Statistical modelling for zoonotic diseases

A thesis presented in partial fulfilment of the requirements for the degree of
Doctor of Philosophy
in
Statistics
at Massey University, Palmerston North,
New Zealand.

Sih-Jing Liao

2020

Contents

List of Tables

List of Figures

List of abbreviations

AIC Akaike information criterion

BA Blood agar

DAG Direct acrylic graph

DIC Deviance information criterion

EIDs Emerging infectious diseases

HPD Highest posterior density

IQR Interquartile range

MLST Multilocus sequence typing

MCMC Markov chain Monte Carlo

NZFSA New Zealand Food Safety Authority

PCR Polymerase chain reaction

pdf Probability density function

pmf Probability mass function

ST Sequence type

wgMLST whole genome multilocus sequence typing

Abstract

Preventing and controlling zoonoses through the design and implementation of public health policies requires a thorough understanding of epidemiology and transmission pathways. A pathogen may have complex transmission pathways that could be affected by environmental factors, different reservoirs and the food chain. One way to get more insight into a zoonosis is to trace back the putative sources of infection. Approaches to attribute the infection to sources include epidemiological observations and microbial subtyping techniques. In order for source attribution from the pathways to human infection to be delineated, this thesis proposes statistical modelling methods with an integration of demographic variables with multilocus sequence typing data derived from human cases and sources. These models are framed in a Bayesian context, allowing for a flexible use of limited knowledge about the illness to make inferences about the potential sources contributing to human infection.

These methods are applied to campylobacteriosis data collected from a surveillance sentinel site in the Manawatu region of New Zealand. A link between genotypes found from sources and human samples is considered in the modelling scheme, assuming genotypes from sources are equal or linked indirectly to that from human cases. Model diagnostics show that the assumption of equal prevalence of genotypes between humans and sources is not tenable, with a few types being potentially more prevalent in humans than in sources, or vice versa. Thus, a model that allows genotypes on humans to differ from those on sources is implemented. In addition, an approximate Bayesian model is also proposed, which essentially cuts the link between human and source genotype distributions when conducting inference.

The final inference from these approaches is the probability for human cases attributable to each source, conditional on the extent to which each case resides in a rural compared to urban environment. Results from the effective models suggest that poultry and ruminants are important sources for human campylobacteriosis. The more rural human cases are located, the higher the likelihood of ruminant-sourced cases is. In contrast, cases are more poultry-associated when their locations are more urban. A little rurality effect is noticed for water and other sources due to small sample sizes compared to that from poultry and ruminants. In addition, animal faeces are believed

to be the primary cause of water contamination via rainfall or runoff coming from farmland and pasture. When water is treated as a medium in the transmission, instead of an end point, water birds are suggested to be the most likely contributor to water contamination.

These findings have implications for public health practice and food safety risk management. A risk management strategy had been carried out in the poultry industry in New Zealand, leading to a marked decrease of urban case rates from a poultry source. However, the findings of this thesis suggest a further step with a focus on rural areas as rural case rates are observed to be relatively higher than urban rates. Further, by exploring the role that water plays in the transmission, it deepens our knowledge of the epidemiology about waterborne campylobacteriosis and highlights the importance of water quality. This opens a potential research direction to study the association of water quality and environmental factors such as higher global temperatures for this disease.

Acknowledgements

The completion of this thesis would not be possible without the supports from people I have met during my PhD study. My profound gratitude first goes to my supervisors: Professor Martin Hazelton, Dr. Jonathan Marshall, and Distinguished Professor Nigel French. Martin helped me establish a clear thesis structure with his great attention to detail. Jonathan advised me on how to improve the efficiency of my codes with his powerful programming skills. Nigel gave me valuable advice on my thesis with his strong expertise in infectious diseases. During the PhD study, Martin and Jonathan explained complex ideas in an intuitive way so that I could pick up easily the concepts of statistical modelling. These ideas would not be carried out properly without Nigel's extensive knowledge about epidemiology of the disease we were studying. I am very grateful for all their help. Their tremendous enthusiasm for statistical inference and infectious diseases made my PhD research enjoyable.

I gratefully acknowledge the contributions of Mid Central Public Health Services, the Molecular Epidemiology and Public Health Laboratory (mEpiLab, Massey University) for data collection, the Ministry for Primary Industries for funding the Manawatu campylobacteriosis sentinel surveillance site, and financial supports from the Infectious Disease Research Centre (IDReC, Massey University), the School of Fundamental Sciences (Massey University) and the New Zealand Food Safety Science Research Centre (NZFSSRC) to this work.

I would like to extend my gratitude to my friends, Olivia Angelin-Bonnet, Ellie Johnson, Nayer Ngametua and Sarah Pirikahu, who always encourage me to be positive in different matters. Special thanks goes to my fellow, Ahmad Mahmoodjanlou, who is generous all the time offering biscuits to nourish my brain while writing the thesis. I appreciate all postgraduates in Maths and Statistics, who were being supportive too and giving good suggestions about my talk practices before a conference presentation. I am also indebted to Anne Lawrence and Debbie Leader, whose passion for tutoring let me realise teaching students statistics can be so fun. This also inspires me to write the thesis from another angle. I also want to thank my art teachers, Dennis Greenwood and Mr. Hu, who showed me how interesting the art world is. Drawing and painting boost my mental power to deal with some days that were harder than others in the

PhD life.

Finally, I would not reach to the end of study without my family. Thanks to my family, and my late father and late granny, who developed my never-give-up attitude to overcome difficulties in my life, and also their endless love supporting me to build my career path in biostatistics.

Publication arising from thesis

Liao, SJ, Marshall J, Hazelton ML, French NP. (2019). Extending statistical models for source attribution of zoonotic diseases : a study of campylobacteriosis. Journal of the Royal Society Interface, 150(16), 20180534. https://doi.org/10.1098/rsif.2018.0534

Chapter 1

Introduction

1.1 Introduction to zoonoses

Infectious diseases are illnesses caused by germs such as viruses and bacteria. From 1940 to 2004, the emergence of 335 infectious pathogens have been reported in the world population. Therefore, 'emerging infectious diseases' (EIDs) is defined for those infections that have a significant impact on global socio-economics and public health (Jones et al., 2008; van Doorn, 2014), for example, avian influenza, advanced HIV infection (AIDs) and severe acute respiratory syndrome (SARS).

Infectious diseases can cause high economic and medical costs due to morbidity and mortality. The annual number of global deaths from infections in recent decades has been shown to be at approximately 15 million and may remain at the same level for the next three decades (WHO, 2013; Dye, 2014). The major cause of infections is associated with more than 1,400 species of human pathogens, 13% of which are classified as 'emerging', with more than half of these species known to be zoonoses-associated (Woolhouse and Gowtage-Sequeria, 2005; Greger, 2007; Jones et al., 2008; Dye, 2014; van Doorn, 2014). Zoonoses are infectious diseases spread between humans and animals that pose a threat to global public health as they account for more than 60% of EIDs (Taylor et al., 2001; Woolhouse and Gowtage-Sequeria, 2005; Greger, 2007; Jones et al., 2008; Reperant and M E Osterhaus, 2013).

In order for such an enormous health burden to be reduced, preventing and controlling infectious diseases becomes extraordinarily important, and our ability to intervene depends on how much we know about the nature of disease transmission. For zoonotic diseases, transmission to humans from animal reservoirs may be complex, involving many sources and exposures linked by different pathways through environmental contamination or direct contact with animals. Infected wild birds and broiler chickens, for example, may contaminate the environment and the food chain, respectively, resulting in hazard in the consumption of drinking water and undercooked chicken through the

water and food supply (Wagenaar et al., 2013). Therefore, knowledge of the potential sources and pathways of infection is key to reducing the burden of disease. Tracing the source of infection also becomes crucial to increasing the ability to implement risk management and intervention (Wilson et al., 2008; Morelli et al., 2012).

A better understanding of the epidemiology of zoonoses can provide useful information to prevent and control the disease spread. For example, a risk management strategy on *Campylobacter* was introduced in New Zealand in 2006. In 2003, New Zealand had the highest incidence rate amongst other developed countries, with a peak at 395.6 per 100,000 population (ESR, 2004; Olson et al., 2008; Lane and Briggs, 2014). Therefore, the New Zealand Food Safety Authority (NZFSA) developed a risk management strategy involving the poultry industry; aimed at limiting the counts of *Campylobacter* on poultry carcases and it was expected to reduce the number of human cases arising from a poultry source. Figure 1.1 shows the annual incidence rates per 100,000 population in New Zealand from 1990 to 2016; a dramatic decline is remarkable during 2007 and 2008. This highlights the importance of identifying the sources of infection.

Figure 1.1: Case rates for campylobacteriosis per 100,000 population in New Zealand from 1990 to 2016. An intervention in the poultry industry conducted in late 2006 resulted in a decreasing incidence of disease in the following years.

1.2 Global epidemiology of campylobacteriosis

Human campylobacteriosis is among the most widespread bacterial gastroenteritis in the world. The disease is mainly caused by *Campylobacter jejuni*, which is in a genus of *Campylobacter* bacteria. This bacteria is a commensal microorganism that exists in the environment and that can be transmitted between animal hosts and humans. The signs of infection are similar to enteritis symptoms (diarrhea, abdominal pain and fever), however, severe complications may develop such as irritable bowel syndrome and Guillain-Barré syndrome (Hahn, 1998). The illnesses are believed to result from: i) food that are already contaminated by faecal matter such as poultry, milk and salad; ii) drinking environmental water that is contaminated by animal or wild bird faeces; or iii) direct-contact with contaminated animals.

In the twentieth century, the cause was not well recognised as techniques of selective isolation were not yet developed (King, 1957; Ryan et al., 2004). However, the increasing incidence over the past two decades has drawn great attention from public health and food safety authorities in developing and developed countries. In 2012, the World Health Organization estimated that the annual notification rates are between 440 and 930 cases per 100,000 population in high-income countries[1]; however, knowledge of the infections in middle- and low- income countries is limited due to insufficient surveillance data (WHO and FAO, 2013). The United States confirmed that the national incidence rate in 2012 was 14.3 cases per 100,000 people, which is a 14% increase on the average annual incidence between 2006 and 2008 (CDC, 2013). In newly industrialised countries, such as China and Africa, concerns about *Campylobacter* infection have been raised in recent years. At a hospital in Beijing, 142 out of 950 (14.9%) patients with acute diarrhea were confirmed to be cases between 2005 and 2009 (Chen et al., 2011), while a meta-analysis from multiple studies of the prevalence in African animals and meat pointed out that the pathogen was detected from more than 35% of 11,828 poultry samples and 25% of 1,975 pig samples across 27 African countries (Thomas et al., 2019).

Over the past decades, campylobacteriosis has been one of the most notifiable food-poisoning diseases in New Zealand. This fact is highlighted in Table 1.1 compiled partly from the annual surveillance report (ESR, 2017); it shows a high incidence of *Campylobacter* infection among all the reported enteric diseases. Further study revealed that the illness occurs in all age groups, while children aged between 0 and 4 and the elderly are more likely to be infected than others owing to immature immunity for children and weak immunity for the elderly (Marshall et al., 2016). The transmission of this

[1] Hing-income countries are countries who have more than US$ 12475 gross national income (GNI) per capita. The threshold of historical classification by income can be found on the web page of The World Bank: http://databank.worldbank.org/data/download/site-content/OGHIST.xls

Disease	2015		2016	
Campylobacteriosis	6218	(135.3)	7456	(158.9)
Cryptosporidiosis	696	(15.1)	1062	(22.6)
Giardiasis	1510	(32.9)	1617	(34.5)
Pertussis	1168	(25.4)	1096	(23.4)
Salmonellosis	1051	(22.9)	1091	(23.2)

Table 1.1: The number of reported cases and the associated rates per 100,000 population in brackets for the most frequently notifiable enteritis in New Zealand in 2015 and 2016.

pathogen is from livestock and wild animals to humans. The pathway can be, for instance, via handling an animal food product that is already contaminated by faecal matter, drinking contaminated water, or eating undercooked chicken. Approaches developed in this thesis will apply to data about campylobacteriosis to gain more insight into the epidemiology of this disease in New Zealand.

1.3 Methods of source attribution for zoonotic diseases

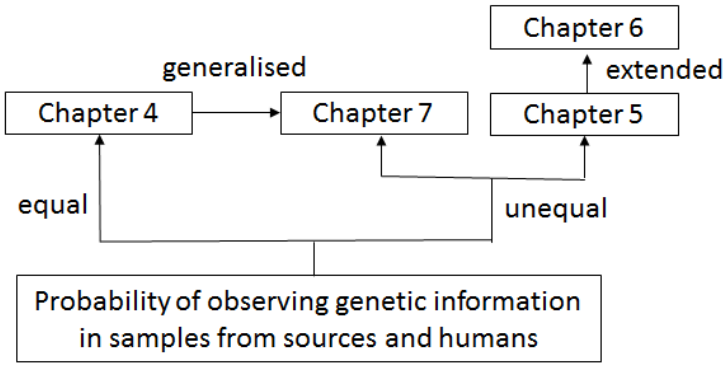
Modelling of disease surveillance data to explore patterns of zoonotic diseases has had a long history in public health. Several methods for attributing human infections to different sources have been addressed in the literature, such as epidemiological studies and microbial subtyping, which is a way to characterise samples (isolates) of a pathogen using phenotypic or genotypic subtyping methods (Andreoletti et al., 2008; Mughini-Gras et al., 2018; Cody et al., 2019).

In recent years, modelling zoonoses requires an advanced approach with the focus changed from just epidemiology or microbiology to a combination of epidemiology, evolutionary genetics and biology (Mullner et al., 2009b; Muellner et al., 2013). Some source attribution models have been proposed by utilising epidemiological observations and the association with genotypes isolated from human and source samples. The Dutch model (van Pelt et al., 1999) and the (modified) Hald model (Hald et al., 2004; Mullner et al., 2009a) are commonly used to estimate the proportion of human cases attributable to sources. These models are based on statistical models using microbial subtypes in human and source samples. Other models from a Bayesian perspective, such as the island genomic model (Wilson et al., 2008) and the HaldDP model (Miller et al., 2017), also determine the contribution of sources to infections. The island genomic model estimates the genetic relationship in samples given the pathogen is undergoing genetic evolution in different environments, while the HaldDP model is an extension of the original and modified Hald models, improving the model identifiability.

The aim of this thesis is to develop new statistical models and methods that can be used in order for the contribution of source populations to zoonotic diseases to be estimated. The type of data usually contains demographic observations and genetic information collected from samples, and hence genetic and non-genetic parameters may be involved in the modelling. In this thesis, statistical models that can deal with such integrated data will be developed for source attribution and their use will be demonstrated by applying them to New Zealand campylobacteriosis data (Marshall et al., 2016). These models can also be applied to zoonotic diseases other than *Campylobacter* infection with suitable tailoring. The fundamental modelling methods use only genetic information to make inferences about the potential sources responsible for the illness caused by the pathogen. The models developed here will also be extended to incorporate variables in order to better understand the epidemiology of the disease so as to inform decision making for future food control and risk management plans.

1.4 Thesis structure

Besides the introductory Chapter 1, this thesis comprises eight chapters. Chapters 2 and 3 present the epidemiological and genetic literature on *Campylobacter* infection, and the introduction to the statistical methods and models required for this PhD research. These chapters include basic knowledge from the process of how to detect genetic differences in samples, the applied probability distributions and Bayesian simulation methods, to the existing source attribution models. The proposed models and the associated applications are covered in Chapters 4 to 7.



The above graph shows the model development in this thesis. Chapter 4 starts with a Bayesian modelling for campylobacteriosis, with the assumption that the proportion split between genotypes found in sources is the same as the proportion split in human cases, indicating that the pathogen may spread directly from host sources to humans. However, *Campylobacter* is undergoing evolution when it is amplifying in the hosts

and also in the disease transmission. This means that the assumption about the equal proportional split between sources and human cases does not capture the differing ability of the pathogen, which can survive in different environments during amplification and transmission. In other words, the genotype distribution observed among human cases assumed to have arisen from a particular source may differ from the genotype distribution on that source, i.e. the genetic information at the source and human levels is mismatched. Therefore, Chapter 5 adopts an evolutionary model in the current model framework to infer the genetic changes between source samples before the final inference about source contribution to the infection. Chapter 6 extends the previous model, which considers four source groups: poultry, ruminants, water and others, to include water birds as an additional source, while treating water as a role of being a vector for the other animal sources. Chapter 7 introduces a novel way to model source attribution, generalising the model developed in Chapter 4 to account for different probabilities of observing genotypes between sources and humans.

Last, Chapters 8 and 9 contain a discussion about the developed models to wider applications and implications in a global context, and conclusions summarising the work of this research as well as possible future study directions. All codes for fitting data to developed models and producing figures are available on GitHub (Liao, 2020). Materials supporting this thesis including a glossary are provided in Appendices A to D.

Chapter 2

Literature review on *Campylobacter* epidemiology

2.1 Insights into human campylobacteriosis

Human campylobacteriosis is by far the most common bacterial gastroenteritis worldwide in the last decade, resulting in a disease burden in developed and developing countries (Ruiz-Palacios, 2007; Kaakoush et al., 2015). The estimated annual cost related to this disease in the Netherlands was €21 million, whereas it amounted to approximately US$1.3 billion in the United States (AH et al., 2005; Batz et al., 2012). This disease is also recognised as the most costly foodborne disease within the health system in New Zealand, amounting to $36 million paid for cases and health services (Gadiel, 2010).

The cause of illness is *Campylobacter* bacteria. It consists of 26 species, of which *Campylobacter jejuni* (*C. jejuni*-) and *C. coli* are responsible for more than 90% of human cases, while other species such as *C. concisus* and *C. upsaliensis* are also associated with gastrointestinal infections (Kaakoush et al., 2015). Symptoms such as muscle pain and diarrhoea occur after exposure to the bacteria on average from 2 to 5 days (Cottam et al., 2008; Gilpin et al., 2020). Despite the fact that the symptoms usually last no more than 7 days, the complications can be life-threatening such as sepsis and Guillain-Barré syndrome (Hahn, 1998; Kaakoush et al., 2015). Any age group is at risk to contract the pathogen. Vulnerable people such as the elderly and the young are highly likely to be infected (Kaakoush et al., 2015; Marshall et al., 2016). This may be because of immunity and human behaviour. People older than 60 years usually have weaker immunity than other age groups of people, while children younger than 5 years have immature immunity and their frequent hand-mouth contact also increases the risk of infection (French et al., 2008).

Campylobacter is ubiquitous in the environment. It colonises in a broad range

of animals such as chicken and sheep, and is transmitted through different routes to infect humans. Specifically, direct contact with farm animals, and consumption of contaminated water, undercooked chicken and unpasteurised milk are believed to be the important risk factors (Mughini Gras et al., 2012; Levesque et al., 2013; Wagenaar et al., 2013; Marshall et al., 2016). The pathogen is also temperature-driven. While the incidence rates peak in spring and summer in most temperate regions, there is little seasonal variation in tropical countries (Sari Kovats et al., 2005a; Strachan et al., 2013). Climatic variables such as rainfall and sunshine may also be contributors to the infection as they are associated with the timing of the seasonal peak (Louis et al., 2005). In addition, the spatial pattern of infection is heterogeneous, with more urban cases associated with poultry and more ruminant-related infections in rural areas (Nylen et al., 2002; Bolwell et al., 2015; Marshall et al., 2016).

Campylobacter is genetically diverse as the species have different genetic characteristics (Colles et al., 2003b; Sheppard et al., 2009; Kittl et al., 2013). A number of studies at the strain level have identified differing sequence types (STs) found in animals and human cases, of which ST-21, ST-45 and ST-48 are the internationally common genotypes of *C. jejuni*. ST-21 and ST-48 are largely found in ovine and bovine sources, whereas ST-45 is regularly discovered in multiple animal species as well as in water (Colles et al., 2003b; French et al., 2008; Sopwith et al., 2008; Carter et al., 2009b; Müllner et al., 2010; Kovanen et al., 2014; Dearlove et al., 2016). However, the distribution of genotypes of *C. jejuni* may vary from country to country and the prevalence of different genotypes is also different between various reservoirs, foods and water (Gillespie et al., 2002). Host-associated variation in the genome is suggested to be a key to determining the likely origin of most strains (Dearlove et al., 2016). STs that are frequently found in poultry samples are also common in human cases but are rarely observed in ruminant samples (Colles et al., 2003b; Müllner et al., 2010). In New Zealand, ST-474 and ST-2381 are unique and rarely found elsewhere in the world. The first is the predominant poultry-associated genotype accounting for approximately 30% of cases, while the latter has been only observed in environmental water and the faeces of wild birds (Carter et al., 2009b; Müllner et al., 2010). In comparison with other countries, New Zealand has a distinct molecular epidemiology of *C. jejuni* as a result of its geographical isolation (Müllner et al., 2010).

The poultry production system in New Zealand is integrated. There are only three poultry companies supplying 90% of chicken meat throughout the country, representing 95% of poultry meat consumption (Müllner et al., 2010). In order for the incidence of human campylobacteriosis to be reduced, NZFSA had implemented a risk management strategy in the poultry industry from 2006 through 2008, monitoring the levels of *Campylobacter* contamination in poultry carcasses at the end of primary processing

(Sears et al., 2011; Duncan, 2014). Through this intervention, hygiene practices such as production and primary processing have been also improved. The intervention is suggested to be a causal effect on the decline in disease incidence (Sears et al., 2011). Figure 1.1 provides the evidence that it resulted in a 58% reduction in the 2008 case rate (166.3 per 100,000), compared to the rate in 2006 (385.6 per 100,000). Additionally, the intervention was also of benefit to the economic cost. Costs for all 2005 foodborne illness in New Zealand were $85.3 million, of which 87% of the total cost was *Campylobacter*-related (Scott et al., 2000). However, the cost for campylobacteriosis in 2009 was notably 60% lower than in 2005; it only accounted for 27% of the total cost ($131 million) of foodborne disease (Gadiel, 2010). The success of the poultry associated intervention in New Zealand highlights the importance of the identification of risk factors and the corresponding strategy to have control and prevention measures.

2.2 Genotyping *Campylobacter*

Molecular typing techniques are common methods to help trace back potential sources of infectious diseases. In order for the sources of *Campylobacter* infection to be identified, genetic data are usually derived from molecular genotyping that enables us to characterise strain types from different samples. The samples in the campylobacteriosis dataset examined in this thesis were from food, environmental water, animals and human cases taken from a sentinel site in the Manawatu region of New Zealand (Bolwell et al., 2015). As *Campylobacter* species colonise the intestines, faecal samples are typically collected from humans and animals. The procedures of sample collection, growth of organisms and DNA extraction are standardised before obtaining the molecular data (Mullner et al., 2009b).

For food like retail chicken samples, samples are washed and massaged in Buffered Peptone Water (BPW) in stomacher bags. The resuspended pellet is then obtained after centrifuging 5 ml of the wash for 35 minutes at 6°C. Next, it is mixed with Bolton Broth and incubated at 42°C for 48 hours before culturing on modified Cefoperazone Charcoal Deoxycholate agar (mCCDA) plates. Because a huge number of bacteria may be found in samples, and *Campylobacter* organisms might be few among them, suitable nutrients are provided to grow and detect easily the species. For faecal samples, they are cultured directly on mCCDA plates with the enrichment Bolton Broth, then incubated 48 hours at 42 °C.

After the incubation, a single colony resembling *Campylobacter* species is isolated and cultured onto Blood Agar (BA) for another 48 hours at 42°C in order to obtain a new growth of *Campylobacter*. To extract DNA sequences, the culture from BA is put in a chemical solution. To get rid of compounds, such as RNA and protein in the cells and to preserve DNA from degradation, the solution is boiled for 10 minutes at 100°C

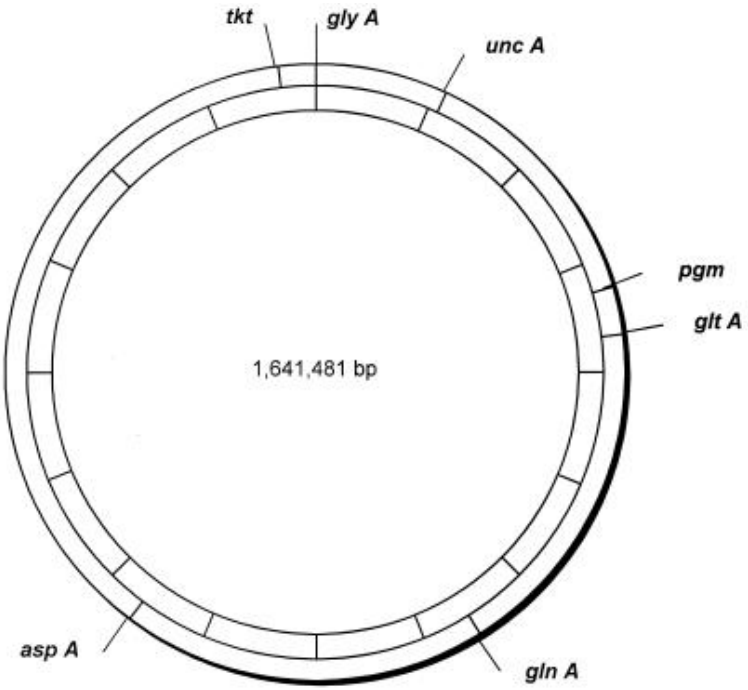


Figure 2.1: The whole genomic sequence of *C. jejuni* consists of 1,641,481 base pairs. MLST identifies seven housekeeping genes (gltA, uncA, aspA, glyA, glnA, pgm and tkt) in the whole genome (Dingle et al., 2001, Fig. 1)

.

(procedures for handling samples in greater detail can be found in the material from Mullner et al. (2009b) and Marshall et al. (2016)).

After DNA extraction, strains are differentiated in species, such as *C. jejuni* and *C. coli*, which are the dominant species of *Campylobacter* bacteria found in human cases of campylobacteriosis. The whole genomic sequence of *C. jejuni* consists of more than 1.6 million base pairs (Dingle et al., 2001; Pearson et al., 2007). At the time the sentinel site was established, a common and economic molecular method, multilocus sequence typing (MLST) (Dingle et al., 2001; Colles et al., 2003a; Urwin and Maiden, 2003), was used for genotyping *Campylobacter*. MLST utilises nucleotide sequences of internal fragments of seven housekeeping genes to index the genetic variation between isolates. These commonly used housekeeping genes are *asp*A (aspartase A), *gln*A (glutamine synthetase), *glt*A (citrate synthase), *gly*A (serine hydroxymethyltransferase), *pgm* (phosphoglucomutase), *tkt* (transketolase), and *unc*A (ATP synthase α subunit). They are an essential part of the core genome of the *C. jejuni* presenting genes responsible for key activities in cells (Urwin and Maiden, 2003; Fearnhead et al., 2014).

Genotyping requires a large amount of DNA sequence. However, only a small amount of DNA fragments are available during the procedure of DNA extraction. A microbial tool, polymerase chain reaction (PCR), is hence widely used to produce

Genotype	*asp*A	*gln*A	*glt*A	*gly*A	*pgm*	*tkt*	*unc*A
403	10	27	16	19	10	5	7
474	2	4	1	2	2	1	5
2026	10	1	16	19	10	5	7
2343	2	4	5	2	10	1	5

Table 2.1: The allelic profiles of a selection of genotypes composed of seven allele numbers at each of the seven housekeeping genes.

many copies of the fragment of DNA sequences by successively heating and cooling the solution. The seven housekeeping genes used in the MLST scheme are relatively conserved under evolution (Dingle et al., 2001; Biggs et al., 2011), but contain sufficient allelic variation for determining distinct genotypes. PCR targets seven housekeeping loci of roughly 400–500 base pairs to enable the sequences to be compared, assigning different numbers to each unique allele.

Figure 2.1 shows a whole genomic sequence of a *Campylobacter* isolate, in which the location of the seven housekeeping genes is depicted. Each unique nucleotide sequence (allele) at each housekeeping gene (locus) representing the variations differing from the other isolates is assigned a number, and the set of numbers across all loci (the allelic profile) is then taken as the genotype, which is assigned a sequence type number. An illustrative example of MLST data for *Campylobacter* is presented in Table 2.1. It shows that the genotypes ST-2026 and ST-474 have completely different allelic combinations across all seven loci, while ST-403 differs from ST-2026 only at the locus *gln*A, and ST-2343 differs from ST-474 at the loci *glt*A and *pgm*. The different allelic profiles enable comparison of genetic similarities or dissimilarities so that an association between sources and infected cases can be made by comparing the distribution of genotypes from human cases with those from potential reservoirs.

2.3 Campylobacteriosis dataset

The data from the campylobacteriosis study include spatial-temporal observations and microbial genotype information from both human and non-human cases. These samples were obtained at a sentinel surveillance site in the Manawatu region of New Zealand from February 2005 to December 2016 (Bolwell et al., 2015; Marshall et al., 2016). Since one isolate was taken from each sample, a total of 4,322 samples (isolates) are in the data.

Case	ST	Sample Date	Rurality	Age	Intervention
1	5	2006-05-16	3	NA	Before
2	5	2012-04-19	3	60	After
3	21	2008-10-21	1	57	After
	...	...	...	...	
1803	10058	2015-10-14	0	16	After
1804	10060	2014-12-12	NA	NA	After

Table 2.2: The data of human cases are comprised of an identified ST, the date of sample collection, rurality level, age and intervention.

2.3.1 Epidemiological data

A total of 1,804 human cases were notified in the surveillance system. The isolates from stool samples were typed from these cases. In the data, the following variables are available: genotype of *Campylobacter*, sample date, rurality category, age, and a binary outcome if the sample was taken before or after January 2008, which is considered to be the time point when the effect of an intervention implemented from 2006 to 2008 is observed (Sears et al., 2011).

A part of the data is displayed in Table 2.2, showing the genotype found from five human cases with the corresponding variables. The age range of cases is between 0 and 95. Modelling age as a continuous variable may be undesirable as several studies pointed out that there is an age-related pattern of human campylobacteriosis (French et al., 2008; Kaakoush et al., 2015; Marshall et al., 2016), indicating the effect of age on infection is nonlinear. This implies that the age variable on a continuous scale would make models heavily parameterised and hence result in more model complexity. Thus, a critical 'tipping point' in *Campylobacter* infection as addressed in the literature occurs when children start at school (Baker et al., 2007; French et al., 2008). To identify the pattern representative of particular age groups, the variable will be divided into 0–4 and 5+ categories and the modelling will be demonstrated in Chapter 4. The rurality scale is in the form of an ordinal classification of urban and rural areas with seven levels coded from -3 to 3, representing highly rural/remote area, rural area with low urban influence, rural area with moderate urban influence, rural area with high urban influence, independent urban area, satellite urban area and main urban area. Approximately 9% and 26.7% of individuals have no information about the location and the age, respectively, which will be assumed missing at random.

The number of typed human cases during 2005 and 2016 categorised by the rurality levels is tabulated in Table 2.3, along with the population from the 2006 and 2013 Census (Statistics New Zealand, a,b). Approximately 80% of cases lived in urban areas (the rurality scale is from 0 to 3), while the remaining cases resided in rural areas.

Rurality scale	Description	Human cases	2006	2013
-3	Highly rural/remote area	20	1,572	1,527
-2	Rural area with low urban influence	133	8,382	8,316
-1	Rural area with moderate urban influence	165	10,392	10,734
0	Rural area with high urban influence	101	6,579	7,155
1	Independent urban area	295	28,611	28,188
2	Satellite urban area	231	19,725	20,526
3	Main urban area	696	76,047	78,108

Table 2.3: The number of human cases in each rurality class during 2005 to 2016 along with the population size in 2006 and 2013 in the Manawatu region of New Zealand.

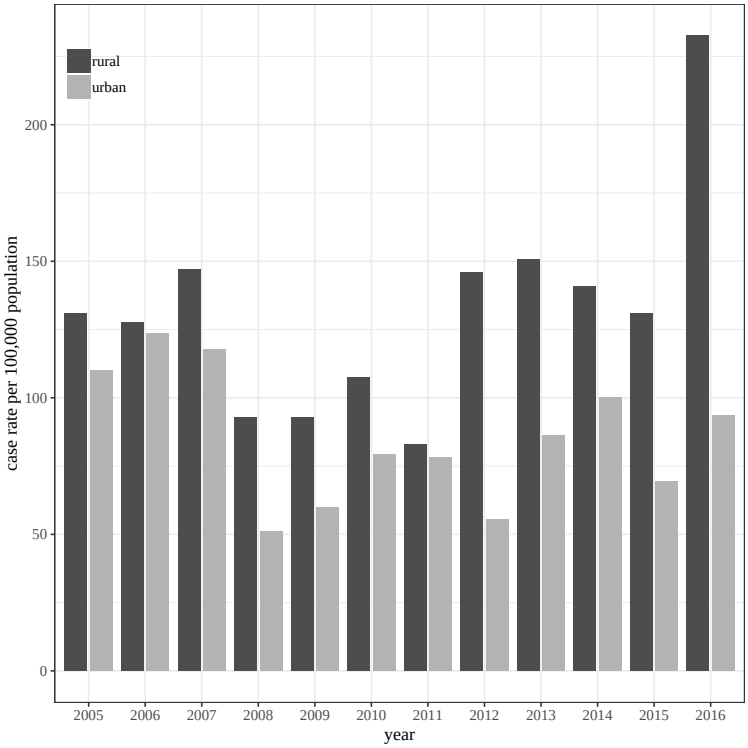


Figure 2.2: Case rates per 100,000 population in urban and rural areas of the Manawatu region of New Zealand from 2005 through to 2016. An intervention in the poultry industry conducted in 2007 and 2008 resulted in a decreasing incidence of campylobacteriosis in the following years, particularly in urban areas.

Genotype	Human	Poultry	Ruminants	Water	Other
42	59	7	53	10	2
45	149	155	10	21	54
474	247	60	15	5	9
2026	28	0	40	5	2
2381	0	0	0	60	3

Table 2.4: The frequency of five genotypes found in isolates from humans and four sources.

As the growth of population between 2006 and 2013 seems marginal, the case rate per 100,000 people based on the population in 2006 is calculated and illustrated in Figure 2.2. There was a large decline in the number of urban cases from 2008 due to the majority of dwellers in the Manawatu region residing in urban areas, which coincides with a national intervention in the poultry industry implemented by the New Zealand Food Safety Authority (NZFSA) in 2007 and 2008. Although the urban rates are improved, a temporary effect on the rural rates is also observed, which has increased from 2012 with a peak in 2016 at 232.5 cases per 100,000 population. In addition, the decrease of urban rates in the Manawatu region also reflects to the annual reports of notifiable diseases in New Zealand. The national case rate of campylobacteriosis per 100,000 people in 2008 is markedly decreased to 156.8, compared to the rate of 302.7 in 2007, and since then the number of notified cases remains stable (ESR, 2009; ESR, 2017).

2.3.2 MLST data

Genotype information was obtained from human cases through analysis of stool samples, and from a pool of non-human samples. A total of 2,518 isolates were typed from non-human samples ranging from chicken carcasses, cattle, sheep, environmental water and wild birds to family pets (cats and dogs), over the same time period and from the same geographical location. These samples were further categorised into four groups, representing major sources of infection: poultry (samples from poultry suppliers), ruminants (cattle and sheep), water (environmental water), and other (cats, dogs, various wild birds, and so on).

The total number of unique MLST genotypes typed from all isolates in the Manawatu is 377, of which 38.5% are found among human cases. In these genotypes typed from human cases, 62 of them are not observed from source isolates. Ideally, if all transmission to humans is assumed to come from animal reservoirs, all genotypes typed from the human population would be observed in source populations. However, it is unlikely to occur due to limited sampling of animal reservoirs. In addition, some genotypes

are highly likely to be typed from some sources, but not arise in humans due to lack of human exposure and/or strain virulence. Five common genotypes are listed in Table 2.4; the first four of which are frequently observed in human cases, whilst the last one is largely found in water. As found in other studies, ST-45 and ST-474 are relatively more common in poultry, whereas ST-42 and ST-2026 are mainly detected in ruminants (Colles et al., 2003a; Mullner et al., 2009b; Muellner et al., 2013). The fifth genotype, ST-2381, is not found among human cases, appearing only in water and other sources, in this case being found in pukeko and takahē birds from the Rallidae family (Carter et al., 2009a; French et al., 2014).

Further, the distribution of the top 40 genotypes typed from human cases and each source is illustrated in Figure 2.3. It shows that the predominant genotype attributed to human infection is ST-474, accounting for 17% of total unique genotypes, followed by ST-45, contributing about 10%. For non-human isolates, some genotypes found in human samples may not be observed in each source group. For example, ST-2026 was detected in humans and all sources except for poultry. In general, the most frequently observed STs in each source group based on the top 40 STs detected in humans are ST-45 and ST-48 for poultry, ST-50 and ST-61 for ruminants, ST-45 and ST-42 for water, and ST-45 for other sources. One should be aware that Figure 2.3 only illustrates the genotypes that are found in human cases, so it is unable to provide information regarding genotypes that are not found in humans but may be frequently observed amongst source isolates. For example, ST-2381 is not observed in humans in the dataset but is relatively common in environmental water and other sources.

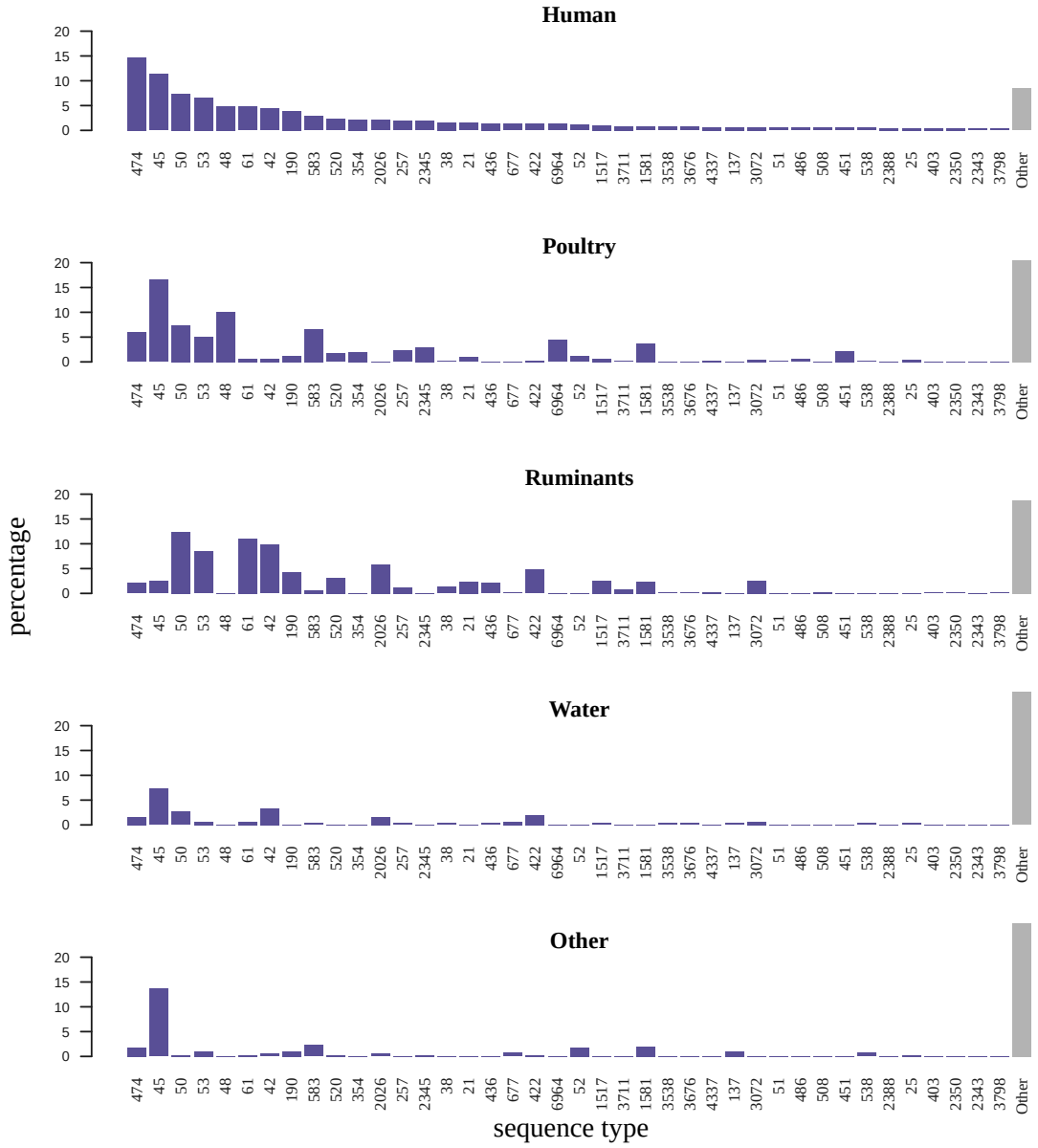


Figure 2.3: The percentage of the top 40 STs detected in human isolates, with the distribution of associated STs found in source isolates. ST-45 is the most frequent genotype found in all isolates, except for the isolates from ruminants contributing to ST-50.

Chapter 3

Background on statistical methods and models

Models incorporating genetic information as well as epidemiological covariates can enhance our understanding about the spread of pathogens so that decision-makers can implement effective plans for risk management and intervention. Methods of developing comprehensive statistical models for such integrated data have been increasingly discussed and applied (Cottam et al., 2008; Mullner et al., 2009a,b, 2010; Morelli et al., 2012). To quantify the relative combination of different sources to *Campylobacter* infection, this thesis proposes models that are in a Bayesian context and are designed to analyse integrated epidemiological and molecular data.

Before going to Chapters 4-7, where the proposed models are introduced, this chapter describes the statistical methods that will be used in modelling and analysing data from the campylobacteriosis study. Section 3.1 covers the statistical methods such as probability distributions and Bayes' theorem. Then, the methods that are used to approximate characteristics of probability distributions from the perspective of Bayesian statistics are introduced in Section 3.2, followed by a section outlining how to fit a regression model with multinomial response, by specifying effects on a logit scale. Lastly, Section 3.4 gives an overview of some existing models quantifying the contribution of sources to zoonotic pathogens.

3.1 Use of Bayesian methods

This section covers background material on mathematical and statistical methods used in the thesis, where Bayesian inference is mainly used to find the posterior attribution probabilities for sources of campylobacteriosis. To begin with, Bayes' theorem is introduced, which leads to the posterior probability of parameters of interest. Then, the probability distributions that are exploited in the analysis are described, followed

with the approach of how to approximate posterior characteristics. Lastly, an overview of statistical models applied to the data are provided, together with a review of some existing source attribution models used within the realm of public health.

3.1.1 Bayes' theorem

Bayes' theorem, which is associated with the multiplicative rule of probability and the law of total probability, is central to a Bayesian statistical inference. It allows us to reverse the direction of likelihood functions when making probabilistic statements.

The multiplicative rule states that, for two events E_1 and E_2 of interest coming from the same sample space, the probability of observing both events that will occur can be derived from,

$$P(E_1 \cap E_2) = P(E_1)P(E_2 \mid E_1),$$

or, alternatively,

$$P(E_1 \cap E_2) = P(E_2)P(E_1 \mid E_2).$$

Since the right-hand side of the first equation is same as the one in the alternative equation, we divide the second equation by $P(E_1)$ to find the conditional probability $P(E_2 \mid E_1)$,

$$P(E_2 \mid E_1) = \frac{P(E_2)P(E_1 \mid E_2)}{P(E_1)}, \quad P(E_1) \neq 0. \tag{3.1}$$

Given that E_2 and E_1 (the complement of E_2) are mutually exclusive, the denominator can be rewritten by applying the law of total probability,

$$\begin{aligned} P(E_1) =& P\Big[(E_1 \cap E_2) \cup (P(E_1 \cap \text{not } E_2)\Big] \\ =& P(E_1 \cap E_2) + P(E_1 \cap \text{not } E_2) \\ =& P(E_2)P(E_1 \mid E_2) + P(\text{not } E_2)P(E_1 \mid \text{not } E_2). \end{aligned} \tag{3.2}$$

After incorporating the above expansion into Equation (3.1), Bayes' theorem tells us that the probability of observing E_2, given that E_1 has already occurred, results in,

$$P(E_2 \mid E_1) = \frac{P(E_2)P(E_1 \mid E_2)}{P(E_2)P(E_1 \mid E_2) + P(\text{not } E_2)P(E_1 \mid \text{not } E_2)}.$$

If E_2 is subdivided into $i = 1, \ldots, n$ categories, which are exhaustive and mutually

exclusive, the above conditional probability for each sub-event can be generalised to,

$$\mathrm{P}(\mathrm{E}_{2i} \mid \mathrm{E}_1) = \frac{\mathrm{P}(\mathrm{E}_{2i})\mathrm{P}(\mathrm{E}_1 \mid \mathrm{E}_{2i})}{\sum_{i=1}^{n} \mathrm{P}(\mathrm{E}_{2i})\mathrm{P}(\mathrm{E}_1 \mid \mathrm{E}_{2i})}. \tag{3.3}$$

In summary, Bayes' theorem provides a way to facilitate an inference by reversing conditional probabilities. The procedure in the above equation makes use of the prior on E_{2i} about E_1, yielding the final belief about the plausibility of E_{2i}, given the observed event E_1.

3.1.2 Probability distributions

The campylobacteriosis data provide the frequency of each genotype found on human and source isolates (see Table 2.4), so a multinomial distribution is used to model the probability of the combination of frequency that each genotype may be typed from humans or from a source of origin.

The multinomial distribution

For I different genotypes that are typed from n samples collected from one source population, let x_i denote the number of times the ith genotype is found in the source isolates and θ_i denote the probability that the ith genotype is observed from these samples, where $i = 1, \ldots, I$. Then, the probability mass function (pmf) of the multinomial distribution follows,

$$p(\boldsymbol{x} \mid \boldsymbol{\theta}) = \frac{n!}{\prod_{i=1}^{I} x_i!} \prod_{i=1}^{I} \theta_i^{x_i}, \qquad x_i \in \{0, 1, \ldots, n\}, \quad 0 < \theta_i < 1, \quad n \in \mathbb{N},$$

subject to,

$$\sum_{i=1}^{I} x_i = n, \quad \sum_{i=1}^{I} \theta_i = 1.$$

It is always interesting to know how likely it is that a particular genotype will turn up in light of what have already been observed. Bayes' theorem can provide such a posteriori information, with Equation (3.3) giving,

$$\begin{aligned} p(\boldsymbol{\theta} \mid \boldsymbol{x}) &= \frac{L(\boldsymbol{\theta}; \boldsymbol{x}) p(\boldsymbol{\theta})}{p(\boldsymbol{x})} \\ &\propto L(\boldsymbol{\theta}; \boldsymbol{x}) p(\boldsymbol{\theta}). \end{aligned} \tag{3.4}$$

To estimate the probability of typing genotypes given the observed frequencies $p(\boldsymbol{\theta} \mid \boldsymbol{x})$, three components are required from the above equation: i) $L(\boldsymbol{\theta}; \boldsymbol{x})$, the

likelihood function of the observed data given the parameter vector $\boldsymbol{\theta}$; ii) $p(\boldsymbol{\theta})$, the prior probability of $\boldsymbol{\theta}$ to express our knowledge of the parameters before the data are collected; and iii) $p(\boldsymbol{x})$, the marginal likelihood of the data by integrating $\boldsymbol{\theta}$ out.

Notice that the marginal likelihood is a normalising constant, ensuring the integral of posterior probability density equals to 1. This means that if the form of Equation (3.4) is recognisable as a known distribution, the correct normalising constant can be deduced without having to directly compute the marginal likelihood. In other words, the posterior probability is proportional to the numerator. However, in practice, the marginal likelihood is usually difficult to compute due to non-closed forms or high dimensional parameter spaces. To simplify the computation and hence calculate the posterior quantities of interest, methods such as numerical integration can be used to integrate the numerator in order to work out the normalising constant.

The Dirichlet distribution

In Bayesian inference, the initial belief about the parameter of interest is updated by the event being observed under prior knowledge, so that the conditional probabilities can be reversed in order for the posterior probability to be calculated. In this process, the use of some probability distributions are required for the calculation of the posterior probability when it can be expressed in a closed form. Therefore, it brings out the idea of conjugacy between the prior and the posterior distributions through the type of likelihood functions.

A convenient way to model the prior belief is to choose a distribution that, along with the likelihood, forms a conjugate pair so that the posterior distribution is of the same type as the prior distribution. If data $\boldsymbol{x}$ have n observations, $x_1, \ldots, x_I$, following a multinomial distribution with parameters $\theta_1, \ldots, \theta_I$, then the conjugate prior for $\boldsymbol{\theta}$ is a Dirichlet distribution. In this distribution, $\boldsymbol{\theta}$ are parameterised by a vector of parameters $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_I)$, so the probability density function (pdf) of the prior on $\boldsymbol{\theta}$ is,

$$f(\boldsymbol{\theta}) = \left[\frac{\prod_{i=1}^{I}\Gamma(\alpha_i)}{\Gamma\left(\sum_{i=1}^{I}\alpha_i\right)}\right]^{-1} \prod_{i=1}^{I}\theta_i^{\alpha_i-1}, \qquad \alpha_i > 0. \tag{3.5}$$

Then, the posterior probability in Equation (3.4) is rewritten to the form,

$$\begin{aligned} p(\boldsymbol{\theta} \mid \boldsymbol{x}) &\propto \prod_{i=1}^{I}\theta_i^{x_i}\prod_{i=1}^{I}\theta_i^{\alpha_i-1} \\ &= \prod_{i=1}^{I}\theta_i^{x_i+\alpha_i-1}, \end{aligned}$$

which is again a Dirichlet distribution with a parameter vector $(\boldsymbol{x}+\boldsymbol{\alpha})$. Therefore, the posterior Dirichlet distribution is derived from the observations from data $\boldsymbol{x}$ and the vector of parameters $\boldsymbol{\alpha}$, which is regarded as pseudo counts from pseudo data.

The gamma distribution

The Dirichlet distribution is related to the gamma distribution (see the derivation in Appendix A.1), which will be used in the model development in Chapter 7. As $\boldsymbol{\theta}$ is assumed to be $\text{Dir}(\boldsymbol{\alpha})$, it can be also expressed by introducing $\boldsymbol{Z}$,

$$\theta_i = \frac{z_i}{\sum_{k=1}^{I} z_k},$$

where $\boldsymbol{Z}$ is a random vector of $Z_1, Z_2, \ldots, Z_I$ from a gamma distribution, $G(\alpha_i, \beta = 1)$, $i = 1, \ldots, I$, whose pdf is of the form,

$$f_{Z_1,\ldots,Z_I}(z_1, \ldots, z_I) = \prod_{i=1}^{I} \frac{1^{\alpha_i}}{\Gamma(\alpha_i)} z_i^{\alpha_i - 1} \exp(-1 \cdot z_i), \quad 0 < z_i < \infty.$$

Thus, a gamma distribution can be used to produce samples from a Dirichlet distribution. This is particularly advantageous to unrestrict the parameter space for the Dirichlet distribution, where the range of $\boldsymbol{\theta}$ is between 0 and 1. The equivalent parameters of the gamma distribution can be alternatively used due to $\boldsymbol{Z} \in [0, \infty)$.

The Cauchy distribution

The last probability distribution to be detailed is the Cauchy distribution. Let W denote a random variable from this distribution, whose pdf is defined as,

$$f(w) = \frac{1}{\pi\gamma \left\{1 + \left(\frac{w-\mu_0}{\gamma}\right)^2\right\}},$$

where μ_0 and γ represent the location of the peak and the dispersion of the distribution. The value of γ usually denotes half of the interquartile range (IQR) as the probable error, specifying that 50% of the values lie within the interval about the centre of the distribution and 50% outside.

The shape of the Cauchy distribution is symmetric with the mode and median at the peak. However, the mean and the variance are undefined due to the non-integrable random variable. An intuitive way to think about the non-convergence of the mean for the Cauchy distribution is its super-heavy tails. This property is suitable to meet the assumptions considered in the model development in Chapter 7, illustrating the indirect link of observing genotypes between sources and humans. The distribution of

genotypes between sources and humans may not be the same as only a small number of genotypes are believed to be more prevalent than others in sources or humans.

3.2 Approximation of posterior characteristics

The posterior density can be found in theory using Equation (3.4). However, in practice estimating the evidence $p(\boldsymbol{x})$ is often intractable due to the high dimensional integral, yielding to unfeasible analytical computation. Nowadays, researchers exploit approximation techniques to simply the computation such as Markov chain Monte Carlo (MCMC) in order for the quantities of interest are calculated.

MCMC methods are centrally built by Markov chains and Monte Carlo integration. For example, the posterior expectation of $\boldsymbol{\theta}$,

$$\mu(\boldsymbol{\theta}) = \int \boldsymbol{\theta} p(\boldsymbol{\theta} \mid \boldsymbol{x}) d\boldsymbol{\theta},$$

can be calculated through MCMC simulation by generating a set of random samples $(\theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(m)})$ from the posterior density $p(\boldsymbol{\theta} \mid \boldsymbol{x})$ to estimate the integral. This is so called Monte Carlo integration, which exploits the Law of Large Numbers to ensure that the sample mean,

$$\hat{\mu}_m(\boldsymbol{\theta}) = \frac{1}{m} \sum_{j=1}^{m} \theta^{(j)},$$

converges almost surely to the true expectation. Markov chains provide a convenient way of sampling from the posterior distribution when the normalising constant is unknown.

A Markov chain is a sequence of random variables $(\theta^{(1)}, \theta^{(2)}, \theta^{(3)}, \ldots,)$ evolving over discrete time that has a finite state space Θ (Robert and Casella, 2004). It has a property that the probability distribution of a chain moving to $\theta^{(j)}$ (the state of θ at step j) is conditionally independent of any state before step $j-1$, given the current state of $\theta^{(j-1)}$,

$$p\Big(\theta^{(j)} \mid \theta^{(j-1)}\Big) = p\Big(\theta^{(j-1)} \mid \theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(j-2)}\Big).$$

Hence, this Markov property shows the evolution of the chain is 'memoryless', no matter the behaviour the sample values before, the chain only considers the most recent state of the sample to decide the destination of the next jump. If it is regular, it will in time converge to generate samples from the target (posterior) distribution. Therefore, the idea for a Markov chain is to construct a chain for which the stationary distribution matches the target distribution. Notice that a chain can start from any

arbitrary point. Nonetheless, to ensure that samples simulated in a chain converge to the target distribution, 'burn-in' is used to discard the initial samples which may start far in the tails of the distribution. On the other hand, the jumps in the Markov chain are conditionally independent, indicating the samples will be serially correlated. To reduce the correlation between samples, 'thinning' is introduced to keep every k samples but throw away the rest of samples; it is also of benefit, particularly to computers with restricted storage.

There are a number of techniques introduced in the MCMC methods for obtaining samples from probability distributions. Two common sampling methods (the Metropolis-Hastings algorithm and the Gibbs sampler) will be used in the thesis.

3.2.1 The Metropolis-Hastings algorithm

Ideally, samples are drawn directly from the posterior density in the MCMC method. However, when the target posterior density is unavailable in a closed form, that is, the normalising constant in Equation (3.4) is unknown, an arbitrary proposal density, $q(\boldsymbol{\theta}^* \mid \boldsymbol{\theta})$ can be used instead, from which a candidate $\boldsymbol{\theta}^*$ is sampled conditional on the current value of $\boldsymbol{\theta}$. This is how the Metropolis-Hastings algorithm produces a Markov chain for $\boldsymbol{\theta}$ with respect to the target density through a proposal density, when the proposal density is easy to simulate.

The Markov chain produced here is stationary and reversible if the condition below is satisfied,

$$\pi(\boldsymbol{\theta^*})q(\boldsymbol{\theta} \mid \boldsymbol{\theta^*}) = \pi(\boldsymbol{\theta})q(\boldsymbol{\theta^*} \mid \boldsymbol{\theta}), \tag{3.6}$$

where $\pi(\boldsymbol{\theta})$ is the posterior density. However, the moves from $\boldsymbol{\theta}$ to $\boldsymbol{\theta}^*$ and the reversed jump may be imbalanced in the course of updating proposals. In other words, one state may move less or more often than the other compared to the reverse. To ensure the system is in equilibrium, an acceptance probability δ is introduced on both sides of Equation (3.6) to adjust the system in case it is imbalanced (Chib and Greenberg, 1995; Gelman et al., 2014a). Thus, the move from one to the other state is accepted with the probability,

$$\delta = \min\left\{1, \underbrace{\frac{\pi(\boldsymbol{\theta}^*)}{\pi(\boldsymbol{\theta})}}_{\text{target ratio}} \times \underbrace{\frac{q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^*)}{q(\boldsymbol{\theta}^* \mid \boldsymbol{\theta})}}_{\text{proposal ratio}}\right\}.$$

In fact, the target ratio is independent of the normalising constant as it is cancelled, and hence only the move of $\boldsymbol{\theta}$ associated with the product of the likelihood function and the prior needs to be evaluated. Here a generic Metropolis-Hastings algorithm is

outlined for T samples of $\boldsymbol{\theta}$:

0. Choose an arbitrary random value of $\theta^{(0)}$ as a starting point. For $t = 0, 1, \ldots, T-1$, proceed to the following steps:

1. Sample a candidate $\boldsymbol{\theta}^*$ from the proposal distribution, $q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)})$.

2. Generate a random uniform value, $u \sim U(0, 1)$.

3. Calculate the acceptance rate δ to update the proposal,

$$\delta = \min\left\{1, \frac{L(\boldsymbol{\theta}^* \mid \boldsymbol{x})}{L(\boldsymbol{\theta}^{(t)} \mid \boldsymbol{x})} \times \frac{p(\boldsymbol{\theta}^*)}{p(\boldsymbol{\theta}^{(t)})} \times \frac{q(\boldsymbol{\theta}^{(t)} \mid \boldsymbol{\theta}^*)}{q(\boldsymbol{\theta}^* \mid \boldsymbol{\theta}^{(t)})}\right\}.$$

4. If $u < \delta$, accept the proposal to be the value in the next state so that $\theta^{(t+1)} = \theta^*$, otherwise, retain the previous accepted value to the next state, that is, $\theta^{(t+1)} = \theta^{(t)}$.

5. Repeat steps 1 to 4 $(T-1)$ times to obtain the posterior samples for $\boldsymbol{\theta}$.

3.2.2 The Gibbs sampler

Another technique to approximate the characteristics of the posterior distribution when it is difficult to sample from is the Gibbs sampler. It is a special case of the Metropolis-Hastings algorithm as a proposal distribution is not required, but full conditionals are known. This means that the conditional probability of each parameter, given all the others, is analytical. Unlike the aforementioned algorithm, the acceptance probability in this sampler equals to 1, meaning that the proposals sampled from each conditional probability are always accepted (Robert and Casella, 2004; Kadane, 2011; Gelman et al., 2014a).

Let $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_K)$ be the parameter vector and $\boldsymbol{\theta}_{-k}$ denote $\boldsymbol{\theta}$ without θ_{-k}, $k = 1, \ldots, K$. Each conditional probability is assumed to be expressed analytically. The parameter vector $\boldsymbol{\theta}$ can be updated with the following steps:

0. Let $t = 0$, choose a set of arbitrary initial values, $\boldsymbol{\theta}^{(t)} = (\theta_1^{(t)}, \ldots, \theta_K^{(t)})$.

