds	**P-value**	**Odds ratio (95% CI)**
Intercept	1.797	0.263	6.030 (0.260-139.68)
Age	-0.253	0.361	0.776 (0.451-1.337)
Being male (vs female)	-0.004	0.993	0.996 (0.402-2.470)
Reactive to HIV testing (vs non-reactive)	-0.215	0.793	0.806 (0.161-4.038)
Unknown HIV status (vs non-reactive)	0.125	0.829	1.133 (0.367-3.502)
Use of amoxicilline in the last 6 months	0.093	0.615	1.097 (0.764-1.575)
Use of cotrimoxazole in the last 6 months	-0.428	0.101	0.652 (0.391-1.088)
Use of other antibiotics in the last 6 months	0.303	0.100	1.353 (0.944-1.940)
Number of people living in the household	-0.598	0.152	0.550 (0.243-1.246)
Average household monthly income	-0.750	0.145	0.473 (0.172-1.295)
Number of days since the first sample	-1.203	0.184	0.300 (0.051-1.768)
Harmonic term (sinday)	2.900	0.106	18.182 (0.541-610.54)
Harmonic term (sinday2)*	3.451	**0.048**	31.516 (1.026-968.56)
Harmonic term (cosday)	5.945	**0.020**	381.97 (2.508-58179.2)
Harmonic term (cosday2)	0.271	0.618	1.312 (0.452-3.809)

*Significant variables highlighted in bold.
Household random effect variance: 1.316

Table 3.21: Model KE2 : ESBL *E. coli* in Chikwawa

	Log-odds	**P-value**	**Odds ratio (95% CI)**
Intercept	-0.604	0.045	0.546 (0.303-0.985)
Harmonic term (cosday)*	1.583	**0.001**	4.868 (1.927-12.294)
Use of cotrimoxazole in the last 6 months	-0.367	0.134	0.693 (0.428-1.120)

*Significant variables highlighted in bold.
Household random effect variance: 1.515

Table 3.22: Model KE3 : ESBL *E. coli* in Chikwawa

	Log-odds	**P-value**	**Odds ratio (95% CI)**
Intercept	-1.462	2e-06	0.232 (0.127-0.424)
In-house prevalence*	0.246	**2e-05**	1.279 (1.142-1.432)
Harmonic term (cosday)	0.827	**0.012**	2.286 (1.198-4.359)
Use of cotrimoxazole in the last 6 months	-0.299	0.164	0.742 (0.487-1.129)

*Significant variables highlighted in bold

3.3.1.3.2 ESBL-producing *K. pneumoniae* in Chikwawa

For ESBL *K. pneumoniae* in Chikwawa, the models raised some convergence issues created by the lack of positive samples for the HIV status variable. Considering HIV status had never appeared to be a significant risk factor in previous models, we made the decision to remove the HIV status variable from this analysis.

The full model did not identify any significant risk factors but found an extremely high variance of 34.05 for the household random effect (Table 3.23). This can be explained by the fact that out of only fifteen individuals colonised with ESBL-producing *K. pneumoniae*, eight of them were living in three specific households. After using our selection algorithm to select the most important variables among the full model, only having had cotrimoxazole in the last six months was selected and still no variables were found significant (Table 3.24). The variance for the household did decrease from 34.05 to 20.65.

For ESBL *K. pneumoniae* in Chikwawa, the best model was the third model (Table 3.25), which identified a unique risk factor, the number of positive people in the house (OR 3.594 [CI:1.868-6.916]). This model again showed that for ESBL *K. pneumoniae*, living in a house with colonised people is a major risk factor. These results are consistent with what we found when looking at all the Malawian areas as one area.

Table 3.23: Model KK1: ESBL *K. pneumoniae* in Chikwawa

	Log-odds	**P-value**	**Odds ratio (95% CI)**
Intercept	-7.905	0.006	0.0004 (1e-06-0.106)
Age	0.304	0.528	1.355 (0.527-3.481)
Being male (vs female)	-0.071	0.944	0.932 (0.131-6.657)
Use of amoxicilline in the last 6 months	0.458	0.191	1.580 (0.796-3.139)
Use of cotrimoxazole in the last 6 months	-0.684	0.125	0.505 (0.211-1.209)
Use of other antibiotics in the last 6 months	0.210	0.538	1.234 (0.633-2.405)
Number of people living in the household	-0.547	0.647	0.579 (0.056-6.014)
Average household monthly income	-0.030	0.964	0.970 (0.262-3.585)
Number of days since the first sample	-1.028	0.489	0.358 (0.019-6.595)

*Significant variables highlighted in bold.
Household random effect variance: 34.05

Table 3.24: Model KK2 : ESBL *K. pneumoniae* in Chikwawa

	Log-odds	**P-value**	**Odds ratio (95% CI)**
Intercept	-5.968	0.006	0.003 (4e-05-0.184)
Use of cotrimoxazole in the last 6 months	-0.562	0.184	0.570 (0.249-1.305)

*Significant variables highlighted in bold.
Household random effect variance: 20.65

Table 3.25: Model KK3: ESBL *K. pneumoniae* in Chikwawa

	Log-odds	**P-value**	**Odds ratio (95% CI)**
Intercept	-2.913	1.4e-15	0.054 (0.027-0.111)
Number of colonised people in the house*	1.279	**0.0001**	3.594 (1.868-6.916)

*Significant variables highlighted in bold

3.3.1.4 Chileka

3.3.1.4.1 ESBL-producing *E. coli* in Chileka

For ESBL *E. coli* in Chileka, some convergence issues emerged due to all seven people having had amoxicillin being non-colonised with ESBL *E. coli*. This issue was resolved by removing the variable from the analysis. The full model identified the harmonic term cosday (OR 1.971 [CI:1.000-3.884]) as significant (Table 3.26). After using our selection algorithm, only the sample date (OR 2.023 [CI:1.107-3.696]) was significant (Table 3.27).

Here, the best model was the prevalence model (Table 3.28), which identified one important risk factor, the prevalence (OR 1.168 [CI:1.069-1.275]). Considering previous models always showed a strong seasonal effect for *E. coli*, we had a closer look at Chileka. After an increase in positive samples during the wet season up to March 2020, the 6 month break before three more samples became available in September (COVID-19) caused the model to overestimate the importance of the sample date variable. Removing these samples returned similar results to the models here but increasing the importance of the harmonic terms and decreasing the importance of the sample date. This confirmed that seasonality is still an important effect in Chileka for ESBL *E. coli* as it is for the other Malawian areas. As previously, this model showed that for ESBL *E. coli* in Chileka, living in a house with a high prevalence puts individuals at a higher risk of colonisation and that risk is increased during the wet season.

Table 3.26: Model CE1 : ESBL *E. coli* in Chileka

	Log-odds	**P-value**	**Odds ratio (95% CI)**
Intercept	0.658	0.367	1.930 (0.463-8.043)
Age	0.094	0.577	1.099 (0.790-1.529)
Being male (vs female)	-0.094	0.777	0.910 (0.474-1.748)
Reactive to HIV testing (vs non-reactive)	0.283	0.723	1.328 (0.278-6.351)
Unknown HIV status	0.011	0.980	1.011 (0.433-2.358)
Use of cotrimoxazole in the last 6 months	0.071	0.725	1.074 (0.723-1.593)
Use of other antibiotics in the last 6 months	-0.189	0.486	0.828 (0.487-1.408)
Number of people living in the household	0.017	0.939	1.017 (0.664-1.558)
Average household monthly income	-0.257	0.225	0.773 (0.510-1.171)
Number of days since the first sample	1.155	0.056	3.173 (0.973-10.354)
Harmonic term (sinday)	1.103	0.083	3.012 (0.865-10.482)
Harmonic term (sinday2)	0.548	0.169	1.730 (0.792-3.776)
Harmonic term (cosday)*	0.678	**0.050**	1.971 (1.000-3.884)
Harmonic term (cosday2)	-0.310	0.368	0.733 (0.374-1.440)

*Significant variables highlighted in bold.
Household random effect variance: 0.644

Table 3.27: Model CE2 : ESBL *E. coli* in Chileka

	Log-odds	**P-value**	**Odds ratio (95% CI)**
Intercept	-0.234	0.371	0.791 (0.473-1.322)
Number of days since the first sample*	0.705	**0.022**	2.023 (1.107-3.696)

*Significant variables highlighted in bold.
Household random effect variance: 0.772

Table 3.28: Model CE3: ESBL *E. coli* in Chileka

	Log-odds	**P-value**	**Odds ratio (95% CI)**
Intercept	-0.905	0.001	0.404 (0.235-0.696)
In-house prevalence*	0.155	**0.001**	1.168 (1.069-1.275)
Number of days since the first sample	0.420	0.064	1.522 (0.976-2.375)

*Significant variables highlighted in bold

3.3.1.4.2 ESBL-producing *K. pneumoniae* in Chileka

For ESBL *K. pneumoniae* in Chileka, the sample date was identified as a significant risk factor (OR 2.463 (CI:[1.225-4.952]) (Table 3.29). The variance of the household random effect is 0.235. After using our selection algorithm to select the most important variables among the full model, the same variable was selected and significant (Table 3.30). The household random effect was also removed after variable selection, likely due to its low variance.

For ESBL *K. pneumoniae* in Chileka, adding the number of people in the house or the in-house prevalence did not affect the model selection and the previously selected model was deemed the best model. This model shows that for ESBL *K. pneumoniae*, having an unknown HIV status puts you at a higher risk of ESBL *K. pneumoniae* colonisation. There is also a linear temporal effect for this particular area of Malawi.

Table 3.29: Model CK1: ESBL *K. pneumoniae* in Chileka

	Log-odds	**P-value**	**Odds ratio (95% CI)**
Intercept	-2.308	0.001	0.100 (0.027-0.370)
Age	0.085	0.681	1.089 (0.726-1.634)
Being male (vs female)	-0.112	0.781	0.894 (0.407-1.966)
Reactive to HIV testing (vs non-reactive)	1.286	0.186	3.617 (0.537-24.352)
Unknown HIV status (vs non-reactive)	1.235	0.065	3.439 (0.926-12.777)
Use of amoxicilline in the last 6 months	0.158	0.538	1.171 (0.709-1.932)
Use of cotrimoxazole in the last 6 months	-0.081	0.752	0.922 (0.557-1.527)
Use of other antibiotics in the last 6 months	-0.010	0.974	0.950 (0.552-1.776)
Number of people living in the household	-0.046	0.832	0.955 (0.623-1.464)
Average household monthly income	0.263	0.214	1.301 (0.859-1.970)
Number of days since the first sample*	0.901	**0.011**	2.463 (1.225-4.952)

*Significant variables highlighted in bold.
Household random effect variance: 0.235

Table 3.30: Model CK2/CK3 : ESBL *K. pneumoniae* in Chileka

	Log-odds	**P-value**	**Odds ratio (95% CI)**
Intercept	-2.236	0.0003	0.107 (0.032-0.360)
Reactive to HIV testing (vs non-reactive)	1.335	0.138	3.799 (0.650-22.188)
Unknown HIV status	1.120	0.080	3.065 (0.876-10.720)
Number of days since the first sample*	0.748	**0.014**	2.113 (1.165-3.833)

*Significant variables highlighted in bold

The best model results for each combination ESBL/area are summarised in Table 3.31.

Table 3.31: Summary results of non-spatial models

		Ndirande	Chikwawa	Chileka	All
E. coli	Significant variables	In-house prevalence - Sample date - Harmonic terms	In-house prevalence - Harmonic terms	In-house prevalence	In-house prevalence - Sample date - Harmonic terms
	Notes		Over-estimation of the harmonic terms due to a 5-month break (COVID-19) in samples	Convergence issue due to all individuals having taken amoxicillin not being colonised / Over-estimation of the sample date and under-estimation of the harmonic terms due to a 6-month break in samples	
K. pneumoniae	Significant variables	Number of colonised people in the house	Number of colonised people in the house	Sample date	Number of colonised people in the house
	Notes		Convergence issue due to a lack of diversity in samples for the HIV status variable	Over-estimation of the importance of HIV status probably due to the high correlation between unknown HIV status and ESBL-positive samples	

3.3.2 Introducing a spatial component to the model

Using a Gaussian process, we introduced a spatial component to our models. For each area and ESBL type, we designed a model using Stan and included the predetermined variables found significant in each respective best model. We also added two variables to include longitude and latitude that we scaled to have a mean of zero and a standard deviation of one.

Looking at Ndirande, the maps in Figure 3.12 showed that the entire area was plain, with an approximate mean of zero for the projected gaussian process. Therefore, it appeared that the original strong spatial correlation we had detected early on in Ndirande for both ESBL *E. coli* (ESBL-Ec) and ESBL *K. pneumoniae* (ESBL-K) was now fully explained by household-level covariates such as in-house prevalence or number of colonised people in the house. The parameter estimates for those variables were similar to the ones found with the best model. All parameter estimates for the spatial models can be found in Appendix B.3.

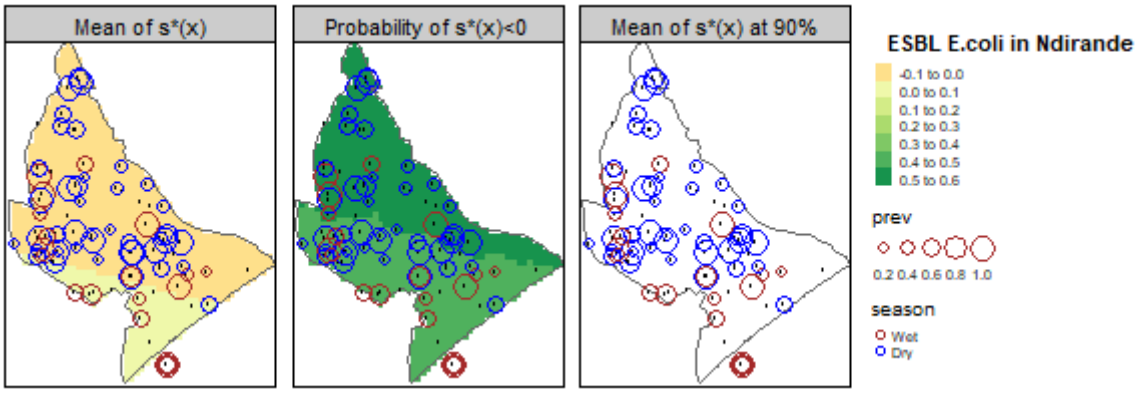


Figure 3.12: Gaussian process maps for Ndirande (ESBL-Ec)

In order to make sure our prior choices were appropriate, we changed ϕ multiple times. Posteriors for our Gaussian process parameters can be found in Table 3.32.

Considering the choice of an inverse Gamma distribution for the prior of ϕ (values close to zero), we noted that the posterior for ϕ was extremely different from its prior. We also saw in Figure 3.13 that the prior and posterior for σ^2 were the same with a mean value of two, therefore the model did not bring out any information on the variance

of the spatial process that was still very low. Moreover, with a median length scale of 849 kilometers, the model concluded that there was no evidence of spatial correlation.

Table 3.32: Estimates for the Gaussian process parameters in Ndirande (ESBL-Ec)

	Median	**Mean**	**Std error**	**2.5%**	**97.5%**
ϕ	849.42	3976.56	6e+04	173.27	1.8e+04
σ^2	1.69	2.01	1.42	0.24	5.58
τ	3.44	3.80	2.18	0.67	9.01

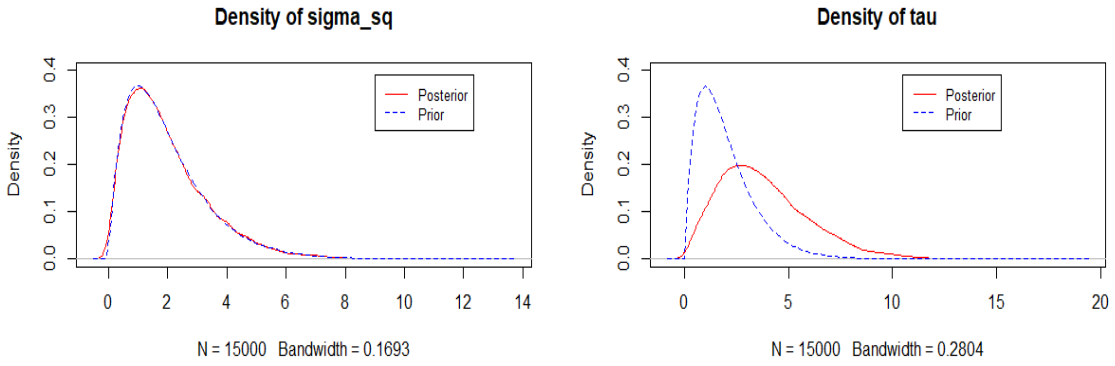


Figure 3.13: Density plots of σ^2 and τ for Ndirande (ESBL-Ec)

Due to the similarity in results for both ESBL types and no sign of spatial correlation for any of the areas, the other parameter estimates, diagnostics and maps can be found in Appendix B.3.

3.4 Discussion

In this chapter, we aimed at investigating the risk factors for human gut mucosal colonisation with ESBL-producing *E. coli* and *K. pneumoniae* in our three Malawian sites: Ndirande (urban), Chileka (peri-urban) and Chikwawa (rural). Using logistic models, we looked into individual-level covariates such as age and gender, HIV status and antibiotic use and household-level covariates such as household income and density. We also explored a potential effect of seasonality and of sharing a household with other colonised individuals by looking at the number of colonised people or the prevalence in the house.

Recent reviews on ESBL-producing Enterobacteriaceae in Africa have estimated a prevalence of 18% to 22%[3, 4]. The prevalence we found for ESBL-producing Enterobacteriaceae, especially for ESBL-producing *E. coli* is much higher than these estimates which is consistent with a continued increase of their prevalence over time[4]. It is also substantially higher than in the few community studies that have taken place in sub-Saharan Africa[101, 174]. However, it is consistent with findings from recent hospital studies in the region[3]. Due to their nosocomial nature, ESBL-producing Enterobacteriaceae were initially thought to originate within the hospital setting and transmit towards the community via patients. However, this increase in prevalence over time in the community suggests that there is a possible reversal of the situation with members of the community getting infections due to their colonisation and bringing it back to the hospital.

Overall, the antibiotic use at the baseline visit was determined by whether participants had been given antibiotics in the last six months. The results show that the reported numbers are relatively low with 6.5% of the participants given amoxicillin, 6.6% given cotrimoxazole and 5.7% given other antibiotics. When looking at a general antibiotic use (any of the antibiotic groups), 15.5% of the participants were given a course in the last six months. We note that the antibiotic use was determined by asking the participants whether or not they were given or received antibiotics. Consequently, there could have been issues with the reporting which could have created a reporting bias. CPT could also have been included in the antibiotic variables, however, while CPT tends to be a permanent treatment, we focused on recent previous use of antibiotics. Moreover, due to the extremely high correlation between CPT and HIV status and the small number of individuals on CPT, we chose to only keep HIV status. We assumed that being reactive to HIV testing would capture the same information as being on CPT.

In the case of ESBL *E. coli*, a marked seasonality was present for all individual study regions in Malawi, and for all study regions combined. The evidence in Figure 3.11 highlighted a higher prevalence during the wet season (November to April) than during the dry season (May to October). We can assume that heavy rain during the rainy season causes more water to be accumulated, creating mud and floodwater, which might lead to more contact between individuals and contaminated soil or water. Additionally, social behaviours caused by heavy rain or flooding, such as indoor crowding, might increase

the risk of transmission.

For ESBL *E. coli*, we consistently observed in Malawi and each specific area that living in a house with a high existing prevalence put individuals at higher risk of colonisation. Similarly, for ESBL *K. pneumoniae*, except for Chileka, we consistently observed that living in a house with a high number of colonised people is the main risk factor. Prevalences for both ESBL-producing *E. coli* and ESBL-producing *K. pneumoniae* vary slightly depending on the area. The prevalence was consistently higher in the urban areas and lower in the rural area. This could simply be due to the higher density of population in the urban areas, and as we have shown, the higher population density within the houses located in urban areas.

We note that a limitation in this analysis is the use of an algorithm that selects one single best model over other models that might be of similar fit. We added the full model as a way to give unbiased parameter estimates for our variables and we acknowledge that the use of such selection procedures is not always ideal. We also acknowledge that multiple testing is a potential concern, especially for coefficients where $p \approx 0.05$. However, all our models (full, after selection and best) give similar results in terms of the importance of seasonality for ESBL-producing *E. coli* and adding the prevalence or number of colonised people in the house consistently show the importance of sharing a household with other colonised people. Additionally, consistent effects all have very low p-values thus the chance of type I error is reduced.

In Chileka, HIV status appeared as a slightly significant effect (<0.1), however Chileka is the only area where the majority of individuals (73%) were of unknown status. Among the participants with positive samples for ESBL *K. pneumoniae*, 83% have unknown HIV status. This might have caused the model to overestimate the importance of HIV status in this area, by declaring that having an unknown HIV status puts you at a higher risk than having a determined status. Further research should use a better estimate of HIV status to explore its association with ESBL-producing Enterobacteriaceae colonisation. Additionally, due to lack of diversity in samples in some variables in specific areas, two variables were removed at different occurrences to allow the model to converge properly. In Chikwawa and Chileka, we note that the 6-month break between sample accessibility and the low amount of added samples caused the model to overes-

timate the importance of the sample date. Overall, the low number of people colonised with ESBL-producing *K. pneumoniae* creates difficulty when trying to explore the relationship between colonisation and other variables in our models.

The geostatistical model showed that no evidence of spatial correlation was found in any of the areas with an unreasonably large length scale of spatial correlation. Adding an informative prior for the length scale did not bring out any further information, suggesting that our study failed to detect the presence of a spatial correlation as it previously has in other similar studies[130, 160]. However, we can not be assured that such correlation is non-existent as it might just have been too short-scaled to be detected by our design. We have found that what happens within the household is important in these areas, therefore the scale of spatial correlation might have been smaller than what we expected.

Finally, when selecting our best model, we found that within-house prevalence or amount of colonised people accounts for a large proportion of the between-house variance, removing the need for the household-level random effect. These results highlight the importance of within-household transmission, and therefore household-level interventions and/or interventions focused on within-household behaviour may be very promising.

Chapter 4

Individual and WASH risk factors of ESBL-producing *E. coli* and *K. pneumoniae* colonisation over time in Malawi

4.1 Introduction

Whilst studies have shown that the prevalence of ESBL-producing Enterobacteriaceae in sSA is high[3, 4], little is known about asymptomatic colonisation with ESBL-producing Enterobacteriaceae. Learning more about asymptomatic colonisation in the community is crucial in order to prevent transmission, and as a consequence, reduce symptomatic infections. Prior to 2016, no studies described risk factors for ESBL-producing Enterobacteriaceae in sSA[3]. Since then, a few studies have described risk factors in the community setting. Recent antibiotic use (in the last weeks to months) has been found as a risk factor in several of these studies[101, 104, 105]. Other risk factors such as older age and previous hospital admission[101] were identified. One study found a positive association between older age and higher prevalence of ESBL colonisation[101], and one found that higher income is associated with a higher prevalence of ESBL colonisation[108]. How-

ever, most of these studies focused on a specific population within the community and not the general population. We cannot be assured that the risk factors detected in these specific populations would be the same throughout the general population. This highlights the need for a community study within the general healthy population that could help confirm or identify risk factors for human gut mucosal colonisation with ESBL-producing Enterobacteriaceae. Moreover, although it is recognised that WASH has an effect on transmission[32], risk factors related to WASH are still not well known. Only one study found that having private inside access to drinking water was positively associated with ESBL colonisation[175].

To be able to reduce colonisation and transmission in East and Southern Africa, we need to explore the dynamics of colonisation in this particular context. An understanding of what happens once an individual is colonised is needed to tailor appropriate interventions. We previously showed the importance of within-household transmission and what is required now is to understand how that transmission is impacted by the duration of the colonisation. If we can determine how long a specific household is at risk of colonisation once one member is colonised, this could help understand better the temporal dynamics of within-household transmission, and help inform the design of public health policies that interrupt transmission of AMR-bacteria. Such interventions are highly likely to be impactful at interrupting transmission of enteric bacteria more broadly.

The aim of this chapter is to perform rigorous examination of the WASH risk factors for ESBL colonisation in different socioeconomic settings in Malawi, accounting for temporal variability. Herein, we describe findings from a year and a half of longitudinal study using microbiological and household surveys. The WASH infrastructure was ascertained via checklist and sanitation inspection forms. This chapter includes repeated measurements in order to provide more precise information about fluctuations in ESBL colonisation status. After the COVID-19 "break", more information, especially about the WASH infrastructure, became available to us which allowed us to evaluate the influence of WASH factors on ESBL colonisation in a community-based setting. Finally, it will allow for a comparison between the risk factors identified in the previous chapter and the ones that will be identified here.

We fit a generalised linear mixed model to identify WASH risk factors and to determine their effect on ESBL colonisation. We also use a squared exponential correlation structure to explore the effect of time on ESBL colonisation status. We showed in the previous chapter that at baseline, the household appears to be important in driving transmission and that is where the interventions should take place. In this chapter, we include a temporal random effect at the household-level that already captures the variation previously captured by the household-level covariates describing the sharing of a household with colonised people. A better understanding of how the WASH context of the different communities impacts ESBL colonisation and transmission could improve public health responses and detect potential strategies for effective intervention and control in similar communities.

4.2 DRUM data

To study the longitudinal aspect of the ESBL distribution in the Malawian study areas, we used three types of data from the DRUM database: individual, household and laboratory results data.

4.2.1 Household data

In order to investigate WASH practices at the household-level, household WASH covariates were collected in the following ways. Reported variables, such as presence of a toilet at the household, were based on questions asked to the study participants during the baseline assessment visit whilst observed variables, such as the type of toilet, were answered by the field teams observing the household infrastructure at multiple time points. These variables were screened for importance by Tracy Morse, Kondwani Chidwisano and Derek Cocker, accounting for pre-existing knowledge on the risks and critical control points for faecal-oral transmission.

Twenty-six variables were selected and went through cleaning and processing with the help of the WASH team. Among these variables, three variables were removed due to a lack of variation. The remaining twenty-three were either kept as is or modified

to better examine the effect of those WASH factors on ESBL colonisation. The final lists of seventeen reported household variables and fourteen observed household variables, with any modifications, can respectively be found in Appendix C.1 and C.2. These variables were then merged into one household dataset consisting of both types of variables including data for three-hundred households. One household was excluded due to having no observed information and twenty-six households were excluded due to having no enrolment information. Household-level covariates such as household income and size were also included.

4.2.2 Individual data

Whilst the DRUM study focuses mainly on the transmission at the household-level[102], individual-level covariates such as age, gender, HIV status and antibiotic use were also selected. We previously found low reported levels of the different types of antibiotics we were interested in therefore we changed the way we include antibiotic use in our analysis. Antibiotic use is now defined as the reported use of any antibiotics in the last six months (at baseline) and subsequently, for each follow-up visit, as the reported use of antibiotics between that visit and the previous one. This variable was constructed by combining all the available antibiotic variables present in the questionnaire. A more detailed examination of the impact of antibiotic variables could have been undertaken however, due to the proportion of antibiotic use found in the previous chapter being close to 15%, it seemed unrealistic to look at the impact of individual groups of antibiotics on ESBL colonisation.