1. Draw a sample of $\theta_1^{(t+1)}$ from $p(\theta_1 \mid \boldsymbol{\theta}_{-1}^{(t)})$.

2. Draw a sample of $\theta_2^{(t+1)}$ from $p(\theta_2 \mid \boldsymbol{\theta}_{-2}^{(t+1)})$.

3. Keep sampling each component of $\boldsymbol{\theta}$ by updating the conditional probability until $\theta_K^{(t+1)} \sim p(\theta_K \mid \boldsymbol{\theta}_{-K}^{(t+1)})$.

4. Repeat steps 1 to 3 until the number of posterior samples $t = T$ is reached.

3.3 Generalised linear regression model

The ultimate goal in this thesis is the estimation of the final attribution probability. To that end, a generalised linear regression model on the logit scale is considered in order to include individual level variables within the probabilities.

Let J denote the number of source categories, h denote the index of each human case, and $\hat{\pi}_{hj}$ denote the posterior attribution probability for h^{th} human case caused by source j. An intuitive way of modelling is to relate the linear combination of variables,

$$\eta_{hj} = \beta_{j0} + \beta_{j1}c_{1h} + \cdots + \beta_{jp}c_{ph},$$

to the probability $\hat{\pi}_{hj}$. However, it has a structural drawback in that the linear combination may lead to a quantity outside the range of the probability between 0 and 1. Therefore, the linear model on a logit scale is used to model the probability, which can be expressed below,

$$\hat{\pi}_{hj} = \frac{\exp(\eta_{hj})}{\sum_{j=1}^{J}\exp(\eta_{hj})},$$

with the guarantee of $\sum_{j=1}^{J}\hat{\pi}_{hj} = 1$.

Note that if there is a single categorical variable with L levels, then $\hat{\pi}_{hj}$ and η_{hj} take no more than L distinct values and at times the h index represents the factor level. The application can be seen in Chapter 5.

3.3.1 Model comparison

There are different ways to explain the data used in this thesis, that is, more than one model that considers different combinations of variables or treats a variable in a different manner such as numerical and categorical. In order for the goodness-of-fit of these generalised linear models to be evaluated, accuracy between models is measured and compared to see which model is best for prediction.

A classical way to quantify the model performance is to use deviance. Suppose data $\boldsymbol{x}$ have a log-likelihood function $l(\boldsymbol{\theta};\boldsymbol{x})$, the deviance is defined as $-2l(\boldsymbol{\theta};\boldsymbol{x})$ (Nelder and Wedderburn, 1972). This method allows the comparison between two nested models in hypothesis testing. The model with the lower deviance is the better model. For non-nested models, Akaike information criterion (AIC) is an alternative and is given by,

$$-2l(\hat{\boldsymbol{\theta}}_{\text{MLE}};\boldsymbol{x}) + 2p,$$

where p denotes the number of parameters used in the model and the log-likelihood

function is evaluated at the maximum likelihood estimate $\hat{\boldsymbol{\theta}}$. An issue to compare the two non-nested models, given the same data, is that parameters are estimated twice in these models, so AIC penalises the double estimation by $2p$ (Spiegelhalter et al., 2014). However, AIC may not work well for Bayesian hierarchical models with priors imposed on parameters as the number of parameters is uncertain.

Consider the source attribution models developed in this thesis are in a Bayesian framework, deviance information criterion (DIC), which is a generalisation of AIC, is applied for model comparison in Chapter 5. Unlike AIC based on the maximum likelihood estimate $\hat{\boldsymbol{\theta}}$, the calculation of DIC requires the posterior samples for the parameter of interest, and the log-likelihood function that relates to the posterior probability as it provides information to measure how a model fits (Gelman et al., 2014b). A model with smaller DIC is considered to be better goodness-of-fit.

Given M posterior samples for the parameter $\boldsymbol{\theta}$, the generic computation of DIC starts from the goodness-of-fit of a model, which can be quantified by calculating the expectation of the deviance $\bar{d}$ specified as,

$$\bar{d} = -2\frac{1}{M}\sum_{m=1}^{M} l(\theta_m^{\text{post}}; \boldsymbol{x}),$$

in which the log-likelihood function evaluated at $\theta_1^{\text{post}}, \ldots, \theta_M^{\text{post}}$ is involved. Then, for the model complexity, the effective number of parameters P_d is considered and is formulated as,

$$\begin{aligned} P_d &= \bar{d} - \hat{d}, \\ \hat{d} &= -2l(\bar{\theta}_{\text{post}}; \boldsymbol{x}), \end{aligned}$$

with the log-likelihood function evaluated at the posterior sample mean, $\bar{\theta}_{\text{post}} = \frac{1}{M}\sum_{m=1}^{M}\theta_m^{\text{post}}$. Therefore, the DIC is a measure of model fit in conjunction of model complexity and is expressed as,

$$\begin{aligned} \text{DIC} &= \bar{d} + P_d \\ &= 2\bar{d} - \hat{d} \\ &= 2l(\bar{\theta}_{\text{post}}; \boldsymbol{x}) - 4\frac{1}{M}\sum_{m=1}^{M} l(\theta_m^{\text{post}}; \boldsymbol{x}). \end{aligned} \tag{3.7}$$

3.4 Existing source attribution models

Studies on source attribution dating back to two decades ago used classical epidemiological approaches to analyse surveillance data with variables in order to assess the

disease burden arising from the environment and food consumption (Adak et al., 1995; Eberhart-Phillips et al., 1997; van Pelt et al., 1999). Some approaches (e.g. the Dutch model) are based on statistical modelling for microbial subtyping data, others (e.g. STRUCTURE) are using genetic relatedness among samples to estimate the proportion of human cases attributed to sources (Mughini-Gras et al., 2018; Cody et al., 2019). Over the last decade Bayesian perspective has also been introduced in the modelling (e.g. the HaldDP model) to quantifying the contribution of sources (Mughini-Gras et al., 2018). Hence, this section summarises some of the source attribution models that have been developed and discussed in the literature.

3.4.1 Conventional perspective

The Dutch model

To begin with, van Pelt et al. (1999) proposed the Dutch model in 1999. This approach can estimate the proportion z_j of human cases from source j via the model,

$$z_j = \frac{\hat{\lambda}_j}{\sum_{s=1}^{J} \hat{\lambda}_s},$$

with the expected number of human cases $\hat{\lambda}_j$ arising from source j, which can be obtained by summing λ_{ij} over all strain types,

$$\lambda_{ij} = \frac{k_{ij}}{\sum_{s=1}^{J} k_{is}} h_i,$$

where k_{ij} is the relative frequency of strain type i found on source j, and h_i is the number of strain type i found on human cases. This model was further developed from the perspective of frequentist to Bayesian statistics in 2004, that is, the Hald model.

3.4.2 Bayesian perspective

The Hald model

The Hald model (Hald et al., 2004) extends the principle of the Dutch model by additionally considering pathogenic ability q_i for survival in humans with strain type i, and a_j for force on source j as food. Let n_i denote the number of human cases with strain type i, and Hald et al. (2004) modelled it through a Poisson distribution,

$$n_i \sim \text{Poisson}\Big(\sum_{j=1}^{J} \lambda_{ij}\Big), \quad \lambda_{ij} = m_j p_{ij} q_i a_j, \quad n_i = 0, 1, \ldots \tag{3.8}$$

with the parameter λ_{ij} representing the expected number of human cases per year with strain type i from source j. This parameter can be derived from the product of: i) m_j, the quantity of consumed food source j; ii) p_{ij}, the prevalence of the strain type i on source j; and iii) the factors influencing the pathogenic ability in the transmission: a_j on food source j and q_i on strain type i.

Nonetheless, this model has limitations (Mullner et al., 2009a; Miller et al., 2017): i) the uncertainty of the source prevalence was not considered; ii) the source and strain type-specific factors were fixed; and iii) the model is under-identified as the underlying parameters were not able to be estimated from the data alone. The number of parameters for sources and strain types is $I + J$, which is more than the data points I.

The modified Hald model

The modified Hald model (Mullner et al., 2009a) improved the original Hald model by not only splitting the data into different time frames to satisfy identifiability, but also incorporating uncertainty in p_{ij}, q_i and a_j, which are specified below,

$$
\begin{aligned}
&p_{ij} = \pi_j r_{ij}, \quad \pi_j \sim \text{Beta}(1,1), \quad r_{ij} \sim \text{Dir}(\mathbf{1});\\
&\log q_i \sim \text{N}(0,\tau), \quad \tau \sim \text{Gamma}(0.01, 0.01);\\
&a_j \sim \text{Exponential}(0.002).
\end{aligned}
$$

This model is similar to the original model (3.8), but with an additional index t to specify each parameter in time interval t except for a_j and q_i. In addition, m_j is removed from the model as the amount of food source consumed is absorbed by the magnitude of a_j describing the force of infection as a vehicle in the transmission.

The novel approach for this model is to incorporate uncertainty into parameters. First, the prevalence p_{ij} is estimated by modelling the process of source sampling through π_j and r_{ij}, where the first parameter is the overall prevalence of positive isolates from source j, while the latter one denotes the proportion of isolates typed from source j with strain type i. Secondly, the factor q_i this time is modelled hierarchically with a gamma distributed prior on τ, allowing the variation characterised in the strain types. Lastly, a constant of uninformative prior 0.002 is assigned to the exponential distributed a_j in order to avoid a large amount of force on food sources. In fact, the sensitivity analysis for this prior shows that there is no significant effect on the final inference between the value of 0.01 and 0.002.

The modified Dutch model

The modified Dutch model (Mughini-Gras et al., 2014) retains the principle of the Dutch model, however, it was developed in a Bayesian framework. Not only does this

model consider the time period of study, but it also incorporates the uncertainty of the prevalence of the strain types on sources as well as the quantity of food consumption.

The parameters involved in this model are similar to that specified in the previous source attribution models (from the Dutch model to the modified Hald model), but with an index t specifying a time period of observation. It estimates the expected number of cases λ_{ijt}, who are of strain type i, given the infection is from source j in a period of time t via,

$$\lambda_{ijt} = \frac{p_{ijt} m_{jt}}{\sum_{j=1}^{J} p_{ijt} m_{jt}} h_{it},$$

where p_{ijt} denotes the prevalence of strain type i on source j in time t, m_{jt} represents the average amount of food consumption per person per day, which is from source j in time t, and h_{jt} specifies the number of cases who are of strain type i in time t. The distribution of $\log(m_{jt})$ is assumed to be normal with parameters μ_{jt} and σ_{jt}, whereas the prevalence p_{ijt} is estimated using the assumption defined in the modified Hald model, i.e. $p_{ijt} = \pi_j r_{ijt}$. Similar to the modified Hald model, this model assigns Beta and Dirichlet priors to π_j and r_{ijt}, respectively, with parameter settings differing from the modified Hald model.

Despite some parameters commonly used in both the modified Hald and modified Dutch models, there are some differences in source attribution results between these two models. Mughini-Gras et al. (2014) discussed that reasons causing the discrepancies could be different methods of computation and lack of parameters a_j and q_i considered in the latter model. In comparison with the modified Hald model, the modified Dutch model does not consider the chance of a source becomes a vehicle, for example, undercooked meat, and it disregards the abilities of strain types in the transmission such as survivability and virulence.

The HaldDP model

Miller et al. (2017) recently published an R package named *source*R based on the developed HaldDP model that refines the properties of the original Hald model combining with the modified one.

This model proposed solutions to the drawbacks inherent in both Hald models. First, the assumption of a log-normal distributed random variable for the type-specific factor $\boldsymbol{q}$ along with a gamma distributed prior on τ may lead to over-dispersion (Gelfand et al., 1995; Miller et al., 2017). To improve this approach, $\boldsymbol{q}$ is instead assumed to be from a Dirichlet process defined below,

$$q_i \sim \mathrm{DP}(c_q, H_0),$$

where H_0 is the base distribution and c_q is a scaling parameter. This model allows for sampling a distribution of $\boldsymbol{q}$ through the Dirichlet process, which is a stochastic process whose range is a set of probability distributions, starting from the base distribution H_0, with a positive real number c_q specifying the variation of the distribution.

Second, the sum of posterior absolute prevalence of strain type i on source j over strain types should be equal to the overall prevalence of strain types from source j, i.e. $\sum_{i=1}^{I} p_{ij} = \pi_j$. Nevertheless, the Beta distributed priors imposed on p_{ij} lead to the probability π_j outside the range of $(0, 1)$. Given $\boldsymbol{p} = \boldsymbol{\pi r}$, this model instead assumes that the number of positive source samples follows a multinomial distribution with the probability r_{ij} (so that $\sum_i r_{ij} = 1$ is retained) and also fixes π_j as an empirical proportion of tested source samples with positive tests. Lastly, this model keeps the temporal variable in the frame, but incorporates additionally a location variable l, leading to,

$$o_{itl} \sim \text{Poisson}\Big(\sum_{j=1}^{J} \lambda_{ijtl}\Big), \quad \lambda_{ijtl} = p_{ijt} q_i a_{jtl},$$

with the assumptions that the spatial and temporal variables such as the case notification date and the location are usually regarded as risk factors of human diseases, and that sources are only considered to be affected by the temporal variables.

The Wilson model

Wilson et al. (2008) proposed an approach that differs from the above models in that it considers not only the epidemiological data but the genetic evolution of genotypes observed in human and non-human isolates.

This method is comprised of two models. The first one is to use a multinomial distribution to model human cases attributed to sources. Let F_{S_h} denote the probability of observing the h^{th} human case who is contracted from a source of origin S. Assuming the source of origin for each individual S_h is known, the pmf can be written as,

$$p(S \mid F) \propto \prod_{h=1}^{H} F_{S_h}.$$

Ideally, the posterior probability of F can be estimated via Bayes' theorem by incorporating an appropriate prior $p(F)$. However, the assumption of known S_h is unrealistic. To know better about how the source of origin contributes to human cases, an evolutionary model, which is the second part of the Wilson model, is therefore developed to find the possible source for each human case based on genotypes G typed from individuals.

Genotype	*asp*A	*gln*A	*glt*A	*gly*A	*pgm*	*tkt*	*unc*A
45	4	7	10	4	1	7	1
8072	4	7	10	4	788	7	1
137	4	7	10	4	42	7	1
583	4	7	10	4	42	51	1
3718	2	4	1	4	1	1	5

Table 3.1: A mutation may be observed at *pgm* for ST-8072 after comparing it to ST-45. ST-137 and ST-3718 could arise through a recombination as allele 42 in ST-137 and may be from ST-583, while ST-45 and ST-3718 only share one common allele at *pgm*.

Hence, the posterior probability becomes $p(F, S \mid G) \propto p(G \mid S, F)p(S \mid F)p(F)$.

The authors analysing MLST data on campylobacteriosis envisage that genetic changes in *Campylobacter* result in different strains typed from samples. These changes are assumed to be mutations and recombinations undergone within sources, and migrations moved between sources as 'islands'.

In this model, the definition of genetic changes such as mutation is not identical to that from the perspective of biology. A mutation here means that an allele at a locus is novel and has never appeared in other genotypes before. This can be seen in Table 3.1 showing allelic profiles for five genotypes; the allelic profile of ST-45 is almost same as ST-8072, with one allele differing at the locus *pgm*. Allele 788 had not seen in any other genotypes in isolates until ST-8072 typed from a sample. Thus, this allele is very likely a new mutant at this locus for ST-8072. Further, a recombination means an allele at a locus that has already been observed in other allelic profiles. The allelic profile of ST-45 in the table is nearly identical to ST-137 except for the allele at the *pgm* locus. Allele 42 at the *pgm* locus for ST-137 is common in other genotypes such as ST-583, and so ST-137 might be the result of a recombination between ancestors of ST-45 and ST-583. In addition, a migration is straightforward: a genotype typed from at least two different sources indicates that it moves from one source to the other.

These evolutionary events can be parameterised given the event happened at each locus is either independent (unlinked locus) or dependent (linked locus). The definition of parameters used in the evolutionary model are listed in Table 3.2. With the assumption of unlinked loci, the likelihood of sampling a genotype g (with alleles at each of seven loci) from source population j which experiences either mutation or migration can be computed by,

$$L_{\text{unlinked}}(G; S) = \prod_{l=1}^{7} \begin{cases} \mu_j, & \text{if } g \text{ is novel} \\ (1 - \mu_j)B^l_{gj}, & \text{otherwise,} \end{cases} \tag{3.9}$$

Parameter	Description
μ_j	the probability of sampling an allele from source population j which is a new mutant among all isolates, $j = 1, \ldots, J$
B^l_{gj}	the probability of sampling g, an allele at locus l in genotype g, from source population j, $l = 1, \ldots, 7$
$M_{jj'}$	the probability of sampling an allele from source population j that has been typed from source population $j^{'}, j^{'} \neq j$
$F^l_{gj'}$	the frequency with which g has been observed in those genotypes typed from source population $j^{'}$
N_g	the total number of genotypes typed from all isolates
R_j	the probability of a genotype typed from source population j that has undergone recombination per locus

Table 3.2: The definition of each parameter used in the evolutionary model in (3.9) and (3.10).

where $B^l_{gj} = \sum_{j'=1}^{J} F^l_{gj'} M_{jj'}$ is in the absence of mutation. By contrast, to model the linked locus, i.e. the seven loci are dependent, two genotypes g and c typed from different source populations are evaluated. The sampling of an allele at a specific locus of these two genotypes can be expressed via mutation, recombination and migration,

$$L_{\text{linked}}(G;S) = \sum_{c=1}^{N_g} \frac{M_{jJ_c}}{N_{J_c}} \prod_{l=1}^{7} \begin{cases} \mu_j, & \text{if } g \text{ is novel,} \\ (1-\mu_j) R_j B^l_{gj}, & \text{if } g \neq c, \\ (1-\mu_j) R_j B^l_{gj} + (1-\mu_j)(1-R_j), & \text{if } g = c. \end{cases} \tag{3.10}$$

With appropriate priors assigned to the evolutionary parameters, the estimates of these parameters can be obtained via the MCMC method. Hence, the proportion of human cases attributable to each source of origin can be approximated through Bayes' theorem,

$$p(F \mid S, G) \propto \prod_i \Big(\sum_j p(S \mid F) L(G;S) \Big) p(F),$$

which sums over the source populations accounting for the overall uncertainty from each source.

3.4.3 Summary

The objective of these representative models for source attribution developed between 1999 and 2017 is to estimate the proportion of cases attributable to sources. The Dutch

model simply assumes the probability that different strain types cause to human infection is the same (Mullner et al., 2009b). However, some strain types may be more dominant than others. For example, some STs in Figure 2.3 are more frequently observed in particular sources such as ST-50. Hence, the assumption of the Dutch model is improved by the Hald and modified Hald models; that the survival and virulence abilities on food and strain types are considered by using Bayesian inference to incorporate the uncertainty on parameters. Recently, these two Hald models were refined by Miller et al. (2017) through jointly modelling the source and human data with respect to the type factor using a non-parametric sampling approach (Mughini-Gras et al., 2018; Liao et al., 2019). These models make use of sampling and simulation to link the association between human cases and sources via the prevalence of strain types. However, the cause of different strain types observed in some sources and humans remains unknown. The Wilson model estimates the genotype distribution on each source by modelling the genetic evolutionary process rather than using only the frequencies of each type. By doing so it can potentially utilise the genetic similarities between types to inform the attribution.

In this thesis, models based on the Bayesian paradigm for source attribution of campylobacteriosis are proposed. The evolutionary part of the Wilson model will be adopted in addition to building a new, simpler model as a base for comparison and extension. The evolutionary model provides valuable insight into the disease transmission from an epidemiological to molecular perspective. Therefore, it will be used as a priori knowledge in the modelling, where the adaptation can be found in Chapters 5 and 6.

Chapter 4

Bayesian modelling for the source of campylobacteriosis

The source and human data that originated from the campylobacteriosis study described in Chapter 2.2 will be modelled using a Bayesian approach in this chapter, with the aim of assigning sources probabilistically to human cases. The development of models is described in Section 4.1. The model fitting with variables and the applied MCMC algorithms can be found in Section 4.2. Section 4.3 shows the final inference of source attribution probability and convergence diagnostics, followed by a section of sensitivity analysis examining how sensitive the results are if some assumptions or settings change. In the last section, a conclusion is made including a discussion about the model being mis-specified.

4.1 Model description

To start with, the number of genotypes typed from each source are regarded as a multinomial experiment, assuming that the observed distribution of genotypes for each source group is representative of the true distribution. Let the realisation x_{ij} of a random variable $\boldsymbol{X}$ denote the number of genotype i found from source j with a probability $\pi_{ij} = p(\mathrm{ST}_i \mid \mathrm{source}_j)$, where $i = 1, \ldots, I$ and $j = 1, \ldots, J$. The multinomial likelihood is of the form,

$$L(\boldsymbol{X}; \boldsymbol{\pi}) \propto \prod_{i=1}^{I} \prod_{j=1}^{J} \pi_{ij}^{x_{ij}}, \quad x_{ij} \in \{0, 1, \cdots, n_j\}, \quad 0 < \pi_{ij} < 1, \tag{4.1}$$

given $n_j = \sum_{i=1}^{I} x_{ij}$ is the total number of genotypes observed on source j, and π_{ij} is subject to $\sum_{i=1}^{I} \pi_{ij} = 1$. Note that x_{ij} can be 0 as it is possible that genotype i is not observed on source j.

For human data, let y_i denote the number of genotype i typed on human cases, which is also assumed to be multinomial distributed with the probability $\hat{\pi}_i$. The likelihood turns out to be,

$$L(\boldsymbol{Y}; \hat{\boldsymbol{\pi}}) \propto \prod_{i=1}^{I} \hat{\pi}_i^{y_i}, \quad y_i \in \{0, 1, \ldots\}, \quad 0 < \hat{\pi}_i < 1, \tag{4.2}$$

where $\sum_{i=1}^{I} y_i = H$ is the total number of observed genotypes from human cases, namely the sample size of human isolates (because each isolate is only typed once).

Assume for now that π_{ij} also applies to human cases, that is, the probability of a human sample arising from source j is also of genotype i. Then, the Law of Total Probability (3.2) is applied to express $\hat{\pi}_i$ in terms of the parameter π_{ij} and the human attribution probabilities. Specifically,

$$p(\mathrm{ST}_i \text{ typed from human cases}) = \sum_{j=1}^{J} p(\mathrm{ST}_i \mid \mathrm{source}_j) p(\mathrm{source}_j), \tag{4.3}$$

or equivalently,

$$\hat{\pi}_i = \sum_{j=1}^{J} \pi_{ij} F_j,$$

where $p(\mathrm{source}_j) = F_j$ is the attribution probability that a random human case is infected from source j. Then, the likelihood (4.2) for the human data leads to,

$$L(\boldsymbol{Y}; \boldsymbol{\pi}, \boldsymbol{F}) \propto \prod_{i=1}^{I} \Big(\sum_{j=1}^{J} \pi_{ij} F_j \Big)^{y_i}. \tag{4.4}$$

As data $\boldsymbol{X}$ and $\boldsymbol{Y}$ are assumed to be independent, the joint likelihood function is simply the product of likelihood (4.1) and (4.4).

For computational convenience to deal with the summation term in the likelihood (4.4), a latent variable $\boldsymbol{Z}$ is introduced, which categorises human cases by sources and genotypes. Specifically, z_{ij} is defined to be the number of human cases attributed to genotype i being found on source j, with a constraint, $\sum_{j=1}^{J} z_{ij} = y_i$. Thus, the total number of genotype i observed on humans is exactly contributed from these sources.

As y_i could result from different combinations of z_{ij}, the multinomial theorem (see Appendix A.2) is used to replace the term, $(\sum_{j=1}^{J} \pi_{ij} F_j)^{y_i}$. Then the multinomial joint

likelihood incorporating with $\boldsymbol{Z}$ turns out to be,

$$L(\boldsymbol{X}, \boldsymbol{Y}; \boldsymbol{\pi}, \boldsymbol{F}) \propto \prod_{i=1}^{I} \sum_{z_{i1}+z_{i2}+\ldots+z_{iJ}=y_i} \binom{y_i}{z_{i1}, z_{i2}, \ldots, z_{iJ}} \prod_{j=1}^{J} \pi_{ij}^{z_{ij}} F_j^{z_{ij}} \prod_{i=1}^{I} \pi_{ij}^{x_{ij}}.$$

In addition, given the priors on π_{ij} and F_j are a Dirichlet distribution $D(\alpha_{ij}^p)$ and some probability distribution $p(F)$, respectively, the posterior for $\boldsymbol{\pi}$ and $\boldsymbol{F}$ are derived as,

$$p(\boldsymbol{\pi}, \boldsymbol{F} \mid \boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z}, \boldsymbol{\alpha}^p) \propto \prod_{i=1}^{I} \sum_{z_{i1}+z_{i2}+\ldots+z_{iJ}=y_i} \binom{y_i}{z_{i1}, z_{i2}, \ldots, z_{iJ}} \prod_{j=1}^{J} \pi_{ij}^{z_{ij}} F_j^{z_{ij}} \prod_{i=1}^{I} \pi_{ij}^{x_{ij}} \prod_{i=1}^{I} \pi_{ij}^{\alpha_{ij}^p - 1} p(F),$$

and the conditional posterior for $\boldsymbol{\pi}$, $\boldsymbol{Z}$ and $\boldsymbol{F}$ can be expressed as,

$$\begin{aligned}
p(\boldsymbol{\pi} \mid \boldsymbol{F}, \boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z}, \boldsymbol{\alpha}^p) &\propto \prod_{i=1}^{I} \prod_{j=1}^{J} \pi_{ij}^{z_{ij}+x_{ij}+\alpha_{ij}^p-1}, \\
p(\boldsymbol{Z} \mid \boldsymbol{F}, \boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{\alpha}, \boldsymbol{\pi}^p) &\propto \prod_{i=1}^{I} \sum_{z_{i1}+z_{i2}+\ldots+z_{iJ}=y_i} \binom{y_i}{z_{i1}, z_{i2}, \ldots, z_{iJ}} \prod_{j=1}^{J} \left(\pi_{ij} F_j\right)^{z_{ij}}, \quad (4.5) \\
p(\boldsymbol{F} \mid \boldsymbol{\pi}, \boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z}, \boldsymbol{\alpha}^p) &\propto \prod_{j=1}^{J} F_j^{\sum_{j=1}^{J} z_{ij}} p(F).
\end{aligned}$$

Notice that the posterior for $\boldsymbol{\pi}$ is $\mathrm{Dir}(\boldsymbol{Z} + \boldsymbol{X} + \boldsymbol{\alpha})$, whereas the posterior for $\boldsymbol{Z}$ is regarded as a multinomial distribution with a vector of parameters $(\boldsymbol{y}, \boldsymbol{\gamma})$, where γ_{ij} is specified as $\frac{\pi_{ij} F_j}{\sum_{k=1}^{J} \pi_{ik} F_k}$ such that $\sum_{j=1}^{J} \gamma_{ij} = 1$. The posterior for $\boldsymbol{F}$ could be analytic, depending on the prior on $\boldsymbol{F}$, which is crucial to deciding the class of MCMC algorithms as it determines whether the posterior $\boldsymbol{F}$ is in a closed form.

4.2 Model fitting and MCMC inference

As described before, for each source, the frequency of genotypes typed from samples follows a multinomial distribution. The probability of typing each genotype has a Dirichlet prior with a parameter α_i^p assumed to be 1, $i = 1, \ldots, 377$, indicating that every isolate is believed to be equally likely *a priori*. Then, the attribution probabilities $\boldsymbol{F}$ can be modelled with or without variables, leading to different MCMC algorithms to obtain the posterior samples.

4.2.1 Model without variables

When the attribution probability F_j for source j is modelled without variables, a Dirichlet distribution with a parameter α_j^F, $j = 1, \ldots, 4$, is considered to depict the uncertainty of the probability. To express the prior belief that the chance of infection arising from each source is equally likely, $\boldsymbol{\alpha}^F = \mathbf{1}$ is assumed. Then, the posterior $\boldsymbol{F}$ in Equation (4.5) results in a closed form of $\text{Dir}(\boldsymbol{Z} + \mathbf{1})$, a Gibbs sampler is hence applied for posterior simulation.

Gibbs sampling algorithm

Algorithm 1 Gibbs sampling when $\boldsymbol{F}$ is modelled without variables

1: Set initial values $\alpha_i^p = 1$, $i = 1, \ldots, 377$
2: Set initial values $\alpha_j^F = 1$, $j = 1, \ldots, 4$
3: Draw samples of $\boldsymbol{\pi}_j^{(0)}$ from $\text{Dir}(\boldsymbol{\alpha}^p)$ for source j, $j = 1, \ldots, 4$
4: Draw samples of $\boldsymbol{F}^{(0)}$ from $\text{Dir}(\boldsymbol{\alpha}^F)$
5: Drawing samples of $\boldsymbol{z}_i^{(0)}$ from a multinomial distribution with parameters $(y_i, \boldsymbol{\gamma}_i^{(0)})$ for genotype i, where

$$\gamma_{ij}^{(0)} = \frac{\pi_{ij}^{(0)} F_j^{(0)}}{\sum_{k=1}^{4} \pi_{ik}^{(0)} F_k^{(0)}}$$

6: **for** iteration $m = 1, 2, \ldots, M$ **do**
7: Draw samples of $\boldsymbol{\pi}_j^{(m)} \sim p(\boldsymbol{\pi}_j \mid \boldsymbol{F}^{(m-1)}, \boldsymbol{z}^{(m-1)})$ for source j, $j = 1, \ldots, 4$
8: Draw samples of $\boldsymbol{z}_i^{(m)} \sim p(\boldsymbol{z}_i \mid \boldsymbol{\pi}^{(m)}, \boldsymbol{F}^{(m-1)})$ for genotype i, $i = 1, \ldots, 377$
9: Draw samples of $\boldsymbol{F}^{(m)} \sim p(\boldsymbol{F} \mid \boldsymbol{\pi}^{(m)}, \boldsymbol{z}^{(m)})$
10: **end for**

The procedure of Gibbs sampling is straightforward as proposals are always accepted in the algorithm, and so the sampling can be directly conducted for each full conditional probability.

Given $\boldsymbol{\alpha}^p$, $\boldsymbol{\alpha}^F$, and data $\boldsymbol{X}$ and $\boldsymbol{Y}$ are fixed, the parameters $\boldsymbol{\pi}$, $\boldsymbol{Z}$ and $\boldsymbol{F}$ are required for sampling in turn from the conditional posterior probabilities. The steps of the sampler are outlined in Algorithm 1. After setting the values to the prior parameters $\boldsymbol{\alpha}^p$ and $\boldsymbol{\alpha}^F$, the parameters $\boldsymbol{\pi}^{(0)}$ and $\boldsymbol{F}^{(0)}$ are initialised in lines 3 and 4 from a Dirichlet distribution. Then, before initialising the pseudo numbers of human cases attributed to genotype i across the four sources (i.e. $\boldsymbol{z}_i^{(0)}$), the probability $\gamma_{ij}^{(0)}$ has to be derived from a function consisting of $\boldsymbol{\pi}$ and $\boldsymbol{F}$. Lastly, the samples of $\boldsymbol{z}_i^{(0)}$ for each genotype are drawn from a multinomial distribution with the number of trails y_i and the probabilities $\boldsymbol{\gamma}_i^{(0)}$.

Once the initial values are obtained, the proceeding steps of drawing samples are similar, but with updated parameters. For each iteration m, $m = 1, 2, \ldots, M$, a new set of $\boldsymbol{\pi}_j^{(m)}$ for each source are sampled from $p(\boldsymbol{\pi}_j \mid \boldsymbol{F}^{(m-1)}, \boldsymbol{z}^{(m-1)})$, which is $\text{Dir}(\boldsymbol{X} + \boldsymbol{z}^{(m-1)} + \mathbf{1})$. Then, similar to line 5, new samples of $\boldsymbol{z}_i^{(m)}$ are drawn from $p(\boldsymbol{z}_i \mid \boldsymbol{\pi}^{(m)}, \boldsymbol{F}^{(m-1)})$, a multinomial distribution with parameters $(y_i, \boldsymbol{\gamma}_i^{(m)})$ after calculating $\gamma_{ij}^{(m)}$, given $\pi_{ij}^{(m)}$ and $F_j^{(m-1)}$. Finally, the sampling for $\boldsymbol{F}^{(m)}$ is conducted by drawing from $p(\boldsymbol{F} \mid \boldsymbol{\pi}^{(m)}, \boldsymbol{z}^{(m)})$, which is again a Dirichlet distribution with a parameter vector $(\mathbf{1} + \sum_{i=1}^{I} z_{ij}^{(m)})$.

4.2.2 Model with variables

When the epidemiological observations are incorporated in the modelling of attribution probability, the fundamental idea is to use a logit function for estimation. This means that the attribution probability F_j for source j is modelled on the logit scale assuming the category J is the baseline,

$$F_j = \frac{\exp(f_j)}{1 + \sum_{l=1}^{J-1} \exp(f_l)}, \quad 0 \leq F_j \leq 1, \quad j = 1, \ldots, J-1, \tag{4.6}$$

where $F_J = 1 - \sum_{j=1}^{J-1} F_j$, and f_j is the log-odds, in which predictors are included.

To include the individual level variables, the calculation of subject-specific attribution probability is required. In other words, the focus of the proportion $\hat{\boldsymbol{\pi}}$ in Equation (4.3) and the attribution probabilities $\boldsymbol{F}$ in Equation (4.6) changes to individual-wise. The dimension of F_j therefore increases to F_{hj}, denoting the attribution probability of source j caused the h^{th} human case illed, where $h = 1, \ldots, 1804$. The probability of observing genoype i in an individual case is then given by,

$$p(\text{ST}_{i[h]} \mid \text{variables}_h) = \sum_{j=1}^{4} p(\text{ST}_{i[h]} \mid \text{source}_j) p(\text{source}_j \mid \text{variables}_h), \tag{4.7}$$

in which the index $i[h]$ identifies genotype i found on human case h, and $p(\text{ST}_{i[h]} \mid \text{source}_j)$ is the probability that genotype i typed from human case h arises from source j. Hence, the likelihood (4.4) can be rewritten in a case-by-case manner,

$$L(\boldsymbol{Y}; \boldsymbol{\pi}, \boldsymbol{F}) \propto \prod_{h=1}^{1804} \sum_{j=1}^{4} \pi_{i[h]j} F_{hj}. \tag{4.8}$$

Consider the case where the genotype data for each individual are supplemented by p additional variables, and the relationship between variables is assumed to be linear. The log-odds f_{hj} for case h who is associated with source j is expressed as a general

linear predictor function,

$$f_{hj} = \beta_{0j} + \beta_{1j}c_{1h} + \cdots + \beta_{pj}c_{ph}, \quad j = 1, \ldots, J-1,$$

and $f_{hJ} = 0$, given the source baseline is the category J. Here, the prior of each regression parameter is assumed to be N(0,1), which is fairly vague when the logit scale is used. However, in such a multicategory logit model, the choice of the baseline may alter the prior specification. For example, the logit regression model with two variables c_1 and c_2 when $J = 3$ is,

$$f_{hj} = \beta_{0j} + \beta_{1j}c_{1h} + \beta_{2j}c_{2h}, \quad j = 1, 2,$$

whereas when the effect between two sources which are not the baseline is of interest, the model is comparing the pair of source categories,

$$\log\left(\frac{F_{h1}}{F_{h2}}\right) = \beta_0^* + \beta_1^* c_{1h} + \beta_2^* c_{2h},$$

where $\beta_0^* = \beta_{01} - \beta_{02}$, $\beta_1^* = \beta_{11} - \beta_{12}$ and $\beta_2^* = \beta_{21} - \beta_{22}$. Then, the prior variance of $\boldsymbol{\beta}^*$ is two times bigger than that of $\boldsymbol{\beta}$. This could have a subtle impact on the final attribution $\boldsymbol{F}$ after the back transformation from the logit scale.

In our case, if ruminants is the source baseline, the prior variance of the regression parameters for comparisons with ruminants on the logit scale will become 1, whereas the parameters modelling the contrast between, say, poultry and water, will have the prior variance as 2. Nonetheless, if poultry is made as the baseline, then those variance results are reversed.

To apply the model of f_{hj} to the campylobacteriosis data, assume c_1 is the numeric variable ranging from -3 to 3 representing the classified rurality of each human case shown in Table 2.2, and c_2 is a binary variable specifying the age of individual being either < 5 or ≥ 5 years. The age covariate was particularly divided into two groups as the prevalence of *Campylobacter* infection in children under 5 years of age is higher than in older children (≥ 5). This represents a distinct category of pre-school children with different exposures (children first attending school in New Zealand is on their fifth birthday) (Baker et al., 2007; French et al., 2008). Then, the model with these two variables and the interaction is algebraically expressed below, given $H = 1804$, $J = 4$, and $P = 4$ denoting the total number of regression parameters considered in the model.

$$\boldsymbol{f} = \boldsymbol{C}\boldsymbol{\beta}, \tag{4.9}$$

where

$$\boldsymbol{f}_{H\times(J-1)} = \begin{bmatrix} f_{11} & f_{21} & \cdots & f_{(J-1)1} \\ f_{12} & f_{22} & \cdots & f_{(J-1)2} \\ \vdots & \vdots & \ddots & \vdots \\ f_{1H} & f_{2H} & \cdots & f_{(J-1)H} \end{bmatrix},$$

$$\boldsymbol{C}_{H\times P} = \begin{bmatrix} 1 & c_{11} & c_{21} & c_{11}c_{21} \\ 1 & c_{12} & c_{22} & c_{12}c_{22} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & c_{1H} & c_{2H} & c_{1H}c_{2H} \end{bmatrix},$$

$$\boldsymbol{\beta}_{P\times(J-1)} = \begin{bmatrix} \beta_{01} & \beta_{02} & \cdots & \beta_{0(J-1)} \\ \beta_{11} & \beta_{12} & \cdots & \beta_{1(J-1)} \\ \beta_{21} & \beta_{22} & \cdots & \beta_{2(J-1)} \\ \beta_{31} & \beta_{32} & \cdots & \beta_{3(J-1)} \end{bmatrix}.$$

The matrix of parameters $\boldsymbol{\beta}$ is comprised of four rows representing the intercept, coefficients for rurality and age variables as well as their interaction term, and the elements c_{1h} and c_{2h} in the design matrix $\boldsymbol{C}$ can be any number of the seven scales and an indicator for age being > 5, respectively, if case h in such an age group was from such a degree of rurality.

Further, $\boldsymbol{Z}$ are simulated through a multinomial distribution after the calculation of $\boldsymbol{F}$ (as the probabilities $\boldsymbol{\gamma}$ depend on the values of $\boldsymbol{F}$). This time, the dimension of $\boldsymbol{Z}$ as well as the parameters are changed from the genotype to the individual level due to variables included in the modelling. That is, the number of trials y_i becomes 1 as only one genotype is typed from each case, and the probability is no longer γ_{ij} but γ_{hj} defined as,

$$\gamma_{hj} = \frac{\pi_{i[h]j}F_{hj}}{\sum_{k=1}^{4}\pi_{i[h]k}F_{hk}}, \tag{4.10}$$

where $i[h]$ is the index identifying h^{th} human case is of genotype i in order for $\boldsymbol{\pi}$ to relate to the associated $\boldsymbol{F}$. However, the dimension of $\boldsymbol{Z}$ has to change back to genotype-wise so that the probability of typing genotype i from source j (i.e. π_{ij}) can be updated from the conditional posterior probability following $\text{Dir}(\boldsymbol{X} + \boldsymbol{Z}^{'} + \boldsymbol{1})$. Therefore, $z_{ij}^{'}$ is obtained via a transformation specified as,

$$z_{ij}^{'} = \sum_{h:i[h]=i} z_{i[h]j}, \tag{4.11}$$

where,

$$z_{i[h]j} = \begin{cases} 1, & \text{if } h^{\text{th}} \text{ case has genotype } i \text{ from source } j \\ 0, & \text{otherwise.} \end{cases}$$

In conclusion, the attribution probability is estimated by a linear regression combination on the logit scale. The posterior attribution probabilities $\boldsymbol{F}$ will then not have a closed form. To simulate the posterior samples, the Metropolis-Hastings-within-Gibbs algorithm is applied as result of the availability of all posterior conditional probabilities except for the posterior $\boldsymbol{F}$.

Metropolis-Hastings-within-Gibbs sampling algorithm

Overall, this sampler outlined in Algorithm 2 has similar steps as Algorithm 1 updating parameters $\boldsymbol{\pi}$ and $\boldsymbol{Z}$, but different procedures for $\boldsymbol{F}$.

Before running the sampler, the number of variables p considered in the model has to be decided, and so the design matrix $\boldsymbol{C}_{H\times P}$ is defined, where $H = 1804$, and $P = 1 + p$. The prior on the probabilities $\boldsymbol{\pi}_j$ for each source is still a Dirichlet distribution with the parameters $\boldsymbol{\alpha}^p = \mathbf{1}$. After initialising $\boldsymbol{\pi}_j^{(0)}$, starting points for $\boldsymbol{\beta}^{(0)}$ and $\boldsymbol{f}^{(0)}$ (and hence the associated $\boldsymbol{F}^{(0)}$ are obtained) are drawn from different probability distributions list from lines 4 to 6. Then, given $\boldsymbol{\pi}^{(0)}$ and $\boldsymbol{F}^0$, the initial number $z_{hj}^{(0)}$ of observing a genotype on human case h, from which the type originates from source j, is simulated via a multinomial distribution with parameters $(1, \boldsymbol{\gamma}_h^{(0)})$, where $\gamma_{hj}^{(0)}$ is computed through Equation (4.10).

For each iteration m starting from line 8, the first thing is to change the dimension of $\boldsymbol{z}^{(m-1)}$ to $\boldsymbol{z}'^{(m-1)}$ via transformation (4.11) in order for $\boldsymbol{\pi}_j^{(m)}$ to be updated using $p(\boldsymbol{\pi}_j \mid \boldsymbol{F}^{(m-1)}, \boldsymbol{Z}'^{(m-1)})$, which is Dirichlet distributed with the parameter vector summing up $x_{ij}, z_{ij}^{'}$ and $\alpha_i^p = 1$ over source j. Then, through Equation (4.10), new samples of $\boldsymbol{z}_h^{(m)}$ for each human case over four sources are drawn from $p(\boldsymbol{z}_h \mid \boldsymbol{\pi}^{(m)}, \boldsymbol{F}^{(m-1)})$ in line 11, which is again a multinomial distribution with the updated probabilities $\boldsymbol{\gamma}_h^{(m)}$. Next, the procedure of random-walk Metropolis sampling is introduced to update indirectly $\boldsymbol{F}$ through regression parameters $\boldsymbol{\beta}$. In line 12, a permutation P_T of $\{1, \ldots, T\}$, where $T = (J-1) \times P = 3P$, is set for updating randomly each element of $\boldsymbol{\beta}$ at a time. For each t, the current value of $\beta_{(t)}^{(m-1)}$ is replaced by an element $\beta_{(t)}^*$ sampled from a proposal distribution $Q(\beta_{(t)}^*, \beta_{(t)}) = N(\beta_{(t)}^{(m-1)}, 1)$. Then, similar to lines 5 and 6, $\boldsymbol{f}^*$ and $\boldsymbol{F}^*$ are calculated, given the new proposal of $\boldsymbol{\beta}^*$. The value of $\beta_{(t)}^*$ is accepted with the probability $\mathcal{A}(\beta_{(t)}^{(m-1)}, \beta_{(t)}^*) = \min\{1, \psi\}$, where ψ is defined in line 18. If $\mathcal{A}$ is larger than a uniform random sample u, the proposal is accepted to be $\beta_{(t)}^{(m)}$ in the next state m, otherwise, the current value is retained to be $\beta_{(t)}^{(m)}$.

Algorithm 2 Metropolis-Hastings-within-Gibbs sampling when $\boldsymbol{F}$ is modelled with variables

1: Define the design matrix $\boldsymbol{C}_{1804\times P}$ depending on the number of parameters P included in the log-odds $\boldsymbol{f}$
2: Set initial values of $\alpha_i^p = 1$, $i = 1, \ldots, 377$
3: Draw initial samples of $\boldsymbol{\pi}_j^{(0)} \sim \text{Dir}(\boldsymbol{\alpha}^p)$ for source j, $j = 1, \ldots, 4$
4: Draw initial values for regression parameters $\boldsymbol{\beta}_{P\times 3}^{(0)}$ from N(0,1)
5: Calculate $\boldsymbol{f}^{(0)} = \boldsymbol{C}\boldsymbol{\beta}^{(0)}$
6: Compute $F_{hj}^{(0)}$ for human case h and source j through Equation (4.6) given $f_{hj}^{(0)}$, $h = 1, \ldots, 1804$, $j = 1, 2, 3$
7: Draw initial combinations of $\boldsymbol{z}_h^{(0)} \sim \text{Multinomial}(1, \boldsymbol{\gamma}_h^{(0)})$ for each human case, where $\gamma_{hj}^{(0)}$ is derived from Equation (4.10) given $\boldsymbol{\pi}^{(0)}$ and $\boldsymbol{F}^{(0)}$
8: **for** iteration $m = 1, 2, \ldots,$ **do**
9: Transform $z_{hj}^{(m-1)}$ to $z_{ij}^{'(m-1)}$ via definition (4.11)
10: Draw samples of $\boldsymbol{\pi}_j^{(m)} \sim p(\boldsymbol{\pi}_j \mid \boldsymbol{F}^{(m-1)}, \boldsymbol{z}^{'(m-1)})$ for source j
11: Draw samples of $\boldsymbol{z}_h^{(m)} \sim p(\boldsymbol{z}_h \mid \boldsymbol{\pi}^{(m)}, \boldsymbol{F}^{(m-1)})$ for case h
12: Sample a permutation P_T of $\{1, \ldots, T\}$, $T = 3P$
13: **for** $t \in P_T$ **do**
14: Let $\boldsymbol{\beta}^* = \boldsymbol{\beta}^{(m-1)}$
15: Propose an element $\beta_{(t)}^*$ from $Q\left(\beta_{(t)}^*, \beta_{(t)}^{(m-1)}\right)$ and substitute it in $\boldsymbol{\beta}^*$
16: Repeat line 5 and 6 to find $\boldsymbol{f}^*$ and the associated $\boldsymbol{F}^*$ given $\boldsymbol{\beta}^*$
17: Generate $u \sim U(0, 1)$
18: Calculate $\mathcal{A}\left(\beta_{(t)}^{(m-1)}, \beta_{(t)}^*\right) = \min\{1, \psi\}$, where

$$\psi = \frac{p\left(\boldsymbol{F}^* \mid \boldsymbol{\pi}^{(m)}, \boldsymbol{z}^{(m)}\right)}{p\left(\boldsymbol{F}^{(m-1)} \mid \boldsymbol{\pi}^{(m)}, \boldsymbol{z}^{(m)}\right)} \frac{\pi\left(\beta_{(t)}^*\right)}{\pi\left(\beta_{(t)}^{(m-1)}\right)}$$

19: **if** $u < \mathcal{A}$ **then**
20: $\beta_{(t)}^{(m)} \leftarrow \beta_{(t)}^*$, implying $\boldsymbol{F}^{(m)} \leftarrow \boldsymbol{F}^*$
21: **else**
22: $\beta_{(t)}^{(m)} \leftarrow \beta_{(t)}^{(m-1)}$, implying implying $\boldsymbol{F}^{(m)} \leftarrow \boldsymbol{F}^{(m-1)}$
23: **end if**
24: **end for**
25: **end for**

Notice that the proposal distribution Q is a random walk proposal. It allows for updating the parameter based on the previous information, i.e. the moves are centred around the current value $\beta_{(t)}^{(m-1)}$ with a random variate of a symmetric distribution such that

$$Q\Big(\beta_{(t)}^{(m-1)} \mid \beta_{(t)}^{*}\Big) = Q\Big(\beta_{(t)}^{*} \mid \beta_{(t)}^{(m-1)}\Big).$$

Thus, the ratio of proposal density equals to 1, simplifying the function of ψ to be a product of the posterior ratio and the prior ratio. In addition, the calculation of the acceptance probability $\mathcal{A}$ is simplified in the programming using the logarithm. Equivalently, the proposal is thus accepted if $\log u < \min\{0, \log \psi\}$, where $\log \psi$ is formed by multiplying the two ratios on the log scale, which are expressed as,

$$\log \frac{p\Big(\boldsymbol{F}^{*} \mid \boldsymbol{\pi}^{(m)}, \boldsymbol{z}^{(m)}\Big)}{p\Big(\boldsymbol{F}^{(m-1)} \mid \boldsymbol{\pi}^{(m)}, \boldsymbol{z}^{(m)}\Big)} = \sum_{h=1}^{1804}\sum_{j=1}^{4} z_{hj}^{(m)}\Big(\log F_{hj}^{*} - \log F_{hj}^{(m-1)}\Big),$$

$$\log \frac{\pi\Big(\beta_{(t)}^{*}\Big)}{\pi\Big(\beta_{(t)}^{(m-1)}\Big)} = \frac{\Big(\beta_{(t)}^{(m-1)}\Big)^2 - \Big(\beta_{(t)}^{*}\Big)^2}{2}.$$

4.3 Results

Posterior attribution of human cases of campylobacteriosis for four sources with the 80% highest posterior density (HPD) credible intervals is presented in this section. An HPD credible interval is used to quantify the uncertainty of posterior attribution probability. It is an interval with the smallest distance between upper and lower bounds among all credible intervals. In this case, the interval for each source is obtained by first ordering the M posterior samples of F_j. Then, given the level is 80%, the ordered samples are divided into two parts under the level of coverage 0.2 (resulted from $1 - 80/100$), meaning that the lower and upper bounds can be the first and last $0.2 \times M + 1$ ordered samples. Finally, the distance between each pair of lower and upper bounds is calculated. The approximate credible interval is the one with the smallest difference, where 80% of posterior probabilities are included.

The results are based on the attribution model both with and without individual level variables, and are obtained after running the algorithm a length of 51,000 iterations. The first 1,000 samples were removed, and every 5^{th} sample was chosen from the sequence. Hence, there is a total number of $M = 10,000$ posterior samples for each parameter.

In addition, convergence diagnostics are also conducted here. The samples drawn

from sequences produced by the algorithms are evaluated whether they are representative of the underlying stationary distribution through graphical assessments. The two aspects to consider here are mixing and convergence. A burn-in period is used in a chain as 'pre-convergence', with the hope that the chain is effectively converged to its stationary (target) distribution thereafter. Mixing is associated with the degree of autocorrelation, even after convergence. Good mixing of a chain means the average and the variation of moves of samples are invariant over time, implying that the sampler efficiently explores the parameter space so samples are being fairly accepted or rejected, and are also less correlated to each other. As samples generated from a Markov chain in the sampler are not completely independent, it may provide less information about the target distribution. However, when the degree of autocorrelation between samples becomes smaller, the amount of information provided by the samples about the posterior distribution increases.

On the other hand, to ensure the convergence of a chain, the number of iterations is also of importance. Throwing away samples in the beginning of a chain may be required when initial points affect the performance of convergence, while discarding samples for every sample that is chosen can reduce the autocorrelation. This can be examined visually by using trace plots, from which multiple chains can be also compared to see if they converge to the similar estimates so that the convergence is detected.

4.3.1 Posterior attribution probability

Results about the posterior attribution probability are displayed, given three ways of estimating the attribution probabilities $\boldsymbol{F}$ demonstrated in the previous section: i) assigning directly a Dirichlet prior on $\boldsymbol{F}$; ii) considering a linear combination on the logit scale for $\boldsymbol{F}$, but no variables are included, i.e. the model only contains the intercept $\boldsymbol{\beta}_0$; and iii) remaining the linear format in ii), but variables such as rurality and age are incorporated.

Attribution probability without variables

The first two methods, which do not include any variables, are not identical due to different prior specifications. When a Dir($\mathbf{1}$) prior is assumed for the attribution probability, it is believed that the infections are contributed evenly by sources. However, when the attribution probability is modelled through $\boldsymbol{f} = \boldsymbol{\beta}_0$, a standard normal prior is assigned to the intercept. This means that N(0,1) is equivalently imposed on the log-odds $\boldsymbol{f}$ when parameters are compared with the source baseline, otherwise the variance becomes double when they are comparing with two non-baseline source categories.

Overall, the final attribution resulting from these two methods are very similar, which can be seen in Figure 4.1. On average, the posterior percentage of human cases

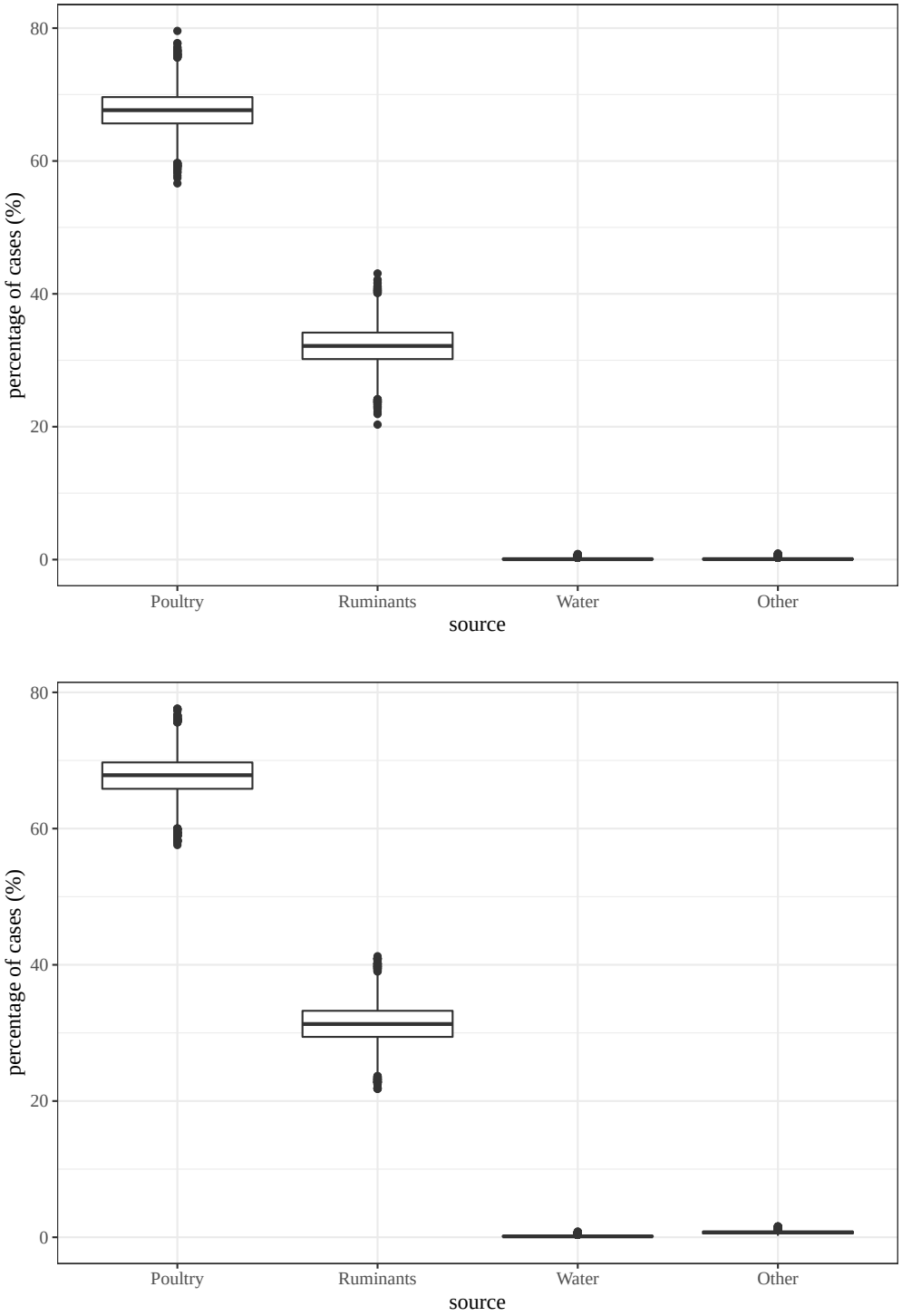


Figure 4.1: The percentage of human cases attributable to four sources given the baseline is other sources, and the attribution probability $\boldsymbol{F}$ has a prior of Dir(1) (upper panel) or $\boldsymbol{F}$ is modelled only with the intercept on the logit scale with a prior of N(0,1) (lower panel).

attributable to poultry is more than 65%, followed by approximately 31% of ruminant-associated cases, whereas water and other sources have little contribution with only 4% of cases in total. It reveals that the posterior attribution strongly favours poultry and ruminants, regardless of whichever prior is used in these two methods.

Attribution probability with variables

When convariates are incorporated to infer the posterior attribution probability, the results are based on the model, in which the rurality variable, or rurality and age variables are included. The linear model is exactly the model (4.9), however, the dimensions of matrices $\boldsymbol{\beta}$ and $\boldsymbol{C}$ changes to 2×3 and $H \times 2$, given $J = 4$ if only the rurality effect is considered.

The posterior attribution for each source across the rurality scales and the associated 80% credible interval with or without the consideration of age are showed in Figure 4.2. It suggests that more cases are on average attributed to poultry in urban areas, while more cases are associated with ruminants in rural areas. If the model only considers the rurality effect, the final attribution indicates that poultry related cases range from 33% in highly remote areas to 75% in main urban centres. By contrast, more cases are attributed to ruminants in rural areas, ranging from 63% in highly remote areas to 23% in main cities. Interestingly, fewer than 4% of human cases across all rurality levels are attributed to sources other than poultry and ruminants, with a slightly higher attribution to other sources in highly rural areas.

On the other hand, when the model also considers the age variable, the final attribution between two age groups (< 5 and ≥ 5) of human cases is distinguishable. The infection in children aged < 5 (82%) in rural areas is much more associated with ruminants than those aged 5 or above (62%), while poultry contributes less to children younger than 5 years old (13%) compared to those aged ≥ 5 (33%). The credible intervals in the age group of < 5 are relatively wider than the counterpart, resulting from a little higher uncertainty caused by the smaller number of young children aged < 5 years, which only accounts for 13% of the data set.

4.3.2 Convergence diagnostics

Attribution probability without variables

For the first modelling with the Dirichlet prior assumption on the attribution probability, the posterior samples for $\boldsymbol{F}$ for each source are displayed in Figure 4.3 (a). It shows that the chains mix well; the convergence is observed after re-running the sampler, for which similar results are showed in Appendix C.1 (Figure C.1).

Next, posterior samples for $\boldsymbol{F}$ based on the second method of modelling $\boldsymbol{F}$ linearly

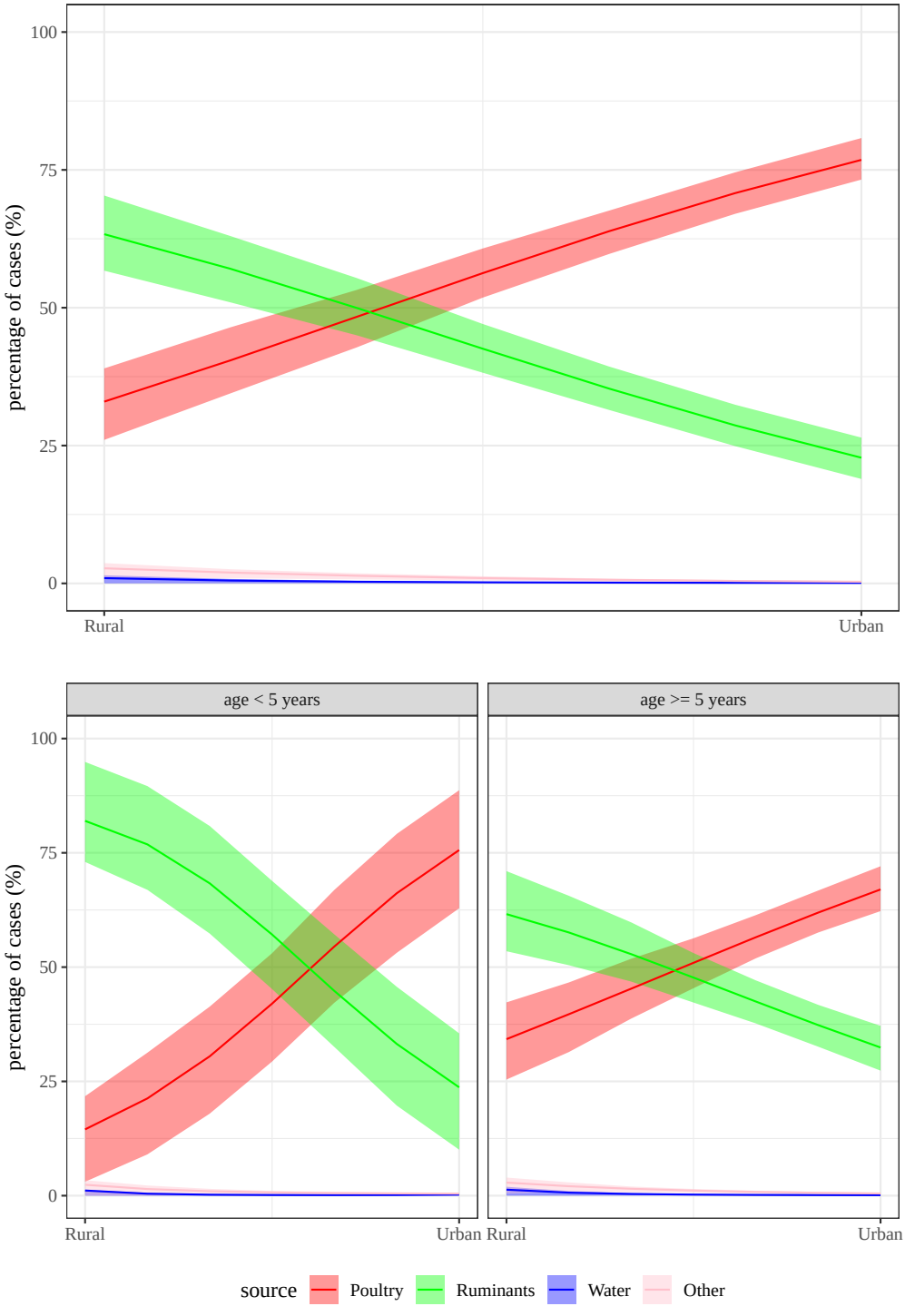
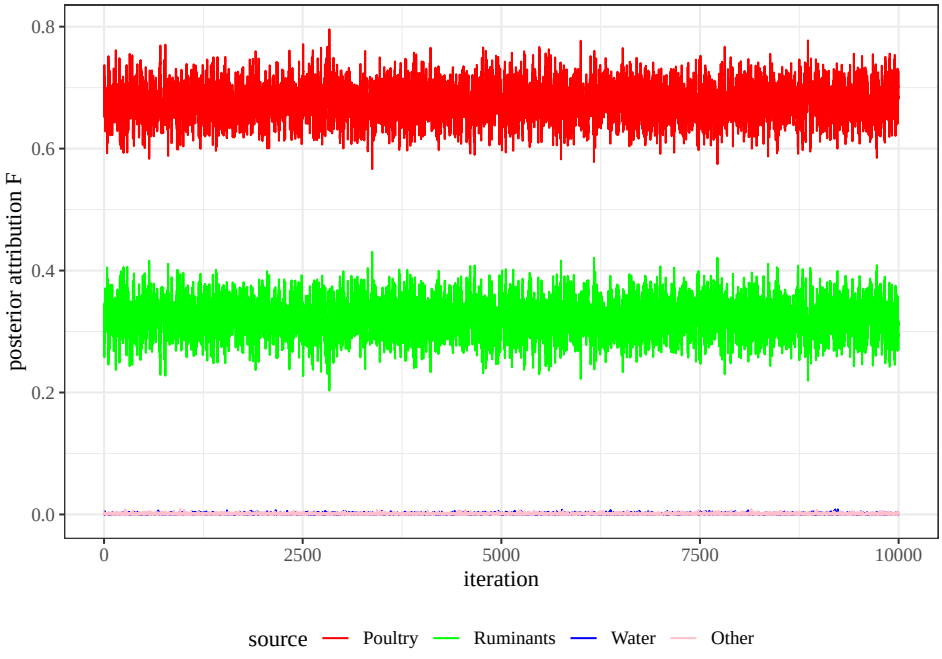
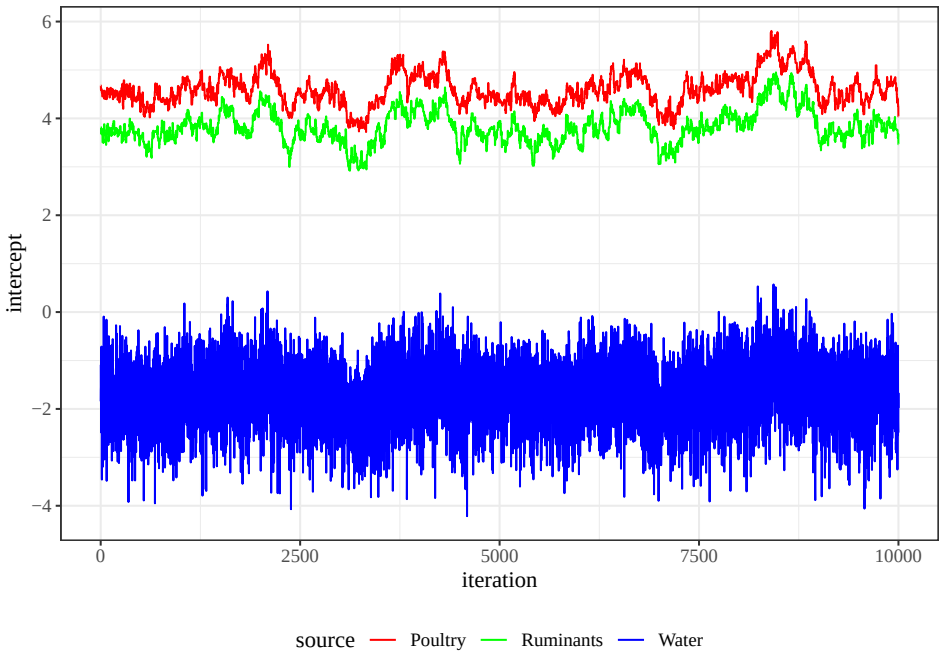


Figure 4.2: The percentage of human cases with 80% HPD credible intervals attributable to four sources, given other sources are the baseline, and only the rurality variable is considered (upper panel) or both rurality and age variables are included (lower panel) in the model.



(a)



(b)

Figure 4.3: The trace plot of (a) the Gibbs sampler for four sources given $\boldsymbol{F}$ was assigned a Dir($\mathbf{1}$) prior, and (b) the Metropolis-Hastings sampling for three sources given $\boldsymbol{F}$ was modelled by $\boldsymbol{f} = \boldsymbol{\beta}_0$ with other sources as the baseline.

on the logit scale without any variables are illustrated in Figure 4.3 (b). Given other sources are the baseline, the Metropolis-Hastings sampling algorithm gives relatively poor mixing for poultry and ruminants as the steps with the random walk sampler is very slow, and so there is a high degree of correlation between samples. In addition, the relationship between poultry and ruminants seems to be correlated as the direction of jumps between these two sources are almost the same. In contrast to these two sources, the chain for water seems to mix a little better as the jumps are around the centre of the movements with relatively constant variation, although it is not good enough.

Despite similar posterior attribution results, the trace plot of these two methods are different. The chain produced from the latter method does not mix well for all sources, reflecting a need for updating the parameter for different sources with different step sizes in the algorithm, which will be assessed in the next section.

Attribution probability with variables

For the model including the rurality variable, the chain of each fitted parameter is displayed in Figure 4.4. Again, the sampler only demonstrates good mixing for water. The mixing of the intercept and the slope of rurality for poultry and ruminants are not as acceptable as that of Figure 4.3 (a), and the autocorrelation remains high. The sampler did not explore enough the parameter space in different directions as the steps for these two sources do not jump randomly. It indicates that the successive iterations are highly correlated for each of these two sources in such a long run with a length of 51,000 samples.

Moreover, the possible correlation between poultry and ruminants is observed in the trace plot; the correlation between these two sources for both parameters are positively strong with the correlation coefficient higher than 0.9. This is confirmed in Appendix C.1 (Figure C.2). In addition to the correlation between sources within the parameter, Figure 4.5 shows that the relationship between the intercept and the slope of rurality for each source is uncorrelated; the samples of each source are clustered without any patterns.

Lastly, when the model also considers the age variable along with the interaction between rurality and age, a lack of overall good mixing is observed in Figure 4.6. A fairly good mixing for three sources is observed both for the slope of age and the interaction. Nonetheless, a poor mixing is noticed again for the intercept and the slope of rurality parameters, particularly for poultry and ruminants. The correlation between sources for each parameter is same as has been observed before. A strong and positive correlation between poultry and ruminants is confirmed in Appendix C.1 (Figures C.3 and C.4).

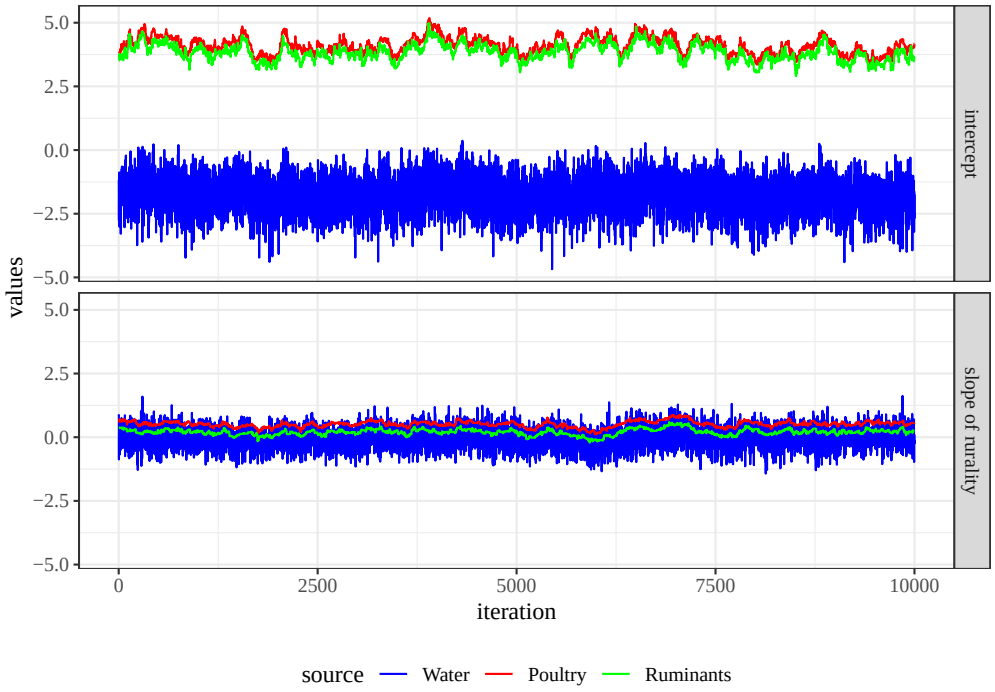


Figure 4.4: The trace plots for the intercept and the slope of rurality parameters, given the rurality variable is considered in the model with other sources as the baseline.

In conclusion, given the baseline is other sources, it seems that poultry and ruminants largely affect the mixing of chains when variables are included in the model.

4.4 Sensitivity analysis

To ensure the robustness of the resulting posterior inferences, the assessment of sensitivity is conducted and demonstrated when modelling with the rurality variable on the logit scale, with respect to the prior, source baseline and model specifications.

4.4.1 Prior specification

The Metropolis-Hastings sampling in Algorithm 2 has been applied to the approach of modelling with variables. Starting points for the proportion of genotypes on sources $\boldsymbol{\pi}$ were first sampled from the prior of Dir($\mathbf{1}$). To see if different starting points would influence the output of our algorithm, a seed is set to generate specific initial values of $\boldsymbol{\pi}$ in the beginning of a chain. It is anticipated that different starting points would not affect the chain when it is converged as the MCMC draws are no longer depending on the initial values. Indeed, this change has no significant impact on the final inferences after re-running the model, compared to the original results in Figures 4.2, 4.4 and 4.5

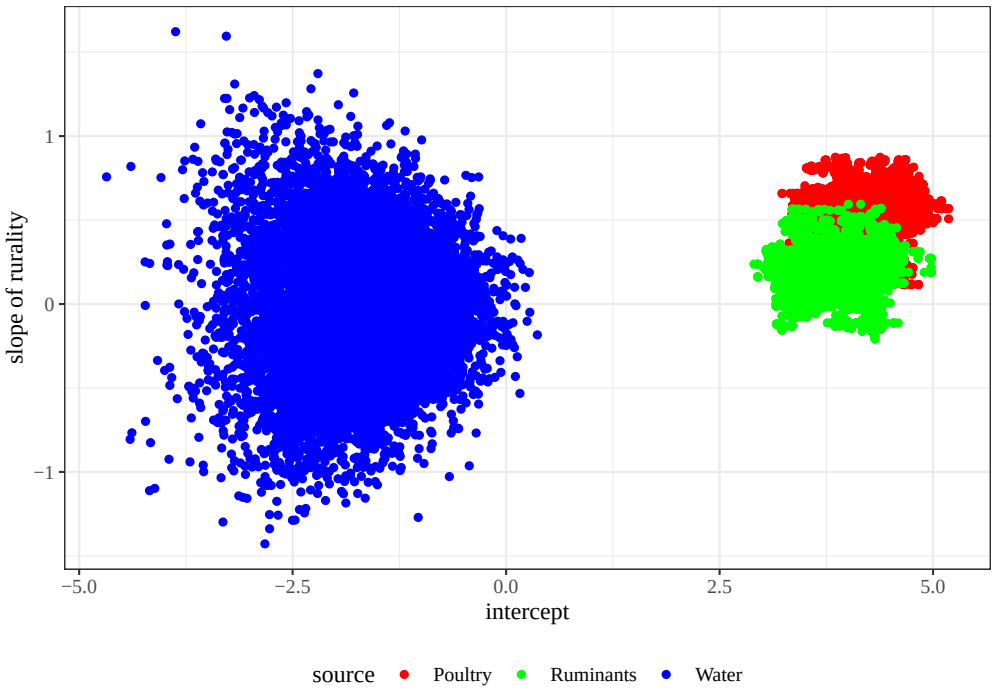


Figure 4.5: The scatter plot of two parameters, given the rurality effect is considered in the model, with other sources as the baseline.

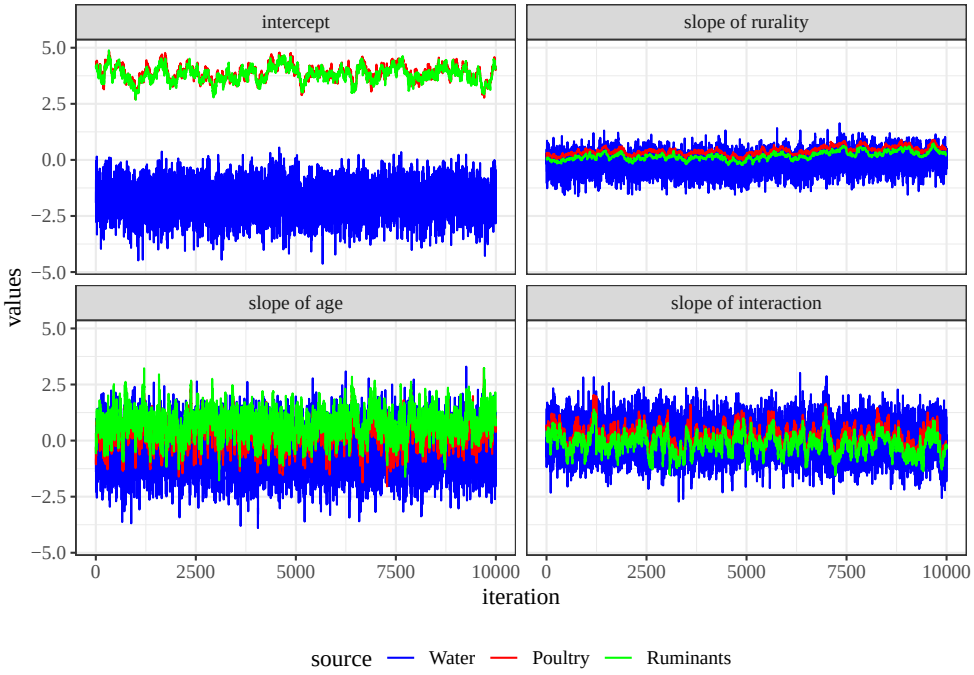


Figure 4.6: The trace plot for each parameter, given the rurality and age variables are considered in the model, with other sources as the baseline category, in which the age variable is regarded as binary with a threshold of 5 years old.

(see supporting figures displayed in Appendix C.1, Figures C.5, C.6 and C.7).

Another way to assess a prior sensitivity is to change the prior variance. A standard normal prior for regression coefficients was originally assumed. This time, the dispersion of the prior is tweaked from 1 to two options: $\sigma^2 = 0.025$ and $\sigma^2 = 4$. It is found that: i) the values of log-odds $\boldsymbol{f}$ tend to be closer to 0 as the prior variance becomes smaller; ii) altering the prior dispersion has no tangible effect on the posterior attribution, largely because the attribution is dominated by poultry and ruminants; and iii) the mixing of the chain for poultry and ruminants becomes better when the underlying prior variance becomes smaller. Figures to support these findings are displayed in Appendix C.1 from Figures C.8 to C.10.

4.4.2 Source baseline

As noted earlier, poultry and ruminants are highly correlated when other sources are the underlying baseline in the model. After re-running the sampler with the baseline changing from other sources to ruminants, it is not surprising that the final attribution is comparable to the original one (refer Figure C.11 in Appendix C.1 against Figure 4.2). Nevertheless, the prior specification is altered for the reasons discussed in Section 4.2.2 due to the change of source baseline. The trend of $\boldsymbol{f}$ for poultry and water changes towards lower values across the rurality levels, however, there is little disparity in posterior attribution resulted from the model with the baseline between other sources and ruminants. Moreover, the mixing of the intercept and slope of rurality for each source is greatly improved, and the problem of high correlation between poultry and ruminants is also eliminated. These can be confirmed with Figures C.12 and C.13 in Appendix C.1.

4.4.3 Model specification

Originally, both source and human data contribute towards estimation of the probability π_{ij} of observing genotype i on source j, with the assumption that the probability of observing genotypes on both human and source isolates is equal. If the assumption is true, the estimates should not be affected too much by the relative amount of data available from sources on the one hand, and humans on the other. To investigate if it is valid, the volume of source data is artificially increased 100 times in order for the human data to be dominated by the source data in terms of determining the estimates of π_{ij}.

After re-running the model, the posterior attribution displayed in Figure 4.7 is found to be sensitive to the rise in the amount of source information. In comparison with the original posterior attribution in Figure 4.2, the percentage of human cases living in rural areas attributable to ruminants becomes higher (approximately increases from

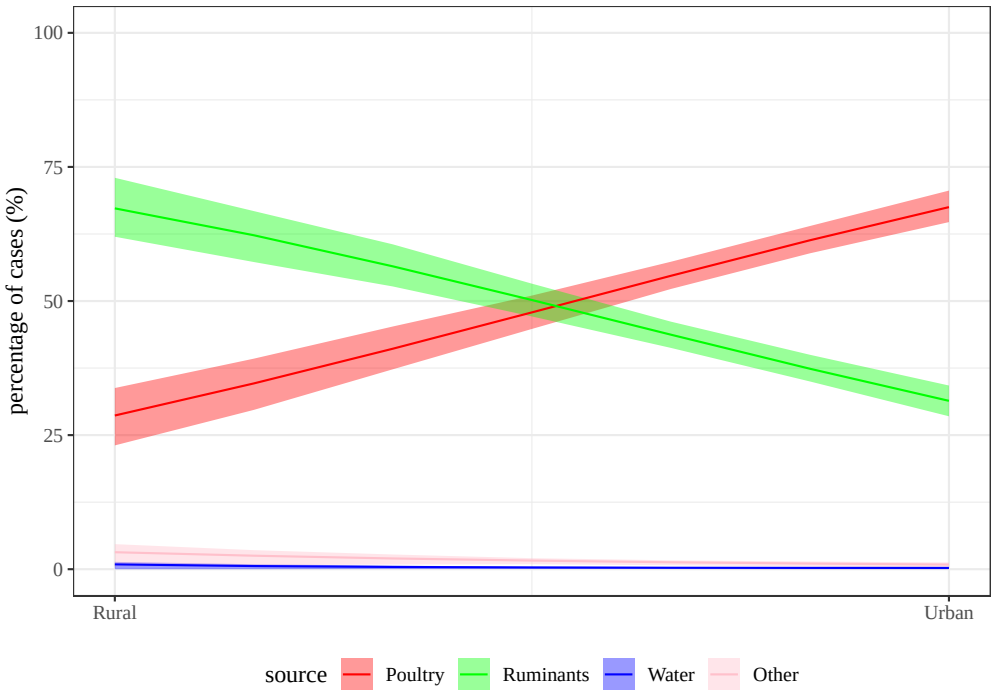


Figure 4.7: Increasing the size of observed source data affects the posterior attribution, given the model only includes the rurality variable with other sources as the baseline.

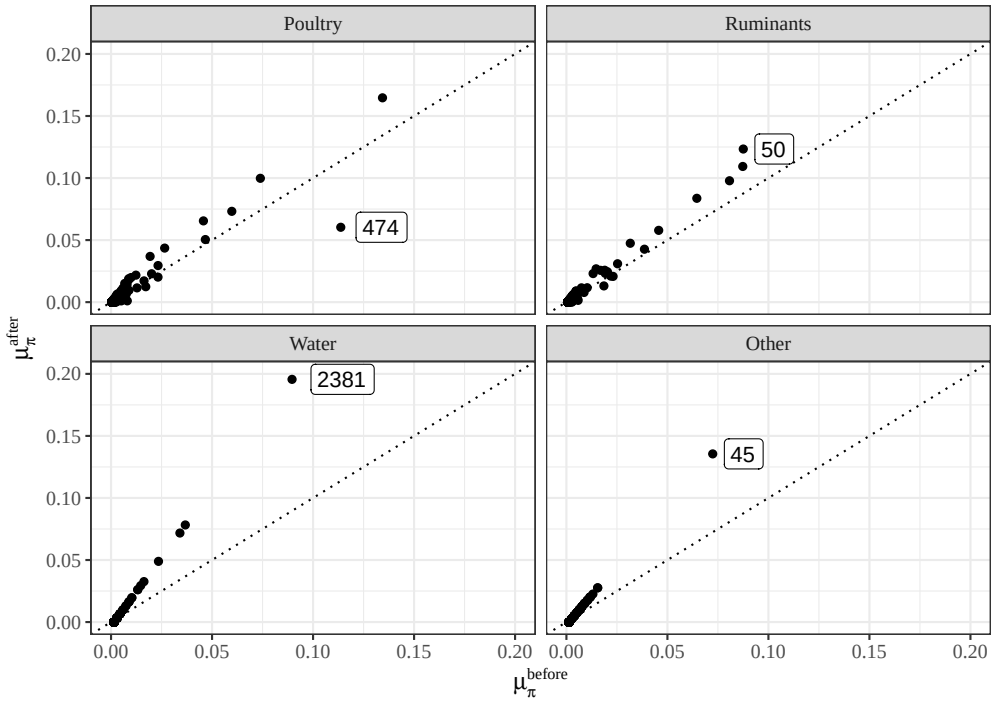


Figure 4.8: The comparison of the posterior mean of $\boldsymbol{\pi}$ before and after increasing source data 100 times, in which a genotype with the maximum difference is labelled for each source.

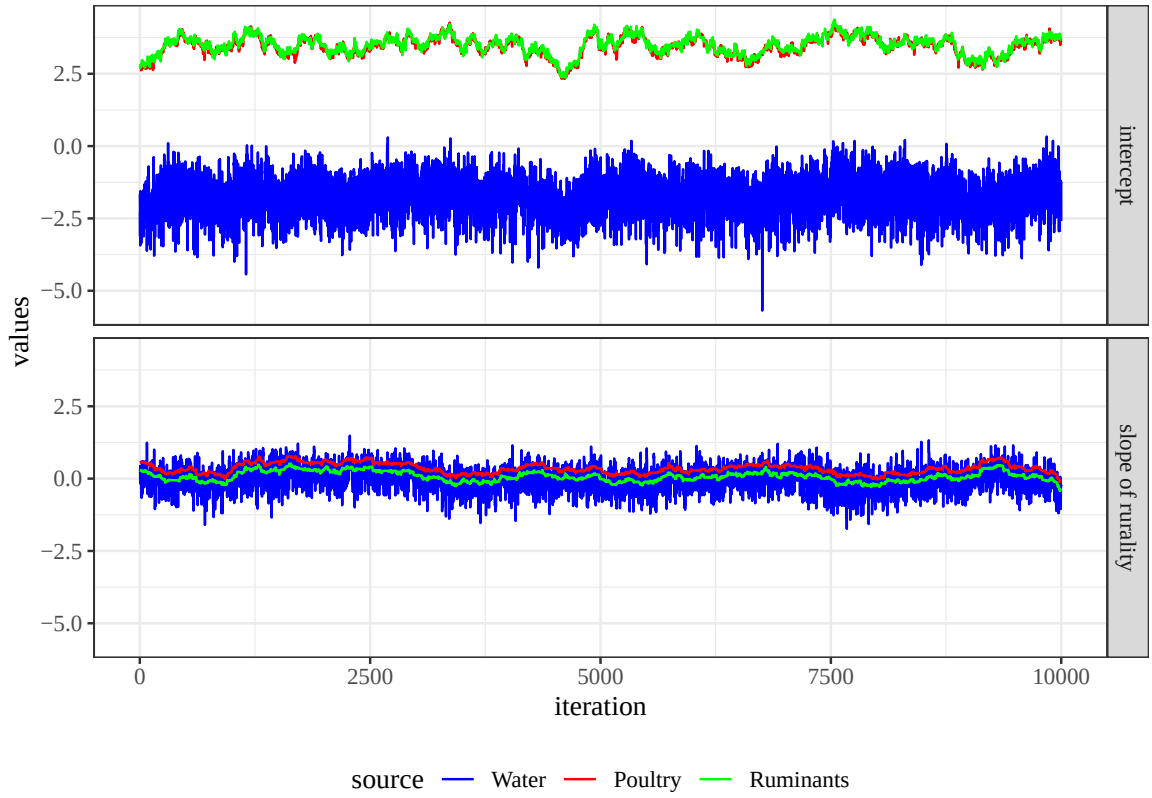


Figure 4.9: The trace plot for regression parameters considered in the model after increasing source data $\boldsymbol{X}$ 100 times.

63% to 69%), and meanwhile the contribution from poultry decreases. By contrast, the percentage of human cases attributable to poultry in urban areas decreases about 8%, leading to higher contribution from ruminants.

Further, the squared difference of the posterior mean of $\boldsymbol{\pi}$ before and after changing the size of source information is calculated. The comparison of the posterior mean of $\boldsymbol{\pi}$ for each genotype before and after the change is displayed in Figure 4.8, in which a genotype with the maximum difference is labelled. If the assumption of common $\boldsymbol{\pi}$ holds, the points should lie approximately on the straight line. When the difference becomes larger, the points move farther away from the line. It tells that the genotypes, for which the values of $\boldsymbol{\pi}$ change most, are among the most commonly observed, and hence the ones, for which the most information is provided. This implies that the assumption is not plausible, the probability of observing genotypes between sources and humans are different. Possible reasons will be further discussed in the next section.

For convergence diagnostics, there is no significant influence over the mixing of parameters and the relationship between sources. Figures 4.9 and 4.10 confirm that the chain still mixes as poorly as the original one, and poultry is still highly related

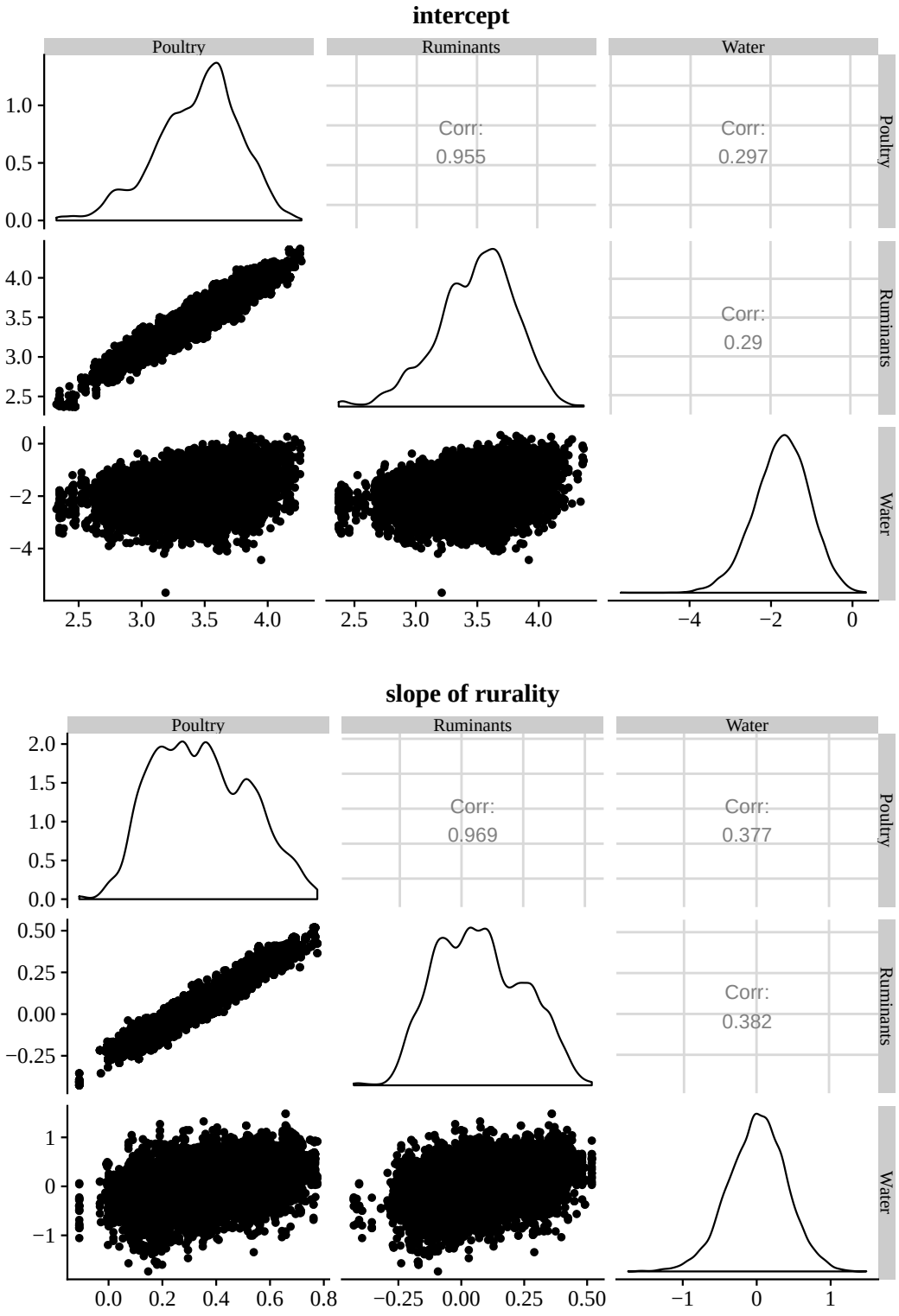


Figure 4.10: The matrix of scatter plots for the intercept and slope of rurality parameters, given the baseline is other sources after increasing source data $\boldsymbol{X}$ 100 times.

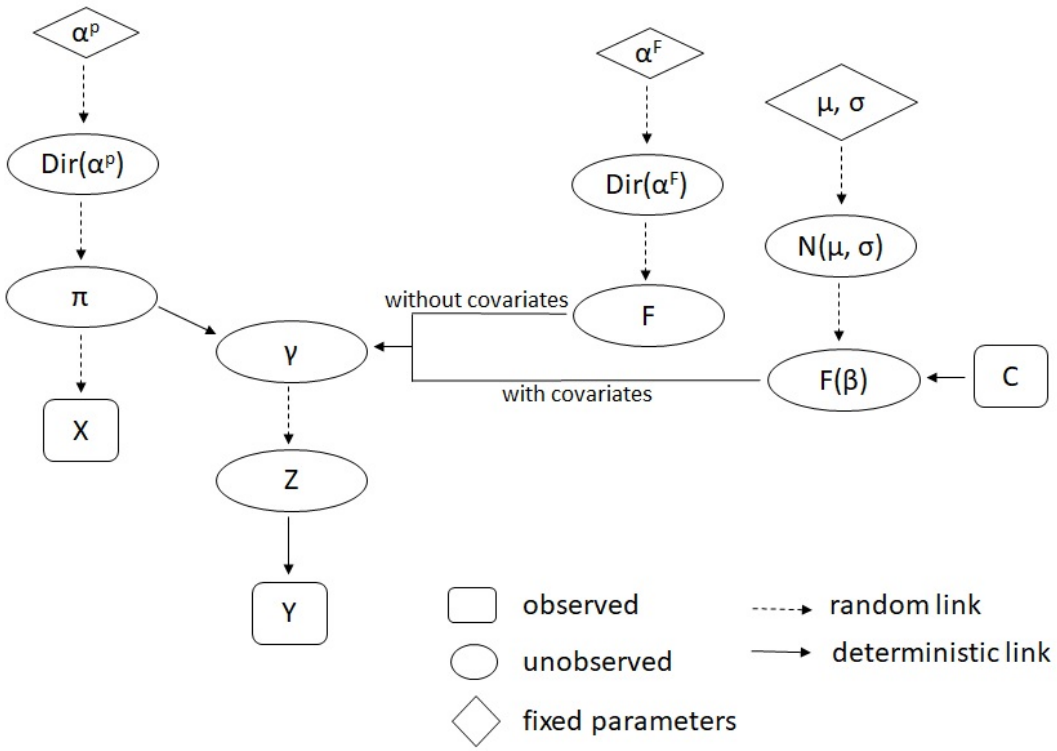


Figure 4.11: The current model framework includes the options of variables being included or not in the modelling for the attribution probability.

to ruminants. Therefore, changing the size of source information has an impact on the final attribution, but has no improvement in the mixing of chains and correlation between poultry and ruminants.

4.5 Model misspecification

From the previous sensitivity assessment, the prior specification and the choice of source baseline seem to have little influence on posterior attribution. Although the trace plots for each aspect of assessments have showed that some chains do not mix rapidly, neither are they mixing very poorly after the long run of the sampler with 51,000 iterations. In addition, the strong correlation between the dominant sources observed in the scatter plots can be circumvented by changing the baseline from other sources to either poultry or ruminants. In this regard, there is no major concern about convergence for the algorithms, but a critical issue about the models arise from changing the size of source data.

A direct acrylic graph (DAG) in Figure 4.11 illustrates the framework of the developed models, in which letters in upper case represent the data or variables, while Greek letters indicate the parameters used in the modelling. The genotype data from

Genotype	Human	Poultry	Ruminants	Water	Others
5	2	0	0	0	0
45	149	155	10	21	54
48	82	100	0	0	0
50	106	68	72	8	1
53	103	50	52	2	4
474	247	60	15	5	9
2381	0	0	0	60	3

Table 4.1: Of seven genotypes, five are very commonly isolated from humans, which are also frequently found on some sources. ST-5 is rare, while ST-2381 is largely found in water, but not in humans.

source isolates $\boldsymbol{X}$ provide information about the frequency of the typed genotype from each source group. Thus, for each source, the data are modelled using a multinomial distribution with unknown probabilities $\boldsymbol{\pi}_j$, to which a Dirichlet prior with a vector of fixed parameters $\boldsymbol{\alpha}^p$ is assigned.

Similarly, a multinomial distribution is also used to describe the genotype data from human isolates $\boldsymbol{Y}$. Ideally, there is a random link between $\boldsymbol{Y}$ and $\boldsymbol{\pi}$ under the assumption that genotypes are equally likely observed on source and human isolates, i.e. $\boldsymbol{X}$ and $\boldsymbol{Y}$ are linked through the common $\boldsymbol{\pi}$. Thus, the likelihood for $\boldsymbol{Y}$ not only depends on $\boldsymbol{\pi}$, but also the attribution probabilities $\boldsymbol{F}$. For computational convenience, the multinomial theorem is applied and a latent variable $\boldsymbol{Z}$ is introduced to calculate the likelihood. It follows a multinomial distribution with probabilities $\boldsymbol{\gamma}$, which are determined by parameters $\boldsymbol{\pi}$ and $\boldsymbol{F}$. Further, when no variables are included in the attribution model, $\boldsymbol{F}$ are assumed to be Dirichlet distributed with fixed parameters $\boldsymbol{\alpha}^F$. If variables C are considered, a normal prior with a fixed mean and variance is assumed for the regression parameters $\boldsymbol{\beta}$ included in the linear model on the logit scale.

The models look well-structured within a Bayesian paradigm in the DAG, however, the investigation of the increased source data shown in the previous section suggests that the assumption of equal $\boldsymbol{\pi}$ is flawed. Possible reasons can be that there might exist sample errors induced by the models, and that types are not as similar as expected. When data were described in Chapter 2.2, it showed that only 145 out of 377 STs are found in human cases. This means that the models, in which the types are assumed to be similar, are likely inappropriate. Additionally, each genotype seems not equally likely to infect humans and cause disease severe enough for the subject to seek medical attention. *C. jejuni* are ubiquitous. They are not only capable of colonising a wide range of hosts, but also can adapt to various hosts (Fitzgerald et al., 2001). It is also likely that some types are better at surviving in different food matrices and some may

be more virulent to humans. Table 4.1 summarises some STs from the data. It shows that genotypes often observed in human cases may be more common on some sources, and some genotypes could be common in certain sources but never identified in humans (e.g. ST-2381). Thus, the current model is mis-specified.

Consequently, the implausible assumption motivates a need for a modified model, in which the probability of typing genotypes from sources and humans are assumed to be different. *Campylobacter* is constantly evolving, the genetic changes lead to some genotypes being related to others closely or distantly (Fearnhead et al., 2005; Wilson et al., 2009). This highlights the importance of considering the molecular-based information when modelling source attribution. In other words, we will be in a position to use evolutionary models to describe the probability of typing genotypes from sources, and these modified models will be proposed and described in the next chapter.

Chapter 5

Approximate Bayesian models

The developed models in the previous chapter were inadequate due to the assumption of equal probability of genotypes observed between human and source isolates. In order for the problem of model mis-specification to be solved, reliance on the flawed assumption should be removed. Here the problem is addressed by introducing an approximate Bayesian model, which allows the probabilities $\boldsymbol{\pi}$ to differ between sources and humans.

To prevent the human data from directly influencing estimation of the probabilities $\boldsymbol{\pi}$ for sources, prior knowledge about the sampling distribution of types observed on sources is introduced. It can express the assumption that the proportion of genotypes typed on source isolates is not directly related to that on humans, but is affected by the underlying genetic evolution in the pathogen. Introducing such a complex genetic model in the model framework opens up the possibility of using a representation of genetic evolution to help model $\boldsymbol{\pi}$ for sources. Nonetheless, it brings a key question, whether more complex genetic models yield superior attribution results or whether a significantly simpler model may suffice, yet few publications in the literature have addressed this point. This becomes more important as model complexity extends to include epidemiological variables. Therefore, it is motivated to develop a simple model to estimate the prior of $\boldsymbol{\pi}$. This allows us to compare it against the genetic model, and the impact of considered evolutionary processes between isolates on the final attribution can then be investigated.

This chapter contains five sections. Section 5.1 defines the new models. It depicts the sampling distribution of genotypes derived from models with or without microbial genetic information in order to find the probability of genotypes arising from sources. Section 5.2 presents the procedure of model fitting with the consideration of the rurality effect, and outlines the steps of applied MCMC algorithms. Then, results are displayed in Section 5.3 including the posterior attribution probability for each source, the investigation between the genetic and simpler models, and the convergence of MCMC chains. Evaluation about model robustness is also demonstrated and presented in Section 5.4,

while discussions about the proposed models and findings are made in Section 5.5.

5.1 Model description

The goal of attribution models is to estimate the probability that the observed human cases arise from each putative source. Given the genotyping information, the first step of estimation is the sampling distribution of genotypes for each source and then the appropriate combinations of those genotype distributions that most likely give rise to the set of genotypes observed among human cases. Specifying the sampling distribution of genotypes found on sources is fundamental for the purpose of not only exploring how it affects the source attribution probability, but also investigating the disparity in attribution effect made between models with different types of variables.

5.1.1 Model with microbial genetic information

To start with, prior knowledge of the probabilities $\boldsymbol{\pi}$ is approximated by models with or without microbial genetic information. The genetic model used here is the linked evolutionary model (3.10) in the Wilson model introduced in Chapter 3.4. The term 'asymmetric Island model' is designated in order to distinguish it from the Wilson model. Originally, the model was adopted in the source attribution study for human campylobacteriosis (Marshall et al., 2016) using the allelic profile information for each genotype in order to estimate mutation and recombination probabilities within, and migration probabilities between, each source 'island'.

In this chapter, this genetic model also incorporates evolutionary processes between isolates so that the sampling distribution of genotypes among each putative source is built. This may range from using the proportion of each observed type (Hald et al., 2004) on each source through to approximating genetic distances to infer evolutionary processes within and between each source isolate. The model can estimate the likelihood of observing a genotype on a source when it has not been previously observed. Thus, the probabilities $\boldsymbol{\pi}$ are estimated indirectly, by first estimating the evolutionary parameters, and then deriving the sampling distributions.

An R package, *island*R, has been already developed according to the asymmetric Island model. The estimation of sampling distributions is thus obtained through the use of this package. Further information is available from the repository on GitHub (Marshall, 2019).

5.1.2 Model without microbial genetic information

For the reader's convenience, the notation used here follows that introduced in the previous chapter. For all isolates in the data, H isolates belong to humans, while the

remaining N isolates are categorised in J groups as the major sources attributed to the infection. There are I genotypes representing the total number of unique types detected from all isolates, and the marginal frequency of types n_j found in source j is subject to $\sum_j^J n_j = N$, where $j = 1, \ldots, J$. Typically, I, the number of detected types is smaller than the sample size of isolates as multiple isolates will be of the same type.

To discover the effect of incorporating genetic information at the allelic profile level as used in the asymmetric Island model, a simple model is developed for the genotype sampling distribution. With the assumption that the observed distribution of genotypes is representative of the true distribution, a multinomial distribution defined in Equation (4.1) is again used to model observed genotypes $\boldsymbol{X}$ arising from source j. The prior of the probabilities $\boldsymbol{\pi}_j$ is hence modelled using a Dirichlet distribution with parameters $\boldsymbol{\alpha}_j^p$ due to the conjugate pair for the multinomial model. Thus, similar to the posterior for $\boldsymbol{\pi}_j$ in Equation (4.5), but without the introduced latent variable $\boldsymbol{Z}$, it takes the same form which is expressed as,

$$p(\boldsymbol{\pi}_j \mid \boldsymbol{x}_j) \propto \prod_{i=1}^{I} \pi_{ij}^{\alpha_{ij}^p + x_{ij} - 1}, \quad \alpha_{ij}^p > 0. \tag{5.1}$$

To express the belief that every isolate is equally likely *a priori*, $\boldsymbol{\alpha}_j^p$ are again assumed to be $\mathbf{1}$. Therefore, the sampling distribution of genotypes can be approximated through the simulation of the posterior $\mathrm{Dir}(\boldsymbol{x}+\mathbf{1})$, which is the so-called simpler model (or the Dirichlet model) as it is relatively simple for estimation compared to the asymmetric Island model.

5.1.3 Likelihood of observing genotypes on humans

Each type i, $i = 1, \ldots, I$, may be found in more than one human case and the rurality variable is also considered in the modelling. The likelihood of observing genotypes in human cases is modelled with respect to individuals using a multinomial distribution as defined in Equation (4.8). That is,

$$\begin{aligned} L(\boldsymbol{Y}; \boldsymbol{\pi}, \boldsymbol{F}) &\propto \prod_{h=1}^{H} \sum_{j=1}^{J} p(\mathrm{ST}_{i[h]} \mid \mathrm{source}_j) p(\mathrm{source}_j \mid \mathrm{variable}_h) \\ &= \prod_{h=1}^{H} \sum_{j=1}^{J} \pi_{i[h]j} F_{hj}. \end{aligned}$$

Given that the sampling distributions $p(\mathrm{ST}_{i[h]} \mid \mathrm{source}_j)$ are known through the approximation methods described before, each sample of $p(\mathrm{source}_j \mid \mathrm{variable}_h)$ may be estimated by optimising the above likelihood, for example, using a Metropolis–Hastings algorithm within a Bayesian context, with suitable priors on $p(\mathrm{source}_j \mid \mathrm{variable}_h)$.

This has the effect of integrating over the uncertainty in F_{hj} while estimating it. Thus, the MCMC is run as a sub-chain for each sample from the sampling distribution of genotypes.

5.2 Model fitting and MCMC inference

In a change to the models proposed in the previous chapter, the source and human likelihoods are not multiplied together in order to prevent the human data from impacting on the probabilities $\boldsymbol{\pi}$ for sources. Figure 5.1 shows the model structure. In comparison with the models proposed in Chapter 4, the latent variable $\boldsymbol{Z}$ is removed. The probabilities $\boldsymbol{\pi}$ are estimated by the asymmetric Island or Dirichlet model, and are involved in the human likelihood function. Unlike the previous chapter, the posterior attribution probabilities are not sampled based on the joint likelihood for the source and human data, but optimised through the human likelihood. Hence, the models are no longer 'full Bayesian' but 'approximate Bayesian', and so the MCMC algorithm also differs from that applied before.

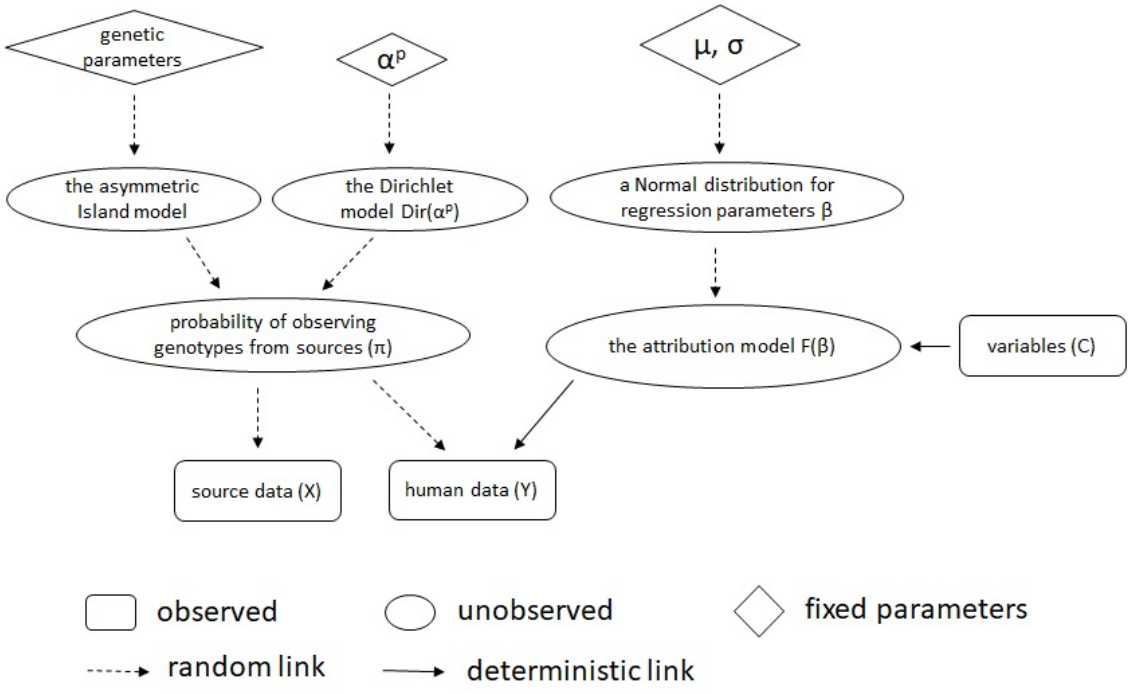


Figure 5.1: The new model framework uses the asymmetric Island model with consideration of genetic evolution, or the Dirichlet model without consideration of genetic evolution, to estimate the probabilities $\boldsymbol{\pi}$ before the inference about human attribution.

5.2.1 Attribution model with a variable

To extend the analysis so as to include the variables at an individual level, it is required to calculate subject-specific attribution conditional probabilities, $p(\text{source } j \mid \text{variable}_h)$. Similar to the definition in Equation (4.6), given $H = 1804$ and $J = 4$, the attribution probability F_{hj} for the h^{th} human case attributed to source j is specified as,

$$F_{hj} = \frac{\exp(f_{hj})}{1 + \sum_{l=1}^{3} \exp(f_{hl})}, \tag{5.2}$$

with constraints $\sum_{j=1}^{4} F_{hj} = 1$ and $0 \leq F_{hj} \leq 1$. The probability is again estimated using a linear combination on the logit scale, where $f_{h4} = 0$ is treated as the source baseline of f_{hj}.

Similar to the model (4.9), when taking the rurality variable c into account, the model of f_{hj} becomes,

$$f_{hj} = \beta_{0j} + \beta_{1j} c_h, \quad j = 1, 2, 3. \tag{5.3}$$

There are two ways to treat the variable c: one is to treat as numeric, and the other as categorical. If c is numeric, f_{hj} is exactly defined as the above model, in which c_h is the numeric rurality level of case h ranged from -3 to 3. Conversely, when c is categorical with seven levels of rurality scale, each of the degrees is regarded as an indicator with a superscript number d, $d = 1, \ldots, 7$, which corresponds with the position of the category. Thus, f_{hj} can be rewritten as,

$$f_{hj} = \beta_{1j} c_{1h} + \beta_{2j} c_{2h} + \ldots + \beta_{dj} c_{dh} + \ldots + \beta_{7j} c_{7h}, \tag{5.4}$$

where

$$c_{dh} = \begin{cases} 1 & \text{if case } h \text{ is in the category } d, \\ 0 & \text{otherwise.} \end{cases}$$

To differentiate the performance between the two types of variable, 'the linear model' and 'the categorical model' are encoded and linked to these two fitted models separately.

5.2.2 The Metropolis-Hastings algorithm

The MCMC method applied in this chapter is an Metropolis-Hastings algorithm with the steps outlined in Algorithm 3.

To begin with, the design matrix $\boldsymbol{C}$ is determined to be 1804×2 given only one variable (rurality) is considered in the attribution model. Initial values for the regression

Algorithm 3 Metropolis-Hastings sampling

Define the design matrix $\boldsymbol{C}_{1804\times 2}$
Draw initial values for regression parameters $\boldsymbol{\beta}^{(0)} \sim$N(0,1)
Calculate $\boldsymbol{f}^{(0)} = \boldsymbol{C}\boldsymbol{\beta}^{(0)}$ either through the linear model (5.3) or the categorical model (5.4)
Compute $F_{hj}^{(0)}$ for human case h and source j through Equation (5.2) given $f_{hj}^{(0)}$
Simulate S times the sampling distribution of genotypes $\boldsymbol{\pi}_j$ for source j, $j = 1, \ldots, 4$ referring the example of how to use the *island*R package ▷ If use the asymmetric Island model
or, draw S samples of $\boldsymbol{\pi}_j \sim \text{Dir}(\boldsymbol{X} + \boldsymbol{\alpha}^p)$ for source j, given $\alpha_i^p = 1$, $i = 1, \ldots, 377$ ▷ If use the Dirichlet model
for simulation $s = 1, \ldots, S$ of $\boldsymbol{\pi}^{(s)}$ **do**
    for iteration $m = 0, \ldots, M-1$ **do**
        Sample t from the permutation P_T of $\{1, \ldots, T\}$
        Propose a vector of parameters $\boldsymbol{\beta}^*$ with an element $\beta_{(t)}^* \sim Q\left(\beta_{(t)}^* \mid \beta_{(t)}^{(m)}\right)$
        Generate $u \sim U(0,1)$
        Calculate $\mathcal{A}\left(\beta_{(t)}^{(m)}, \beta_{(t)}^*\right) = \min\{1, \psi\}$, where

$$\psi = \frac{L\left(\boldsymbol{Y}; \boldsymbol{\pi}, F(\boldsymbol{\beta}^*)\right)}{L\left(\boldsymbol{Y}; \boldsymbol{\pi}, F(\boldsymbol{\beta}^{(m)})\right)} \frac{\pi\left(\beta_{(t)}^*\right)}{\pi\left(\beta_{(t)}^{(m)}\right)}$$

        if $u < \mathcal{A}$ **then**
            $\beta_{(t)}^{(m)} \leftarrow \beta_{(t)}^*$, implying $\boldsymbol{F}^{(m)} \leftarrow \boldsymbol{F}^*$
        else
            $\beta_{(t)}^{(m)} \leftarrow \beta_{(t)}^{(m-1)}$, implying $\boldsymbol{F}^{(m)} \leftarrow \boldsymbol{F}^{(m-1)}$
        end if
    end for
end for

parameters $\boldsymbol{\beta}^{(0)}$ are then sampled from a standard normal distribution. Note that the dimension of $\boldsymbol{\beta}$ varies upon the type of rurality variable. If it is numeric, $\boldsymbol{\beta}$ is a 2 by 3 matrix comprised of $\{\beta_{01}, \beta_{02}, \ldots, \beta_{13}\}$. In contrast, it becomes a 7×3 matrix with values $\{\beta_{11}, \beta_{12}, \ldots, \beta_{77}\}$ when the variable is categorical. Next, $\boldsymbol{f}^{(0)}$ are calculated, given $\boldsymbol{\beta}^{(0)}$ via Equation (5.3) or Equation (5.4), depending on the type of rurality, and so the associated $\boldsymbol{F}^{(0)}$ are computed through Equation (5.2). After simulating $S = 100$ genotype distributions using either the asymmetric Island or Dirichlet model, an MCMC chain is run for each sample of $\boldsymbol{\pi}$, and each chain is generated after running the MCMC sampler M times. An example of using the asymmetric Island model for the simulation of genotype distributions has displayed on the GitHub repository of the *island*R package (Marshall, 2019). For the Dirichlet model, the initial sampling distributions are simply simulated through $\text{Dir}(\boldsymbol{X} + \boldsymbol{1})$ for each source.

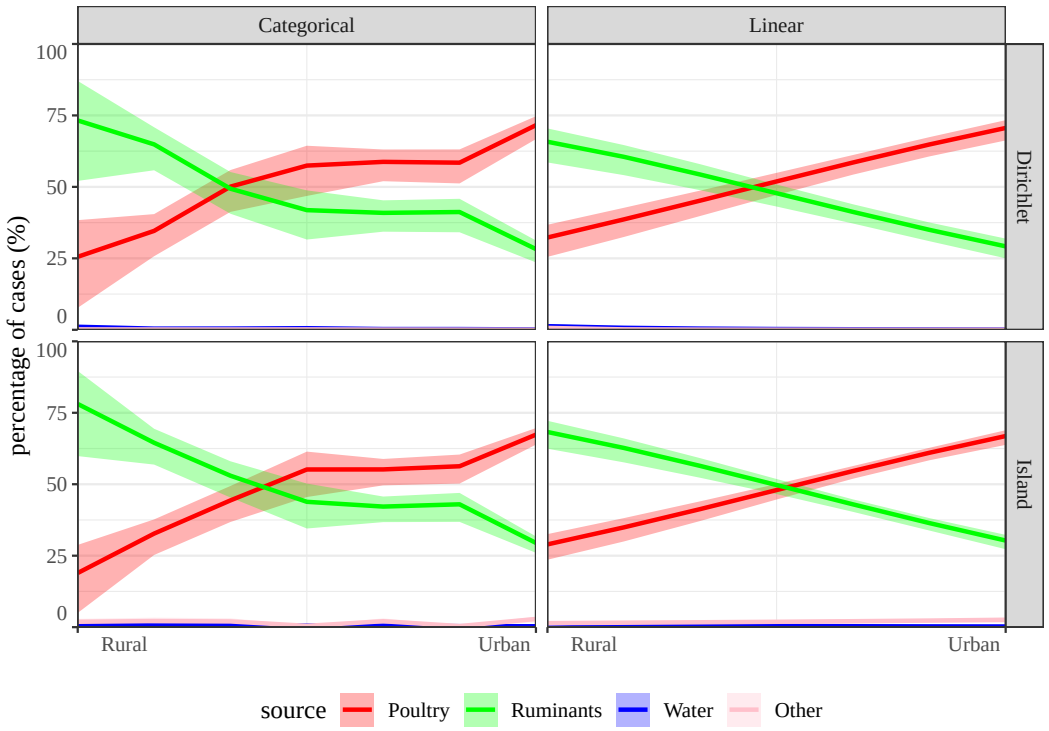


Figure 5.2: The percentage of human cases with 80% credible intervals for poultry, ruminants, water and other sources over the rurality scales. The attribution is generated from both the linear and the categorical models, with the underlying estimated sampling distribution of genotypes with evolutionary information (the asymmetric Island model) or without any genetic information (the Dirichlet model).

In each iteration m, $m = 0, \ldots, M-1$, each element of $\boldsymbol{\beta} = \{\beta_{(t)}\}_{t=1}^{T}$ is updated at a time. An element $\beta^*_{(t)}$ is sampled from a proposal distribution Q, which is again, a normal density with $(\mu_\beta, \sigma_\beta) = (\beta_{(t)}, 1)$. To decide if the move of $\beta_{(t)}$ (from the current value to the proposed one) is accepted, a random sample u is drawn from $\mathrm{U}(0,1)$ for evaluation. If the acceptance probability $\mathcal{A}\left(\beta^{(m)}_{(t)}, \beta^*_{(t)}\right) = \min\{1, \psi\}$ is smaller than u, the proposal is accepted and used to the next state of the chain. The value of ψ results from the product of likelihood and prior ratios (the proposal ratio equals to 1, and so it is cancelled out). Similar to the calculation before, ψ is on the logarithm scale in the programming. Hence, the log-prior ratio is similar to what has been displayed previously, which is of the form,

$$\log \frac{\pi\left(\beta^*_{(t)}\right)}{\pi\left(\beta^{(m)}_{(t)}\right)} = \frac{{\beta^{(m)}_{(t)}}^2 - {\beta^*_{(t)}}^2}{2},$$

Fitted models	Genotype models	
	Dirichlet	Island
Linear	13,464.4	14,491.3
Categorical	13,471.1	14,499.7

Table 5.1: The DIC values for the linear and categorical model applied to the data from 2005 to 2016, given the sampling distribution of genotypes is derived from either the asymmetric Island or Dirichlet model.

and the log-likelihood ratio turns out to be,

$$\log \frac{L\Big(\boldsymbol{Y}; \boldsymbol{\pi}, F(\boldsymbol{\beta}^*)\Big)}{L\Big(\boldsymbol{Y}; \boldsymbol{\pi}, F(\boldsymbol{\beta}^{(m)})\Big)} = \sum_{h=1}^{1804} \Big(\log \sum_{j=1}^{4} \pi_{i[h]j} F^*_{hj} - \log \sum_{j=1}^{4} \pi_{i[h]j} F^{(m)}_{hj} \Big),$$

where F_{hj} is derived from Equation (5.2) given $f_{hj}(\boldsymbol{\beta})$ is defined as a linear model (5.3) or a categorical model (5.4). Finally, the sampler keeps updating each t of the permutation P_T until M iterations are reached.