Additional variables such as household ID, individual ID and visit number were also included in this dataset. Four observations were removed due to missing data on age and gender of the individuals, eighteen duplicates were removed and visit date corrections were applied on 337 observations. After cleaning and processing, 2908 observations were included in the dataset, containing information at baseline and at follow-up visits for the individuals.

4.2.3 Human laboratory results

Out of 2852 human samples collected by the field teams over time, twenty-seven did not have a returned sample, two had inconsistencies with their household ID and fifty-one duplicates in individual ID and visit number were removed. Therefore, human laboratory results for ESBL *E. coli* and *K. pneumoniae* colonisation were available for 2772 samples. Additional covariates such as sample ID, individual ID, household ID, visit number and sample date were kept for the purpose of linking the datasets together.

4.2.4 Data linking

Considering our response variable was human colonisation with ESBL-producing bacteria, we only kept the samples for which we have laboratory results and complete household and individual information. When merging the three previously described datasets, a hundred and ninety-five samples were removed due to a lack of information on the individual at various times. Eighty-four samples, coming from households that were situated over 200 meters outside of the polygon limits, were also removed. This threshold was kept in order to include households that were subsequently chosen by the field teams due to a refusal from the original sampled household. After joining these three datasets, the combined dataset, which has been used for the following analysis, contained 2493 samples from 894 individuals in 259 households and 50 variables. The complete list of variables with the rationale for including them can be found in Table 4.1 and the list of variables with their detailed description can be found in Appendix C.3.

Table 4.1: Rationale for the covariates

Variable	**Rationale**
HIV status	Risk factor: Immunosuppression due to HIV infection can lead to gut dysbiosis and long term antibiotic prophylactic therapy and consequently susceptibility to acquisition and long-term carriage of resistant bacteria
Recent use of antibiotics	Risk factor: Antibiotic use leads to the depletion of drug-susceptible gut-commensal bacteria, providing a selective advantage to resistant bacteria behind to reproduce.
Age	Risk factor: Infants are more likely to enter in contact with potentially contaminated environments. Older adults require access to healthcare more frequently, increasing the risk of exposure to contaminated healthcare settings.
Sex	Protective: Women are more likely to get into contact with contaminated environment while doing the housework and taking care of the children.
Number of people living in the household	Risk factor: More people living in the household, more contact and proximity, more opportunities for within-household transmission.
Average household monthly income	Other: Income allows for better access to WASH infrastructure (protective) but also to antibiotics (risk factor).
Having children of school age	Risk factor: Children in school have more opportunities for spreading as they have more contacts.
Presence of a toilet in the household	Protective: The presence of a toilet avoids the use of a shared/public toilet and separation of people from faecal matter

Open defecation	Risk factor: Open defecation relates to inadequate WASH practice, increasing risk for transmission through the environment.
Sharing the toilet with non-household members	Risk factor: Sharing the toilet with non-household members increase the number of contacts and therefore the risk of colonisation.
Presence of a disposal mechanism for animal waste	Protective: If no disposal mechanism is available, disposal in the environment is likely, leading to potential environmental contamination.
Eating street food	Risk factor: Increases the risk of transmission through contaminated food products.
Eating from shared plates	Risk factor: Increases the number of contacts through sharing of the same plate.
Having a pipe as drinking water source	Risk factor / Protective: Having access to piped water reduces the risk brought by the use of a communal tap, which allows for more contacts between the tap and multiple users.
Having a communal tap as drinking water source	
Having a tube well/borehole as drinking water source	
Use of alternative water for cleaning utensils	Risk factor: The use of alternative water from a contaminated water source such as a river could increase the risk.
Owning birds	Risk factor: Animals defecate freely in the environment, increasing the risk of contact between individuals and faeces. However, pigs also eat waste, therefore could show a protective effect.
Owning cattle, goats or sheep	
Owning dogs or cats	
Owning pigs	

Keeping animals inside	Risk factor: Permitting animals inside allows for contamination of the household environment with animal faeces.
Contact with river water	Risk factor: Contact with river water increases the risk of transmission if the river water is contaminated.
Contact with drains	Risk factor: Contact with drains increases the chance of contact between individuals and faeces.
Toilet type	Risk factor / Protective: Having a pit latrine allows for separation of people from faecal matter while having a shared toilet increases your risk of contact with other individuals faeces.
Toilet floor material	Risk factor: Concrete or wood offer better protection against contaminated soil.
Having a drop hole cover on the toilet	Protective: A drop hole cover allows for reducing the flies access to faeces, therefore reducing further contamination of the environment.
Presence of toilet paper in the toilet	Protective: Cleaning materials allow for better WASH practice, reducing the risk of faecal-oral transmission.
Presence of newspaper/paper in the toilet	Protective: Cleaning materials allow for better WASH practice, reducing the risk of faecal-oral transmission.
Visible human faeces around the household	Risk factor: Seeing human faeces around the household suggests inappropriate WASH behavioral practices that might increase the risk of colonisation.
Presence of handwashing facilities (hwf) in the household	Protective: The presence of such facilities allows for handwashing after going to the toilet, decreasing the risk of faecal-oral transmission.
Frequency of soap presence in handwashing facilities	Protective: Soap allows for better protection against bacteria when washing hands.

Storing water covered	
Storing water uncovered	Risk factor / Protective: Storing water uncovered allows for potential contamination through animal contact, while covering it reduces that risk.
Storing water in a container with lid/tap	
Contact between animals and food areas	Risk factor: Increases the risk of transmission through contaminated food.
Visible animal faeces around the household	Risk factor: environmental contamination with faecal matter suggest a higher chance of contact with faeces.
Presence of standing water around the household	Risk factor: Resistant bacteria are frequently detected in water sources, therefore presence of standing water around the household could increase the risk.
Number of days since the first sample	Risk factor: Rising levels of AMR around the world suggest a higher chance of getting colonised with AMR bacteria with time.
Harmonic terms (sinday, cosday, sinday2, cosday2)	Risk factor / Protective: Higher risk during the wet season than the dry season.
Study area	Risk factor: Chikwawa is the rural area, with a higher presence of animals and potentially antibiotic use for animals. Ndirande is a high-density settlement, with potential for higher within-household transmission.

4.3 Methods

4.3.1 Exploratory analysis

Although the number of households and individuals in Ndirande and Chileka was higher than in Chikwawa, the number of available samples was fairly similar with 36% in Chikwawa and Chileka and 28% in Ndirande. The distribution of samples, individuals and households per polygon can be seen in Table 4.3.

Table 4.3: Distribution of the number of households, individuals and samples per polygon

	Ndirande	**Chikwawa**	**Chileka**
Households	96	64	99
Individuals	285	259	350
Samples	709	891	893

We initially aimed for the microbiological sampling to take place four times over six months, although compromises had to be made due to the COVID-19 break and some samples had to be delayed. Therefore, after data cleaning, each individual had one to four samples. The distribution of available samples per individual can be found in Table 4.4.

Table 4.4: Distribution of the number of samples available per individual

Samples	**1**	**2**	**3**	**4**
Individuals	233	96	192	373

More than half of the participants had three to four samples while only 10% had only two samples. The remaining two hundred and thirty-three individuals only had one sample. Among these samples, seventy-six percent were first visit samples that were added at the end of the study after the COVID-19 break. Out of 894 individuals, 95%

had a first visit sample, 69% had a second visit sample, 64% had a third visit sample and 51% had a fourth visit sample as can be seen in Table 4.5. The distribution of these samples over time and by polygon can be seen on Figure 4.1. We noted that those late added first visit samples came mainly from Chileka and Ndirande, which explains their higher number of first visit samples compared to Chikwawa.

Table 4.5: Distribution of the number of samples per visit

Visit time	**First visit**	**Visit 2**	**Visit 3**	**Visit 4**
Samples	851	616	570	456

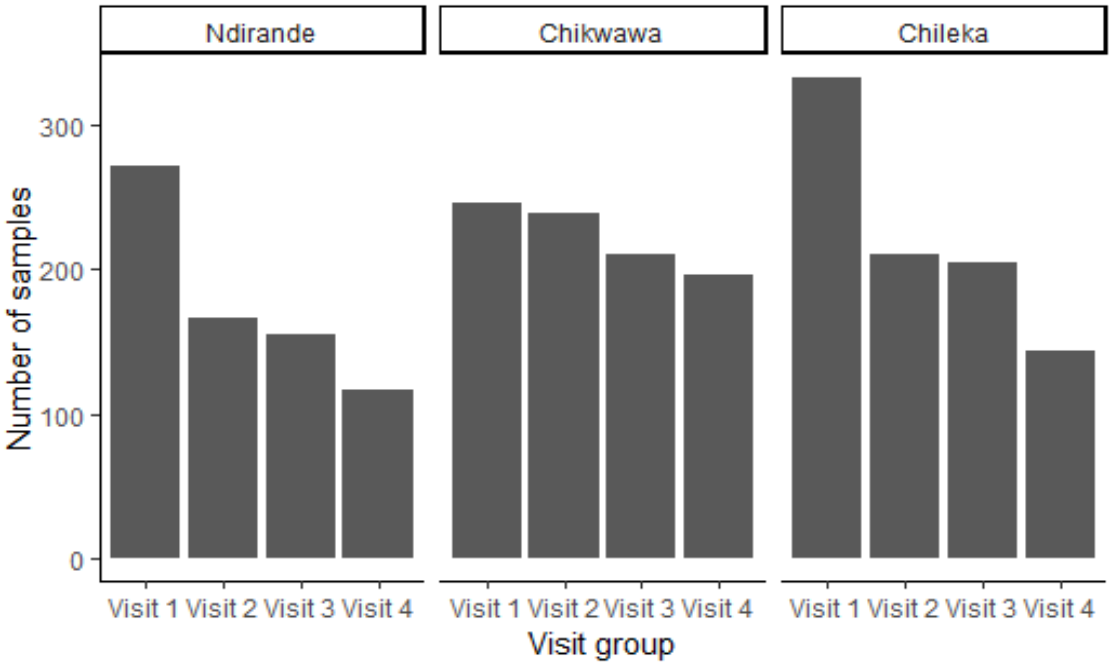


Figure 4.1: Distribution of samples per visit time and polygon

As Figure 4.2 shows, 55.1% of the individuals were under the age of 20 years old, 36.2% were between 21 and 50 years old, 7.3% were between 51 and 70 years old and the last 1.3% were over 70 years old. It should be noted that adults were considered to be over 15 years old (54% of the individuals), and school age was considered to be older than 5 years old and up to 15 years old (27.1%). Over all the individuals, 57% were

female with slightly varying proportions in each age group. 65.2% of the adults (>15) were female and 46.7% of the school age children were female.

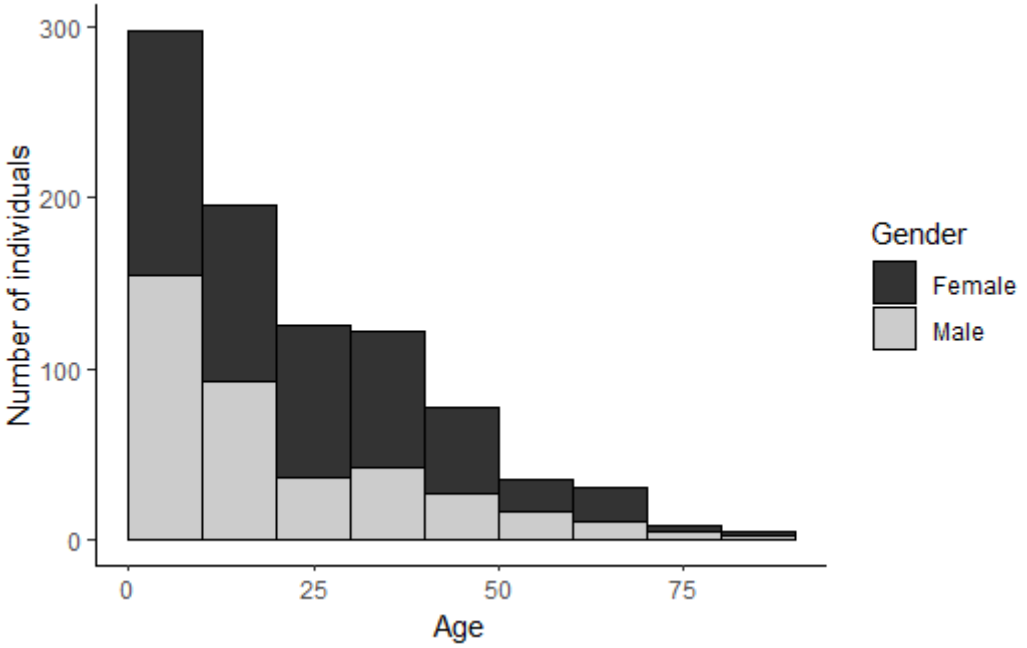


Figure 4.2: Distribution of age and gender for the 894 individuals

At the first visit, 15.2% of participants reported having taken at least a course of antibiotics in the last six months, while between subsequent visits, 6%, 9.4% and 8.3% reported at least a course of antibiotics. Overall, in Chikwawa, participants who reported having received at least a course of antibiotics were 15.4% whilst in Ndirande and Chileka, only 9.3% and 6.2% of participants respectively reported it.

Overall, the prevalence of ESBL-producing *E. coli* in our samples was 37% and the prevalence of ESBL-producing *K. pneumoniae* was 11.9%. At the first visit, 310 were positive for ESBL-producing *E. coli* (36.4%) and 100 were positive for ESBL-producing *K. pneumoniae* (11.8%). At the second visit, 203 were positive for ESBL-producing *E. coli* (33%) and 72 were positive for ESBL-producing *K. pneumoniae* (11.7%). At the third visit, 216 were positive for ESBL-producing *E. coli* (37.9%) and 57 were positive for ESBL-producing *K. pneumoniae* (10%). At the last visit, 193 were positive for ESBL-producing *E. coli* (42.3%) and 67 were positive for ESBL-producing *K. pneumoniae* (14.7%). The

prevalences can be found in Table 4.6.

Correlations between the WASH (numerical) variables can be visualised on the heatmap in Figure 4.3. A strong positive correlation was noticeable between multiple animal factors. Bird owners appeared to be more likely to keep animals inside, therefore also more likely to have said animals come into contact with food areas. Keeping animals inside also increased the likelihood of visible animal faeces around the household area.

In terms of sanitation, the data suggested that the higher the income is, the more likely the household's water drinking source is coming from a pipe and not from a tube or a well. Moreover, the higher the income of a household is, the more likely hand washing facilities and soap were present in the house, and cleaning materials such as toilet paper were present near the toilet.

Food factors such as eating from shared plates appeared to be negatively correlated with the previously mentioned sanitation factors which were themselves positively correlated with income. It seemed that the higher the income of the household is, the less chance individuals used shared plates when eating.

Table 4.6: Prevalence of ESBL-producing *E. coli* and ESBL-producing *K. pneumoniae* over time

Visit	ESBL	Positive (P)	Total (T)	(P/T)*100
First visit	*E. coli*	310	851	36.4%
	K. pneumoniae	100	851	11.8%
Visit 2	*E. coli*	203	616	33.0%
	K. pneumoniae	72	616	11.7%
Visit 3	*E. coli*	216	570	37.9 %
	K. pneumoniae	57	570	10.0%
Visit 4	*E. coli*	193	456	42.3%
	K. pneumoniae	67	456	14.7%

In conclusion, these correlations suggested that the socioeconomic status of the household greatly influences the WASH situation of the household. Having a higher income allows for a better access to cleaner water and easier availability of sanitation and hygiene products.

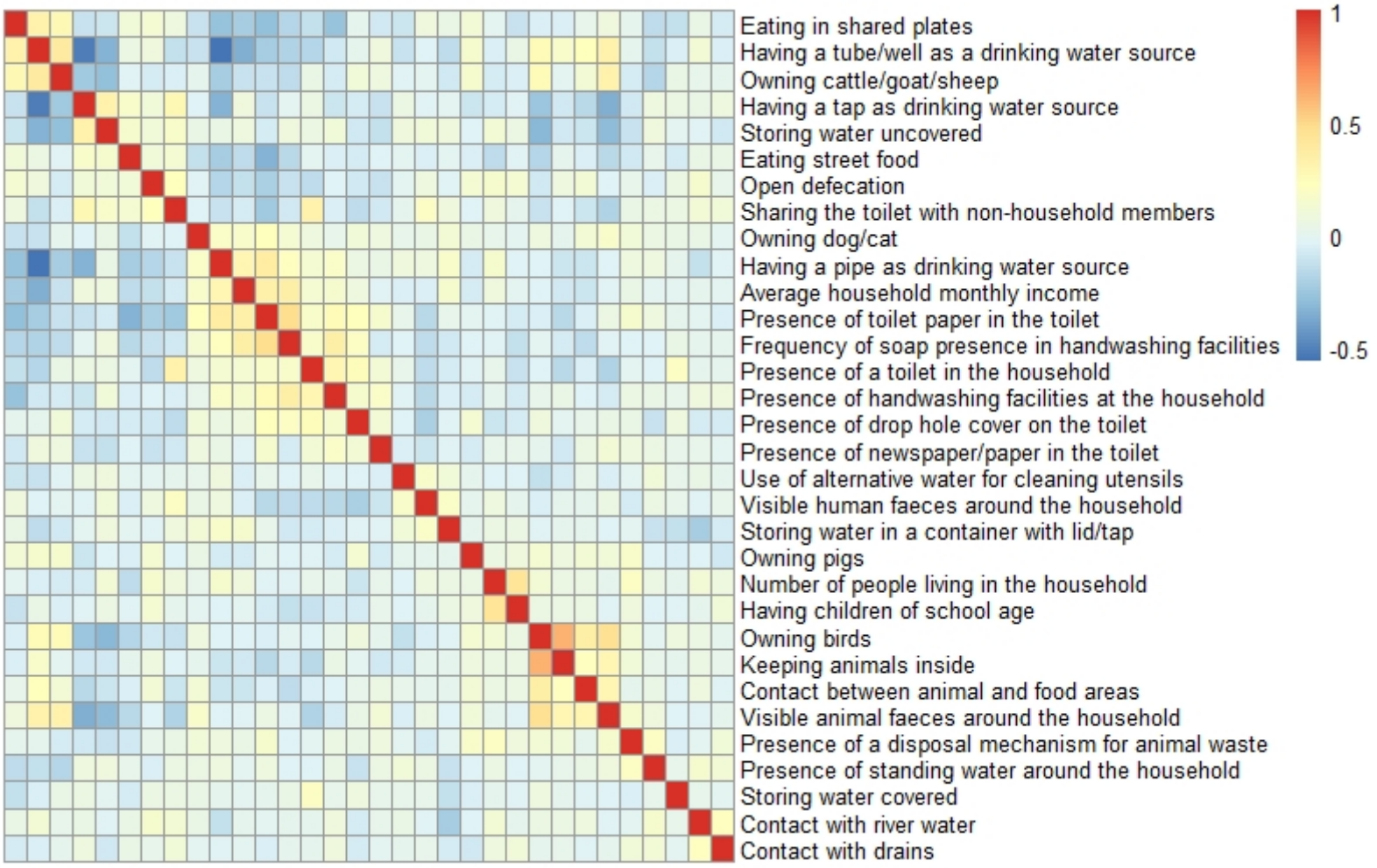


Figure 4.3: Correlation heatmap of household-level covariates

4.3.2 Impact of WASH on ESBL colonisation

In order to start exploring how WASH variables affect the participants colonisation with either ESBL-producing *E. coli* or ESBL-producing *K. pneumoniae*, we first ran univariable models for each variable. However, we noticed that the study area appeared to be affecting the result of this analysis. By acting as a confounder, WASH variables that vary depending on the study area the participant was in had a different signal if the study area was included as a variable in the analysis. Therefore, we first ran a generalised linear model including only the study area for both ESBL-producing *E. coli* and ESBL-producing *K. pneumoniae*. These results can be visualised in Tables 4.7 and 4.8. The reference level for each categorical variable with more than 2 levels throughout this chapter can be found in Appendix C.3. We found that there was a significant effect from the study area on ESBL-producing *E. coli*, with a higher risk of being colonised in Ndirande (OR 1.39 CI[1.13-1.70]) compared to Chileka. However for ESBL-producing *K. pneumoniae*, this was not the case. There was no sign of a significant effect of the study area on the colonisation status. This was verified by adding the study area to see how it affected the univariable models for ESBL *K. pneumoniae*. Consequently, univariable analysis was run differently depending on the bacterial species. The study area was added as a covariate when running univariable models for ESBL-producing *E. coli* but was not included for ESBL-producing *K. pneumoniae*. The harmonic terms were always run as a single term throughout this analysis. The numerical variables were standardised to allow an easier comparison with the temporal model, therefore the odds ratio should be interpreted as a change for each increase in standard deviation. For both the ESBL-producing *E. coli* models and the ESBL-producing *K. pneumoniae* models, we set a p-value threshold of 0.2 to select which variables to include in the temporal models. It was a pragmatic decision made to avoid missing the identification of important variables.

Table 4.7: Relationship between the study area and the ESBL-producing *E. coli* colonisation status

	Log-odds	**P-value**	**Odds ratio (95% CI)**
Intercept	-0.607	<2e-16	0.545 (0.475-0.625)
Living in Chikwawa (vs Chileka)	-0.061	0.540	0.941 (0.774-1.144)
Living in Ndirande (vs Chileka)*	**0.326**	**0.002**	**1.385 (1.132-1.696)**

*Significant variables highlighted in bold

Table 4.8: Relationship between the study area and the ESBL-producing *K. pneumoniae* colonisation status

	Log-odds	**P-value**	**Odds ratio (95% CI)**
Intercept	-2.105	<2e-16	0.122 (0.099-0.151)
Living in Chikwawa (vs Chileka)	0.196	0.183	1.216 (0.912-1.622)
Living in Ndirande (vs Chileka)	0.098	0.536	1.103 (0.809-1.504)

4.3.3 Modelling ESBL colonisation over time

Let y_{ijt} be 1 if the individual i tested at household j at time t is colonised. Our model can be expressed as:

$$\text{logit}(p_{ijt})) = \alpha + (x_{ij})^T\beta + \theta_1\cos(\frac{2\pi t}{365T}) + \theta_2\sin(\frac{2\pi t}{365T}) + \theta_3\cos(\frac{2\pi t}{365T}) + \theta_4\sin(\frac{2\pi t}{365T}) + u_{jt}$$

where Y_{ijt} follows a Bernoulli distribution with probability p_{ijt} i.e.

$$Y_{ijt} \sim \text{Bernoulli}(p_{ijt})$$

α is the intercept, x_{ij} are the household-level and individual-level explanatory variables, β are the regression coefficients for the fixed effects, $\theta_1\cos(\frac{2\pi t}{365T}) + \theta_2\sin(\frac{2\pi t}{365T}) + \theta_3\cos(\frac{2\pi t}{365T}) + \theta_4\sin(\frac{2\pi t}{365T})$ the annual and bi-annual harmonic terms, with $T = (1, \frac{1}{2})$ the period in years. In order to look at the temporal correlation between different time points, we included a temporally-correlated random effect u_{jt} at the household-level with covariance structure:

$$\text{cov}(u_{jt}, u_{jt+s}) = \sigma^2 e^{-\frac{s^2}{\phi^2}} + \tau^2 \tag{4.1}$$

where s is the distance (in time) between two time points, ϕ is the scale of temporal correlation, σ^2 is the variance of the temporal process, τ^2 is the nugget effect. As in Chapter 3, we made the decision to use a Bayesian framework to allow for more flexibility in designing our probability model. The model was fitted using the standard implementation of the No-U-Turn Sampler (NUTS) in the Stan modelling language [167] on R through the package RStan [168]. The code for implementing this model in Stan can be found in Appendix C.4. The model was run with twenty thousand iterations for each of the three Markov chains. Convergence was evaluated by inspection of traceplots and the Gelman-Rubin statistic being close to 1. Posterior estimates of parameters were expressed as medians with 95% credible intervals. Prior distributions were chosen as follows:

$$\alpha \sim \text{Normal}(0, 100)\ \beta \sim \text{Normal}(0, 10)$$

$$\sigma^2 \sim \text{Gamma}(2, 1)\ \tau \sim \text{Gamma}(2, 1)$$

$$\phi \sim \text{Gamma}(4, 0.125)$$

The prior distribution for ϕ was based on recent work on the dynamics of gut mucosal colonisation with ESBL-producing Enterobacteriaceae in Malawi where they found an estimated mean time of 43 days between the ESBL-E colonised and uncolonised states[173].

We initially considered adding the random effect at the individual-level, however our reasoning was that we previously suggested the importance of within-household transmission, therefore we were more interested to see if there is temporal variability at the household-level. Moreover, we had a look at adding an individual-level random effect in the model first but it resulted in a low variance (<0.4) and the model found a resulting correlation range of approximately 3 days. Considering that for this analysis, our individuals tend to have at least a month between two visits, these results showed that we could consider samples from one individual as independent. We then decided to focus this analysis on the hundred and twenty-nine households with available sample results for all four time points.

4.4 Results

4.4.1 Human gut mucosal colonisation with ESBL-producing *E. coli* over time

Having a water drinking source coming from a tube or a well, having a drophole cover on the toilet and animals being able to enter into contact with the food areas all appeared to be highly significant. Whilst a positive association was detected between the drinking water source coming from a tube or a well and the ESBL-producing *E. coli* colonisation status, the opposite can be said for the drinking water source coming from a pipe. This was highlighted by the negative correlation between those two water variables noticed previously in the correlation heatmap. Animals being able to enter into contact with food areas was positively associated with the outcome, whilst having a drophole cover on the toilet appeared to have a protective effect, being negatively associated with colonisation status. Additionally, variables such as keeping animals inside the house, having a toilet floor that is not made of concrete or wood and having clean paper in the toilet were also quite highly significant. Having clean paper in the toilet was negatively associated with colonisation status while the others showed a positive association. Other variables such as older age, the presence of open defecation in the area, owning cattle, sheep or goats, entering into contact with river water all were significant (<0.05) and were positively associated with the colonisation status. Contrarily, being male, having a higher income, having a disposal mechanism for animal waste, having a piped water drinking source, storing water in a container with lid and tap were all significant but negatively associated with colonisation status. In terms of temporal variables, the harmonic terms were also highly significant and the sample date appeared to be significant and positively associated, suggesting an increase in colonised samples over time. The results for the selected variables can be found in Table 4.9 and the full results of the univariable models can be found in Appendix C.5.

We subsequently ran what we called the temporal model, which investigates the temporal flux of ESBL-producing *E. coli* colonisation while looking at how the effect of WASH variables changes after adding the household-level random effect correlated in time.