5.3 Results

5.3.1 Posterior attribution probability

The final attribution of human cases of campylobacteriosis ($\boldsymbol{F}$) with 80% credible intervals for each source is displayed in Figure 5.2, given the rurality effect is considered. The graphs are categorised by the types of model (asymmetric Island or Dirichlet) estimating the sampling distribution of genotypes on sources and the manner in which the rurality variable is modelled (categorical or linear on the logit scale).

Overall, the attribution pattern is similar to what has been observed in the previous chapter. The results are relatively stable irrespective of the types of model or how rurality is represented in the attribution model. The majority of human cases are attributed to ruminants and poultry, with more cases attributed to ruminants in rural areas and more cases attributed to poultry in urban areas. For the Dirichlet and asymmetric Island models, the linear and categorical models of rurality show broadly the same trend, suggesting that the additional flexibility given by the categorical model is not required, and that the shift in attribution as the level of rurality changes is adequately modelled by a linear trend on the logit scale. The linear model has the advantage of tighter credible intervals as it can share data across the seven levels of rurality, resulting in a clearer separation of ruminant and poultry attribution, particularly in highly rural areas where the data are sparse.

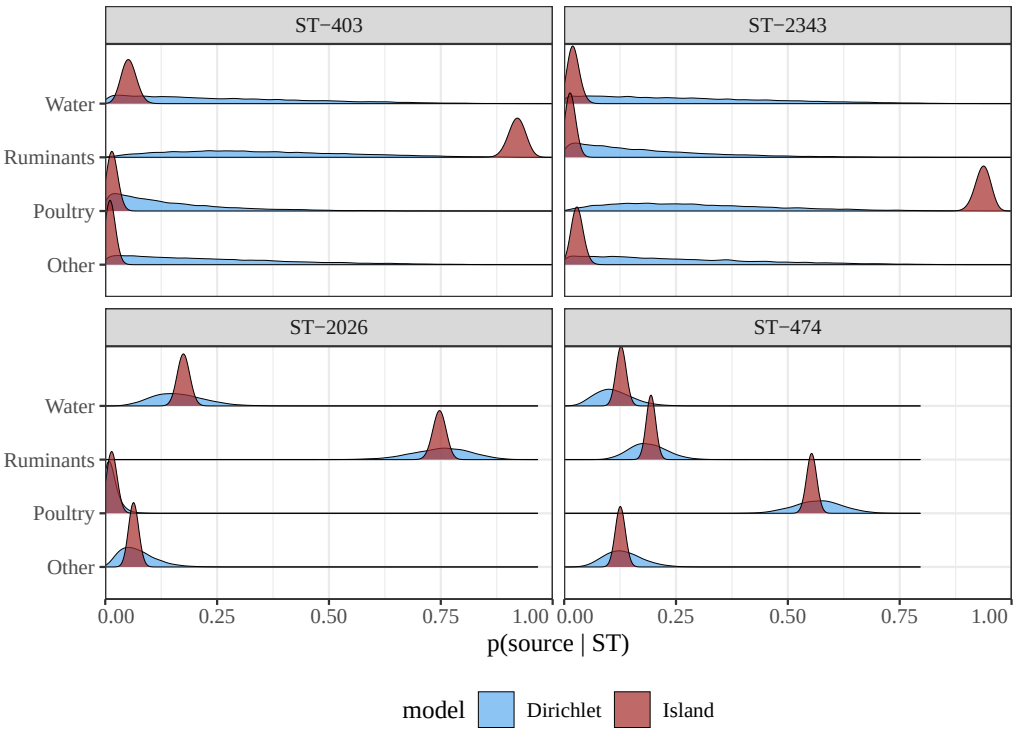


Figure 5.3: Posterior probability for each source for four sequence types from the asymmetric Island and Dirichlet models, assuming that each source is *a priori* equally likely.

There are some small differences between the genotype models, with the Dirichlet model showing a greater attribution to poultry (ranging from 40% in highly rural areas to 75% in main urban centres) than the asymmetric Island model (ranging from 30% in rural areas to 65% in urban centres). This also occurs similarly in the categorical model. Interestingly, the asymmetric Island model attributes approximately 7% of human cases across all rurality levels to sources other than poultry, ruminants and water, and gives a small attribution to water in highly rural areas, whilst the Dirichlet model indicates that both these sources are unimportant.

5.3.2 Model selection

For model comparison, the DIC were calculated with the outcome displayed in Table 5.1 after using Equation (3.7) based on the likelihood (4.8), given $\boldsymbol{F}$ are estimated by posterior samples of $\boldsymbol{\beta}$. Overall, there is a clear signal that a linear representation of rurality (on the logit scale) is adequate due to relatively small values compared to the categorical model. Note that the asymmetric Island and Dirichlet models are not directly comparable by the DIC as the likelihoods are on different scales. That is, for each source the Dirichlet model assumes all potential sequence types have been observed

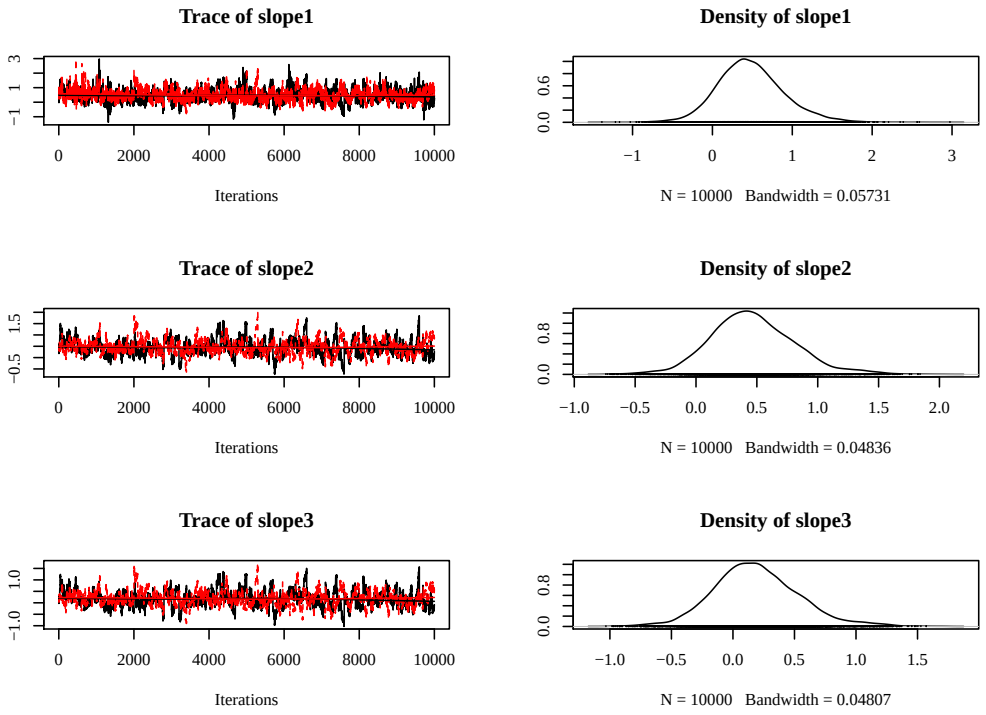


Figure 5.4: The trace plot (left panel) and the density plot (right panel) for the slope of rurality in the linear model, given the sampling distribution of genotypes is estimated by the asymmetric Island model. Each row corresponds to the parameter with the number 1, 2 and 3 representing the source, other, poultry and ruminants respectively.

so that $\sum_{i=1}^{377} \pi_{ij} = 1$, whereas the asymmetric Island model allows for unobserved sequence types so that $\sum_{i=1}^{377} \pi_{ij} < 1$.

5.3.3 Comparison of genotype models

The small differences in attribution observed between the asymmetric Island and Dirichlet models may be due to the additional genetic information available to the first model. To investigate this, it is of interest to know the probability $p(\text{source}_j \mid \text{ST}_i)$ that source j is of type i, assuming *a priori* that each source was equally likely. As described before, $\boldsymbol{\pi}_j$ for source j are estimated after simulating the sampling distribution of genotypes using the asymmetric Island or Dirichlet model. This time, $\boldsymbol{\pi}$ are generated 10,000 times. For each iteration, the conditional probability $p(\text{source}_j \mid \text{ST}_i)$ was calculated through $\frac{\pi_{ij}}{\sum_{j=1}^{4} \pi_{ij}}$ such that the summation across sources gives 1 for each type.

The conditional probabilities given a selection of four genotypes are illustrated in Figure 5.3. Genotypes ST-403 and ST-2343 are observed primarily in humans (7 and 6 cases, respectively). Each of them is only observed once either in ruminants or poultry so that the Dirichlet model has little information available to distinguish between

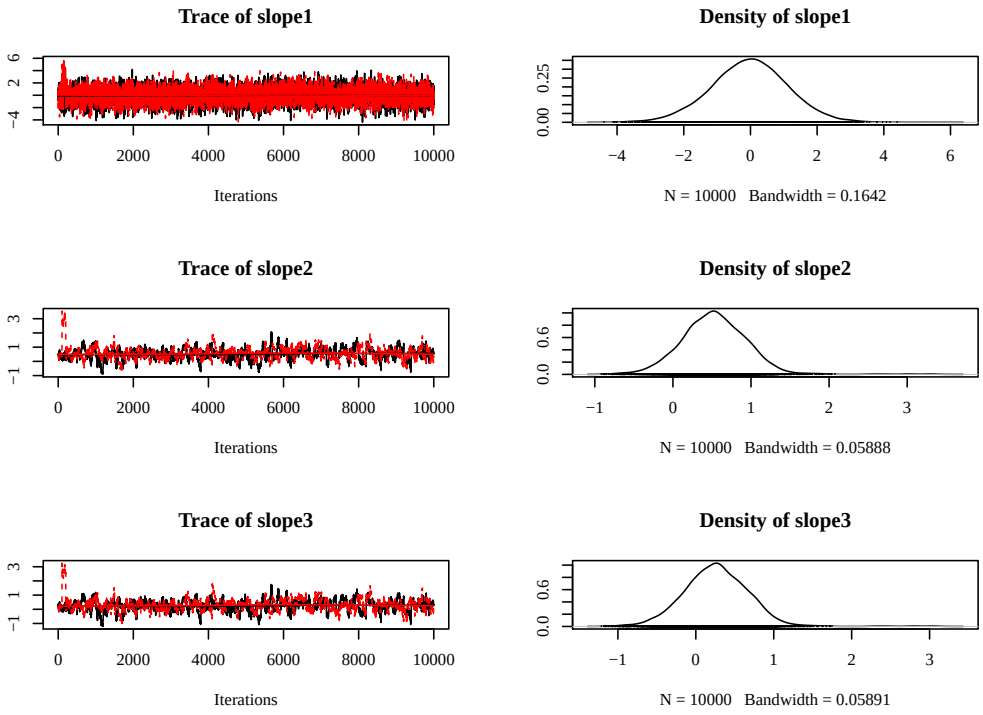


Figure 5.5: The trace plot (left panel) and the density plot (right panel) for the slope of rurality in the linear model, given the genotype distribution is estimated by the Dirichlet model. Each row corresponds to the parameter with the number 1, 2 and 3 representing the source, other, poultry and ruminants, respectively.

sources. The asymmetric Island model, however, can exploit the genetic relationship between genotypes and hence gives narrower probability distributions than the Dirichlet model. As shown in Table 2.1, ST-403 differs at just one locus from ST-2026, which is a type observed frequently in human cases and ruminant isolates. Whereas, ST-2343 differs at two loci from ST-474, observed commonly in human cases and poultry isolates. Thus, the asymmetric Island model can clearly assign ST-403 to ruminants and ST-2343 to poultry, whilst the Dirichlet model cannot distinguish between sources. In contrast, both models provide similar probabilities for ST-2026 and ST-474, which are both observed frequently.

5.3.4 Convergence diagnostics

The convergence assessment in this section will only focus on the linear model on the logit scale since it is suggested to be more adequate than the categorical model as pointed out in Table 5.1.

Figure 5.4 shows the trace and density plots for the slope of rurality in the linear model, given the genetic evolution between isolates is considered. It suggests that the

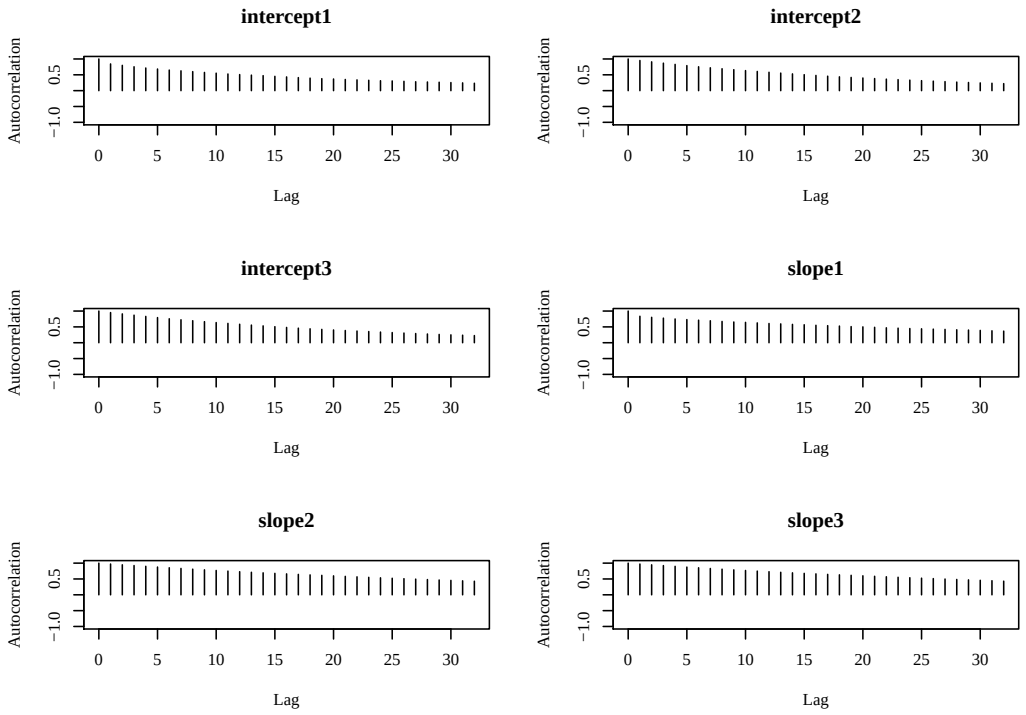


Figure 5.6: The autocorrelation of samples from the first chain for the regression parameters included in the linear model, given the genotype model is the asymmetric Island model. Each graph corresponds to the number 1, 2 and 3 representing the source, other, poultry and ruminants respectively.

sampler explores the parameter space quite efficiently. The variation of moves for each source between two chains (one in black, the other in red) in the trace plot is about constant and centred at the average line. This is also confirmed by the density plot that each posterior distribution is adequately normal distributed. On the other hand, Figure 5.5 presents the trace and density plots based on the model without genotypic information. It has comparable results as the asymmetrc Island model in that the two chains for each source mix well, especially for other sources. The first few samples with higher values in the second chain might require a little more burn-in, however, the variability of moves is approximately constant. It indicates that the sampler may converge faster as the dependence decays in successive iterations.

Further, autocorrelation between samples for these two genotype models are displayed in Figures 5.6, and 5.7. The autocorrelation in these two figures is bound to drop off as the lag increases, suggesting that samples could be considered to be independent in the chain.

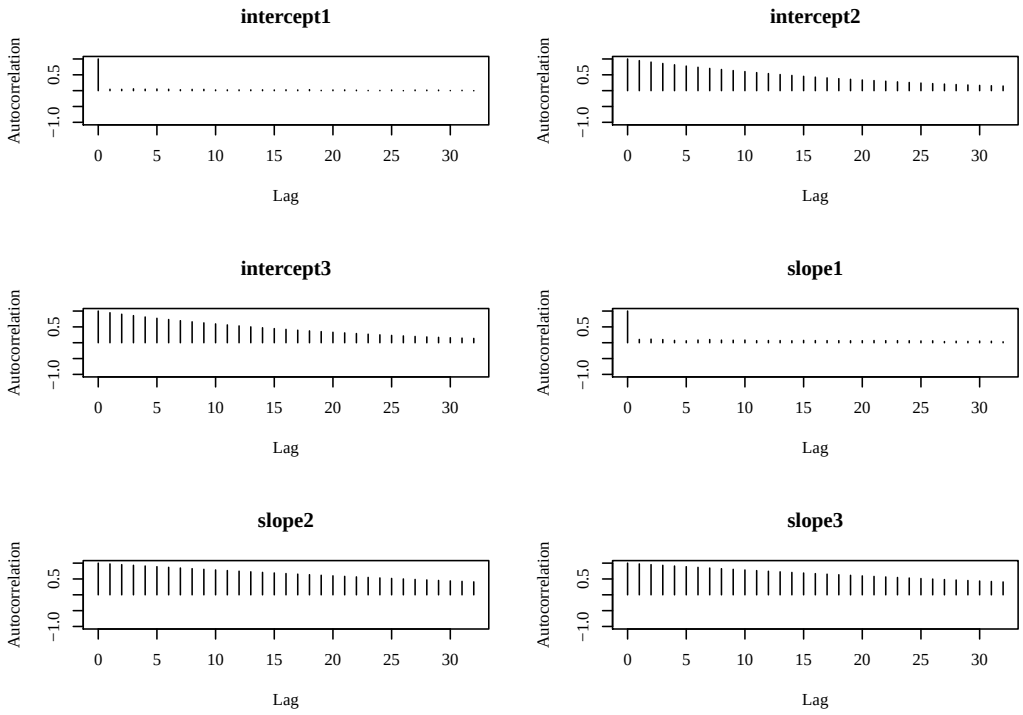


Figure 5.7: The autocorrelation of samples from the first chain for the regression parameters considered in the linear model, given the genotype model is the Dirichlet model. Each graph corresponds to the number 1, 2 and 3 specifying the source, other, poultry and ruminants respectively.

5.4 Sensitivity analysis

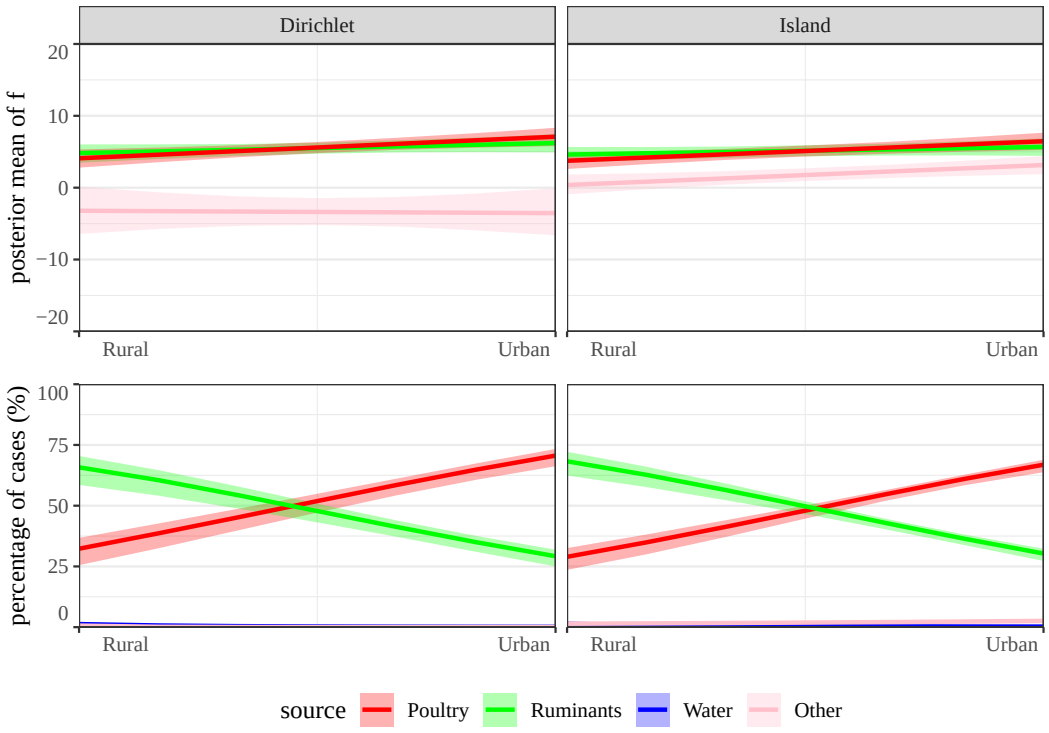
To ensure that the final inference about source attribution is robust to any changed settings, there are three aspects for examination: data with two different time periods, choices of prior distribution for the regression parameters included in the attribution model, and scales of the Dirichlet prior parameters $\boldsymbol{\alpha}^p$.

As mentioned in Chapter 2.2, a major public health initiative in 2007 led to a significant reduction in the number of cases of campylobacteriosis in New Zealand. In order to examine the effects of this change on the attribution probabilities, the analysis is repeated by including an interaction, with two time periods, 2005–2007 and 2008–2016. The general trend in attribution with the effect of rurality for each of the time periods is presented in Figure 5.8, given the linear trend is on the logit scale. There is a clear difference, with a significantly lower attribution to poultry (and correspondingly higher attribution to ruminants) in all but the most rural of areas, being strongest in highly urban areas. Thus, although the intervention did not eliminate infection arising from poultry (Muellner et al., 2011), the reduction highlights a significant improvement

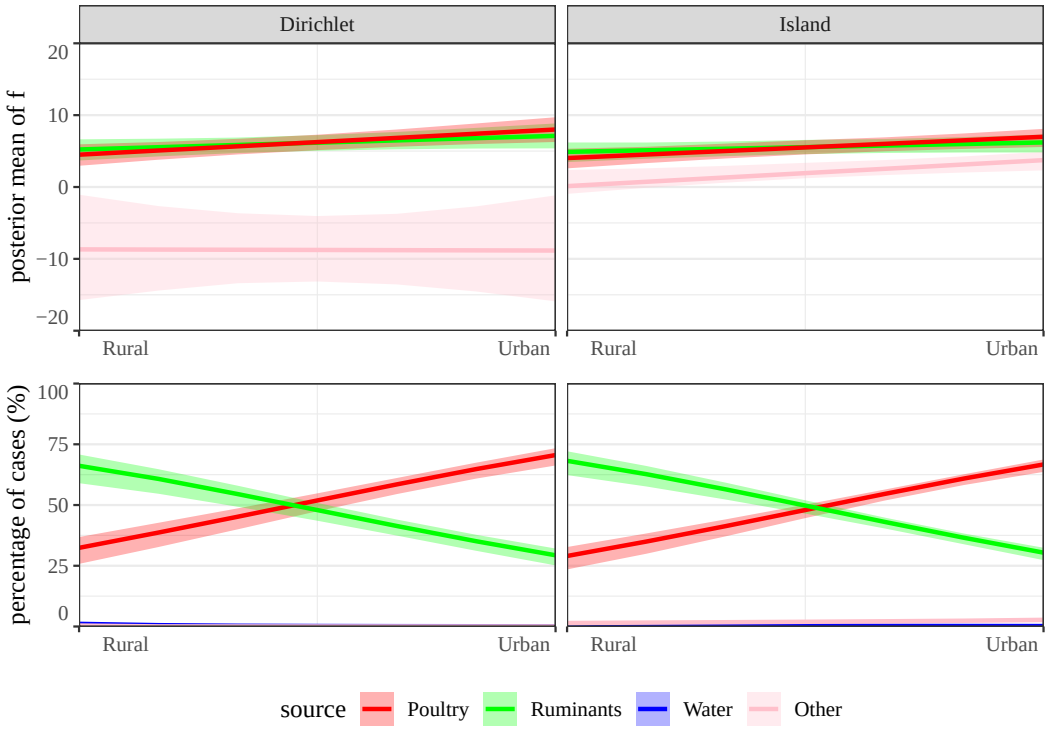
Figure 5.8: Posterior attribution ($\boldsymbol{F}$) of human cases during 2005–2007 and 2008–2016 with 80% credible intervals for poultry, ruminants, water, and other sources over the rurality scales. The attribution is generated using the linear model, given the genotype distribution is estimated with evolutionary information (the asymmetric Island model) or without any genetic information (the Dirichlet model).

in contribution of poultry to disease, particularly in urban areas where most cases were located.

As with any Bayesian analysis, it is of interest to examine the sensitivity of the results to the choice of prior distributions. Originally, standard normal priors for regression coefficients on the logit scale were used. This time, the variation of normal priors changes from $\sigma^2 = 1$ to $\sigma^2 = 64$; the resulting posterior mean of $\boldsymbol{f}$ and the associated attribution $\boldsymbol{F}$ are displayed in Figure 5.9. It suggests that $\boldsymbol{f}$ and the corresponding attribution probabilities $\boldsymbol{F}$ do not significantly change, even though $\boldsymbol{f}_{\mathsf{other}}$ marginally drifts, with wider credible interval when the variation of normal prior becomes larger. This is largely because the attribution is dominated by poultry and ruminant sources, with the water source in particular being close to zero. Thus, for the asymmetric Island and Dirichlet models, $\boldsymbol{f}_{\mathsf{poultry}}$ and $\boldsymbol{f}_{\mathsf{ruminants}}$ remain positive, while $\boldsymbol{f}_{\mathsf{other}}$ is also positive for the first model, but being negative in the latter model. However, the magnitude of these can increase without making much difference to their corresponding $\boldsymbol{F}$'s. The prior on $\boldsymbol{f}$ thus tends to restrict this ill-behaviour rather than acting as a



(a) σ^2=1



(b) σ^2=64

Figure 5.9: The posterior mean of $\boldsymbol{f}$ and the associated attribution $\boldsymbol{F}$ with 80% credible interval for each source across each level of rurality, given the variation of normal priors of $\boldsymbol{\beta}$ is (a) $\sigma^2 = 1$, or (b) $\sigma^2 = 64$.

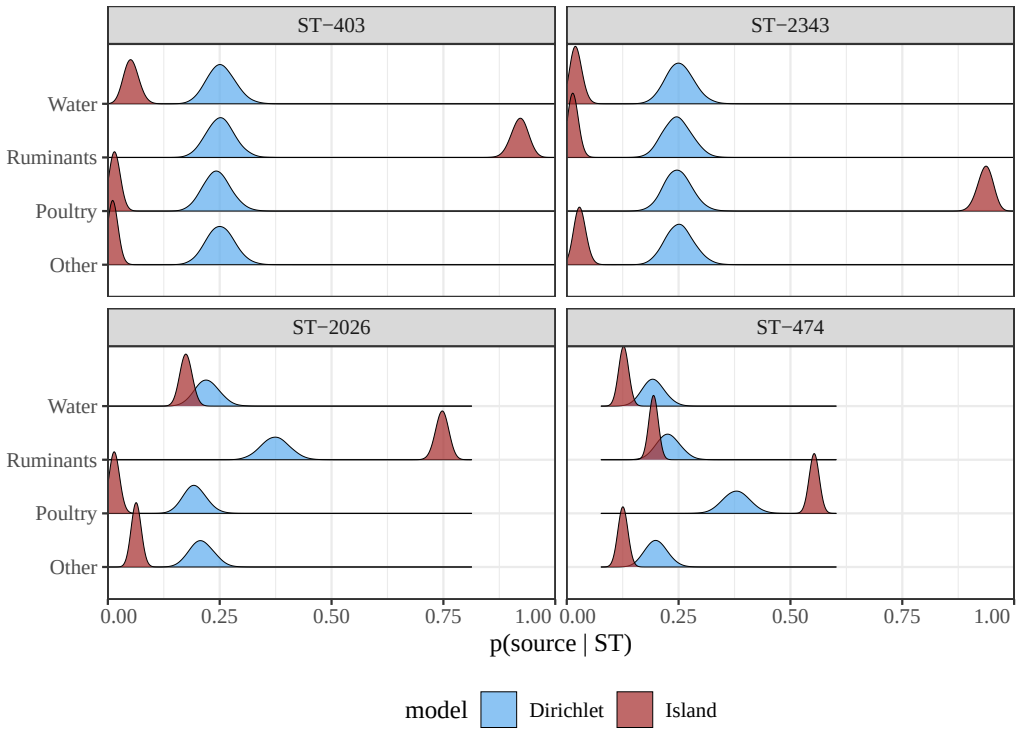


Figure 5.10: Posterior probability for each source for four sequence types after changing the scale of prior $\boldsymbol{\alpha}^p$ considered in the Dirichlet model from $\mathbf{1}$ to $\mathbf{50}$.

strong constraint on the attribution probabilities.

Moreover, the priors $\boldsymbol{\alpha}^p$ in the Dirichlet model (5.1) also makes little difference if it keeps small as it most strongly effects genotypes that are rare, which do not contribute significantly to the overall attribution. The prior can be thought of as data augmentation such that $\boldsymbol{\alpha}^p = \mathbf{1}$ is equivalent to adding a single observation of each genotype to source j. Thus, large values of $\boldsymbol{\alpha}^p$ will cause the genotype distributions across sources to look more similar, which is confirmed by Figure 5.10. The sampler has been re-run with $\boldsymbol{\alpha}^p = \mathbf{50}$, resulting in equal attribution to each source, particularly for those genotypes being rarely observed (ST-403 and ST-2343). For genotypes ST-2026 and ST-474 that are commonly found in ruminants and poultry, respectively, the distribution from the Dirichlet model moves slightly towards the asymmetric Island model due to greater observations on such sources.

5.5 Discussion

Models that determine the source of human infection, particularly for zoonotic pathogens that originate in animal populations, are of considerable value to public health policy

makers. However, such models may be complex, particularly when utilising evolutionary models. An outstanding question is whether such complexity is required, or whether a simpler model may work as effectively. Here a model with a focus on genotype frequency, which is less complex than evolutionary models, is developed to estimate the attribution probability for each source of *Campylobacter* infection. This model differs from the asymmetric Island model, in that it does not model evolution, opting instead to infer the sampling distribution of genotypes directly from the observed count data.

The results show that the Dirichlet and the asymmetric Island models give largely similar final attribution probabilities, with both models demonstrating a clear effect of rurality on attribution: cases in rural areas are more likely to have originated from ruminants, while those in main urban centres are more likely to be of poultry origin. As most people (>80%, refer Table 2.3) in the Manawatu region live in urban centres, this highlights the importance of poultry as a reservoir for campylobacteriosis, which is well established in the literature (Mullner et al., 2009b, 2010; Levesque et al., 2013; Marshall et al., 2016).

When running models allowing attribution probabilities to differ before and after the intervention in the poultry industry in 2007, a clear difference was observed, with much lower poultry (and higher ruminant) attribution, particularly in main urban centres after the intervention during 2008–2016. Considering that most people in the Manawatu region live in urban areas, and that case rates in urban areas decreased from 2008 onwards, it is clear that this intervention coincided with a dramatic reduction in poultry attributed illness as reported elsewhere (Sears et al., 2011). Given that campylobacteriosis cases in rural areas are mostly attributed to ruminant sources, and that case rates in these areas have been relatively higher than those in urban centres since 2008 (see Figure 2.2), there is a clear need for public health interventions to focus in this area.

While the overall attribution was consistent between the Dirichlet and asymmetric Island models, it would be expected that the conditional probabilities for a given genotype might differ markedly. For those genotypes observed infrequently (or not at all) among the sources, the Dirichlet model has little information, while the asymmetric Island model can exploit information from cases with similar (but not identical) genetic profiles. In the case of MLST data with just 7 loci, the majority of human cases and source isolates come from a relatively small number of sequence types which are observed often. Thus the Dirichlet model performs well, as it has sufficient observations to estimate the genotype distribution well where the bulk of the data lie. It is only those genotypes that are rarely observed where it performs poorly, but as they are rarely observed, they do not contribute significantly to the overall attribution. In other circumstances, such as where we have many more than 7 loci, we would expect to

have many more rare genotypes, so that the Dirichlet model might provide little useful information. At the extreme example of whole genome MLST (wgMLST) where each isolate would typically be unique, it would provide essentially no information at all. In such circumstances, however, the asymmetric Island model would be expected to still perform well, assuming that information could still be transferred between similar genotypes.

The Dirichlet model is less complex than the asymmetric Island model that utilises the prevalence of genotypes in sources to derive the sampling distribution of genotypes. It is similar to the HaldDP model introduced in Chapter 3.4, that jointly models the source and human cases, accounting for uncertainty in the sampling process. However, this model is an extension of the Hald and modified Hald models, which model human cases using a Poisson rather than a multinomial distribution, and instead of estimating the proportion of cases attributed to each source directly, model source effects as well as genotype effects (Miller et al., 2017; Mughini-Gras et al., 2018).

In the models developed so far water is assumed to be a source of human campylobacteriosis infection, but water differs from the other food and environmental sources in that it is not an amplifying reservoir for *Campylobacter* (Wagenaar et al., 2013). In contrast, genotypes found in water might be expected to originate in the other sources present here, particularly ruminants and wild birds, but also potentially from humans as well via discharge of unprocessed human waste. Hence, water acts as a transmission pathway from sources to humans, being both an end point (reduced water quality from faecal contamination) and a source (human consumption of water, either recreationally or through untreated water supplies). Despite the fact that there is presently little evidence that water is an important source for campylobacteriosis from the current models, it is of importance to expand the role of water in the source attribution models, which will be discussed in the next chapter.

Chapter 6

The role of water in the transmission process

The consumption of water, either recreationally or through poor-quality water supplies, has been considered as one of potential sources of *Campylobacter* infection in humans – both sporadic cases and outbreaks (Mcbride, 2012). The pathogen can spread through water, which can be contaminated with faeces, or from streams and rivers near where animals graze. *Campylobacter* outbreaks caused by water contamination have been consistently reported all over the world, from one in Vermont, the United States in 1978 to a recent one in Havelock North, New Zealand in 2016 (CDC, 1978; DIA, 2017). The first caused 20% of residents to become ill, while the latter was estimated to impact approximately 40% of residents and contributed to three deaths. A recent investigation of the latter outbreak reveals that faeces from sheep were the most likely pathway contaminating water supplies and that heavy rainfall is believed to be the environmental factor resulting in the occurrence of the outbreak (Gilpin et al., 2020).

Despite the results in the previous chapters suggesting no more than 5% of sporadic cases attributable to environmental water, the magnitude of a waterborne outbreak can be enormous. This brings us to a key question: what role does water play in the transmission? Water as a source, being a mixing vessel receiving strains from sources such as wildlife and water birds (e.g. ducks), is not an amplifying reservoir for the pathogen (Mullner et al., 2009a,b; Mughini Gras et al., 2012; Mughini-Gras et al., 2016; Shrestha et al., 2019). It also acts as a proxy for sources other than four source categories used in this research, i.e. poultry, ruminants and other sources such as family pets. In this regard, the inclusion of water in the modelling is advisable, especially if the sampling of the other sources is suboptimal, as it allows for an estimate of the likely contribution of these unknown sources beyond the ones included in the data.

Therefore, the aim of this chapter is to extend the role of water in the modelling. A new method of treating water as a medium will be proposed based on the model

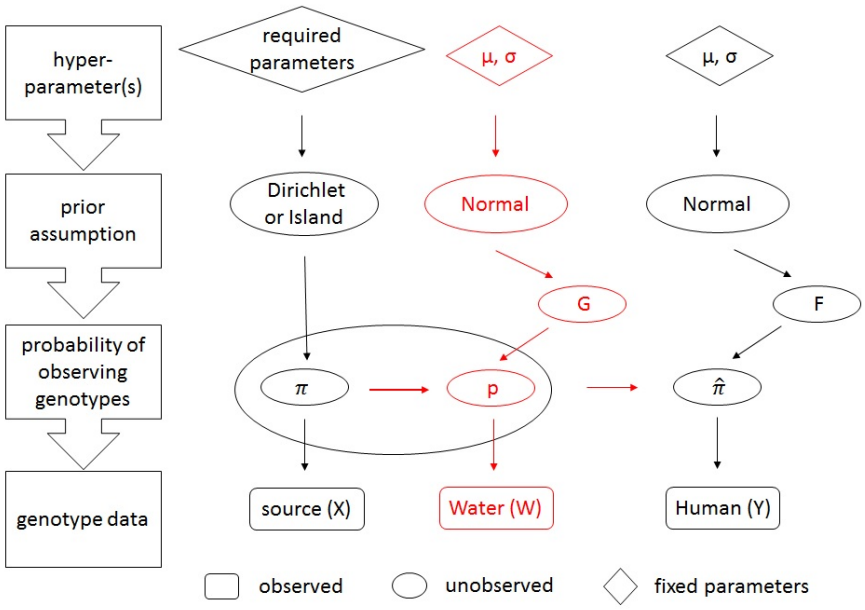


Figure 6.1: The framework is similar to that described in the previous chapters, but it introduces a method of modelling water attribution coloured in red, in which the probability p for water data is related to the probability π for source data. After p is estimated, it provides water information together with source information to the probability $\hat{\pi}$ for inference about human attribution.

framework built in the previous chapter. In previous analyses, samples from water birds were originally categorised in the group of other sources in the data. To better understand the role that water birds play in water contamination, and hence human infection, they will be separated from the category of other sources and regarded as an additional source group in the analysis.

In this chapter, the newly developed models will be depicted in Section 6.1. The general idea of modelling here is similar to before. Nonetheless, this time, data for sources other than water will be modelled to attribute the contamination of water. Then, the findings from the water data will be considered together in the modelling for human attribution. Section 6.2 describes how to fit data to estimate water and human attribution along with applied MCMC methods. Results and convergence diagnostics are presented in Sections 6.3 and 6.4, followed by discussions in the last section.

6.1 Modelling for water and human attribution

The framework of modelling is illustrated in Figure 6.1. The flow chart on the left side explains the general idea of the framework. It starts from a model assumption about the probability of observing genotypes on samples from sources $\boldsymbol{X}$ (excluding water),

water $\boldsymbol{W}$ and human cases $\boldsymbol{Y}$. As it is a Bayesian approach, prior assumptions are assigned to the parameters used in the models with constant hyperparameters.

To distinguish source attribution from water, data collected from water samples $\boldsymbol{W}$, which is a vector of the ST counts found in water, are separated from the source data. Similar to the models proposed before, a multinomial probability distribution is retained to describe each data set. First, modelling for source data $\boldsymbol{X}$ other than water is unchanged, but the number of source categories reduces to $J-1$. The sampling distribution of genotypes $\boldsymbol{\pi}$ for sources is still estimated by either the asymmetric Island or Dirichlet model (see Chapter 5.1). However, the focus of final attribution is changed from humans to water, i.e. the methods developed from the previous chapter are applied with water data $\boldsymbol{W}$ to replace human ones. Next, $\boldsymbol{W}$ are assumed multinomial distributed with the probability p_i of observing sequence type i from water samples, where $i = 1, \ldots, I$, given I unique genotypes typed from all isolates. The parameters $\boldsymbol{p}$ are also determined by the probabilities $\boldsymbol{\pi}$ and $\boldsymbol{G}$, where $\boldsymbol{G} = \{G_j\}_{j=1}^{J-1}$ specifies the probability of water observations falling in the source category j, $j = 1, \ldots, J-1$, through log-odds g_j, whose prior is assumed to follow a normal distribution. Lastly, the probability $\hat{\boldsymbol{\pi}}$ of observing genotypes on human isolates is not only estimated by a linear model $\boldsymbol{F}$ on the logit scale as before, but also by integrating the information from sources and water.

To start with, the likelihood of observing water data is equivalent to the one defined for the source data in Equation (4.1) with different notations. Let w_i denote the number of genotype i typed from water samples n_w and denoted by p_i, which is the probability that any given water sample is of sequence type i. The likelihood for water data is hence of the form,

$$L(\boldsymbol{W}; \boldsymbol{p}) \propto \prod_{i=1}^{I} p_i^{w_i}, \quad w_i \in \{0, 1, \cdots, n_w\}, \quad 0 < p_i < 1.$$

For convenient notations of the category of sources and water, the last category of J sources is assumed to be water, then the probability p_i for water data is determined by,

$$p_i = \sum_{j=1}^{J-1} \pi_{ij} G_j, \tag{6.1}$$

where π_{ij} is the probability of genotype i typed from source j, which has been specified in Chapter 5 estimated by the asymmetric Island or Dirichelt model, and G_j is defined as,

$$G_j = \frac{\exp(g_j)}{\sum_{l=1}^{J-1} \exp(g_l)}, \quad j = 1, \ldots, J-1, \tag{6.2}$$

in which the log-odds g_j represents the probability that a water sample is attributed to source j, given the source baseline is the category $J-1$. This means that $g_{J-1}=0$, and $G_{J-1}=1-\sum_{j=1}^{J-2}G_j$.

For human data $\boldsymbol{Y}$, they are also assumed to be multinomial distributed with the notations of y_i denoting the number of human isolates of type i, and q_i denoting the corresponding probability that any human isolate is of that type. If there are no variables considered in the attribution model, the likelihood for $\boldsymbol{Y}$ is defined as,

$$L(\boldsymbol{Y};\hat{\boldsymbol{\pi}})\propto\prod_{i=1}^{I}\hat{\pi}_i^{y_i},\quad 0<\hat{\pi}_i<1,\quad i=1,\ldots,I,$$

where

$$\hat{\pi}_i=p_iF_J+\sum_{j=1}^{J-1}\pi_{ij}F_j, \tag{6.3}$$

and F_j has already been specified in Equation (4.6). Otherwise, the likelihood function in the case-specific form (rather than type-specific) can be written as,

$$L(\boldsymbol{Y};\hat{\boldsymbol{\pi}})\propto\prod_{h=1}^{H}\hat{\pi}_h,\quad 0<\hat{\pi}_h<1,\quad h=1,\ldots H,$$

where

$$\hat{\pi}_h=p_{i[h]}F_{hJ}+\sum_{j=1}^{J-1}\pi_{i[h]j}F_{hj}, \tag{6.4}$$

in which

$$F_{hj}=\frac{\exp\left(f_{hj}\right)}{\sum_{l=1}^{J}\exp\left(f_{hl}\right)},\quad j=1,\ldots,J, \tag{6.5}$$

given the source category J as the baseline such that $f_{hJ}=0$ and $F_{hJ}=1-\sum_{j=1}^{J-1}F_{hj}$, and the definition of $\boldsymbol{f}$ has already been specified in Equation (4.9). Therefore, $\hat{\pi}_h$ is not only derived from the attribution probability F_{hj}, but also affected by the water attribution probability $p_{i[h]}$, with an h index specifying the sequence type i observed in water is also found on the h^{th} individual.

6.2 Model fitting and MCMC simulation

Let the order of four source categories j be fixed as poultry, other, ruminants and water, the following sections demonstrate model fitting for water and human attribution, given whether the rurality variable is considered in the analysis. In spite of the role of water sometimes being different from a source in this chapter, it is still categorised in the source groups for convenient descriptions of model fitting.

6.2.1 Estimates for water attribution

When water birds are not considered as an additional source of infection, the probabilities $\boldsymbol{G}$ for the first three sources, given the source baseline is ruminants, are specified as,

$$\begin{aligned} G_1 &= \frac{\exp(g_1)}{1+\exp(g_1)+\exp(g_2)} \\ G_2 &= \frac{\exp(g_2)}{1+\exp(g_1)+\exp(g_2)} \\ G_3 &= \frac{1}{1+\exp(g_1)+\exp(g_2)}. \end{aligned}$$

Then, given the sampling distribution of genotypes π_{ij}, $i = 1, \ldots, 377$, $j = 1, \ldots, 4$, has been simulated using one of the two types of model demonstrated in Chapter 5.1, the probability of observing genotype i in water is thus estimated through,

$$p_i = \pi_{i1}G_1 + \pi_{i2}G_2 + \pi_{i3}G_3.$$

6.2.2 Estimates for human attribution

Assume the log-odds $\boldsymbol{f}$ for each human case can be expressed as a linear predictor. When there are no variables included in the model, the log-odds are of the form,

$$f_j = \beta_{0j}, \quad j = 1, 2, 3, 4,$$

where $f_3 = 0$. Then, the attribution probabilities on the logit scale for each source are calculated through Equation (4.6). Next, the proportion $\hat{\pi}_i$ of genotype i found on humans is expressed as,

$$\hat{\pi}_i = \Big(\pi_{i1}G_1 + \pi_{i2}G_2 + \pi_{i3}G_3\Big)F_4 + \pi_{i1}F_1 + \pi_{i2}F_2 + \pi_{i3}F_3.$$

By contrast, when the rurality variable c is considered, the model is defined algebraically,

$$\boldsymbol{f}_{1804\times 3} = \begin{bmatrix} 1 & c_1 \\ 1 & c_2 \\ \vdots & \vdots \\ 1 & c_{1804} \end{bmatrix} \begin{bmatrix} \beta_{01} & \beta_{02} & \beta_{04} \\ \beta_{11} & \beta_{12} & \beta_{14} \end{bmatrix}.$$

For each individual, the attribution probability for each source leads to,

$$\begin{aligned} F_{h1} &= \frac{\exp\left(\beta_{01} + \beta_{h1}c_h\right)}{1 + \exp\left(\beta_{01} + \beta_{h1}c_h\right) + \exp\left(\beta_{02} + \beta_{h2}c_h\right) + \exp\left(\beta_{04} + \beta_{h4}c_h\right)} \\ F_{h2} &= \frac{\exp\left(\beta_{02} + \beta_{h2}c_h\right)}{1 + \exp\left(\beta_{01} + \beta_{h1}c_h\right) + \exp\left(\beta_{02} + \beta_{h2}c_h\right) + \exp\left(\beta_{04} + \beta_{h4}c_h\right)} \\ F_{h3} &= \frac{1}{1 + \exp\left(\beta_{01} + \beta_{h1}c_h\right) + \exp\left(\beta_{02} + \beta_{h2}c_h\right) + \exp\left(\beta_{04} + \beta_{h4}c_h\right)} \\ F_{h4} &= \frac{\exp\left(\beta_{04} + \beta_{h4}c_h\right)}{1 + \exp\left(\beta_{01} + \beta_{h1}c_h\right) + \exp\left(\beta_{02} + \beta_{h2}c_h\right) + \exp\left(\beta_{04} + \beta_{h4}c_h\right)}, \end{aligned}$$

and the probability of observing genotype i on human isolates can be rewritten using an index $i[h]$ to link the probability of observing a genotype both on a source and the h^{th} human sample. That is,

$$\hat{\pi}_h = p_{i[h]}F_{h4} + \pi_{i[h]1}F_{h1} + \pi_{i[h]2}F_{h2} + \pi_{i[h]3}F_{h3}.$$

When water birds as an additional source are separated from the category of other sources, the method of model fitting for water and human attribution is fairly similar to the above description. The demonstration can be found in Appendix A.3. Note that the data set used here differs from the original one in that water birds have been filtered out of the category of other sources to be an additional source, and that the time period of sampling has extended from the end of 2016 to 2017, i.e. the sample dates of data now range from February 2005 to December 2017. Therefore, the total number of isolates increases from 4,322 to 4,600, and the number of unique STs that have been typed from human and source isolates becomes 406. There is little difference in the number of isolates falling in each source category and the amount of human cases living in each class of rurality between the two data sets. The comparison of these features is tabulated in Appendix B.1 (Tables B.1 and B.2).

6.2.3 MCMC simulation

The MCMC method applied in this chapter is still a Metropolis-Hastings algorithm, which has been run in four different settings, combining the inclusion of the the numeric

rurality variable c and/or of water birds as an additional source of infection.

Algorithm 4 Metropolis-Hastings algorithm, given $J = 4$

1: Define the design matrix $\boldsymbol{C}_{1804\times P}$, where $P = 1$, or 2
2: Draw starting points of $\boldsymbol{g}^{(0)} = \{g_v\}_{v=1}^{V}$ and $\boldsymbol{\beta}^{(0)} = \{\beta_t\}_{t=1}^{T}$ from N(0,1), where $(V, T) = (2, 3)$, given $P = 1$, or $(3, 6)$, given $P = 2$
3: Calculate $\boldsymbol{G}^{(0)}$, given $\boldsymbol{g}^{(0)}$ via Equation (6.2)
4: Find $\boldsymbol{F}^{(0)}$, given $\boldsymbol{\beta}^{(0)}$ and $\boldsymbol{C}$ via Equation (4.6) when $P = 1$, or (6.5) when $P = 2$
5: Simulate S times the sampling distribution of genotypes $\boldsymbol{\pi}_j$ for source j, $j = 1, \ldots, 4$ as applied in line 5 or 6 of Algorithm 3 in the previous chapter
6: **for** simulation s, $s = 1, \ldots, S$ of $\boldsymbol{\pi}^{(s)}$ **do**
7: Compute $\boldsymbol{p}^{(0)}$ using Equation (6.1)
8: Compute $\hat{\boldsymbol{\pi}}^{(0)}$ using equations (6.3) and (6.4) when $P = 1$ and 2
9: **for** iteration m, $m = 0, \ldots, M-1$ **do**
10: Sample v from the permutation P_V of $\{1, \ldots, V\}$ for the parameters $\boldsymbol{g}$
11: Sample t from the permutation P_T of $\{1, \ldots, T\}$ for the parameters $\boldsymbol{\beta}$
12: Propose $\boldsymbol{g}^*$ with the value of g_v^* sampled from $Q\left(g_v^* \mid g_v^{(m)}\right)$
13: Propose $\boldsymbol{\beta}^*$ with the value of β_t^* drawn from $Q\left(\beta_t^* \mid \beta_t^{(m)}\right)$
14: Find $\boldsymbol{G}^*$, $\boldsymbol{F}^*$, $\boldsymbol{p}^*$ and $\hat{\boldsymbol{\pi}}^*$ repeating steps of lines 3, 4, 7 and 8
15: Generate $u \sim U(0, 1)$
16: Compute $\mathcal{A}\left(g_v^{(m)}, g_v^*, \beta_t^{(m)}, \beta_t^*\right) = \min\{1, \psi\}$, where

$$\psi = \frac{L\left(\boldsymbol{W}, \boldsymbol{Y}; g_v^*, \beta_t^*\right)}{L\left(\boldsymbol{W}, \boldsymbol{Y}; g_v^{(m)}, \beta_t^{(m)}\right)} \frac{\pi\left(g_v^*\right)}{\pi\left(g_v^{(m)}\right)} \frac{\pi\left(\beta_t^*\right)}{\pi\left(\beta_t^{(m)}\right)}$$

17: **if** $u < \mathcal{A}\left(g_v^{(m)}, g_v^*, \beta_t^{(m)}, \beta_t^*\right)$ **then**
18: $g_v^{(m+1)} \leftarrow g_v^{(*)}$ and $\beta_t^{(m+1)} \leftarrow \beta_t^{(*)}$
19: **else**
20: $g_v^{(m+1)} \leftarrow g_v^{(m)}$ and $\beta_t^{(m+1)} \leftarrow \beta_t^{(m)}$
21: **end if**
22: **end for**
23: **end for**

The steps of simulation are outlined in Algorithm 4, in which water birds are not considered as a source group, i.e. the number of source categories is $J = 4$, containing poultry, ruminants, water and other sources. First, the dimension of the design matrix $\boldsymbol{C}$ is 1804 $\times P$, where P is the number of regression parameters included in the attribution model. In our case, P equals to 1 or 2, indicating the model only contains an intercept, or an intercept and the slope of the rurality variable, respectively. Next, vectors of log-odds $\boldsymbol{g}$ and regression parameters $\boldsymbol{\beta}$ are initialised by sampling from N(0,1) as stated in line 2. (V, T) indicates the length of elements contained in $\boldsymbol{g}$ and $\boldsymbol{\beta}$, which are determined by $J - 2$ and $P \times (J - 1)$, respectively. In this case, $(V, T) = (2, 3)$ as

$P = 1$. After initialising the values of $\boldsymbol{g}$ and $\boldsymbol{\beta}$, the parameters $\boldsymbol{G}$ are de-logited via Equation (6.2), and so are the attribution probabilities $\boldsymbol{F}$ through Equation (4.6) when $P = 1$, or Equation (6.5) when $P = 2$. Further, similar to that described in Chapter 5, the probability π_{ij} that isolates typed from source j are of sequence type i is simulated $S = 100$ times using the asymmetric Island or Dirichlet model in order to mix over multiple samples of $\boldsymbol{\pi}$ in the final inference.

For every simulated $\boldsymbol{\pi}^{(s)}$, $s = 1, \ldots, 100$, the starting points of $\boldsymbol{p}$ and $\hat{\boldsymbol{\pi}}$ are calculated, given $\boldsymbol{\pi}^{(s)}$, $\boldsymbol{G}^{(0)}$ and $\boldsymbol{F}^{(0)}$. Then, the procedure of the Metropolis-Hastings sampler is applied to update the elements of $\boldsymbol{g}$ and $\boldsymbol{\beta}$ at the same time. For each iteration m, where $m = 0, \ldots, M-1$, permutations $P_V = \{1, \ldots, V\}$ and $P_T = \{1, \ldots, T\}$ are sampled in order to decide which element is going to be replaced. Given v and t, the new elements g_v^* and β_t^* are proposed using a random walk sampling with normal distributed steps as stated in lines 12 and 13. The proposed $\boldsymbol{G^*}, \boldsymbol{F^*}$, $\boldsymbol{p^*}$ and $\hat{\boldsymbol{\pi}}^*$ are thus obtained after repeating lines 3, 4, 7, and 8. Lastly, to decide if the move of proposals (g_v^*, β_t^*) is accepted in a chain, a random sample u is drawn from U(0,1) and an acceptance rate $\mathcal{A} = \min\{1, \psi\}$ is computed for evaluation. If u is smaller than $\mathcal{A}$, the proposals are accepted to be the elements in the next state $m + 1$. Otherwise, the proposals are rejected and the elements with the current state m are retained to the state of $m + 1$.

In fact, the calculation of ψ involves the ratio of joint likelihood for water and human data, proposal densities $Q(g_v^* \mid g_v^{(m)})$ and $Q(\beta_t^* \mid \beta_t^{(m)})$, and prior densities for g_v and β_t. However, the ratio of each proposal density is cancelled out due to the symmetric random walk. As ψ is on a logarithmic scale in the programming, the criterion of acceptance turns out to be $\log u < \min\{0, \log \psi\}$, where each term of $\log \psi$ is thus specified as,

$$\log \frac{\pi\left(g_v^*\right)}{\pi\left(g_v^{(m)}\right)} = \frac{{g_v^{(m)}}^2 - {g_v^*}^2}{2}$$
$$\log \frac{\pi\left(\beta_v^*\right)}{\pi\left(\beta_v^{(m)}\right)} = \frac{{\beta_t^{(m)}}^2 - {\beta_t^*}^2}{2},$$

and given $J = 4$, the joint likelihood on the log scale when P =1 or 2 leads to,

Algorithm 5 Metropolis-Hastings algorithm, given $J = 5$

1: Define the design matrix $\boldsymbol{C}_{1834\times P}$, where $P = 1$, or 2
2: Draw starting points of $\boldsymbol{g}^{(0)} = \{g_v\}_{v=1}^{V}$ and $\boldsymbol{\beta}^{(0)} = \{\beta_t\}_{t=1}^{T}$ from N(0,1), where $(V, T) = (3, 4)$, given $P = 1$, or $(3, 8)$, given $P = 2$
3: Calculate $\boldsymbol{G}^{(0)}$, given $\boldsymbol{g}^{(0)}$ via Equation (6.2)
4: Find $\boldsymbol{F}^{(0)}$, given $\boldsymbol{\beta}^{(0)}$ and $\boldsymbol{C}$ via Equation (4.6) when $P = 1$, or (6.5) when $P = 2$
5: Simulate S times the sampling distribution of genotypes $\boldsymbol{\pi}_j$ for source j, $j = 1, \ldots, 5$ as applied in line 5 or 6 of Algorithm 3 in the previous chapter
6: Same steps as listed in Algorithm 4 from lines 6 to 23

$$\log \frac{L\Big(\boldsymbol{W}, \boldsymbol{Y}; g_v^*, \beta_t^*\Big)}{L\Big(\boldsymbol{W}, \boldsymbol{Y}; g_v^{(m)}, \beta_t^{(m)}\Big)} = \sum_{i=1}^{377} w_i \Big(\log p_i^* - \log p_i^{(m)} \Big) + \log \Big(p_i^* F_J^* + \sum_{j=1}^{J-1} \pi_{ij} F_j^* \Big) - \log \Big(p_i^{(m)} F_J^{(m)} + \sum_{j=1}^{J-1} \pi_{ij} F_j^{(m)} \Big),$$

or,

$$\log \frac{L\Big(\boldsymbol{W}, \boldsymbol{Y}; g_v^*, \beta_t^*\Big)}{L\Big(\boldsymbol{W}, \boldsymbol{Y}; g_v^{(m)}, \beta_t^{(m)}\Big)} = \sum_{h=1}^{1804} w_{i[h]} \Big(\log p_{i[h]}^* - \log p_{i[h]}^{(m)} \Big) + \log \Big(p_{i[h]}^* F_{Jh}^* + \sum_{j=1}^{J-1} \pi_{i[h]j} F_{jh}^* \Big) - \log \Big(p_{i[h]}^{(m)} F_{Jh}^{(m)} + \sum_{j=1}^{J-1} \pi_{i[h]j} F_{jh}^{(m)} \Big).$$

On the other hand, if water birds are treated as an additional source of infection, then the MCMC steps are as listed in Algorithm 5. The procedure is a little different from the previous algorithm in that the length of elements for $\boldsymbol{\beta}$ changes from 3 to 4 if $P = 1$, or from 6 to 8 if $P = 2$, and that the number of source categories becomes $J = 5$. The remaining steps are exactly the same and the way to calculate the acceptance rate $\mathcal{A}$ is unchanged.