Table 4.9: Univariable analysis results between ESBL-producing *E. coli* colonisation status and each variable accounting for the study area (Selected variables (<0.2))

	Log-odds	**P-value**	**Odds ratio (95% CI)**
Unknown HIV status (vs non-reactive)	-0.168	0.074	0.846 (0.704-1.017)
Recent use of antibiotics	0.062	0.137	1.063 (0.981-1.153)
Age*	**0.092**	**0.026**	**1.097 (1.011-1.189)**
Being male (vs female)	**-0.175**	**0.039**	**0.840 (0.711-0.992)**
Average household monthly income	**-0.107**	**0.019**	**0.899 (0.822-0.983)**
Open defecation	**0.089**	**0.038**	**1.093 (1.005-1.189)**
Presence of a disposal mechanism for animal waste	**-0.090**	**0.046**	**0.914 (0.837-0.998)**
Eating from shared plates	-0.079	0.072	0.924 (0.848-1.007)
Having a pipe as drinking water source	**-0.099**	**0.022**	**0.906 (0.832-0.985)**
Having a well as drinking water source	**0.191**	**3.8e-04**	**1.210 (1.089-1.345)**
Use of alternative water for cleaning utensils	0.057	0.174	1.059 (0.975-1.149)
Owning cattle, goats or sheep	**0.093**	**0.044**	**1.097 (1.003-1.201)**
Keeping animals inside	**0.122**	**0.005**	**1.129 (1.038-1.228)**
Contact with river water	**0.092**	**0.044**	**1.097 (1.003-1.200)**
Toilet floor material: no toilet (vs concrete/wood)	**0.275**	**0.050**	**1.317 (1.000-1.736)**
Toilet floor material: other (vs concrete/wood)	**0.330**	**0.002**	**1.391 (1.130-1.712)**
Having a drop hole cover on the toilet	**-0.164**	**1.7e-04**	**0.849 (0.779-0.925)**
Presence of newspaper/paper in the toilet	**-0.141**	**0.002**	**0.868 (0.796-0.947)**
Frequency of soap presence in handwashing facilities	-0.082	0.079	0.921 (0.841-1.009)
Storing water covered	-0.067	0.098	0.935 (0.863-1.013)
Contact between animals and food areas	**0.187**	**1.3e-05**	**1.205 (1.108-1.311)**
Presence of standing water around the household	-0.069	0.121	0.933 (0.855-1.018)
Number of days since the first sample	**0.118**	**0.005**	**1.125 (1.037-1.221)**
Harmonic term (sinday)	**-0.181**	**0.004**	**0.834 (0.738-0.943)**
Harmonic term (cosday)	**0.305**	**2e-04**	**1.357 (1.156-1.593)**

*Significant variables highlighted in bold

We found that men are less at risk of becoming colonised with ESBL-producing *E. coli* (OR 0.786 CI[0.678-0.910]) and that having a tube or a well as a water drinking source highly increases your risk of becoming colonised (OR 1.550 CI[1.003-2.394]). Coming into contact with standing water also appeared to be negatively associated with colonisation status (OR 0.749 CI[0.574-0.978]). Finally, there was still an apparent signal of annual seasonality noticeable from the presence of part of the harmonic term. These results are presented in Table 4.10.

Using the covariance structure, we found a range of temporal correlation estimated at 77.85 days (CrI [30.85-140.60]), thus samples that have been sampled in the same household more than 77 days apart are effectively uncorrelated. The estimates for the variance of the temporal process and the nugget effect were both close to 1, respectively 1.25 (CI [0.57-1.73]) and 1.29 (CI [0.89-1.69]). These results are presented in Table 4.11. The densities of the priors and posteriors of all three parameters can be found in Figure 4.4. Convergence was verified by looking at the traceplots in Figure 4.5 and we confirmed that the Gelman-Rubin statistic was close to 1 for all parameter estimates.

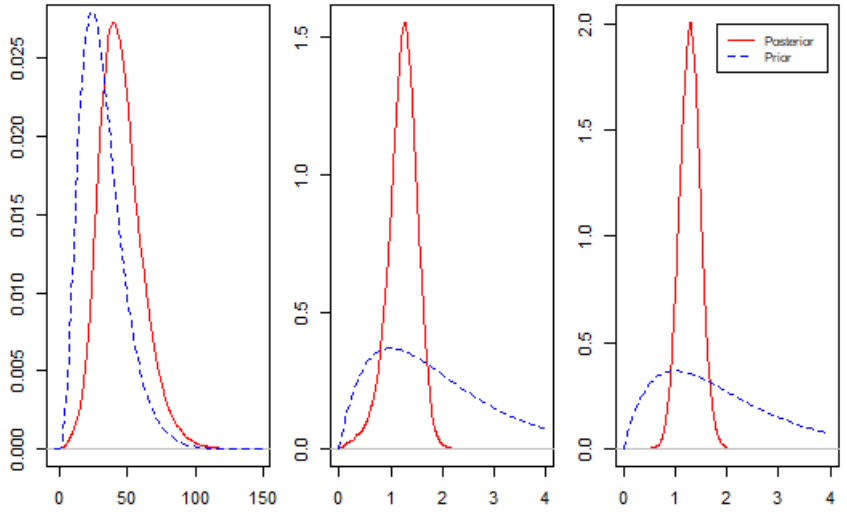


Figure 4.4: Prior and posterior density of ϕ, σ and τ (left to right, without warm-up) for the ESBL *E. coli* temporal model

Table 4.10: Temporal model results for ESBL-producing *E. coli* colonisation status

	Log-odds	**Odds ratio (95% CrI)**
Intercept	-0.716	0.489 (0.360-0.663)
Reactive to HIV testing (vs non-reactive)	0.027	1.027 (0.863-1.223)
Unknown HIV status (vs non-reactive)	-0.031	0.969 (0.808-1.163)
Recent use of antibiotics	0.093	1.097 (0.946-1.274)
Age	0.132	1.141 (0.970-1.343)
Being male (vs female)*	**-0.241**	**0.786 (0.678-0.910)**
Average household monthly income	0.226	1.254 (0.916-1.715)
Open defecation	0.054	1.055 (0.826-1.349)
Presence of a disposal mechanism for animal waste	0.104	1.110 (0.857-1.437)
Eating from shared plates	-0.245	0.783 (0.598-1.024)
Having a pipe as drinking water source	0.132	1.141 (0.818-1.592)
Having a tube/well as drinking water source	**0.438**	**1.550 (1.003-2.394)**
Use of alternative water for cleaning utensils	0.014	1.014 (0.802-1.283)
Owning cattle, goats or sheep	0.139	1.149 (0.892-1.480)
Keeping animals inside	0.075	1.078 (0.852-1.364)
Contact with river water	0.048	1.049 (0.791-1.391)
Toilet floor material: none (vs concrete/wood)	0.123	1.131 (0.799-1.600)
Toilet floor material: other (vs concrete/wood)	0.131	1.140 (0.820-1.585)
Presence of drop hole cover on the toilet	-0.202	0.817 (0.626-1.067)
Presence of newspaper/paper in the toilet	-0.155	0.856 (0.670-1.094)
Frequency of soap presence in handwashing facilities	-0.000	1.000 (0.640-1.563)
Storing water covered	-0.179	0.836 (0.537-1.302)
Storing water in a container with lid/tap	-0.034	0.967 (0.754-1.240)
Contact between animal and food areas	0.218	1.244 (0.983-1.573)
Presence of standing water around the household	**-0.289**	**0.749 (0.574-0.978)**
Number of days since the first sample	0.167	1.182 (0.869-1.608)
Harmonic term (sinday)	**-0.466**	**0.628 (0.453-0.869)**
Harmonic term (cosday)	0.371	1.449 (0.958-2.191)
Harmonic term (sinday2)	-0.084	0.919 (0.644-1.314)
Harmonic term (cosday2)	0.018	1.018 (0.711-1.457)
Living in Chikwawa (vs Chileka)	-0.276	0.759 (0.527-1.093)
Living in Ndirande (vs Chileka)	0.386	1.471 (0.980-2.207)

*Significant variables highlighted in bold

Table 4.11: Estimates for ϕ, σ and τ in the ESBL-Ec temporal model

	Estimate	Std	2.5%	97.5%
phi	43.52	15.87	18.82	81.36
sigma	1.25	0.29	0.57	1.73
tau	1.29	0.20	0.89	1.69

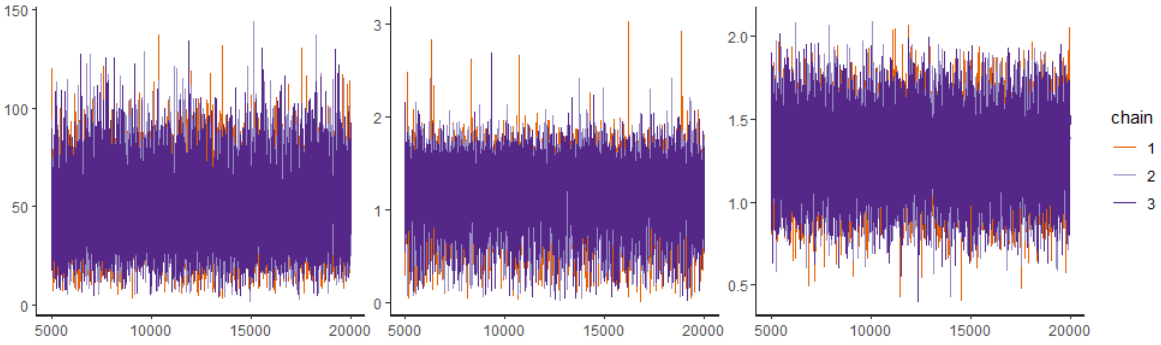


Figure 4.5: Trace plots of ϕ, σ and τ (left to right, without warm-up) for the temporal model for ESBL *E. coli*

4.4.2 Human gut mucosal colonisation with ESBL-producing *K. pneumoniae* over time

In the case of ESBL-producing *K. pneumoniae*, the size of the household was the only highly significant variable except for the harmonic terms, showing that the more people live in a household, the greater the risk of being colonised. Variables such as eating street food, eating from shared plates, owning birds and coming into contact with drains were also significant. Eating street food and eating from shared plates surprisingly appeared to have a protective effect on the ESBL-producing *K. pneumoniae* colonisation status. Owning birds and entering into contact with drains were both positively associated with colonisation status. In terms of temporal variables, the harmonic terms were also highly significant. The selected variables can be found in Table 4.12. The full results of the univariable models can be found in Appendix C.6.

We subsequently ran the temporal model for ESBL *K. pneumoniae*, which investigates the temporal flux of colonisation while looking at how the effect of WASH variables changes after adding the household-level random effect correlated in time.

Table 4.12: Univariable analysis results between ESBL-producing *K. pneumoniae* colonisation status and each variable accounting for the study area

	Log-odds	**P-value**	**Odds ratio (95% CI)**
Recent use of antibiotics	0.107	0.058	1.112 (0.996-1.242)
Number of people living in the household*	**0.229**	**6.1e-05**	**1.257 (1.124-1.406)**
Presence of a toilet in the household	0.085	0.193	1.088 (0.958-1.236)
Eating street food	**-0.125**	**0.029**	**0.882 (0.788-0.987)**
Eating from shared plates	**-0.152**	**0.016**	**0.859 (0.759-0.972)**
Having a pipe as drinking water source	0.106	0.077	1.112 (0.989-1.250)
Having a tap as drinking water source	-0.120	0.072	0.887 (0.778-1.011)
Use of alternative water for cleaning utensils	-0.087	0.187	0.917 (0.806-1.043)
Owning birds	**0.138**	**0.026**	**1.148 (1.016-1.296)**
Owning dogs or cats	0.086	0.153	1.089 (0.969-1.225)
Owning pigs	0.085	0.131	1.089 (0.975-1.216)
Contact with drains	**0.137**	**0.011**	**1.147 (1.032-1.275)**
Toilet type: pit latrine (vs no toilet)	0.287	0.107	1.333 (0.940-1.889)
Visible human faeces around the household	0.109	0.074	1.115 (0.990-1.256)
Storing water uncovered	-0.102	0.093	0.903 (0.801-1.017)
Number of days since the first sample	-0.097	0.122	0.907 (0.802-1.026)
Harmonic term (sinday)	**-0.345**	**3.7e-04**	**0.708 (0.586-0.856)**
Harmonic term (cosday)	**0.243**	**0.038**	**1.275 (1.014-1.605)**
Harmonic term (cosday2)	0.166	0.095	1.181 (0.972-1.434)
Living in Chikwawa (vs Chileka)	0.196	0.183	1.216 (0.912-1.622)

*Significant variables highlighted in bold

We found that having previously used antibiotics (in the last six months or in-between visits) increased your risk of being colonised with ESBL-producing *K. pneumoniae* (OR 1.281 CI[1.049-1.565]). We also saw a negative association between eating in shared plates and colonisation (OR 0.672 CI[0.460-0.980]). Finally, there was a signal of annual seasonality noticeable from the presence of part of the harmonic term, similar to the one we found for ESBL *E. coli*. These results are presented in Table 4.13.

We found a range of temporal correlation for ESBL *K. pneumoniae* colonisation estimated at 54.29 days (CrI [12.91-130.43]), thus samples that have been sampled in the same household more than 54 days apart are effectively uncorrelated. The estimates for the variance of the temporal process and the nugget effect were both close to 1, respectively 1.17 (CI [0.24-2.79]) and 1.63 (CI [1.05-2.28]). These results are presented in Table 4.14. The densities of the priors and posteriors of all three parameters can be found in Figure 4.6. Convergence was verified by looking at the traceplots in Figure 4.7 and we confirmed that the Gelman-Rubin statistic was close to 1 for all parameter estimates.

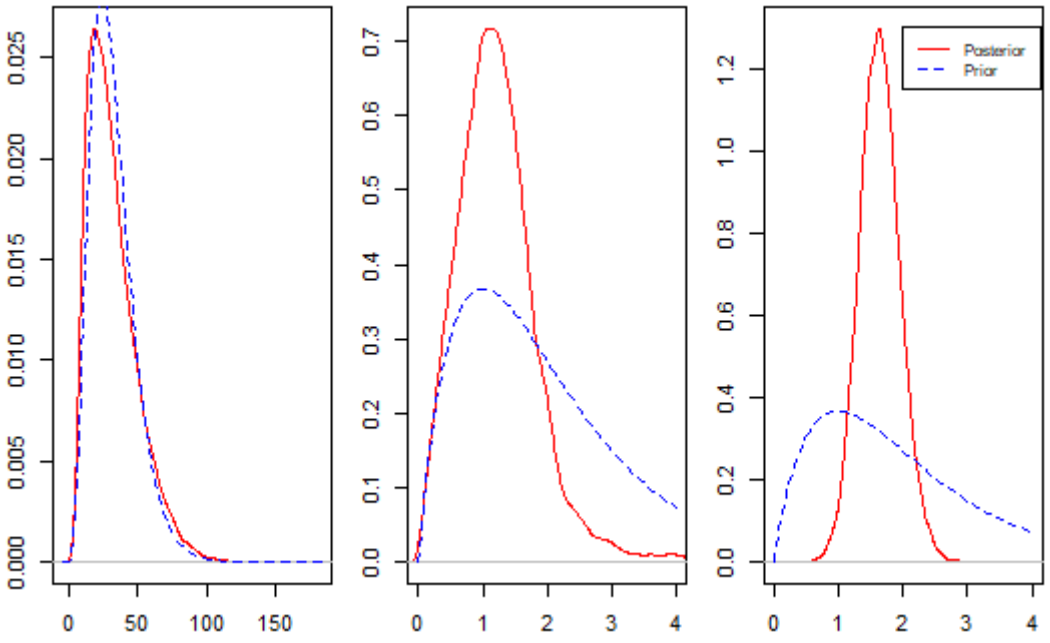


Figure 4.6: Prior and posterior density of ϕ, σ and τ (left to right) for the ESBL *K. pneumoniae* temporal model

Table 4.13: Temporal model results for ESBL-producing *K. pneumoniae* colonisation status

	Log-odds	**Odds ratio (95% CrI)**
Intercept	-3.432	0.032 (0.017-0.060)
Recent use of antibiotics*	**0.248**	**1.281 (1.049-1.565)**
Number of people living in the household	0.298	1.347 (0.932-1.947)
Presence of a toilet in the household	-0.136	0.873 (0.468-1.628)
Eating street food	0.091	1.095 (0.797-1.505)
Eating from shared plates	**-0.398**	**0.672 (0.460-0.980)**
Having a pipe as drinking water source	-0.253	0.776 (0.506-1.190)
Having a tap as drinking water source	-0.423	0.655 (0.408-1.051)
Use of alternative water for cleaning utensils	-0.241	0.786 (0.563-1.097)
Owning birds	0.015	1.015 (0.706-1.459)
Owning dogs or cats	0.024	1.024 (0.735-1.427)
Owning pigs	0.157	1.170 (0.853-1.604)
Contact with drains	0.219	1.245 (0.906-1.710)
Toilet type: other (vs no toilet)	0.203	1.225 (0.752-1.996)
Toilet type: pit latrine (vs no toilet)	0.204	1.226 (0.550-2.734)
Toilet type: shared toilet (vs no toilet)	-0.395	0.674 (0.395-1.148)
Visible human faeces around the household	0.224	1.251 (0.881-1.777)
Storing water uncovered	-0.326	0.722 (0.478-1.089)
Number of days since the first sample	-0.066	0.936 (0.587-1.493)
Harmonic term (sinday)	**-0.753**	**0.471 (0.289-0.767)**
Harmonic term (cosday)	0.448	1.565 (0.883-2.774)
Harmonic term (sinday2)	-0.304	0.738 (0.434-1.255)
Harmonic term (cosday2)	0.483	1.621 (0.961-2.736)
Living in Chikwawa (vs Chileka)	-0.038	0.963 (0.606-1.529)
Living in Ndirande (vs Chileka)	0.274	1.315 (0.812-2.130)

*Significant variables highlighted in bold

Table 4.14: Estimates of ϕ, σ and τ for the ESBL *K. pneumoniae* temporal model

	Estimate	**Std**	**2.5%**	**97.5%**
phi	27.82	17.91	7.78	75.21
sigma	1.17	0.66	0.24	2.79
tau	1.63	0.31	1.05	2.28

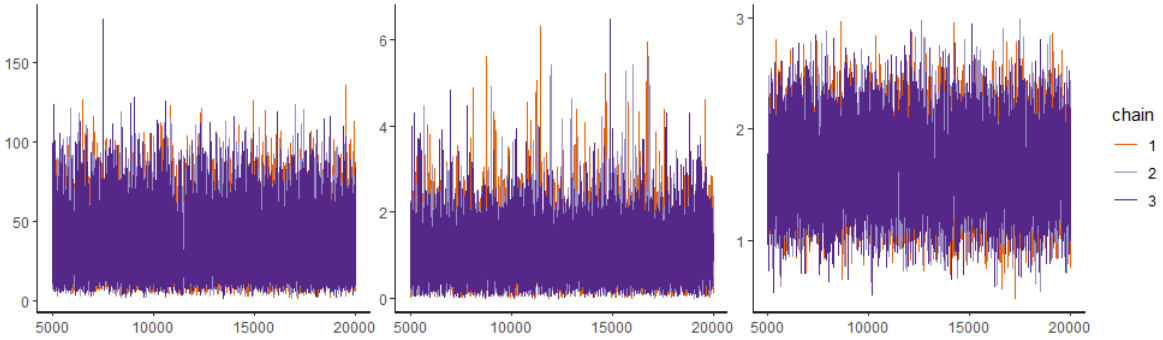


Figure 4.7: Trace plots of ϕ, σ and τ (left to right, without warm-up) for the temporal model for ESBL *K. pneumoniae*

4.5 Discussion

In this chapter, using logistic models, we looked into various WASH, demographic and household risk factors that could impact the gut mucosal colonisation of humans with ESBL-producing Enterobacteriaceae in community-based settings in Malawi, accounting for temporal variability. The availability of repeated measurements for approximately half of our households (after data cleaning) allowed us to investigate whether colonisation status was correlated in time within a household. The objective was to see if samples taken a certain time apart, e.g. 2 months, were more likely to be colonised than if taken with a longer wait, e.g. 6 months. Recent research in Malawi has shown that there appears to be some temporal correlation in the human gut mucosal colonisation with ESBL-producing Enterobacteriaceae thus we based our temporal correlation range prior distribution on their findings[173].

Due to the extent of the DRUM study, the amount of available variables from various questionnaires was extremely wide. For that reason, we had to pre-select vari-

ables depending on their perceived importance by the local experts and DRUM WASH team. This was a pragmatic decision on our side, however this could have led to some important variables being missed in the analysis. Additionally, we used univariable analysis to select variables for our temporal model. Similarly, this could lead to the non-identification of some important variables in our data, which is why we decided to set the significance level at 0.2 for the variable selection through univariable analysis.

The COVID-19 pandemic also played a role in derailing the microbiological sampling for our study and potentially impacting our results. The pandemic caused the sampling and microbiological testing to be suspended for a period of approximately four to six months, which caused a serious delay in our data collection and in finding out what the samples who were already collected were showing. Moreover, considering the temporal aspect of our analysis, this break caused a five to six months break in the middle of our samples, which added a difficulty in understanding what is really happening in the data at the temporal level. While initially the sampling visits were supposed to be relatively regular (0,1,3,6 months), this was rendered impossible by the pandemic and caused irregularities both between visit times between households and even between individuals within households at times.

The way we consider antibiotic use in our analysis should also be noted. Previously at baseline (Chapter 3), we wished to see if antibiotic groups showed an effect on either ESBL-producing *E. coli* colonisation or ESBL-producing *K. pneumoniae*. However that analysis not only showed that reported levels of specific antibiotic groups were quite low, it also did not show any effect on ESBL colonisation. Therefore, we decided to modify this variable for this analysis to only look at general antibiotic use as it appeared to be a more realistic strategy. Another concern about the use of these antibiotic-related variables relates to the way they were determined. These were ascertained by asking the participants whether or not they were given or received antibiotics. Consequently, there could have been issues with the reporting, and no verification was possible on our side.

The highest level of antibiotic use reported in time was at the baseline visit, with 15.2% of participants having been given at least a course of antibiotics in the last six months. In the subsequent visits, 6%, 9.4% and 8.3% were respectively reported. This might be explained by the shorter amount of time between subsequent visits com-

pared to the initial six months. Chikwawa had the highest reported antibiotic use with 15.4% while Ndirande and Chileka had respectively 9.3% and 6.2%. Chikwawa being the rural area in which we have previously found a lower average income, participants may have encountered more often organisations such as non-governmental organisations that might have been able to offer them treatment or antibiotics. They also might have increased access to antibiotics due to the greater presence of animal farming[176].

We found an overall prevalence of 37% for ESBL-producing *E. coli*, with varying proportions over time (33%-42.3%). For ESBL-producing *K. pneumoniae*, we found an overall prevalence of 11.9% ranging from 10% to 14.7% over time. We note that the highest prevalence for both of them in term of visits is during the fourth visit. Because of the variation in dates even within a visit group, it is hard to conjecture, however we can confirm an increasing trend for both over time, also noticed in the univariable analysis for ESBL-producing *E. coli*.

Antibiotic use was identified as a risk factor for ESBL-producing *K. pneumoniae*, which is consistent with previous studies in Sub-saharan Africa[3, 101, 104, 105]. Moreover, eating in shared plates rather than in separate plates appeared to have a protective effect. We suggest that on top of being a cultural practice, this is related to the social status, and that if people have less income, they have less access to food and therefore, are more likely to eat in shared plates. This is supported by the negative correlation found between those variables.

Through the correlation heatmap, we found that the social context seems to affect a lot of the WASH variables. As expected, the higher the income is, the more people are able to have access to better WASH infrastructure and products. Moreover, the income is positively correlated with using a water drinking source coming from a pipe while it is negatively correlated with having a drinking water source coming from a tube or a well. This suggests a positive association between using a tube or a well and being colonised, which is identified in the temporal model for ESBL-producing *E. coli*. Being female was also identified as a risk factor for ESBL-producing *E. coli*, which could be explained by the fact that traditionally, women tend to do the housework and the laundry or cook food and take care of the children. Therefore, this would put them at higher risk of being in contact with contaminated faeces or environment. No direct association

was found between income and gut mucosal colonisation in the model. Further work is needed to understand the association between income and gut mucosal colonisation.

Furthermore, at the univariable level, other variables were identified as highly significant such as the study area, with a higher risk of being colonised with ESBL-producing *E. coli* in Ndirande. Having animals inside and animals being in contact with food areas were also positively associated with gut mucosal colonisation. In contrast, having paper in the toilet and a drophole cover on the toilet were both negatively associated with ESBL-producing *E. coli* colonisation. This could suggest that implementation of such WASH infrastructure could help reduce transmission.

For ESBL-producing *K. pneumoniae*, at the univariable level, except from the harmonic terms, the household size was the only risk highly significant risk factor. This signal is consistent with what we found in the previous chapter. Considering that the variables we consistently found as most significant in the previous chapter, the in-house prevalence and number of colonised people in the house, were both capturing all the variation in the household-level random effect, we decided to not include them in this chapter as we wanted to focus on including this temporal correlation at the household-level. The temporal models for both bacterial species have shown that there is a temporal correlation range of between eight and eleven weeks, which suggests that within-household transmission occurs within such a time frame. Therefore, two samples taken within that time frame are more likely to both be colonised than if spread apart in time any further. Additionally, both models also showed a marked annual seasonality that is also consistent with what we found at baseline.

Although we found temporal correlation at the household-level, we could not find any at the individual-level, which suggested that an individual's samples could be seen as independent. This was somewhat surprising as we expected to find that an individual stays colonised for a certain amount of time. We highlight potential explanations for this lack of temporal correlation at the individual level. We have chosen to use stool samples however, different types of samples could have been used for testing, such as rectal swabs, which might be better for screening. Additionally, the laboratory protocol for testing is inherently digital, as we test for presence or absence of ESBLs and not quantity. This drove us to reconsider the way we are testing for gut mucosal colonisation

and whether or not it is the best way to analyse the samples. Further work is needed to investigate the specificity and sensitivity of the test.

A recent study on carriage of presumptive *E. coli* in a high-density informal settlement in Kenya showed that antibiotic use had little explanatory power for the prevalence of AMR and suggested that WASH factors are likely more important in driving transmission in these settings[177]. Here, our results also point towards transmission through contaminated water and/or inappropriate WASH infrastructure when looking at colonisation with ESBL-producing *E. coli*, with a high prevalence in the community and various identified WASH risk factors. Additionally, seasonality and gender also suggest the importance of WASH and the environment in driving ESBL-producing *E. coli* transmission. However, for ESBL-producing *K. pneumoniae*, previous antibiotic use was identified as a risk factor, therefore emphasizing the importance of antimicrobial exposure in driving ESBL-producing *K. pneumoniae* transmission.

Chapter 5

Discussion

This thesis explored the individual, household and WASH risk factors, across space and time, for ESBL-producing *E. coli* and *K. pneumoniae* colonisation in Southern Malawi. The MRC funded DRUM project started at a similar time as my MRC Doctoral Training Fellowship. As part of my thesis, I was able to develop the spatial sampling design for the households in the DRUM study in both Uganda and Malawi. Subsequently, I worked with the data collected by the field teams in Malawi using said spatial design to help answer questions that were drawn up when the project first started. I focused mainly on exploring the risk factors for ESBL-producing *E. coli* and *K. pneumoniae* colonisation in order to inform our understanding of transmission in these various community settings. This allowed me to split this question up into two categories, respectively focusing on the spatial and temporal aspect of the DRUM study and data.

5.1 Overview of the chapters

Chapter 2 focused on finding a suitable spatial sampling design for the DRUM study areas. Looking for spatial correlation using geostatistical models required an efficient sampling design that maximises variability in terms of the covariates while also making sure enough variability is present between sampling locations in order to inform the spatial parameters of the model. Hence, we based our sampling design on the recent "inhibitory

design with close pairs" method[139], which allows for a compromise between regular sampling and complete spatial randomness. Moreover, due to the One Health aspect of the DRUM study, study areas were selected from a variety of settings, i.e. urban/peri-urban/rural, thus were not likely to be homogeneous in terms of population density and accessibility. This led us to modify and extend the inhibitory design with close pairs[139] to include the possibility of using additional freely available geographical information, such as population density rasters or OpenStreetMap data, or pre-existing geolocation data to further inform the design and thus develop a pragmatic tool for expanding this method to various densities in multiple areas.

Next, Chapter 3 analysed the first part of the data collected in Malawi using the spatial sampling design. Due to the COVID-19 pandemic, this analysis was solely performed on the data collected at baseline visit, the sampling and/or laboratory results for the subsequent visits not having been performed at the time. Risk factors for human gut mucosal colonisation with ESBL-producing *E. coli* and *K. pneumoniae* were explored in all three Malawian sites, i.e. Ndirande, Chileka and Chikwawa, in order to learn more about their transmission patterns in different settings. Among all three areas combined and subsequently for each individual area, sharing a household with colonised people was identified as the main risk factor for ESBL colonisation. For ESBL-producing *E. coli*, the main risk factor was the ESBL prevalence in the household which indicates a potential dilutional effect when the household density increases, while for ESBL-producing *K. pneumoniae*, the main risk factor was the total number of colonised people in the house (except for Chileka). However, both consistently showed the importance of sharing a household with other colonised people. Additionally, a marked seasonality was detected for ESBL *E. coli* in all the study areas, which suggests a higher prevalence during the wet season than during the dry season. Finally, the geostatistical model did not detect any spatial correlation, suggesting that either there is no spatial correlation, or that the correlation was too short-scaled to be detected by the design.

Lastly, following findings from Chapter 3, Chapter 4 detailed an investigation into various WASH, demographic and household risk factors that could further explain the transmission of ESBL-producing *E. coli* and *K. pneumoniae* leading to human gut mucosal colonisation, accounting this time for temporal variability. More data became available after restrictions caused by the pandemic were lifted and we were able to use

longitudinal data to perform this analysis. Over the whole time period, an overall prevalence of 37% was found for ESBL-producing *E. coli* and 11.9% for ESBL-producing *K. pneumoniae*. In the multivariable analysis, being female was identified as a risk factor for colonisation with ESBL-producing *E. coli* as well as using a water drinking source such as a tube or a well. At the univariable level, living in Ndirande was identified as a risk factor as well as animal-related factors, such as having animals inside the house and animals being in contact with food areas. Contrarily, having access to cleaning materials like paper in the toilet and having a drop-hole cover on the toilet had a protective effect. For ESBL-producing *K. pneumoniae*, previous antibiotic use was identified as a risk factor in the multivariable analysis. Eating from shared plates instead of separate plates showed a protective effect, that we suggested could be related to its negative correlation with income. This would suggest a potential positive association between income and ESBL-producing *K. pneumoniae* colonisation. We also found that income is correlated with many of the WASH variables in the study, thus we conjectured that the social context affects greatly the WASH variables in our study. At the univariable level, household density was identified as a risk factor, which is consistent with what we found in the previous chapter. Temporal correlation was detected for both ESBL-producing *E. coli* and *K. pneumoniae* with an approximate range of eight to eleven weeks. A marked annual seasonality was also detected in both models, which is consistent with what we found in the previous chapter.

5.2 Implications for ESBL transmission leading to human colonisation

This thesis suggests the importance of within-household transmission for ESBL-producing *E. coli* and *K. pneumoniae*. In Chapter 3, we found that sharing the household with other colonised people was overwhelmingly the main risk factor for colonisation with either bacterial species. Subsequently, in Chapter 4, we showed that there is a temporal correlation of two to three months, not at the individual level but at the household level. This highlights the importance of the household in driving ESBL transmission.

Furthermore, the use of a water drinking source such as a tube or well was also

identified as a risk factor for ESBL-producing *E. coli* in the multivariable analysis. The strong negative correlation between a tube or well water drinking source and a piped drinking source also indicates that piped water would have an opposite effect and offer a protective effect against colonisation. Access to cleaning materials and having a drop-hole cover were found to have a protective effect while factors related to animal ownership were identified as risk factors in the univariable analysis. Additionally, being female was identified as a risk factor in the multivariable analysis. Women traditionally tend to be the ones taking care of housework such as cooking, laundry and taking care of the children. Consequently, this might be the reason why they are more at risk, as they have more chances of entering in contact with multiple potential sources of contamination. These results point towards transmission through contaminated water and/or inappropriate WASH infrastructure. This underlines the need for improved access to water and suggests that WASH behavioural practice might be beneficial. It also shows that better WASH conditions should help in decreasing transmission.

The high prevalence of ESBL-producing *E. coli* we detected in our human samples suggests that the transmission of this particular ESBL-producing bacterial species might no longer be mainly originating from the hospital environment, but also suggests a significant presence in the community. Human gut mucosal colonisation has been identified as a risk factor for subsequent ESBL infection[99, 100], therefore this high prevalence might be leading to more infections, causing community members to bring the resistant bacteria back to the hospital when getting sick.

Previous antibiotic use was identified as a risk factor for ESBL-producing *K. pneumoniae* in the multivariable analysis. This is consistent with previous research on ESBL colonisation[101, 173]. This emphasizes the importance of antimicrobial exposure in driving ESBL colonisation, thus highlighting the need for more responsible antibiotic consumption. Factors related to food sharing such as eating in shared plates were also identified as significant. Whilst plate sharing is a cultural practice, it is also a marker of lower income, as suggested by the correlation heatmap in Chapter 4. This suggests a potential association between income and colonisation with ESBL-producing *K. pneumoniae*. However, no direct association was found between income and ESBL colonisation in this analysis. Additionally, the household density was identified as a risk factor in both the univariable analysis in Chapter 4 and throughout the analyses of Chapter 3,

highlighting again the importance of within-household transmission.

In terms of seasonality, there is a constant annual seasonality with higher prevalence during the rainy season detected for ESBL-producing *E. coli* throughout the thesis, and also detected for ESBL-producing *K. pneumoniae* when the longitudinal data was included. This highlights again the need for better WASH infrastructure and behavioural practice. Additionally, a positive association between the sample date and ESBL colonisation is also often seen throughout the analyses, suggesting an increase in ESBL prevalence over time. This is highlighting the importance of this study in trying to prevent transmission in order to stop infections before the problem grows any more.

5.3 Novel contribution of the work

Chapter 2 details the development and proposes a new modified spatial sampling design, that permits the integration of more precise information on the study area or the population into the design. It allows for a more comprehensive sampling of households and a more efficient design in terms of field work efficiency, making the access to households on the field easier and faster. It can also be applied to various diseases, countries and settings. It is currently in talks to either be integrated into the geosample package available through the Comprehensive R Archive Network (CRAN) on R software or to be made into its own package.

Little is known about asymptomatic colonisation with ESBL *E. coli* and *K. pneumoniae*, which occurs before symptomatic infection in sub-Saharan Africa. The DRUM study is the first study investigating risk factors of ESBL colonisation among asymptomatic individuals in various community settings in sub-Saharan Africa. Additionally, Chapter 3 is the first spatial analysis of the prevalence of ESBL colonisation in humans. We have shown that at this scale, there does not appear to be any evidence of spatial correlation. However, we suggest that there might be spatial correlation at a smaller scale, as evidenced by the importance of within-household transmission.

Chapter 3 and 4 show that the prevalence of ESBL-producing *E. coli* colonisation in the community in Southern Malawi is within the ranges that have been described

previously in other studies on ESBL colonisation in sub-Saharan Africa [3, 4]. However, while a big proportion of these studies were either from the hospital setting or on a specific population, our study focused on the general population (asymptomatic individuals) in community settings in sub-Saharan Africa. Comparing to community studies exclusively, we found a prevalence of 37% which is much higher than the estimate of 18% [95% CI11–28%] found for community members[3]. This prevalence is close to some of the highest reported in the world[4]. Additionally, Chapter 4 was the first analysis to identify gender as a risk factor for ESBL-producing *E. coli* colonisation in this setting.

5.4 Limitations

Prior to the study, little was known about the potential scale of spatial correlation for ESBL colonisation since this is the first spatial study on the prevalence of ESBL colonisation. Therefore, we were not able to use pre-existing knowledge on the scale of correlation and had to select weakly informative priors for the spatial model. Furthermore, the design of the study did not allow us to detect any evidence of spatial correlation. However, it can not be confirmed at this point whether there really is a lack of spatial correlation, or if that correlation was just too short-scaled to be detected by the design. Future studies can use this work, and use the level of resolution provided by whole genome sequencing to test for shorter scale correlation when looking at spatial variation.

When looking at the longitudinal data, since we expected colonisation to last for some time, we thought to find some temporal correlation at the individual level for the ESBL colonisation status. However we did not find any, yet we found it at the household level. There are multiple reasons why we suggest this could have happened: colonisation could not last as long as we would have thought in an individual or the way we test for colonisation could be less sensitive than initially thought. Further work is needed to understand why this correlation was not found at the individual level and determine if the way we test for ESBL colonisation is appropriate.

We note that the methods of variable selection have their own limitations. There is no agreement on what the best way to select variables is in such models, therefore we decided to use different strategies in different chapters. Using a forwards se-

lection procedure implies selecting one best fitting model over other models that could have been of similar fit. Moreover, it is possible that variables that appear as not significant, might have been significant when combined with other variables in the same model, which mainly happens in the presence of correlated variables. We included the full models to allow for an unbiased estimation of the parameters. Using univariable analysis also presents some limitations such as potentially missing important variables when keeping a selection threshold of 0.05, which is why we set ours to 0.2. We also preselected the variables using local expert knowledge due to the extremely wide dataset of covariates. This might have created a selection bias on the covariates. Overall, our goal was to find the best fitting model, that made sense at the epidemiological level, and we did find consistent results among the different areas and when adding more data.

COVID-19 impacted the study in various ways. It first impacted the data collection and microbiological testing, because of restrictions. This not only delayed access to the data and laboratory results but also changed the way we look at our longitudinal data. What should have been regular time points with pre-decided time intervals became very irregular. This impacted the model and the results, initially in the spatial model by inserting a "pause" between samples, which affected the results by making it more difficult to model variables like the sample date and harmonic terms. Secondly, because we do not know what the colonisation status would have been for samples that would have been sampled during those few months, it affects our understanding of how the seasonality really affects ESBL colonisation. Although we found evidence of seasonality, the results would have been stronger in terms of interpretation if we had had access to that data.

5.5 Future work

We have found multiple risk factors for ESBL-producing *E. coli* and *K. pneumoniae*. It would be interesting to see if these risk factors vary in the different socioeconomic settings when including the temporal effect we detected for all three areas combined. Moreover, combining spatial and temporal approaches would allow for further investigation of the transmission patterns and a comparison with what we have discovered throughout

this thesis. Future spatial studies on ESBL colonisation should either increase the number of households to be sampled within areas of similar size as ours, however that is not always practical due to cost-constraints, or reduce the size of the study area in order to increase point density within the area and potentially detect spatial correlation, if it does exist. Furthermore, looking at a smaller number of individuals but at more frequent and regular time points could help confirm the findings of this thesis and explore whether temporal correlation can be detected at the individual level in this context.

Using the same spatial and temporal approaches to investigate the risk factors for ESBL-producing *E. coli* and *K. pneumoniae* in the DRUM study areas in Uganda will also permit a comparison between countries, which could help in confirming the findings of this thesis. They might also highlight other risk factors and/or transmission pathways and help us better understand the transmission patterns of ESBL colonisation in different settings in sub-Saharan Africa.

Potential transmission routes have been identified in this thesis. Luckily, the DRUM study was designed to include many types of field sampling, including environmental and water sampling. These samples will allow us to determine an estimate of the levels of contamination with ESBL *E. coli* and *K. pneumoniae* in the environment and potentially ascertain that they support our findings.

Evaluating WASH behaviours and infrastructure focusing on households where all individuals are colonised at one time point and comparing it with households where no individuals are colonised could also potentially be helpful in trying to understand which behaviours are impacting ESBL transmission. In order to interrupt transmission, there is a need for understanding which WASH behaviours are impacting the ESBL colonisation the most and how to correct or modify that specific behaviour.

We have found various risk factors and correlations suggesting that income might be associated with ESBL colonisation. However, the relationship between income and ESBL colonisation is likely complex, as previous studies have found opposite results on the effect of socio-economic status on colonisation[107, 108]. Additionally, the relationship between income and WASH factors is also complex, as two families with similar income might still have different WASH infrastructure. Thus future work is needed on

the impact of income and/or social context on the WASH infrastructure and how it relates to ESBL colonisation. Such an association would impact the way interventions are designed at a national level.

Finally, our findings show that community-based surveillance is required in sSA with a prevalence of colonisation with the most common ESBL-producing Enterobacteriaceae reaching some of the highest levels detected across the world[4]. Preventing further transmission leading to asymptomatic colonisation would likely reduce ESBL infections and subsequently, the overall burden of disease.

5.6 Conclusion

In conclusion, this thesis started with the development of a modified spatial sampling design, allowing for inclusion of geographical data or pre-existing knowledge into the design. This design can be used for various diseases, countries and settings and will soon be published as a paper with accompanying R package (modifying the geosample package or creating a new package). Subsequently, using the broad questions the DRUM study planned to answer, we have developed spatial and longitudinal approaches to explore risk factors for ESBL colonisation in various settings in Southern Malawi. This thesis findings suggest the importance of within-household transmission route in driving ESBL colonisation in the community. It also highlights how complex transmission in this setting is and the potential importance of the environmental and faecal-oral routes through the identification of WASH factors and gender as risk factors. We recommend that interventions aimed at preventing transmission, targeted at the household level, focusing on modifying the within-household behaviour relating to WASH factors and/or improving the WASH infrastructure, may have great potential. Furthermore, risk factors differing for ESBL-producing *E. coli* and *K. pneumoniae* drive the question of whether their main transmission pathways are the same. While ESBL-producing *E. coli* appears to be mainly affected by WASH factors, ESBL-producing *K. pneumoniae* shows an association only with antibiotic exposure and factors that were correlated with income. Further work is needed to explore this question, but if this is the case, interventions might need to vary depending on the social context and the bacterial species. Due to the identifica-

tion of antibiotic use as a risk factor and *K. pneumoniae* being the archetype nosocomial pathogen, we suggest that improved IPC measures and antibiotic usage and stewardship training might help in preventing transmission of ESBL *K. pneumoniae*. The transmission patterns of ESBL asymptomatic colonisation remain difficult to describe as the overall body of research in the subject in sub-Saharan Africa is still recent and not well detailed. However, this work has taken part and allowed for a new way of designing spatial sampling as well as new spatial and longitudinal approaches to look at transmission patterns for ESBL colonisation in various settings. In conclusion, this work highlights the value of international and interdisciplinary collaborations between various fields and how new methodological approaches can be developed and applied to real-life contexts.

References

[1] World Health Organization. (2021). *WHO integrated global surveillance on ESBL-producing E. coli using a "One Health" approach: implementation and opportunities.* https://apps.who.int/iris/handle/10665/340079 License: CC BY-NC-SA 3.0 IGO

[2] Musicha, P., Cornick, J. E., Bar-Zeev, N., French, N., Masesa, C., Denis, B., Kennedy, N., Mallewa, J., Gordon, M. A., Msefula, C. L., Heyderman, R. S., Everett, D. B. and Feasey, N. A. (2017). *Trends in antimicrobial resistance in bloodstream infection isolates at a large urban hospital in Malawi (1998-2016): a surveillance study.* The Lancet. Infectious diseases, 17(10), 1042–1052. https://doi.org/10.1016/S1473-3099(17)30394-8

[3] Lewis, J.M., Lester, R., Garner, P. and Feasey, N.A. (2019). *Gut mucosal colonisation with extended-spectrum beta-lactamase producing Enterobacteriaceae in sub-Saharan Africa: a systematic review and meta-analysis.* Wellcome open research, 4.

[4] Karanika, S., Karantanos, T., Arvanitis, M., Grigoras, C. and Mylonakis, E. (2016). *Fecal colonization with extended-spectrum beta-lactamase–producing Enterobacteriaceae and risk factors among healthy individuals: a systematic review and metaanalysis.* Reviews of Infectious Diseases, 63(3), pp.310-318.

[5] World Health Organization. (2015). *Global action plan on antimicrobial resistance.* World Health Organisation. https://apps.who.int/iris/handle/10665/193736

[6] United Nations. (2015). *United Nation's Sustainable Development Goals – 17 Goals to Transform Our World.* https://www.un.org/sustainabledevelopment.

[7] United Nations. (2016). *UN Draft political declaration of the high-level meeting of the General Assembly on antimicrobial resistance.* https://www.un.org/pga/71/wp-content/uploads/sites/40/2016/09/Draft-AMR-Declaration.pdf

[8] O'Neill J. (2016). *Tackling drug-resistant infections globally: Final report and recommendations.*

[9] Who.int. (2019). *Antimicrobial resistance. [online]* World Health Organisation. https://www.who.int/en/news-room/fact-sheets/detail/antimicrobial-resistance

[10] World Health Organization. (2014). *Antimicrobial resistance: global report on surveillance.* World Health Organisation. https://apps.who.int/iris/handle/10665/112642

[11] Michael, C.A., Dominey-Howes, D., and Labbate, M. (2014). *The antimicrobial resistance crisis: causes, consequences, and management.* Frontiers in public health, 2, 145. https://doi.org/10.3389/fpubh.2014.00145

[12] Moore-Gillon, J. (2001). *Multidrug-resistant tuberculosis: this is the cost.* Annals of the New York Academy of Sciences, 953(1), pp.233-240.

[13] Jonas, O.B., Irwin, A., Berthe, F.C.J., Le Gall, F.G., Marquez, P.V. et al. (2017). *Drug-resistant infections : a threat to our economic future (Vol. 2) : final report (English).* HNP/Agriculture Global Antimicrobial Resistance Initiative Washington, D.C. : World Bank Group. http://documents.worldbank.org/curated/en/323311493396993758/final-report

[14] Freire-Moran, L., Aronsson, B., Manz, C., Gyssens, I.C., So, A.D., Monnet, D.L., Cars, O. and ECDC-EMA working group. (2011). *Critical shortage of new antibiotics in development against multidrug-resistant bacteria—Time to react is now.* Drug resistance updates, 14(2), pp.118-124.

[15] World Health Organization. (2021). *2020 antibacterial agents in clinical and preclinical development: an overview and analysis.* World Health Organisation. https://apps.who.int/iris/handle/10665/340694 License: CC BY-NC-SA 3.0 IGO

[16] Wester, A.L., Gopinathan, U., Gjefle, K., Solberg, S.Ø. and Røttingen, J.R. (2017). *Antimicrobial resistance in a one health and one world perspective mechanisms and solutions* International Encyclopedia of Public Health. pp. 140–153. Elsevier.

[17] Singer, A.C., Kirchhelle, C. and Roberts, A.P. (2019). *Reinventing the antimicrobial pipeline in response to the global crisis of antimicrobial-resistant infections.* F1000Research, 8.

[18] Beyer, P., Cueni, T.B., Knox, J. and Riedl, F. (2020). *A research and development fund for new treatments for bacterial infections.* Bulletin of the World Health Organization, 98(12), p.822.

[19] Abraham, E.P. and Chain, E. (1940). *An enzyme from bacteria able to destroy penicillin.* Nature, 146(3713), pp.837-837.

[20] Davies, J. and Davies, D. (2010). *Origins and evolution of antibiotic resistance.* Microbiology and molecular biology reviews : MMBR, 74(3), 417–433. https://doi.org/10.1128/MMBR.00016-10

[21] Ventola, C.L. (2015). *The antibiotic resistance crisis: part 1: causes and threats.* Pharmacy and therapeutics, 40(4), p.277.

[22] Mindlin, S.Z. and Petrova, M.A. (2017). *On the origin and distribution of antibiotic resistance: permafrost bacteria studies.* Molecular Genetics, Microbiology and Virology, 32(4), pp.169-179.

[23] Laxminarayan, R., Duse, A., Wattal, C., Zaidi, A.K., Wertheim, H.F., Sumpradit, N., Vlieghe, E., Hara, G.L., Gould, I.M., Goossens, H. and Greko, C. (2013). *Antibiotic resistance — the need for global solutions.* The Lancet infectious diseases, 13(12), pp.1057-1098.

[24] Magiorakos, A.P., Srinivasan, A., Carey, R.B., Carmeli, Y., Falagas, M.E., Giske, C.G., Harbarth, S., Hindler, J.F., Kahlmeter, G., Olsson-Liljequist, B. and Paterson, D.L. (2012). *Multidrug-resistant, extensively drug-resistant and pandrug-resistant bacteria: an international expert proposal for interim standard definitions for acquired resistance.* Clinical microbiology and infection, 18(3), pp.268-281.

[25] Nikaido, H. (2009). *Multidrug resistance in bacteria.* Annual review of biochemistry, 78, pp.119-146.

[26] Kuehn, B.M. (2013). *IDSA: Better, Faster Diagnostics for Infectious Diseases Needed to Curb Overtreatment, Antibiotic Resistance.* JAMA, 310(22), 2385–. https://doi.org/10.1001/jama.2013.283828

[27] World Health Organization. (2015). *Worldwide country situation analysis: response to antimicrobial resistance.* World Health Organisation. https://apps.who.int/iris/handle/10665/163468

[28] Delepierre, A., Gayot, A. and Carpentier, A. (2012). *Update on counterfeit antibiotics worldwide; public health risks*. Med. Mal. Infect. 42 (6), 247–255.

[29] World Health Organization, Food and Agriculture Organization of the United Nations & World Organisation for Animal Health. (2021). *Monitoring global progress on antimicrobial resistance: tripartite AMR country self-assessment survey (TrACSS) 2019–2020: global analysis report.* World Health Organisation. https://apps.who.int/iris/handle/10665/340236 License: CC BY-NC-SA 3.0 IGO

[30] Laxminarayan, R., Van Boeckel, T., Frost, I., Kariuki, S., Khan, E.A., Limmathurotsakul, D., Larsson, D.J., Levy-Hara, G., Mendelson, M., Outterson, K. and Peacock, S.J. (2020). *The Lancet Infectious Diseases Commission on antimicrobial resistance: 6 years later.* The Lancet Infectious Diseases, 20(4), pp.e51-e60.

[31] Van, T.T.H., Yidana, Z., Smooker, P.M. and Coloe, P.J. (2020). *Antibiotic use in food animals worldwide, with a focus on Africa: Pluses and minuses.* Journal of global antimicrobial resistance, 20, pp.170-177.

[32] Medlicott, K., Wester, A., Gordon, B., Montgomery, M., Tayler, E., Sutherland, D., Schmoll, O., De-Souza, M., Koo-Oshima, S., Da-Balogh, K. and Pinto-Ferreira, J. (2020). *Technical brief on water, sanitation, hygiene and wastewater management to prevent infections and reduce the spread of antimicrobial resistance.* WHO/FAO/OIE Recommendations Report.

[33] Wellington, E.M., Boxall, A.B., Cross, P., Feil, E.J., Gaze, W.H., Hawkey, P.M., Johnson-Rollings, A.S., Jones, D.L., Lee, N.M., Otten, W. and Thomas, C.M. (2013). *The role of the natural environment in the emergence of antibiotic resistance in Gram-negative bacteria.* The Lancet infectious diseases, 13(2), pp.155-165.

[34] Segura, P. A., François, M., Gagnon, C., and Sauvé, S. (2009). *Review of the occurrence of anti-infectives in contaminated wastewaters and natural and drinking waters.* Environmental health perspectives, 117(5), 675–684. https://doi.org/10.1289/ehp.11776

[35] Kinney, C.A., Furlong, E.T., Zaugg, S.D., Burkhardt, M.R., Werner, S.L., Cahill, J.D. and Jorgensen, G.R. (2006). *Survey of organic wastewater contaminants in biosolids destined for land application.* Environmental science & technology, 40(23), pp.7207-7215.

[36] FAO. (2016). *Drivers, dynamics and epidemiology of antimicrobial resistance in animal production.* http://www.fao.org/feed-safety/resources/resources-details/en/c/452608/

[37] Huijbers, P.M., Blaak, H., de Jong, M.C., Graat, E.A., Vandenbroucke-Grauls, C.M. and de Roda Husman, A.M. (2015). *Role of the environment in the transmission of antimicrobial resistance to humans: a review.* Environmental science & technology, 49(20), pp.11993-12004.

[38] Wuijts, S., van den Berg, H.H., Miller, J., Abebe, L., Sobsey, M., Andremont, A., Medlicott, K.O., van Passel, M.W. and de Roda Husman, A.M. (2017). *Towards a research agenda for water, sanitation and antimicrobial resistance.* Journal of Water and Health, 15(2), pp.175-184.

[39] Baquero, F., Martínez, J.L. and Cantón, R. (2008). *Antibiotics and antibiotic resistance in water environments.* Current opinion in biotechnology, 19(3), pp.260-265.

[40] Finley, R.L., Collignon, P., Larsson, D.J., McEwen, S.A., Li, X.Z., Gaze, W.H., Reid-Smith, R., Timinouni, M., Graham, D.W. and Topp, E. (2013). *The scourge of antibiotic resistance: the important role of the environment.* Clinical infectious diseases, 57(5), pp.704-710.

[41] Novo, A., André, S., Viana, P., Nunes, O.C. and Manaia, C.M. (2013). *Antibiotic resistance, antimicrobial residues and bacterial community composition in urban wastewater.* Water research, 47(5), pp.1875-1887.

[42] WWAP (United Nations World Water Assessment Programme). (2017). *The United Nations World Water Development Report 2017: Wastewater, The Untapped Resource.* Paris, UNESCO.

[43] Corcoran, E., Nellemann, C., Baker, E., Bos, R., Osborn, D., et al. (2010). *Sick Water? The Central Role of Wastewater Management in Sustainable Development. A Rapid Response Assessment.* United Nations Environment Programme and UN-HABITAT. ISBN:978-82-7701-075-5.