6.3 Results

6.3.1 Posterior water attribution

After simulating the sampling distribution of genotypes multiple times using the asymmetric Island and Dirichlet models, posterior distributions of water contamination attributed to four source categories are illustrated in the upper panel of Figure 6.2, given the fitted attribution model only has an intercept. It shows that most water contamination is due to other sources, accounting for more than 75% of samples, followed by

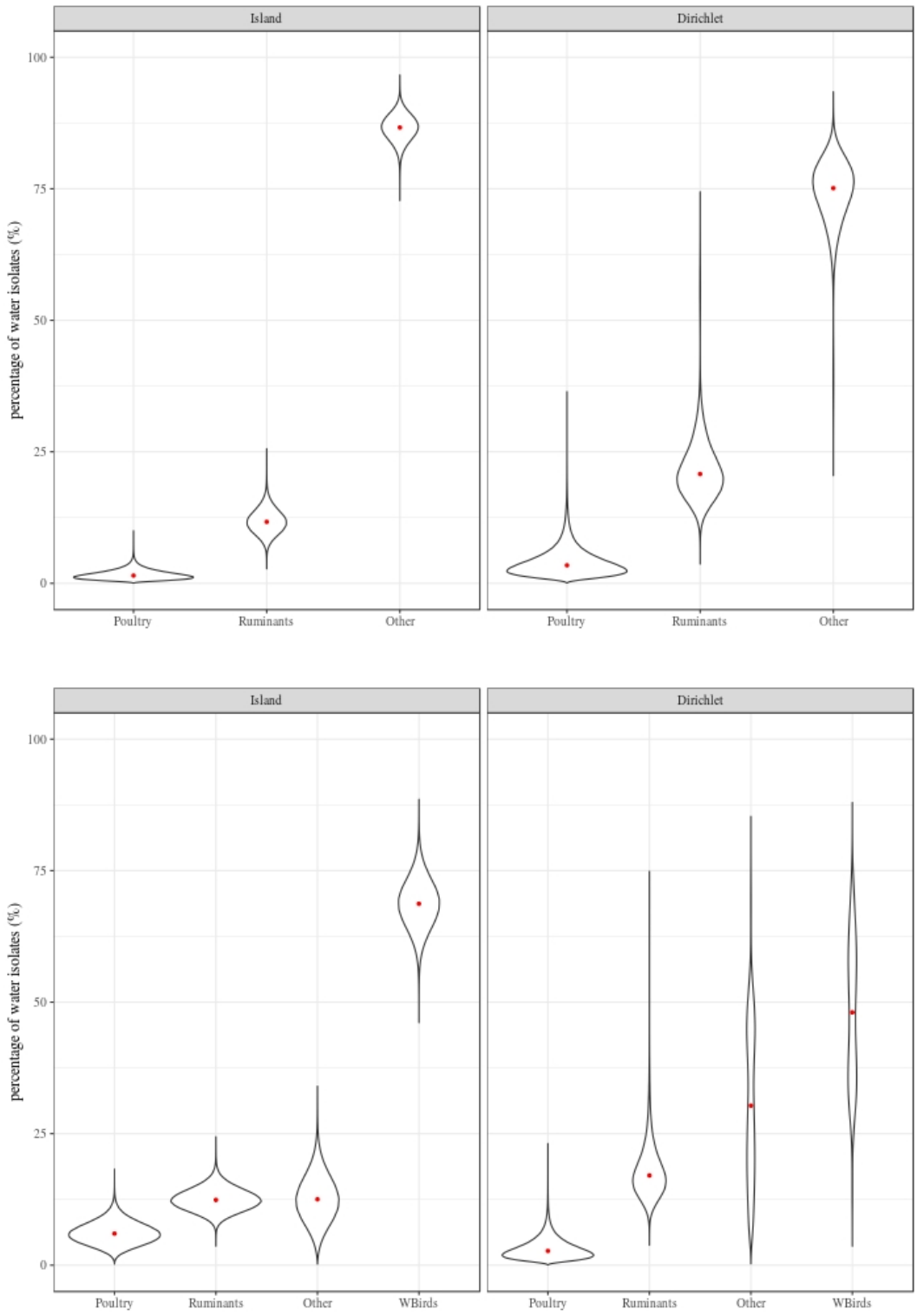


Figure 6.2: The percentage of water isolates attributable to each source with (lower panel) or without (upper panel) the source of water birds after mixing over 100 simulated sampling distributions of genotypes $\boldsymbol{\pi}$, estimated by the asymmetric Island model against the Dirichlet model in the analysis. The median for each category is marked in red. The inference is based on the fitted model when $c = 0$ and ruminants are treated as the source baseline.

ruminants and poultry whatever the type of model and sampling distribution. However, the Dirichlet model tends to have a larger variation in the attribution than the asymmetric Island model.

When water birds are regarded as a source and filtered from the category of other sources, the lower panel of Figure 6.2 shows that water birds indeed play an important role when $\boldsymbol{\pi}$ are estimated by the asymmetric Island model. On average, approximately 70% of water samples are contaminated by this source. The median marked in red is about 20% higher than that suggested by the Dirichlet model. The latter model shows that water birds have a higher median than other sources, although these two categorised sources have relatively higher uncertainty than poultry and ruminants observed from the flat shape of distribution. As a result of the small number of water and other source isolates, the latter model can not specifically distinguish the contamination caused by these two sources, while the first model provides more precise information to specify the effect of water birds due to the estimation of genetic distances between isolates.

In the source data, samples from other sources account for 15% of all samples, of which nearly half of samples (45%) are from water birds. From the lower panel of the figure, it is found that the suggested 75% of contribution from other sources from the upper panel of the figure may be overestimated if the effect of water birds is ignored. On the other hand, it is expected that there is no impact on water attribution whether epidemiological variables are included in the attribution model $\boldsymbol{F}$. Results of $c \neq 0$ are indeed very similar to that of $c = 0$, which are evident in Figure C.14 of Appendix C.2.

6.3.2 Posterior human attribution

Posterior attribution for human campylobacteriosis with and without the source of water birds considered in the application is displayed in Figure 6.3, given $c = 0$ is considered in the fitted model. The results before and after including water birds are comparable in that poultry and ruminants are still the key sources of infection (accounting for more than 90% of cases) and that human illness caused by water birds seems rather uncommon. Outliers are also observed for each source category, which may be because of sampling variation of $\boldsymbol{\pi}$.

When the rurality effect is considered in the linear model on the logit scale, the final attribution between two types of model and the associated 80% credible intervals with and without considering the source of water birds are presented in Figure 6.4. Again, including the source of water birds has no impact on the final inference. Poultry and ruminants are still the major causes of human infection. Both types of model have similar results, with a little more variation spread in the credible intervals in rural areas due to far fewer rural cases than urban cases.

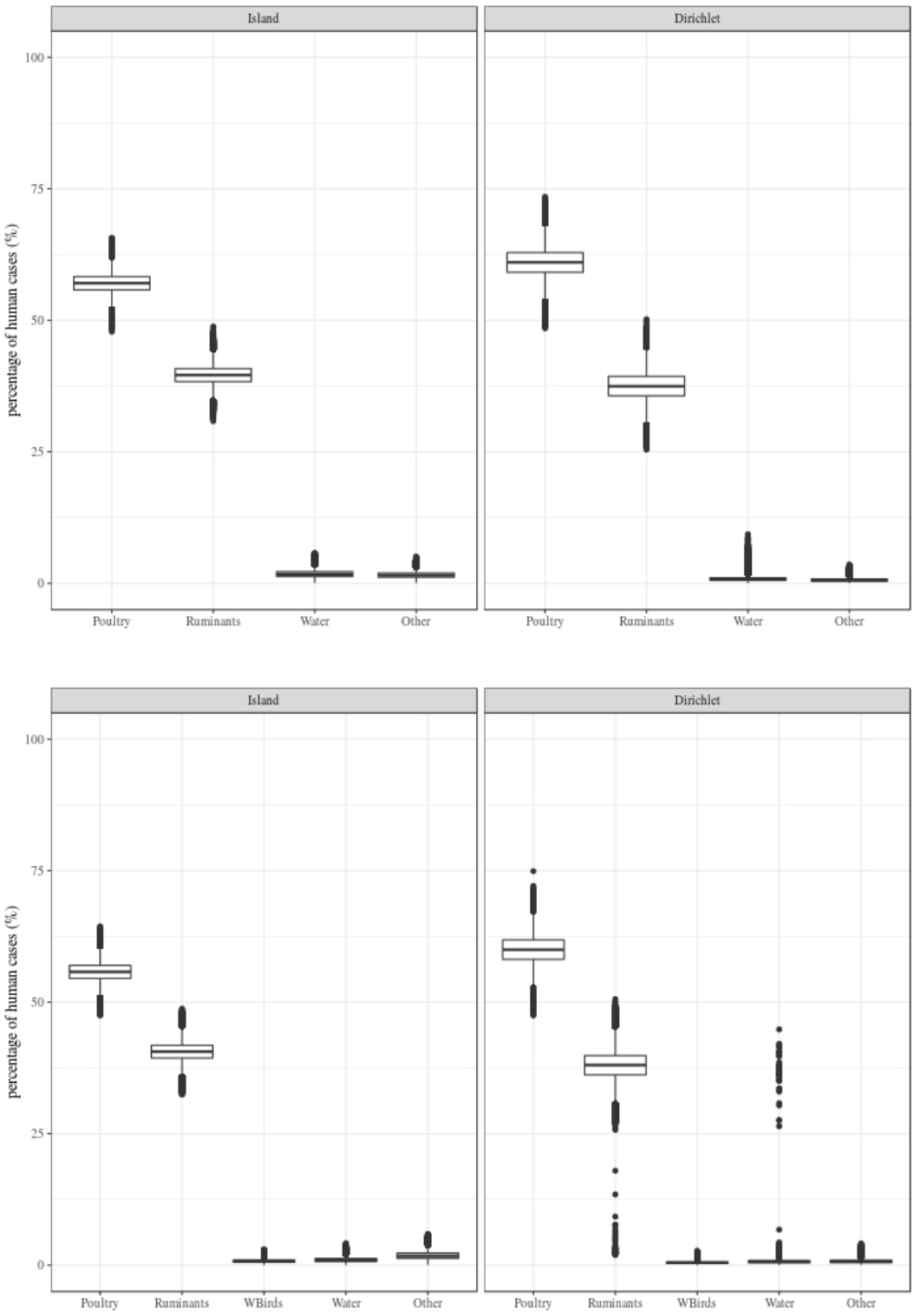


Figure 6.3: Posterior human attribution for the analysis including (lower panel) or not including (upper panel) the source of water birds, given no variables are considered in the modelling with ruminants as the source baseline and the sampling distribution of genotypes $\boldsymbol{\pi}$ are simulated 100 times by the asymmetric Island model against the Dirichlet model.

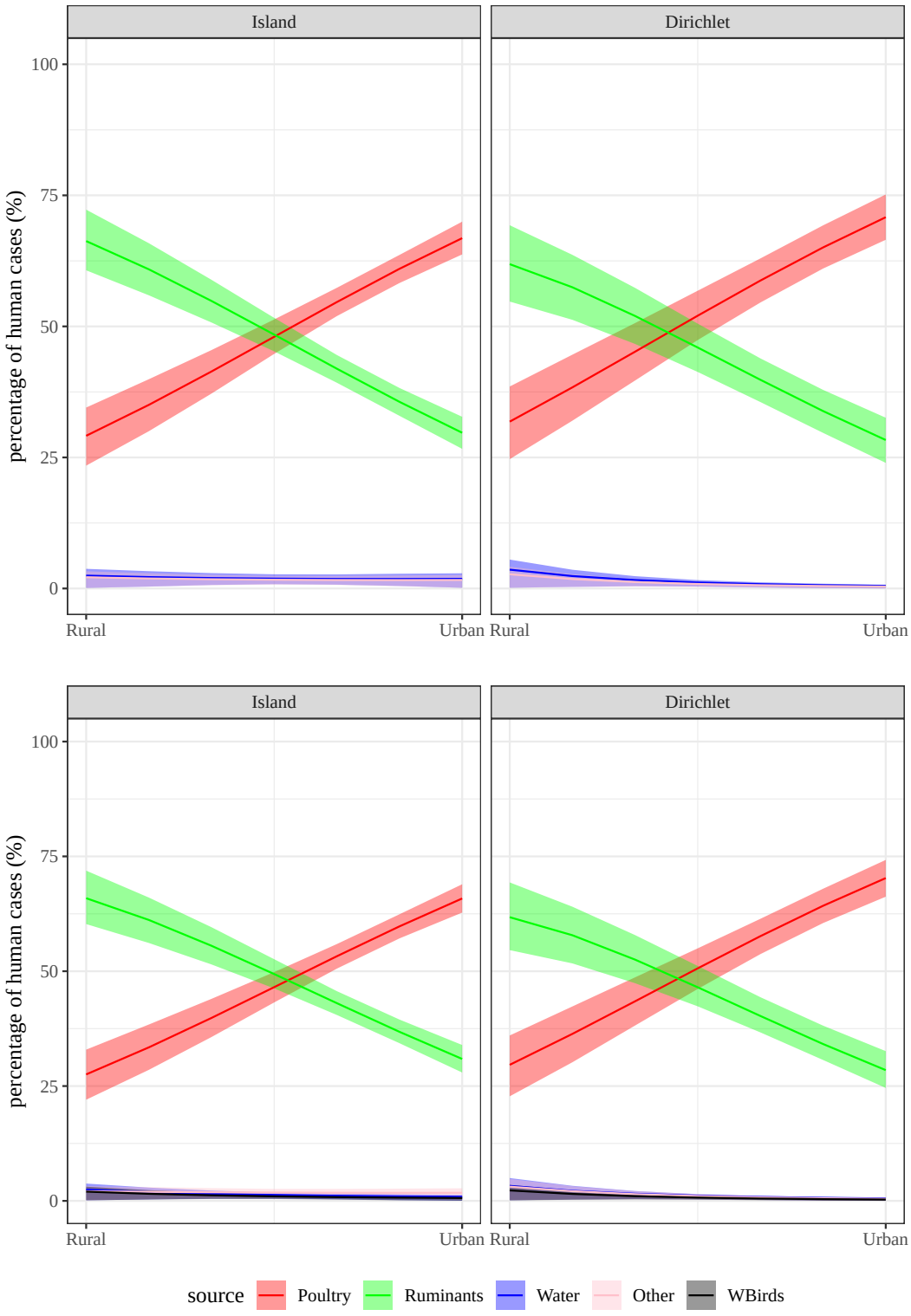


Figure 6.4: Posterior human attribution and 80% credible intervals for sources excluding water birds (upper panel), and for sources including water birds (lower panel), given the rurality variable is considered in the modelling with ruminants as the source baseline, and $\boldsymbol{\pi}$ are estimated 100 times by the asymmetric Island model against the Dirichlet model.

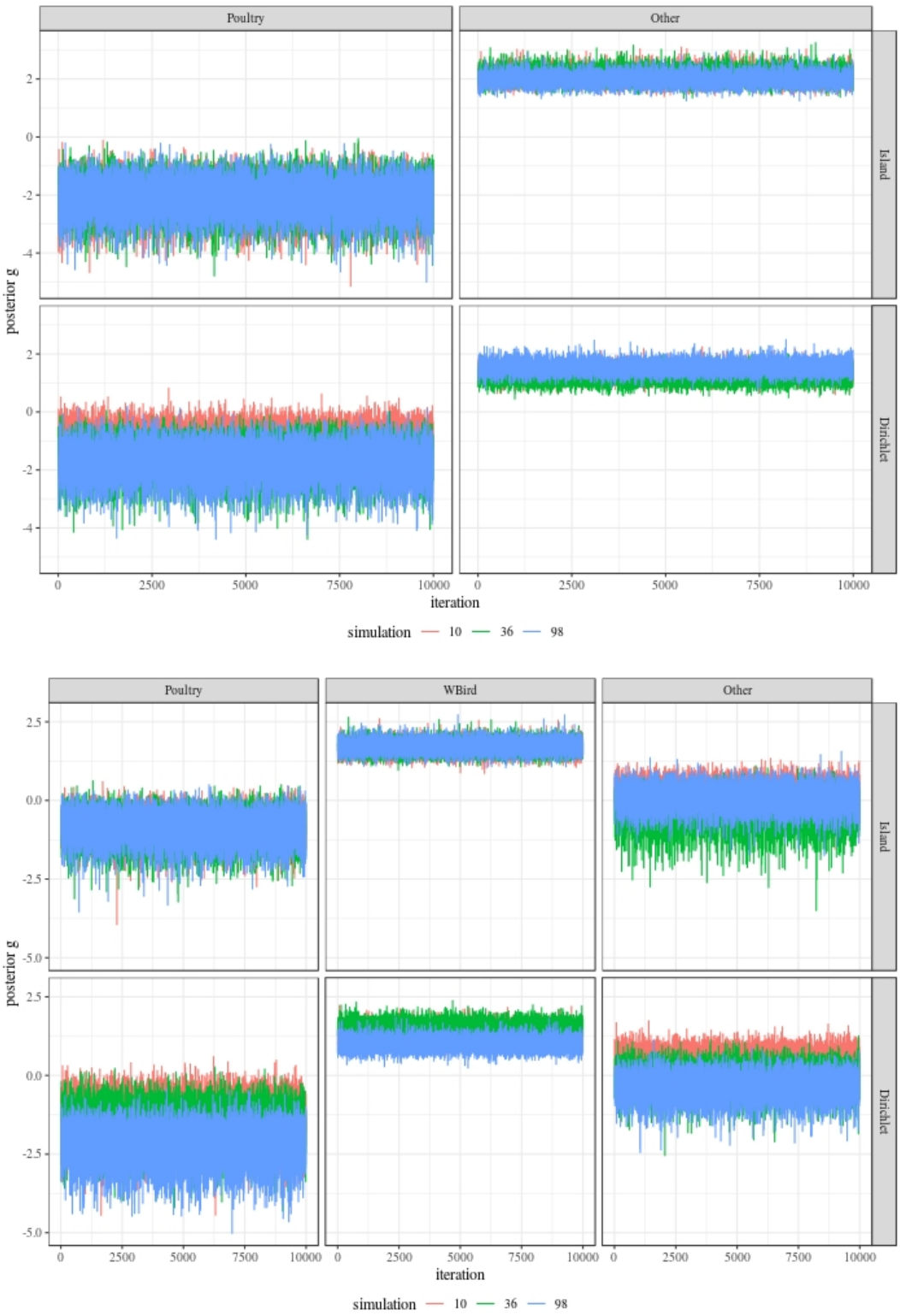


Figure 6.5: Trace plots for the water parameters $\boldsymbol{g}$ when the attribution model only includes the intercept in the analysis, in which the source of water birds is not included (upper panel) or included (lower panel), given the source baseline is ruminants and three simulations of $\boldsymbol{\pi}$ estimated by the asymmetric Island model against the Dirichlet model.

6.4 Diagnostics

6.4.1 Convergence for parameters

The MCMC sampler has been run for four different settings with different perspectives about variables and soure groups. Each time, the algorithm has produced 10,000 samples after discarding the first 3,000 samples and selecting every 100^{th} simulation in a chain, given one simulated $\boldsymbol{\pi}$. This means that the total number of posterior samples is 1,000,000 for 100 simulated $\boldsymbol{\pi}$. To visualise better the trace plot of parameters for convergence diagnostics, chains are displayed only for a few simulated $\boldsymbol{\pi}$, which are randomly picked from 100 simulations.

The trace plots for the water parameters $\boldsymbol{g}$ when the rurality effect is not included in the modelling is presented in Figure 6.5, given three randomly selected simulations of $\boldsymbol{\pi}$. The upper panel of the figure shows the results based on the analysis without the source of water birds. The chains for each of three simulations of genotype sampling distribution mix well and fast. They overlap for the most part, but as expected differ a little for differing samples of $\boldsymbol{\pi}$. For the analysis with water birds as a source, the trace plots presented in the lower panel of the figure are very similar to that without the source of water birds. Good mixing and well converged chains are also observed. This is same as to the chains for the intercept regression parameter for each source, which are displayed in Figure 6.6.

It is also observed that the chains have good mixing and convergence for water and regression parameters when the attribution model considers the rurality variable. However, an increase of the burn-in period may be required as the starting points of some chains are far away from where most of sequences centre (see Figures from C.15 to C.17 in Appendix C.2).

Regardless of whether the source of water birds is considered and whether the rurality variable is included in the attribution model, it does not affect the convergence of chains resulting from both types of model simulating the sampling distribution of genotypes.

6.4.2 Uncertainty of the sampling distribution of genotypes $\boldsymbol{\pi}$

The sampling distribution of genotypes $\boldsymbol{\pi}$ was originally simulated $S = 100$ times by the two types of model. This may lead more uncertainty when inferring the final attribution for water contamination and human infection. Thus, in order for the effect of integrating over all $\boldsymbol{\pi}$ to be explored, results concerning water and human attribution are compared by three different combinations of the number of iterations m and simulations s.

A comparison of uncertainty contains three final inferences based on three combinations of (s, m): $(1, 1)$, $(1, 100)$, and $(100, 100)$. Given one simulated $\boldsymbol{\pi}$ and one iteration,

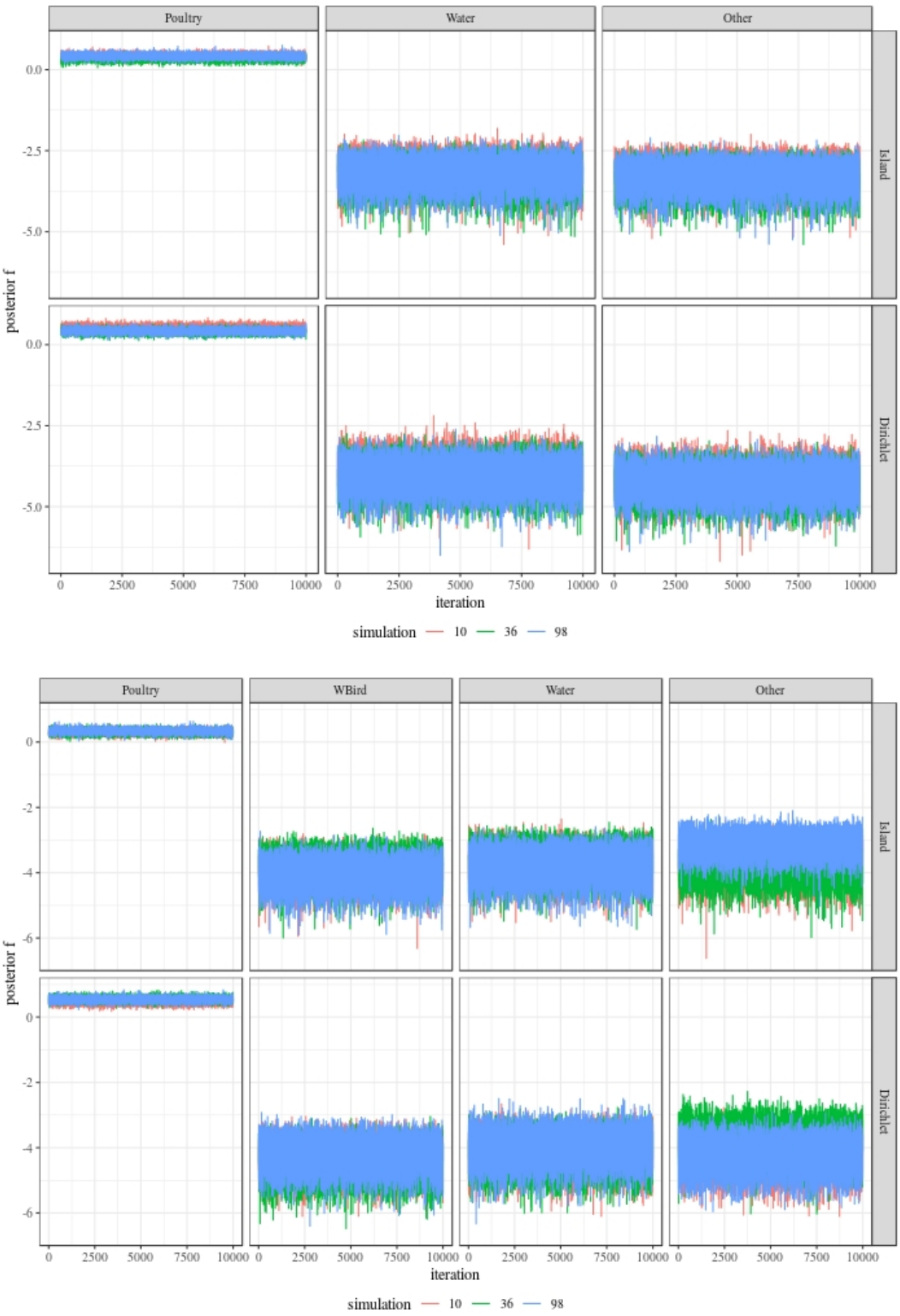


Figure 6.6: Trace plots for parameters $\boldsymbol{f}$ (or equivalently the intercept parameters) when the rurality variable is not considered in the attribution model for the analysis without (upper panel) or with (lower panel) the source of water birds, given the source baseline is ruminants and three simulations of $\boldsymbol{\pi}$ estimated by the asymmetric Island model against the Dirichlet model.

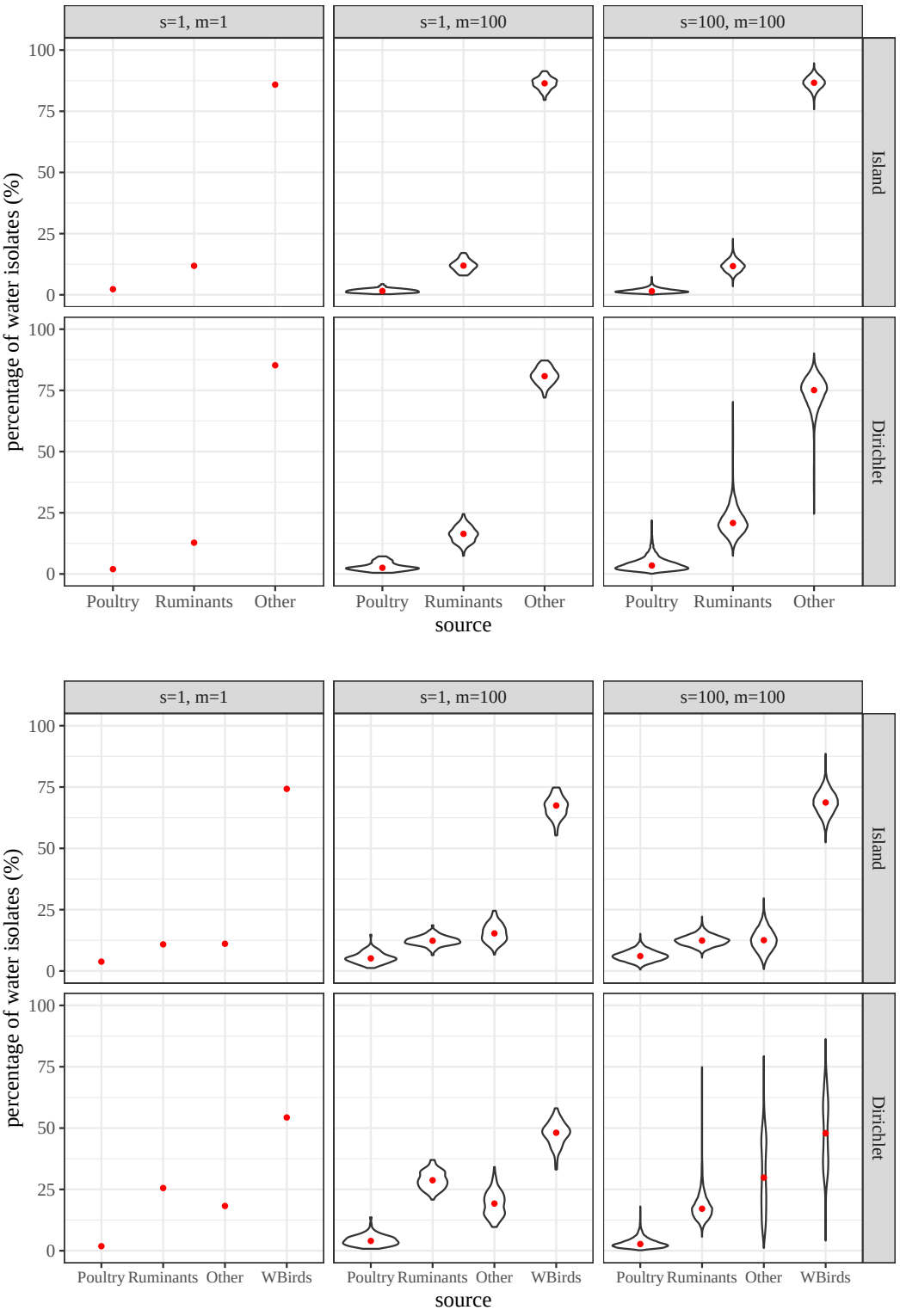


Figure 6.7: The percentage of water isolates attributable to each source before (upper panel) and after (lower panel) considering water birds in the analysis, given $\boldsymbol{\pi}$ are estimated by the two types of model, the source baseline is ruminants, and no variables are included in the attribution model. Each panel contains three outcome resulted from different combinations of the number of simulations s for $\boldsymbol{\pi}$ and the number of iteration m. The associated median for each source is marked in red, besides the left side of each panel, which is the point estimate for each source.

the result is a point estimate for each source contributing to water contamination or human infection. Then, the number of iterations m increases from 1 to 100, given the same simulated $\boldsymbol{\pi}$ ($s = 1$). This can illustrate how the final attribution changes after updating the water parameters and the regression parameters. Lastly, the number of simulations s increases from 1 to 100, while m retains 100. This shows the variation of the results from the simulations of $\boldsymbol{\pi}$.

When there is no variable considered in the attribution model, the posterior water attribution before and after including the source of water birds in the analysis are displayed in the upper and lower panels of Figure 6.7, given $\boldsymbol{\pi}$ are estimated by two types of model and the source baseline is ruminants. Each panel shows the results from three combinations of (s, m). Regardless of which type of model is used for estimating $\boldsymbol{\pi}$, both panels show an increase of uncertainty from the parameters when the iteration m increases from 1 to 100, given $s = 1$. After mixing over 100 simulations of $\boldsymbol{\pi}$, both panels display different variation for the two types of model. The right side of the upper panel shows that without water birds as a source, the Dirichlet model tends to have relatively larger uncertainty than the asymmetric Island model when inferring the sources of water contamination, while the latter model also has a slightly wider range for each source, compared to $(s = 1, m = 100)$. After including water birds in the analysis, the asymmetric Island model in the right side of the lower panel is similar as before. It points out the importance of including water birds as a separate category, with a little more uncertainty for each source. However, the Dirichlet model struggles to identify the contribution from water birds and other sources as their uncertainty becomes much larger. The uncertainty for ruminants is also increased, compared to $(s = 1, m = 100)$.

Meanwhile, the comparison of human attribution is also presented in Figure 6.8, with or without water birds considered in the analysis. Overall, both panels have similar variation for each type of model. Nonetheless, the Dirichlet model might have more uncertainty after integrating over all $\boldsymbol{\pi}$ when water birds are included in the analysis. This is evident in the right side of the lower panel, with more smaller and/or larger outliers observed for ruminants and water.

Further, when the attribution model takes the rurality effect into account, the comparison of water and human attributions with three combinations of (s, m) are displayed in Figures 6.9 and 6.10. For water attribution, it is similar to the findings without considering the rurality effect. The right side of both panels confirms again that the asymmetric Island model tends to have smaller uncertainty than the Dirichlet model. However, the point estimates for water birds and other sources from the Dirichlet model in the lower panel of Figure 6.9 are not similar to that of $(s = 100, m = 100)$

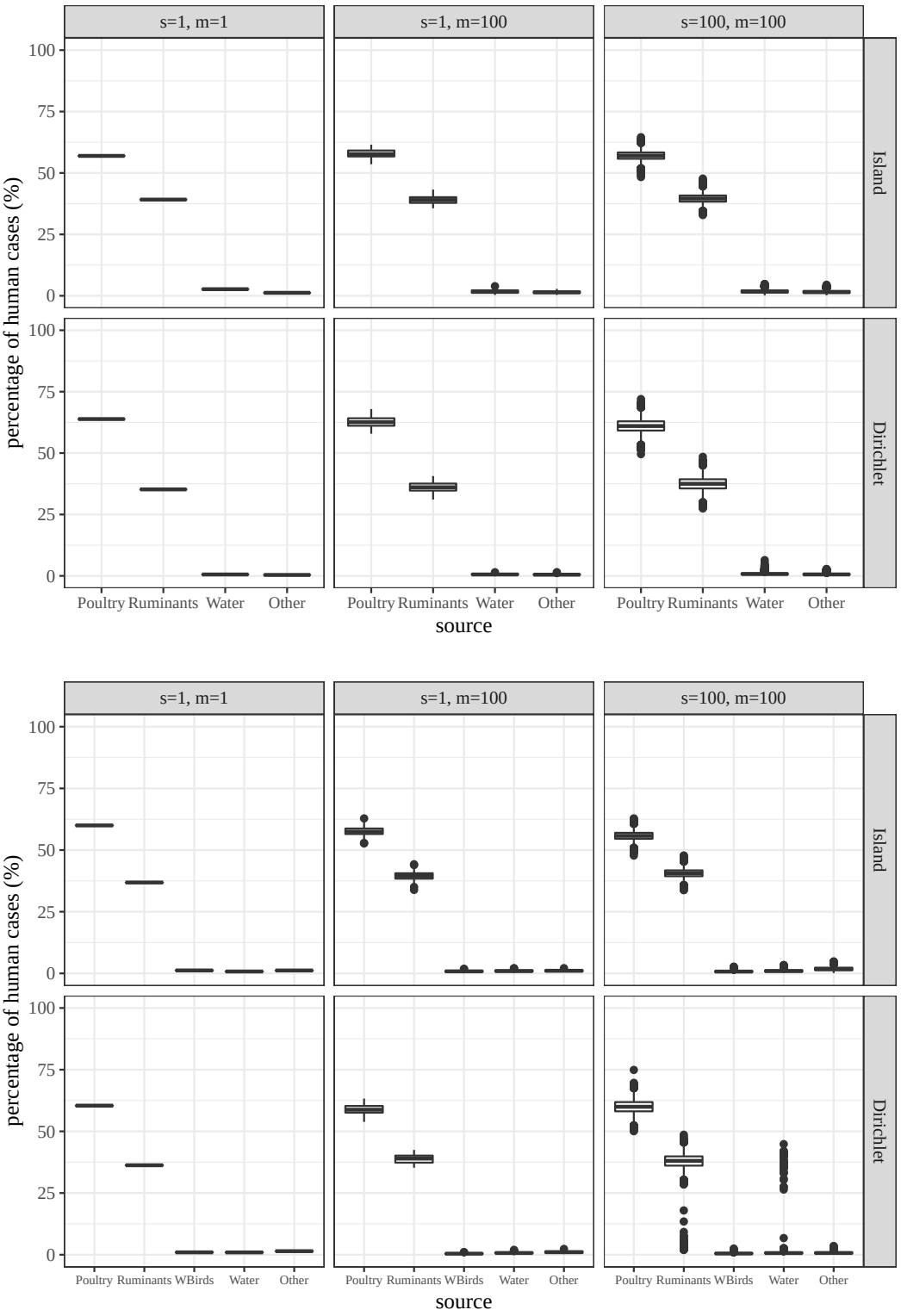


Figure 6.8: Posterior human attribution before (upper panel) and after (lower panel) considering water birds in the analysis, given $\boldsymbol{\pi}$ are estimated by the two types of model, the source baseline is ruminants, and no variables are included in the attribution model. Each panel contains three outcome resulted from different combinations of (s, m). The posterior means and the associated 80% credible intervals are illustrated, except for the left side of each panel, which is just the point estimate for each source.

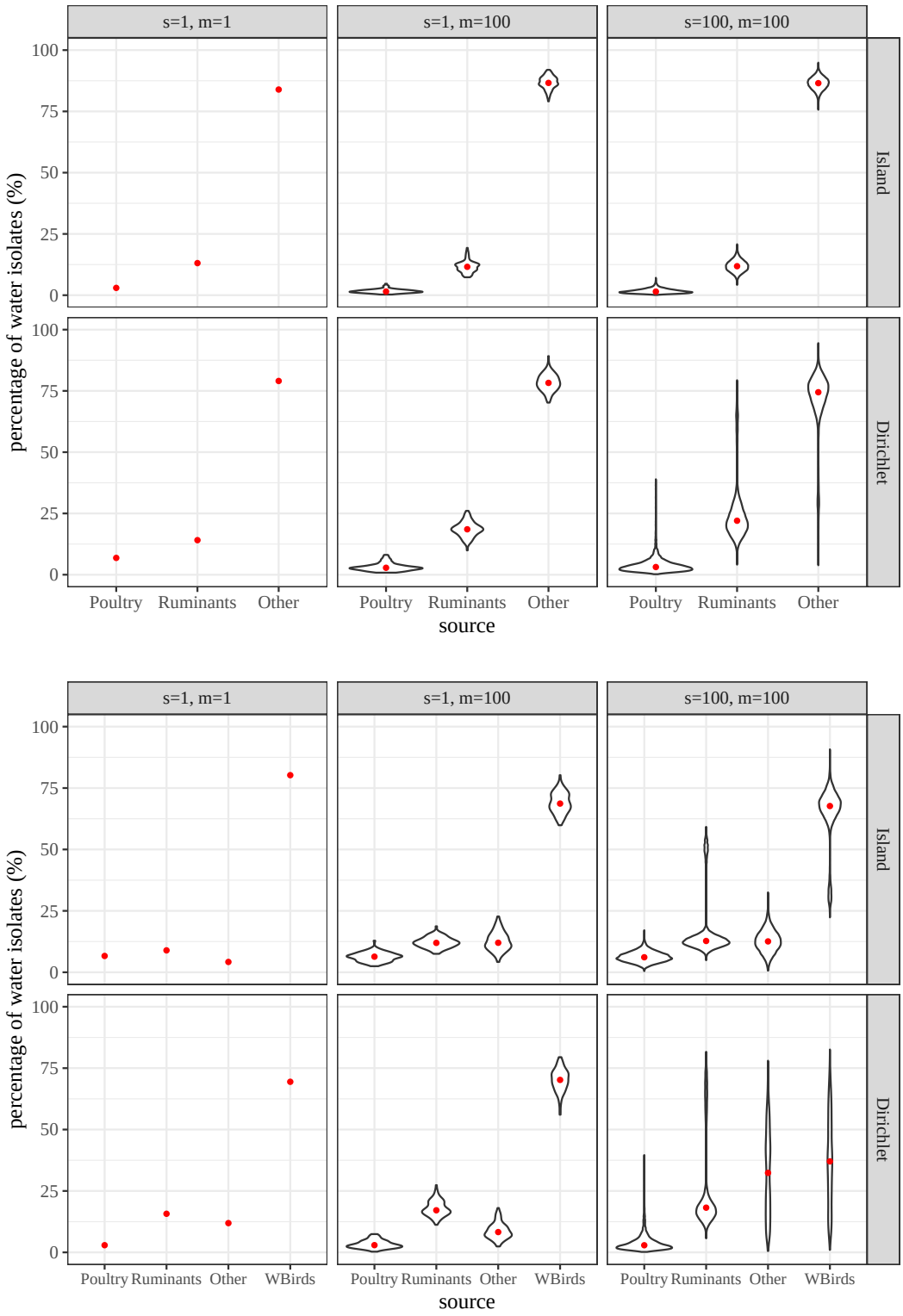


Figure 6.9: The percentage of water isolates attributable to each source with (upper panel) and without (lower panel) separating water birds from the 'other' sources in the analysis, given $\boldsymbol{\pi}$ are estimated by the two types of model, the source baseline is ruminants, and the rurality effect is included in the attribution model. Each panel contains three outcome resulted from different combinations of (s, m). The associated median for each source is marked in red, besides the left side of each panel, which is the point estimate for each source.

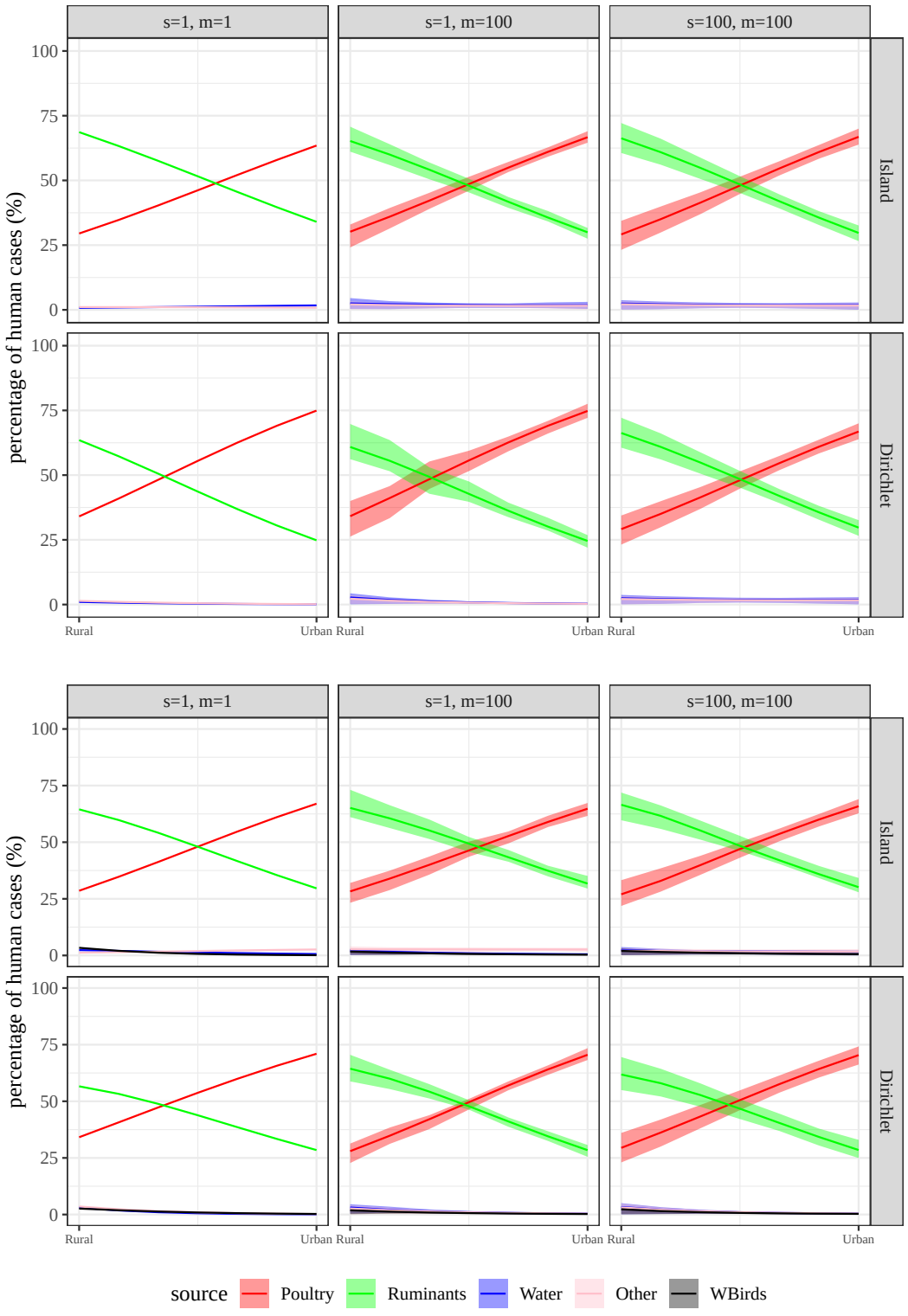


Figure 6.10: Posterior human attribution before (upper panel) and after (lower panel) considering water birds in the analysis, given $\boldsymbol{\pi}$ are estimated by the two types of model, the source baseline is ruminants, and the rurality effect is included in the attribution model. Each panel contains three outcome resulted from different combinations of (s, m). Posterior samples are graphically described by box plots, except for the left side of each panel, which is a line chart connecting the point estimate for each source across the rurality levels.

due to more variation incorporated from the simulations of $\boldsymbol{\pi}$. Regarding human attribution, both types of model do not show significant differences in uncertainty between $s = 1$ and $s = 100$, given 100 iterations. Although a little wider range is observed for the Dirichlet model with $(s = 1, m = 100)$, when the source of water birds is not included in the analysis.

6.5 Conclusions

In this chapter, the models determining the source attribution of human campylobacteriosis follow a similar framework to what has been proposed in the previous chapter. However, the previous models have not yet treated water as a medium in the transmission pathway. Human infection may be caused directly or indirectly by water through, for example, drinking contaminated water or recreational exposure through activities such as swimming in surface water (Kapperud et al., 1993; Pearson et al., 1993; Waage et al., 1999). As the pathogen would not grow in water, but amplify in animal hosts, the role of water here is regarded as a vector (instead of as a reservoir), carrying the isolates and sequence types from faeces coming from these hosts including birds. To gain more insight into the role of water in the transmission process, the current approach expands it using the previous models in order for the probability of each source contaminating water to be estimated. Then, such inference about water is combined with the modelling to attribute sources of human infection.

As the current approach is associated with the models developed before, the inferences about water contamination and human illness are made from four perspectives: i) whether the sampling distribution of genotypes $\boldsymbol{\pi}$ are simulated by the asymmetric Island or the Dirichlet model; ii) whether water birds are an important source of *Campylobacter* in relation to water contamination; and iii) whether the source of water birds and the rurality effect have impacts on human infection.

When attributing hosts as sources of contamination of water (whether or not the rurality effect is considered), the asymmetric Island model suggests that water birds become the predominant source, with at least 60% of samples being attributed to this source, regardless of the number of simulations s for $\boldsymbol{\pi}$. Whereas, the Dirichlet model may work ineffectively as it is not able to differentiate between water birds and other sources, but with larger spread of attribution for these two sources after mixing over all $\boldsymbol{\pi}$.

For human attribution (whether or not the rurality effect is considered), the results for both types of model are very similar. Poultry and ruminants are always the dominant sources of infection, while the other sources only contribute less than 5% of illness for whether or not water birds are being considered as a source. Thus, including water birds in the attribution does not affect the final inference between both types of model

as only about 2% of human infections are attributed to this source. In addition, the uncertainty from simulations of $\boldsymbol{\pi}$ looks similar between both types of model, despite of fact that the Dirichlet model might attribute sources with a littler wider range due to sampling variation.

In conclusion, the uncertainty in $\boldsymbol{\pi}$ would be expected to be potentially smaller in the asymmetric Island model compared to the Dirichlet model due to the use of information from the allelic profiles. However, the variation appears to be very small in comparison to the latter model, which might indicate that the uncertainty from the asymmetric Island model is underestimated. On the other hand, the Dirichlet model does not infer the attribution as precisely as the asymmetric Island model owing to inefficient identification of sources attributed to water contamination. Even though it estimates attribution of water contamination with more uncertainty, it still provides similar human case attributions as what the asymmetric Island model suggests. Thus, this approach provides almost the same human attribution as illustrated in previous chapters.

In addition, water birds are revealed to be an important source of contamination of surface water. This deepens our understanding of the role of water birds compared to other animals as a source of contamination of surface water and as a source of human infection. Some recent studies also suggested that many types isolated from water samples are related to wild birds, followed with a few types attributable to poultry and ruminants (Mughini-Gras et al., 2016; Shrestha et al., 2019). The authors of these studies used the Wilson model (introduced in Chapter 3.2) to attribute types found in water samples to putative sources of origin. This model estimates the attribution probabilities $\boldsymbol{F}$ with an assumption that the observations are multinomial distributed, whereas the asymmetric Island model transforms $\boldsymbol{F}$ to be on a logit scale with normal priors on the regression coefficients. Despite the different way to estimate $\boldsymbol{F}$, the evolutionary parameterisation of these two models is identical. This means that the final inference about source attribution from the asymmetric Island model will be the same as the Wilson model.

Chapter 7

Towards a new approach for source attribution

The models developed in Chapter 4 use multinomial distributions to depict source and human data in the principles of Bayesian inference, and are under the assumption that the probability of observing a genotype on human cases arising from a source is same as the probability of isolating the type from the source. In other words, the probability π_{ij} is commonly used in the likelihood (4.1) for sources and the likelihood (4.8) for humans, which brings an advantage of computational convenience. However, such an assumption is unlikely to be true in reality, as it is believed that the chance of typing a genotype from source or human isolates may be unequal.

Like all bacteria, *Campylobacter* evolves and adapts to different environments, including different hosts, leading to different degrees of prevalence of genotypes isolated from animals and humans (Dearlove et al., 2016). This implies that the probability of observing genotypes from human cases defined in Equation (4.3) is not valid. This problem was circumvented in Chapter 5 by using an approximate Bayesian method, cutting the probability linked between source and human data. However, to retain a formal Bayesian framework, this chapter aims to generalise the models from Chapter 4, developing an approach to describe the relationship of genotype prevalence between the two types of isolates.

This chapter contains five sections. Section 7.1 depicts how the models are generalised, and how π_{ij} is distinguished using a newly developed Bayesian method to account for the difference between sources and humans. Section 7.2 illustrates these models using the rurality effect as the only variable and outlines the applied MCMC algorithms. Then, the final inference about human attribution and the investigation of π_{ij} for sources and humans are presented in Section 7.3. Lastly, convergence diagnostics and conclusions are included in Sections 7.4 and 7.5.

7.1 Model structure

7.1.1 Likelihoods for data

The method of modelling source and human data is the same as what has been demonstrated in Chapters 4–6. Both sorts of data that are comprised of the frequency of each genotype found on each category are assumed to arise from a multinomial distribution, but the probability of observing genotypes on source isolates $p(\text{ST}_i \mid \text{source}_j)$ may differ from that for humans. In order for the two probabilities to be distinguished, π_{ij} is replaced by $\pi_{ij}^{\mathbb{S}}$ and $\pi_{ij}^{\mathbb{H}}$ for the data $\boldsymbol{X}$ and $\boldsymbol{Y}$.

The new version of the likelihood for sources based on the likelihood (4.1) becomes,

$$L(\boldsymbol{X}; \boldsymbol{\pi}^{\mathbb{S}}) \propto \prod_{i=1}^{I} \prod_{j=1}^{J} {\pi_{ij}^{\mathbb{S}}}^{x_{ij}}, \quad x_{ij} \in \{0, 1, \cdots, n_j\}, \quad 0 < \pi_{ij}^{\mathbb{S}} < 1.$$

The prior of the original π_{ij} for data $\boldsymbol{X}$ is retained here, which follows a Dirichlet distribution with the parameter α_{ij}^{p}. As a Dirichlet distribution is related to a gamma distribution (see the derivation in Appendix A.1), $\pi_{ij}^{\mathbb{S}}$ is equivalent to the form,

$$\pi_{ij}^{\mathbb{S}} = \frac{\eta_{ij}^{\mathbb{S}}}{\sum_{k=1}^{I} \eta_{kj}^{\mathbb{S}}}, \quad 0 < \eta_{ij}^{\mathbb{S}} < \infty, \tag{7.1}$$

where $\eta_{ij}^{\mathbb{S}}$ is gamma distributed with the pdf specified as,

$$f(\boldsymbol{\eta}^{\mathbb{S}}) \propto \prod_{i=1}^{I} \prod_{j=1}^{J} {\eta_{ij}^{\mathbb{S}}}^{(\alpha_{ij}^{g} - 1)} \exp\left(-\eta_{ij}^{\mathbb{S}}\right), \quad \alpha_{ij}^{g} > 0.$$

Thus, the likelihood for sources can be rewritten as,

$$L(\boldsymbol{X}; \boldsymbol{\eta}^{\mathbb{S}}) \propto \prod_{i=1}^{I} \prod_{j=1}^{J} \left(\frac{\eta_{ij}^{\mathbb{S}}}{\sum_{k=1}^{I} \eta_{kj}^{\mathbb{S}}} \right)^{x_{ij}}.$$

The likelihood for humans (4.2) is also altered as $p(\text{ST}_i \mid \text{source}_j)$ in the definition of $\hat{\boldsymbol{\pi}}$ (4.3) is no longer valid. Instead, the probability that human cases are of a genotype arising from a source is assumed to be different to the type observed from the source, namely, $p^{\mathbb{H}}(\text{ST}_i \mid \text{source}_j) \neq p^{\mathbb{S}}(\text{ST}_i \mid \text{source}_j)$, and thus it leads to,

$$p(\text{ST}_i \text{ typed from human cases}) = \sum_{j=1}^{J} p^{\mathbb{H}}(\text{ST}_i \mid \text{source}_j) p(\text{source}_j)$$

such that the likelihood for humans turns out to be,

$$L(\boldsymbol{Y};\boldsymbol{\pi}^{\mathbb{H}},\boldsymbol{F}) \propto \prod_{i=1}^{I}\Big(\sum_{j=1}^{J}\pi_{ij}^{\mathbb{H}}F_j\Big)^{y_i}, \quad y_i \in \{0,1,\ldots\}, \quad 0<\pi_{ij}^{\mathbb{H}}<1. \tag{7.2}$$

In the same way as $\pi_{ij}^{\mathbb{S}}$ is defined, it is useful to reparameterise $\pi_{ij}^{\mathbb{H}}$ in terms of a positive variable $\eta_{ij}^{\mathbb{H}} : \pi_{ij}^{\mathbb{H}} = \frac{\eta_{ij}^{\mathbb{H}}}{\sum \eta_{kj}^{\mathbb{H}}}$, which is a function of $\eta_{ij}^{\mathbb{S}}$ defining the relationship between cases and sources (see Chapter 7.1.2 for further details). In addition, epidemiological variables can also be incorporated in the modelling for $\boldsymbol{F}$ as has been done before. The likelihood from the perspective of individuals is hence expressed as,

$$L(\boldsymbol{Y};\boldsymbol{\pi}^{\mathbb{H}},\boldsymbol{F}) \propto \prod_{h=1}^{H}\sum_{j=1}^{J}\pi_{i[h]j}^{\mathbb{H}}F_{hj},$$

where the index $i[h]j$ refers to a genotype i from the h^{th} case corresponding to source j.

7.1.2 The relationship of genotyping between sources and humans

As addressed before, some types are found more (or less) in human cases than the frequencies found in sources. An obvious example is ST-474 and ST-2381. The former is observed more often in humans than in any sources, while the latter is only found in water and wild birds. However, it is expected that $\boldsymbol{\pi}^{\mathbb{S}}$ are similar to $\boldsymbol{\pi}^{\mathbb{H}}$ for most of types that are rarely observed in sources and humans, with just a small number of them potentially being different. Thus, in order for different degrees of genotype prevalence between humans and sources to be described, a discrepancy variable O that follows a Cauchy distribution with parameters (μ_0, γ_0) is introduced, representing the link between $\boldsymbol{\eta}^{\mathbb{S}}$ and $\boldsymbol{\eta}^{\mathbb{H}}$.

The realisation o_{ij} expresses the link of typing a genotype i from isolates between source j and human cases. A convenient way of computation involving this variable is to use an exponential scale. Hence, the relationship is defined as $\eta_{ij}^{\mathbb{H}} = \eta_{ij}^{\mathbb{S}} \times e^{o_{ij}}$, where the pdf for O is of the form,

$$L(O;\gamma_0) = \prod_{i=1}^{I}\prod_{j=1}^{J}\frac{1}{\pi\gamma_0\left\{1+\left(\frac{o_{ij}}{\gamma_0}\right)^2\right\}}, \quad \gamma_0>0, \quad -\infty<o_{ij}<\infty,$$

given the scale parameter γ_0 specified as half of the IQR of the distribution. Note that the distribution typically takes values very close to 0 (as the location parameter $\mu_0 = 0$) when the value of γ_0 is small, which is a property of the Cauchy distribution due to its very heavy tails. In this case, it indicates that types where $\boldsymbol{\eta}^{\mathbb{H}}$ and $\boldsymbol{\eta}^{\mathbb{S}}$ are

almost identical, but at the same time it also accommodates 'outlying' types where there is a big difference. In the results that follow, the value of IQR will set to be 0.01 as a starting point in order for the shape of the distribution for types to be highly concentrated at 0, allowing for only a few types to have large values of O.

Therefore, given the defined relationship between $\eta_{ij}^{\mathbb{H}}$ and $\eta_{ij}^{\mathbb{S}}$, the probability $\pi_{ij}^{\mathbb{H}}$ that a type i found on human isolates arises from source j used in the likelihood (7.2) can also be specified as,

$$\begin{aligned}\pi_{ij}^{\mathbb{H}} &= \frac{\eta_{ij}^{\mathbb{H}}}{\sum_{k=1}^{I}\eta_{kj}^{\mathbb{H}}} \\ &= \frac{\eta_{ij}^{\mathbb{S}}\exp\left(o_{ij}\right)}{\sum_{k=1}^{I}\eta_{kj}^{\mathbb{S}}\exp\left(o_{kj}\right)}.\end{aligned} \tag{7.3}$$

This model will become the model with a common π_{ij} in Chapter 4 when $\boldsymbol{\pi}^{\mathbb{H}} = \boldsymbol{\pi}^{\mathbb{S}}$. It means that there is no difference between sources and human cases in relation to genotyping, i.e. $\boldsymbol{o} = \mathbf{0}$, and so $\boldsymbol{\eta}^{\mathbb{H}} = \boldsymbol{\eta}^{\mathbb{S}}$.

7.1.3 Model fitting with the rurality effect

In this chapter, the rurality effect will be considered as the only variable for convenience of demonstration. Given four source categories are of concern in the analysis, and the rurality variable c is numeric, the attribution model is of the form,

$$F_{hj} = \frac{\exp\left(f_{hj}\right)}{1+\sum_{l=1}^{3}\exp\left(f_{hl}\right)}, \tag{7.4}$$

where

$$f_{hj} = \beta_{0j} + \beta_{1j}c_h, \quad j = 1,2,3,$$

and $f_{h4} = 0$, given $j = 4$ is the source of ruminants and is treated as the source baseline.

As the model fitting for variables has been demonstrated in the previous chapters, for more details about estimation of F_{hj} using a numeric variable, see Chapters 4.2.2 and 5.2.

7.2 MCMC algorithms

The algorithms applied in this chapter are again the Metropoilis-Hastings sampler. The procedure of MCMC simulation is almost identical whether or not the link variable O is considered in the modelling. When the link effect O is random, i.e. $\boldsymbol{\pi}^{\mathbb{S}} \neq \boldsymbol{\pi}^{\mathbb{H}}$, the variable has to be updated in the algorithm. By contrast, the model is similar to the

one developed in Chapter 4 when $O = 0$, meaning that the equal probability of typing genotypes is assumed between source and human isolates.

The steps of running a MCMC chain are outlined in Algorithm 6. At first, the parameters required to update are initialised as below:

1. The starting points of $\eta_{ij}^{\mathbb{S}}$ are sampled from a gamma distribution, given $(\alpha^g, \beta^g) = (1, 1)$, where $i = 1, \ldots, 377$, $j = 1, \ldots, 4$.

2. The starting points of β_p are sampled from a standard normal distribution, where $p = 1, \ldots, P$, and P denotes the total number of regression parameters $\boldsymbol{\beta}$ required in the model of $\boldsymbol{F}$. In this case, $P = 6$ as the rurality variable c is considered, given $J = 4$.

3. When the link effect O is considered to be random, the starting points of o_{ij} are drawn from a Cauchy distribution, given $(\mu_0, \gamma_0) = (0, \frac{\text{IQR}}{2})$, where IQR=0.01.

Then, the corresponding $\boldsymbol{F}^{(0)}$, $\pi_{ij}^{\mathbb{S}\,(0)}$ and $\pi_{ij}^{\mathbb{H}(0)}$ are obtained using equations (7.4), (7.1) and (7.3).