[44] Mceachran, A. D., Blackwell, B.R., Hanson, J.D., Wooten, K.J., Mayer, G.D., Cox, S. B. and Smith, P. N. (2015). *Antibiotics, Bacteria, and Antibiotic Resistance Genes: Aerial Transport from Cattle Feed Yards via Particulate Matter.* Environ Health Perspect, 123: 337-43.

[45] Otte, J., Roland-Holst, D., Pfeiffer, D., Soares-Magalhaes, R., Rushton, J., Graham, J. and Silbergeld, E. (2007). *Industrial livestock production and global health risks.* Food and Agriculture Organization of the United Nations, Pro-Poor Livestock Policy Initiative Research Report.

[46] Wardyn, S.E., Forshey, B.M., Farina, S.A., Kates, A.E., Nair, R., Quick, M.K., Wu, J.Y., Hanson, B.M., O'Malley, S.M., Shows, H.W. and Heywood, E.M. (2015). *Swine farming is a risk factor for infection with and high prevalence of carriage of multidrug-resistant Staphylococcus aureus.* Clinical infectious diseases, 61(1), pp.59-66.

[47] McEwen, S.A. and Collignon, P.J. (2018). *Antimicrobial resistance: a one health perspective.* Microbiology spectrum, 6(2), pp.6-2.

[48] Collignon, P. (2015). *Antibiotic resistance: are we all doomed?* Internal medicine journal, 45(11), pp.1109-1115.

[49] Joint Tripartite (FAO, OIE, WHO) and UNEP. (2021). *One Health High Level Expert Panel (OHHLEP).* World Health Organisation. `https://www.who.int/groups/one-health-high-level-expert-panel`

[50] Joint Tripartite and UNEP. (2021). *Tripartite and UNEP support OHHLEP's definition of "One Health".* World Health Organisation. `https://www.who.int/news/item/01-12-2021-tripartite-and-unep-support-ohhlep-s-definition-of-one-health`

[51] Geneva: World Health Organization and the United Nations Children's Fund. (2021). *Progress on household drinking water, sanitation and hygiene 2000-2020: five years into the SDGs.* World Health Organisation. https://washdata.org/sites/default/files/2021-07/jmp-2021-wash-households.pdf

[52] World Health Organization. (2019). *No time to Wait: Securing the future from drug-resistant infections.* World Health Organisation. https://www.who.int/publications/i/item/no-time-to-wait-securing-the-future-from-drug-resistant-infections

[53] OECD (2018). *Stemming the Superbug Tide: Just A Few Dollars More.* OECD Health Policy Studies, OECD Publishing, Paris. https://doi.org/10.1787/9789264307599-en.

[54] Lacotte, Y., Årdal, C. and Ploy, M.C. (2020). *Infection prevention and control research priorities: what do we need to combat healthcare-associated infections and antimicrobial resistance? Results of a narrative literature review and survey analysis.* Antimicrobial Resistance & Infection Control, 9(1), pp.1-10.

[55] Frost, I., Van Boeckel, T.P., Pires, J., Craig, J. and Laxminarayan, R. (2019). *Global geographic trends in antimicrobial resistance: the role of international travel.* Journal of travel medicine, 26(8), p.taz036.

[56] World Tourism Organization. (2019). *International Tourism Highlights, 2019 Edition.* UNWTO, Madrid, https://doi.org/10.18111/9789284421152.

[57] Ruppé, E., Andremont, A. and Armand-Lefèvre, L. (2018). *Digestive tract colonization by multidrug-resistant Enterobacteriaceae in travellers: an update.* Travel medicine and infectious disease, 21, pp.28-35.

[58] Nordmann, P., Naas, T. and Poirel, L. (2011). *Global spread of carbapenemase-producing Enterobacteriaceae.* Emerging infectious diseases, 17(10), p.1791.

[59] Grami, R., Mansour, W., Mehri, W., Bouallègue, O., Boujaâfar, N., Madec, J.Y. and Haenni, M. (2016). *Impact of food animal trade on the spread of mcr-1-mediated colistin resistance, Tunisia, July 2015.* Eurosurveillance, 21(8), p.30144.

[60] Jansen, W., Mueller, A., Grabowski, N.T., Kehrenberg, C., Muylkens, B. and Al Dahouk, S. (2019). *Foodborne diseases do not respect borders: zoonotic pathogens and*

antimicrobial resistant bacteria in food products of animal origin illegally imported into the European Union. The Veterinary Journal, 244, pp.75-82.

[61] Batura, N., Cuevas, C., Khan, M. and Wiseman, V. (2018). *How effective and cost-effective are behaviour change interventions in improving the prescription and use of antibiotics in low-income and middle-income countries? A protocol for a systematic review.* BMJ open, 8(5), p.e021517.

[62] IHME, GBD Results Tool | GHDx, (2018). http://ghdx.healthdata.org/gbd-results-tool.

[63] Klein, E.Y., Van Boeckel, T.P., Martinez, E.M., Pant, S., Gandra, S., Levin, S.A., Goossens, H. and Laxminarayan, R. (2018). *Global increase and geographic convergence in antibiotic consumption between 2000 and 2015.* Proceedings of the National Academy of Sciences, 115(15), pp.E3463-E3470.

[64] Paterson, D.L. and Bonomo, R.A. (2005). *Extended-spectrum beta-lactamases: a clinical update.* Clinical microbiology reviews, 18(4), 657–686. https://doi.org/10.1128/CMR.18.4.657-686.2005

[65] Shaikh, S., Fatima, J., Shakil, S., Rizvi, S.M.D. and Kamal, M.A. (2015). *Antibiotic resistance and extended spectrum beta-lactamases: Types, epidemiology and treatment.* Saudi journal of biological sciences, 22(1), pp.90-101.

[66] Li, Q., Chang, W., Zhang, H., Hu, D. and Wang, X. (2019). *The role of plasmids in the multiple antibiotic resistance transfer in ESBLs-producing Escherichia coli isolated from wastewater treatment plants.* Frontiers in microbiology, 10, p.633.

[67] Reygaert, W.C. (2018). *An overview of the antimicrobial resistance mechanisms of bacteria.* AIMS microbiology, 4(3), p.482.

[68] San Millan, A., Escudero, J.A., Gifford, D.R., Mazel, D. and MacLean, R.C. (2016). *Multicopy plasmids potentiate the evolution of antibiotic resistance in bacteria.* Nature ecology evolution, 1(1), pp.1-8.

[69] Tacconelli, E., Carrara, E., Savoldi, A., Harbarth, S., Mendelson, M., Monnet, D. L., Pulcini, C., Kahlmeter, G., Kluytmans, J., Carmeli, Y., Ouellette, M., Outterson, K., Patel, J., Cavaleri, M., Cox, E. M., Houchens, C. R., Grayson, M. L., Hansen,

P., Singh, N., Theuretzbacher, U., WHO Pathogens Priority List Working Group. (2018). *Discovery, research, and development of new antibiotics: the WHO priority list of antibiotic-resistant bacteria and tuberculosis.* The Lancet. Infectious diseases, 18(3), 318–327. https://doi.org/10.1016/S1473-3099(17)30753-3

[70] Storberg, V. (2014). *ESBL-producing Enterobacteriaceae in Africa – a non-systematic literature review of research published 2008–2012.* Infection Ecology & Epidemiology, 4:1, DOI: 10.3402/iee.v4.20342

[71] Musicha, P., Msefula, C.L., Mather, A.E., Chaguza, C., Cain, A.K., Peno, C., Kallonen, T., Khonga, M., Denis, B., Gray, K.J. and Heyderman, R.S. (2019). *Genomic analysis of Klebsiella pneumoniae isolates from Malawi reveals acquisition of multiple ESBL determinants across diverse lineages.* Journal of Antimicrobial Chemotherapy, 74(5), pp.1223-1232.

[72] Centers for Disease Control and Prevention, National Center for Emerging and Zoonotic Infectious Diseases (NCEZID), Division of Foodborne, Waterborne, and Environmental Diseases (DFWED). *Escherichia coli.* https://www.cdc.gov/ecoli/index.html

[73] Podschun, R. and Ullmann, U. (1998). *Klebsiella spp. as nosocomial pathogens: epidemiology, taxonomy, typing methods, and pathogenicity factors.* Clinical microbiology reviews, 11(4), 589–603. https://doi.org/10.1128/CMR.11.4.589

[74] Moradigaravand, D., Martin, V., Peacock, S.J. and Parkhill, J. (2017). *Evolution and epidemiology of multidrug-resistant Klebsiella pneumoniae in the United Kingdom and Ireland.* MBio, 8(1), pp.e01976-16.

[75] Malande O.O., Nuttall J., Pillay V., Bamford C., Eley B. (2019). *A ten-year review of ESBL and non-ESBL Escherichia coli bloodstream infections among children at a tertiary referral hospital in South Africa.* PLOS ONE 14(9): e0222675. https://doi.org/10.1371/journal.pone.0222675

[76] Tansarli, G.S., Poulikakos, P., Kapaskelis, A. and Falagas, M.E. (2014). *Proportion of extended-spectrum β-lactamase (ESBL)-producing isolates among Enterobacteriaceae in Africa: evaluation of the evidence—systematic review.* Journal of Antimicrobial Chemotherapy, 69(5), pp.1177-1184.

[77] Sonda, T., Kumburu, H., van Zwetselaar, M., Alifrangis, M., Lund, O., Kibiki, G. and Aarestrup, F.M. (2016). *Meta-analysis of proportion estimates of Extended-Spectrum-Beta-Lactamase-producing Enterobacteriaceae in East Africa hospitals.* Antimicrobial Resistance & Infection Control, 5(1), pp.1-9.

[78] Seni, J., Moremi, N., Matee, M., Van der Meer, F., DeVinney, R., Mshana, S.E. and D Pitout, J.D. (2018). *Preliminary insights into the occurrence of similar clones of extended-spectrum beta-lactamase-producing bacteria in humans, animals and the environment in Tanzania: A systematic review and meta-analysis between 2005 and 2016.* Zoonoses and public health, 65(1), pp.1-10.

[79] Pitout, J.D., Nordmann, P., Laupland, K.B. and Poirel, L. (2005). *Emergence of Enterobacteriaceae producing extended-spectrum β-lactamases (ESBLs) in the community.* Journal of Antimicrobial Chemotherapy, 56(1), pp.52-59.

[80] Mathers, A.J., Peirano, G. and Pitout, J.D. (2015). *The role of epidemic resistance plasmids and international high-risk clones in the spread of multidrug-resistant Enterobacteriaceae.* Clinical microbiology reviews, 28(3), pp.565-591.

[81] Seni, J., Najjuka, C.F., Kateete, D.P., Makobore, P., Joloba, M.L., Kajumbula, H., Kapesa, A. and Bwanga, F. (2013). *Antimicrobial resistance in hospitalized surgical patients: a silently emerging public health concern in Uganda.* BMC research notes, 6(1), pp.1-7.

[82] Onduru, O.G., Mkakosya, R.S., Aboud, S. and Rumisha, S.F. (2021). *Genetic Determinants of Resistance among ESBL-Producing Enterobacteriaceae in Community and Hospital Settings in East, Central, and Southern Africa: A Systematic Review and Meta-Analysis of Prevalence.* Canadian Journal of Infectious Diseases and Medical Microbiology, 2021.

[83] Ampaire, L., Muhindo, A., Orikiriza, P., Mwanga-Amumpaire, J., Boum, Y. and Bebell, L. (2016). *A review of antimicrobial resistance in East Africa.* African journal of laboratory medicine, 5(1), pp.1-6.

[84] Lester, R., Musicha, P., Van Ginneken, N., Dramowski, A., Hamer, D.H., Garner, P. and Feasey, N.A. (2020). *Prevalence and outcome of bloodstream infections due to*

third-generation cephalosporin-resistant Enterobacteriaceae in sub-Saharan Africa: a systematic review. Journal of Antimicrobial Chemotherapy, 75(3), pp.492-507.

[85] Mukonzo, J.K., Namuwenge, P.M., Okure, G., Mwesige, B., Namusisi, O.K. and Mukanga, D. (2013). *Over-the-counter suboptimal dispensing of antibiotics in Uganda.* Journal of multidisciplinary healthcare, 6, p.303.

[86] Chikowe, I., Bliese, S. L., Lucas, S., and Lieberman, M. (2018). *Amoxicillin Quality and Selling Practices in Urban Pharmacies and Drug Stores of Blantyre, Malawi.* The American journal of tropical medicine and hygiene, 99(1), 233–238. https://doi.org/10.4269/ajtmh.18-0003

[87] WaterAid Uganda. https://www.wateraid.org/where-we-work/uganda

[88] WaterAid Malawi. https://www.wateraid.org/mw/health

[89] Moses, A., Bwanga, F., Boum, Y., and Bazira, J. (2014). *Prevalence and Genotypic Characterization of Extended-Spectrum Beta-Lactamases Produced by Gram Negative Bacilli at a Tertiary Care Hospital in Rural South Western Uganda.* British microbiology research journal, 4(12), 1541–1550. https://doi.org/10.9734/BMRJ/2014/9792

[90] Ampaire, L., Nduhura, E., and Wewedru, I. (2017). *Phenotypic prevalence of extended spectrum beta-lactamases among enterobacteriaceae isolated at Mulago National Referral Hospital: Uganda.* BMC research notes, 10(1), 448. https://doi.org/10.1186/s13104-017-2786-3

[91] Andrew, B., Kagirita, A. and Bazira, J. (2017). *Prevalence of Extended-Spectrum Beta-Lactamases-Producing Microorganisms in Patients Admitted at KRRH, Southwestern Uganda.* International journal of microbiology, 2017, 3183076. https://doi.org/10.1155/2017/3183076

[92] Kateregga, J. N., Kantume, R., Atuhaire, C., Lubowa, M. N. and Ndukui, J. G. (2015). *Phenotypic expression and prevalence of ESBL-producing Enterobacteriaceae in samples collected from patients in various wards of Mulago Hospital, Uganda.* BMC pharmacology & toxicology, 16, 14. https://doi.org/10.1186/s40360-015-0013-1

[93] Najjuka, C. F., Kateete, D. P., Kajumbula, H. M., Joloba, M. L. and Essack, S. Y. (2016). *Antimicrobial susceptibility profiles of Escherichia coli and Klebsiella pneumoniae isolated from outpatients in urban and rural districts of Uganda.* BMC research notes, 9, 235. https://doi.org/10.1186/s13104-016-2049-8

[94] Onduru, O. G., Mkakosya, R. S., Rumisha, S. F., and Aboud, S. (2021). *Carriage Prevalence of Extended-Spectrum β-Lactamase Producing Enterobacteriaceae in Outpatients Attending Community Health Centers in Blantyre, Malawi.* Tropical medicine and infectious disease, 6(4), 179. https://doi.org/10.3390/tropicalmed6040179

[95] Lautenbach, E., Patel, J.B., Bilker, W.B., Edelstein, P.H. and Fishman, N.O. (2001). *Extended-spectrum β-lactamase-producing Escherichia coli and Klebsiella pneumoniae: risk factors for infection and impact of resistance on outcomes.* Clinical Infectious Diseases, 32(8), pp.1162-1171.

[96] Søraas, A., Sundsfjord, A., Sandven, I., Brunborg, C. and Jenum, P.A. (2013). *Risk factors for community-acquired urinary tract infections caused by ESBL-producing enterobacteriaceae–a case–control study in a low prevalence country.* PloS one, 8(7), p.e69581.

[97] Ben-Ami, R., Rodríguez-Baño, J., Arslan, H., Pitout, J.D., Quentin, C., Calbo, E.S., Azap, Ö.K., Arpin, C., Pascual, A., Livermore, D.M. and Garau, J. (2009). *A multinational survey of risk factors for infection with extended-spectrum β-lactamase-producing Enterobacteriaceae in nonhospitalized patients.* Clinical Infectious Diseases, 49(5), pp.682-690.

[98] Buys, H., Muloiwa, R., Bamford, C. and Eley, B. (2016). *Klebsiella pneumoniae bloodstream infections at a South African children's hospital 2006–2011, a cross-sectional study.* BMC infectious diseases, 16(1), pp.1-10.

[99] Denis, B., Lafaurie, M., Donay, J.L., Fontaine, J.P., Oksenhendler, E., Raffoux, E., Hennequin, C., Allez, M., Socie, G., Maziers, N. and Porcher, R. (2015). *Prevalence, risk factors, and impact on clinical outcome of extended-spectrum beta-lactamase-producing Escherichia coli bacteraemia: a five-year study.* International journal of infectious diseases, 39, pp.1-6.

[100] Gorrie, C.L., Mirceta, M., Wick, R.R., Judd, L.M., Wyres, K.L., Thomson, N.R., Strugnell, R.A., Pratt, N.F., Garlick, J.S., Watson, K.M. and Hunter, P.C. (2018). *Antimicrobial-resistant Klebsiella pneumoniae carriage and infection in specialized geriatric care wards linked to acquisition in the referring hospital.* Clinical infectious diseases, 67(2), pp.161-170.

[101] Mshana, S.E., Falgenhauer, L., Mirambo, M.M., Mushi, M.F., Moremi, N., Julius, R., Seni, J., Imirzalioglu, C., Matee, M. and Chakraborty, T. (2016). *Predictors of bl a CTX-M-15 in varieties of Escherichia coli genotypes from humans in community settings in Mwanza, Tanzania.* BMC infectious diseases, 16(1), pp.1-9.

[102] Cocker, D., Sammarro, M., Chidziwisano, K., Elviss, N., Jacob, S.T., Kajumbula, H., Mugisha, L., Musoke, D., Musicha, P., Roberts, A.P. and Rowlingson, B. (2022). *Drivers of resistance in Uganda and Malawi (DRUM): a protocol for the evaluation of one-health drivers of extended spectrum beta lactamase (ESBL) resistance in low-middle income countries (LMICs).* Wellcome Open Research, 7(55), p.55.

[103] Moremi, N., Claus, H., Rutta, L., Frosch, M., Vogel, U. and Mshana, S.E. (2018). *High carriage rate of extended-spectrum beta-lactamase-producing Enterobacteriaceae among patients admitted for surgery in Tanzanian hospitals with a low rate of endogenous surgical site infections.* Journal of Hospital Infection, 100(1), pp.47-53.

[104] Tellevik, M.G., Blomberg, B., Kommedal, Ø., Maselle, S.Y., Langeland, N. and Moyo, S.J. (2016). *High prevalence of faecal carriage of ESBL-producing Enterobacteriaceae among children in Dar es Salaam, Tanzania.* PloS one, 11(12), p.e0168024.

[105] Sanneh, B., Kebbeh, A., Jallow, H.S., Camara, Y., Mwamakamba, L.W., Ceesay, I.F., Barrow, E., Sowe, F.O., Sambou, S.M., Baldeh, I. and Jallow, A. (2018). *Prevalence and risk factors for faecal carriage of Extended Spectrum β-lactamase producing Enterobacteriaceae among food handlers in lower basic schools in West Coast Region of The Gambia.* PLoS One, 13(8), p.e0200894.

[106] Olaru, I.D., Tacconelli, E., Yeung, S., Ferrand, R.A., Stabler, R.A., Hopkins, H., Aiken, A.M. and Kranzer, K. (2021). *The association between antimicrobial resistance and HIV infection: a systematic review and meta-analysis.* Clinical Microbiology and Infection, 27(6), pp.846-853.

[107] Herindrainy, P., Randrianirina, F., Ratovoson, R., Ratsima Hariniana, E., Buisson, Y., Genel, N., Decre, D., Arlet, G., Talarmin, A. and Richard, V. (2011). *Rectal carriage of extended-spectrum beta-lactamase-producing gram-negative bacilli in community settings in Madagascar.* PLoS One, 6(7), p.e22738.

[108] Farra, A., Frank, T., Tondeur, L., Bata, P., Gody, J.C., Onambele, M., Rafaï, C., Vray, M. and Breurec, S. (2016). *High rate of faecal carriage of extended-spectrum β-lactamase-producing Enterobacteriaceae in healthy children in Bangui, Central African Republic.* Clinical Microbiology and Infection, 22(10), pp.891-e1.

[109] Hilty, M., Betsch, B.Y., Bögli-Stuber, K., Heiniger, N., Stadler, M., Küffer, M., Kronenberg, A., Rohrer, C., Aebi, S., Endimiani, A. and Droz, S. (2012). *Transmission dynamics of extended-spectrum beta-lactamase–producing Enterobacteriaceae in the tertiary care hospital and the household setting.* Clinical infectious diseases, 55(7), pp.967-975.

[110] Lo, W.U., Ho, P.L., Chow, K.H., Lai, E.L., Yeung, F. and Chiu, S.S. (2010). *Fecal carriage of CTXM type extended-spectrum beta-lactamase-producing organisms by children and their household contacts.* Journal of infection, 60(4), pp.286-292.

[111] Haverkate, M.R., Platteel, T.N., Fluit, A.C., Stuart, J.C., Leverstein-van Hall, M.A., Thijsen, S.F., Scharringa, J., Kloosterman, R.C., Bonten, M.J. and Bootsma, M.C. (2017). *Quantifying within-household transmission of extended-spectrum beta-lactamase-producing bacteria.* Clinical Microbiology and Infection, 23(1), pp.46-e1.

[112] Martinez, E.P., Cepeda, M., Jovanoska, M., Bramer, W.M., Schoufour, J., Glisic, M., Verbon, A. and Franco, O.H. (2019). *Seasonality of antimicrobial resistance rates in respiratory bacteria: A systematic review and meta-analysis.* PloS one, 14(8), p.e0221133.

[113] Collignon, P., Beggs, J. J., Walsh, T. R., Gandra, S. and Laxminarayan, R. (2018). *Anthropological and socioeconomic factors contributing to global antimicrobial resistance: a univariate and multivariable analysis.* The Lancet. Planetary health, 2(9), e398–e405.

[114] Ercumen, A., Pickering, A.J., Kwong, L.H., Arnold, B.F., Parvez, S.M., Alam, M., Sen, D., Islam, S., Kullmann, C., Chase, C. and Ahmed, R. (2017). *Animal feces*

contribute to domestic fecal contamination: evidence from E. coli measured in water, hands, food, flies, and soil in Bangladesh. Environmental science & technology, 51(15), pp.8725-8734.

[115] Lindeberg, Y.L., Egedal, K., Hossain, Z.Z., Phelps, M., Tulsiani, S., Farhana, I., Begum, A. and Jensen, P.K.M. (2018). *Can Escherichia coli fly? The role of flies as transmitters of E. coli to food in an urban slum in Bangladesh.* Tropical Medicine & International Health, 23(1), pp.2-9.

[116] Pessinaba, C.N., Landoh, D.E., Dossim, S., Bidjada, B., Kere-Banla, A., Tamekloe, T.A., Doumbia, T., Douti, K., Bakonde, B.V. and Segbena, A.Y. (2018). *Screening for extended-spectrum beta-lactamase-producing Enterobacteriaceae intestinal carriage among children aged under five in Lomé, Togo.* Medecine et maladies infectieuses, 48(8), pp.551-554.

[117] Zambrano, L.D., Levy, K., Menezes, N.P. and Freeman, M.C. (2014). *Human diarrhea infections associated with domestic animal husbandry: a systematic review and meta-analysis.* Transactions of the Royal Society of Tropical Medicine and Hygiene, 108(6), pp.313-325.

[118] Navab-Daneshmand, T., Friedrich, M.N., Gächter, M., Montealegre, M.C., Mlambo, L.S., Nhiwatiwa, T., Mosler, H.J. and Julian, T.R. (2018). *Escherichia coli contamination across multiple environmental compartments (soil, hands, drinking water, and handwashing water) in urban Harare: correlations and risk factors.* The American journal of tropical medicine and hygiene, 98(3), p.803.

[119] Monira, S., Bhuyian, M.S.I., Parvin, T., Uddin, I.M., Zohura, F., Hasan, M.T., Biswas, S.K., Hasan, K., Masud, J., Rashid, M.U. and Rahman, Z. (2020). *Child mouthing of soil and presence of animals in child sleeping spaces are associated with growth faltering among young children in Dhaka, Bangladesh (CHoBI7 Program).* Tropical Medicine & International Health, 25(8), pp.1016-1023.

[120] Diggle, P.J., and Ribeiro, P. (2007). *Model-Based Geostatistics* Springer New York, 2007.

[121] Diggle, P.J., and Giorgi, E. (2019). *Model-Based Geostatistics for Global Public Health : Methods and Applications* CRC Press LLC.

[122] Galvin, S., Bergin, N., Hennessy, R., Hanahoe, B., Murphy, A.W., Cormican, M. and Vellinga, A. (2013). *Exploratory spatial mapping of the occurrence of antimicrobial resistance in E. coli in the community.* Antibiotics, 2(3), pp.328-338.

[123] Murray, C.J., Ikuta, K.S., Sharara, F., Swetschinski, L., Aguilar, G.R., Gray, A., Han, C., Bisignano, C., Rao, P., Wool, E. and Johnson, S.C. (2022). *Global burden of bacterial antimicrobial resistance in 2019: a systematic analysis.* The Lancet, 399(10325), pp.629-655.

[124] Liu, Y., Jiang, S., Liu, Y., Wang, R., Li, X., Yuan, Z., Wang, L. and Xue, F. (2011). *Spatial epidemiology and spatial ecology study of worldwide drug-resistant tuberculosis.* International Journal of Health Geographics, 10(1), pp.1-10.

[125] Kiffer, C.R., Camargo, E.C., Shimakura, S.E., Ribeiro, P.J., Bailey, T.C., Pignatari, A.C. and Monteiro, A. (2011). *A spatial approach for the epidemiology of antibiotic use and resistance in community-based studies: the emergence of urban clusters of Escherichia coli quinolone resistance in Sao Paulo, Brasil.* International journal of health geographics, 10(1), pp.1-10.

[126] Norman, P., Kemp, T. and Minton, J. (2016). *Antibiotic resistance: estimating the population level distribution of Extended-Spectrum Beta-Lactamases (ESBLs) in West Yorkshire, UK.*

[127] George, E.A., Sankar, S., Jesudasan, M.V., Sudandiradoss, C. and Nandagopal, B. (2015). *Molecular characterization of CTX-M type Extended Spectrum Beta Lactamase producing E. coli isolated from humans and the environment.* Indian journal of medical microbiology, 33, pp.S73-S79.

[128] Miller, E.A., Ponder, J.B., Willette, M., Johnson, T.J. and VanderWaal, K.L. (2020). *Merging metagenomics and spatial epidemiology to understand the distribution of antimicrobial resistance genes from Enterobacteriaceae in wild owls.* Applied and environmental microbiology, 86(20), pp.e00571-20.

[129] Birkegård, A.C., Ersbøll, A.K., Halasa, T., Clasen, J., Folkesson, A., Vigre, H. and Toft, N. (2017). *Spatial patterns of antimicrobial resistance genes in a cross-sectional sample of pig farms with indoor non-organic production of finishers.* Epidemiology Infection, 145(7), pp.1418-1430.

[130] French, N., Barrigas, M., Brown, P., Ribiero, P., Williams, N., Leatherbarrow, H., Birtles, R., Bolton, E., Fearnhead, P. and Fox, A. (2005). *Spatial epidemiology and natural population structure of Campylobacter jejuni colonizing a farmland ecosystem.* Environmental microbiology, 7(8), pp.1116-1126.