Let the chain iterate M times starting from line 5 of Algorithm 6. To update the parameters with two dimensions $(I \times J)$ one at a time, a random order of permutations P_I of $\{1, \ldots, 377\}$ and P_J of $\{1, 2, 3, 4\}$ is sampled. When the link effect O is random, o_{ij} and $\eta_{ij}^{\mathbb{S}}$ are first updated at the same time as they are likely to be quite correlated. For each source j, values of $\eta_{ij}^{\mathbb{S}\,*}$ are newly proposed from a log-normal distribution centred on the current value of $\eta_{ij}^{\mathbb{S}}$ with the standard deviation 1, and o_{ij}^* is updated using $\text{N}(o_{ij}, 0.05)$. The corresponding proposals of $\pi_{ij}^{\mathbb{S}\,*}$ and $\pi_{ij}^{\mathbb{H}*}$ are again obtained using equations (7.1) and (7.3). Then, a random sample u_1 is drawn from U(0,1) as a threshold to see if the proposals are accepted after evaluating the acceptance probability $\mathcal{A}_1$ in line 14. Last, the values of $\boldsymbol{\eta}^{\mathbb{S}*}$, $\boldsymbol{\pi}^{\mathbb{S}*}$, $\boldsymbol{o}^*$ and $\boldsymbol{\pi}^{\mathbb{H}*}$ are accepted and regarded as the next state $m+1$ if $u_1 < \mathcal{A}_1$; otherwise, the current state m of these parameters are copied to be the new one.

The acceptance probability $\mathcal{A}_1$ is comprised of five terms: i) the likelihood ratio for source data $\boldsymbol{X}$; ii) the likelihood ratio for human data $\boldsymbol{Y}$; iii) the proposal ratio for $\boldsymbol{\eta}$; iv) the prior ratio for $\boldsymbol{\eta}$; and v) the prior ratio for $\boldsymbol{o}$. In the programming, a logarithm

Algorithm 6 Metropolis-Hastings algorithm with or without the link effect O

Draw initial ${\eta_{ij}^{\mathbb{S}}}^{(0)}$ from G(1,1), $i = 1, \ldots, 377$, $j = 1, 2, 3, 4$
Draw initial ${\beta_p}^{(0)}$ from N(0,1), $p = 1, \ldots, 6$
Draw initial $o_{ij}^{(0)}$ from Cauchy(0, $\gamma_0 = \frac{0.01}{2}$) ▷ If O is random
Find ${\pi_{ij}^{\mathbb{S}}}^{(0)}$, $\pi_{ij}^{\mathbb{H}(0)}$ and $\boldsymbol{F}^{(0)}$ through Equation (7.1), (7.3) and (7.4)
for $m = 0, \ldots, M-1$ **do**
    Sample a random order of permutations P_I of $\{1, \ldots, 377\}$ and P_J of $\{1, 2, 3, 4\}$
    for $j \in P_J$ **do**
        for $i \in P_I$ **do**
            Propose ${\eta_{ij}^{\mathbb{S}}}^*$ sampling from lognormal(${\eta_{ij}^{\mathbb{S}}}^{(m)}, 1$)
            Propose o_{ij}^* sampling from N($o_{ij}^{(m)}, 0.05$) ▷ If O is random
            Compute ${\pi_{ij}^{\mathbb{S}}}^*$ and $\pi_{ij}^{\mathbb{H}^*}$ through Equation (7.1) and (7.3)
            Generate $u_1 \sim U(0, 1)$
            Calculate $\mathcal{A}_1\left(\boldsymbol{\eta}^{\mathbb{S}(m)}, \boldsymbol{\eta}^{\mathbb{S}*}\right) = \min\{1, \psi_1\}$, where

$$\psi_1 = \frac{L\left(\boldsymbol{X}; \boldsymbol{\pi}^{\mathbb{S}*}\right)}{L\left(\boldsymbol{X}; \boldsymbol{\pi}^{\mathbb{S}(m)}\right)} \frac{L\left(\boldsymbol{Y}; \boldsymbol{\pi}^{\mathbb{H}*}, \boldsymbol{\beta}\right)}{L\left(\boldsymbol{Y}; \boldsymbol{\pi}^{\mathbb{H}(m)}, \boldsymbol{\beta}\right)} \frac{Q_\eta\left(\boldsymbol{\eta}^{\mathbb{S}(m)} \mid \boldsymbol{\eta}^{\mathbb{S}*}\right)}{Q_\eta\left(\boldsymbol{\eta}^{\mathbb{S}*} \mid \boldsymbol{\eta}^{\mathbb{S}(m)}\right)} \frac{\pi_\eta\left(\boldsymbol{\eta}^{\mathbb{S}*}\right)}{\pi_\eta\left(\boldsymbol{\eta}^{\mathbb{S}(m)}\right)}$$

            or, $\mathcal{A}_1\left(\boldsymbol{\eta}^{\mathbb{S}(m)}, \boldsymbol{\eta}^{\mathbb{S}*}, \boldsymbol{o}^{(m)}, \boldsymbol{o}^*\right) = \min\{1, \psi_1\}$, where ▷ If O is random

$$\psi_1 = \frac{L\left(\boldsymbol{X}; \boldsymbol{\pi}^{\mathbb{S}*}\right)}{L\left(\boldsymbol{X}; \boldsymbol{\pi}^{\mathbb{S}(m)}\right)} \frac{L\left(\boldsymbol{Y}; \boldsymbol{\pi}^{\mathbb{H}*}, \boldsymbol{\beta}\right)}{L\left(\boldsymbol{Y}; \boldsymbol{\pi}^{\mathbb{H}(m)}, \boldsymbol{\beta}\right)} \frac{Q_\eta\left(\boldsymbol{\eta}^{\mathbb{S}(m)} \mid \boldsymbol{\eta}^{\mathbb{S}*}\right)}{Q_\eta\left(\boldsymbol{\eta}^{\mathbb{S}*} \mid \boldsymbol{\eta}^{\mathbb{S}(m)}\right)} \frac{\pi_\eta\left(\boldsymbol{\eta}^{\mathbb{S}*}\right)}{\pi_\eta\left(\boldsymbol{\eta}^{\mathbb{S}(m)}\right)} \frac{\pi_c\left(\boldsymbol{o}^*\right)}{\pi_c\left(\boldsymbol{o}^{(m)}\right)}$$

            if $u_1 < \mathcal{A}_1$ **then**
                $\boldsymbol{\eta}^{\mathbb{S}(m+1)} \leftarrow \boldsymbol{\eta}^{\mathbb{S}*}$
                $\boldsymbol{\pi}^{\mathbb{S}(m+1)} \leftarrow \boldsymbol{\pi}^{\mathbb{S}*}$
                $\boldsymbol{o}^{(m+1)} \leftarrow \boldsymbol{o}^*$ ▷ If O is random
                $\boldsymbol{\pi}^{\mathbb{H}(m+1)} \leftarrow \boldsymbol{\pi}^{\mathbb{H}*}$
            else
                $\boldsymbol{\eta}^{\mathbb{S}(m+1)} \leftarrow \boldsymbol{\eta}^{\mathbb{S}(m)}$
                $\boldsymbol{\pi}^{\mathbb{S}(m+1)} \leftarrow \boldsymbol{\pi}^{\mathbb{S}(m)}$
                $\boldsymbol{o}^{(m+1)} \leftarrow \boldsymbol{o}^{(m)}$ ▷ If O is random
                $\boldsymbol{\pi}^{\mathbb{H}(m+1)} \leftarrow \boldsymbol{\pi}^{\mathbb{H}(m)}$
            end if
        end for
    end for

28: Draw a random sample of permutation P_P of $\{1, \ldots, 6\}$ ▷ If c is considered
29: **for** $p \in P_P$ **do**
30: Sample a proposal of $\boldsymbol{\beta}^*$ with $\beta_p^* \sim N(\beta_p^{(m)}, 1)$
31: Find the corresponding $\boldsymbol{F}^*$ using Equation (7.4)
32: Generate $u_2 \sim U(0, 1)$
33: Calculate $\mathcal{A}_2\left(\beta_p^{(m)}, \beta_p^*\right) = \min\{1, \psi_2\}$, where

$$\psi_2 = \frac{L\left(\boldsymbol{Y}; \boldsymbol{\pi}^{\mathbb{H}}, \beta_p^*\right)}{L\left(\boldsymbol{Y}; \boldsymbol{\pi}^{\mathbb{H}}, \beta_p^{(m)}\right)} \frac{\pi_\beta\left(\beta_p^*\right)}{\pi_\beta\left(\beta_p^{(m)}\right)}$$

34: **if** $u_2 < \mathcal{A}_2\left(\beta_p^{(m)}, \beta_p^*\right)$ **then**
35: $\beta_p^{(m+1)} \leftarrow \beta_p^*$
36: **else**
37: $\beta_p^{(m+1)} \leftarrow \beta_p^{(m)}$
38: **end if**
39: **end for**
40: **end for**

scale was used in order to simplify the calculation for these terms. Hence, they become,

$$\log \frac{L\left(\boldsymbol{X}; \boldsymbol{\pi}^{\mathbb{S}*}\right)}{L\left(\boldsymbol{X}; \boldsymbol{\pi}^{\mathbb{S}(m)}\right)} = \sum_{i=1}^{377} x_{ij} \left(\log \pi_{ij}^{\mathbb{S}\,*} - \log \pi_{ij}^{\mathbb{S}\,(m)} \right)$$

$$\log \frac{L\left(\boldsymbol{Y}; \boldsymbol{\pi}^{\mathbb{H}*}, \boldsymbol{\beta}\right)}{L\left(\boldsymbol{Y}; \boldsymbol{\pi}^{\mathbb{H}(m)}, \boldsymbol{\beta}\right)} = \sum_{h=1}^{1804} \left(\log \sum_{j=1}^{4} F_{hj} \pi_{i[h]j}^{\mathbb{H}\,*} - \log \sum_{j=1}^{4} F_{hj} \pi_{i[h]j}^{\mathbb{H}\,(m)} \right)$$

$$\log \frac{Q_\eta\left(\boldsymbol{\eta}^{\mathbb{S}(m)} \mid \boldsymbol{\eta}^{\mathbb{S}*}\right)}{Q_\eta\left(\boldsymbol{\eta}^{\mathbb{S}*} \mid \boldsymbol{\eta}^{\mathbb{S}(m)}\right)} = \sum_{i=1}^{377} \sum_{j=1}^{4} \left(\log \eta_{ij}^{\mathbb{S}\,*} - \log \eta_{ij}^{\mathbb{S}\,(m)} \right)$$

$$\log \frac{\pi_\eta\left(\boldsymbol{\eta}^{\mathbb{S}*}\right)}{\pi_\eta\left(\boldsymbol{\eta}^{\mathbb{S}(m)}\right)} = \sum_{i=1}^{377} \sum_{j=1}^{4} (\alpha_{ij}^g - 1) \left(\eta_{ij}^{\mathbb{S}\,(m)} - \eta_{ij}^{\mathbb{S}\,*} \right)$$

$$\log \frac{\pi_o\left(\boldsymbol{o}^*\right)}{\pi_o\left(\boldsymbol{o}^{(m)}\right)} = \sum_{i=1}^{377} \sum_{j=1}^{4} \log \left[1 + \left(\frac{o_{ij}^{(m)}}{0.005} \right)^2 \right] - \log \left[1 + \left(\frac{o_{ij}^*}{0.005} \right)^2 \right],$$

where F_{hj} in the second term is derived from Equation (7.4), given $\boldsymbol{\beta} = \{\beta_{01}, \ldots, \beta_{13}\}$.

Once these parameters are updated, the sampler continues updating the regression parameters β_p, $p = 1, \ldots, 6$, starting from line 28. These steps are similar as before. A random order from permutation P_P of $\{1, \ldots, 6\}$ is sampled in order for the elements of $\boldsymbol{\beta}$ to be proposed one at a time, and a random walk is also used to propose a move

of β_p. Next, a random sample u_2 is drawn from U(0,1) to see if it is smaller than the acceptance rate $\mathcal{A}_2$ in order to decide the values of the next state. This time, $\mathcal{A}_2$ only consists of the likelihood ratio for human data and the prior for $\boldsymbol{\beta}$. These terms in the logarithmic version are,

$$\log \frac{L\left(\boldsymbol{Y}; \boldsymbol{\pi}^{\mathbb{H}}, \beta_p^*\right)}{L\left(\boldsymbol{Y}; \boldsymbol{\pi}^{\mathbb{H}}, \beta_p^{(m)}\right)} = \sum_{h=1}^{1804} \left(\log \sum_{j=1}^{4} F_{hj}^* \pi_{i[h]j}^{\mathbb{H}} - \log \sum_{j=1}^{4} F_{hj}^{(m)} \pi_{i[h]j}^{\mathbb{H}} \right)$$

$$\log \frac{\pi_\beta\left(\beta_p^*\right)}{\pi_\beta\left(\beta_p^{(m)}\right)} = \frac{\beta_p^{(m)2} - \beta_p^{*2}}{2}.$$

Lastly, the algorithm yields a chain of $M = 5,000$ samples for each parameter, after disregarding the first 2,000 samples, and thinning by retaining only every 10^{th} sampled value.

When $O = 0$, there is no need to update O and hence lines 3, 10, 18 and 23 of the algorithm are ignored. The acceptance rate $\mathcal{A}_2$ is changed and specified in line 13 as the term of prior ratio for O is removed.

7.3 Results

This section will focus on the final inference about human attribution, given the rurality variable is included in the attribution model. The posterior percentage of human cases attributed to sources, given $\boldsymbol{\pi}^{\mathbb{H}} = \boldsymbol{\pi}^{\mathbb{S}}$, will be compared to not only that of $\boldsymbol{\pi}^{\mathbb{H}} \neq \boldsymbol{\pi}^{\mathbb{S}}$, but also that from Chapter 4. In addition, the relationship between $\boldsymbol{\pi}^{\mathbb{H}}$ and $\boldsymbol{\pi}^{\mathbb{S}}$ is also illustrated here along with the STs that have the most different effect linked between humans and sources.

7.3.1 Posterior proportion of cases attributable to sources of infection

When $O = 0$, the final attribution is expected to be identical to the Chapter 4 results. This is because, the common probability π_{ij} used in Chapter 4 has the same prior as the approach proposed here, with a gamma formulation adopted for $\eta_{ij}^{\mathbb{S}}$. Figure 7.1 confirms that the result on the left panel from Chapter 4 and the one on the right from Chapter 7 are almost the same, with a slightly difference due to sampling variation in the algorithm. However, as discussed before, results from Chapter 4 may not be valid due to model misspecification, leading to different final inference after increasing the amount of source data. This means that the approach in Chapter 4 is incorrect such that O should not be 0.

When the link variable O is random, Figure 7.2 presents a comparison of the final

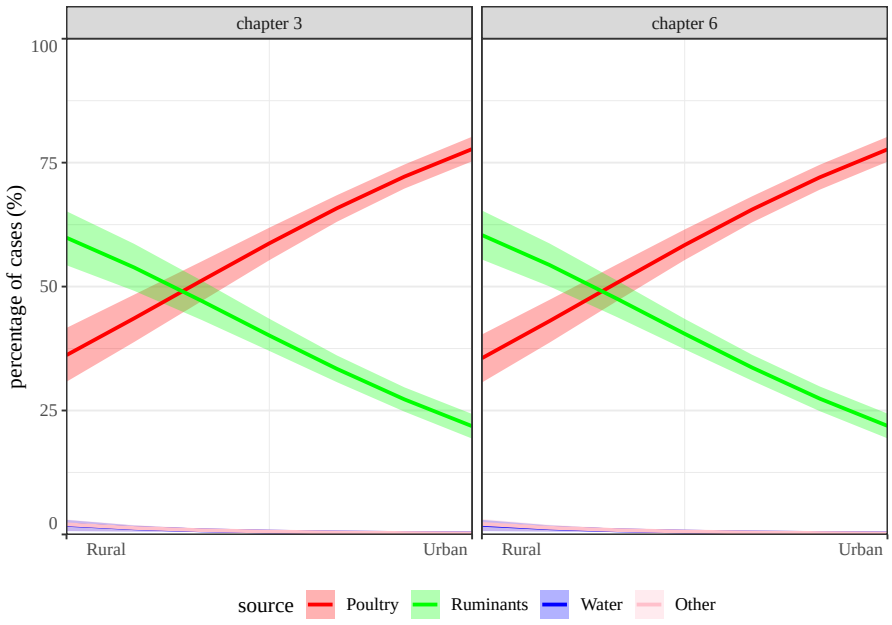


Figure 7.1: The percentage of human cases attributable to sources from rural to urban areas, with 80% credible intervals. Given 5,000 iterations have been obtained after the burin-in period of 2,000 and selecting every 10^{th} sample, the left panel is resulted from the models developed in Chapter 4, while the right one is based on the current approach with $\boldsymbol{\pi}^{\mathbb{S}} = \boldsymbol{\pi}^{\mathbb{H}}$.

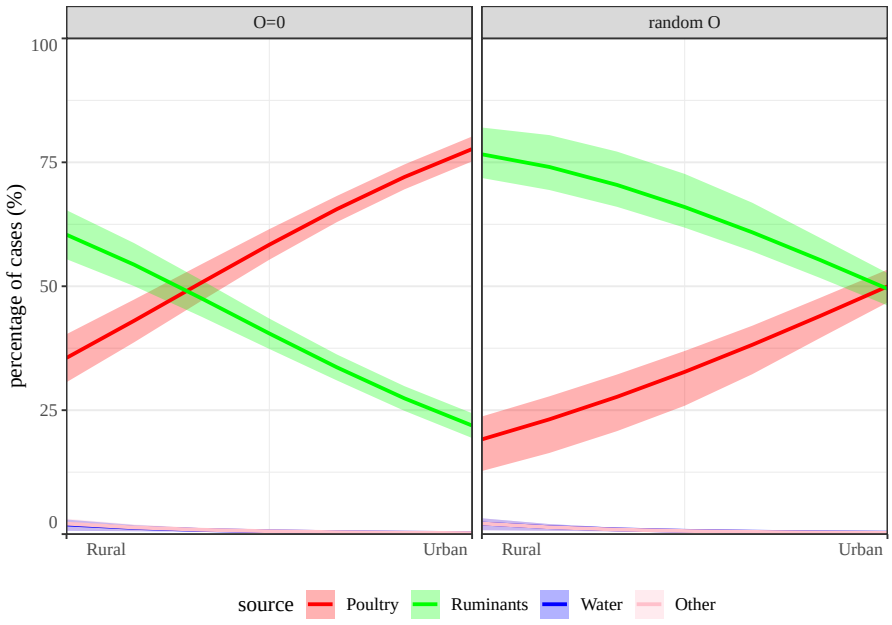


Figure 7.2: The percentage of human cases attributable to sources from rural to urban areas, with 80% credible intervals, using the current approach. The result on the left is based on $\boldsymbol{\pi}^{\mathbb{S}} = \boldsymbol{\pi}^{\mathbb{H}}$ ($O = 0$), while the one on the right results from $\boldsymbol{\pi}^{\mathbb{S}} \neq \boldsymbol{\pi}^{\mathbb{H}}$ (O is random).

inference between $O = 0$ and $O \neq 0$. After considering the random link effect to the probability between humans and sources, the result becomes unexpected and differs from the one on the left. It suggests that: i) when the cases live in highly remote areas, ruminants are responsible for more than 75% of cases, leading to much lower contribution from poultry (<25%); ii) the less rural areas the cases live, the more likely it is that their illnesses are equally attributed to poultry and ruminants; and iii) the location where the cases live is irrelevant to the infection caused by water and other sources (these two sources have a little impact on the disease as observed before). These findings except for iii) do not correspond with the results suggested from Chapters 5 and 6. This reflects a need on further investigations about posterior parameters $\boldsymbol{\pi}^{\mathbb{S}}$, $\boldsymbol{\pi}^{\mathbb{H}}$ and $\boldsymbol{o}$ to figure out reasons why they are so different from that of $\boldsymbol{o} = \mathbf{0}$.

7.3.2 Posterior probability of typing genotypes

Under the assumption that the prevalence of genotypes isolated from samples differs between humans and sources, Figure 7.3 displays the posterior probabilities $\boldsymbol{\pi}^{\mathbb{S}}$ and $\boldsymbol{\pi}^{\mathbb{H}}$ for STs that have the average probability larger than 0.02. The upper panel of the figure shows that only a few STs whose probability is on average higher than the threshold in each category, implying that these STs found on sources are potentially more prevalent than others. However, for human cases, the prevalent STs suggested from the lower panel of the figure are not exactly the same as the upper panel. This tells that the prevalence of genotypes between humans and sources is different. For example, the most frequent type isolated from poultry is ST-45. However, for humans, ST-474 arising from poultry is the most likely observed type, followed by ST-45.

The mean squared difference in the prevalence for each type between humans and sources is calculated. Figure 7.4 confirms that the prevalence of some genotypes differs between humans and sources. This graph illustrates the comparison of posterior mean of $\boldsymbol{\pi}$ for each type between humans and each of four sources, in which the top 3 types with the highest mean squared difference are labelled. When $\boldsymbol{\pi}^{\mathbb{S}}$ and $\boldsymbol{\pi}^{\mathbb{H}}$ are very close to each other, the points would sit on the diagonal. This means that the chance of observing a genotype from the source is similar to that from human cases who are of the type. This could link the infection may be simply due to the same source of population. From Figure 7.4, it is also observed that only a few STs may have different prevalence between humans and sources. For the types associated with poultry such as ST-474 labelled in the top panel, it accounts for about 35% of human cases, whilst it only accounts for fewer than 5% of poultry isolates. The panel for ruminants also gives three types, however, the difference is small. Approximately 10% of human cases are associated with ST-53, which may only account for 6% of ruminant isolates, for example. The prevalence of genotypes related to water and other samples seems to be

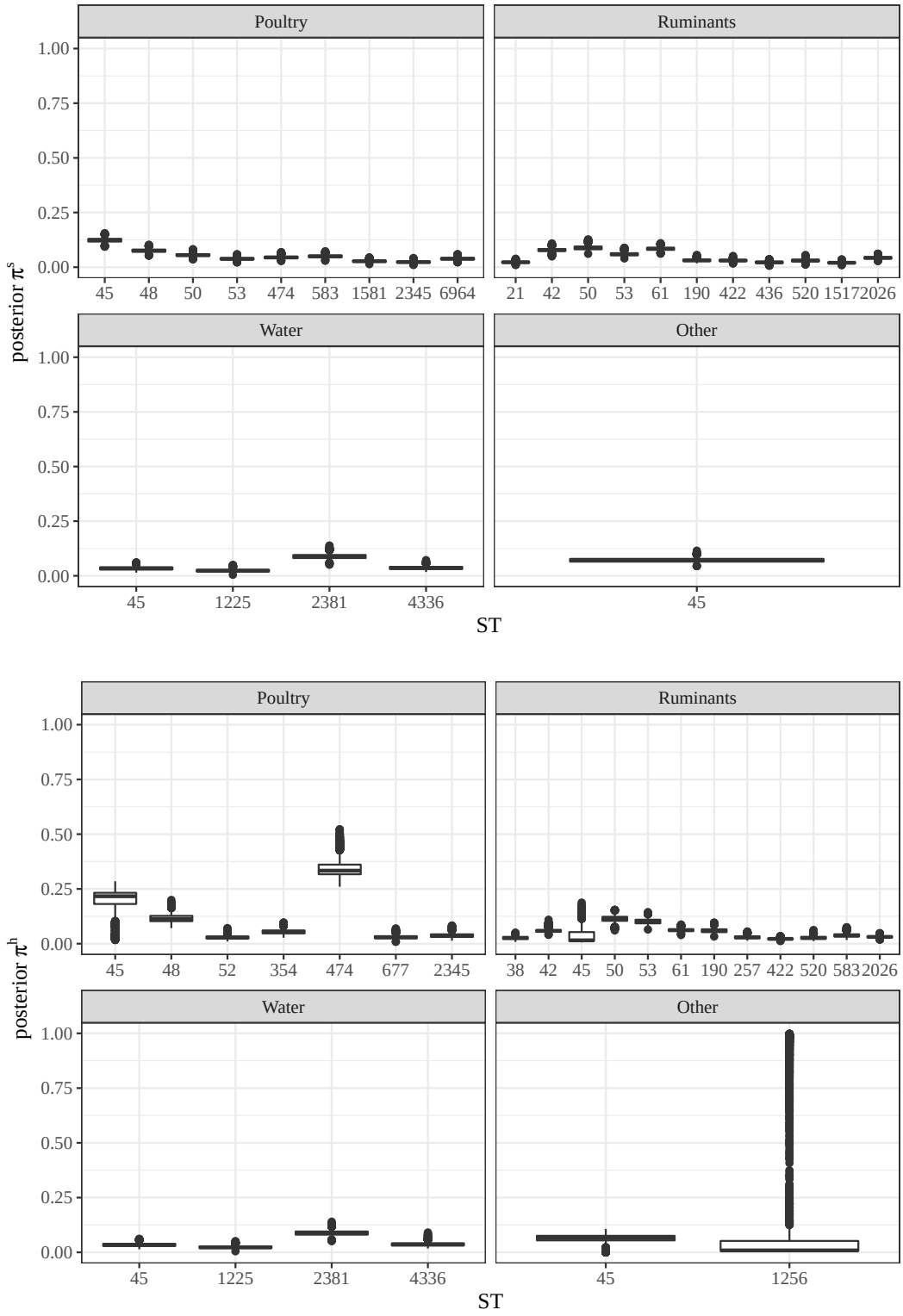


Figure 7.3: Posterior $\boldsymbol{\pi}^{\mathbb{S}}$ (upper panel) and $\boldsymbol{\pi}^{\mathbb{H}}$ (lower panel) for STs that have $\mu_{\boldsymbol{\pi}^{\mathbb{S}}} > 0.02$ and $\mu_{\boldsymbol{\pi}^{\mathbb{H}}} > 0.02$ for each source.

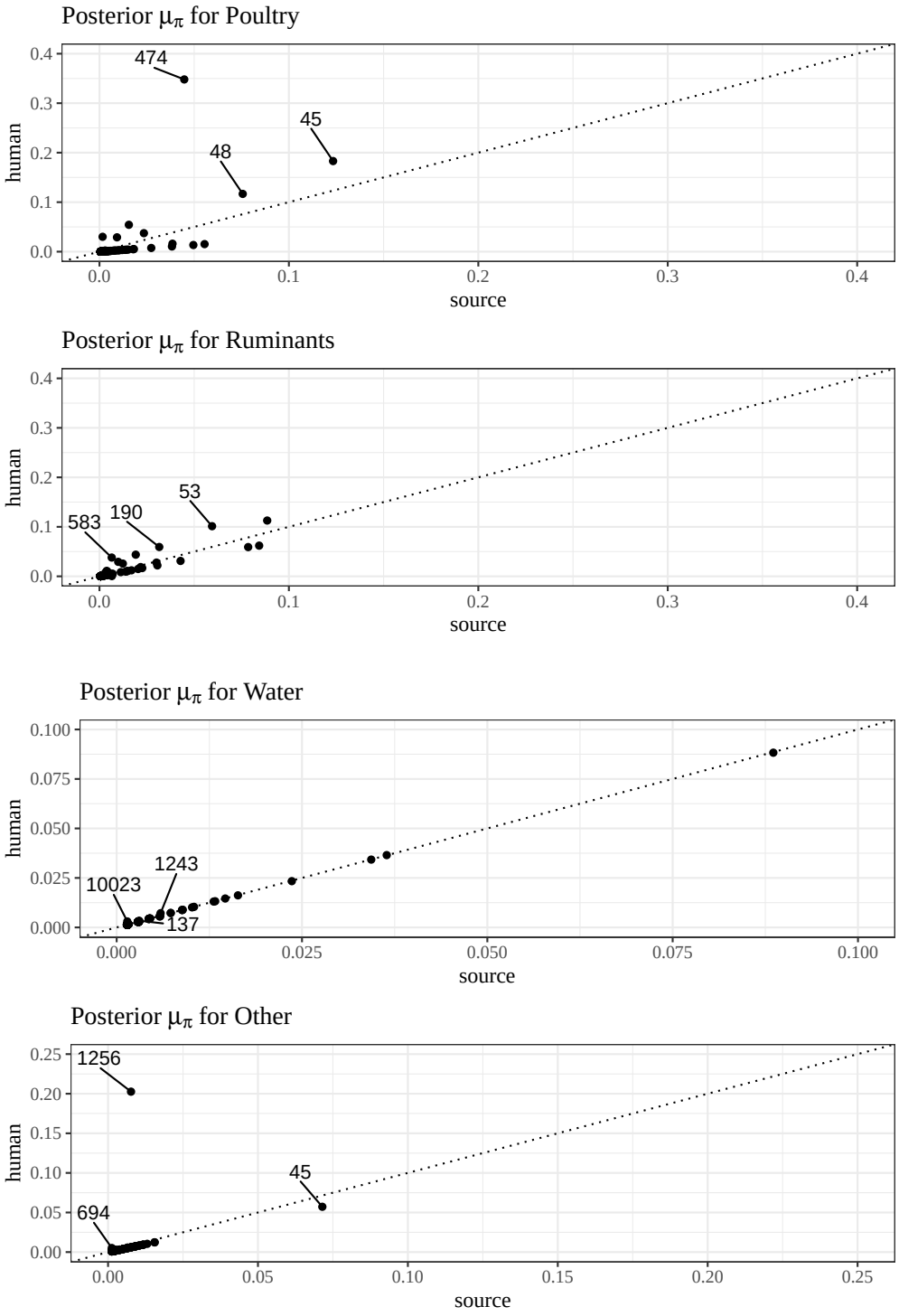


Figure 7.4: A comparison of the posterior expected probability of typing genotypes from isolates between human cases and each of the four sources. The labels in each category represent the STs with the top 3 highest mean squared difference between $\boldsymbol{\pi}^{\mathbb{S}}$ and $\boldsymbol{\pi}^{\mathbb{H}}$.

fairly similar to that for humans except for ST-1256, which is responsible for 20% of human cases associated with other sources, but only for $< 1.25\%$ of other isolates.

To assess how the link effect O works behind the relationship of genotype prevalence between humans and sources, the posterior median of O is evaluated for each ST across the four sources. In general, the posterior median of O for most of types is centred around 0 and ranged between (-0.1, 0.1). However, there are some STs (from poultry and ruminants) that have a relatively larger median than other STs. This is evident in the upper panel of Figure 7.5 that displays the density of posterior medians, disregarding the absolute value of median larger than 0.4. The lower panel of Figure 7.5 further displays the median and the associated 80% HPD intervals for those types that have $|\text{median}| \geq 0.4$. It is noted that some types have a significant link effect such as ST-474 and ST-677 for poultry, and ST-1115 for ruminants. The types for poultry in general have a small interval, meaning that it is 80% sure that the population median is very close to the estimated one. Similarly, small uncertainty about median for some ruminant-associated types is also observed, but not for ST-1115.

These types with a significant link effect for poultry and ruminants do not match up with those in Figure 7.4 having the most distinguishable prevalence. The difference in prevalence for humans and sources may be induced by a multiplicative effect on the gamma scale, this can be seen in Equation (7.3). While the link effect of ST-45 associated with poultry (with the median of 1.807 and a smaller interval) might not be as large in magnitude, it might just be because it is already highly prevalent in poultry and hence it does not need much of a multiplicative effect from $e^{1.807}$ to scale up to the prevalence in humans. In addition, types for water and other sources all seem to have small effects with $|\text{median}| < 0.4$. This might be because not only a small number of samples are observed among all source isolates, but also the contribution from these two sources to human infection is small such that there is a small impact on human cases in terms of genotype prevalence.

7.4 Diagnostics

7.4.1 Convergence for regression parameters

The trace plot for the regression parameters $\boldsymbol{\beta}$ considered in the fitted model is displayed in Figure 7.6 for three chains, given ruminants are the source baseline. The samples of the intercept and the slope of rurality for three sources are compared in a different manner of the link effect, i.e. $O = 0$ and $O \neq 0$.

When $\boldsymbol{\pi}^{\mathbb{H}}$ and $\boldsymbol{\pi}^{\mathbb{S}}$ are the same, the chains on the upper panel of Figure 7.6 for each parameter and each source are mixing fairly well and converging to the same values. Sequences for poultry tend to be centred around (0, 0.5) for the intercept and the slope,

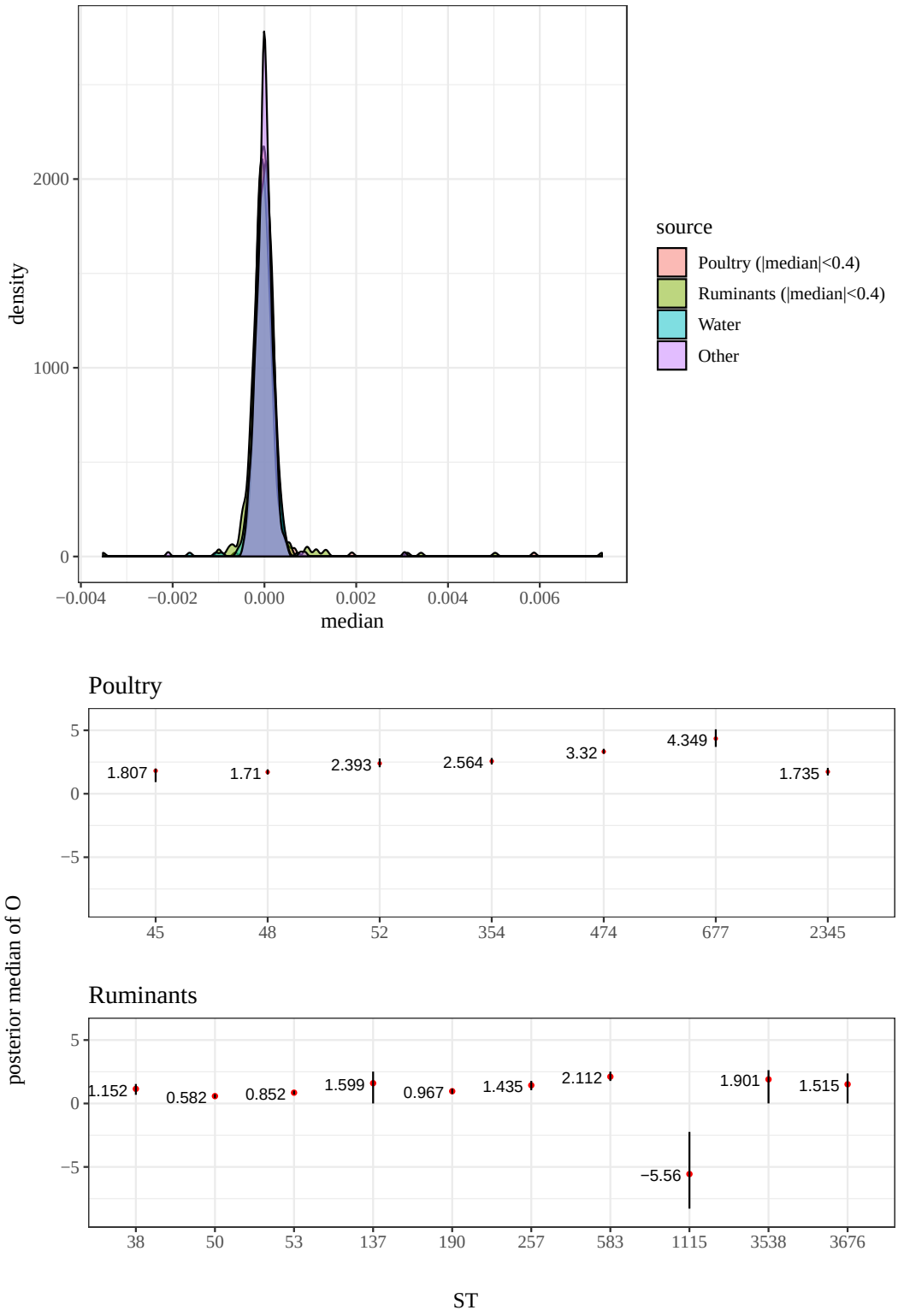
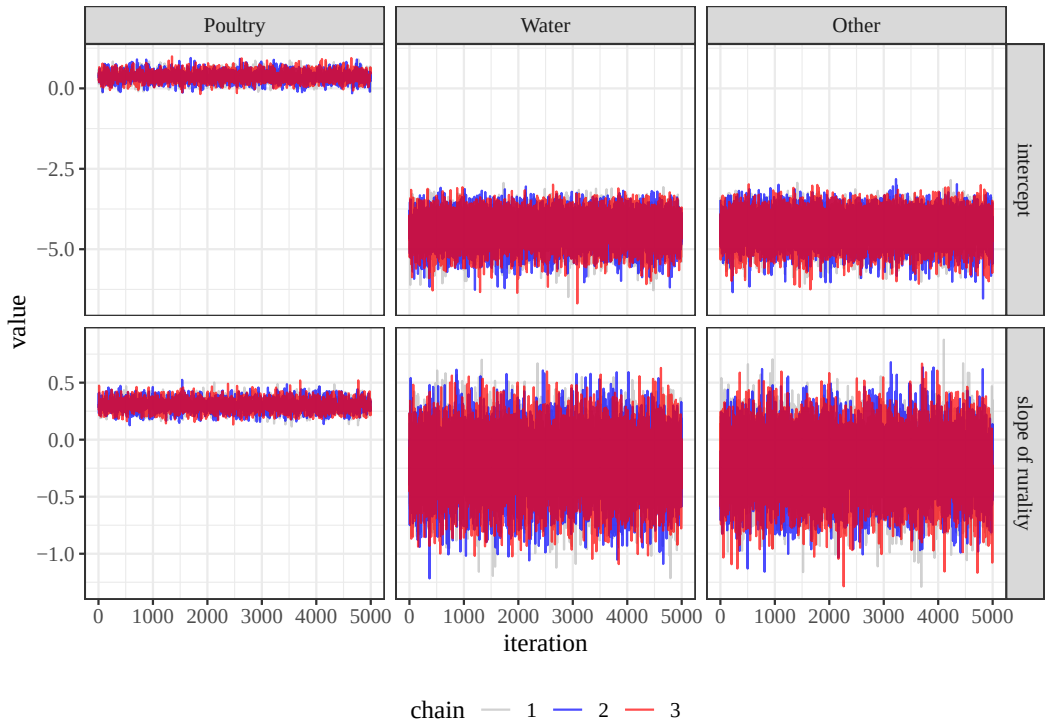
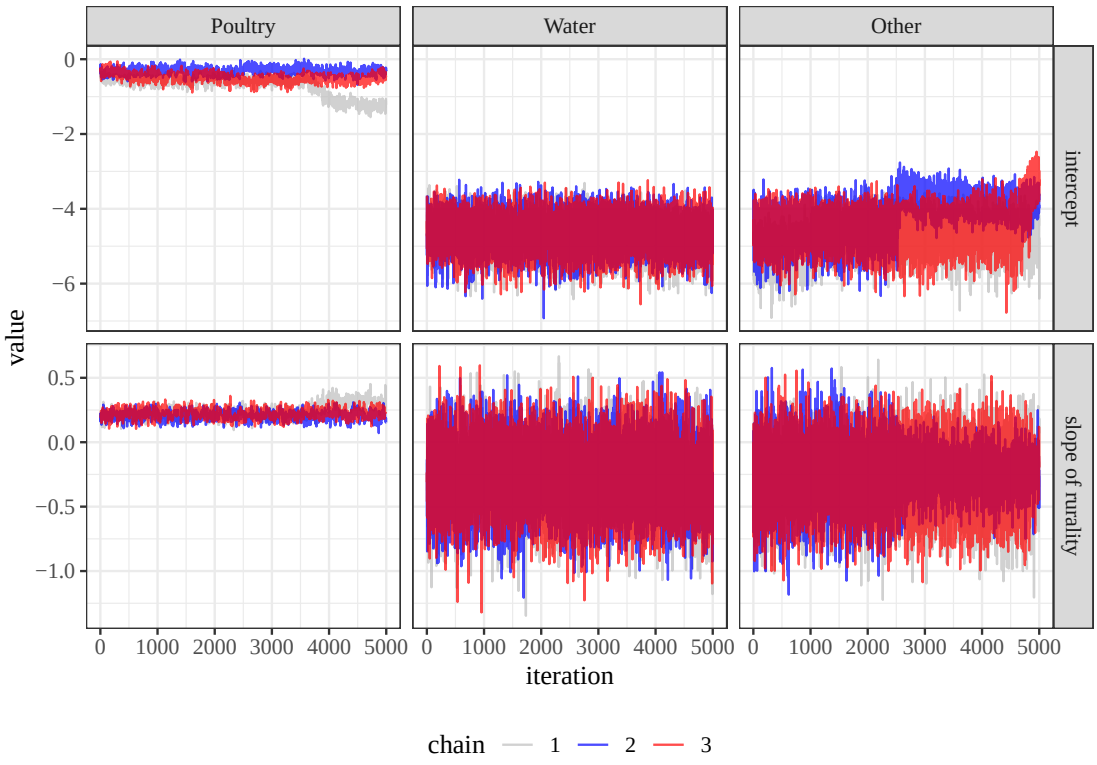


Figure 7.5: Posterior median of O for all types across four sources. The upper panel is the density plot for |medians| < 0.4. The lower panel shows the STs that have |median| ≥ 0.4 with the associated 80% HPD interval.



(a) $\boldsymbol{o} = \mathbf{0}$



(b) $\boldsymbol{o} \neq \mathbf{0}$

Figure 7.6: The trace plot with three chains for regression parameters $\boldsymbol{\beta}$ considered in the attribution model $\boldsymbol{F}$, given the baseline source is ruminants and the link variable O is (a) not considered, or (b) considered to be random.

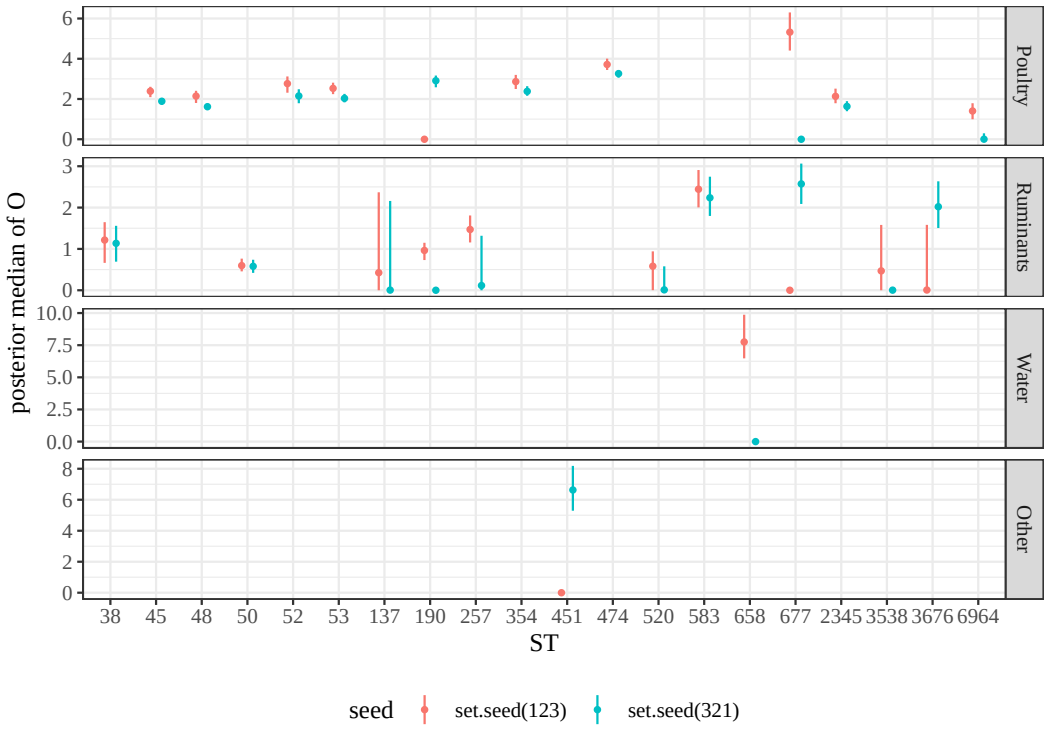


Figure 7.7: Posterior median of O and the associated 80% HPD interval for types that have |median| ≥ 0.4 from two different initial conditions of seed setting to the starting points of O.

while water and other sources have relatively larger values ranged between (-6,-3) for the intercept and (-1, 0.5) for the slope. When the link effect is introduced and the components $\boldsymbol{o}$ are updated by a random walk of $N(0, 0.05)$, the chains on the bottom of Figure 7.6 show that they are not converged well for poultry and other sources for both parameters. This is evident by looking at the last 2,000 iterations of the three chains, of which they are not centred around the same range of values. This suggests that the sampler does not mix and converge well.

7.4.2 Robustness for the link effect

To assess if the link variable provides consistent information, the MCMC algorithm given the link and rurality effects are considered has been re-run using two different initial conditions of seed setting for the components $\boldsymbol{o}$.

In general, most of types have a posterior median of O ranged between (-0.1, 0.1) as observed before. However, there are still some types with |median| ≥ 0.4. The posterior median of O for these types after setting two different seeds (set.seed(123) and set.seed(321)) to initialise the starting points of O are presented in Figures 7.7. It

is expected that the sampler with different initial conditions should not give different results about the random link effect. However, types suggested from the first seed setting are not exactly same as the second one. This implies that the performance of the random link effect may not be robust as the sampler does not converge well.

7.5 Conclusions

In this chapter, a novel approach of Bayesian source attribution model is developed. This model differs from those proposed in Chapter 4 in that it covers the modelling for the difference in the degree of genotype prevalence in sources and human cases. The probability of typing genotypes from isolates between these two types of samples was assumed to be equal before. Here it is assumed to be different as it is believed that some genotypes are more frequently observed on some sources than humans or vice versa. The new approach hence introduces a Cauchy variable O, describing the discrepancy effect on genotype prevalence.

Results about human attribution across the rurality levels show that the assumption about no relationship of genotype prevalence between sources and humans ($O = 0$) yields almost identical findings as what has obtained from Chapter 4 due to the same prior assumption with different forms of pdf, i.e. Dirichlet and gamma. When a random link effect O is introduced in the prevalence of genotypes, the final inference is not comparable with that of $O = 0$. It suggests that the chance of being infected due to ruminants or poultry is equal when cases live in main city centres. However, this outcome is not in concordance with other studies that have identified higher poultry-associated infection in urban areas (Mullner et al., 2010; Levesque et al., 2013; Marshall et al., 2016).

The relationship between $\boldsymbol{\pi}^{\mathbb{H}}$ and $\boldsymbol{\pi}^{\mathbb{S}}$ for each type in each source category is then investigated. It is found that only a few STs have distinguishable prevalence differing between sources and humans, particularly for types associated with poultry and ruminants. This may be because of the small number of observations of water and other sources. The assessment of posterior median of O further reveals that a highly prevalent type does not need to scale up its importance to human cases through the multiplicative effect defined for the relationship between $\boldsymbol{\pi}^{\mathbb{S}}$ and $\boldsymbol{\pi}^{\mathbb{H}}$. From the diagnostics, a noticeable problem of this model is that the sampler does not always explore the posterior very well. This is evident in the MCMC chains that are not well converged, and the dissimilar results of posterior median of O after changing the initial conditions.

Causes of the ineffective sampler could be many. Using a Cauchy variable to depict different degrees of genotype prevalence is presumably appropriate. The tails of the distribution represent only a few STs that are markedly different in prevalence between humans and sources as extreme values, while the majority of STs are centred at 0 as

their prevalence contributions do not change much. However, the model with Cauchy differences might be too complex to fit reliably. It seems that high dimensional data ($I \times J$) do not help to clarify how the genotype prevalence in sources relates to humans through the inference about the link effect O. In addition, it is possible that there may be strong correlation between regression parameters and the link effect, which would not help chain convergence.

To sum up, this chapter attempts to develop a source attribution model including statistical inference about different genotype prevalence between humans and sources. Although this approach does not succeed in capturing the disparity in prevalence for some types between the two types of samples, it highlights a need for genetic information to be considered in the modelling. A better understanding of which types are likely to have a very different prevalence may be required to enhance this approach, for example, expert opinions or use of some genetic models. This could be a possible direction for future research.

Chapter 8

Research findings and models: A discussion

8.1 Epidemiology of human campylobacteriosis

Human campylobacteriosis in New Zealand shows a spatial pattern in the estimated proportion of cases attributed to sources of infection, with urban cases being highly likely to be of poultry origin and rural cases more associated with ruminants. Food and water consumption is a reason to cause the different epidemiology of the disease between urban and rural areas. In comparison with urban cases who usually acquire their infection from undercooked chicken, rural residents tend to consume more ruminant products such as beef and raw milk than poultry, drink (especially untreated) water from their own private water supplies and swim in lakes and rivers (Spencer et al., 2012). When the age of cases is considered, rural children aged <5 years are more likely to be infected by ruminants than their urban counterparts. Besides the direct contact with farm animals, preschool children tend to be exposed to a variety of strains in the environment, such as exposure in playgrounds in parks with the faeces of wild birds (which are identified as carriers of *Campylobacter*), while having frequent hand-mouth contact (French et al., 2008).

While similar findings have been reported in other studies (Strachan et al., 2009; Müllner et al., 2010; Mughini Gras et al., 2012; Marshall et al., 2016), to what extent can our inferences from the New Zealand campylobacteriosis data be applied more widely? The issues around such a generalisation are not always straightforward. The difference in urban and rural areas is not true for the adult population, but only for young children in countries like the Netherlands and Scotland (Strachan et al., 2013). Trips to adjacent countries, particularly in summer (the peak season for campylobacteriosis), might result in a similar pattern of source attribution in adults living in urban and rural areas. Another explanation is that not each infected case has the same level of

illness owing to different immune-responses (Havelaar et al., 2009). Thus, it is possible that a person with mild symptoms does not seek medical service and hence is excluded from the surveillance system. Further, the finding of young children being at higher risk of infection could be a reflection of a possible diagnostic bias as they are subject to closer medical scrutiny and are characterised by higher healthcare system consumption. In other words, the chance that they seek healthcare service is higher than other age groups of people.

Regarding the model development, this research raises questions for the use of more complex models for source attribution. For conventional MLST data, the Dirichelt option, being a model dealing with the frequency of genotypes observed in isolates, provides considerable attribution inference. It is only inferior to the asymmetric Island model for rare genotypes isolated in samples. The increasing use of genetic data is likely to make complex models superior to the Dirichlet option. This solution may not be applicable to high-resolution genomic data such as wgMLST, whereas the asymmetric Island model may still work to associate infections with sources even if the genetic profile of each isolate is not matched perfectly due to a high level of detail in genetic information. However, is it always necessary to use complex models, even high-resolution genomic data, for source attribution per se? Indeed, it is certainly important to use models with more complexity for source attribution in relation to evolution, phylogenetics, outbreaks etc, but it might not be necessary to itself given the attributions based on conventional MLST and wgMLST are actually similar.

8.2 Wider applications of the models

This research is primarily about the development of new models that can be used for source attribution. The methods developed here integrate statistical epidemiology into genomics (genetic data in the form of MLST) in the flexible Bayesian frameworks. This work also has the potential for extension to whole genome methods. The use of the asymmetric Island model in the approaches would allow one to attribute sources to the infection even for wgMLST, for example.

While these models have been applied to examine the origins of *Campylobacter* infection in New Zealand, they can in principle be applied to a host of other zoonoses elsewhere, perhaps with appropriate tailoring to the application at hand. For example, the methods could be applied to salmonellosis, another kind of food poisoning caused by the bacteria *Salmonella*. This is because salmonellosis data may be equivalent to the data used here as they could contain subtypes (sub-classifications of strains) isolated from both human cases and the reservoirs. Therefore, the models developed in this thesis are able to apply to similar kinds of data sets, e.g. campylobacteriosis and

salmonellosis data collected in the Netherlands (Mughini Gras et al., 2012; Mughini-Gras et al., 2014).

In addition, there are only two traditional epidemiological covariates considered in this thesis (rurality and age), since these are the variables available in the Manawatu dataset. Nevertheless, the modelling frameworks developed are sufficiently broad to admit any relevant spatial or temporal covariates, either numeric or categorical. In other words, this thesis has been successful in the 'proof of concept' for this modelling approach. This encourages researchers to adopt the methods and apply them to data for comparable diseases from other countries.

Chapter 9

Conclusions and future work

9.1 Conclusions

This thesis has developed novel approaches to model source attribution, and they have been applied to the study of human campylobacteriosis. These models allow genetic and epidemiological data to be combined in the analysis, and hence they have the potential for widespread application. Their application to surveillance data from a sentinel site in the Manawatu region of New Zealand has certainly shed new light on the epidemiology of *Campylobacter* infection in this country.

A full Bayesian model is proposed to estimate the contribution of putative sources of infection in Chapter 4. Without considering any demographic variables, it is observed that poultry and ruminants are predominant sources for the infection. After adjusting the model with the rurality effect, cases are suggested to be more associated with poultry when living in urban areas, whilst they are more ruminant-associated when residing in rural areas. However, this model is mis-specified as the final attribution differs after artificially increasing the amount of source data. The relative amount of data from source and human data should not greatly affect the posterior attribution if the assumption of equal prevalence of genotypes isolated from samples between humans and sources is true. This means that this faulty assumption should be removed from the modelling.

As the assumption of a common $\boldsymbol{\pi}$ between sources and humans is not true, Chapter 5 demonstrates newly developed Bayesian models, introducing prior knowledge about the sampling distribution of genotypes observed on sources. This model framework differs from the previous one, in that the simulation of posterior attribution probability is not based on the joint likelihood for sources and human cases. Estimation of the attribution probability may be found by optimising the human likelihood, using a Metropolis-Hastings algorithm with suitable priors on the attribution probability. Given the rurality variable is considered in the model, the final inference is the

same as what has been suggested before: rural cases are more likely to be of ruminant origin, whilst urban cases are more likely to have originated from poultry. In addition, before the inference about human attribution, the sampling distribution of genotypes is estimated by two different types of model: the asymmetric Island and Dirichlet models. When data provide sufficient information (i.e. genotypes that are frequently observed), it shows that the Dirichlet model gives comparable final attribution with that of the asymmetric Island model. Nonetheless, the latter model outperforms the former for genotypes that are infrequently observed due to the estimation of genetic evolution between isolates.

The convergence and robustness of models developed in Chapter 5 are also confirmed. When the time of intervention in the poultry industry is included in the analysis, a dramatic reduction in poultry attributed to urban cases is observed, which is consistent with the literature (Sears et al., 2011). Meanwhile, the higher number of ruminant-associated cases in rural areas highlights a need for public health interventions.

From the above two chapters (Chapters 4 and 5), it is confirmed that poultry and ruminants are by far the most attributable sources for *Campylobacter* infection, while only a small contribution comes from water and other sources. These approaches treat water as a source of infection. However, the role that water plays in a transmission pathway can be more than as an origin. Water differs from animal reservoirs in that it is not an amplification vessel, but a vehicle for transmitting the pathogen. This means that water could be not only an end point, but also a medium contaminated by human waste or faeces from animals colonised by the bacteria, via a discharge of unprocessed water, runoff from farm fields, or water birds. This presents risks of exposure to recreational and drinking water when they are contaminated. Thus, to expand the role of water in a transmission route, it is motivated to propose other models in this regard.

Therefore, Chapter 6 utilises the models from Chapter 5 to estimate the probability of sources contaminating water. The results are then included in a novel approach to infer the final attribution for human cases. Originally, four source categories (i.e. poultry, ruminants, water and others) are of concern in the analysis. However, an additional source – water birds – is also included in the modelling here, considering they may be an intermediate host, spreading the pathogen to animals some distance away through water (Mullner et al., 2009a,b; Mughini Gras et al., 2012; Mughini-Gras et al., 2016). The inference about water attribution that resulted from the asymmetric Island model shows that water birds may be the major cause of the contamination, while the Dirichlet model suggests that both water birds and other sources are equally important. However, introducing the source of water birds in the modelling has no

impact on the final human attribution. Similar to the lesser contribution of water and other sources to the infection, the effect of water birds is uncommon. Further, the uncertainty in the simulation of $\boldsymbol{\pi}$ for the asymmetric Island model is smaller than for the Dirichlet model. The first model may be underestimating the variation of mixing over all $\boldsymbol{\pi}$ in the algorithm due to the use of information from the allelic profiles.

As addressed before, the full Bayesian model proposed in Chapter 4 has found to be mis-specified due to the inappropriate assumption about the prevalence of genotypes. One way to improve this model was to use an approximate Bayesian approach, which is not in a full Bayesian context, as demonstrated in Chapter 5. In order for the full Bayesian framework to be retained, Chapter 7 generalised the model from Chapter 4, in which the relationship of observing genotypes differing between human cases and sources is considered, using a Cauchy distribution to describe the different degrees of genotype prevalence.

In comparison to the results from Chapter 4, this generalisation yields identical final inference when there is no association between genotype prevalence and the type of samples (humans and sources). This is because these two approaches use the same prior assumption with different forms of pdf (Dirichlet and gamma). However, when the association is assumed, i.e. $O \neq 0$, the result from the generalised approach is unexpected: not only are urban cases highly poultry-associated, but also highly ruminant-associated. This is in disagreement with other studies that observe that infection in urban areas tends to be poultry rather than ruminant associated (Mullner et al., 2010; Levesque et al., 2013; Marshall et al., 2016). After investigating the relationship of genotype prevalence between humans and sources and the Cauchy link effect, it is noted that the model defining the relationship with Cauchy differences might be too heavily parameterised to fit reliably. Types with distinguishable prevalence differing between sources and humans are not consistently identified through the magnitude of the link effect. In addition, the sampler is found to be ineffective as it does not explore the posterior well and hence the MCMC chains are not converged well.

In conclusion, new approaches have been proposed in this thesis with a focus on Bayesian principles, and have been applied to data combining the demographic variables with the molecular information from both individuals and sources. It is observed that the models from Chapters 5 and 6 perform very well. They can identify not only the major sources attributed to human cases, but also the origins of water contamination, despite the fact that they are not in a full Bayesian framework, avoiding the flawed assumption that the probability of observing gentoypes in sources equals to that in humans via simulation of the sampling distribution of genotypes. It also shows that the genetic-free model (the Dirichlet model) is at least as effective as the genetic model (the asymmetric Island model) for final inference. Thus, genetic models may not be

always required for modelling source attribution; genetic-free models are an alternative as long as data provide enough information for inference.

There are some potential benefits of using the genetic-free models. They are less complex and use fewer parameters compared to genetic models. Hence, they are more convenient and easy to implement. Further, they can be applied to both phenotyping and genotyping data in principle; for example, serotyping or phage typing as used in the original Hald model (Hald et al., 2004). However, one limitation is that they may not be applicable to large-scale data such as wgMLST, but genetic models may still work adequately due to the specification of the genetic distance between isolates.

9.2 Future work

There are some areas arising from this thesis that may be of interest for future work. One is an improvement for the model proposed in Chapter 7, the other is an in-depth analysis on water contamination in relation to *Campylobacter* infection.