[131] Gauld, J.S. (2020). *Rivers, rainfall, and risk factors:geostatistical and epidemiological approaches to disentangle potential transmission routes of typhoid fever.* PhD thesis. https://doi.org/10.17635/lancaster/thesis/949

[132] Gómez-Barroso, D., García-Carrasco, E., Herrador, Z. et al. (2017). *Spatial clustering and risk factors of malaria infections in Bata district, Equatorial Guinea.* Malar J 16, 146. https://doi.org/10.1186/s12936-017-1794-z

[133] Tildesley, M.J., House, T.A., Bruhn, M.C., Curry, R.J., O'Neil, M., Allpress, J.L., Smith, G. and Keeling, M.J. (2010). *Impact of spatial clustering on disease transmission and optimal control.* Proceedings of the National Academy of Sciences, 107(3), pp.1041-1046.

[134] Cordes, J. and Castro, M.C. (2020). *Spatial analysis of COVID-19 clusters and contextual factors in New York City.* Spatial and spatio-temporal epidemiology, 34, p.100355.

[135] Ramírez-Aldana, R., Gomez-Verjan, J.C. and Bello-Chavolla, O.Y. (2020). *Spatial analysis of COVID-19 spread in Iran: Insights into geographical and structural transmission determinants at a province level.* PLoS neglected tropical diseases, 14(11), p.e0008875.

[136] Gomes, D.S., Andrade, L.A., Ribeiro, C.J.N., Peixoto, M.V.S., Lima, S.V.M.A., Duque, A.M., Cirilo, T.M., Góes, M.A.O., Lima, A.G.C.F., Santos, M.B. and Araújo, K.C.G.M. (2020). *Risk clusters of COVID-19 transmission in northeastern Brazil: prospective space–time modelling.* Epidemiology Infection, 148.

[137] Andrade-Pacheco, R., Rerolle, F., Lemoine, J., Hernandez, L., Meïté, A., Juziwelo, L., Bibaut, A. F., van der Laan, M. J., Arnold, B. F. and Sturrock, H. (2020). *Finding hotspots: development of an adaptive spatial sampling approach.* Scientific reports, 10(1), 10939. https://doi.org/10.1038/s41598-020-67666-3

[138] Gilks, W.R., Richardson, S. and Spiegelhalter, D. eds., (1995). *Markov chain Monte Carlo in practice.* CRC press.

[139] Chipeta, M., Terlouw, D., Phiri, K. and Diggle, P. (2016). *Inhibitory geostatistical designs for spatial prediction taking account of uncertain covariance structure.* Environmetrics.

[140] Delmelle, E. (2009). *Spatial sampling.* The SAGE handbook of spatial analysis, 183, p.206.

[141] Cochran, W.G. (1977). *Sampling Techniques* 3rd ed. Wiley, New York

[142] Wang, J.F., Jiang, C.S., Li, L.F. and Hu, M.G. (2009). *Spatial Sampling and Inferences*, Science Press, Beijing.

[143] Berry, B.J.L. and Baker, A.M. (1968). *Geographic sampling.*, In: Berry B.J.L. and Marble D.F. (eds), Spatial Analysis: a Reader in Statistical Geography; pp. 91–100. Prentice-Hall: Englewood Cliffs, N.J.

[144] Zimmerman, D. L. (2006). *Optimal network design for spatial prediction, covariance parameter estimation, and empirical prediction.* Environmetrics 17 (6), 635–652.

[145] Müller, W.G., 2007. *Collecting spatial data: optimum design of experiments for random fields.* Springer Science Business Media.

[146] Martin, R. J. (2001). *Comparing and contrasting some environmental and experimental design problems.* Environmetrics 12, 303-317

[147] Diggle, P. and Lophaven, S. (2006). *Bayesian Geostatistical Design.* Scandinavian Journal of Statistics, 33(1), pp.53-64.

[148] Olea, R.A. (1984). *Sampling design optimization for spatial functions.* Mathematical Geology 16, 369–392. https://doi.org/10.1007/BF01029887

[149] Lark, R.M. (2002). *Optimized spatial sampling of soil for estimation of the variogram by maximum likelihood* Geoderma, Volume 105, Issues 1–2, Pages 49-80, ISSN 0016-7061. https://doi.org/10.1016/S0016-7061(01)00092-1

[150] Müller, W.G. and Stehlík, M. (2010). *Compound optimal spatial designs.* Environmetrics, 21: 354-364. https://doi.org/10.1002/env.1009

[151] Chipeta, M. G., Rowlingson B. and Diggle, P. J. (2019). *geosample: An R package for geostatistical sampling designs.*

[152] Porto de Albuquerque, J., Yeboah, G., Pitidis V. and Ulbrich P. (2019). *osmgeosample: Construction of Geostatistical Sampling Designs with OSM Data.* doi:10.13140/RG.2.2.13710.20804.

[153] Engel, C. (2017). *Introduction to Spatial Data Types in R. [online]* `https://cengel.github.io/rspatial/2_spDataTypes.nb.html`.

[154] Sadler, J. (2018). *Introduction to GIS with R. [online]* `https://www.jessesadler.com/post/gis-with-r-intro/`

[155] Brown, C. (2018). *sf or sp for spatial R programming [online]* `http://www.seascapemodels.org/rstats/2018/03/23/should-I-learn-sp-or-sf.html`

[156] Huber, A. and Mosler, H. (2013). *Determining behavioral factors for interventions to increase safe water consumption: a cross-sectional field study in rural Ethiopia.* International Journal of Environmental Health Research, 23(2), pp.96-107.

[157] Morse, T., Chidziwisano, K., Musoke, D., Cocker, D. and Feasey, N. (2019). *Development of a protocol for assessing the role of WASH in AMR distribution in the environment.* In: Water and Health: Where Science Meets Policy, 2019-10-07 - 2019-10-11, University of North Carolina.

[158] Darton, T.C., Meiring, J.E., Tonks S. on behalf of the STRATAA Study Consortium, et al. (2017). *The STRATAA study protocol: a programme to assess the burden of enteric fever in Bangladesh, Malawi and Nepal using prospective population census, passive surveillance, serological studies and healthcare utilisation surveys.* BMJ Open 2017;7:e016283. doi: 10.1136/bmjopen-2017-016283.

[159] Lautenbach E., Strom B.L., Bilker W.B., Patel J.B., Edelstein P.H. and Fishman N.O. (2001). *Epidemiological Investigation of Fluoroquinolone Resistance in Infections Due to Extended-Spectrum β-Lactamase—Producing Escherichia coli and Klebsiella pneumoniae.* Clinical Infectious Diseases, Volume 33, Issue 8, 15 October 2001, Pages 1288–1294, https://doi.org/10.1086/322667

[160] Gauld, J.S., Olgemoeller, F., Heinz, E., Nkhata, R., Bilima, S., Wailan, A.M., Kennedy, N., Mallewa, J., Gordon, M.A., Read, J.M. and Heyderman, R.S. (2021).

Spatial and genomic data to characterize endemic typhoid transmission. Clinical Infectious Diseases.

[161] UNFPA Malawi. (2019). *Malawi 2018 Population and Housing Census Main Report.* Retrieved August 20, 2022, from https://malawi.unfpa.org/en/resources/malawi-2018-population-and-housing-census-main-report

[162] World Bank. (2019). *Poverty & Equity Brief. Malawi.* Retrieved August 20, 2022, from https://databankfiles.worldbank.org/data/download/poverty/33EF03BB-9722-4AE2-ABC7-AA2972D68AFE/Archives-2019/Global_POVEQ_MWI.pdf

[163] Hastie, T. and Hastie, M.T. (2020). *Package 'gam'*. GAM Package CRAN, https://cran.r-project.org/web/packages/gam/gam.pdf

[164] Bozdogan H. (1987). *Model Selection and Akaike's Information Criterion (AIC): The General Theory and Its Analytical Extensions.* Psychometrika, 52, 345–370.

[165] Bates D., Mächler M., Bolker B. and Walker S. (2015). *Fitting Linear Mixed-Effects Models Using lme4.* Journal of Statistical Software, 67(1), 1–48. doi: 10.18637/jss.v067.i01.

[166] Kuznetsova A., Brockhoff P.B. and Christensen R.H.B. (2017). *lmerTest Package: Tests in Linear Mixed Effects Models.*, Journal of Statistical Software, 82(13), 1–26. doi: 10.18637/jss.v082.i13.

[167] Stan Development Team. (2020). *Stan Modeling Language User's Guide and Reference Manual*, 2.27. https://mc-stan.org

[168] Stan Development Team (2020). *RStan: the R interface to Stan.*, R package version 2.21. http://mc-stan.org/

[169] Neal, R. (2011). *MCMC Using Hamiltonian Dynamics.* Handbook of Markov Chain Monte Carlo, edited by Steve Brooks, Andrew Gelman, Galin L. Jones, and Xiao-Li Meng, 116–62. Chapman; Hall/CRC.

[170] Hoffman, M. D., and Gelman A. (2014). *The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo.* Journal of Machine Learning Research 15: 1593–623. http://jmlr.org/papers/v15/hoffman14a.html

[171] Stan Development Team (2020). *Brief Guide to Stan's Warnings.* https://mc-stan.org/misc/warnings.html

[172] Gelman, A., Carlin, J.B., Stern, H.S. and Rubin, D.B. (1995). *Bayesian data analysis.* Chapman and Hall/CRC.

[173] Lewis, J.M., Mphasa, M., Banda, R., Beale, M.A., Heinz, E., Mallewa, J., Jewell, C., Faragher, B., Thomson, N.R. and Feasey, N.A. (2021). *Dynamics of gut mucosal colonisation with extended spectrum beta-lactamase producing Enterobacterales in Malawi.* medRxiv.

[174] Albrechtova, K., Dolejska, M., Cizek, A., Tausova, D., Klimes, J., Bebora, L. and Literak, I. (2012). *Dogs of nomadic pastoralists in northern Kenya are reservoirs of plasmid-mediated cephalosporin-and quinolone-resistant Escherichia coli, including pandemic clone B2-O25-ST131.* Antimicrobial agents and chemotherapy, 56(7), pp.4013-4017.

[175] Chereau, F., Herindrainy, P., Garin, B., Huynh, B.T., Randrianirina, F., Padget, M., Piola, P., Guillemot, D. and Delarocque-Astagneau, E. (2015). *Colonization of extended-spectrum-β-lactamase-and NDM-1-producing Enterobacteriaceae among pregnant women in the community in a low-income country: a potential reservoir for transmission of multiresistant Enterobacteriaceae to neonates.* Antimicrobial agents and chemotherapy, 59(6), pp.3652-3655.

[176] MacPherson, E., Reynolds, J., Sanudi, E., Nkaombe, A., Mankhomwa, J., Dixon, J. and Chandler, C.I. (2021). *Understanding antimicrobial use in subsistence farmers in Chikwawa District Malawi, implications for public awareness campaigns.* https://doi.org/10.31235/osf.io/e7b6n

[177] Omulo, S., Lofgren, E.T., Lockwood, S., Thumbi, S.M., Bigogo, G., Ouma, A., Verani, J.R., Juma, B., Njenga, M.K., Kariuki, S. and McElwain, T.F. (2021). *Carriage of antimicrobial-resistant bacteria in a high-density informal settlement in Kenya is associated with environmental risk-factors.* Antimicrobial Resistance & Infection Control, 10(1), pp.1-12.

Appendix A

Appendix: Spatial design for the DRUM study areas

A.1 R code for implementation of the design

Libraries and notes

```
# Libraries
# library(raster)
# library(rgdal)
# library(sf)
# library(sp)
# library(tmap)
# library(ggplot2)
# library(osmdata)

# The final icpSample function needs:
# @param n: total number of points
# @param k: number of close pairs
# @param delta: minimum distance between points
# @param zeta: radius for close pairs
# @param poly: polygon "sfc_POLYGON" with a crs in Arc 1960 / UTM zone
# @param proposal: choose between ("density","census","osm","")
# ("" is for general sampling)

# If multiple polygons (Kampala), use the separate icpSampleM function
# @param poly: union of the polygons in this case

# Specific:
# Raster => "yourraster" raster of choice, tif file (usually)
# Multiple => "pol" list of polygons (not the same as @param poly)
# Census => "yourcensus" matrix of census households gps coordinates
#           (1st and 2nd column :long/lat)
# OSM => 2 char "crs" and "crs2": respectively crs of the polygon
#       in WGS84 and Arc 1960
```

SamplePoint function

```
# For population density-weighted sites (K)

sampleWeightedPoints <- function(r,
                                 size,
                                 kappa,
                                 method = c("power", "mixture"),
                                 doPlot = T) {
  pts <- raster::rasterToPoints(r, spatial = T)
  names(pts) <- c("density")
  tosample <- NULL
  if (method[1] == "power") {
    tosample <- sample(
      nrow(pts),
      size = size,
      prob = pts$density^kappa,
      replace = T
    )
```

```
  }
  else if (method[1] == "mixture") {
    p <- pts$density / sum(pts$density)
    p <- 1 - (1 - p * kappa) * (1 - (1 - kappa) / nrow(pts))
    tosample <- sample(nrow(pts),
      size = size,
      prob = p,
      replace = T
    )
  }
  else {
    stop("method must be one of c('power','mixture')")
  }

  rv <- data.frame(density = pts[tosample, ]$density)
  xy <- sp::coordinates(pts[tosample, ])
  xy[, 1] <- jitter(xy[, 1], amount = raster::res(r)[1] / 2)
  xy[, 2] <- jitter(xy[, 2], amount = raster::res(r)[2] / 2)
  sp::coordinates(rv) <- xy
  sp::proj4string(rv) <- sp::proj4string(pts)
  if (doPlot) {
    plot(r)
    plot(rv, pch = "+", add = T)
  }
  rv
}

samplePt <- function(poly, yourraster) {
  while (TRUE) {
    pt <- sampleWeightedPoints(
      r = yourraster, size = 1, kappa = 1.0,
      method = "mixture", doPlot = F
    )
    if (sf::st_within(sf::st_point(pt@coords), poly, sparse = FALSE)) {
      return(sf::st_point(pt@coords))
    }
  }
}

# General

samplePoint <- function(poly) {
  box <- sf::st_bbox(poly)
  xcoord <- stats::runif(1, min = box$xmin, max = box$xmax)
  ycoord <- stats::runif(1, min = box$ymin, max = box$ymax)
  pt <- sf::st_point(matrix(c(xcoord, ycoord), nrow = 1))
  if (sf::st_within(pt, poly, sparse = FALSE) == FALSE) {
    samplePoint(poly)
  } else {
    return(pt)
  }
}
```

```r
# Census
sampPt <- function(poly, censusxy) {
  samp <- censusxy[sample(nrow(censusxy), 1), ]
  pt <- sf::st_point(matrix(c(samp[1], samp[2]), nrow = 1))
  if (sf::st_within(pt, poly, sparse = FALSE) == FALSE) {
    sampPt(poly)
  } else {
    return(pt)
  }
}

# OSM

get_buildings <- function(poly) {
  osmd <- osmdata::opq(bbox = sf::st_bbox(sf::st_transform(poly, 4326)))
  osmd <- osmdata::add_osm_feature(osmd, key = "building")
  osmd <- osmdata::osmdata_sf(osmd)

  bldgs <- sf::st_centroid(osmd$osm_polygons)

  bldgs <- sf::st_transform(bldgs, sf::st_crs(poly))
  bldgs <- sf::st_coordinates(bldgs)
  names(bldgs) <- c("longitude", "latitude")
  return(bldgs)
}


sampPtH <- function(poly, bldgs) {
  samp <- bldgs[sample(nrow(bldgs), 1), c("longitude", "latitude")]
  pt <- sf::st_point(matrix(c(as.numeric(samp[1]), as.numeric(samp[2])), nrow = 1))
  if (sf::st_within(pt, poly, sparse = FALSE) == FALSE) {
    sampPtH(poly, bldgs)
  } else {
    return(pt)
  }
}
```

Minimum distance

```r
minDistance <- function(pt, pts) {
  d <- c()
  for (j in 1:dim(pts)[1]) {
    d[j] <- sf::st_distance(pt, sf::st_point(pts[j, ]))
  }
  return(min(d))
}
```

InhibSample

```
# Population-density weighted

inhibSampleD <- function(n, k, poly, rdensity) {
  # Sample first point
  pts <- samplePt(poly, rdensity)
  # Sample the rest
  i <- 1
  while (i < n) {
    pt <- samplePt(poly, rdensity)
    if (minDistance(pt, pts) > k) {
      pts <- rbind(pts, pt)
      i <- i + 1
    }
  }
  return(pts)
}

# General

inhibSample <- function(n, k, poly) {
  # Sample first point
  pts <- samplePoint(poly)
  # Sample the rest
  i <- 1
  while (i < n) {
    pt <- samplePoint(poly)
    if (minDistance(pt, pts) > k) {
      pts <- rbind(pts, pt)
      i <- i + 1
    }
  }
  return(pts)
}

# Multiple polygons

whichpol <- function(pt, pol) {
  a <- c()
  for (j in 1:3) {
    a[j] <- sp::point.in.polygon(sf::st_coordinates(pt)[, 1],
      sf::st_coordinates(pt)[, 2],
      sf::st_coordinates(pol[[j]])[, 1],
      sf::st_coordinates(pol[[j]])[, 2],
      mode.checked = FALSE
    )
  }
  return(which(a == 1))
}
```

```
inhibSampleM <- function(n, k, poly, pol) {

  ### Sample first point
  nump <- rep(0, length(pol))
  pts <- samplePoint(poly)
  x <- whichpol(pts, pol)[1]
  nump[x] <- nump[x] + 1
  ## Sample the rest

  while (nrow(pts) < n) {
    pt <- samplePoint(poly)
    x <- whichpol(pt, pol)
    x <- x[1] # Guard against overlapping polygons
    if ((minDistance(pt, pts) > k) & (nump[x] < (n / length(pol)))) {
      pts <- rbind(pts, pt)
      nump[x] <- nump[x] + 1
    } else {
      samplePoint(poly)
    }
    print(nump)
  }
  return(pts)
}

# Census

inhibSampleC <- function(n, k, poly, censusxy) {
  # Sample first point
  pts <- sampPt(poly, censusxy)
  # Sample the rest
  i <- 1
  while (i < n) {
    pt <- sampPt(poly, censusxy)
    if (minDistance(pt, pts) > k) {
      pts <- rbind(pts, pt)
      i <- i + 1
    }
  }
  return(pts)
}

# OSM

inhibSampleO <- function(n, k, poly, bldgs) {
  # Sample first point
  pts <- sampPtH(poly, bldgs)
  # Sample the rest
  i <- 1
  while (i < n) {
    pt <- sampPtH(poly, bldgs)
    if (minDistance(pt, pts) > k) {
      pts <- rbind(pts, pt)
      i <- i + 1
```

```
    }
  }
  return(pts)
}
```

SamplePtinRadius

```
samplePtInRadius <- function(pt, zeta) {
  print(pt)
  theta <- stats::runif(1, 0, 2 * pi)
  r <- zeta
  kPt <- sf::st_coordinates(pt) + r * c(cos(theta), sin(theta))
  print(kPt)
  return(sf::st_point(kPt, sp::CRS(sp::proj4string(pt))))
}
```

Final function: icpSample

```
icpSample <- function(n, k, delta, zeta, poly, proposal, ...)
{
  {    if (proposal == "density") {
    inhibS <- inhibSampleD # needs rdensity raster
  } else if (proposal == "census") {
    inhibS <- inhibSampleC # needs censusxy matrix of xy coords
  } else if (proposal == "osm") {
    inhibS <- inhibSampleO # needs bldgs OSM building data
  } else {
    inhibS <- inhibSample
  }  }

  inhibPts <- inhibS(n - k, delta, poly, ...)
  kPts <- inhibPts[sample(nrow(inhibPts), k), ]
  cpPts <- lapply(1:dim(kPts)[1], function(i) {
    while (TRUE) {
      kpt <- samplePtInRadius(sf::st_point(kPts[i, ]), zeta)
      if (sf::st_within(kpt, poly, sparse = FALSE)) {
        return(kpt)
      }
    }
  })
  cpPts <- do.call("rbind", cpPts)
  rbind(inhibPts, cpPts)
}

# If multiple consecutive polygons (ex. Kampala) use this function instead
# Checks for same number of inhibitory points and close pairs in each polygon
# Based on 3 polygons
```

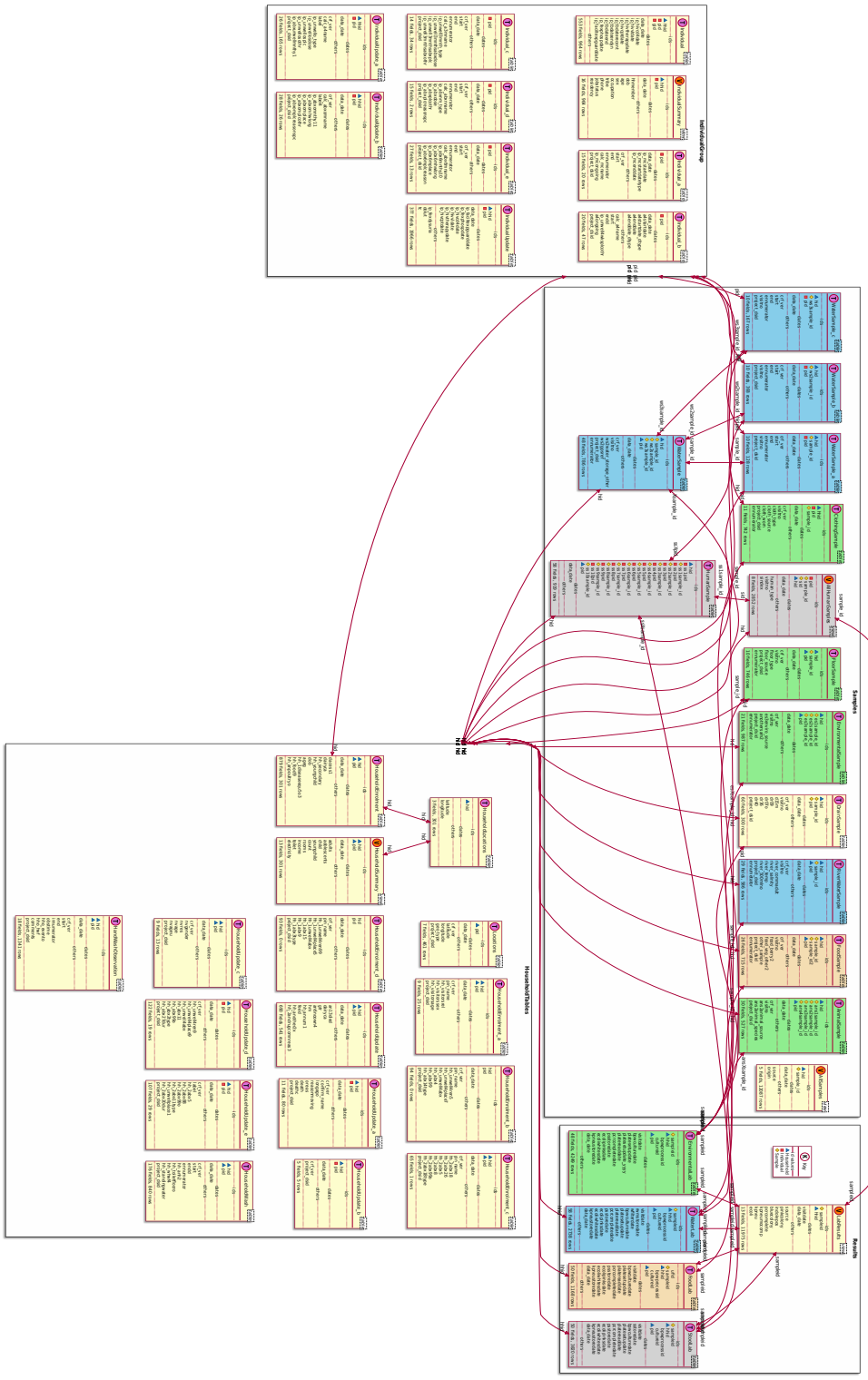
```r
icpSampleM <- function(n, k, delta, zeta, poly, pol) {
  inhibPts <- inhibSampleM(n - k, delta, poly, pol)
  vec <- c()
  v <- c()
  for (i in 1:(n - k)) {
    for (j in 1:length(pol)) {
      vec[j] <- sp::point.in.polygon(sf::st_coordinates(sf::st_point(inhibPts[i, ]))[1],
        sf::st_coordinates(sf::st_point(inhibPts[i, ]))[2],
        sf::st_coordinates(pol[[j]])[, 1],
        sf::st_coordinates(pol[[j]])[, 2],
        mode.checked = FALSE
      )
    }
    v[i] <- as.numeric(which(vec == 1))
  }
  inhibPts <- cbind(inhibPts, v)
  # change kPts if number of polygons != 3
  kPts <- inhibPts[c(
    sample(which(inhibPts[, length(pol)] == 1), k / length(pol)),
    sample(which(inhibPts[, length(pol)] == 2), k / length(pol)),
    sample(which(inhibPts[, length(pol)] == 3), k / length(pol))
  ), 1:2]

  cpPts <- lapply(1:dim(kPts)[1], function(i) {
    while (TRUE) {
      kpt <- samplePtInRadius(sf::st_point(kPts[i, ]), zeta)
      if (sf::st_within(kpt, poly, sparse = FALSE)) {
        return(kpt)
      }
    }
  })
  cpPts <- do.call("rbind", cpPts)
  rbind(inhibPts[, 1:2], cpPts)
}


# output:matrix of n points (n-k inhibitory points and k close pairs)
```

Appendix B

Appendix: Transmission patterns of ESBL-producing *E. coli* and *K. pneumoniae* leading to human gut mucosal colonisation in the community in Southern Malawi

B.1 DRUM database schema

©Barry Rowlingson



B.2 R code for the forwards selection algorithm

AIC calculation

```
calcAIC = function(data, f_string) {
  if (grepl("|",f_string,fixed=TRUE))
    fit = with(data,
               expr=glmer(as.formula(f_string), family=binomial("logit"),
                          control=glmerControl(optimizer="bobyqa",
                                               optCtrl = list(maxfun = 100000))))
  else fit = with(data, expr=glm(as.formula(f_string), family=binomial("logit")))
  AIC(fit)
}
```

Formula building

```
build_formula = function(y, x) {
  paste(y, paste(x, collapse='+'), sep='~')
}
```

Forwards selection algorithm

```
stepwise = function(data, y_name, initx, x_names) {

  initf = build_formula(y_name, initx)
  message("Fitting formula: ", initf)
  aic = calcAIC(data, initf)
  new_aics=c()
  for (i in 1:length(x_names)){
    f=build_formula(y_name, c(initx,x_names[i]))
    print(f)
    new_aics[i] = calcAIC(data, f)}

  delta = aic - new_aics
  if(all(delta < 0)) return(initf)

  initx=c(initx,x_names[which.max(delta)])
  x_names = x_names[-which.max(delta)]
  stepwise(data, y_name, initx, x_names)
}
```

Adapted from Chris Jewell's backwards selection algorithm for GLM

B.3 Gaussian Processes: parameters, maps and diagnostic plots

B.3.1 ESBL-producing *E. coli* in Ndirande

Table B.1: Parameter estimates for the spatial model in Ndirande (ESBL-Ec)

	Log-odds	**Std error**	**2.5%**	**97.5%**
Intercept	-1.097	1.480	-4.113	1.895
In-house prevalence	0.224	0.068	0.106	0.373
Harmonic term (cosday)	1.368	0.519	0.442	2.486
Harmonic term (cosday2)	1.083	0.483	0.218	2.109
Number of days since the first sample	0.373	0.200	0.004	0.797

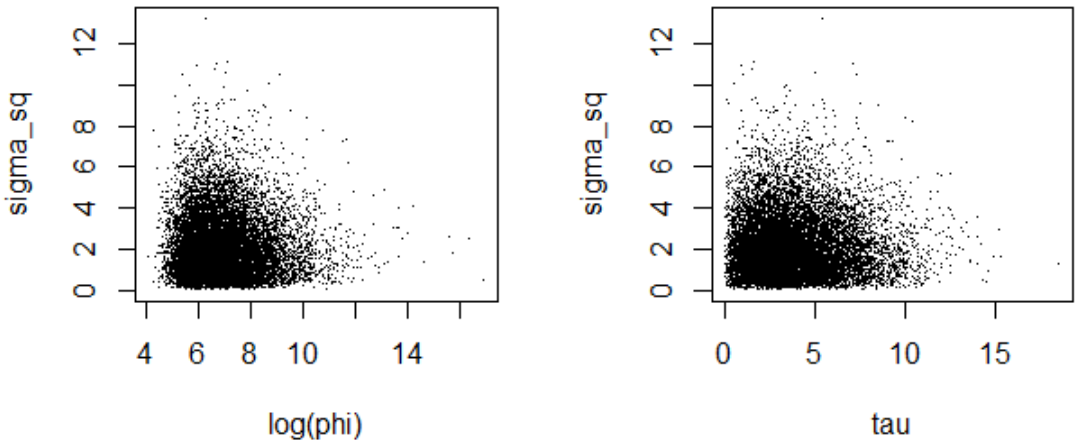


Figure B.1: Pairwise correlation plots for Ndirande (ESBL-Ec)

B.3.2 ESBL-producing *K. pneumoniae* in Ndirande

Table B.2: Parameter estimates for the spatial model in Ndirande (ESBL-K)

	Median	Mean	Std error	2.5%	97.5%
ϕ	862.05	4066.09	50864.09	176.03	18456.09
σ^2	1.65	1.98	1.41	0.24	5.57
τ	2.49	2.86	1.85	0.41	7.43

	Log-odds	Std error	2.5%	97.5%
Intercept	-3.074	1.507	-6.195	-0.025
Number of colonised people in the house	0.673	0.333	0.038	1.370

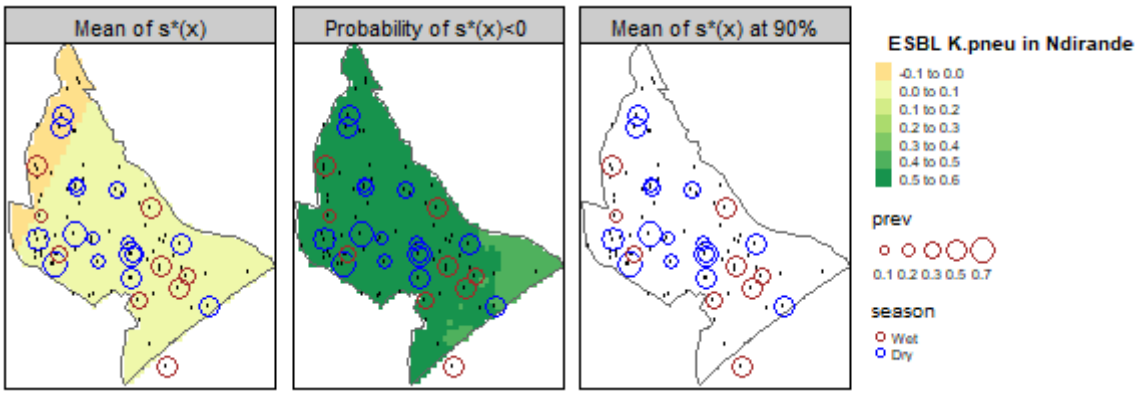


Figure B.2: Gaussian process maps for Ndirande (ESBL-K)

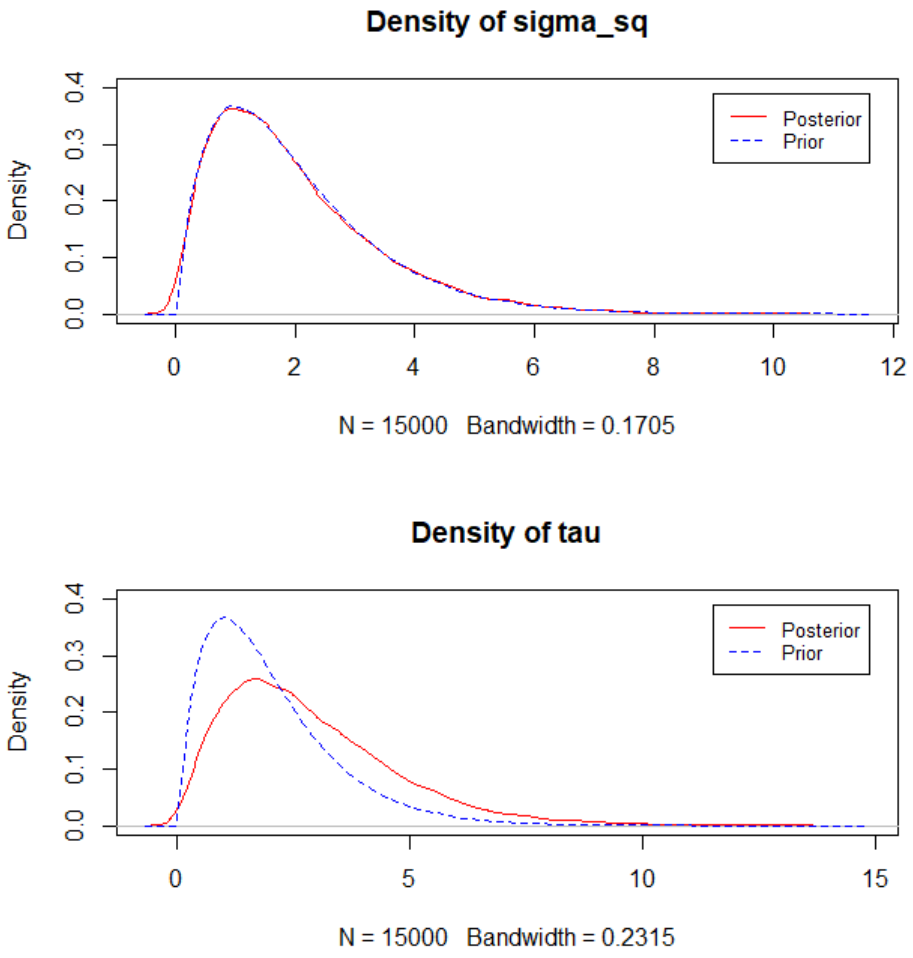


Figure B.3: Density plots of σ^2 and τ for Ndirande (ESBL-K)

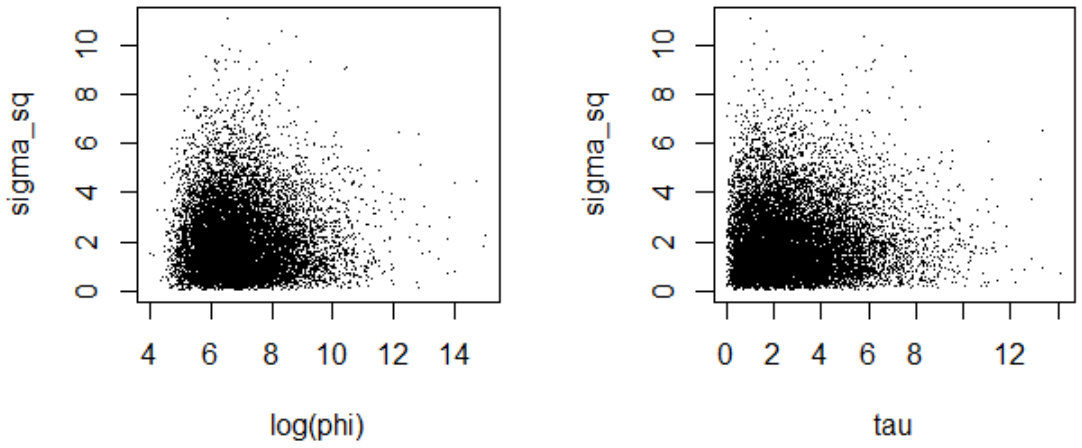


Figure B.4: Pairwise correlation plots for Ndirande (ESBL-K)

B.3.3 ESBL-producing *E. coli* in Chikwawa

Table B.3: Parameter estimates for the spatial model in Chikwawa (ESBL-Ec)

	Median	Mean	Std error	2.5%	97.5%
ϕ	792.22	3942.6	6.2e+04	167.92	1.6e+04
σ^2	1.72	2.04	1.42	0.25	5.62
τ	2.94	3.32	2.06	0.49	8.31

	Log-odds	Std error	2.5%	97.5%
Intercept	-2.005	1.511	-4.991	1.084
In-house prevalence	0.374	0.107	0.191	0.611
Harmonic term (cosday)	1.301	0.619	0.174	2.633
Use of cotrimoxazole in the last 6 months	-0.454	0.322	-1.155	0.116

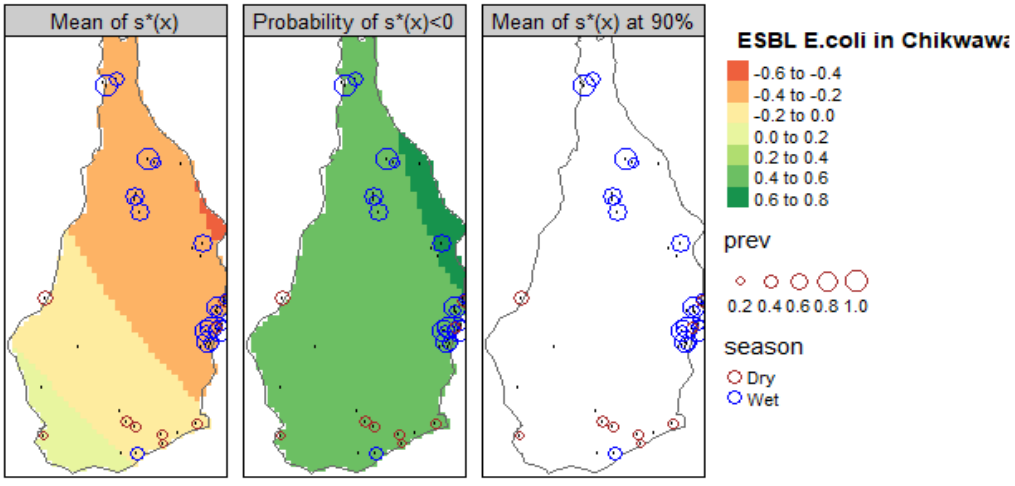


Figure B.5: Gaussian process maps for Chikwawa (ESBL-Ec)

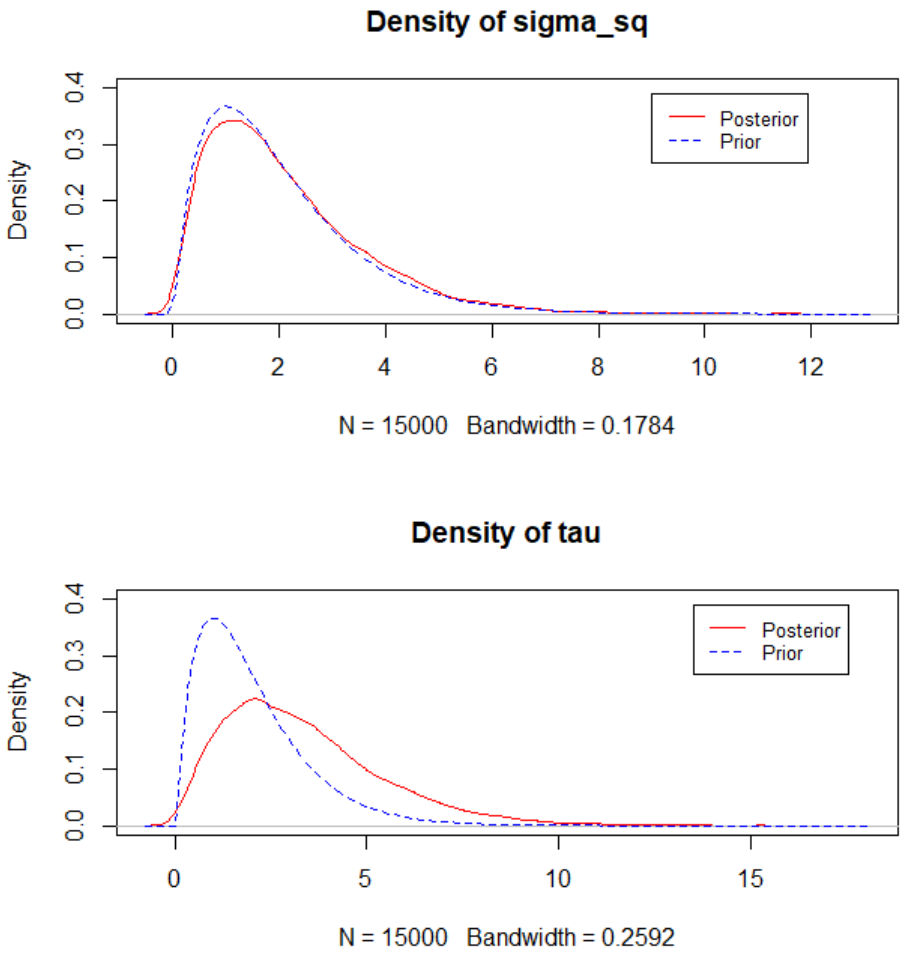


Figure B.6: Density plots of σ^2 and τ for Chikwawa (ESBL-Ec)

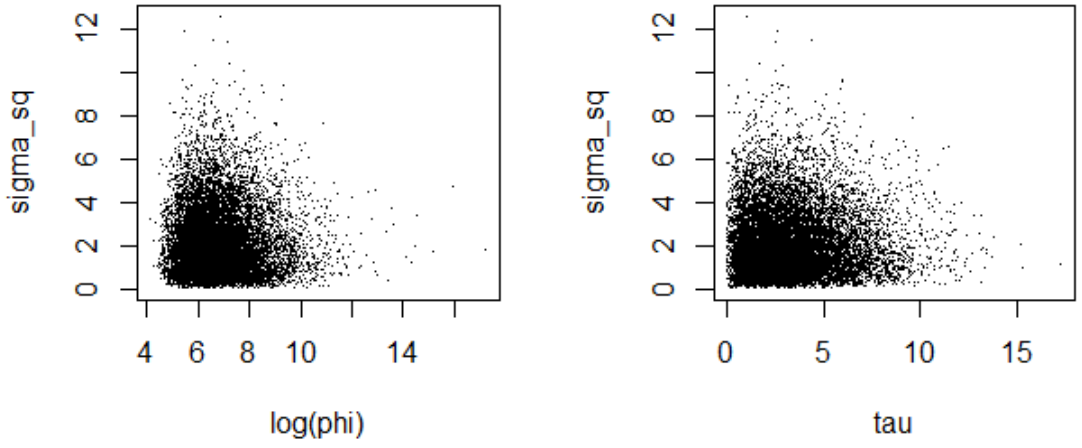


Figure B.7: Pairwise correlation plots for Chikwawa (ESBL-Ec)

B.3.4 ESBL-producing *K. pneumoniae* in Chikwawa

Table B.4: Parameter estimates for the spatial model in Chikwawa (ESBL-K)

	Median	**Mean**	**Std error**	**2.5%**	**97.5%**
ϕ	811.46	3431.77	39039.85	178.21	16002.80
σ^2	1.73	2.06	1.43	0.27	5.70
τ	2.38	2.75	1.83	0.35	7.23

	Log-odds	**Std error**	**2.5%**	**97.5%**
Intercept	-4.113	1.579	-7.304	-0.962
Number of colonised people in the house	1.737	0.539	0.764	2.898

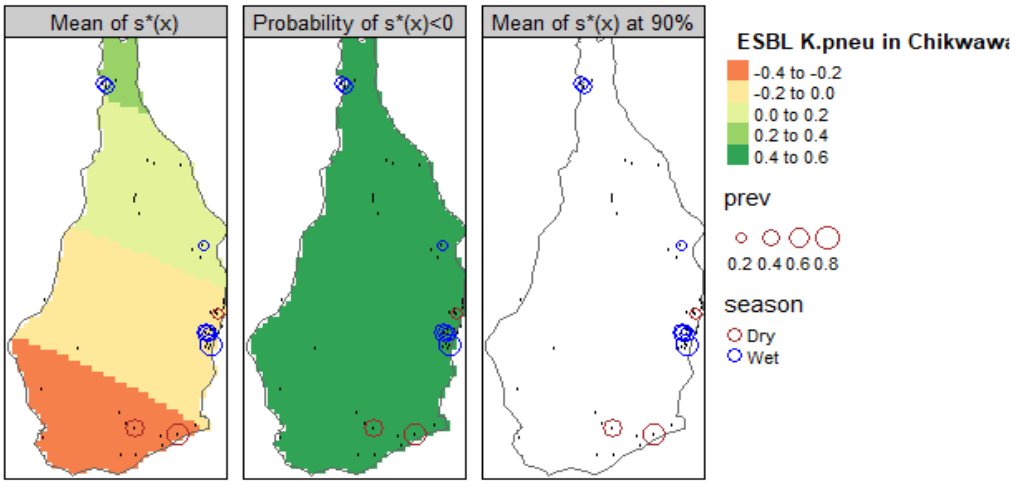


Figure B.8: Gaussian process maps for Chikwawa (ESBL-K)

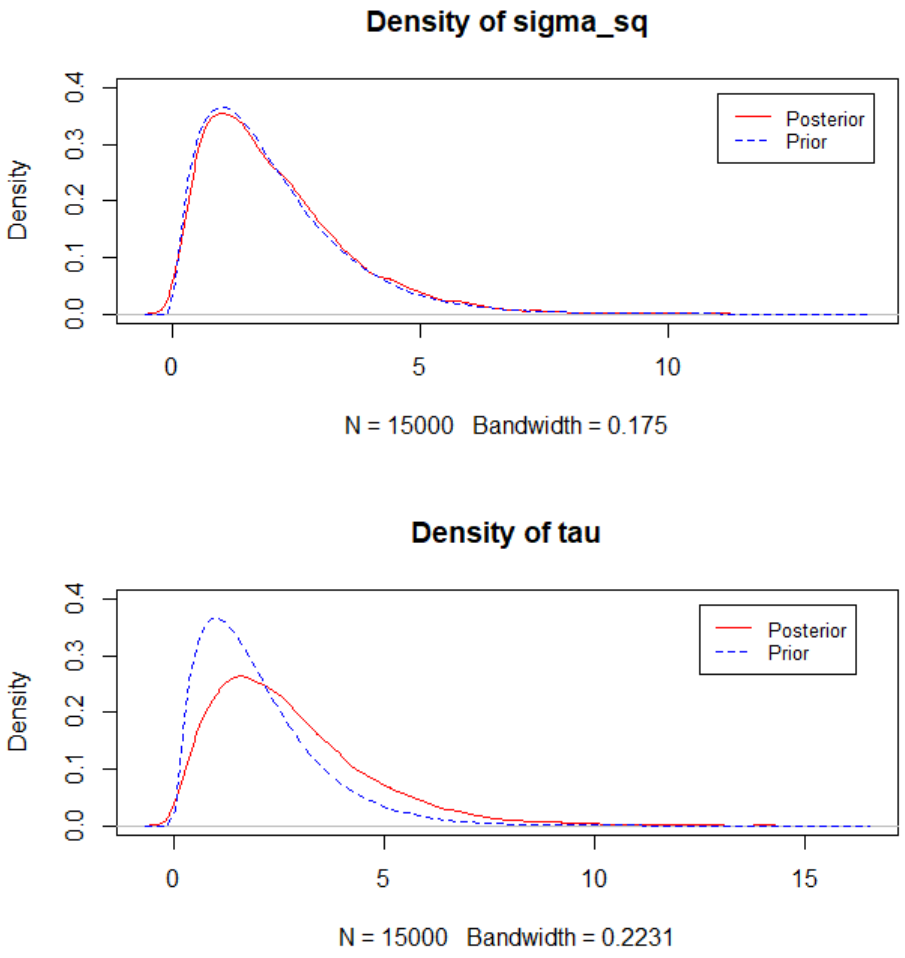


Figure B.9: Density plots of σ^2 and τ for Chikwawa (ESBL-K)

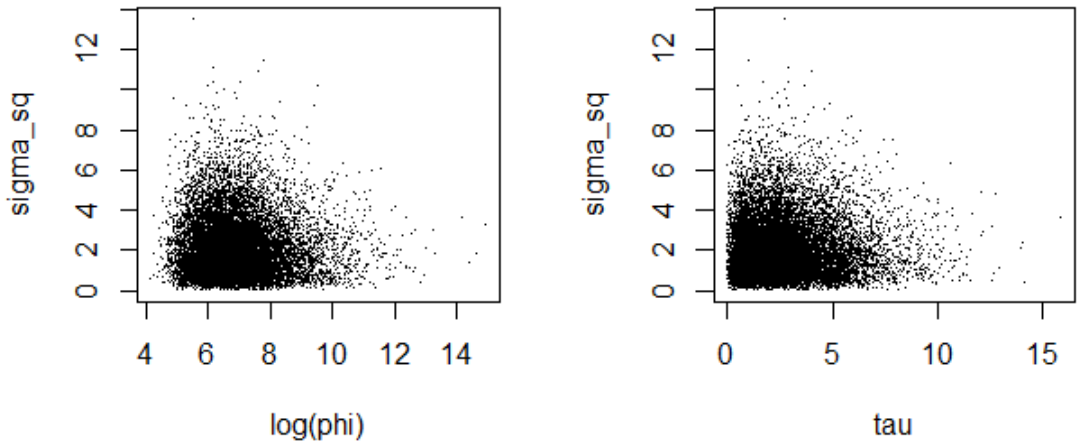


Figure B.10: Pairwise correlation plots for Chikwawa (ESBL-K)

B.3.5 ESBL-producing *E. coli* in Chileka

Table B.5: Parameter estimates for the spatial model in Chileka (ESBL-Ec)

	Median	Mean	Std error	2.5%	97.5%
ϕ	886.48	3776.68	4e+04	183.8	1.8e+04
σ^2	1.63	1.94	1.37	0.23	5.39
τ	2.90	3.25	1.96	0.52	7.97

	Log-odds	Std error	2.5%	97.5%
Intercept	-1.384	1.459	-4.357	1.573
In-house prevalence	0.234	0.078	0.096	0.404
Number of days since the first sample	0.635	0.366	-0.038	1.402

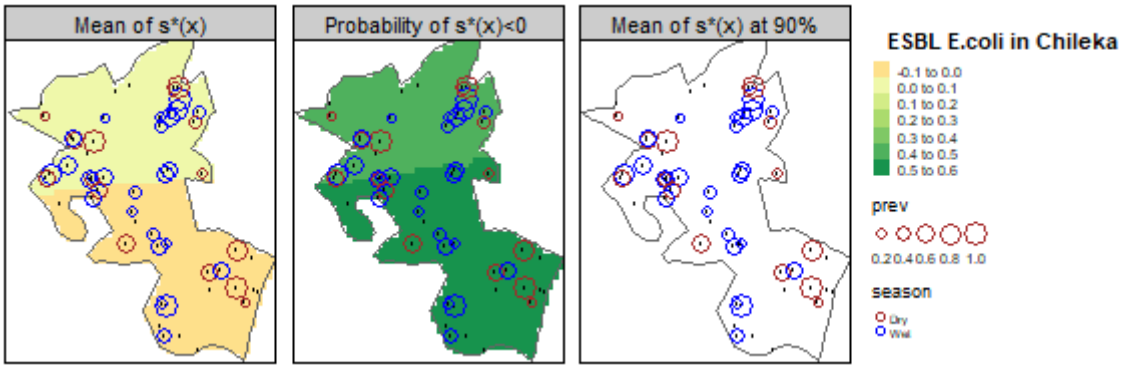


Figure B.11: Gaussian process maps for Chileka (ESBL-Ec)

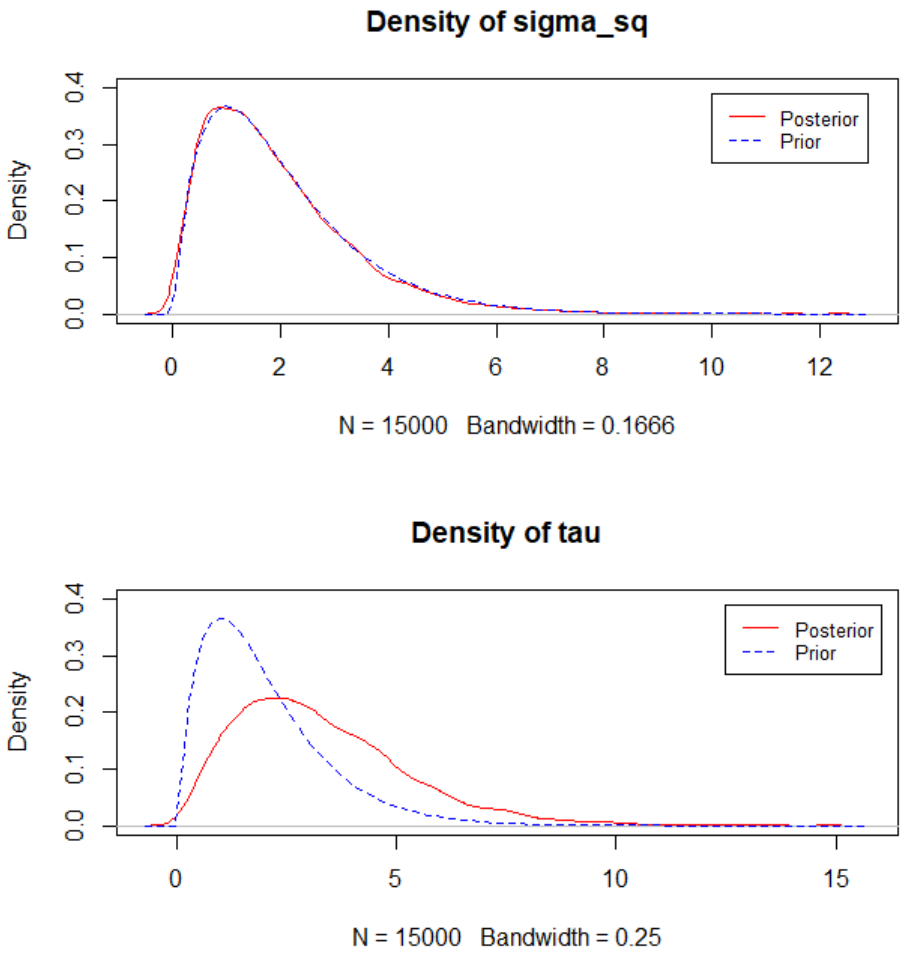


Figure B.12: Density plots of σ^2 and τ for Chileka (ESBL-Ec)