Although the effort of model generalisation for Chapter 4 is demonstrated in Chapter 7, the inference is unsuccessful after introducing a Cauchy random effect to link the relationship of observing genotypes between cases and sources. As investigated before, the posterior median of O for many of them is centred around 0, indicating that the prevalence of many types from sources may be close to that from humans. A possible way to improve the modelling is to divide types with the link effect O into three groups: $O = 0$ and whether the absolute value of O is in the range of (0, 0.04). This might be helpful to identify the relationship of genotype prevalence between humans and sources. Another possibility for model improvement is to use expert opinions as prior knowledge of genotype prevalence. It is believed that most of the genotypes have a property of generalist lifestyle with highly diverse genetic lineages (e.g. a common type of ST-45), such that humans and broad host ranges are flexibly colonised (Gripp et al., 2011; Levesque et al., 2013; Dearlove et al., 2016; Mughini-Gras et al., 2016). Host-specific types are also noticed, such as ST-2381, which is only found largely in water and water birds. Therefore, it could be helpful to understand the relationship by containing knowledge of generalist and specialist types as a part of modelling.

On the other hand, the approach in Chapter 6 potentially opens a discussion about transmission of *Campylobacter* through waterways. Faecal matter may be the primary cause of water contamination via rainfall or runoff coming from farmland and pasture (Wagenaar et al., 2013). This highlights the importance of water quality. It is observed that climate change has presented risks to affect water quality due to extreme events such as higher temperatures (Bates et al., 2008).

In recent decades, rainfall and temperature have been extensively studied in the literature, addressing the association between climate change and zoonotic diseases

(Sari Kovats et al., 2005b; Lake et al., 2009; Rind and Pearce, 2010; Spencer et al., 2012; Lal et al., 2013; Djennad et al., 2019). This means that studying the impact of water quality on the disease requires a broad view of climate change. For example, changes in rainfall patterns and temperature may project to affect land use and hence influence water quality (Lal et al., 2013). In addition, this thesis reveals that water birds may be the most likely contributor to water contamination. It is also possible that ruminant strains can be found in freshwater and may be related to the infection via recreational water (Shrestha et al., 2019). Therefore, in order for the epidemiology of waterborne campylobacteriosis to be understood well, modelling pathway attribution with the use of higher resolution typing data such as core genome and wgMLST may be a possible research direction, considering factors ranging from land use, flood and droughts to wild bird activities.

Appendix A

Supporting materials

A.1 Derivation of a Dirichlet distribution from a gamma distribution

As we assumed that the transformation between $\boldsymbol{Z}$ and $\boldsymbol{\theta}$ is one-to-one, we can inverse the above random variable to,

$$z_j = \theta_j d_I, \quad j = 1, \ldots, I-1$$
$$z_I = \theta_I d_I = (1 - \theta_1 - -\theta_{I-1}) d_I, \quad d_I = \sum_{i=1}^{I} z_i.$$

When $I > 2$, the transformation of random variables in the probability distribution involves the Jacobian matrix, which is defined below,

$$f_{\Theta_1,\ldots,\Theta_{I-1},D_I}(\theta_1, \ldots, \theta_{I-1}, D_I) = f_{Z_1,\ldots,Z_I}(z_1, \ldots, z_I) \cdot |J|,$$

where the Jacobian transformation J can be calculated via the matrix below,

$$J = \begin{bmatrix} \frac{\partial z_1}{\partial \theta_1} & \frac{\partial z_1}{\partial \theta_2} & \cdots & \frac{\partial z_1}{\partial \theta_{I-1}} & \frac{\partial z_1}{\partial d_I} \\ \frac{\partial z_2}{\partial \theta_1} & \frac{\partial z_2}{\partial \theta_2} & \cdots & \frac{\partial z_2}{\partial \theta_{I-1}} & \frac{\partial z_2}{\partial d_I} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \frac{\partial z_{I-1}}{\partial \theta_1} & \frac{\partial z_{I-1}}{\partial \theta_2} & \cdots & \frac{\partial z_{I-1}}{\partial \theta_{I-1}} & \frac{\partial z_{I-1}}{\partial d_I} \\ \frac{\partial z_I}{\partial \theta_1} & \frac{\partial z_I}{\partial \theta_2} & \cdots & \frac{\partial z_I}{\partial \theta_{I-1}} & \frac{\partial z_I}{\partial d_I} \end{bmatrix}$$

$$= \begin{bmatrix} d_I & 0 & \ldots & 0 & \theta_1 \\ 0 & d_I & \ldots & 0 & \theta_2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \ldots & d_I & \theta_{I-1} \\ -d_I & -d_I & \ldots & -d_I & (1-\theta_1-\ldots-\theta_{I-1}) \end{bmatrix} = d_I^{I-1}.$$

Therefore, the context of the joint pdf of $(\Theta_1, \ldots, \Theta_{I-1}, D_I)$ is,

$$f_{\Theta_1,\ldots,\Theta_{I-1},D_I}(\theta_1, \ldots, \theta_{I-1}, d_I) = \frac{1}{\Gamma(\alpha_1)\cdots\Gamma(\alpha_I)} \theta_1^{\alpha_1-1} \cdots \theta_{I-1}^{\alpha_{I-1}-1} (1-\theta_1-\ldots-\theta_{I-1})^{\alpha_I-1} d_I^{\sum_{i=1}^{I}\alpha_i-1} \exp(-d_I);$$

then, it is found that the joint pdf of $(\Theta_1, \ldots, \Theta_{I-1})$ is exactly the Dirichlet distribution with parameters $\boldsymbol{\alpha}$ in equation (3.5) after integrating d_I out,

$$\begin{aligned} f(\theta_1, \ldots, \theta_{I-1}) &= \int_0^\infty f(\theta_1, \ldots, \theta_{I-1}, d_I)\, dd_I \\ &= \frac{1}{\Gamma(\alpha_1)\cdots\Gamma(\alpha_I)} \theta_1^{\alpha_1-1} \cdots \theta_{I-1}^{\alpha_{I-1}-1} (1-\theta_1-\ldots-\theta_{I-1})^{\alpha_I-1} \\ &\quad \int_0^\infty d_I^{\sum_{i=1}^{I}\alpha_i-1} \exp(-d_I)\, dd_I \\ &= \frac{\Gamma(\alpha_1+\ldots+\alpha_I)}{\Gamma(\alpha_1)\cdots\Gamma(\alpha_I)} \theta_1^{\alpha_1-1} \cdots \theta_{I-1}^{\alpha_{I-1}-1} (1-\theta_1-\ldots-\theta_{I-1})^{\alpha_I-1}. \end{aligned}$$

A.2 The Multinomial theorem

The multinomial theorem is the generalisation of binomial theorem that allows expanding a power of many real numbers into a sum of power. The binomial theorem consists of two terms (a_1, a_2) with binomial coefficients, $\binom{n}{k} = \frac{n!}{k!(n-k)!}, n = 1, \ldots, k$, and can be written as,

$$\begin{aligned} (a_1+a_2)^n &= \sum_{k=0}^{n} \binom{n}{k} a^{n-k} b^k \\ &= \binom{n}{0} a_1^n a_2^0 + \binom{n}{1} a_1^{n-1} a_2^1 + \cdots + \binom{n}{n} a_1^0 a_2^n, \quad n \in \mathbb{N}. \end{aligned}$$

If the real numbers are more than two terms ($m > 2$), the binomial theorem extends to the multinomial theorem in a form of,

$$(a_1 + a_2 + \ldots + a_m)^n = \sum_{n_1+n_2+\ldots+n_m=n} \left(\begin{array}{c} n \\ n_1, n_2, \ldots, n_m \end{array} \right) \prod_{m=1}^{n} a_m^{n_m},$$

with multinomial coefficients below to express the number of possible combinations of different things out of n for each term,

$$\left(\begin{array}{c} n \\ n_1, n_2, \ldots, n_m \end{array} \right) = \frac{n!}{n_1! n_2! \ldots n_m!}.$$

The above theorem will be used when a genotype typed from human cases is assumed to be associated with one of target sources of origin. Therefore, the notation n can be regarded as the frequency of the genotype typed from human cases; $n_1, \ldots, n_m$ indicate the total number of the genotype typed from humans must be from one of m source categories with an appropriate probability $a_m^{n_m}$.

A.3 Model fitting with water birds as an additional source

The model fitting demonstrated here is for the circumstance when water birds are also considered in the source groups with the underlying framework developed in Chapter 5. The total number of source categories hence changes from 4 to 5, which are ordered as poultry, other, water birds, ruminants and water.

A.3.1 Estimates for water attribution

After simulating π_{ij}, $i = 1, \ldots, 406$, $j = 1, \ldots, 5$, using the asymmetric Island or Dirichlet model, the probabilities $\boldsymbol{p}$ and $\boldsymbol{G}$ given the source baseline is ruminants ($j = 4$) turn out to be,

$$p_i = \pi_{i1} G_1 + \pi_{i2} G_2 + \pi_{i3} G_3 + \pi_{i4} G_4,$$

where,

$$G_1 = \frac{\exp(g_1)}{1 + \exp(g_1) + \exp(g_2) + \exp(g_3)}$$
$$G_2 = \frac{\exp(g_2)}{1 + \exp(g_1) + \exp(g_2) + \exp(g_3)}$$
$$G_3 = \frac{\exp(g_3)}{1 + \exp(g_1) + \exp(g_2) + \exp(g_3)}$$

$$G_4 = \frac{1}{1 + \exp(g_1) + \exp(g_2) + \exp(g_3)}.$$

A.3.2 Estimates for human attribution

When no variables are included in the attribution model, the log-odds f_j are defined as,

$$f_j = \beta_{0j}, \quad j = 1, 2, 3, 5,$$

and $f_4 = 0$, given the source baseline is ruminants. Then, the attribution probabilities on the logit scale are expressed as,

$$\begin{aligned}
F_1 &= \frac{\exp(\beta_{01})}{1 + \exp(\beta_{01}) + \exp(\beta_{02}) + \exp(\beta_{03}) + \exp(\beta_{05})} \\
F_2 &= \frac{\exp(\beta_{02})}{1 + \exp(\beta_{01}) + \exp(\beta_{02}) + \exp(\beta_{03}) + \exp(\beta_{05})} \\
F_3 &= \frac{\exp(\beta_{03})}{1 + \exp(\beta_{01}) + \exp(\beta_{02}) + \exp(\beta_{03}) + \exp(\beta_{05})} \\
F_4 &= \frac{1}{1 + \exp(\beta_{01}) + \exp(\beta_{02}) + \exp(\beta_{03}) + \exp(\beta_{05})} \\
F_5 &= \frac{\exp(\beta_{05})}{1 + \exp(\beta_{01}) + \exp(\beta_{02}) + \exp(\beta_{03}) + \exp(\beta_{05})}.
\end{aligned}$$

Hence, the proportion $\hat{\pi}_i$ of genotype i found on humans becomes,

$$\hat{\pi}_i = \Big(\pi_{i1}G_1 + \pi_{i2}G_2 + \pi_{i3}G_3 + \pi_{i4}G_4\Big)F_5 + \pi_{i1}F_1 + \pi_{i2}F_2 + \pi_{i3}F_3 + \pi_{i4}F_4.$$

When the rurality variable c is considered in the attribution model, the algebraic form of f_j changes to,

$$\boldsymbol{f}_{1804\times 4} = \begin{bmatrix} 1 & c_1 \\ 1 & c_2 \\ \vdots & \vdots \\ 1 & c_{1804} \end{bmatrix} \begin{bmatrix} \beta_{01} & \beta_{02} & \beta_{03} & \beta_{05} \\ \beta_{11} & \beta_{12} & \beta_{13} & \beta_{15} \end{bmatrix}.$$

Thus, the attribution probabilities for five sources from the perspective of individuals lead to,

$$F_{h1} = \frac{\exp(\beta_{01} + \beta_{h1}c_h)}{1 + \exp(\beta_{01} + \beta_{h1}c_h) + \exp(\beta_{02} + \beta_{h2}c_h) + \exp(\beta_{03} + \beta_{h3}c_h) + \exp(\beta_{05} + \beta_{h5}c_h)}$$

$$F_{h2} = \frac{\exp(\beta_{02} + \beta_{h2}c_h)}{1 + \exp(\beta_{01} + \beta_{h1}c_h) + \exp(\beta_{02} + \beta_{h2}c_h) + \exp(\beta_{03} + \beta_{h3}c_h) + \exp(\beta_{05} + \beta_{h5}c_h)}$$

$$F_{h3} = \frac{\exp(\beta_{03} + \beta_{h3}c_h)}{1 + \exp(\beta_{01} + \beta_{h1}c_h) + \exp(\beta_{02} + \beta_{h2}c_h) + \exp(\beta_{03} + \beta_{h3}c_h) + \exp(\beta_{05} + \beta_{h5}c_h)}$$

$$F_{h4} = \frac{1}{1 + \exp(\beta_{01} + \beta_{h1}c_h) + \exp(\beta_{02} + \beta_{h2}c_h) + \exp(\beta_{03} + \beta_{h3}c_h) + \exp(\beta_{05} + \beta_{h5}c_h)}$$

$$F_{h5} = \frac{\exp(\beta_{05} + \beta_{h5}c_h)}{1 + \exp(\beta_{01} + \beta_{h1}c_h) + \exp(\beta_{02} + \beta_{h2}c_h) + \exp(\beta_{03} + \beta_{h3}c_h) + \exp(\beta_{05} + \beta_{h5}c_h)},$$

and the probability of observing genotype i on human isolates is expressed as,

$$\hat{\pi}_h = p_{i[h]}F_{h5} + \pi_{i[h]1}F_{h1} + \pi_{i[h]2}F_{h2} + \pi_{i[h]3}F_{h3} + \pi_{i[h]4}F_{h4}.$$

Appendix B

Supporting tables

B.1 Chapter 6

Data period	Human	Poultry	Ruminants	Water	Other	Water Birds
2005-2017	1834	1163	856	341	223	183
2005-2016	1804	1047	773	303	395	NA

Table B.1: Data sets collected during 2005-2016 and 2005-2017 show the difference in the number of isolates sampled from each category.

Data period	-3	-2	-1	0	1	2	3
2005-2017	20	134	168	103	302	236	707
2005-2016	20	133	165	101	295	231	696

Table B.2: The number of human cases dwelling in each rurality class (ranged from -3 to 3) from data collected during 2005-2016 and 2005-2017.

Appendix C

Supporting figures

C.1 Chapter 4

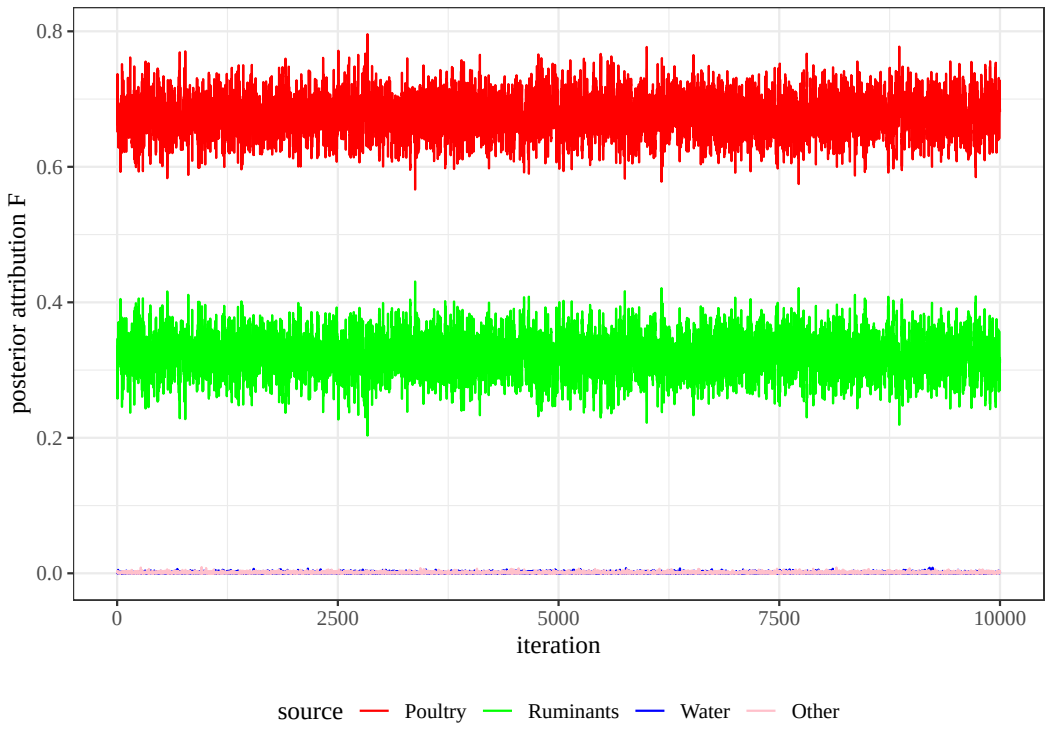


Figure C.1: The trace plot for posterior attribution probability after re-running the Gibbs sampler, given the attribution probability is modelled with Dir(1) prior.

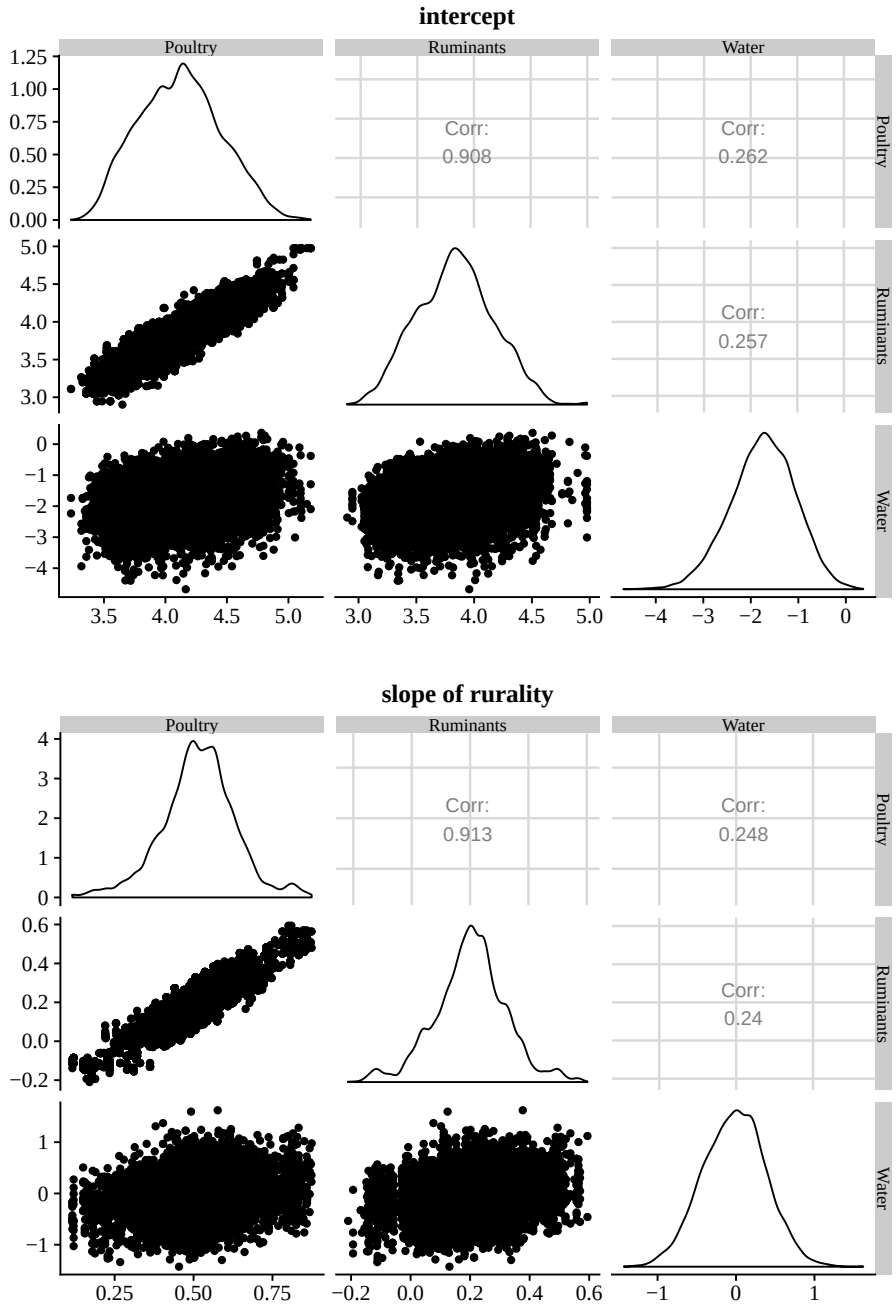


Figure C.2: The matrix of scatter plots for the parameters: intercept (top), and slope of rurality (bottom). Each matrix visualizes: i) the density plot for the parameter for each source on the diagonal; ii) the correlation coefficient of the parameter between sources in the upper panel; and iii) the scatter plot for the parameter between sources in the lower panel.

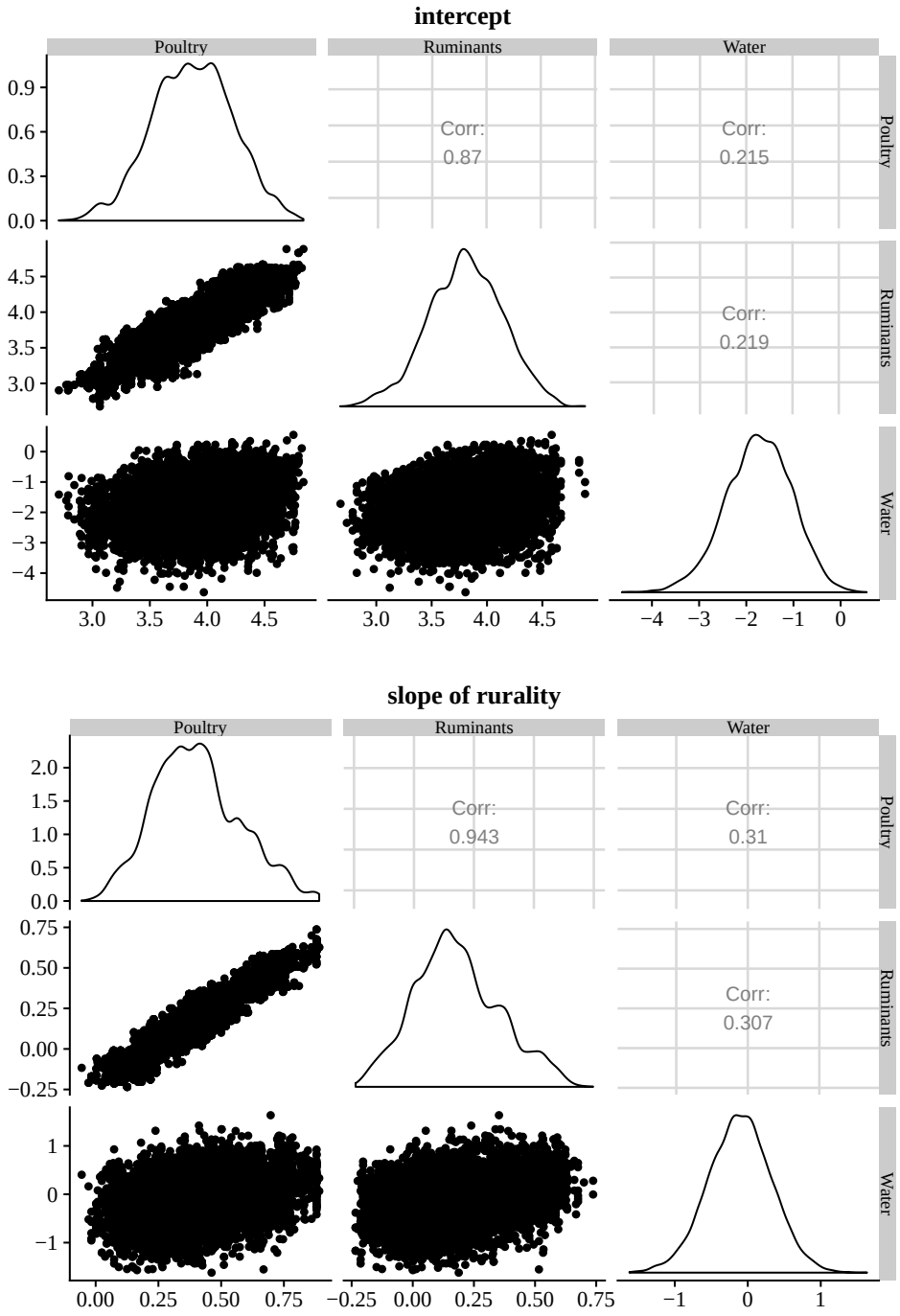


Figure C.3: The matrix of scatter plots for the fitted parameters: intercept (top), slope of rurality (bottom), given the baseline is other sources.

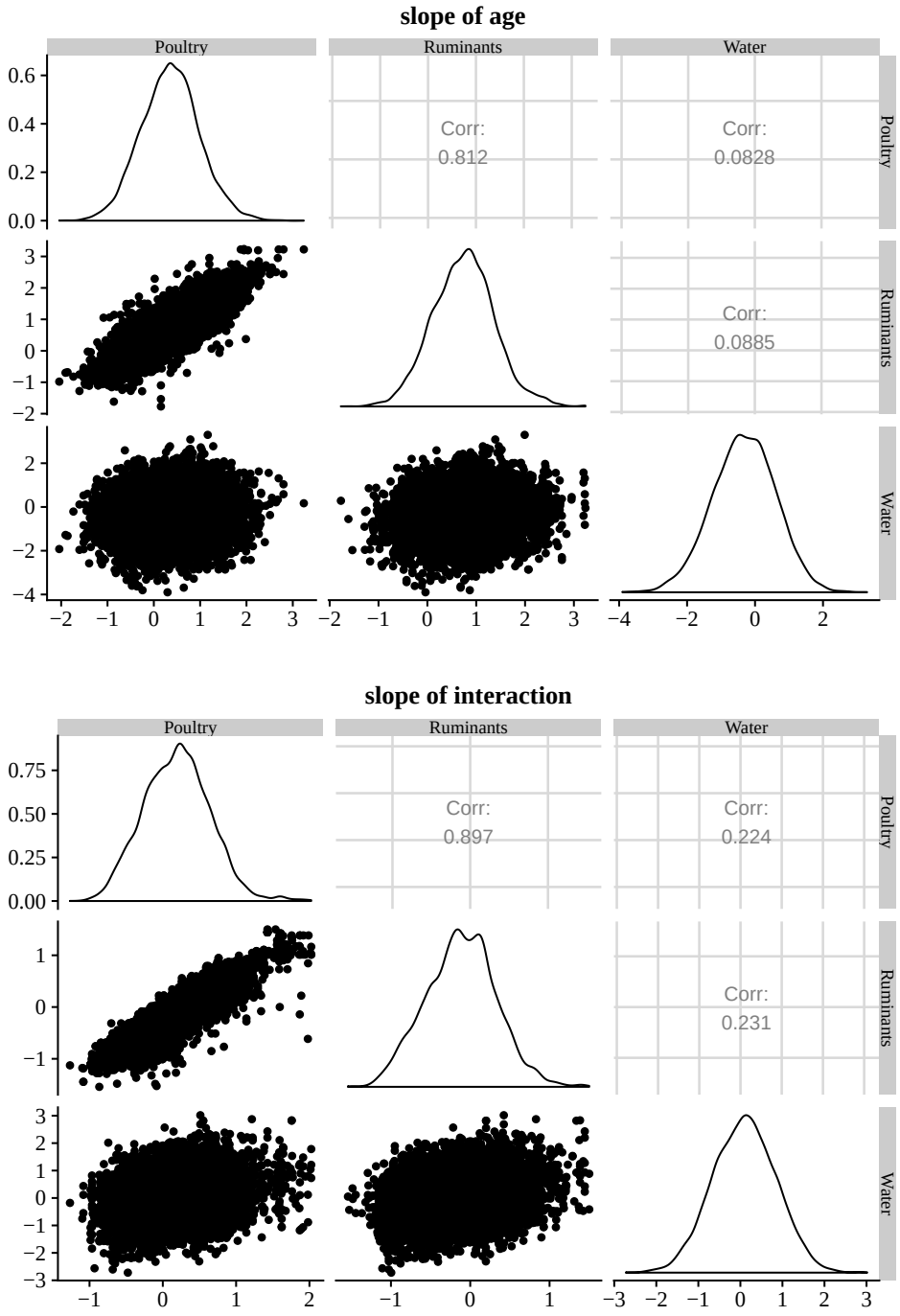


Figure C.4: The matrix of scatter plots for the fitted parameters: slope of age (top) and the interaction between rurality and age (bottom), given the baseline is other sources.

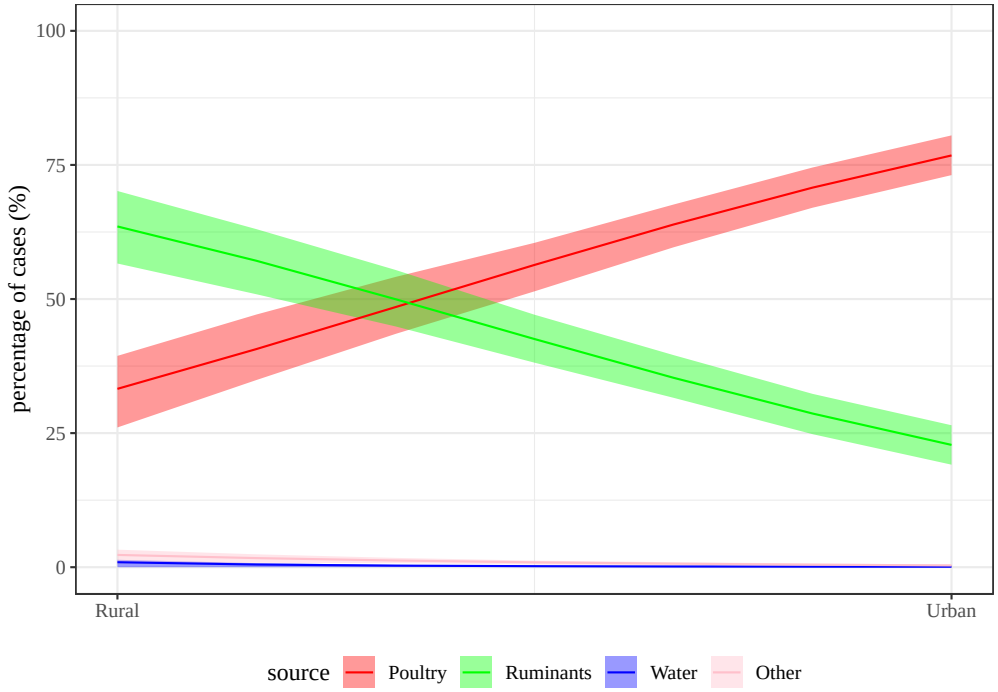


Figure C.5: The posterior attribution after setting a seed to starting values of $\boldsymbol{\pi}$ in the chain, given the rurality variable is considered in the model with other sources as the baseline.

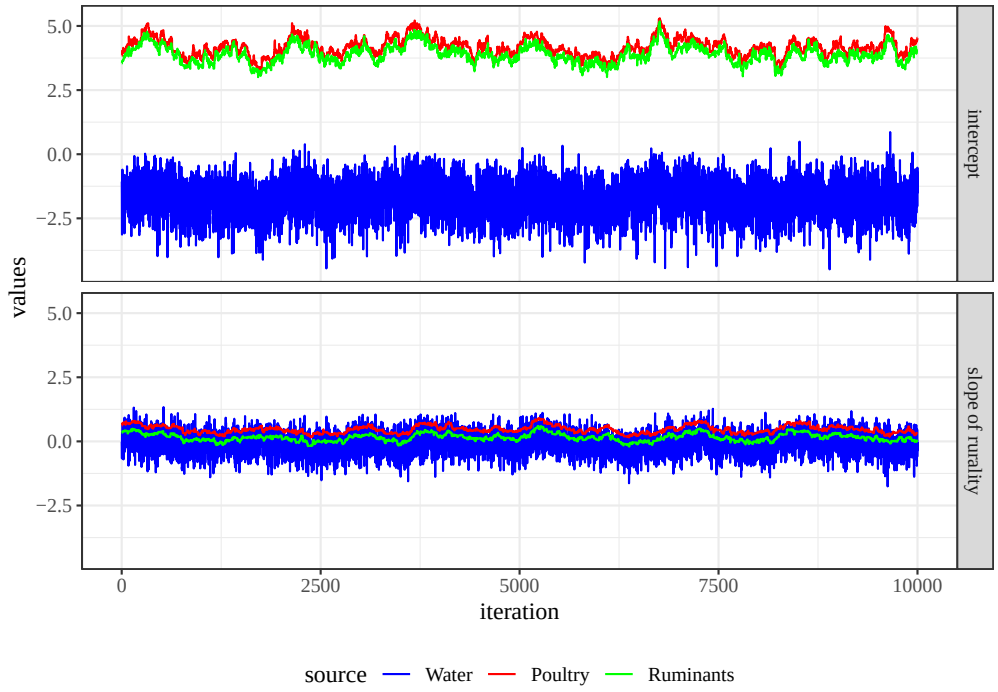


Figure C.6: The trace plots for parameters considered in the model after a seed is set to starting values of $\boldsymbol{\pi}$ in the chain, given other sources are the baseline.

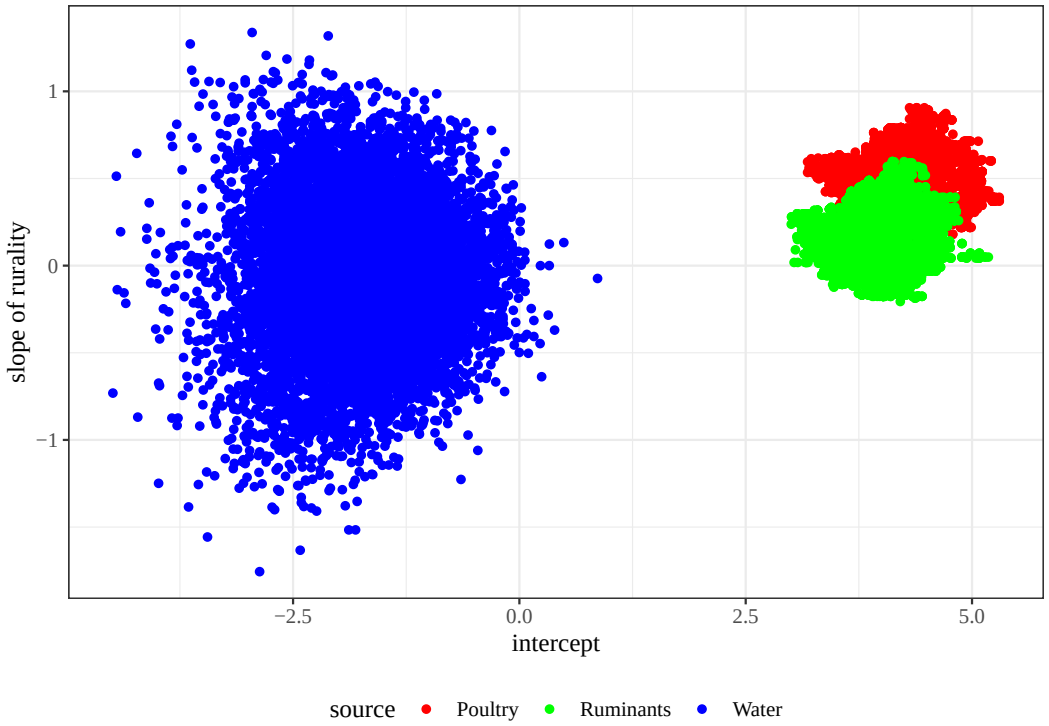


Figure C.7: The scatter plot of two parameters, given the baseline is other sources in the model.

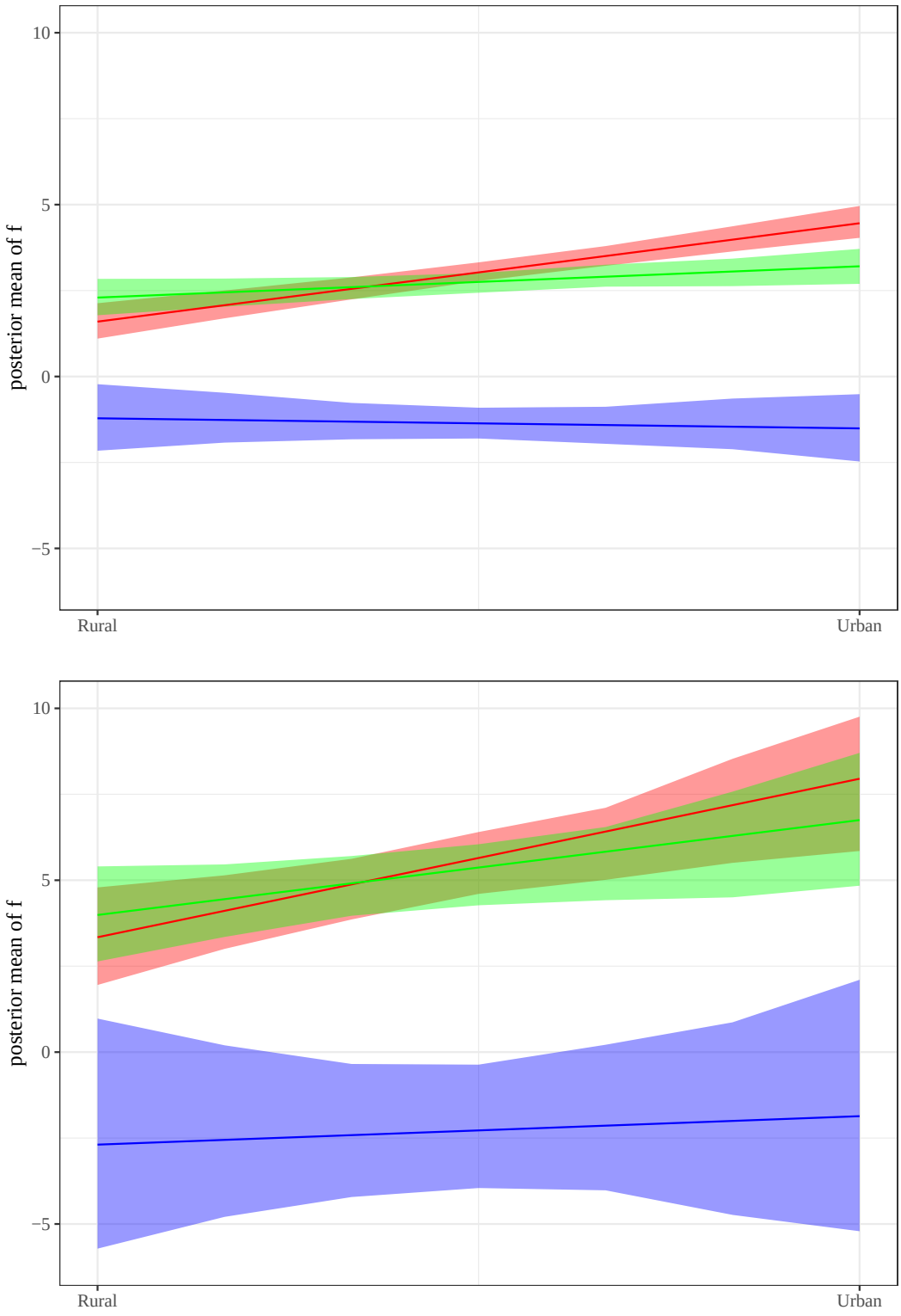


Figure C.8: The posterior mean of $\boldsymbol{f}$ for each source, given the baseline is other sources and the variance of normal prior for regression parameters $\boldsymbol{\theta}$ changes from 1 to 0.025 (top), or from 1 to 4 (bottom).

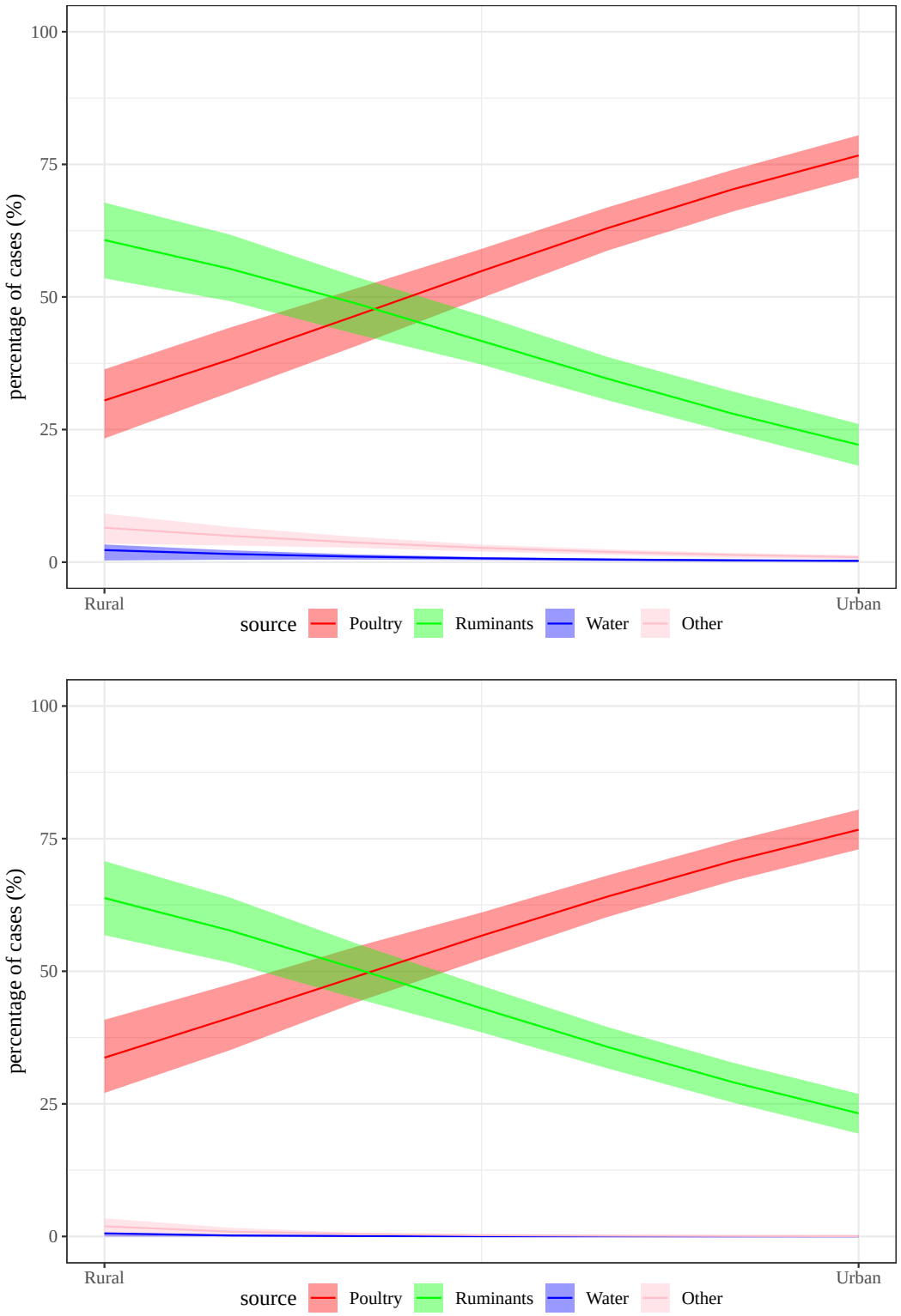
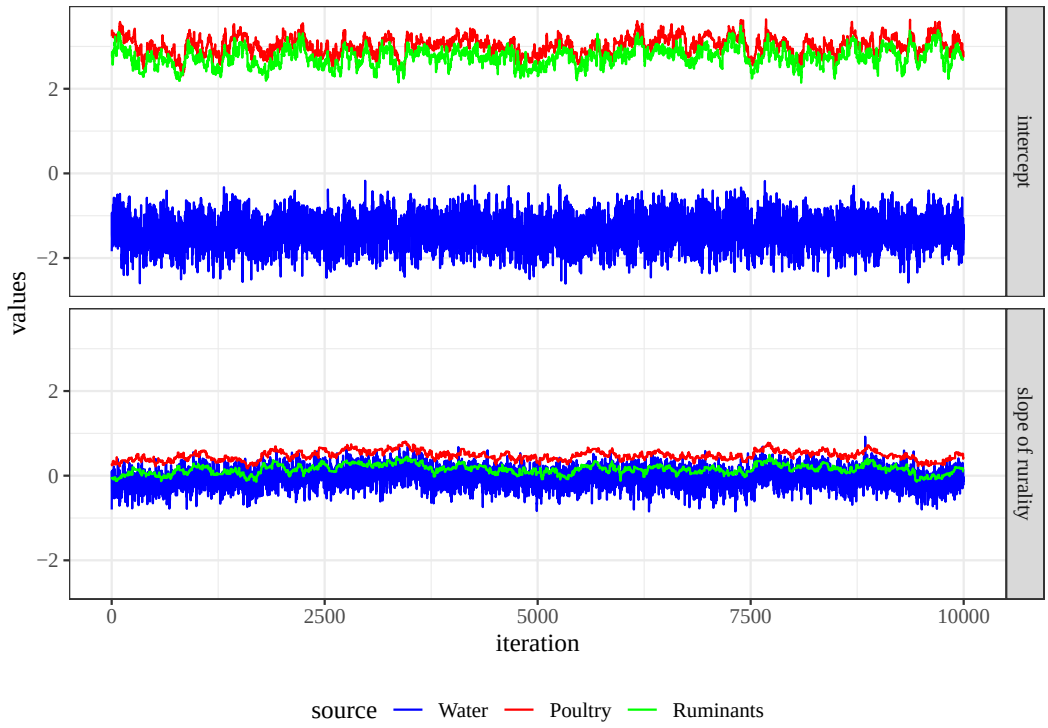
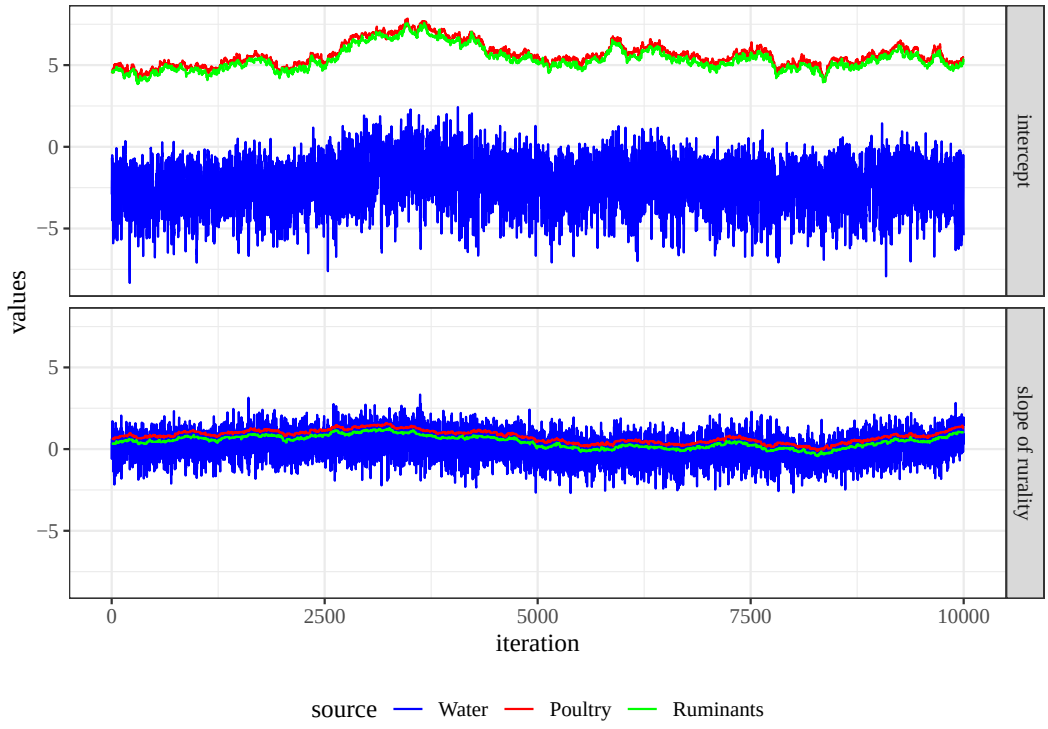


Figure C.9: The percentage of human cases attributable to each source, given the variance of normal prior for regression parameters in the algorithm changes from 1 to 0.025 (top), or from 1 to 4 (bottom).



(a)



(b)

Figure C.10: The trace plot for considered regression parameters when the variance of normal prior for parameters (a) decreased from 1 to 0.025 (top), or (b) increased from 1 to 4 (bottom), given the baseline is other sources.

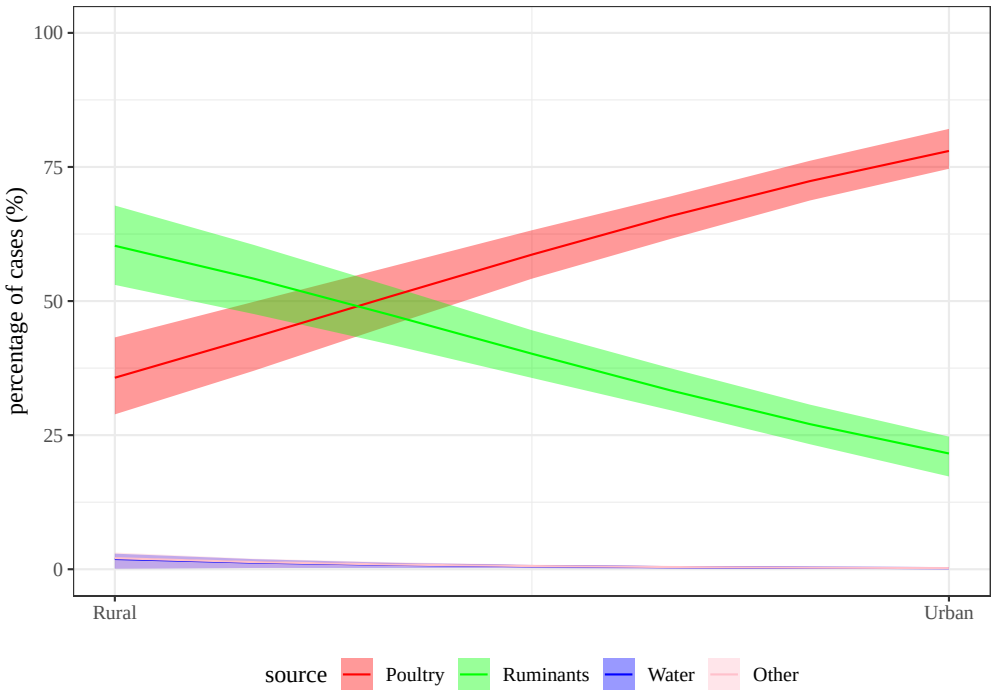


Figure C.11: The percentage of human cases attributable to each source, given the baseline changed from other sources to ruminants.

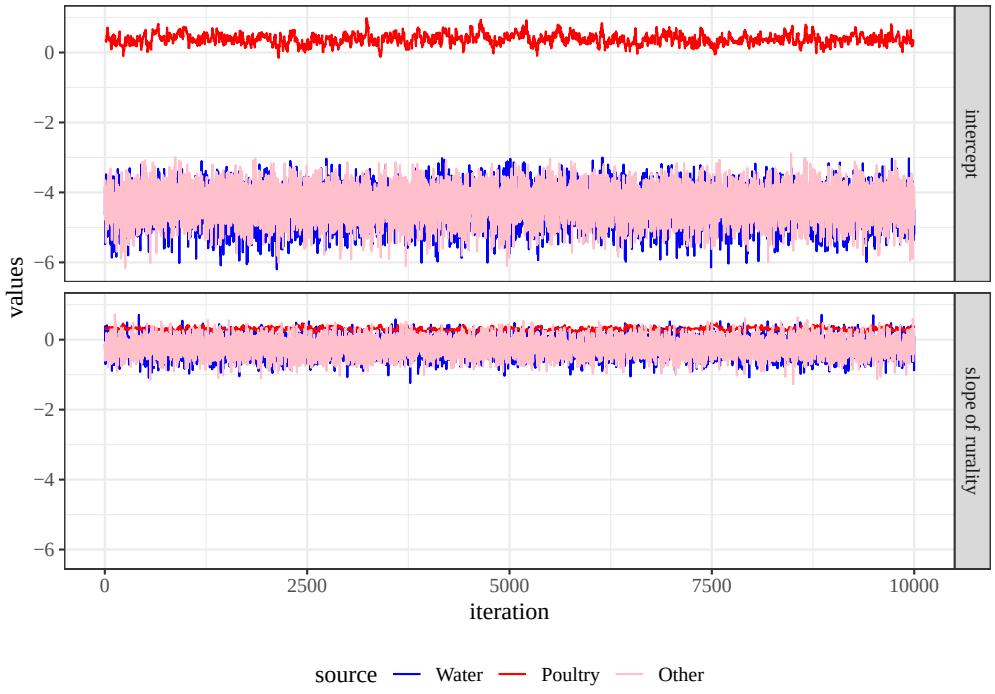


Figure C.12: The trace plot for regression parameters considered in the model after changing the baseline from other sources to ruminants.

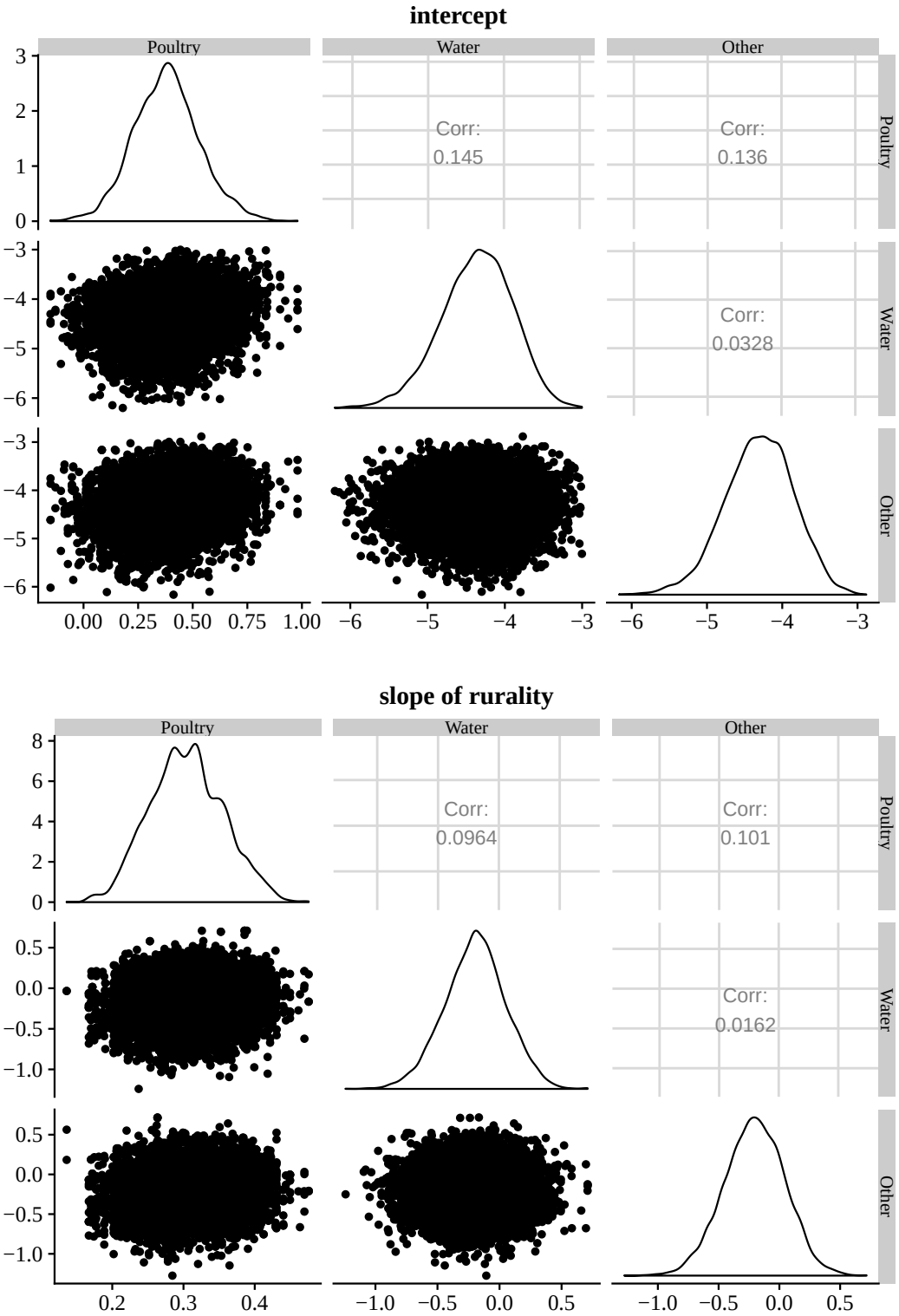


Figure C.13: The matrix of scatter plots for the fitted intercept and slope of rurality parameters, given the baseline changed from other sources to ruminants.

C.2 Chapter 6

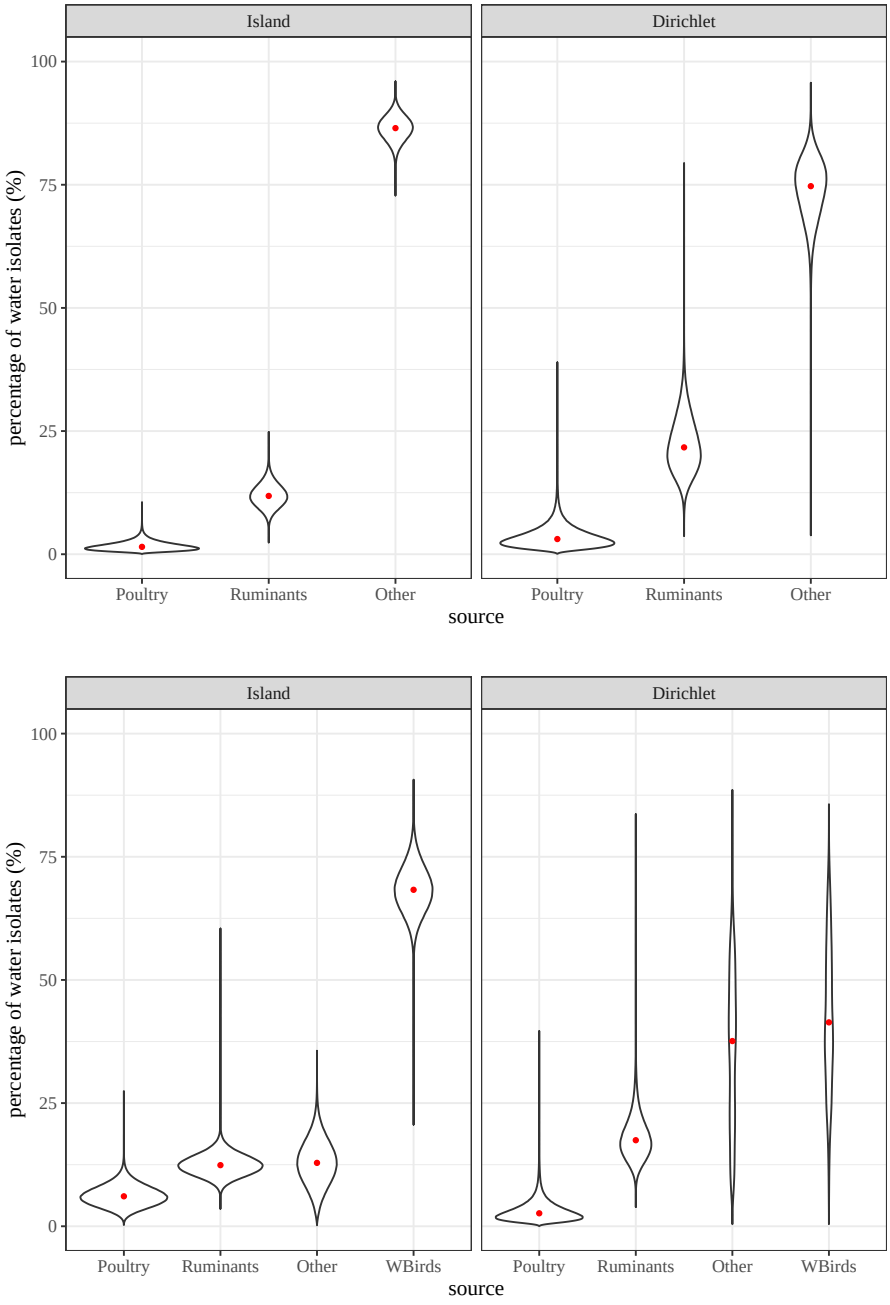


Figure C.14: Percentage of water isolates attributable to each source when water birds are included (lower panel) or not included (upper panel) in the analysis, after mixing over 100 simulated $\boldsymbol{\pi}$, estimated by the asymmetric Island model against the Dirichlet model. The inference is based on the fitted model with the rurality variable c, given ruminants are the source baseline.

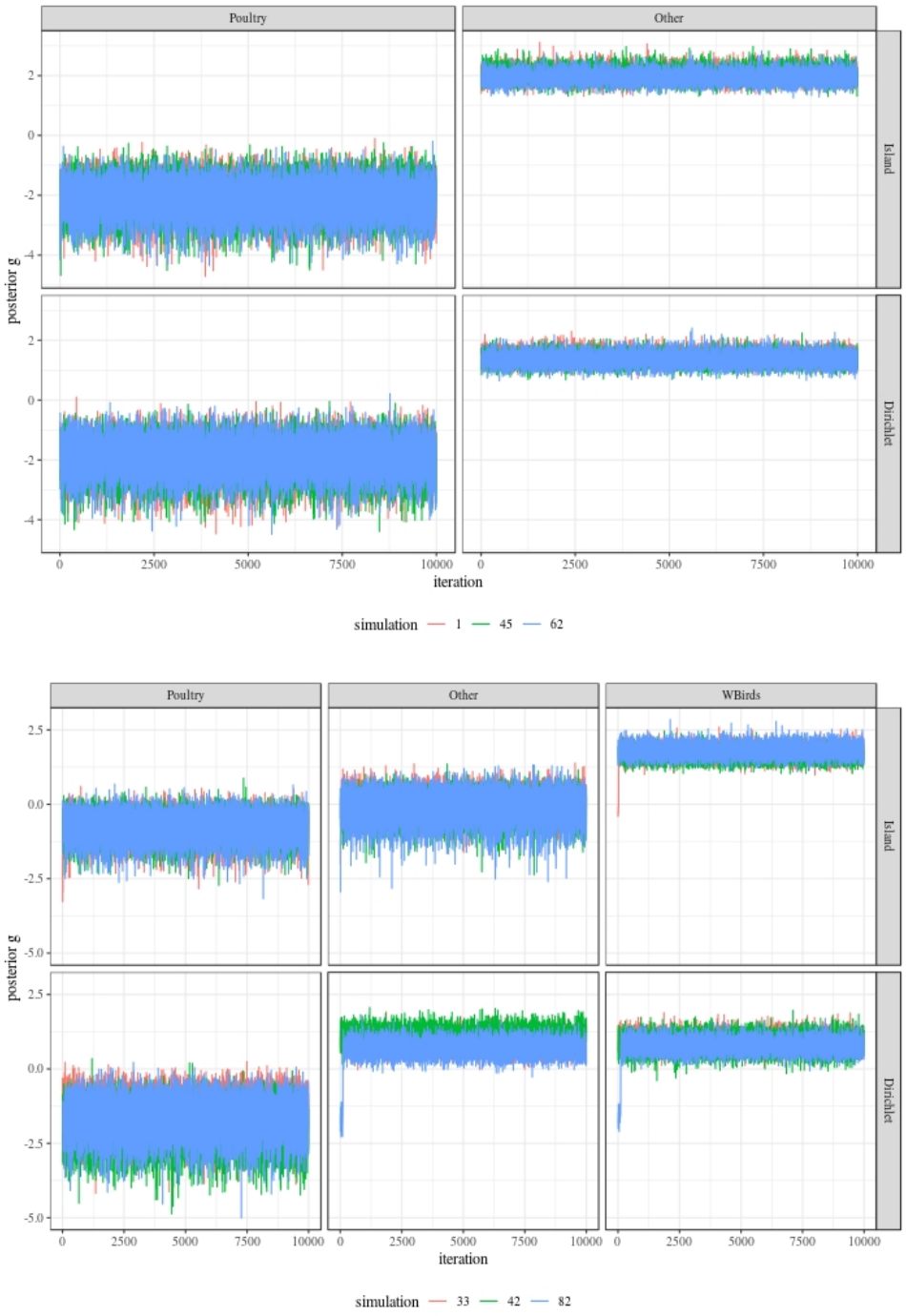


Figure C.15: Trace plots for the water parameters $\boldsymbol{g}$ when the rurality variable is included in the attribution model for the analysis without (upper panel) or with (lower panel) the source of water birds, given the source baseline is ruminants and three simulations of $\boldsymbol{\pi}$ estimated by the asymmetric Island model against the Dirichlet model.

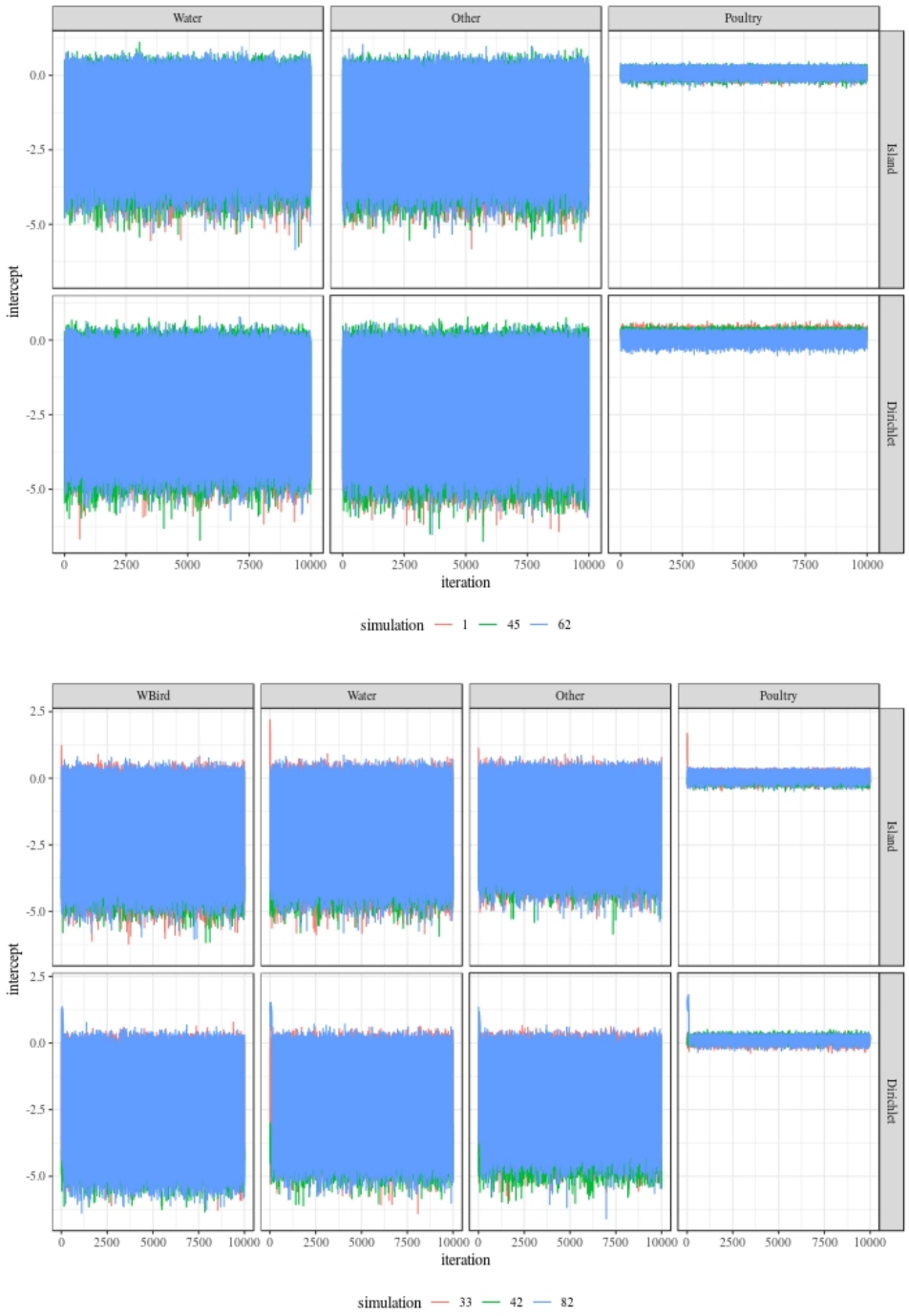


Figure C.16: Trace plots for the intercept regression parameters when the rurality variable is included in the attribution model for the analysis without (upper panel) or with (lower panel) the source of water birds, given ruminants are the source baseline and three simulations of $\boldsymbol{\pi}$ estimated by the asymmetric Island model against the Dirichlet model.

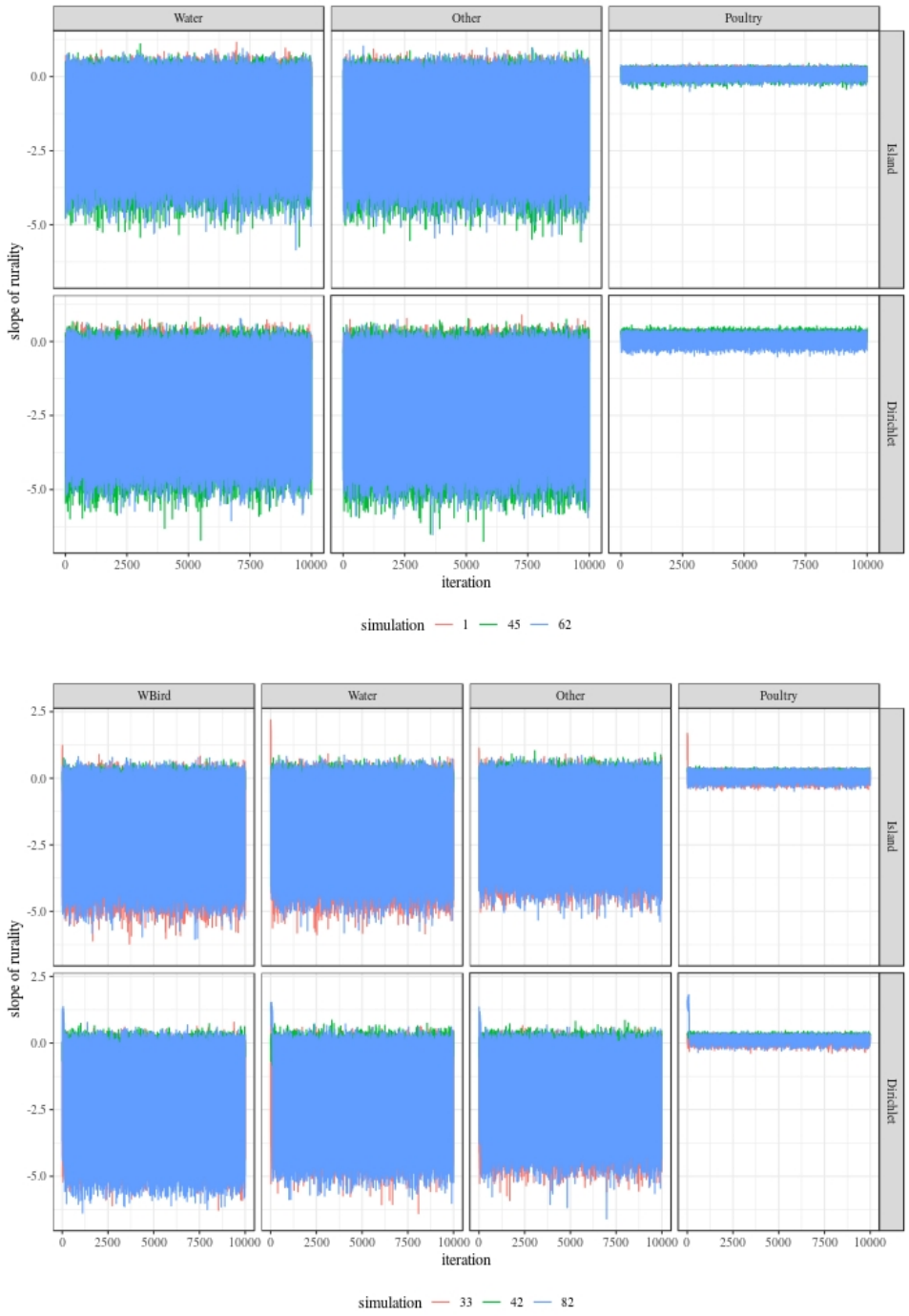


Figure C.17: Trace plots for the slope of rurality parameters when the rurality variable is included in the attribution model for the analysis without (upper panel) or with (lower panel) the source of water birds, given ruminants are the source baseline and three simulations of $\boldsymbol{\pi}$ estimated by the asymmetric Island model against the Dirichlet model.

Appendix D

Glossary

Allele A specific variation of a gene or locus, which may lead to different observable traits within a species.

Amplification An increase in the amount of target DNA by using a Polymerase Chain Reaction (PCR).

Attribution probability A probability of a random human case being caused by a source of infection.

Base pair A pair of chemical bases or nucleotides bonded together, forming the building blocks of the DNA structure.

Core genome The genes present in all/most strains of a species, including the housekeeping genes.

Food matrices A term to express complex structures of foods from their nutrient components to molecular relationships to each other.

Gene A sequence of nucleotides at some position of DNA, providing code to determine the characteristics transferred from a parent to offspring.

Generalist A property, describing a genotype that can be frequently found in a wide range of host animals.

Genome All the genetic information of an organism.

Genotype A term referring to isolates that have the same unique combination of alleles.

Gibbs sampler One of MCMC algorithms, also a special case of Metropolis-Hasting algorithm with a probability of acceptance of one, obtaining approximated samples from a specified probability distribution when direct sampling from a target distribution is difficult.

Guillain-Barre syndrome A rare disorder of the human nerve system, leading to paralysis in the limbs and compromised respiration.

Highest posterior density An interval within which a parameter of interest falls with a certain probability.

Housekeeping gene A gene responsible for the existence of a cell, maintaining the basic and essential cellular functions.

Interquartile range A statistical measure that points out where the middle 50% of data locates.

Irritable bowel syndrome An intestinal disorder causing cramping, bloating, diarrhoea and constipation.

Isolate A single colony representing a bacterial clone isolated from agar plates when microorganisms are cultured.

Likelihood function A function to express the likeliness of different values of a parameter given observed data.

Locus A part of a gene that may be used for typing.

Markov chain Monte Carlo A sampling method, being comprised of a class of algorithms, allows us to find the posterior distribution of parameters of interest through accepting simulated samples at a certain probability.

Metropolis-Hastings algorithm One of MCMC algorithms, obtaining samples from a proposal distribution using an acceptance-rejection step when direct sampling from a target distribution is difficult.

Multilocus sequence typing A molecular technique for the typing of several loci (usually seven), providing an allelic profile of an isolate.

Nucleotide A chemical base that forms the building block of a strand of DNA (or RNA).

Pathogen A microorganism that can cause diseases such as bacteria and viruses.

Posterior probability A probability describing how likely it is that an event will happen after considering observations and prior beliefs.

Prior probability A probability expressing the chance an event will happen prior to actual observations.

Sequence type A designated genetic term, referring to a unique allelic profile.

Squared difference A method to measure the variation between actual and estimated data through squaring the difference between the observed and estimated values.

Whole genome multilocus sequence typing A molecular typing technique with higher resolution than MLST due to the inclusion of all available genes, rather than the 7 used in conventional MLST.

Bibliography

Adak, G. K., Cowden, J. M., Nicholas, S., and Evans, H. S. (1995). The Public Health Laboratory Service national case-control study of primary indigenous sporadic cases of *Campylobacter* infection. *Epidemiology and Infection*, 115(1):15–22. (doi:10.1017/S0950268800058076).

AH, H., MJ, N., MJJ, M., AG, d. K., M-J, B., EG, E., WF, J.-R., W, v. P., JA, W., GA, d. W., and H, v. d. Z. (2005). Cost and benefits of controlling Campylobacter in the Netherlands-Integrating risk analysis, epide- miology and economics. Technical report, National Institute for Public Health and the Environment.

Andreoletti, O., Budka, H., Buncic, S., Colin, P., Collins, J. D., De, A., Griffin, J., Havelaar, A., Hope, J., Klein, G., Kruse, H., Magnino, S., López, A. M., Mclauchlin, J., Nguyen-thé, C., Noeckler, K., Noerrung, B., Maradona, M. P., Roberts, T., Vågsholm, I., and Vanopdenbosch, E. (2008). Overview of methods for source attribution for human illness from food- Scientific Opinion of the Panel on Biological Hazards Adopted on 9 July 2008. *The EFSA Journal*, 764:1–43.

Baker, M. G., Sneyd, E., and Wilson, N. A. (2007). Is the major increase in notified campylobacteriosis in New Zealand real? *Epidemiology and Infection*, 135(1):163–170. (doi:10.1017/S0950268806006583).

Bates, B., Kundzewicz, Z., Wu, S., and Palutikof, J. (2008). Climate Change and Water - IPCC Technical Paper VI. Technical report, Intergovernmental Panel on Climate Change Secretariat, Geneva.

Batz, M. B., Hoffmann, S., and Morris, J. G. (2012). Ranking the disease burden of 14 pathogens in food sources in the united states using attribution data from outbreak investigations and expert elicitation. *Journal of Food Protection*, 75(7):1278–1291.

Biggs, P. J., Fearnhead, P., Hotter, G., Mohan, V., Collins-Emerson, J., Kwan, E., Besser, T. E., Cookson, A., Carter, P. E., and French, N. P. (2011). Whole-genome comparison of two *Campylobacter jejuni* isolates of the same sequence

type reveals multiple loci of different ancestral lineage. *PLoS ONE*, 6(11). (doi:10.1371/journal.pone.0027121).

Bolwell, C. F., Gilpin, B. J., Campbell, D., and French, N. P. (2015). Evaluation of the representativeness of a sentinel surveillance site for campylobacteriosis. *Epidemiology and Infection*, 143(9):1990–2002. (doi:10.1017/S0950268814003173).

Carter, P. E., McTavish, S. M., Brooks, H. J., Campbell, D., Collins-Emerson, J. M., Midwinter, A. C., and French, N. P. (2009a). Novel clonal complexes with an unknown animal reservoir dominate *Campylobacter jejuni* isolates from river water in new zealand. *Appl Environ Microbiol*, 75(19):6038–46. (doi:10.1128/AEM.01039-09).

Carter, P. E., McTavish, S. M., Brooks, H. J., Campbell, D., Collins-Emerson, J. M., Midwinter, A. C., and French, N. P. (2009b). Novel clonal complexes with an unknown animal reservoir dominate *Campylobacter jejuni* isolates from river water in New Zealand. *Applied and Environmental Microbiology*, 75(19):6038–6046.

Centers for Disease Control and Prevention (1978). Waterborne *Campylobacter* Gastroenteritis — Vermont. *Morbidity and Mortality Weekly Report*, 27(25):207–207.

Centers for Disease Control and Prevention (2013). Incidence and trends of infection with pathogens transmitted commonly through food - foodborne diseases active surveillance network, 10 U.S. sites, 1996-2012. *Morbidity and Mortality Weekly Report*, 62(15):283–287.

Chen, J., Sun, X.-t., Zeng, Z., and Yu, Y.-y. (2011). *Campylobacter* enteritis in adult patients with acute diarrhea from 2005 to 2009 in Beijing, China. *Chinese Medical Journal*, 124(10):1508–1512. (doi:10.3760/cma.j.issn.0366-6999.2011.10.013).

Chib, S. and Greenberg, E. (1995). Understanding the metropolis-hastings algorithm. *The American Statistician*, 49(4):327–335.

Cody, A. J., Maiden, M. C., Strachan, N. J., and McCarthy, N. D. (2019). A systematic review of source attribution of human campylobacteriosis using multi-locus sequence typing. *Eurosurveillance*, 24(43). (doi:10.2807/1560-7917.ES.2019.24.43.1800696).

Colles, F. M., Jones, K., Harding, R. M., and Maiden, M. C. (2003a). Genetic diversity of *Campylobacter jejuni* isolates from farm animals and the farm environment. *Appl Environ Microbiol*, 69(12):7409–13.

Colles, F. M., Jones, K., Harding, R. M., and Maiden, M. C. (2003b). Genetic Diversity of *Campylobacter jejuni* Isolates from Farm Animals and the Farm Environment. *Applied and Environmental Microbiology*, 69(12):7409–7413.

Cottam, E. M., Thebaud, G., Wadsworth, J., Gloster, J., Mansley, L., Paton, D. J., King, D. P., and Haydon, D. T. (2008). Integrating genetic and epidemiological data to determine transmission pathways of foot-and-mouth disease virus. *Proc Biol Sci*, 275(1637):887–95. (doi:10.1098/rspb.2007.1442).

Dearlove, B. L., Cody, A. J., Pascoe, B., Méric, G., Wilson, D. J., and Sheppard, S. K. (2016). Rapid host switching in generalist *Campylobacter* strains erodes the signal for tracing human infections. *ISME Journal*, 10(3):721–729. (doi:10.1038/ismej.2015.149).

Dingle, K. E., Colles, F. M., Wareing, D. R., Ure, R., Fox, A. J., Bolton, F. E., Bootsma, H. J., Willems, R. J., Urwin, R., and Maiden, M. C. (2001). Multilocus sequence typing system for *Campylobacter jejuni*. *J Clin Microbiol*, 39(1):14–23. (doi:10.1128/JCM.39.1.14-23.2001).

Djennad, A., Iacono, G. L., Sarran, C., Lane, C., Elson, R., Höser, C., Lake, I. R., Colón-gonzález, F. J., Kovats, S., Semenza, J. C., Bailey, T. C., Kessel, A., Fleming, L. E., and Nichols, G. L. (2019). Seasonality and the effects of weather on Campylobacter infections. *BMC Infectious Diseases*, 19(255):1–10. (doi:10.1186/s12879-019-3840-7).

Duncan, G. E. (2014). Determining the health benefits of poultry industry compliance measures: The case of campylobacteriosis regulation in New Zealand. *New Zealand Medical Journal*.

Dye, C. (2014). After 2015: infectious diseases in a new era of health and development. *Phil Trans R Soc B*, 369:20130426. (doi:10.1098/rstb.2013.0426).

Eberhart-Phillips, J., Walker, N., Garrett, N., Bell, D., Sinclair, D., Rainger, W., and Bates, M. (1997). Campylobacteriosis in New Zealand: results of a case-control study. *Journal of Epidemiology & Community Health*, 51:686–691. (doi:10.1136/jech.51.6.686).

Fearnhead, P., Biggs, P. J., and French, N. P. (2014). Learning about recombination in *Campylobacter*. In Sheppard, S. and Méric, G., editors, Campylobacter *Ecology and Evolution*, pages 9–22. Norfolk, UK: Caister Academic Press.

Fearnhead, P., Smith, N. G., Barrigas, M., Fox, A., and French, N. (2005). Analysis of recombination in *Campylobacter jejuni* from MLST population data. *Journal of Molecular Evolution*. (doi:10.1007/s00239-004-0316-0).

Fitzgerald, C., Stanley, K., Andrew, S., and Jones, K. (2001). Use of Pulsed-Field Gel Electrophoresis and Flagellin Gene Typing in Identifying Clonal Groups of *Campylobacter jejuni* and *Campylobacter coli* in Farm and Clinical Environments. *Applied*

and Environmental Microbiology, 67(4):1429–1436. (doi:10.1128/AEM.67.4.1429-1436.2001).

French, N., Yu, S., Patrick, P. B., Holland, B., Fearnhead, P., Binney, B., Fox, A., Grove-White, D., Leigh, J. W., Miller, W., and Müllner, P. (2014). Evolution of *Campylobacter* species in New Zealand. In Sheppard, S. and Méric, G., editors, Campylobacter *Ecology and Evolution*, pages 221–40. Norfolk, UK: Caister Academic Press.

French, N. P., Midwinter, A., Holland, B., Collins-Emerson, J., Pattison, R., Colles, F., and Carter, P. (2008). Molecular Epidemiology of *Campylobacter jejuni* Isolates from Wild-Bird Fecal Material in Children's Playgrounds. *Applied and Environmental Microbiology*, 75(3):779–783. (doi:10.1128/aem.01979-08).

Gadiel, D. (2010). Applied Economics Pty Ltd: The economic cost of foodborne disease in New Zealand. Report for the New Zealand Food Safety Authority. Accessed 27 August 2020. (hppts://https://www.mpi.govt.nz/dmsdocument/25814/direct).

Gelfand, A. E., Sahu, S. K., and Carlin, B. P. (1995). Efficient parametrisations for normal linear mixed models. *Biometrika*, 82(3):479–488. (doi:10.1093/biomet/82.3.479).

Gelman, A., Carlin, J. B. B., Stern, H. S. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. B. (2014a). Basics of Markov chain simulation. In *Bayesian data analysis*, pages 275–281. Chapman and Hall/CRC Texts in Statistical Science (01 November 2013).

Gelman, A., Carlin, J. B. B., Stern, H. S. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. B. (2014b). Evaluating, comparing, and expanding models. In *Bayesian data analysis*, pages 167–173. Chapman and Hall/CRC Texts in Statistical Science (01 November 2013).

Gillespie, I. A., O'Brien, S. J., Frost, J. A., Adak, G. K., Horby, P., Swan, A. V., Painter, M. J., and Neal, K. R. (2002). A case-case comparison of Campylobacter coli and Campylobacter jejuni infection: A tool for generating hypotheses. *Emerging Infectious Diseases*.

Gilpin, B. J., Walker, T., Paine, S., Sherwood, J., Mackereth, G., Wood, T., Hambling, T., Hewison, C., Brounts, A., Wilson, M., Scholes, P., Robson, B., Lin, S., Cornelius, A., Rivas, L., Hayman, D. T., French, N. P., Zhang, J., Wilkinson, D. A., Midwinter, A. C., Biggs, P. J., Jagroop, A., Eyre, R., Baker, M. G., and Jones, N. (2020). A large scale waterborne Campylobacteriosis outbreak, Havelock North, New Zealand. *Journal of Infection*, 81(3):390–395. (doi:10.1016/J.JINF.2020.06.065).

Greger, M. (2007). The human/animal interface: Emergence and resurgence of zoonotic infectious diseases. (doi:10.1080/10408410701647594).

Gripp, E., Hlahla, D., Didelot, X., Kops, F., Maurischat, S., Tedin, K., Alter, T., Ellerbroek, L., Schreiber, K., Schomburg, D., Janssen, T., Bartholomäus, P., Hofreuter, D., Woltemate, S., Uhr, M., Brenneke, B., Grüning, P., Gerlach, G., Wieler, L., Suerbaum, S., and Josenhans, C. (2011). Closely related *Campylobacter jejuni* strains from different sources reveal a generalist rather than a specialist lifestyle. *BMC Genomics*. (doi:10.1186/1471-2164-12-584).

Hahn, A. F. (1998). Guillain-barre syndrome. *Lancet*, 352(9128):635–41. (doi:10.1016/S0140-6736(97) 12308-X).

Hald, T., Vose, D., Wegener, H. C., and Koupeev, T. (2004). A bayesian approach to quantify the contribution of animal-food sources to human salmonellosis. *Risk Anal*, 24(1):255–69. (doi:10.1111/j.0272-4332.2004.00 427.x).

Havelaar, A. H., Van Pelt, W., Ang, C. W., Wagenaar, J. A., Van Putten, J. P., Gross, U., and Newell, D. G. (2009). Immunity to *Campylobacter*: Its role in risk assessment and epidemiology *Campylobacter* immunity in epidemiology and risk assessment A. H. Havelaar et al. *Critical Reviews in Microbiology*, 35(1):1–22. (doi:10.1080/10408410802636017).

Institute of Environmental Science and Research Limited (2004). Notifiable and Other Diseases in New Zealand Annual Report. Technical Report April.

Institute of Environmental Science and Research Limited (2009). Notifiable and Other Diseases in New Zealand 2008 Annual Surveillance Report. Technical Report June.

Institute of Environmental Science and Research Limited (2017). Notifiable Diseases in New Zealand Annual Report 2016 Environmental Science and Research Limited. Technical Report September.

Jones, K. E., Patel, N. G., Levy, M. A., Storeygard, A., Balk, D., Gittleman, J. L., and Daszak, P. (2008). Global trends in emerging infectious diseases. *Nature*, 451(7181):990–993. (doi:10.1038/nature06536).

Kaakoush, N. O., Castaño-Rodríguez, N., Mitchell, H. M., and Man, S. M. (2015). Global epidemiology of campylobacter infection. *Clinical Microbiology Reviews*.

Kadane, J. B. (2011). *Principles of uncertainty*. CRC Press, 1st edition.

Kapperud, G., Skjerve, E., Vik, L., Hauge, K., Lysaker, A., Aalmen, I., Ostroff, S. M., and Potter, M. (1993). Epidemiological investigation of risk factors for *Campylobacter*

colonization in Norwegian broiler flocks. *Epidemiology and Infection*, 111(2):245–256. (doi:10.1017/S0950268800056958).

King, E. O. (1957). Human infections with vibrio fetus and a closely related vibrio. *Journal of Infectious Diseases*. (doi:10.1093/infdis/101.2.119).

Kittl, S., Heckel, G., Korczak, B. M., and Kuhnert, P. (2013). Source attribution of human *Campylobacter* isolates by MLST and Fla-typing and association of genotypes with quinolone resistance. *PLoS ONE*, 8(11):1–8.

Kovanen, S. M., Kivistö, R. I., Rossi, M., Schott, T., Kärkkäinen, U. M., Tuuminen, T., Uksila, J., Rautelin, H., and Hänninen, M. L. (2014). Multilocus sequence typing (MLST) and whole-genome MLST of *Campylobacter jejuni* isolates from human infections in three districts during a seasonal peak in Finland. *Journal of Clinical Microbiology*.

Lake, I., Gillespie, I., Bentham, G., Nichols, G., Lane, C., Adak, G., and Threlfall, E. (2009). A re-evaluation of the impact of temperature and climate change on foodborne illness. *Epidemiology and Infection*, 137(11):1538–1547. (doi:10.1017/S0950268809002477).

Lal, A., Ikeda, T., French, N., Baker, M. G., and Hales, S. (2013). Climate variability, weather and enteric disease incidence in New Zealand: Time series analysis. *PLoS ONE*, 8(12). (doi:10.1371/journal.pone.0083484).

Lane, R. and Briggs, S. (2014). Campylobacteriosis in New Zealand: room for further improvement. *Journal of New Zealand Medical Association*, 127(1391):6–9.

Levesque, S., Fournier, E., Carrier, N., Frost, E., D. Arbeit, R., and Michaud, S. (2013). Campylobacteriosis in urban versus rural areas: A case-case study integrated with molecular typing to validate risk factors and to attribute sources of infection. *PLoS ONE*, 8(12):e83731.

Liao, S. J. (2020). R codes to support the PhD research available from GibHub. Accessed 31 March 2020 (https://github.com/jingliao/PhD_research).

Liao, S. J., Marshall, J., Hazelton, M. L., and French, N. P. (2019). Extending statistical models for source attribution of zoonotic diseases : a study of campylobacteriosis. *Journal of the Royal Society Interface*, 150(16):20180534. (doi:10.1098/rsif.2018.0534).

Louis, V. R., Gillespie, I. A., O'Brien, S. J., Russek-Cohen, E., Pearson, A. D., and Colwell, R. R. (2005). Temperature-driven *Campylobacter* seasonality in England and Wales. *Applied and Environmental Microbiology*, 71(1):85–92.

Marshall, J., French, N., Thornley, C., and van der Logt, P. (2016). Source attribution january to december 2014 of human *Campylobacter jejuni* cases from the manawatu. Technical report, Ministry for Primary Industries. Accessed 21 March 2018. (https://www.mpi.govt.nz/dmsdocument/15385/loggedIn).

Marshall, J. C. (2019). An R package available to install from GibHub: islandR. Accessed 03 December 2019 (https://github.com/jmarshallnz/islandR).

Mcbride, G. (2012). Issues in Setting Secondary Contact Recreation Guidelines for New Zealand Freshwaters. Technical Report September.

Miller, P., Marshall, J., French, N., and Jewell, C. (2017). sourcer: Classification and source attribution of infectious agents among heterogeneous populations. *PLoS Comput Biol*, 13(5):e1005564. (doi:10.1371/journal.pcbi.1005564).

Morelli, M. J., Thebaud, G., Chadoeuf, J., King, D. P., Haydon, D. T., and Soubeyrand, S. (2012). A bayesian inference framework to reconstruct transmission trees using epidemiological and genetic data. *PLoS Comput Biol*, 8(11):e1002768. (doi:10.1371/journal.pcbi.1002768).

Muellner, P., Marshall, J. C., Spencer, S. E., Noble, A. D., Shadbolt, T., Collins-Emerson, J. M., Midwinter, A. C., Carter, P. E., Pirie, R., Wilson, D. J., Campbell, D. M., Stevenson, M. A., and French, N. P. (2011). Utilizing a combination of molecular and spatial tools to assess the effect of a public health intervention. *Prev Vet Med*, 102(3):242–53. (doi:10.1016/j.prevetmed.2011.07.011).

Muellner, P., Pleydell, E., Pirie, R., Baker, M. G., Campbell, D., Carter, P. E., and French, N. P. (2013). Molecular-based surveillance of campylobacteriosis in new zealand–from source attribution to genomic epidemiology. *Euro Surveill*, 18(3).

Mughini-Gras, L., Barrucci, F., Smid, J. H., Graziani, C., Luzzi, I., Ricci, A., Barco, L., Rosmini, R., Havelaar, A. H., Van Pelt, W., and Busani, L. (2014). Attribution of human Salmonella infections to animal and food sources in Italy (2002-2010): Adaptations of the Dutch and modified Hald source attribution models. *Epidemiology and Infection*, 142(5):1070–1082.

Mughini-Gras, L., Kooh, P., Augustin, J. C., David, J., Fravalo, P., Guillier, L., Jourdan-Da-Silva, N., Thébault, A., Sanaa, M., Watier, L., Carlin, F., Leclercq, A., Hello, S. L., Pavio, N., and Villena, I. (2018). Source attribution of foodborne diseases: Potentialities, hurdles, and future expectations. *Frontiers in Microbiology*, 9:1983. (doi:10.3389/fmicb.2018.01983).

Mughini-Gras, L., Penny, C., Ragimbeau, C., Schets, F. M., Blaak, H., Duim, B., Wagenaar, J. A., de Boer, A., Cauchie, H. M., Mossong, J., and van Pelt, W. (2016). Quantifying potential sources of surface water contamination with *Campylobacter jejuni* and *Campylobacter coli*. *Water Research*. (doi:10.1016/j.watres.2016.05.069).

Mughini Gras, L., Smid, J. H., Wagenaar, J. A., de Boer, A. G., Havelaar, A. H., Friesema, I. H., French, N. P., Busani, L., and van Pelt, W. (2012). Risk factors for campylobacteriosis of chicken, ruminant, and environmental origin: A combined case-control and source attribution analysis. *PLoS ONE*, 7(8). (doi:10.1371/journal.pone.0042599).

Müllner, P., Collins-Emerson, J. M., Midwinter, A. C., Carter, P., Spencer, S. E., Van Der Logt, P., Hathaway, S., and French, N. P. (2010). Molecular epidemiology of *Campylobacter jejuni* in a geographically isolated country with a uniquely structured poultry industry. *Applied and Environmental Microbiology*, 76(7):2145–2154.

Mullner, P., Jones, G., Noble, A., Spencer, S. E., Hathaway, S., and French, N. P. (2009a). Source attribution of food-borne zoonoses in New Zealand: A modified hald model. *Risk Analysis*, 29(7):970–984. (doi:10.1111/j.1539-6924.2009.01224.x).

Mullner, P., Shadbolt, T., Collins-Emerson, J. M., Midwinter, A. C., Spencer, S. E., Marshall, J., Carter, P. E., Campbell, D. M., Wilson, D. J., Hathaway, S., Pirie, R., and French, N. P. (2010). Molecular and spatial epidemiology of human campylobacteriosis: source association and genotype-related risk factors. *Epidemiol Infect*, 138(10):1372–83. (doi:10.1017/S0950268809991579).

Mullner, P., Spencer, S. E., Wilson, D. J., Jones, G., Noble, A. D., Midwinter, A. C., Collins-Emerson, J. M., Carter, P., Hathaway, S., and French, N. P. (2009b). Assigning the source of human campylobacteriosis in new zealand: a comparative genetic and epidemiological approach. *Infect Genet Evol*, 9(6):1311–9. (doi:10.1016/j.meegid.2009.09.003).

Nelder, J. and Wedderburn, R. W. M. (1972). Generalized Linear Models. *Journal of the Royal Statistical Society. Series A (General)*, 135(3):370–384. (doi:10.2307/2344614).

New Zealand Department of Internal Affairs (2017). Report of the Havelock North drinking water inquiry: Stage 1. Technical Report May.

Nylen, G., Dunstan, F., Palmer, S. R., Andersson, Y., Bager, F., Cowden, J., Feierl, G., Galloway, Y., Kapperud, G., Megraud, F., Molbak, K., Petersen, L. R., and Ruutu, P. (2002). The seasonal distribution of *Campylobacter* infection in nine European countries and New Zealand. *Epidemiology and Infection*, 128(3):383–390.

Olson, C. K., Ethelberg, S., van Pelt, W., and Tauxe, R. V. (2008). Epidemiology of *Campylobacter jejuni* Infections in Industrialized Nations. In In Nachamkin I, Szymanski C, B. M. e., editor, *Campylobacter, Third Edition*, chapter 9, pages 163–189. ASM Press, Washington, DC., third edition. (doi:10.1128/9781555815554.ch9).

Pearson, A. D., Greenwood, M., Healing, T. D., Rollins, D., Shahamat, M., Donaldson, J., and Colwell, R. R. (1993). Colonization of broiler chickens by waterborne *Campylobacter jejuni*. *Applied and Environmental Microbiology*, 59(4):987–996.

Pearson, B. M., Gaskin, D. J. H., Segers, R. P. A. M., Wells, J. M., Nuijten, P. J. M., and Van Vliet, A. H. M. (2007). The complete genome sequence of *Campylobacter jejuni* strain 81116 (NCTC11828). *Journal of Bacteriology*. (doi:10.1128/JB.01404-07).

Reperant, L. A. and M E Osterhaus, A. D. (2013). The Human-Animal Interface. *Microbiology spectrum*. (doi:10.1128/microbiolspec.OH-0013-2012).

Rind, E. and Pearce, J. (2010). The spatial distribution of campylobacteriosis in New Zealand, 1997-2005. *Epidemiology and Infection*, 138(10):1359–1371. (doi:10.1017/S095026881000018X).

Robert, C. P. and Casella, G. (2004). *Monte Carlo Statistical Methods*. Springer, 2nd ed edition. (doi:10.2307/1270959).

Ruiz-Palacios, G. M. (2007). The health burden of *Campylobacter* infection and the impact of antimicrobial resistance: Playing chicken. *Clinical Infectious Diseases*, 44(5):701–703.

Ryan, K. J., Ray, C. G., and Sherris, J. C. (2004). Vibrio, Campylobacter, and Helicobacter. In *Sherris medical microbiology: an introduction to infectious diseases*, chapter 22, page 379. McGraw-Hill, New York, four edition.

Sari Kovats, R., Edwards, S. J., Charron, D., Cowden, J., D'Souza, R. M., Ebi, K. L., Gauci, C., Gerner-Smidt, P., Hajat, S., Hales, S., Hernández Pezzi, G., Kriz, B., Kutsar, K., McKeown, P., Mellou, K., Menne, B., O'Brien, S., Van Pelt, W., and Schmid, H. (2005a). Climate variability and *Campylobacter* infection: An international study. *International Journal of Biometeorology*, 49(4):207–214.

Sari Kovats, R., Edwards, S. J., Charron, D., Cowden, J., D'Souza, R. M., Ebi, K. L., Gauci, C., Gerner-Smidt, P., Hajat, S., Hales, S., Hernández Pezzi, G., Kriz, B., Kutsar, K., McKeown, P., Mellou, K., Menne, B., O'Brien, S., Van Pelt, W., and Schmid, H. (2005b). Climate variability and campylobacter infection: An international study.

International Journal of Biometeorology, 49(4):207–214. (doi:10.1007/s00484-004-0241-3).

Scott, W., Scott, H., Lake, R., and MG, B. (2000). Economic cost to New Zealand of foodborne infectious disease. *The New Zealand medical journal*, 113(1113):281.

Sears, A., Baker, M. G., Wilson, N., Marshall, J., Muellner, P., Campbell, D. M., Lake, R. J., and French, N. P. (2011). Marked campylobacteriosis decline after interventions aimed at poultry, new zealand. *Emerg Infect Dis*, 17(6):1007–15. (doi:10.3201/eid/1706.101272).

Sheppard, S. K., Dallas, J. F., Strachan, N. J., MacRae, M., McCarthy, N. D., Wilson, D. J., Gormley, F. J., Falush, D., Ogden, L. D., Maiden, M. C., and Forbes, K. J. (2009). *Campylobacter* genotyping to determine the source of human infection. *Clinical Infectious Diseases*, 48(8):1072–1078.

Shrestha, R. D., Midwinter, A. C., Marshall, J. C., Collins-emerson, J. M., and Pleydell, E. J. (2019). Campylobacter jejuni Strains Associated with Wild Birds and Those Causing Human Disease in Six High-Use Recreational Waterways in New Zealand. *Applied and Environmental Microbiology*, 85(24):1–15. (doi:10.1128/AEM.01228-19).

Sopwith, W., Birtles, A., Matthews, M., Fox, A., Gee, S., Painter, M., Regan, M., Syed, Q., and Bolton, E. (2008). Identification of potential environmentally adapted *Campylobacter jejuni* strain, United Kingdom. *Emerging Infectious Diseases*.

Spencer, S. E., Marshall, J., Pirie, R., Campbell, D., Baker, M. G., and French, N. P. (2012). The spatial and temporal determinants of campylobacteriosis notifications in new zealand, 2001-2007. *Epidemiol Infect*, 140(9):1663–77. (doi:10.1017/S0950268811002159).

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and Van der Linde, A. (2014). The deviance information criterion: 12 years on. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*. (doi:10.1111/rssb.12062).

Statistics New Zealand. Data from: 2006 census data meshblock dataset. Accessed 15 March 2018. (http://archive.stats.govt.nz/Census/2006-census/meshblock-dataset.aspx).

Statistics New Zealand. Data from: 2013 census data meshblock dataset. Accessed 15 March 2018. (http://archive.stats.govt.nz/Census/2013-census/data-tables/meshblock-dataset.aspx).

Strachan, N., Rotariu, O., Smith-Palmer, A., Cowden, J., Sheppard, S., O'Brien, S., Maiden, M., Macrae, M., Bessell, P., Matthews, L., Reid, S., Innocent, G., Ogden, I.,

and Forbes, K. (2013). Identifying the seasonal origins of human campylobacteriosis. *Epidemiology and Infection*, 141(6):1267–1275.

Strachan, N. J., Gormley, F. J., Rotariu, O., Ogden, I. D., Miller, G., Dunn, G. M., Sheppard, S. K., Dallas, J. F., Reid, T. M., Howie, H., Maiden, M. C., and Forbes, K. J. (2009). Attribution of campylobacter infections in northeast scotland to specific sources by use of multilocus sequence typing. *Journal of Infectious Diseases*, 199(8):1205–1208.

Taylor, L. H., Latham, S. M., and Woolhouse, M. E. (2001). Risk factors for human disease emergence. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 356(1411):983–989. (doi:10.1098/rstb.2001.0888).

Thomas, K. M., de Glanville, W. A., Barker, G. C., Benshop, J., Buza, J. J., Cleaveland, S., Davis, M. A., French, N. P., Mmbaga, B. T., Prinsen, G., Swai, E. S., Zadoks, R. N., and Crump, J. A. (2019). Prevalence of *Campylobacter* and *Salmonella* in African food animals and meat: A systematic review and meta-analysis. *International Journal of Food Microbiology*. (doi:10.1016/j.cellsig.2019.109410).

Urwin, R. and Maiden, M. C. (2003). Multi-locus sequence typing: a tool for global epidemiology. *Trends Microbiol*, 11(10):479–87.

van Doorn, H. R. (2014). Emerging infectious diseases. *Medicine (Abingdon, England : UK ed.)*, 42(1):60–63. (doi:10.1016/j.mpmed.2013.10.014).

van Pelt, W., van de Giessen, A., van Leeuwen, J., Wannet, W., Henken, A., Evers, E., de Wit, M., and van Duynhoven, Y. (1999). Oorsprong, omvang en kosten van humane salmonellose. deel 1. oorsprong van humane salmonellose met betrekking tot varken, rund, kip, ei en overige bronnen. *Infectieziekten Bulletin*, 10(12):240–3.

Waage, A. S., Vardund, T., Lund, V., and Kapperud, G. (1999). Detection of small numbers of *Campylobacter jejuni* and *Campylobacter coli* cells in environmental water, sewage, and food samples by a seminested PCR assay. *Applied and Environmental Microbiology*, 65(4):1636–1643.

Wagenaar, J. A., French, N. P., and Havelaar, A. H. (2013). Preventing *Campylobacter* at the source: why is it so difficult? *Clin Infect Dis*, 57(11):1600–6. (doi:10.1093/cid/cit555).

Wilson, D. J., Gabriel, E., Leatherbarrow, A. J., Cheesbrough, J., Gee, S., Bolton, E., Fox, A., Fearnhead, P., Hart, C. A., and Diggle, P. J. (2008). Tracing the source of campylobacteriosis. *PLoS Genetics*, 4(9):e1000203. (doi:10.1371/journal.pgen.1000203).

Wilson, D. J., Gabriel, E., Leatherbarrow, A. J., Cheesbrough, J., Gee, S., Bolton, E., Fox, A., Hart, C. A., Diggle, P. J., and Fearnhead, P. (2009). Rapid evolution and the importance of recombination to the gastroenteric pathogen *Campylobacter jejuni*. *Molecular Biology and Evolution*, 26(2):385–397. (doi:10.1093/molbev/msn264).

Woolhouse, M. E. and Gowtage-Sequeria, S. (2005). Host range and emerging and reemerging pathogens. *Emerging Infectious Diseases*, 11(12):1842–1847. (doi:10.3201/eid1112.050997).

World Health Organization (2013). Mortality and global health estimates. , Geneva, Switzerland: World Health Organization.

World Health Organization, Food and Agriculture Organization of the United Nations and World Organisation for Animal Health (2013). The global view of campylobacteriosis: report of an expert consultation, utrecht, netherlands, 9-11 july 2012. , World Health Organization.

Human Ethics Notification - 4000015018

humanethics@massey.ac.nz 14:43 (31 minutes ago), 6 Oct.

to Sih-Jing.Liao.1, J.C.marshall, M.Hazelton, N.P.French

HoU Review Group

Ethics Notification Number: 4000015018
Title: Statistical modelling for the analysis of human Campylobacter disease from Manawatu

Thank you for your notification which you have assessed as Low Risk.

Your project has been recorded in our system which is reported in the Annual Report of the Massey University Human Ethics Committee.

The low risk notification for this project is valid for a maximum of three years.

If situations subsequently occur which cause you to reconsider your ethical analysis, please log on to http://rims.massey.ac.nz and register the changes in order that they be assessed as safe to proceed.

Please note that travel undertaken by students must be approved by the supervisor and the relevant Pro Vice-Chancellor and be in accordance with the Policy and Procedures for Course-Related Student Travel Overseas. In addition, the supervisor must advise the University's Insurance Officer.

A reminder to include the following statement on all public documents:

"This project has been evaluated by peer review and judged to be low risk. Consequently it has not been reviewed by one of the University's Human Ethics Committees. The researcher(s) named in this document are responsible for the ethical conduct of this research.
If you have any concerns about the conduct of this research that you want to raise with someone other than the researcher(s), please contact Dr Brian Finch, Director (Research Ethics), email humanethics@massey.ac.nz. "

Please note that if a sponsoring organisation, funding authority or a journal in which you wish to publish require evidence of committee approval (with an approval number), you will have to complete the application form again answering yes to the publication question to provide more information to go

before one of the University's Human Ethics Committees. You should also note that such an approval can only be provided prior to the commencement of the research.

You are reminded that staff researchers and supervisors are fully responsible for ensuring that the information in the low risk notification has met the requirements and guidelines for submission of a low risk notification.

If you wish to print an official copy of this letter, please login to the RIMS system, and under the Reporting section, View Reports you will find a link to run the LR Report.

Yours sincerely

Dr Brian Finch
Chair, Human Ethics Chairs' Committee and
Director (Research Ethics)

MASSEY UNIVERSITY
TE KUNENGA KI PŪREHUROA
UNIVERSITY OF NEW ZEALAND

Human Ethics Application

Application ID :	4000015018
Application Title :	Statistical modelling for the analysis of human Campylobacter disease from Manawatu
Date of Submission :	06/10/2015
Primary Investigator :	Miss Shi-Jing Liao
Other Investigators :	Prof Martin Hazelton
	Prof Nigel French
	Dr Jonathan Marshall

1. Risk Assessment

Project Detail

1 **Application Title**
*Maximum 2000 characters**

Statistical modelling for the analysis of human Campylobacter disease from Manawatu

2 **Recruitment / Data collection start date.**
*This date must be in the future.**

02/11/2015

3 **Projected end of project date.***

02/11/2017

4 **Project Type ***

- ○ Academic Staff Research
- ○ General Staff Research
- ◉ Postgraduate Student Research
- ○ Undergraduate Student Research
- ○ Evaluation
- ○ Teaching
- ○ Other

5 **Project Summary**
*Please outline in no more than 2000 characters in lay language**

The developments in statistical modelling in epidemiology have focused greatly the attentions on spatial-temporal data in recent years. However, to incorporate the genetic information with epidemiological data will allow us to capture biological complications in the disease models, despite the price of increased model complexity and hence difficulties in conducting inference. The purpose of this project is to build more comprehensive models which can combine the spatial-temporal data with multilocus sequence typing data and to apply these models to attribute human Campylobacter disease in Manawatu. To combat some inevitable problems with modelling fitting, a flexible technique, approximate Bayesian computation method, will be examined. By applying this approach to any model, it can provide a possible means of doing inference for models with complex distributions. This study should result in an improvement in the Campylobacter disease monitoring and allow the decision-maker to implement the early intervention and prevention to this disease and to better explore the transmission between human and sources.

6 **Describe the peer review process that has been used to discuss and analyse the ethical issues present in this project.***

To assess any ethical issues that may raise in this study, my supervisors and I have met with Dr. Jackie Benschop, a senior staff at mEpiLab. By combining the spatial (location) data with genetic information, the individuals could become potentially identifiable to the researchers. I have read through the Code of Ethical Conduct for Research, Teaching and Evaluations Involving Human Participants and the Scope of HDEC Review on the Massey University Human Ethics Committee website. This study is at low risk and fits the exclusion of "use or disclosure authorised" (section 27.3, SOPs for HDECs digest) since the inform consent from individuals will be received and the anonymous notified cases will be collected.

and met ,

7 **List the ethical issues considered and explain how they have been addressed***

To circumvent the potential issue that the individuals may become identifiable to the researchers, the inform consent from all individuals will be received.
All collected data will be pre-filtered and the individuals will be anonymous such that the information will not be disclosed to the researchers.

8 **With whom did you peer review your research?***

Dr. Jackie Benschop from mEpiLab, IVABS

Applicant

1 Department Hidden*

Institute of Fundamental Sciences

2 Ethics Category Hidden*

Human

Campus of Chief Applicant
*(or Campus of Supervisor for Student)**

- (•) Manawatu
- ( ) Wellington
- ( ) Albany

3 Personnel
You can add any additional team members here. Click on 'More criteria' below to access the advanced search function.

1	Surname	Hazelton
	Given Name	Martin
	Full Name	Prof Martin Hazelton
	Position	Co-Applicant
	Primary?	No
	Work Number	
	Email Address	M.Hazelton@massey.ac.nz
	Department	121
	College	50
2	Surname	French
	Given Name	Nigel
	Full Name	Prof Nigel French
	Position	Co-Applicant
	Primary?	No
	Work Number	
	Email Address	N.P.French@massey.ac.nz
	Department	126
	College	50
3	Surname	Liao
	Given Name	Sih-Jing
	Full Name	Miss Shi-Jing Liao
	Position	Chief Applicant
	Primary?	Yes
	Work Number	
	Email Address	S.J.Liao@massey.ac.nz
	Department	121
	College	50
4	Surname	Marshall
	Given Name	Jonathan
	Full Name	Dr Jonathan Marshall
	Position	Co-Applicant
	Primary?	No
	Work Number	
	Email Address	J.C.marshall@massey.ac.nz
	Department	121
	College	50

Please add name of co researchers if unable to locate above

This question is not answered.

Risk Assessment

Health and Disability Research

1 **Is Health and Disability Ethics Committee review required for this study?***

- (•) No
- ( ) Yes

Link to Scope

Link HDEC Flowchart

2 **Does your research include:**

a **Situations where the researcher may be at risk of harm***

- (•) No
- ( ) Yes

b **Use of a questionnaire or interview, whether or not it is anonymous, which might reasonably be expected to cause discomfort, embarrassment or psychological or spiritual harm to the participants. ***

- (•) No
- ( ) Yes

c **Processes that are potentially disadvantageous to a person or group, such as the collection of information which may expose a person / group to discrimination.***

- (•) No
- ( ) Yes

d **Collection of information of illegal behavior(s) gained during the research which could place the participants at risk of criminal or civil liability or be damaging to their financial standing, employability, professional or personal relationships.***

- (•) No
- ( ) Yes

e **Collection of blood, body fluid, tissue samples or other samples.***

- (•) No
- ( ) Yes

f **Any form of exercise regime, physical examination, deprivation.***

- (•) No
- ( ) Yes

g **The administration of any form of drug, medicine (other than in the course of standard medical procedure), or placebo. ***

- [x] No
- [ ] Yes

h **Physical pain, beyond mild discomfort.***

- [x] No
- [ ] Yes

i **Any Massey University teaching which involves the participation of Massey University students for a demonstration of procedures or phenomena which have potential for harm.***

- [x] No
- [ ] Yes

j **Participants whose identities are known to the researcher giving oral consent rather than written consent, other than for cultural reasons .***

- [x] No
- [ ] Yes

k **Participants who are unable to give informed consent.***

- [x] No
- [ ] Yes

l **Research on your own students / pupils.***

- [x] No
- [ ] Yes

m **The participation of children (seven (7) years old or younger).***

- [x] No
- [ ] Yes

n **The participation of children under sixteen (16) years old where active parental consent is not being sought.***

- [x] No
- [ ] Yes

o **Participants who are in a dependant situation, such as nursing home or prison, or patients highly dependent on medical care.***

- [x] No
- [ ] Yes

p **Participants who are vulnerable.***

- [x] No
- [ ] Yes

q **The use of previously collected identifiable personal information or research data for which there was no explicit consent for this research.***

- [x] No
- [ ] Yes

r **The use of previously collected biological samples for which there was no explicit consent for this research.***

- [x] No
- [ ] Yes

s **Any evaluation of organisational services or practices where information of a personal nature may be collected and where participants or the organisation may be identified.***

- [x] No
- [ ] Yes

t **Deception of the participants, including concealment or covert observations.***

- (•) No
- ( ) Yes

u **Conflict of interest situation for the researcher.**
*e.g. Is the project funded or supported in any way that might result in a conflict of interest, do any of the researchers have a financial interest in the outcome, or is there a professional or other relationship between the researcher and the participants?**

- (•) No
- ( ) Yes

v **Payments or other financial inducements (other than reasonable reimbursement of travel expenses or time) to participants.***

- (•) No
- ( ) Yes

w **A requirement by an outside organisation (e.g. a funding organisation or a journal in which you wish to publish) for Massey University Human Ethics Committee approval.***

- (•) No
- ( ) Yes

x **I wish to submit a full application for Training / Education purposes***

- (•) No
- ( ) Yes

2. Sign off

Applicant Sign Off

To submit this application please select the check box below, then using the Actions Tab click on the Submit action.

As Chief Applicant I have read the Code of Ethical Conduct for Research, Teaching and Evaluation involving Human Participants. If there are co-researchers I have confirmed that they have read the Code and I have obtained their approval for the content of this application. I/We understand my/our obligations and the rights of the participants. I/We agree to undertake the research as set out in the Code of Ethical Conduct for Research, Teaching and Evaluation involving Human Participants. The information contained in this application is to the very best of my / our knowledge accurate and not misleading.*

- ☑ I agree

MASSEY UNIVERSITY
GRADUATE RESEARCH SCHOOL

STATEMENT OF CONTRIBUTION
DOCTORATE WITH PUBLICATIONS/MANUSCRIPTS

We, the candidate and the candidate's Primary Supervisor, certify that all co-authors have consented to their work being included in the thesis and they have accepted the candidate's contribution as indicated below in the *Statement of Originality*.

Name of candidate:	Sih-Jing Liao
Name/title of Primary Supervisor:	Martin Hazelton
Name of Research Output and full reference:	
Liao, S. J., Marshall, J., Hazelton, M. L., & French, N. P. (2019). Extending statistical models for source attribution of zoonotic diseases: a study of campylobacteriosis. Journal of the Royal Society Interface, 16(150), 20180534	
In which Chapter is the Manuscript /Published work:	chapter 4
Please indicate:	
• The percentage of the manuscript/Published Work that was contributed by the candidate:	75%
and	
• Describe the contribution that the candidate has made to the Manuscript/Published Work:	
Sih-Jing Liao conceived of the modelling with supports from supervisors, carried out the statistical analyses and drafted the manuscript with inputs from supervisors.	
For manuscripts intended for publication please indicate target journal:	
Candidate's Signature:	Sih-Jing Liao 數位簽署者：Sih-Jing Liao 日期：2020.03.31 18:53:33 +13'00'
Date:	31.03.2020
Primary Supervisor's Signature:	Martin Hazelton Digitally signed by Martin Hazelton DN: cn=Martin Hazelton, c=NZ, o=University of Otago, ou=Department of Mathematics and Statistics, email=martin.hazelton@otago.ac.nz Date: 2020.03.31 19:09:50 +13'00'
Date:	31/3/2020

(This form should appear at the end of each thesis chapter/section/appendix submitted as a manuscript/ publication or collected as an appendix at the end of the thesis)