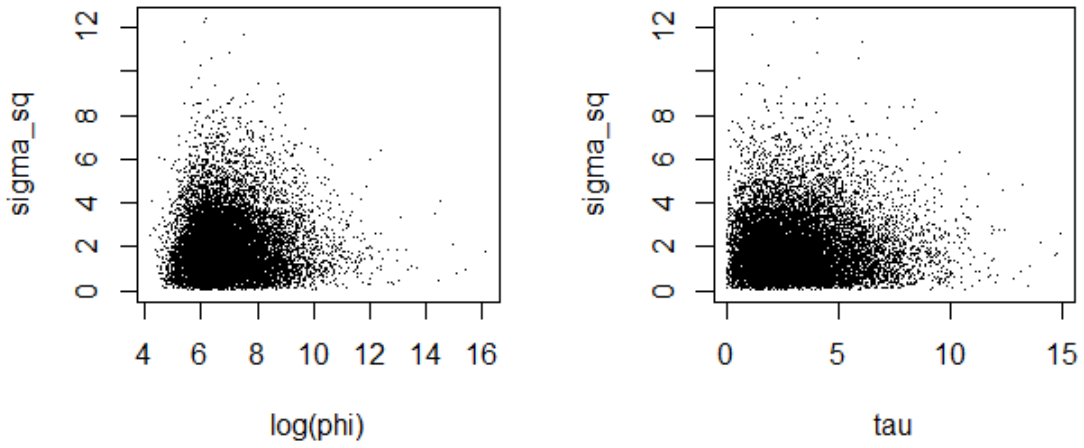


Figure B.13: Pairwise correlation plots for Chileka (ESBL-Ec)

B.3.6 ESBL-producing *K. pneumoniae* in Chileka

Table B.6: Parameter estimates for the spatial model in Chileka (ESBL-K)

	Median	Mean	Std error	2.5%	97.5%
ϕ	912.27	5713.28	126608.72	183.72	18705.21
σ^2	1.62	1.93	1.37	0.24	5.40
τ	3.05	3.41	2.06	0.55	8.33

	Log-odds	Std error	2.5%	97.5%
Intercept	-2.095	1.448	-5.070	0.868
Reactive to HIV testing (vs non-reactive)	0.427	0.320	-0.188	1.080
Unknown HIV status (vs non-reactive)	0.731	0.410	0.019	1.618
Number of days since the first sample	1.058	0.469	0.206	2.065

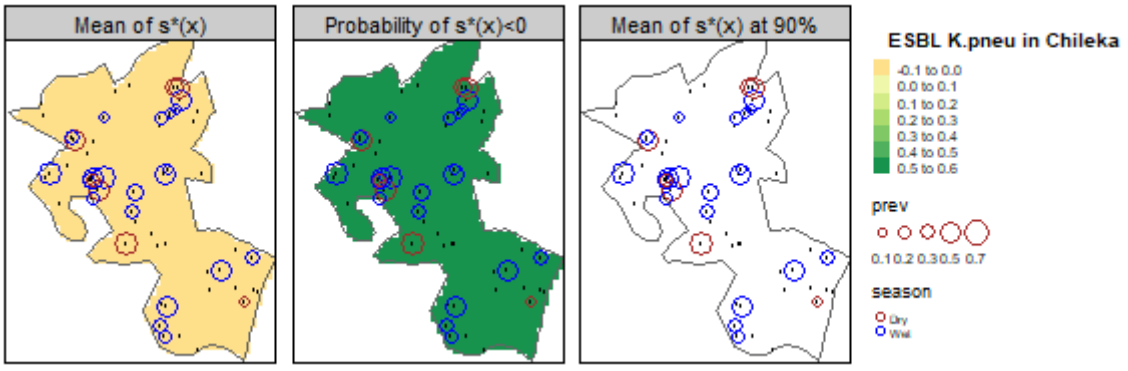


Figure B.14: Gaussian process maps for Chileka (ESBL-K)

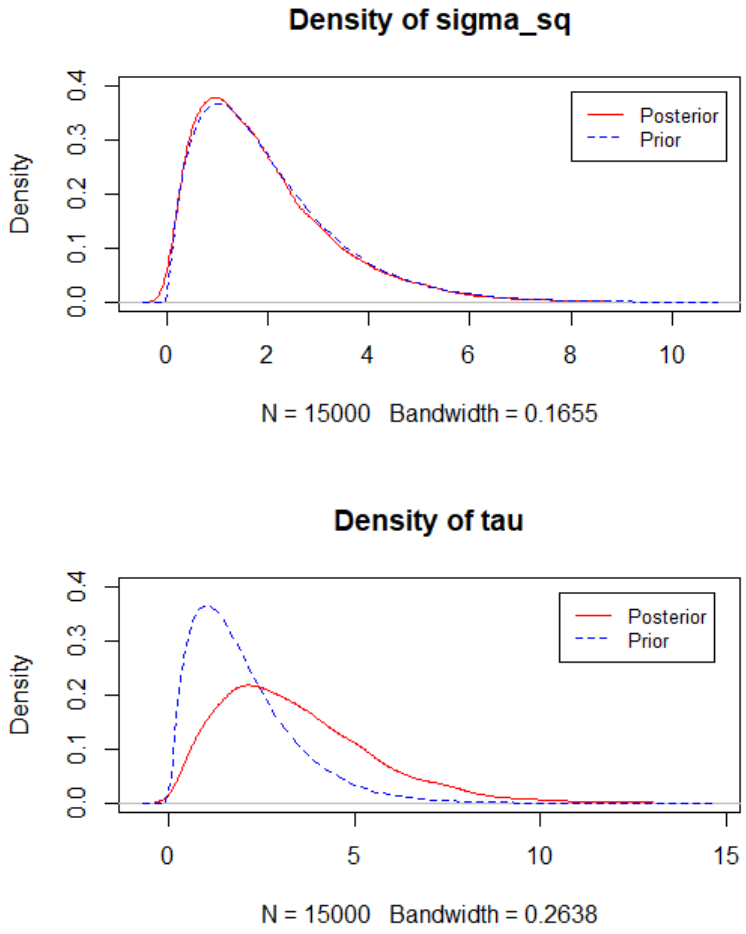


Figure B.15: Density plots of σ^2 and τ for Chileka (ESBL-K)

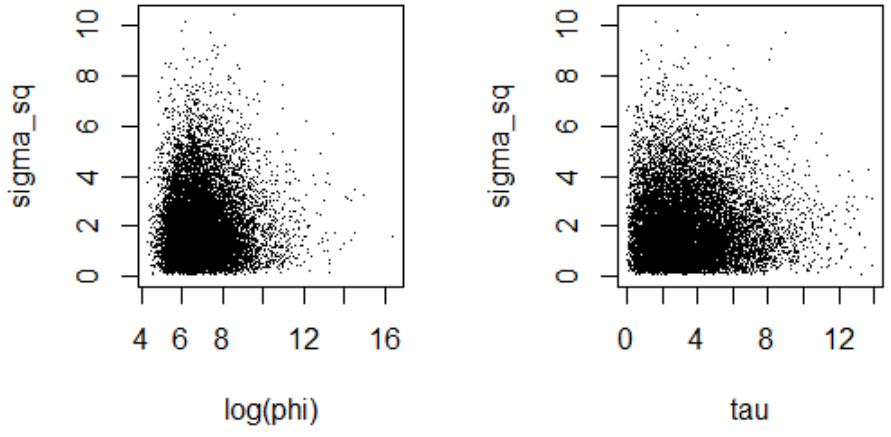


Figure B.16: Pairwise correlation plots for Chileka (ESBL-K)

Appendix C

Appendix: Individual and WASH risk factors of ESBL-producing *E. coli* and *K. pneumoniae* colonisation over time in Malawi

C.1 Household reported WASH variables

Table C.1: Household reported WASH variables

Variable name	Description	Value	Original variable	Modifications
toilet	Toilet presence	Binary (y/n)		
opendefecation	Open human defecation	Binary (y/n)		"never/a few times a year" to no and rest to yes
toiletshare	Sharing hh toilet with non-hh members	Binary (y/n)		
disposal	Disposal mechanism for animal waste	Binary (y/n)	hh_manureanimal hh_manureanimalo	Modified (local expertise)
streetfood	Eat street food	Binary (y/n)		"never" to no and the rest to yes
sharedplates	Eat from shared plates	Binary (y/n)		
pipewater	Drinking water source (pipe)	Binary (y/n)	hh_drinkwater2-3	If either yes, then yes
tapwater	Drinking water source (tap)	Binary (y/n)	hh_drinkwater4	
tubewellwater	Drinking water source (well)	Binary (y/n)	hh_drinkwater5-6	If either yes, then yes
utensilwater	Alternative water used for cleaning utensils	Binary (y/n)		
birds	Birds ownership	Binary (y/n)	hh_an5,8,10,11,12	If any yes, then yes
cattlegoatsheep	Cattle/Goats/Sheep ownership	Binary (y/n)	hh_an1,2,3	If any yes, then yes
dogcat	Dogs/Cats ownership	Binary (y/n)	hh_an7,13	If either yes, then yes
pigs	Pigs ownership	Binary (y/n)	hh_an4	
animalinside	Animals kept inside the house	Binary (y/n)	hh_ancattle/angoats/anpigs/anchickens/andogs	"In the house" to yes and rest to no
riverwater	Hh member interaction with river water	Binary (y/n)	hh_riverwaterchild hh_riverwateradult	If either yes, then yes
drains	Hh member interaction with drains	Binary (y/n)	hh_drainschild hh_drainsadult	If either yes, then yes

C.2 Household observed WASH variables

Table C.2: Household observed WASH variables

Variable name	Description	Value	Original variable	Modifications
toilettype	Type of toilet (Construction)	Categorical	hh_toilet	First observation in time only
toiletfloor	Type of toilet floor	Categorical		First observation in time only
drophole	Presence of drop hole cover	Binary (y/n)		First observation in time only
cleantoiletpaper	Presence of toilet paper	Binary (y/n)	hh_toiletcleanm1	Rule1: If any yes at any time point, then yes
cleanpaper	Presence of newspaper	Binary (y/n)	hh_toiletcleanm2	Rule1
humanfaeces	Visible human defecation	Binary (y/n)	hh_toiletfaeces and hh_hfa	Rule1
handwashfacil	Facilities for hand washing (hwf) at household	Binary (y/n)	hh_toif1-4	Rule1
soapiness	Presence of soap at any hwf	Numeric	hh_ahwfm1-3 - hh_ehwfm1-3	Mean of hwf with soap in the house at each time point over number of obs
coveredst	Water storage method (covered)	Binary (y/n)	hh_drinkwaterst1,3,6,8,11,13	Rule1
uncoveredst	Wsm (uncovered)	Binary (y/n)	hh_drinkwaterst2,5,7,12,14	Rule1
lidandtapst	Wsm (lid and tap)	Binary (y/n)	hh_drinkwaterst4,10	Rule1
animalcontact	Animal interaction with food areas	Binary (y/n)		Rule1
animalfaeces	Animal faeces seen around the area	Binary (y/n)	hh_anfam	Rule1
standingwater	Accumulation of wastewater	Binary (y/n)	Rule1	

C.3 Complete list of variables

Table C.3: Complete list of variables

Variable name	**Description**
HIV status	Categorical, {Reactive/Unknown/Non-reactive} (Reference: Non-reactive)
Recent use of antibiotics	Binary, {1 = at least a course of antibiotic taken in the last 6 months or in-between visits, 0 = none}
Age	Continuous, age at enrolment
Sex	Categorical, {Male/Female} (Reference: Female)
Number of people living in the household	Continuous, number of people living in the household at baseline
Average household monthly income	Continuous, average monthly household income (MWK) at baseline
Having children of school age	Binary, {1 = children of school age living in the household, 0 = no children of school age living in the household}
Presence of a toilet in the household	Binary, {1 = toilet present, 0 = toilet absent}
Open defecation	Binary, {1 = open defecation reported, 0 = no open defecation reported}
Sharing the toilet with non-household members	Binary, {1 = share toilet, 0 = do not share toilet}
Presence of a disposal mechanism for animal waste	Binary, {1 = disposal mechanism available, 0 = no disposal mechanism available}
Eating street food	Binary, {1 = eat street food at times, 0 = never eat street food}
Eating from shared plates	Binary, {1 = shared plates used, 0 = separate plates used}
Having a pipe as drinking water source	Binary, {1 = yes, 0 = no}

Having a communal tap as drinking water source	Binary, {1 = yes, 0 = no}
Having a tube well/borehole as drinking water source	Binary, {1 = yes, 0 = no}
Use of alternative water for cleaning utensils	Binary, {1 = use of different water than the one used for drinking, 0 = use of same water than the one used for drinking}
Owning birds	Binary, {1 = yes (owns one or more), 0 = no}
Owning cattle, goats or sheep	Binary, {1 = yes (owns one or more), 0 = no}
Owning dogs or cats	Binary, {1 = yes (owns one or more), 0 = no}
Owning pigs	Binary, {1 = yes (owns one or more), 0 = no}
Keeping animals inside	Binary, {1 = yes, 0 = no}
Contact with river water	Binary, {1 = any adult or child at the household interact with river water, 0 = no adult or child at the household interact with river water}
Contact with drains	Binary, {1 = any adult or child at the household interact with drains, 0 = no adult or child at the household interact with drains}
Toilet type	Categorical, {Pit latrine/Shared toilet/No toilet/Other} (Reference: No toilet)
Toilet floor material	Categorical, {Concrete or wood/Soil/No toilet} (Reference: Concrete or wood)
Having a drop hole cover on the toilet	Binary, {1 = drop hole cover present, 0 = drop hole cover absent}
Presence of toilet paper in the toilet	Binary, {1 = toilet paper present, 0 = toilet paper absent}
Presence of newspaper/paper in the toilet	Binary, {1 = newspaper or paper present, 0 = newspaper or paper absent}
Visible human faeces around the household	Binary, {1 = visible human stool, 0 = no visible human stool}

Presence of handwashing facilities (hwf) in the household	Binary, {1 = present anywhere within the household, 0 = absence within the household}
Frequency of soap presence in handwashing facilities	Continuous, number of hwf with soap in the house over total number of hwf present in the house
Storing water covered	Binary, {1 = water stored at the house covered, 0 = no water stored at the house covered}
Storing water uncovered	Binary, {1 = water stored at the house uncovered, 0 = no water stored at the house uncovered}
Storing water in a container with lid/tap	Binary, {1 = water stored at the house in a container with lid and/or tap, 0 = no water stored at the house in a container with lid and/or tap}
Contact between animals and food areas	Binary, {1 = animal seen in contact with food areas, 0 = no animal seen in contact with food areas}
Visible animal faeces around the household	Binary, {1 = animal faeces seen around the household, 0 = no animal faeces seen around the household}
Presence of standing water around the household	Binary, {1 = standing water seen, 0 = standing water not seen}
Number of days since the first sample	Continuous, number of days since the first sample was taken
Harmonic terms (sinday, cosday, sinday2, cosday2)	Continuous, described in the modelling framework
Study area	Categorical, {Chikwawa/Ndirande/Chileka} (Reference: Chileka)

C.4 R/STAN code for implementation of the temporal model

```
data {
  int Nhh;                              //Number of households
  int Nindiv;                           //Number of individuals
  int Ntimes;                           //Number of visits
  int N;                                //Total number of observations
  int hh[N];                            // Long format integer id
  int<lower=1, upper=4> followup[N];    // Long format integer id
  int indiv[N];                         // Long format integer id
  int y[N];                             // Long format observations

  matrix[Nhh*Ntimes,Ntimes] distc;      //Distance matrix (Time between visits)
  int K;                                //Number of covariates
  matrix [N, K] X;                      //Matrix of covariates
}

parameters{
  vector[Ntimes] u[Nhh];
  real alpha;
  real sigma;
  vector[K] beta;
  real<lower=0>  phi;
  real<lower=0> tau;
}

transformed parameters{
  vector[N] mu;
  matrix[Ntimes,Ntimes] sub[Nhh];
  vector[Ntimes] s[Nhh];
  matrix[Ntimes,Ntimes] dist[Nhh];
  {
    matrix[Ntimes,Ntimes] L;
    for (i in 1:N) {
      mu[i] = alpha + X[i]*beta ;
    }

    for(j in 1:Nhh){
      dist[j,,]=distc[(3*(j-1)+j):((3*(j-1)+j)+3),];
    }

    for(j in 1:Nhh){
      for(t in 1:Ntimes){
        for (k in 1:Ntimes){
          if (t==k) sub[j,t,k]=1+tau^2;
            else sub[j,t,k]=(sigma^2)*exp(-dist[j,t,k]^2/phi^2);
        }
      }
    }

    for(j in 1:Nhh){
      L = cholesky_decompose(sub[j,,]);
      s[j] = L*u[j] ;
    }
  }
}
```

```
model{
  sigma ~ gamma(2, 1);
  phi ~ gamma(4,0.125);
  alpha ~ normal(0,100);
  tau ~ gamma(2,1);

  for (household in 1:Nhh) {
    u[household] ~ normal(0, 1);
  }

  for (cov in 1:K){
    beta[cov] ~ normal(0,10);
  }

  for (i in 1:N) {
    y[i] ~ bernoulli_logit(mu[i] + s[hh[i], followup[i]]);
  }
}
```

C.5 Univariable results for ESBL-producing *E. coli*

Table C.5: Full univariable analysis results between ESBL-producing *E. coli* colonisation status and each variable accounting for the study area

	Log-odds	**P-value**	**Odds ratio (95% CI)**
Reactive to HIV testing (vs non-reactive)	-0.063	0.700	0.939 (0.682-1.293)
Unknown HIV status (vs non-reactive)	-0.168	0.074	0.846 (0.704-1.017)
Recent use of antibiotics	0.062	0.137	1.063 (0.981-1.153)
Age*	**0.092**	**0.026**	**1.097 (1.011-1.189)**
Being male (vs female)	**-0.175**	**0.039**	**0.840 (0.711-0.992)**
Number of people living in the household	0.003	0.935	1.003 (0.924-1.089)
Average household monthly income	**-0.107**	**0.019**	**0.899 (0.822-0.983)**
Presence of a toilet in the household	-0.021	0.626	0.979 (0.900-1.066)
Open defecation	**0.089**	**0.038**	**1.093 (1.005-1.189)**
Sharing the toilet with non-household members	0.017	0.699	1.017 (0.935-1.105)
Presence of a disposal mechanism for animal waste	**-0.090**	**0.046**	**0.914 (0.837-0.998)**
Eating street food	0.054	0.214	1.055 (0.970-1.148)
Eating from shared plates	-0.079	0.072	0.924 (0.848-1.007)
Having a pipe as drinking water source	**-0.099**	**0.022**	**0.906 (0.832-0.985)**
Having a tap as drinking water source	-0.035	0.463	0.965 (0.879-1.061)
Having a well as drinking water source	**0.191**	**3.8e-04**	**1.210 (1.089-1.345)**
Use of alternative water for cleaning utensils	0.057	0.174	1.059 (0.975-1.149)
Owning birds	4.1e-05	0.999	1.000 (0.915-1.094)
Owning cattle, goats or sheep	**0.093**	**0.044**	**1.097 (1.003-1.201)**
Owning dogs or cats	-0.020	0.628	0.980 (0.903-1.064)
Owning pigs	-0.038	0.397	0.962 (0.881-1.052)
Keeping animals inside	**0.122**	**0.005**	**1.129 (1.038-1.228)**
Contact with river water	**0.092**	**0.044**	**1.097 (1.003-1.200)**
Contact with drains	-0.034	0.430	0.967 (0.890-1.051)
Toilet type: other (vs no toilet)	-0.145	0.440	0.865 (0.600-1.249)
Toilet type: pit latrine (vs no toilet)	-0.026	0.826	0.974 (0.773-1.228)
Toilet type: shared toilet (vs no toilet)	-0.062	0.760	0.940 (0.630-1.402)

Toilet floor material: no toilet (vs concrete/wood)	**0.275**	**0.050**	**1.317 (1.000-1.736)**
Toilet floor material: other (vs concrete/wood)	**0.330**	**0.002**	**1.391 (1.130-1.712)**
Having a drop hole cover on the toilet	**-0.164**	**1.7e-04**	**0.849 (0.779-0.925)**
Presence of toilet paper in the toilet	-0.053	0.234	0.948 (0.868-1.035)
Presence of newspaper/paper in the toilet	**-0.141**	**0.002**	**0.868 (0.796-0.947)**
Visible human faeces around the household	-0.009	0.830	0.991 (0.911-1.078)
Presence of handwashing facilities in the household	0.037	0.405	1.038 (0.951-1.132)
Frequency of soap presence in handwashing facilities	-0.082	0.079	0.921 (0.841-1.009)
Storing water covered	-0.067	0.098	0.935 (0.863-1.013)
Storing water uncovered	0.034	0.481	1.035 (0.941-1.137)
Storing water in a container with lid/tap	**-0.112**	**0.013**	**0.894 (0.819-0.976)**
Contact between animals and food areas	**0.187**	**1.3e-05**	**1.205 (1.108-1.311)**
Visible animal faeces around the household	-0.006	0.902	0.994 (0.904-1.093)
Presence of standing water around the household	-0.069	0.121	0.933 (0.855-1.018)
Having children of school age	-0.015	0.716	0.985 (0.908-1.069)
Number of days since the first sample	**0.118**	**0.005**	**1.125 (1.037-1.221)**
Harmonic term (sinday)	**-0.181**	**0.004**	**0.834 (0.738-0.943)**
Harmonic term (cosday)	**0.305**	**2e-04**	**1.357 (1.156-1.593)**
Harmonic term (sinday2)	-0.032	0.626	0.969 (0.852-1.102)
Harmonic term (cosday2)	-0.029	0.679	0.972 (0.848-1.113)

*Significant variables highlighted in bold

C.6 Univariable results for ESBL-producing *K. pneumoniae*

Table C.7: Full univariable results for ESBL-producing *K. pneumoniae* colonisation status

	Log-odds	**P-value**	**Odds ratio (95% CI)**
Reactive to HIV testing (vs non-reactive)	0.104	0.661	1.109 (0.697-1.765)
Unknown HIV status (vs non-reactive)	0.041	0.750	1.042 (0.809-1.342)
Recent use of antibiotics	0.107	0.058	1.112 (0.996-1.242)
Age	-0.019	0.758	0.981 (0.868-1.108)
Being male (vs female)	-0.159	0.207	0.853 (0.666-1.092)
Number of people living in the household*	**0.229**	**6.1e-05**	**1.257 (1.124-1.406)**
Average household monthly income	-0.014	0.830	0.987 (0.872-1.116)
Presence of a toilet in the household	0.085	0.193	1.088 (0.958-1.236)
Open defecation	0.019	0.755	1.019 (0.904-1.150)
Sharing the toilet with non-household members	0.055	0.371	1.056 (0.937-1.191)
Presence of a disposal mechanism for animal waste	0.057	0.330	1.058 (0.944-1.186)
Eating street food	**-0.125**	**0.029**	**0.882 (0.788-0.987)**
Eating from shared plates	**-0.152**	**0.016**	**0.859 (0.759-0.972)**
Having a pipe as drinking water source	0.106	0.077	1.112 (0.989-1.250)
Having a tap as drinking water source	-0.120	0.072	0.887 (0.778-1.011)
Having a tube/well as drinking water source	-0.009	0.891	0.992 (0.878-1.119)
Use of alternative water for cleaning utensils	-0.087	0.187	0.917 (0.806-1.043)
Owning birds	**0.138**	**0.026**	**1.148 (1.016-1.296)**
Owning cattle, goats or sheep	0.054	0.370	1.056 (0.938-1.188)
Owning dogs or cats	0.086	0.153	1.089 (0.969-1.225)
Owning pigs	0.085	0.131	1.089 (0.975-1.216)
Keeping animals inside	0.041	0.506	1.042 (0.924-1.174)
Contact with river water	-0.001	0.986	0.999 (0.885-1.128)
Contact with drains	**0.137**	**0.011**	**1.147 (1.032-1.275)**
Toilet type: other (vs no toilet)	0.276	0.299	1.317 (0.783-2.216)
Toilet type: pit latrine (vs no toilet)	0.287	0.107	1.333 (0.940-1.889)

Toilet type: shared toilet (vs no toilet)	-0.396	0.261	0.673 (0.338-1.343)
Toilet floor material: no toilet (vs concrete/wood)	-0.148	0.454	0.863 (0.586-1.270)
Toilet floor material: other (vs concrete/wood)	0.154	0.270	1.167 (0.887-1.534)
Having a drop hole cover on the toilet	0.042	0.488	1.043 (0.926-1.176)
Presence of toilet paper in the toilet	0.025	0.676	1.026 (0.910-1.156)
Presence of newspaper/paper in the toilet	-0.047	0.459	0.954 (0.842-1.081)
Visible human faeces around the household	0.109	0.074	1.115 (0.990-1.256)
Presence of handwashing facilities in the household	0.060	0.341	1.062 (0.939-1.201)
Frequency of soap presence in handwashing facilities	-0.064	0.344	0.938 (0.822-1.071)
Storing water covered	0.080	0.280	1.083 (0.937-1.252)
Storing water uncovered	-0.102	0.093	0.903 (0.801-1.017)
Storing water in a container with lid/tap	-0.007	0.910	0.993 (0.879-1.121)
Contact between animals and food areas	-0.050	0.429	0.952 (0.842-1.076)
Visible animal faeces around the household	-0.025	0.678	0.975 (0.865-1.099)
Presence of standing water around the household	-0.047	0.463	0.954 (0.843-1.081)
Having children of school age	0.009	0.882	1.009 (0.893-1.140)
Number of days since the first sample	-0.097	0.122	0.907 (0.802-1.026)
Harmonic term (sinday)	**-0.345**	**3.7e-04**	**0.708 (0.586-0.856)**
Harmonic term (cosday)	**0.243**	**0.038**	**1.275 (1.014-1.605)**
Harmonic term (sinday2)	-0.053	0.592	0.948 (0.781-1.152)
Harmonic term (cosday2)	0.166	0.095	1.181 (0.972-1.434)
Living in Chikwawa (vs Chileka)	0.196	0.183	1.216 (0.912-1.622)
Living in Ndirande (vs Chileka)	0.098	0.536	1.103 (0.809-1.504)

*Significant variables highlighted in bold